

UNIVERSITA' DEGLI STUDI DI PARMA

Dottorato di ricerca  
in  
Progettazione e sintesi di composti biologicamente attivi

Ciclo XXVI

The challenging world  
of *in silico* drug design:  
tools development and applications

Coordinatore:  
Chiar.mo Prof. Marco Mor

Tutor:  
Chiar.mo Prof. Gabriele Costantino

Dottoranda: Claudia Beato



# ABSTRACT

In the attempt to reduce time and costs of the drug discovery process, computational strategies have been looked as the possible solution. Despite still far from being the answer of all the problems, the use of computational approaches is now well established in the drug discovery pipeline. In the course of the years a lot of techniques have been developed and now are applied to several phases of drug discovery and development. In the present thesis is summarized the work conducted in three different projects carried out during my PhD, that allowed me to exploit different computational strategies.

The goal of the main project was the optimization and the validation of the performance of a new drug discovery software, LiGen, result of the collaboration between the Italian pharmaceutical company Dompé, the Italian supercomputing center CINECA and our research group. LiGen was developed to perform protein surface analysis, molecular docking and de novo design; during this project we focused our attention mainly on the first two tools, LiGenPocket, aimed at the binding site analysis and structure-based pharmacophore definition, and LiGenDock, the molecular docking engine. Even if seldom used in computational chemistry, we decided to apply the Design of Experiments (DoE) methodology to optimize parameters controlling LiGenPocket and LiGenDock. At first we applied a fractional factorial design to screen the set of user-adjustable parameters to identify those having the largest influence on the accuracy of the results and then we optimize their values, to ensure the best performance in pose prediction and in virtual screening. Afterwards the results have been also compared with those obtained by two popular docking programs, namely Glide and AutoDock, for pose prediction, Glide and DOCK6 for Virtual Screening.

The second project was the investigation of the binding mode of a series of compounds based on the 2-aminonicotinic 1-oxide scaffold and developed by our synthetic laboratory, to inhibit the 3-hydroxyanthranilic acid dioxygenase (3-HAO), an enzyme of the kynurenine pathway. 3-HAO is

responsible for the production of the neurotoxic tryptophan metabolite quinolinic acid (QUIN); elevated brain levels of QUIN has been connected to several neurodegenerative diseases therefore 3-HAO inhibition may be a useful strategy for Huntington's diseases and Alzheimer's diseases among the others. To predict the most probable binding mode, compounds and the binding site have been characterized at quantum mechanical level, due to the presence of a catalytic iron atom in the binding site. Molecular docking was then used to predict the binding mode of the compounds and to investigate the effects of the substituents at the pyridine ring.

The third project was related to the creation of a database of functional groups to screen chemical libraries, in order to reject or to flag these functionalities in libraries used for virtual screening purposes, in relation to their potential toxicity. The functional groups have been collected from different sources and have been classified according to the type of risk they may be related. This collection of compounds has been enriched also with compounds that have been identified as "frequent hitters", indicating compounds often interfering in vitro assays, especially in HTS. The final database is therefore divided in three group, one collecting the intrinsically reactive moieties, one with functional groups susceptible to biotransformation into reactive metabolites and one containing substructures frequently identified as false positives in experimental tests.

# LIST OF PAPERS

Bruno, A.; Beato, C.; Costantino, G., Molecular dynamics simulations and docking studies on 3D models of the heterodimeric and homodimeric 5-HT(2A) receptor subtype. *Future Medicinal Chemistry* **2011**, *3*, 665-681

Beccari, A. R.; Cavazzoni, C.; Beato, C.; Costantino, G., LiGen: A high performance workflow for chemistry driven de novo design. *J. Chem. Inf. Model.* **2013**, *53*, 1518-1527.

Beato, C.; Beccari, A. R.; Cavazzoni, C.; Lorenzi, S.; Costantino, G., Use of experimental design to optimize docking performance: The case of LiGenDock, the docking module of ligen, a new de novo design program. *J. Chem. Inf. Model.* **2013**, *53*, 1503-1517

Vallerini, G. P.; Amori, L.; Beato, C.; Tararina, M.; Wang, X. D.; Schwarcz, R.; Costantino, G., 2-Aminonicotinic acid 1-oxides are chemically stable inhibitors of quinolinic acid synthesis in the mammalian brain: A step toward new antiexcitotoxic agents. *J. Med. Chem.* **2013**, *56*, 9482-9495



# CONTENTS

<b>Abstract</b>	i
<b>List of Papers</b>	iii
<b>Preface</b> .....	1
<b>CHAPTER 1: Development and optimization of LiGen, a new drug design software</b>	
1.1 Introduction.....	5
1.1.1 Molecular Docking.....	6
1.1.2 Virtual Screening.....	10
1.1.3 <i>De novo</i> Drug Design.....	11
1.1.4 Benchmarking.....	12
1.2 LiGen.....	17
1.2.1 LiGenPASS.....	18
1.2.2 LiGenPocket.....	19
1.2.3 LiGenDock.....	22
1.2.4 LiGenBuilder .....	24
1.3 Aims.....	26
1.4 Materials and Methods.....	28
1.4.1 Dataset /Benchmark composition .....	28
1.4.2 Protein Preparation.....	30
1.4.3 Ligand Preparation.....	30
1.4.4 Experimental Designs.....	31
1.4.5 Screening and Optimization Workflow .....	35
1.4.6 Docking with Glide and AutoDock.....	37
1.4.7 Evaluation Of Self Docking And Virtual Screening Results.....	37
1.5 Results and Discussion.....	39
1.5.1 Pose Prediction Optimization.....	39
1.5.2 Virtual screening Optimization .....	41
1.5.3 Pose Prediction Validation.....	55
1.5.4 Virtual Screening Validation .....	58
1.6 Conclusions.....	60
1.7 Bibliography .....	63

## **CHAPTER 2: Elucidation of the binding mode of a series of 3-HAO inhibitors**

2.1	Introduction.....	71
2.1.1	Kynurenine pathway.....	71
2.1.2	Targeting the kynurenine pathway in the brain.....	77
2.1.3	3-HAO.....	79
2.1.3.1	Crystal structures.....	80
2.1.3.2	Binding Site Analysis .....	85
2.1.3.3	Proposed Mechanism of action and Inhibition .....	87
2.2	Aims.....	90
2.3	Materials and Methods.....	91
2.3.1	Compounds .....	91
2.3.2	Docking to Metalloproteins.....	94
2.3.3	Workflow.....	95
2.4	Results and Discussion.....	97
2.5	Conclusions.....	107
2.6	Bibliography.....	109

## **CHAPTER 3: Creation of a new database of Structural Alerts**

3.1	Introduction.....	119
3.2	Aims.....	124
3.3	Materials and Methods.....	125
3.3.1	Source of Information.....	125
3.3.2	Structural Alerts definitions using SMARTS...	125
3.3.3	Database Structure.....	127
3.4	Results and Discussion.....	129
3.4.1	Rank3.....	129
3.4.2	Rank2.....	132
3.4.3	Rank1.....	138
3.4.4	Profiling of Best Selling Drugs.....	140
3.5	Conclusions.....	142
3.6	Bibliography.....	143

## **APPENDIX A**

## **APPENDIX B**



# PREFACE

Computational approaches are commonly used in drug discovery projects to assist the development of new bioactive molecules. Structure based and ligand based techniques have been used for a long time to help the discovery of new hit or lead compounds, through rational design or virtual screening. Nowadays the field of application of these approaches has definitely widened, flanking several other classical approaches of the drug discovery pipeline. Bioinformatics, that in the post-genomic era became an essential tool to analyze DNA and protein sequences, or computational toxicology, triggered by the discovery that ADMET profile was the main cause of drug failures, are just two of many possible examples.

In this context, during my PhD I applied computational techniques in different projects, that are reflected by the main structure of the thesis, divided in three main chapters, one for each project. Every chapter starts with a brief introduction, introducing the reader into the topic of the project, followed by a declaration of the aims of the study. Afterwards the experimental section illustrates the main features of the adopted techniques; then the results achieved during the project are explained in the results section and summarized in the conclusions at the end of the chapter. The main project, presented here in *Chapter 1*, was the optimization and validation of a new docking program included in LiGen, a new drug discovery software, created by the collaboration of the pharmaceutical industry Dompé, the Italian supercomputing center CINECA and our research group. *Chapter 2* is about the work done in the attempt to give a rational explanation of the activity of a series of 3-hydroxyanthranilic acid dioxygenase inhibitors developed by the synthetic chemists of our research group. The last chapter, *Chapter 3*, is dedicated to the project I carried on during the five months I spent at prof. Abagyan laboratory at University of California, San Diego, and concerns the building of a new database of filtering rules for chemical libraries.



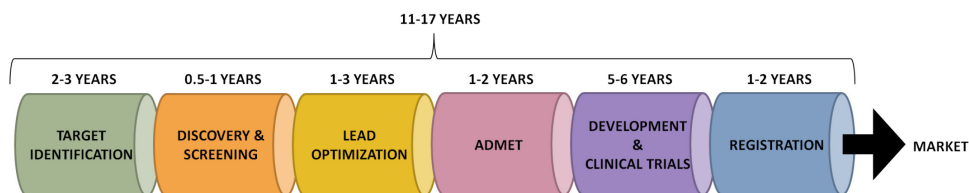
# **CHAPTER 1:**

## **Development and optimization of LiGen, a new drug design software**

CHAPTER 1 Development and optimization of LiGen, a new drug design software

## 1.1 INTRODUCTION

The development of a new drug, from the identification of a new active compound to the final marketed product, is a multi-step process that can be roughly split into pre-clinical development and clinical trials (see Figure 1). The main goal pursued during the pre-clinical development is to optimize a newly discovered compound to improve its biological activity towards the pharmacologically relevant target and its safety. This requires the investigation of the pharmacodynamic and pharmacokinetic properties. But first of all a new chemical entity should be identified and the starting point is always more frequently an *in silico* study. In fact, starting from the beginning of the nineties, in the attempt to control the rise of times and costs, the use of computational techniques flanked the classical experimental research. Drug development is a long and expensive process. The average time from synthesis of a new chemical entity (NCE) to approval of a new drug application (NDA) has increased significantly, from an average 7.9 years in the 1960s to 12.8 years in the 1990s.<sup>1</sup> Likewise the costs for the discovery and development have grown, and now the estimated cost to bring a NCE to market is approximately \$1.2-1.8 USD. Therefore computational approaches are more and more often integrated into drug discovery programs, aiming at accelerating hit identification and lead optimization.<sup>2</sup> At the beginning mainly ligand-based methods were used but then during the last twenty years the number of publicly available 3D protein structures has greatly increased, surpassing ninety thousand structures in 2013, boosting the development of a plethora of structure-based methods. The most widely used approach is molecular docking, which aims to predict the spatial arrangement assumed by a ligand in protein–ligand complexes, where the protein is a receptor or enzyme, but can also be a nucleic acid molecule.



**Figure 1.** The Drug Discovery Process is schematically represented here. The first step consists in the identification of the target, using *in vitro* target expression, knock-out experiments and bioinformatics; the second phase is dedicated to hit identification for example using combinatorial chemistry, structure based drug design, HTS and other *in vitro* and *in vivo* tests; the following step is the optimization of the hit previously discovered into a lead compound, by means of traditional medicinal chemistry techniques and rational drug design; the fourth step is the assessment of the ADMET properties of the lead compound; after that the clinical testing of the compound, together with the development of large-scale synthesis of the product starts, to conclude in the last step, during which all the documentation collected in the clinical phases is evaluated by FDA to decide if the product can reach or not the market.

Moreover docking is a common tool exploited by many different computational techniques, of which the most frequently used in drug discovery are virtual screening, which aims to identify new active ligands among a chemical library, and *de novo* design.

### 1.1.1 Molecular Docking

*In silico* molecular docking is a valuable tool in drug discovery, attempting to predict the structure of a ligand, usually a small molecule, in complex with its biological target. Molecular docking simulations are used for reproducing *in silico* experimental data on protein-ligand interactions, to find a rational explanation for biological activity of compounds and build structure-activity relationships. From a theoretical perspective, molecular docking searches a global optimum in an energy landscape defined by the scoring function, protein, ligand, and degrees of freedom to be explored.<sup>3</sup> To address this complex problem, all docking programs are divided into two main parts, namely, the search of the ligand disposition, guided by the docking algorithm, and the scoring function, that tries to make a prediction of the interaction energy between the ligand and the target, and thus also an estimation of the biological activity.<sup>4</sup>

Therefore a first challenge is the size of the search space, which grows exponentially with the number of degrees of freedom of the system. Sampling the degrees of freedom is not a trivial task because even relatively small organic molecules can contain many conformational degrees of freedom, and the process must be performed with a certain accuracy to identify the conformation that best matches the receptor active site.<sup>2</sup> The first molecular docking programs treated proteins and ligand molecules as rigid bodies, fixing all the internal degrees of freedom, except for the three translations and three rotations.<sup>5</sup> Within this concept, the only way to address the conformational flexibility of the ligands was to pre-generate a library of ligand conformations that were formally treated as separated molecular entities. Examples are the first versions of FRED<sup>6</sup> and DOCK4.0.<sup>7,8</sup> These approaches were quickly replaced by algorithms able to explicitly handle the conformational degrees of freedom during docking simulations. Several ligand flexibility algorithms have been proposed and can be divided into three main families according to the type of search: approaches derived from a so-called systematic search, stochastic methods and molecular dynamics simulation techniques.<sup>2, 9</sup> Systematic search methods group together those algorithms that try to explore all the degrees of freedom in a molecule, as QXP, that carries out full conformational searches for flexible cyclic and acyclic molecules<sup>10</sup> with extremely good results.<sup>11</sup> The main issue of this kind of approach is the combinatorial explosion of ligand conformations number. Thus some approaches have been developed trying to overcome this problem, such as the combined use with filtering techniques, as in Glide,<sup>12</sup> or the use of ligand fragmentation and subsequent incremental reconstruction within the binding site,<sup>13</sup> as in FlexX,<sup>14</sup> Surflex<sup>15</sup> and eHITS.<sup>16</sup> Stochastic-based algorithms make random changes, generally with a limited physical basis, usually changing one degree of freedom at time;<sup>4</sup> examples are Monte Carlo (MC) methods and evolutionary algorithms applied, for example, by ICM<sup>17</sup> and AutoDock.<sup>18</sup> The most used deterministic approach is molecular dynamic simulation. However, molecular dynamics are time consuming and often a main concern regarding this kind of method is that they are not able to cross high free-energy barriers within accessible (usually short) simulation time; therefore, a possible risk is that the system gets trapped in local minima.<sup>2, 4</sup> To avoid this problem, several attempts have been proposed, such as starting the simulation from different ligand positions or simulate the system at different temperatures; however, these efforts are quite expensive in terms of calculation time, limiting the application of molecular dynamic to docking

only one or few compounds. Irrespective of the way the docking poses are generated, the evaluation, in terms of interaction energy, of the ligand conformations inside the binding site represents a second challenge, addressed by adopting a scoring function that tries to pinpoint the experimental (real) binding mode among all those that have been generated.

Scoring functions are mathematical approximating methods for evaluating binding affinity. Using as input the atomic 3D coordinates of the ligand-target complex, scoring functions give an estimation of the free energy of binding or binding constant.<sup>19</sup> The free energy of binding is obtained with the Gibbs-Helmholtz equation:

$$\Delta G = \Delta H - T\Delta S$$

where  $\Delta G$  is the free energy of binding,  $\Delta H$  is the enthalpy,  $T$  is the temperature in Kelvin and  $\Delta S$  the entropy. The binding constant  $K_i$  is related to  $\Delta G$  by the equation:

$$\Delta G = -RT\ln K_i$$

Scoring functions can be classified into three families: empirical scoring functions, knowledge-based, and force field based scoring functions. Empirical scoring functions are weighted sum of several intermolecular interaction terms, where the weighting factors are calibrated through a linear regression procedure, in which theoretical values are fitted to be closest as possible to experimental data. The different terms reflect the different types of interaction established between ligand and target, such as hydrogen bonds, ionic and van der Waals interactions.<sup>20, 21</sup> Knowledge based scoring functions are based on statistical observations of intermolecular contacts identified from large datasets of experimental 3D structures.<sup>22, 23</sup>

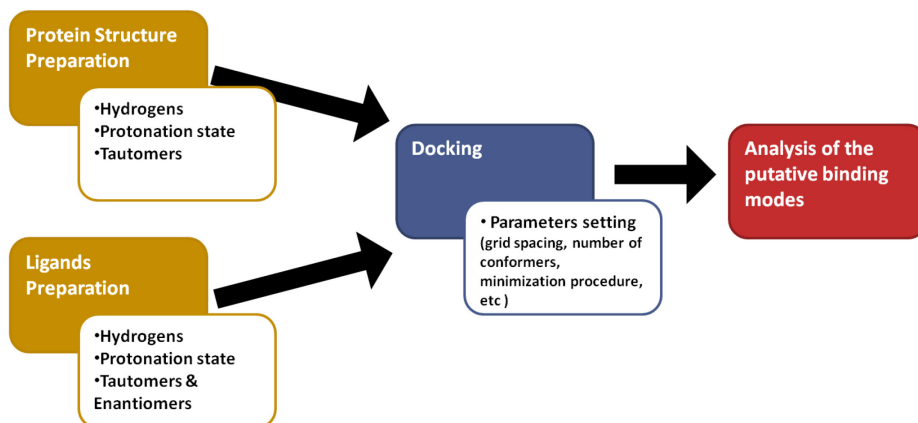
The last family of scoring functions uses non-bonded terms of classical mechanics force fields, summing the interaction energy and the internal energies of both partners, and ideally taking into account the solvent effect. These scoring functions are usually sensitive to atomic coordinates, limiting their applications in cross-docking experiments. Softened van der Waals potentials, in which the contribution of the repulsive term is limited to allow some steric clashes without penalizing too much the corresponding binding mode, have the advantage of being less sensitive to atomic coordinates, but suffer from being less selective.<sup>24</sup>

Unfortunately the perfect scoring function able to accurately estimate the protein-ligand binding energy, does not exist yet, in fact, most docking failures can be attributed to scoring functions.<sup>25</sup> By accurate investigations,



it turned out that accurate prediction of binding affinities, especially for a diverse set of molecules, is genuinely difficult. The problem of the unreliable calculation of  $\Delta G$  arise from the inaccuracies made in the calculation of ligand and protein energies (very big numbers), that are subtracted to give the free energy of binding (usually a small number).<sup>26</sup> Moreover, Tirado-Rives and Jorgensen pointed out that another problem is the small “ window o activity”.<sup>27</sup> In Virtual Screening experiments the free energy difference between the best active compound (around 50 nM) that one might expect to find and the experimental detection limits (100  $\mu$ M) is only about 4.5 kcal/mol. Hence to provide a successful ranking of compounds, more accurate methods may be considered. Some approaches to improve the ranking of the scoring function have been investigated. One method consists in using consensus scoring , i.e. to rank docking results with multiple scoring functions, that has been shown to definitely improve the docking results.<sup>28</sup> Other approaches that aim to produce a more reliable estimation of protein–ligand binding free energies are Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) and the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) methods, which calculate binding free energies by combining molecular mechanics calculations and continuum solvation models. Both of them have been widely exploited, in particular MM/GBSA, in free energy calculations because of their computational efficiency if compared with rigorous methods such as free energy perturbation (FEP) and thermodynamic integration (TI) methods.<sup>29-31</sup> But even MM/GBSA and MM/PBSA are not completely reliable, in fact they are useful mainly to compare the binding energies of close analogues.<sup>26</sup>

A workflow representing the commons steps in molecular docking is represented in Figure 2 . A first essential step is protein preparation. For instance, the presence of crystal contacts in X-ray structures should be verified, as well as the impact of the presence/absence of other interacting partners such as cofactors and metal ions. Moreover, the quality of the structure should be verified at atomic level, for example to analyze the presence of unresolved atoms/residues and steric clashes in the binding region. Finally the assignment of the protonation state is a necessary step, because in PDB crystal structures hydrogen atoms are generally not solved. The latter step is common also to the ligand structures preparation.



**Figure 2. Docking workflow. Schematic representation of the main common steps in a docking experiment.**

## 1.1.2 Virtual screening

Probably the most popular application of molecular docking is virtual screening (VS), which aims to identify new active ligands from a chemical library collecting compounds of unknown activity for the target under investigation. Of course, VS should be used in combination with traditional HTS, as a biological assay is always needed to validate computational methods. However generally in retrospective VS a poor correlation between the accuracy of a binding mode and the enrichment has been noted.<sup>25, 26</sup> This statement is consistent with the fact that the docking problem itself is not resolved yet, and that there are big issues to address in the prediction and recognition of a correct binding mode, as well as in the evaluation of the binding free energy. It is interesting to note that nearly all groups that reported successful examples of VS application,<sup>32-34</sup> performed pre-filtering using two-dimensional similarity methods and shape or drug-like filters to reduce the number of database compounds for the time-consuming steps of flexible docking; they also reported complex scoring procedures and visual analysis, to overcome the limitation of simple molecular docking.

### 1.1.3 *De novo* drug design

Another recurrent method that uses molecular docking is receptor-based *de novo* design, that is aimed at generating novel chemotypes endowed with a particular set of desired properties, generally biological or pharmacological properties.<sup>35</sup> *De novo* design work starts with small molecular fragments, called building blocks, and attempts to find novel drug-like molecules expanding them (the so called “growing” process) by connecting one another directly (“joining”) or through a linker (“linking”).<sup>36</sup> Several *de novo* drug design programs have been developed such as LUDI,<sup>37</sup> LEGEND,<sup>38</sup> LeapFrog,<sup>39</sup> LigBuilder,<sup>40, 41</sup> SPROUT,<sup>42</sup> HOOK,<sup>43</sup> PROLIGANDS,<sup>44-46</sup> and DOGS,<sup>47</sup> which differentiate from each other for how they explore the chemical space, for how they assemble the candidate compounds and for the evaluation of the candidates quality (using a scoring function).

A secondary but widespread application of *de novo* design is lead optimization. In this case, the binding mode of the core structure of the lead has usually already been validated and the scope of the *de novo* approach is to find the best substituents for the scaffold, to increase affinity and/or improve ADME properties.<sup>48</sup> Therefore the number of compounds to generate is reasonably low, and the selection of the fragments is driven by the optimization process and not by the reagents accessibility, as in case of hits discovery.

The main issue that all *de novo* design tools have to deal with, is the feasibility of the newly designed compounds, that should occupy useful chemical space.<sup>49</sup> Furthermore other problems have been reported in recent reviews:<sup>35, 50, 51</sup> i. low structural diversity;<sup>49</sup> ii. low potential for parallel synthesis applications (except when combinatorial chemistry is directly addressed);<sup>52</sup> iii. generally low throughput if compared to docking programs. As a result, these problems have prevented molecular *de novo* design approach to become an established tool in drug design, and, in fact, methods like virtual screening and molecular docking received higher attention, either in terms of routine use or successful examples, despite some fruitful applications having been reported also in case of *de novo* design.<sup>49, 53</sup>

The problem of synthetic accessibility is the most concerning one, determining the usefulness of the results. This matter is heavily related to the library of building blocks and to the algorithm used in the growing process. Two approaches try to address the synthetic feasibility problem: the first one is to use a retrosynthetic analysis of the final ligand, and this is

generally done by the implementation of bond breaking rules. One example is represented by LigBuilder<sup>40, 54</sup> which uses an internal database of building blocks. The second and most popular approach, is the use of a restricted sets of chemical rules, encoding for the most common chemical reactions. Some of the most relevant reaction driven *de novo* software programs with an extended set of chemical reactions are SYNOPSIS<sup>55</sup> and DOGS.<sup>47</sup>

### 1.1.4 Benchmarking

Benchmarking of docking programs is a well established practice.<sup>25, 28, 56, 57</sup> However setting up a fair comparison between different docking software packages is not trivial<sup>58</sup> and in the past there were no standard procedures or guidelines on how the evaluation should be conducted; thus every research group came up using its own slightly different benchmarking protocol, making a genuine comparison of the results almost impossible.<sup>59</sup> Recently several authors published some guidelines on how to compare docking programs.<sup>3, 59-61</sup> and also some dataset prepared ad hoc for software benchmarking were made publicly available,<sup>62-64</sup> encouraging scientist to use them. Of course, to ensure a fair programs comparison, the same number of poses and, in the evaluating the calculation time, also the same number of CPUs for all the softwares should be used. Benchmarking of software should consider the two main applications of molecular docking, the prediction of the binding mode of compounds known to be active at the specific receptor and the identification of new active molecule from a chemical library.

The first application is evaluated using the self-docking, also called cognate docking or re-docking, approach. In this process a ligand is extracted from a crystal structure in which it is bounded to its target protein and the program is challenged to pose the ligand as closely as possible to its experimentally identified structure. It may be argued that cognate re-docking is not a task commonly faced in the normal use of docking tools, since cross-docking (docking of a ligand into a structure with which it was not crystallized) is the actual application of a docking tool.<sup>65</sup> However the use of cross-docking is less common, due to the absence of a robust dataset, collecting different targets and several compounds with experimentally verified binding mode, to be used as a benchmark dataset. The universal measure used in cognate-docking benchmarks is the RMSD (root mean square deviation) between the heavy atoms positions seen in

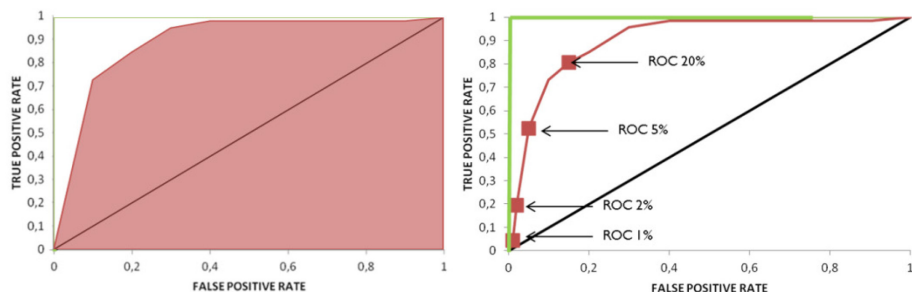
the crystal structure and those predicted with the docking simulation.<sup>65</sup> What changes among different studies is the way it is reported. A regrettably common method is to compare the average RMSDs across a set of structures; however it can be skewed by small number of poorly predicted poses with large RMSD values, leading to possible problems of interpretation of the results.<sup>61</sup> The most accepted method is to report the proportion of successes at a particular threshold of RMSD, that is commonly set to 2 Å, but generally results at different thresholds are also reported.<sup>66, 67</sup> Comparing docking programs for their self-docking ability presents some difficulties, the most evident but often forgot one, is that is meaningless to compute a property with greater precision than the accuracy of the experiment that measured that property. In the specific case of cognate docking this perfectly applies to the attempts of predicting binding poses with greater accuracy than the crystal structure resolution, but unfortunately this is often ignored by scientists involved in benchmarking projects.<sup>3, 61</sup>

The second task to be evaluated in benchmarking docking software is the virtual screening ability, measured in a retrospective examination of the ability to discriminate experimentally known active molecules from inactive compounds in a chemical library. The assessment of VS performances presents more hurdles than cognate docking, in terms of the method itself, datasets and metrics to report results. The first issue is related to the inability of the scoring function to correctly predict the interaction energy, as highlighted before. In fact, in an ideal application of VS, all the active compounds of the chemical library show a more favorable score, i.e. binding energy, compared to the inactive ones, and therefore are ranked as firsts. However, due to the inaccuracy of the scoring functions, the correct ranking based on the compounds affinity is still a far goal.

A second major problem concern the chemical library to be used. The active compounds included in the library should not be all chemically similar, because operationally finding chemically similar molecules as being potentially new active compounds is of little value and in real applications, where ligands that are obvious analogs of existing lead compounds will not be included in libraries to be screened for new possible leads.<sup>68</sup> The choice of decoys, the “inactive” compounds of the dataset, is of extreme importance and greatly influence the results obtained in the software validation.<sup>65</sup> First of all the decoys are supposed to be inactive at the specific target, but generally they have not been experimentally tested to verify the inactivity. In an ideal case the decoys are all experimentally verified inactive

compounds, but in real world it is impossible to retrieve this information; therefore the fact that decoys are not real inactive compounds but only supposed to be not active should be always kept in mind. Secondly, molecules that are completely different from the active ones, bias the evaluation towards better performances than the real ones, whereas compounds too similar to actives can pose a challenge beyond the ability of differentiation of the scoring function. The most widely used dataset for VS assessment is now the Directory of Useful Decoys (DUD), collecting forty targets, with the corresponding sets of active compounds and decoys. In DUD decoys were selected to match the same simple molecular properties of the active compounds, so that the decoys are not trivially separable from the actives, in order to have a more realistic representation of the discriminating ability of the docking method.

Another main issue concerning VS is about the best metric to use in the assessment of screening enrichment. Around this topic there is still a passionate debate and no definitive solution has been given.<sup>3, 59-61, 69, 70</sup> The standard method has been for a long time the enrichment, defined to be the ratio of the observed fraction of active compounds in the top few percent of a virtual screen to that expected by random selection.<sup>59</sup> It is still quite used because it is easy to calculate and promptly understandable, despite it presents some drawbacks, in particular its dependence on the ratio of actives to inactives, which makes enrichment a property of the method and the experimental set-up rather than just an intrinsic property of method.<sup>59</sup> Moreover it gives no weight to where in the ranked list a known active compound appears. Thus to calculate enrichment at 1% in a virtual screen of 10,000 compounds, the number of actives (N) in the top ranked 100 compounds is needed. However the enrichment at 1% is the same whether the N active compounds are ranked at the very top of the list or at the very bottom of the top ranked 100. A metric widely used to determine success in detecting a signal in a background of noise is the receiver operator characteristic (ROC). The ROC curve is derived by plotting noise (fraction of false positives) on the x-axis versus signal (fraction of true positives) on the y-axis. The area under the ROC curve (AUC) is a widely used measure in a variety of fields and in the case of VS shows the performance of a given tool when screening across the entire database is examined, not just at fixed, early points in the screen as enrichment does. The theoretically perfect performance of a virtual screening application gives the maximum area under a ROC curve (1.0), while random performance of a tool gives an AUC of 0.5. Areas under the curve of less than 0.5 imply a systematic ranking of



**Figure 3. Graphic representation of metrics used to evaluate VS results. In red is represented the ROC curve. A) In light red highlighted the area under the ROC curve. B) The red squares along the curve represent the points where the enrichment is evaluated. The straight green line along the vertical axis and on the top of the graph in both A) and B) represents the ideal ROC curve, whereas the black one at 45° represents the random performance.**

decoys higher than known actives. The AUC assesses virtual screening performance across the entire database, known as “global enrichment”, but many practitioners of virtual screening are, rightly, most concerned about early performance of the tools they use, since active compounds are supposed to be ranked better than decoys; moreover the common size of screened libraries is of some thousands to millions molecules, making impossible to visually inspect all the results. This is one reason why enrichment is still commonly used to measure success. The metric of early performance based on the ROC curve is the true positive rate at fixed false positive rates. The true positive rate at a false positive rate of, for example, 1% is a much more robust measure than the enrichment at 1% and provides similar information about the early performance of a tool.<sup>59, 69</sup>

Another important aspect is the use of an appropriate dataset, in terms of proteins and ligands. This is crucial not only for VS benchmark but also for cognate docking. The ideal test set should not be biased toward a given protein family. Instead, it should be large enough to span representative high-resolution complexes, ensure the absence of errors such as steric clashes or crystal contacts; also complexes with missing residues or covalent bonds should be avoided. Binding data (such as  $K_D$  or  $IC_{50}$ ) should be available for each complex, whenever it is possible. CCDC/Astex<sup>63</sup>, PDBbind<sup>64</sup> and DUD<sup>62</sup> are the most popular databases published with the intent of being used as standards for software benchmarking, the first two for molecular docking evaluation, the latter for VS. However even using these well-known datasets some differences in results occur, depending, for example, on the different procedures applied to prepare the dataset for

CHAPTER 1 Development and optimization of LiGen, a new drug design software

molecular docking, which were demonstrated to have a great impact on the results, as pointed out by Corbeil et al..<sup>71</sup>



## 1.2 LiGen

LiGen (Ligand GENERator) is a suite of drug discovery programs, developed by Dompé, CINECA and our research group at University of Parma.<sup>72</sup> Its main applications concern molecular docking and *de novo* design. LiGen consists of a set of tools which can be combined in a user-defined manner to generate project-centric workflows (Figure 3). In a standard application, the modules work sequentially, from the generation of the input constraints (either structure-based, through active site identification, or ligand-based, through pharmacophore definition), to

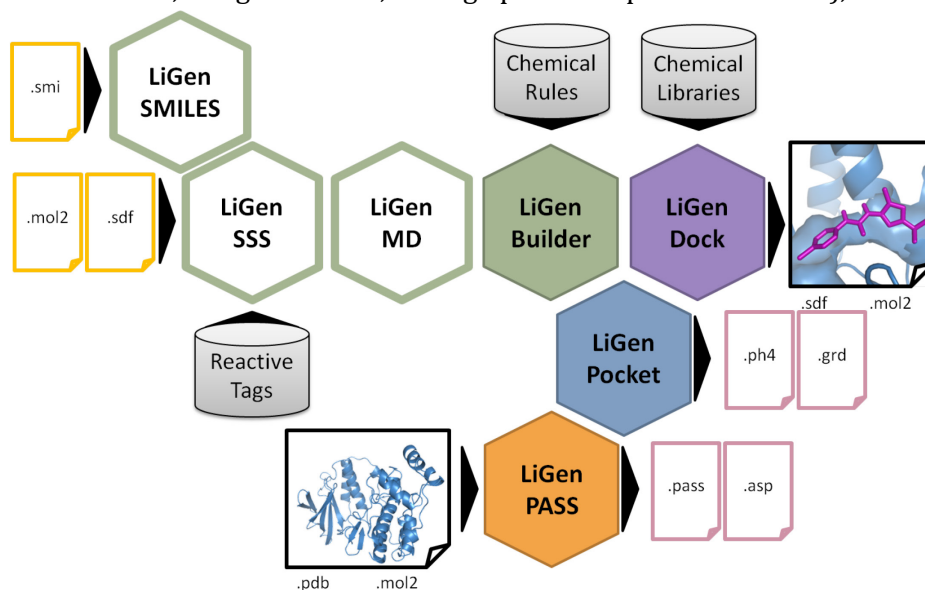
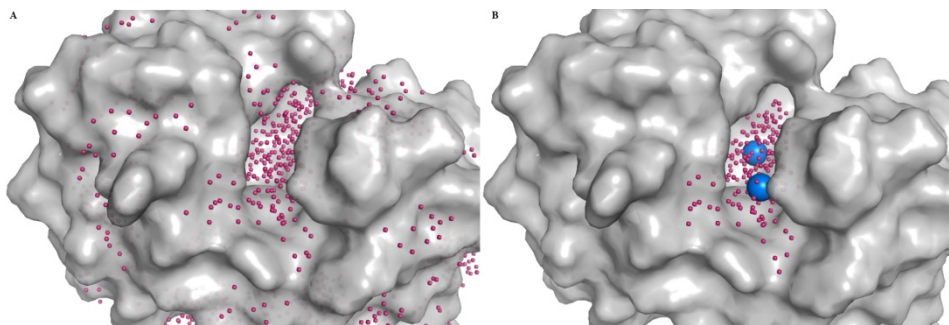


Figure 4. Representation of LiGen modules. LiGenSMILES reads files containing SMILES and generates 3D molecular structure of the molecules or of the fragments. LiGenSSS is the tool that tags the reactive groups of molecules, previously generated by LiGenSMILES or directly read from a mol2 or sdf file. LiGenMD minimizes the molecules or fragments. LiGenPASS analyzes the protein surface to find possible binding sites. LiGenPocket analyzes the binding sites building a grid of the pocket and a pharmacophore reflecting the pocket characteristics (hydrogen bond donor, acceptor and hydrophobic regions). LiGenDock is the tool responsible for molecular docking that uses the pharmacophore to drive the docking process. It is used to dock molecules and by LiGenBuilder to dock fragments, that are further combined according to chemical rules.

docking and *de novo* generation. LiGen main functionalities are LiGenPass, that recognizes possible binding sites on protein surface, LiGenPocket, which is responsible for grid and pharmacophore generation, LiGenDock, the docking engine, and LiGenBuilder, the module in charge of the *de novo* design process. All of them will be described in the following paragraphs.

### 1.2.1 LiGenPass

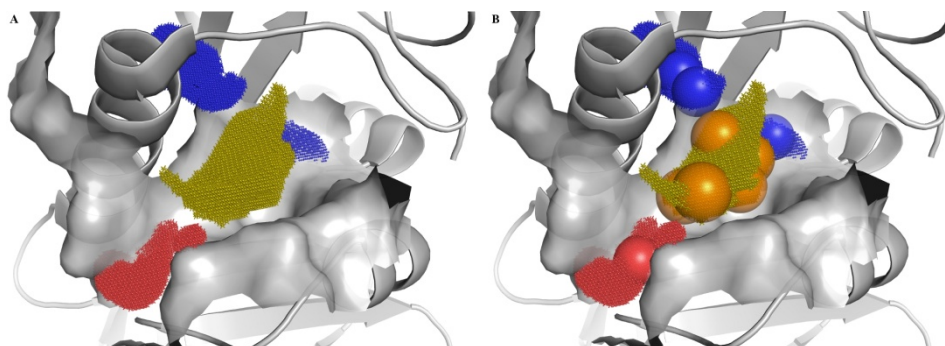
LiGenPass is based upon the algorithm used by PASS (putative active site with spheres), software developed by Brady et al.<sup>73</sup> to identify cavities in target proteins. LiGenPass characterizes regions of buried volume in the target protein and identifies potential binding sites based upon the size, shape, and burial extent of these volumes (Figure 5). Briefly, it analyzes protein surface, and fills cavity with probes; if the cavity contains a number of probes greater than a predetermined threshold, the center of mass of the probes inside the cavity is calculated and is represented by an Active Site Point (ASP). The ASPs are then ranked according to the probability of being a drugable binding site.



**Figure 5 LiGenPASS.** A) Protein surface is analyzed with probes. Every cavity on protein surface is filled with probes but only those cavities big enough to contain a number of probes higher than the threshold value are considered in the end as possible binding sites. B) In figure B it is shown a that the big cavity is recognized as a possible binding site. Actually, given the elongated shape of the pocket, two different sets of probes identify the two sub-pockets forming the whole cavity. The center of mass of the probes inside the cavity is calculated and represented by ASPs (in blue). The two subpockets can be merged together in one bigger pocket by LiGenPocket.

## 1.2.2 LiGenPocket

LiGenPocket computes volume, shape, and physicochemical properties (donor, acceptor, hydrophobic, etc.) of the binding pocket and proposes a pharmacophore model based on these characteristics. LiGenPocket accepts as input a three-dimensional structure of the protein of interest in PDB or MOL2 file format. The basic algorithm of LiGenPocket is a variant of the one proposed in 2000 by Wang et al.<sup>40</sup> Briefly, LiGenPocket creates a regular Cartesian grid (grid spacing 0.5 Å) around the co-crystallized ligand, if there is one, or around the active site point (ASP) generated by LiGenPass, that indicates approximately the center of mass of the binding cavity. In the case of the co-crystallized ligand, the software first defines a sphere around the ligand (with a user defined radius) and then creates the grid inside it. In the first step, a hydrogen atom is used as a probe to check the accessibility of the grid points. If the probe bumps into the protein, that grid point will be labeled as “not free”; otherwise, it will be labeled as “free”. A bump is counted when the interatomic distance is less than the sum of the van der Waals radii reduced by 0.5 Å. In the second step, the possible interaction sites will be derived using three types of probes to map the main interactions usually occurring in binding sites: a positively charged sp<sup>3</sup> nitrogen atom (ammonium cation), representing a hydrogen bond donor; a negatively charged sp<sup>2</sup> oxygen atom (as in a carboxyl group), representing a hydrogen bond acceptor; and a sp<sup>3</sup> carbon atom (methane), representing a hydrophobic group. A score representing the binding energy between the probe and the protein will be calculated. To calculate the scores, LiGen uses an in-house developed scoring function based on the paper of Wang et al.<sup>54</sup> In this way, all the grid points are mapped and assigned three scores, representing the three binding energies of the interactions with the three kinds of probes (Figure 6). In general, however, not all of them would be worthy of being considered for the pharmacophore model definition. For this reason, only those grid points having at least one of the three score higher than the average score for that kind of interaction are retained. Then, the survived grid points are labeled as “H-bond donor”, “H-bond acceptor”, or “hydrophobic”, according to the highest score reached. The number of “neighbors”, defined as the number of grid points with the same label falling within a certain user-defined distance, is computed for every survived grid point. The average number of neighbors of the same type is calculated for each grid point, and only those having a number of neighbors higher than the average are retained for the further step and defined as “key sites”.



**Figure 6. LiGenPocket. A)**The binding site is analyzed using three types of probes, hydrogen bond donor, hydrogen bond acceptor and hydrophobic and regions with favorable interactions are mapped accordingly. **B)** Pharmacophore features are derived from the grid, clustering the probes into pharmacophore points, so that every point represents the center of mass of the probes.

Finally, the survived grid points forming the key sites are clustered, and the geometric center of each cluster represents a pharmacophoric feature. In the binding site analysis, several parameters can be tuned by users according to their own needs. For example, the minimal distance among the pharmacophoric features, or the grid spacing, can be modified through the grid accuracy parameter, selecting the desired degree of accuracy in the pocket description. All the user adjustable pocket parameters are reported in Table 1, along with a brief explanation of their functionalities.

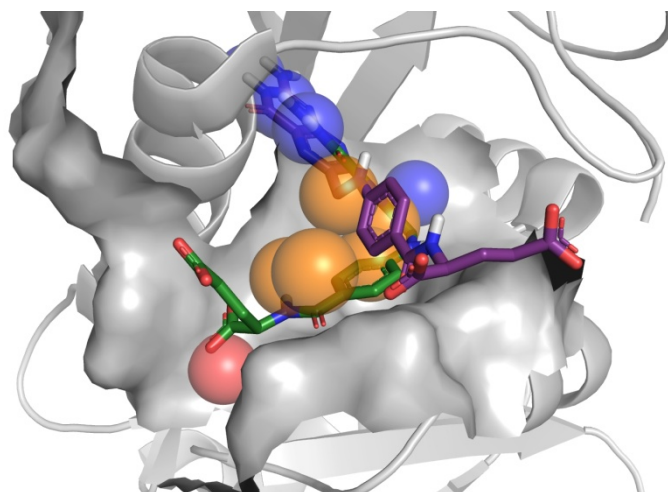
**Table 1. LiGenPocket parameters. Parameters written in *italics* are those included in the experimental designs.**

Parameter	Abbreviation	Notes
<i>Minimal Feature Distance</i>	Min F dist	Sets the minimal distance between the pharmacophore features identified in the binding pocket. (1 to 5 Å)
<i>Maximal Feature Number</i>	Max F num	Maximal number of features that can be considered in the pharmacophore that describes the binding pocket.
<i>Distance Cutoff</i>	Dist CO	Sets the cut-off radius to search for pocket atoms around a ligand. (When pocket is computed around a co-crystallized ligand)
<i>Van der Waals Bumps</i>	Prot VDW B	Defines how much the binding site surface is smoothed: it represents the fraction of the Van der Waals radius considered to tag the grid points.
<i>Grid accuracy</i>	Grid Acc	This parameter specify the grid spacing used in grid generation through the expression : grid spacing = 1.0 Å / Grid accuracy
<i>Ligand neighbor threshold</i>	Lig Neig Thr	Represents the ligand-ligand distance threshold used to count ligand neighbor atoms. It is used to coarse graining ligand atoms.
<i>Score distance threshold</i>	Score Dist Thr	This parameter is used to assign a score to the identified pharmacophore points. It specifies the maximum distance to take into account grid points around the pharmacophore point. The sum of each single grid point score gives the score of the pharmacophore point.
<i>Grid distance threshold</i>	Grid Dist Thr	This value defines for every grid point the area where to count the number of grid points of the same type. The number of grid points found is used to compute the score of the grid point taken into account.
Coarse Grain Ligand		If specified, enables to apply a filter to coarse grain the ligand or the cluster of probes generated by LiGen-Pass, to speed-up the calculation.
Include H bumps		If specified, allows to consider hydrogen atoms during the calculation of grid points that bumps the receptor.
Include water		This keyword allows to include water molecules in the calculation of the receptor grid.

### 1.2.3 LiGenDock

The main feature of LiGenDock is the use of the pharmacophore scheme generated by LiGenPocket as the driving force for the docking procedure, including a nonsystematic flexible docking algorithm. A simple description of the framework of the docking algorithm is the following:

1. One ligand is taken into account, and ligand features (e.g., H-bond donor site) are computed.
2. The docking process starts matching a ligand's feature (i.e., a hydrogen bond donor site) with the previously identified pharmacophoric features of the same type.
3. The docked ligand is rotated by an appropriate angle to match a second pharmacophore feature with a second ligand's feature. Because it is unlikely that the second pair will overlap perfectly, a user defined tolerance cutoff is used to evaluate the goodness of the match.
4. The ligand is then rotated by an appropriate angle around the axis passing between the two pharmacophore features trying to match a third feature (not necessary). Then torsional angles are unlocked, and ligand conformers are generated in situ trying to match as many features as possible (some torsional angles may be selectively locked by the user).
5. At every step, the pose's score, related to the estimated binding energy of the ligand-protein complex, is computed and compared with the scores of previously generated poses. If this actual score is better than the worst score of the already generated poses, the new pose is retained instead of the previous worst pose (the one with the lowest score). The risk of getting trapped in a local minimum can be minimized by imposing a high RMSD difference between two poses to be considered different and further processed.



**Figure 7.** LiGenDock uses the pharmacophore of the binding site as a guide for the docking process. Several ligand conformations inside the binding site are generated to match as many features as possible.

**Table 2.** LiGenDock parameters. Parameters reported in *italics* are those included in the experimental designs.

<b>Parameter</b>	<b>Abbreviation</b>	<b>Notes</b>
<i>Neighbor threshold</i>	Neig Thr	Indicates the number of neighbors of grid points of the same type necessary for a pharmacophore point to be considered as a candidate for ligand docking.
<i>Distance threshold</i>	Dist Thr	Is the maximum distance to consider a ligand functional group superposing on a pharmacophore feature.
Number of poses		Defines the number of output poses.
<i>Pose overlap</i>	Pose Over	Represents the maximum degree of overlap between two poses
Pose diversity		Is used to set a limit to the number of poses of a molecule.
<i>Hydrophobic threshold</i>	Hyd Thr	The value of atomic LogP above which an atomic site is considered hydrophobic. This parameter is used in scoring the interaction between the receptor and the ligand.
<i>Angle delta</i>	Ag Delta	Defines the extent of the angle used to rotate the ligands inside the pocket around the axis of the two matched pharmacophoric features
<i>Conformer Van der Waals bumps</i>	Conf VDW B	Defines the degree of ligand volume smoothing. It represents the fraction of the Van der Waals radius to be considered when computing bumping between fragment during conformer generation.
<i>Conformer angle delta</i>	Conf Ag Delta	Defines the extent of the angle used to rotate rotatable ligand bonds during conformer generation.

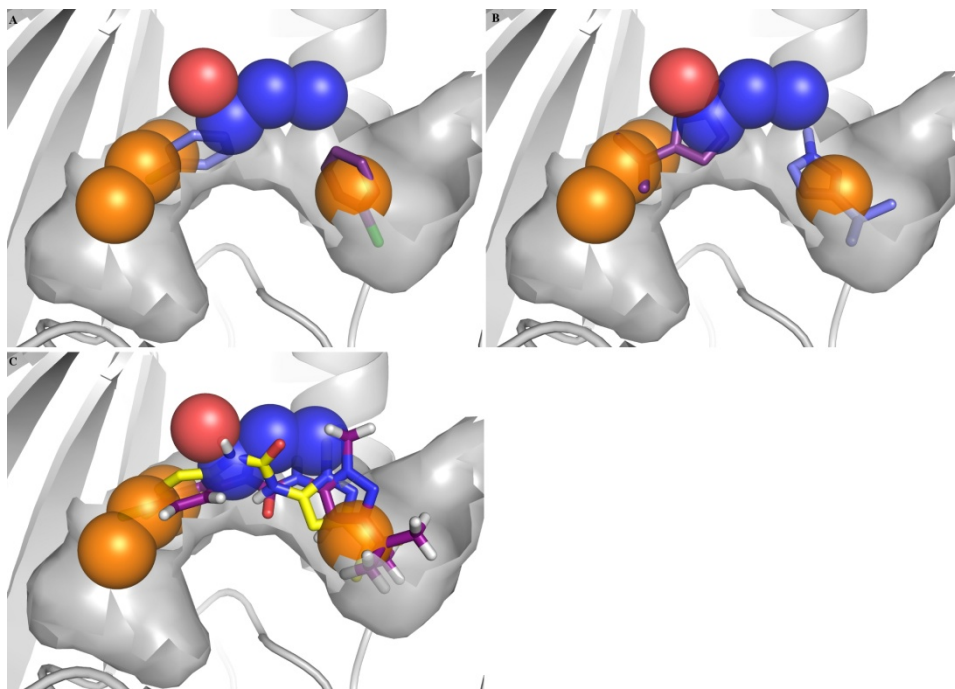
6. Finally, the score is optimized with a simple score minimization algorithm that treats the docked ligand as a rigid body inside the pocket. The ligand is rotated in space around the docking pose until a minimum score, representing the most favorable binding energy, is reached. This minimization is the steepest descendent minimization during which the ligand position inside the binding site is changed by a discrete value of 0.25 Å in seven directions of the three-dimensional space (3 axes and 4 quadrant bisectors), and then the direction of the diminishing score is taken.

Several parameters allow users to tune the docking algorithm according to their needs. A complete list of the parameters with a brief explanation of their meaning is given in Table 2. The values of these parameters can have a large influence on the outcome of the docking process and a suitable set of parameters is necessary for gaining good results. Docking outputs are a collection of ligand poses in mol2, pdb, or sdf file format, and a table summarizing scores and ligand/pharmacophore feature matches.

## 1.2.4 LiGenBuilder

LiGenBuilder is the structures generator module of LiGen. LiGenBuilder first places fragments into the active site. Atoms potentially involved in structure generation are tagged and a set of chemical rules is applied to link fragments together. LiGenBuilder does not need a library of pre-tagged fragments. Rather, any arbitrary database of fragments can be used without need of user manipulation in the *de novo* process. A throughout explanation of the algorithm of LiGenBuilder is reported in the paper of Beccari et al.<sup>72</sup>





**Figure 8.** LiGenBuilder performs the *de novo* design step. A) and B) As a first step fragments are docked in the binding site. Further, they are assembled inside the binding site using real chemical reaction, stored in the chemical rules database. C) Hence the poses generated are minimized inside the binding site and scored.

## 1.3 AIMS

Computational chemistry is now a well-established approach used throughout all the drug discovery process, from hit identification, to lead optimization and ADMET prediction. Among all the diverse techniques grouped together by the computational chemistry definition, one of the most popular in the drug discovery pipeline is molecular docking, which aims to predict the interactions occurring between a ligand and a protein, in terms of spatial arrangement and energy variation.

LiGen, is a new drug discovery software, based on a different docking approach compared to the classical docking algorithms. In fact it uses pharmacophore models of the binding pocket to guide the molecular docking process, limiting the number of calculations and conformations to be generated. Once the algorithm steering the docking process was written, as all new software, the need of parameters optimization became evident.

In addition, all drug discovery softwares, regardless of their final aim and how they manage the problem, are controlled by the values of several user-adjustable parameters, and an appropriate choice of these parameters is a prerequisite to obtain meaningful results.<sup>74</sup> However generally, users tend to adopt default settings, assuming that they will yield reasonably good results, regardless of the specific problem they are involved. Thus, providing the best ensemble of “default settings” is crucial for optimizing the program’s output under standard conditions. Furthermore, having an optimized set of default parameters enables benchmarking of the performance of the program at the best of its possibility.

To find the optimal set of parameters for our new docking program we decided to use experimental designs. This kind of approach is very popular in other field of drug discovery, for example to optimize reaction conditions or excipients content during the formulation processes, however they are seldom apply in computational chemistry. As previously reported by others (see, for example, Andersson et al.), experimental designs are useful also to computational chemistry related matters, providing the best way to find the

optimal ensemble of parameters by changing all of them simultaneously in a controlled and systematic way.<sup>75, 76</sup>

Therefore we decided to apply the Design of experiments (DoE) methodology to the LiGen case, to identify which user-adjustable parameters are influencing the results most and then to optimize their values.

The whole project is hence divided into three main subsequent steps, reflected by the subdivision of the results section of the chapter:

- 1) At first a DoE was set and performed to identify the parameters having the greatest impact on the docking results.
- 2) Afterwards, the values of the previously identify parameters were optimized using another DoE.
- 3) Finally a two-step validation was conducted:
  - i) Performances of LiGen in predicting the binding pose of a known active compound, i.e. self-docking, were compared to those of Glide and AutoDock
  - ii) LiGen ability to identify active compounds among inactives, i.e. virtual screening, were compared to published performances of DOCK6 and again of Glide.

## 1.4 MATERIALS AND METHODS

### 1.4.1 Data Set/Benchmark Composition.

During this study, three different data sets were considered, one for each step of the study. For the optimization of the cognate docking, i.e., the reproduction of the crystallographic ligand conformation, we used as a training set 100 crystallographic complexes taken from the CCDC Astex clean data set, consisting of 224 entries.<sup>63</sup> Trying to obtain a dataset the more representative as possible of the real situation, proteins were selected according to their relative family abundance in the PDB database.<sup>77</sup> First, the 224 complexes were grouped according to the protein family they belong to, and then the percentage abundance of those families in the whole PDB database was calculated. Finally, in agreement with the family representation percentage in the PDB, a hundred protein were selected out of the 224.

To test the optimized parameters for cognate docking, as well as to compare LiGen performance with those of commonly available docking software, we selected 171 complexes taken from the PDBbind CORE SET database (2010 release)<sup>64, 78</sup> excluding entries containing a ligand with molecular weight higher than 500 Da, in agreement with the definition of drug-like molecules given by Lipinsky.<sup>79</sup>

The third benchmark, used to optimize and test the VS ability of the LiGenDock algorithm, was obtained by selecting 36 high-quality crystal structures from the first version (the only one available at the beginning of the study) of directory of useful decoys (DUD).<sup>62</sup> All complexes collected in this third dataset are listed in Table 3. In the present study, four targets originally present in DUD were excluded from the selection: i. the human vascular endothelial growth factor receptor 2 kinase domain (VEGFr2, PDB code: 1vr2) because it lacks a co-crystallized ligand that we used to center the binding site grid (even though we could have used LiGenPass to define the binding site location, we preferred to use the same approach, i.e., using the co-crystallized ligand to center all the binding site grids; therefore, we

**Table 3: DUD complexes taken into account in this project. Underlined complexes are those randomly chosen for VS optimization**

<i>Protein</i>	<i>PDB code</i>	<i>resolution Å</i>	<i>no. Ligands</i>	<i>no. Decoys</i>
<b>Nuclear Hormone Receptors</b>				
<i>ERagonist</i>	1l2i	1.9	67	2361
<u><i>ERantagonist</i></u>	<u>3ert</u>	1.9	39	1399
<u><i>GR</i></u>	<u>1m2z</u>	2.5	78	2804
<i>MR</i>	2aa2	1.9	15	535
<i>PPARg</i>	1fm9	2.1	81	2910
<i>PR</i>	1sr7	1.9	27	967
<i>RXRa</i>	1mvc	1.9	20	708
<b>Kinases</b>				
<i>CDK2</i>	1ckp	2.1	50	1780
<i>EGFr</i>	1m17	2.6	416	14914
<u><i>FGFr1</i></u>	<u>1agw</u>	2.4	118	4216
<u><i>HSP90</i></u>	<u>1uy6</u>	1.9	24	861
<i>P38 MAP</i>	1kv2	2.8	234	8399
<i>SRC</i>	2src	1.5	162	5801
<u><i>TK</i></u>	<u>1kim</u>	2.1	22	785
<b>Serine Proteases</b>				
<u><i>FXa</i></u>	<u>1f0r</u>	2.7	142	5102
<i>Thrombin</i>	1ba8	1.8	65	2294
<i>Trypsin</i>	1bjv	1.8	43	1545
<b>Metalloenzymes</b>				
<i>ACE</i>	1o86	2.0	49	1728
<i>COMT</i>	1h1d	2.0	12	430
<i>PDE5</i>	1xp0	1.8	51	1810
<b>Folate Enzymes</b>				
<i>DHFR</i>	3dfr	1.7	201	7150
<i>GART</i>	1c2t	2.1	21	753
<b>Other Enzymes</b>				
<i>AChE</i>	1eve	2.5	105	3732
<i>ALR2</i>	1ah3	2.3	26	920
<u><i>AmpC</i></u>	<u>1xgi</u>	2.0	21	734
<i>COX-1</i>	1p4g	2.1	25	850
<i>COX-2</i>	1cx2	3.0	349	12491
<i>GPB</i>	1a8i	1.8	52	1851
<i>HIVPR</i>	1hpx	2.0	53	1888
<i>HIVRT</i>	1rt1	2.6	40	1439
<u><i>HMGR</i></u>	<u>1hw8</u>	2.1	35	1242
<i>InhA</i>	1p44	2.7	85	3043
<i>NA</i>	1a4g	2.2	49	1745
<i>PARP</i>	1efy	2.2	33	1178
<u><i>PNP</i></u>	<u>1b8o</u>	1.5	25	884
<u><i>SAHH</i></u>	<u>1a7a</u>	2.8	33	1159

excluded this complex); ii. the platelet-derived growth factor receptor kinase (PDGFRb) because it is an homology model; iii. the androgen receptor (AR, PDB code: 1xq2) because it was superseded in the PDB by 2ao6; iv. the adenosine deaminase (ADA, PDB code: 1vdw) because there was a mismatch with the PDB code published in the original paper.

In the optimization phase, we randomly select 10 targets among those selected from the DUD to reduce the computational efforts needed (Table 3, underlined entries). To evaluate the improvements gained through the optimization procedure, the experiments were performed with default and optimized parameters, using all 36 targets selected from the DUD.

## 1.4.2 Protein Preparation

Protein preparation is a necessary step before every study involving PDB structures, due to the information missing in the PDB file but necessary for docking, for example hydrogen atoms, Water molecules were removed from all the considered proteins. The protein protonation state of complexes taken from the CCDC Astex database was retained as provided by Astex. Protein structures of complexes taken from PDBbind and DUD were prepared using Protein Preparation Wizard,<sup>80</sup> contained in the Maestro suite, undergoing the following preparation steps: (a) hydrogen atoms were added according to the protonation state at pH 7.0, (b) ions and crystallization cofactors were removed, (c) atom and bond types were assigned, and (d) an energy minimization in OPLS2005 was run to refine the structure.

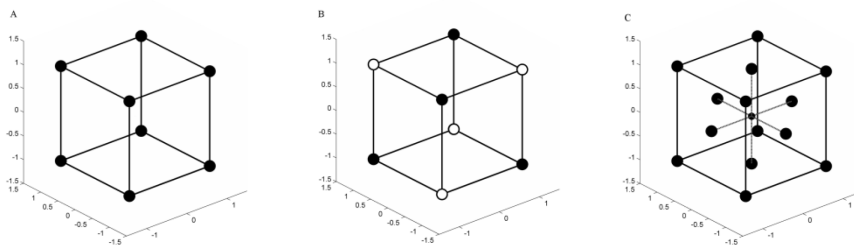
## 1.4.3 Ligands Preparation

Ligand molecules were prepared using LigPrep of Maestro.<sup>81</sup> Cognate docking involves the redocking of the co-crystallized ligand to see whether the docking process is able to reproduce the crystallographic conformation (also called self-docking). Using a starting conformation different from the co-crystallized one is therefore important to evaluate the ability of the software to reproduce the crystallographic ligand conformation. Hence, for every ligand of the selected CCDC Astex and PDBbind CORE SET complexes, a set of conformers, not containing the co-crystallized conformation, was

generated using ConfGen,<sup>82, 83</sup> also included in Maestro suite. Then, a conformer was randomly chosen from this set as a starting conformation for docking. For rigid ligands, no conformers can be generated, so the only option was to assign 3D coordinates different from those in the PDB.

#### 1.4.4 Experimental Designs

Experimental designs are tools used to systematically examine different types of problems, as the identification of the variables affecting the results at most, or to find optimal values of some of the variables in order to obtain better results. Once selected the variables to be investigated (also called “factors”) and the response to be evaluated, experimental designs are used to plan the experiments to obtain the maximum of information from the minimum number of experiments. Several types of experimental designs are available, covering different types of problems. A throughout explanation of experimental designs can be found in references <sup>84-86</sup>. They can be roughly divided into two families, screening designs and optimization designs. To the first belong those design useful to explore the experimental variables, the interactions subsisting among them and their effect on the results, and the two most popular type of designs are full factorial designs (FDs) and fractional factorial designs (FFDs). In FDs the influence of all experimental variables and the interaction effects on the responses are investigated. If the combination of  $k$  factors is investigated at two levels, a factorial design will consist of  $2^k$  experiments. Therefore a design accounting for two 2 variables will contain  $2^2 = 4$  experiments, for three  $2^3 = 8$ , for four  $2^4 = 16$  etc. Some center experiments “center points” should always be included to avoid the risk of not recognize non-linear relationships and to allow the determination of confidence intervals through their repetitions. FFDs are a subset (fraction) of the experimental runs of a FDs, and are used to reduce the number of experiments to perform compared with the original FDs. For example, if an experiment has eight variables, trying to explore all the effects of these variables with a FD, means that 256 ( $2^8$ ) experiments should be performed. Generally in most investigations it is reasonable to assume that the influence of the interactions of third order or higher are very small or negligible and can then be excluded, and this is actual the purpose of using FFDs. The general expression of FFDs, defining the number of experiments to be performed,



**Figure 9. Schematic representations of some experimental designs, in which every sphere represents an experiment. A) A  $2^3$  Full factorial design: the exponent represents the number of investigated variables (factors) and 2 is the number of levels (values) for each factor. B) A  $2^{3-1}$  Fractional Factorial Design, in which 2 is once again the number of levels for each factor, 3 is the total number of factors and -1 represents the level of fractionation; the resulting number ( $2^{3-1}=2^2=4$ ) is the number of the experiments to be performed. C) Faced Central Composite Design for two parameters with three levels.**

is  $2^{k-p}$ , where  $k$  is the number of variables and  $p$  the size of the fraction. The size of the fraction will, of course, define the number of effects and interactions between variables to be estimated and the number of experiments needed. The effects of variables interaction that are not estimated are called *confounded*. An important characteristic of a fractional design is the defining relation, i.e. the so-called *generator*,  $I$ , which gives the set of interaction columns equal in the design matrix to a column of plus signs,  $I$ , that contains the information about how the different columns can be multiplied to obtain it. From the generator one can evince which interaction effects are confounded. The amount of confounded interaction effects defines FFDs are characterized by a property called resolution, the is defined as the shortest “word” (derived from the combination of columns) in the set of generators. The most popular FFDs are:

- Resolution III: the main effects are confounded with two-variables interaction effects; this type of designs, that does not estimate the interaction effects, are particularly useful in a first screening phase, where the most significant set of factors are sought.<sup>84, 87</sup>
- Resolution IV: the main effects are confounded with three-variable interaction effects, and the two-variable interaction effects are confounded with each other;



- Resolution V: the main effects are confounded with four-variable interaction effects, and the two-variable interaction effects are confounded with the three-variable interaction effects.

The most popular type of optimization design is Response Surface Methodology (RSM). RSM is a collection of mathematical and statistical techniques based on the fit of empirical models to experimental data obtained in relation to the experimental designs.<sup>88, 89</sup> Studying variables at least using three different values allows to determine first- and second-order effects and possibly also critical points (maximum, minimum, or saddle). Two types of RSM exist, the Box-Wilson Central Composite Design (CCD) and the Box-Behnken design (BBD). BBD is an independent quadratic design that does not contain an embedded factorial or fractional factorial design. In this design the treatment combinations are at the midpoints of edges of the process space and at the center. Conversely CCDs contain an imbedded factorial or fractional factorial design with center points that is augmented with a group of 'star points' that allow estimation of curvature. There are three types of CCD, depending on where the points of the star design ("star points") are located: (1) circumscribed, with star points located outside the factorial space, (2) inscribed, with star points located inside the factorial design space, used when points of the factorial design are real experimental limits, and (3) faced, with star points located on each face of the factorial space.<sup>84</sup>

Different equation can be used to fit the models, the easiest one is Multiple linear regression (MLR), that is therefore the first choice, and in case it does not work, are applied more complicated equation, as quadratic or polynomial equation. Luckily, in our case MLR was able to explain the effects of the variables, therefore is was used to investigate the relationship between the docking parameters (independent variables or predictors) and the results (dependent variables or predictand). MLR is based on least squares<sup>90</sup>: the model is fitted such that the sum-of-squares of differences of observed and predicted values is minimized. The general equation of the model is :

$$\begin{matrix}
 \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} & = & \begin{pmatrix} f_1(x_1) & \dots & f_p(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \dots & f_p(x_n) \end{pmatrix} & \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} & + & \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\
 y & & X & \beta & & \varepsilon
 \end{matrix}$$

in which  $y_i$  is the *response* or *dependent variable*,  $x_i$  is the *input* or *independent variable* and  $f_j(x_i)$  is a function of the input variable  $x_i$  (sometimes the variable  $x_i$  can be a function of the data). The global response  $y$  is expressed as a linear combination of model terms  $f_j(x)$  ( $j=1, \dots, p$ ) at each of the observations  $(x_1, y_1), \dots, (x_n, y_n)$ ,  $\beta$  is the coefficient of the parameters and  $\varepsilon$  is the residual term associated to the experiments. The function  $f_1(x)=1$  is included among the  $f_j$ , so that the model contains a constant term (the intercept). Coefficients resulting from the design model, were used to interpret parameters' influence on the docking performance.

During model fitting, some statistics of the models can be calculated. In our study to evaluate the accuracy of the model were computed:<sup>90</sup>

- R2 value, which represents the explanatory power of the regression model, computed from the sum-of-squares as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where SST is the total sum of squares, SSE is the sum of error squares and SSR is the sum of squares due to the regression computed respectively as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- F-ratio expressed by

$$F = \frac{MSR}{MSE}$$

where MSR is mean squared regression and MSE the residual mean square. F-ratio represents the explanatory power of the model but the advantage over R2 is that F-ratio takes into account the degrees of freedom, which depend on the sample size and the number of predictors in the model. In this way F-ratio incorporates sample size

and number of independent variables in the assessment of significance of the relationship.

- Adjusted R<sup>2</sup> attempts to compensate for the fact that R<sup>2</sup> for a regression can be made arbitrarily high by including more predictors in the model. Adjusted R<sup>2</sup> is given by

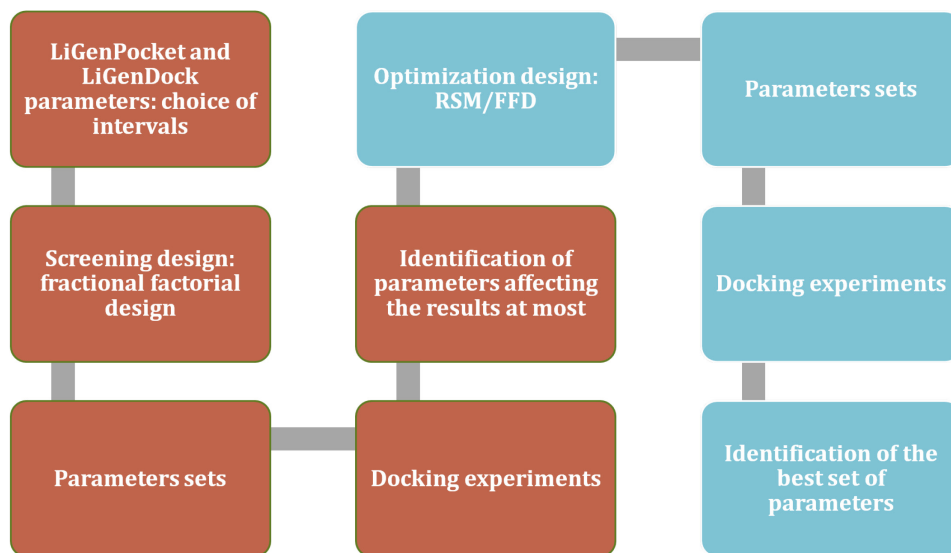
$$\bar{R}^2 = 1 - \frac{\text{MSE}}{\text{MST}}$$

where MST is total mean square.

- P value was calculated for each parameter, and its value represents the level of significance of the parameter. A value smaller than 0.05 means that coefficient calculated by the model is significant with a confidence interval of 95%

### 1.4.5 Screening and Optimization Workflow

In this study, we were interested in analyzing how the docking performance of LiGen is influenced by the different set of parameters and in finding an optimal set of parameters that allows to obtain the best results, both in pose prediction and virtual screening. Experimental designs represented a good solution for our needs because they offer the possibility to vary all the parameters under investigation at the same time. Our goal was first to find out which parameters affect the results most and then to establish the values for optimal docking performance. Therefore we needed at first to screen LiGen parameters to find out which of them were important for the outcome, and then we went into a second-phase of optimization, to assign the optimal values to the previously identified parameters. The general workflow applied is summarized in Figure 6.



**Figure 10.** Flowchart of the main steps of the study. The first five blocks (in brown) represent the steps of the screening procedure to identify the parameters affecting the results at most. The second four blocks in cyan are the steps concerning the optimization phase.

As reported by Leach et al.<sup>26</sup> docking algorithms are generally used for pose prediction and for virtual screening. Pose prediction and virtual screening have different goals: the goal of the first one is to predict how a ligand may bind, assuming that ligand can bind, whereas the aim of the second one is to predict whether a ligand can bind or not. Because they have different aims, the parameters to use can be slightly different; therefore, we decided to optimize parameters for the two main docking applications independently, after having verified that parameters optimized for pose prediction did not gain the sought improvement in VS results when compared with the starting parameters.

Responses evaluated in the models were, for cognate docking, the number of poses with RMSD from the co-crystallized conformation less than 2 Å, whereas for virtual screening, the early enrichment, assessed by the value of the area under the receiver operating characteristic curve (ROC) measured at 1% of the screened database (ROC(1%)).

For our first aim, the identification of the most relevant parameters affecting the docking results, we applied a fractional factorial design (FFD) (Figure 2B). All experimental designs were prepared using MATLAB software.<sup>91</sup> For every quantitative parameter (i.e., parameters for which a

numerical value can be assigned), high and low values were selected, together with a center point, representing the three levels in the design procedure. The range of values for the parameters' intervals was selected in order to be large enough to be sure to capture the effect of the parameter, if there is one. We decided to exclude qualitative parameters (i.e., parameters for which on/off values can be assigned) from our analysis. After the key factors were identified, an optimization phase was performed using full factorial design FD (in the case of virtual screening optimization) or response surface method (RSM) design (in the case of cognate docking optimization; Figure 2A,C, respectively). For our work, we chose a faced CCD and not an inscribed or circumscribed CCD, because for some of the parameters only three levels were possible in the desired design space and the other two type of CCD requires five levels(Figure 2C).

#### **1.4.6 Docking with Glide and AutoDock**

The performances of LiGenDock in cognate docking were compared to the ones of two commonly used programs, namely, Glide and AutoDock. In the comparison, both accuracy and speed were considered and analyzed. Docking with Glide was performed using standard precision (SP) mode, using the default set of parameters, except for the number of required poses, which was changed to 10, so that the same number of poses was used for all programs (Glide, LiGen, and AutoDock). AutoDock uses the Lamarkian version of the genetic algorithm to generate the ligand poses inside the protein active site.<sup>18</sup> In our test, we used the default parameter set. Both in Glide and AutoDock, as well as with LiGen, the active site selection was based on the position of the native ligand in the crystallographic complex.

#### **1.4.7 Evaluation of Self-Docking and Virtual Screening Results**

Along with the increased number of scientific papers reporting new docking software and/or docking software evaluation,<sup>13, 56, 67, 92, 93</sup> recommended guidelines for docking evaluation appeared<sup>3, 59</sup> in the past years. Consistent with those recommendations, we used as response during the optimization of pose prediction the percentage of best-predicted poses

with RMSD less than 2.0 Å from the experimentally-solved ligand structure. In comparing LiGen results with those of other software, we calculated also the percentage of poses with RMSD less than 3.0 Å, both for best-predicted pose (the one with the lowest RMSD, irrespective of the ranking position) and best scoring pose (the pose ranked first by the scoring function). Moreover, the computational time needed for docking was also calculated to better evaluate software performance. Among all the possible metrics previously illustrated to evaluate VS accuracy, we decided to use the area under the receiver operating characteristic (ROC) curve to measure the global enrichment; to evaluate the early enrichment, we applied the values of the AUC under the ROC curve at 1%, 2%, 5%, and 20% of the x-axis, referred to hereafter as ROC(1%), ROC(2%), ROC(5%), and ROC(20%), respectively, as suggested by Repasky et al.<sup>94</sup> BEDROC, with a value of 20 for parameter  $\alpha$ , as suggested by other works<sup>69, 70</sup> was also calculated.

We decided also to evaluate the ability of LiGen to recognize different chemotypes in VS experiments, as the identification of diverse chemical series is extremely important in drug discovery. To assess the recognized chemotypes during VS experiments we used Typed Graph Triangle (TGT) fingerprints.<sup>95</sup> Fingerprints are a very abstract representation of structural features, 2D and/or 3D, of a molecule, used to describe compound similarity or to retrieve a particular class of molecules from a large database. TGT fingerprints are conformation-independent and can be calculated from a two-dimensional representation of the molecule. Each fingerprint is the set of all tuples of the form  $(u,v,w,d,e,f)$ , where  $u, v$  and  $w$  are atom types and  $d, e$  and  $f$  are graph distances between the atoms. The graph distance is defined as the number of bonds in the shortest path between the atoms in the chemical graph. Each atom is assigned one of the following types: 1. *D*, Hydrogen bond donor or Base; 2. *A*, Hydrogen bond acceptor or Acid; 3. *P* Hydrogen bond acceptor and donor; 4. *H* Hydrophobic. Distances are binned into categories so that there is higher resolution in the smaller distances and less in the larger distances.

TGT fingerprints were calculated for every active ligand of each target. Afterwards every set of active compounds was clustered according to TGT similarity using Tanimoto coefficient. Structure were considering belonging to the same cluster if the similarity between fingerprints was higher than 0.88.

## 1.5 RESULTS AND DISCUSSION

LiGen, is a new drug discovery software that uses pharmacophore models of the binding pocket to guide the molecular docking process. As a new software, LiGen, its parameters, needed to be optimized. Instead of varying all the parameters one at a time, we decided to apply a different approach to the optimization. We used experimental designs first to identify which user-adjustable parameters are influencing the results most and then to optimize their values. Once completed the optimization procedure a two-step validation was performed to assess LiGen performances in pose prediction and virtual screening.

### 1.5.1 Pose Prediction Optimization

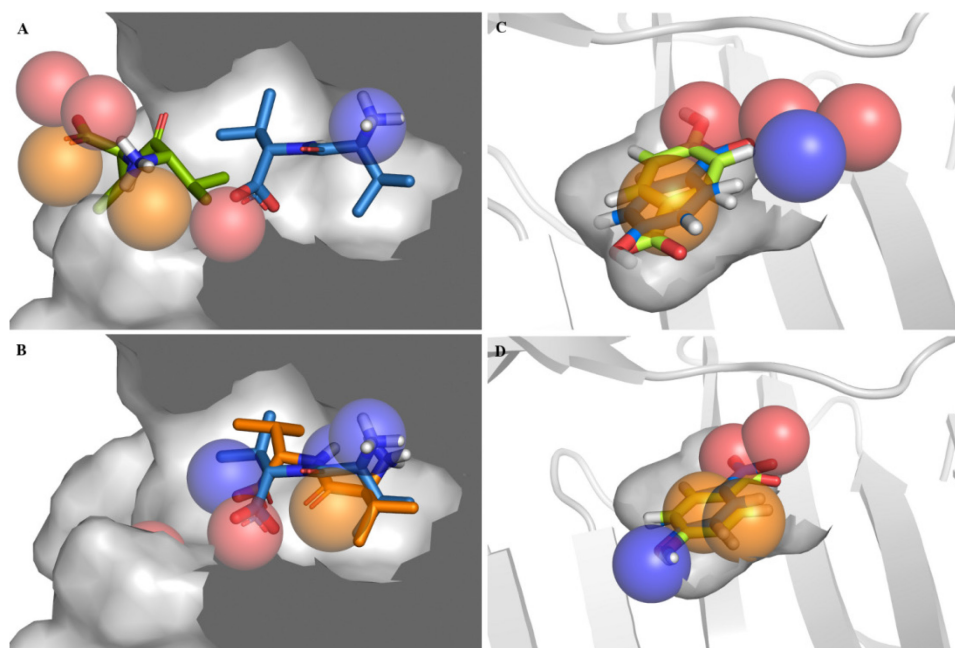
As a first step before proceeding in the optimization, we assessed the original LiGen performance by using the set of parameters assigned during the code development. As shown in Table 4 (column “RMSD original results”), results with the original set of parameters were not very good. Only in six cases out of 100 the best predicted pose has an RMSD less than 2 Å from the co-crystallized ligand. In 15 cases LiGen failed to find a pose. Visual inspection of results showed that many poses are located outside the binding site, for example for proteins 1bgo, 1bmq, 1cqp. Moreover in many cases also the pharmacophore models were not completely contained within the binding site, but pharmacophoric features were also present on the protein surface outside the cavity (1cbs, 1d4p, 1ett, 4tpi etc.), highlighting the need for “smoothing” molecular surface by scaling the Van der Waals radii of receptor atoms (tuning the value of the parameter Van der Waals bumps), to reduce penalties for ligand-protein close contacts (Figure 11A). In other experiments, for example for 2phh and 1fkg, we found that even if the pharmacophore was placed inside the binding site, the functional groups needed for binding were not or not completely positioned correctly (Figure 11C).

**Table 4. Original results and after the optimization. Results for the 100 proteins taken from the Astex database. The RMSD here reported is the RMSD of the best predicted**

pose with respect to the co-crystallized ligand. At the bottom of the table results are summarized, and the average docking time is reported.

PDB code	RMSD Orig.	RMSD After Optim.	PDB code	RMSD Orig.	RMSD After Optim.	PDB code	RMSD Orig.	RMSD After Optim.
1a4g	5.61	2.37	1fkg	6.07	5.01	1tph	2.12	1.79
1a9u	8.20	1.15	1frp	7.58	3.07	1tpp		0.73
1acj		0.76	1ghb	10.65	5.03	1trk		4.31
1acm	2.45	1.74	1glp	7.42	6.21	1tyl	4.50	1.16
1apu		5.13	1gpy	5.16	2.37	1ukz	7.11	1.46
1aqw	8.00	2.10	1hdc	13.99	2.18	1ulb	5.11	1.21
1ase	6.29	1.24	1hfc	6.01	2.52	1ydr		1.26
1b59	9.58	2.18	1imb	1.93	1.47	1yee	7.96	2.17
1bgo	13.28	3.30	1ivb	4.98	2.1	2ak3	6.60	2.6
1bl7	6.43	2.29	1ivq	12.83	5.84	2cht	1.98	0.98
1blh	4.03	1.74	1ldm	5.45	0.92	2cmd	2.99	0.73
1bmq	8.75	6.49	1mld	2.59	1.85	2cpp	1.47	1.09
1byb		4.15	1mmq	4.87	3.02	2dbl	2.91	2.72
1byg	9.63	0.71	1okl	4.22	2.27	2fox	6.09	1.99
1cbs	10.99	1.41	1pbd	5.91	0.36	2h4n	9.53	1.83
1cdg		3.90	1pdz	1.58	1.73	2phh	4.05	0.26
1cil	6.1	2.16	1pgp	4.8	4.33	2qwk	9.61	2.53
1cle	23.54	3.25	1phd	4.32	1.4	2r07	11.91	2.17
1coy	1.12	1.15	1phg	5.78	0.72	2tsc	6.52	2.65
1cqp	8.95	2.09	1ppi		4.87	2yhx	5.88	4.35
1cvu		2.42	1pso	12.88	6.77	3cla	6.07	2.69
1d4p	12.33	1.58	1qbr	14.35	4.28	3cpa	6.65	2.62
1dd7	6.59	5.71	1rbp	12.62	2.22	3ert	11.25	1.82
1dhf	7.83	2.40	1rds	7.87	3.00	3hvt		3.28
1die	3.66	2.35	1rob	10.41	1.97	4aah		1.06
1dy9	7.90	3.88	1rt2	12.16	1.42	4cox	10.41	4.14
1ejn	11.42	6.71	1slt	4.99	1.9	4cts	1.04	1.05
1elc	9.50	4.15	1snc	7.12	1.8	4er2	13.1	8.61
1eta	4.77	2.33	1tdb	3.04	2.21	4fab	4.52	1.17
1ets	8.35	5.10	1tka		2.42	4phv	14.25	1.53
1ett	8.11	5.43	1tmn	9.45	5.08	4tpi	6.74	1.16
1f0s	5.20	1.65	1tng		4.95	5abp	2.05	2.83
1fen	12.32	2.09	1tni		2.94	7tim	10.07	1.39
1fgi		1.96						
		RMSD < 1 Å	RMSD < 2 Å		RMSD < 3 Å		missing results	Time (s)
ORIGINAL RESULTS		0%	6%		12%		15	26.49
AFTER OPTIMIZATION		9%	41%		70%		0	27.23





**Figure 11.** Examples of improvements in pose prediction gained through the optimization protocol. **A)** Using the original parameters pose (in lime green) is placed outside the binding site with respect to the experimentally determined ligand pose (in blue, complex PDB code: 4tpi). **B)** With the optimized set of parameters, especially decreasing a little the Van der Waals volume of the protein atoms forming the binding site, it is possible to produce a pose (in orange) that overlaps quite well the crystallographic one. **C)** At the beginning pose (light green) is flipped respect to the crystal complex (in cyan, PDB code 2phh) due to a non optimal position of the H bond donor/acceptor features of the pharmacophore, whereas **D)** after the optimization the pharmacophore allows to generate a pose that completely overlaps the original one. Figures are prepared with PyMol

All the quantitative LiGen's parameters involved in the binding site characterization and in the docking process itself were selected to perform an experimental design to assess which among them had most impact on results. A total of 15 parameters, eight from LiGenPocket and seven from LiGenDock were selected. Selected parameters are those reported in *italics* in Table 1 and Table 2. In building the binding site grid, non polar hydrogen atoms were not considered, therefore the parameter *include H bumps* was excluded from the experimental design, as well as the *coarse grain ligand* parameter, which allows to speed-up the grid-defining process (it was not included in the study because the investigation of the speed-up process was beyond the scope of this paper). The fifteen selected parameters were used to generate a fractional factorial design of resolution III, with 128

experiments, plus one center point. The parameter intervals were chosen in order to have the value of the central point set at the original value for all parameters, except for those having as original value the lowest or the highest possible value. The resulting experiments, with the set of parameters and results, are reported in Table II of Appendix A. We obtained a significant model, whose statistics are reported in Table 5. The model is not of outstanding quality, due to the fact that many combinations of parameters did not yield any results. A deeper analysis of these results correlates experiments with no or very poor results to the lowest grid accuracy value. However, some important conclusions can be drawn from statistics in Table 5. It suggests that parameters that influence the cognate docking experiments most, are i) the minimal distance between two pharmacophoric features (minimal feature distance), ii) the maximal number of features identified in the binding site (maximal feature number), iii) how much the protein's Van der Waals volume is smoothed (Van der Waals bumps), iv) the grid spacing (grid accuracy) and v) the tolerance in considering a ligand functional group superposed to the feature (distance threshold). All these parameters have a p value lower than 0.05 (Table 5). The p value for neighbor threshold, that indicates which pharmacophoric points should be considered during docking, is just a bit above 0.05, the threshold for being statistically meaningful, so since it is on a border line it should probably be taken into account. For this parameter and also for the other parameters having low influence according to the regression model, the experimental design was repeated investigating two extra-levels outside the values range of the first design, to ensure the low impact was not due to the previously selected ranges. The new parameters' ranges were chosen by extending the previous extreme values by 25% of the difference between them. Results obtained from the two extended levels were compared to those obtained with the original high and low ones respectively, using the non parametric Kolmogorov-Smirnov test with a 0.05 level of significance (Table 6).<sup>96, 97</sup> The test confirms that among the excluded values only neighbor threshold influences the quality of the results, as we already supposed, and should be considered for future analysis.

The six parameters thus identified were then used to perform a RSM, to investigate additional value levels and to study the interaction effect between parameters. The other parameters, which the previous analysis showed to be less influential from the previous analysis, were assigned the central value of the screening design. A faced CCD was identified as the

**Table 5. Designs statistics**

	<b>R<sup>2</sup></b>	<b>R<sup>2</sup> adjust</b>	<b>F</b>
<b>Design 1 (FFD)</b>	0.7113	0.6727	184.007
<b>Design 2 (RSM)</b>	0.9450	0.8971	197.258

<b>Design 1(FFD)</b>	<b>p value</b>	<b>Design 2 (RSM)</b>	<b>p value</b>	<b>P value</b>
Min F Dist	2.63E-04	Min F Dist	0.607	Vdw B Prot * Grid Acc 0.776
Max F Num	3.02E-08	Max F Num	0.001	Vdw B Prot * Dist Thr 0.13
Dist C.O.	0.984	Vdw B Prot	0.344	Vdw B Prot * Neib Thr 0.89
Vdw Bumps P	2.95E-05	Grid Acc	0.047	Grid Acc * Dist Thr 0.78
Grid Acc	7.87E-28	Dist Thr	0.336	Grid Acc * Neib Thr 1.11E-12
Lig Neib Thr	0.984	Neib Thr	0.157	Dist Thr * Neib Thr 0.89
Score Dist Thr	0.953	Min F Dist * Max F Num	0.002	Min F Dist * Min F Dist 0.91
Gris Dist Thr	0.116	Min F Dist * Vdw B Prot	0.323	Max F Num * Max F Num 0.01
Hyd Thr	0.381	Min F Dist * Grid Acc	0.396	Vdw B Prot * Vdw B Prot 0.27
Dist Thr	0.022	Min F Dist * Dist Thr	0.887	Grid Acc * Grid Acc 2.77E-03
Pose Over	0.800	Min F Dist * Neib Thr	0.125	Dist Thr * Dist Thr 0.67
Ag Delta	0.521	Max F Num * Vdw B Prot	0.396	
Conf Vdw B	0.953	Max F Num * Grid Acc	0.479	
Neib Thr	0.066	Max F Num * Dist Thr	0.054	
Conf Ag Delta	0.682	Max F Num * Neib Thr	0.015	

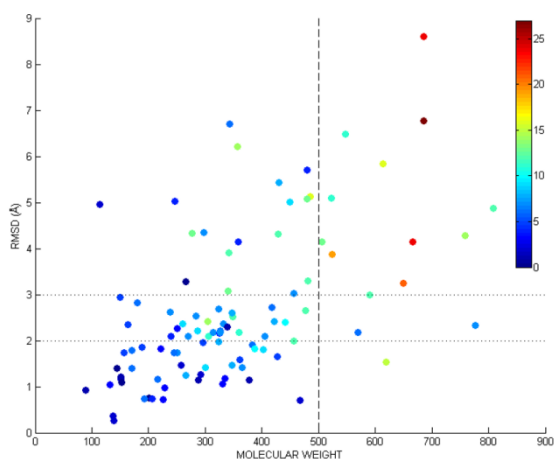
design that best suits our needs. Parameters and results of this experimental design are given in Table III of Appendix A. With respect to the FFD, there was a significant enhancement in the percentage of poses with RMSD less than 2Å for all the experiments. Parameters having the greatest influence on results were: i) the maximal number of identified pharmacophoric features; ii) grid accuracy, i.e. how fine is the grid spacing used in the analysis of the binding site and iii) the interaction between the previous two with neighbor threshold (the number of grid points needed to consider a pharmacophore feature suitable for docking). Best experiments produced poses with RMSD < 2Å in 43 cases (experiments 12, 34 and 37,

**Table 6. Results of the Kolmogorov-Smirnov test from additional dockings using parameter values outside the design range (expanded setting) compared with the high and low parameter values in the design (design setting). Difference in docking results using the expanded settings compared to the designed ones was defined as Kolmogorov-Smirnov  $d$  value  $\geq 0.240$  which corresponds to a 0.05 level of significance for a sample size of 64**

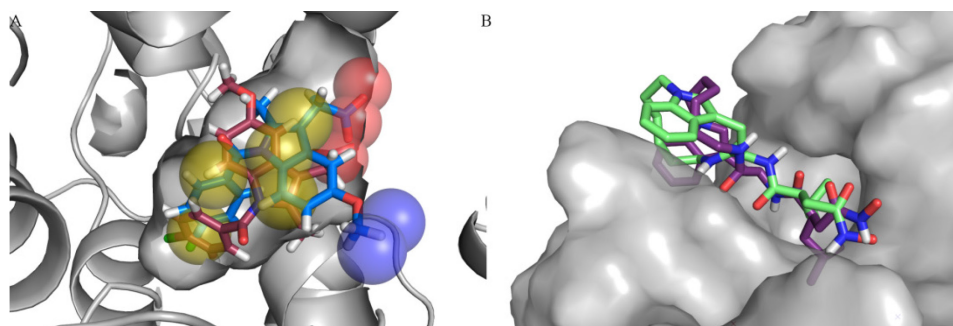
Parameter	settings		D value
	design	expanded	
ligand neighbor threshold	0.5	---	---
	3	3.875	
grid distance threshold	1	0.5	---
	3	3.5	0.167
distance cut off	4	3.5	0.187
	6	6.5	0.100
score distance threshold	1	0.5	0.029
	3	3.5	0.000
conformer angle delta	3	1	0.111
	10	12	0.167
neighbor threshold	50	25	0.292
	150	175	0.281
hydrophobic threshold	0.1	0.05	0.000
	0.3	0.35	0.222
angle delta	10	---	---
	50	60	0.114
conformer Van der Waals Bumps	0.5	0.125	0.081
	1	---	---
pose overlap	0.5	0.125	0.220
	1	---	---

Table III A), and with RMSD  $< 3\text{\AA}$  for 70 complexes out of 100 (experiments 10 and 28, Table III Appendix A). An exhaustive analysis of results revealed that bad results occur more frequently in cases of non-drug-like ligands. As shown in the scatter plot reported in Figure 12, in most cases the poses with higher values of RMSD involve ligands with molecular weight higher than 500Da and more than ten rotatable bonds. Among the exceptions of badly predicted drug like compounds, localized in the upper left part of the scatter plot, are ligands that feature bad pharmacophore-ligand matching due to unsampled binding conformations, as in the cases of 1ejn and 4cox, or due to a binding site partially exposed to the solvent, as in the cases of 1tng and 1mmq (Figure 13).

These results indicate that this round of parameter optimization allowed us to significantly improve the performance of LiGenDock with respect to the initial set of parameters. Yet the obtained results may seem



**Figure 12.** Scatter plot of the RMSD of the best predicted poses of the best experiment of the RSM (experiment 28). Abscissa shows the molecular weight of the ligands. The color encodes the number of rotatable bonds. The best results were observed for drug-like molecules.



**Figure 13.** Examples of badly predicted drug like compounds with the optimized parameters. A) Unsampld ligand binding conformation (PDBcode:4cox, crystallographic ligand pose in blue). B) Partially solvent exposed binding site (PDBcode:1mmq,crystallographic ligand pose in pale green)

not exceptional in terms of absolute metrics, given the number of poses predicted within 2Å from the co-crystallized ligand.

However, it should be noticed that LiGenDock has been originally derived within a *de novo* design suite of programs, where the main objective is the identification of novel chemotypes able to interact with the partner macromolecule. The pharmacophore driven docking procedure used by LiGenDock, based on a non-systematic conformational sampling, results in very high speed, performing docking experiments in an average time of only 27 seconds per protein. Apparently, the cost for speed is paid by a reduced accuracy, although deep visual inspection of the results strongly suggests that poses within 3Å from the co-crystallized ligands are still quite

**Table 7. Virtual Screening results of preliminary test with original parameters (on the left) and with parameters optimized for self docking (right)**

PDB code	<i>Results with original parameters</i>					<i>Results with parameters optimized for cognate docking</i>				
	<i>ROC</i>	<i>% ROC (1%)</i>	<i>% ROC (2%)</i>	<i>% ROC (5%)</i>	<i>% ROC (20%)</i>	<i>ROC</i>	<i>% ROC (1%)</i>	<i>% ROC (2%)</i>	<i>% ROC (5%)</i>	<i>% ROC (20%)</i>
<b>1a7a</b>	0.30	3.00	3.00	3.00	3.00	0.95	7.00	11.00	18.00	33.00
<b>1agw</b>	0.55	0.00	1.70	4.20	11.90	0.63	0.00	2.00	4.00	33.00
<b>1b8o</b>	0.82	8.00	20.00	28.00	64.00	0.77	0.00	0.00	1.00	12.00
<b>1f0r</b>	0.67	2.80	3.50	10.60	45.10	0.62	6.00	7.00	11.00	36.00
<b>1hw8</b>	0.55	0.00	2.90	2.90	17.10	0.76	1.00	1.00	3.00	13.00
<b>1kim</b>	0.31	13.60	27.30	27.30	27.30	0.54	0.00	0.00	0.00	3.00
<b>1uy6</b>	0.46	0.00	0.00	0.00	16.70	0.39	0.00	0.00	0.00	0.00
<b>1xgj</b>	0.31	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00
<b>3ert</b>	0.58	0.00	0.00	2.60	23.10	0.68	2.00	2.00	5.00	20.00
<b>1m2z</b>	0.30	1.30	1.30	1.30	1.30	0.20	0.00	0.00	0.00	0.00
<b>mean</b>	0.49	2.87	5.97	7.99	20.95		0.57	1.60	2.30	4.20
<b>median</b>	0.51	0.65	2.30	2.95	16.90		0.63	0.00	0.50	2.00
<b>sd</b>	0.17	4.30	9.07	10.23	19.35		0.24	2.54	3.55	5.65

accurate; the high RMSD value is due to a different position of some ligand functional groups respect to those of the crystallized ligand, however this slightly different orientation is justified by matching a pharmacophoric feature not matched by the ligand in the crystallographic complex.

## 1.5.2 Virtual Screening Optimization

Virtual screening experiments are conceptually different from pose prediction experiments, though they are both based on the same molecular docking approach. With a retrospective virtual screening experiment we want to discriminate ligands that can bind to a receptor from the decoys that are expected not to bind. When the docking process is guided by pharmacophore models, as in the case of LiGenDock, an important issue is the strictness of the pharmacophore model. Indeed, on the one hand, very strict settings would lead to poor structural diversity in the compounds retrieved from VS, whereas on the other a very fuzzy pharmacophore is more likely to return a large number of false positives.<sup>60</sup> As previously done with the optimization of parameters involved in cognate docking, we sought a set of parameters, with optimized values for VS. Since a preliminary trial with parameters optimized for pose prediction gave modest results, performing slightly better than original parameters in terms of global enrichment but a little worse in case of early enrichment (Table 7), we performed a full factorial design (reported in Table V Appendix A) with parameters that were previously shown to be important: 1) grid accuracy (related with grid spacing), 2) the maximal number of pharmacophoric features, because these two parameters came out as the most important ones from the first part of this study, 3) the minimal distance between two features and 4) angle delta, the angle used in rotating ligand inside the binding site to match as many pharmacophoric features as possible. The maximum number of features was allowed to vary between 8 and 15, a range lower than the one used in the optimization of cognate docking; these values were chosen to avoid recognizing too many decoys as good binders; this choice was driven by the allowed partial ligand-pharmacophore match during virtual screening; in this sense a too large number of pharmacophoric features will have permitted to generate features also for less important characteristic of the binding site, or, for example, for the boundary regions of the pocket.

**Table 8. VS results before and after the optimization for all the DUD complexes, using the entire DUD dataset.**

	PDB code	ORIGINAL						AFTER OPTIMIZATION						
		ROC	ROC (1%)	ROC (2%)	ROC (5%)	ROC (20%)	BEDROC ( $\alpha=20$ )	ROC	ROC (1%)	ROC (2%)	ROC (5%)	ROC (20%)	BEDROC ( $\alpha=20$ )	
<b>serine</b>														
	<i>FXa</i>	1f0r	0.67	2.8	3.5	10.6	45.1	0.1	0.56	1.4	2.1	2.1	17.6	0.03
	<i>thrombin</i>	1ba8	0.63	0	0	6.2	27.7	0.06	0.67	1.5	1.5	4.6	35.4	0.07
	<i>trypsin</i>	1bjv	0.48	0	0	0	20.5	0.02	0.99	0	0	2.3	11.4	0.04
<b>kinase</b>														
	<i>FGFr1</i>	1agw	0.55	0	1.7	4.2	11.9	0.04	0.81	0.8	2.5	10.2	47.5	0.11
	<i>CDK2</i>	1ckp	0.83	0	6	6	24	0.06	0.39	2	2	2	8	0.02
	<i>EGFr</i>	1m17	0.64	0.2	0.5	1.1	32	0.04	0.95	2.3	5.6	41.7	87.4	0.32
	<i>HSP90</i>	1uy6	0.46	0	0	0	16.7	0.01	0.51	0	0	0	12.5	0.01
	<i>SRC</i>	2src	0.93	1.9	3.9	10.3	47.1	0.12	0.63	0	1.9	7.1	25.8	0.07
	<i>TK</i>	1kim	0.31	13.6	27.3	27.3	27.3	0.23	0.71	0	0	4.5	27.3	0.04
	<i>p38</i>	1kv2	0.8	5.1	10.2	13.7	30.9	0.13	0.81	12.1	21.1	37.1	62.9	0.3
<b>metalloenzym</b>														
	<i>ACE</i>	1o86	0.9	0	2	4.1	16.3	0.05	0.73	2	2	2	22.4	0.04
	<i>COMT</i>	1h1d	0.21	0	0	0	9.1	0.02	0.78	0	0	27.3	81.8	0.23
	<i>PDE5</i>	1xp0	0.56	0	3.9	5.9	21.6	0.06	0.63	3.9	5.9	17.6	31.4	0.12
<b>nuclear hormone</b>														
	<i>ERagonist</i>	1l2i	0.89	0	0	6	67.2	0.1	0.92	4.5	19.4	52.2	85.1	0.37
	<i>ERantagonis</i>	3ert	0.58	0	0	2.6	23.1	0.05	0.62	5.1	5.1	7.7	12.8	0.08
	<i>GR</i>	1m2z	0.3	1.3	1.3	1.3	1.3	0.01	0.97	41	52.6	57.7	79.5	0.53
	<i>MR</i>	2aa2	0.88	0	0	0	40	0.04	0.93	0	0	0	33.3	0.04
	<i>PPARg</i>	1fm9	0.64	0	0	0	6.2	0.01	0.62	0	0	0	3.7	0.01
	<i>PR</i>	1sr7	0	0	0	0	0	0	0.9	11.1	25.9	29.6	70.4	0.26



CHAPTER 1 Development and optimization of LiGen, a new drug design software

<i>RXR</i>	1mvc	0	0	0	0	0	0	0.99	60	65	70	75	0.65
<b>folate enzyme</b>													
<i>DHFR</i>	3dfr	0.27	0	0	0	5	0	0.48	0	0.5	1	12.9	0.02
<i>GART</i>	1c2t	0.14	0	0	0	0	0	0.68	4.8	4.8	14.3	28.6	0.1
<b>other enzyme</b>													
<i>COX-2</i>	1cx2	0.77	0.9	1.1	1.1	1.1	0.01	0.88	22.4	37.9	49.7	74.4	0.43
<i>PARP</i>	1efy	0.6	3	3	3	33.3	0.08	0.73	9.1	18.2	24.2	54.5	0.23
<i>AChE</i>	1eve	0.63	0	1	1	10.5	0.02	0.62	0	0	5.7	37.1	0.06
<i>HIVPR</i>	1hpx	0.67	7.5	9.4	13.2	37.7	0.14	0.62	3.8	3.8	11.3	30.2	0.09
<i>HMGR</i>	1hw8	0.55	0	2.9	2.9	17.1	0.03	0.96	0	0	8.6	20	0.06
<i>InhA</i>	1p44	0.63	1.2	2.4	10.6	34.1	0.09	0.65	3.5	5.9	11.8	31.8	0.11
<i>COX-1</i>	1p4g	0.66	0	4	4	4	0.03	0.8	0	4	4	32	0.05
<i>HIVRT</i>	1rt1	0.54	0	0	2.5	10	0.03	0.64	0	0	2.5	27.5	0.04
<i>AmpC</i>	1xgj	0.31	0	0	0	0	0	0.37	0	0	0	0	0
<i>SAHH</i>	1a7a	0.3	3	3	3	3	0.03	0.98	45.5	66.7	93.9	97	0.73
<i>GPB</i>	1a8i	0.57	7.7	11.5	11.5	11.5	0.1	0.92	0	3.8	21.2	92.3	0.24
<i>ALR2</i>	1ah3	0.59	0	0	3.8	23.1	0.03	0.56	0	0	3.8	7.7	0.02
<i>PNP</i>	1b8o	0.82	8	20	28	64	0.26	0.85	0	0	28	72	0.22
<i>NA</i>	1a4g	0.98	4.1	6.1	18.4	38.8	0.14	0.9	2	12.2	36.7	83.7	0.31
<b>mean</b>		0.54	1.3	2.88	4.27	18.29	0.05	0.73	8.07	12.1	20.47	41.82	0.18
<b>median</b>		0.58	0	1.1	2.6	16.3	0.03	0.73	2	3.8	10.2	31.8	0.09
<b>sd</b>		0.26	2.9	5.33	5.87	16.59	0.05	0.17	14.91	19.17	23.9	28.01	0.2

Thus, in principle, some decoys matching only unimportant, or low-important, features could have been retrieved. We decided to include also the minimum distance between the features, to assure the pharmacophore resembles in an accurate way the binding site. Angle delta was included in the design because we wanted to exclude the possibility that lack of recognition of some active compounds (as seen with the original set of parameters) was due to a bad (not fine enough) ligand-pharmacophore match. Parameters not under investigation were set to the optimal values found before. Since the number of experiments to run with all the complexes of the DUD database would have required too much computational time, we randomly selected ten complexes, (underlines in Table 3) to run the optimization procedure, whereas for the evaluation of the improvements resulting from the optimization procedure, the optimized performance was assessed using all the 36 selected targets. Results of LiGen's virtual screening performance, reported in Table 8 and Table VI Appendix A and summarized in Table 9, with the original set of parameters were modest, especially regarding the early enrichment: the average ROC(1%) and ROC(5%) were respectively 1.30 % and 4.27%. Also the global AUC presents a mean value just above random (0.54 with respect to 0.50 for random performance), with several structures showing very low ligand recognition and in two cases actives were completely discarded (PDB code 1sr7 and 1mvc). Parameters and results of the 16 experiments of the full factorial design are listed in Table V Appendix A. Average ROC values have previously been used to assess virtual screening performance using the DUD database.<sup>57, 98</sup> Thus, the average ROC(1%) was used to fit the design and design statistics are given in Table 10. A deep analysis of the results showed that the best results both in terms of global and early enrichment were obtained with parameters of experiment number 1. The high standard deviation is due to the very low values of some virtual screening experiments. In particular the AmpC  $\beta$ -lactamase (AmpC, PDB code: 1xgj) , the thymidine kinase (TK, PDB code: 1kim) and the Human Heat Shock Potein 90 (HSP90, PDB code: 1uy6) showed early enrichment (ROC(1%) and ROC(2%)) almost always equal to zero, regardless of the values assigned to parameters. Very poor enrichment for these structures was already reported in literature in a comparable experiment by Repasky and Murphy.<sup>94</sup> Virtual screening experiments using parameters of the first experiment of the full factorial design were run also for the other structures selected from the DUD database against all the decoys and against self-decoys (only decoys with scaffolds similar to active ligands), and the results

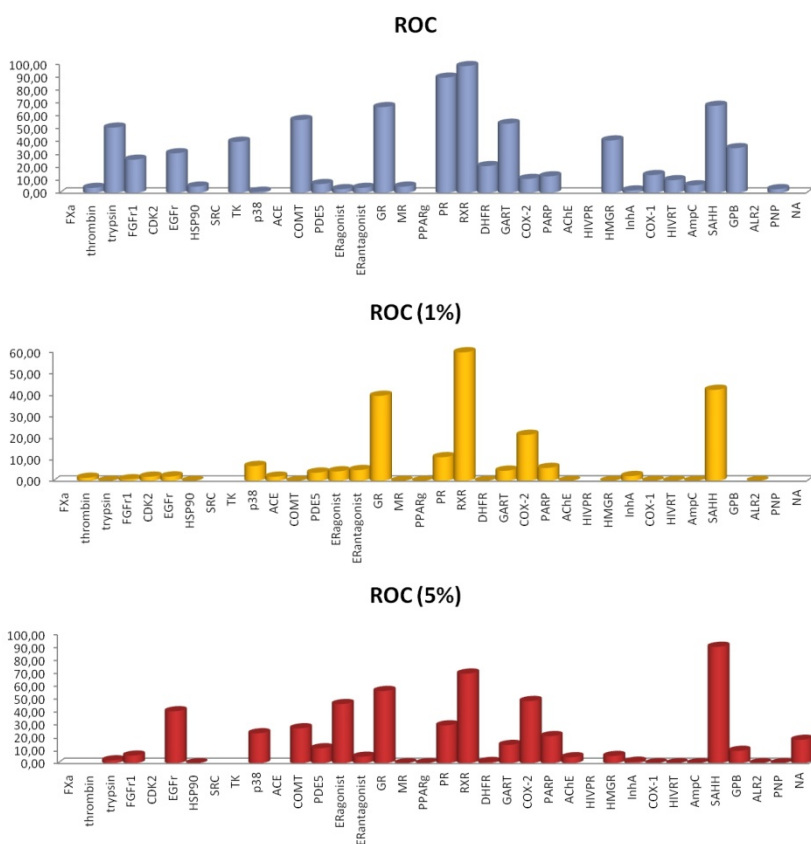
**Table 9. Comparison of VS results considering, in the first part, the entire DUD database and, in the second part of the table, the “own decoys” subset of DUD. Results are reported for the original set of parameters (on the left side of the table) and for the optimized ones (on the right)**

<i>original</i>						
	<i>ROC</i>	<i>ROC (1%)</i>	<i>ROC (2%)</i>	<i>ROC (5%)</i>	<i>ROC (20%)</i>	<i>BEDROC <math>\alpha=20.0</math></i>
<b><i>entire DUD</i></b>						
mean	0.54	1.30	2.88	4.27	18.29	0.05
median	0.58	0.00	1.10	2.60	16.30	0.03
sd	0.26	2.90	5.33	5.87	16.59	0.05
<b><i>self-decoys</i></b>						
mean	0.56	0.92	1.85	4.8	22.38	0.06
mediana	0.61	0.00	0.75	3.00	22.80	0.05
sd	0.24	1.45	2.36	5.15	16.84	0.07
<b>after optimization</b>						
	<i>ROC</i>	<i>ROC (1%)</i>	<i>ROC (2%)</i>	<i>ROC (5%)</i>	<i>ROC (20%)</i>	<i>BEDROC (<math>\alpha=20.0</math>)</i>
<b><i>entire DUD</i></b>						
mean	0.73	8.07	12.10	20.47	41.82	0.18
median	0.73	2.00	3.80	10.20	31.80	0.09
sd	0.17	14.91	19.17	23.90	28.01	0.20
<b><i>self-decoys</i></b>						
mean	0.71	5.94	9.58	16.75	39.07	0.16
mediana	0.69	1.70	3.05	8.05	33.95	0.12
sd	0.16	11.27	16.73	21.6	25.18	0.15

**Table 10. Virtual Screening design statistics**

<i>VIRTUAL SCREENING</i>			
	<b>R<sup>2</sup></b>	<b>R<sup>2</sup> adjust</b>	<b>F</b>
<b>Design 3</b>	0.7929	0.7176	10.5281
<b>Parameters</b>	<b>p value</b>		
Grid accuracy	0.0397		
Maximal features number	0.2796		
Minimal features distance	0.0002		
Angle delta	0.0371		

are reported in Table 7 and Table VI of Appendix A, respectively. The optimized set of parameters improved results for almost half of the structures of the dataset, both for early and global enrichment, as shown in Figure 14. Poor early enrichment are found in case of some kinases, like TK (PDB code 1kim), and HSP90 (PDB code 1uy6). In particular TK is reported to be a challenging target due to receptor flexibility, solvent exposed binding site and the importance of water bridge interactions,<sup>62</sup> not taken into account in our experiments.



**Figure 14. Histogram plot to show the improvements of the ROC, ROC(1%) and ROC(5%) values for all the DUD complexes considered in the study. The bars represent the percentage of improvement gained through the optimization procedure. Complexes with missing bars are those for which no improvement was registered.**

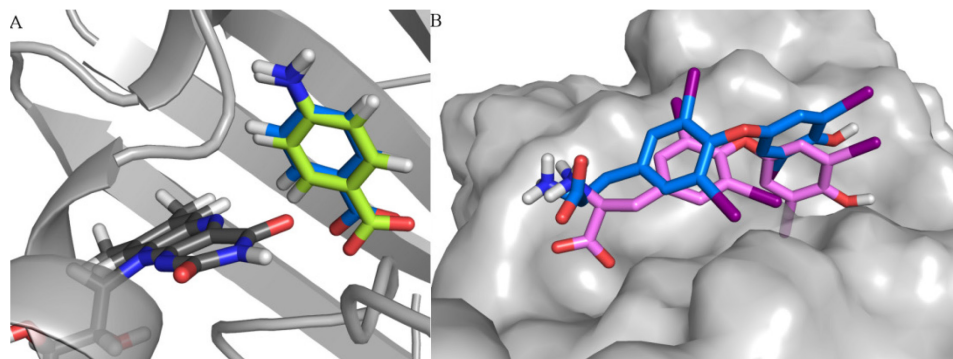
When looking for new inhibitors using VS experiments, one important aspect concerns the structural diversity of the possible new active ligands. VS hits sharing all the same scaffold are not particularly useful for drug discovery purposes, to identify new leads to develop new chemical series. To address LiGen ability of recognizing different chemotypes, we analyzed VS results also from this aspect. For each target, the active ligands were clustered using TGT fingerprints. Then the number of recognized clusters has been calculated for the original performance and for the optimized parameters. The number of recognized chemotypes in the first 1% and 2% of the ROC curve is generally quite low if compared to the total number of chemotypes (Table 11). However a slightly improvement is registered with the optimized parameters. Notably in the case of COX2 and EGFR ligands a great improvement from the original results was found. The not outstanding performance in terms of chemotypes has probably its main reason in the docking protocol and in the scoring process. LiGen does not consider receptor flexibility, whereas probably to recognize active ligands with different scaffold a certain degree of pocket flexibility is required. In fact rearrangement of the side chains can allow a better accommodation of ligands with diverse scaffolds inside the pocket. therefore with the actual algorithm, that ignores the receptor flexibility, ligands, even if are recognized by the pharmacophore, are badly scored by the scoring function due to missing interaction with some of the key residues in the pocket. Nevertheless this problem is not only a LiGen problem but is common to all software packages not considering target flexibility while performing VS experiments. In fact taking into account receptor flexibility during VS terribly increases the calculation time, so that it could not be used to screen large compounds library, that is instead its main purpose.

**Table 11. Number of chemotypes recognized at 1% and 2% of the ROC curve during VS experiments with original and optimized parameters.**

PDB	Numer of ligands	Number of chemotypes	ORIGINAL		OPTIMIZED		
			Chemot. ROC1%	Chemot ROC 2%	Chemot. ROC1%	Chemot ROC 2%	
<b>serine proteasi</b>							
<i>FXa</i>	1f0r	146	67	2	4	2	3
<i>thrombin</i>	1ba8	72	17	0	1	1	1
<i>trypsin</i>	1bjv	49	11	0	0	2	2
<b>kinase</b>							
<i>FGFr1</i>	1agw	120	46	0	3	2	3
<i>CDK2</i>	1ckp	72	38	0	0	1	1
<i>EGFr</i>	1m17	475	180	2	2	7	16
<i>HSP90</i>	1uy6	37	8	0	0	0	0
<i>SRC</i>	2src	159	68	3	5	2	2
<i>TK</i>	1kim	22	9	0	0	0	0
<i>p38</i>	1kv2	454	116	11	19	19	26
<b>metalloenzyme</b>							
<i>ACE</i>	1o86	49	21	1	2	1	1
<i>COMT</i>	1h1d	11	9	0	0	0	0
<i>PDE5</i>	1xp0	88	49	0	2	2	3
<b>nuclear hormone receptor</b>							
<i>ERagonist</i>	1l2i	67	24	0	0	3	5
<i>ERantagonist</i>	3ert	39	9	0	0	2	2
<i>GR</i>	1m2z	78	17	1	1	7	9
<i>MR</i>	2aa2	15	8	0	0	2	4
<i>PPARg</i>	1fm9	85	9	0	0	0	0
<i>PR</i>	1sr7	27	7	0	0	3	4
<i>RXR</i>	1mvc	20	5	0	0	4	4
<b>folate enzyme</b>							
<i>DHFR</i>	3dfr	410	173	0	0	0	1
<i>GART</i>	1c2t	40	3	0	0	0	0
<b>other enzyme</b>							
<i>COX-2</i>	1cx2	349	60	2	3	12	18
<i>PARP</i>	1efy	35	17	1	1	2	3
<i>AChE</i>	1eve	107	33	1	1	1	2
<i>HIVPR</i>	1hpx	62	21	2	3	2	2
<i>HMGR</i>	1hw8	35	10	1	1	0	0
<i>InhA</i>	1p44	86	28	0	2	3	3
<i>COX-1</i>	1p4g	25	12	0	1	1	1
<i>HIVRT</i>	1rt1	43	31	0	0	0	2
<i>AmpC</i>	1xgj	21	7	0	0	0	0
<i>SAHH</i>	1a7a	33	8	1	1	3	4
<i>GPB</i>	1a8i	52	30	0	1	0	2
<i>ALR2</i>	1ah3	26	16	1	1	1	1
<i>PNP</i>	1b8o	50	14	1	1	0	0
<i>NA</i>	1a4g	49	19	3	3	1	3
			<b>ROC 1%</b>	<b>ROC 2%</b>			
<b>Improved with optimized parameters</b>			15	19			
<b>Not changed with optimized parameters</b>			18	11			
<b>Worsened with optimized parameters</b>			3	6			

### 1.5.3 Pose Prediction Validation

To further test the performance of LiGenDock with an optimized set of parameters, we decided to validate on a different dataset the set of parameters from experiment number 28 of the RSM (Table III, Appendix A). The choice of this set of parameters was due to the following reasons: i) it is one of the few sets of parameters for which we were able to obtain poses for all the complexes, thus indicating it is suitable for complexes having different characteristics, for example it is good both for very small (1pbd, ligand MW 137.38, RMSD 0.36) and big ligand size (1eta, ligand MW 776.87, RMSD 2.33) (Figure 15); ii) with this set we were able to obtain the highest number of poses with RMSD less than 3 Å (70%). The number of poses with RMSD less than 2 Å was a little smaller than with other best performing set of parameters, but visual inspection of results suggested that the small differences among best performing experiments in terms of RMSD < 2 Å and < 3 Å should not be emphasized too much, and the ability of producing an higher number of poses was the consideration that guided our choice. The validation was carried out by using the CORE PDBbind database as an external dataset (not used during the optimization study), excluding those complexes with a ligand molecule with molecular weight higher than 500. The same test set was also used to perform docking with two other docking programs, namely Glide and AutoDock. Results are detailed in Table IV of Appendix A and summarized in Table 12. Poses predicted within a RMSD range of 3 Å from the crystallographic pose are 85%, more than AutoDock and only one point less than Glide. As expected the number of poses with a RMSD within 2 Å is smaller for LiGen with respect to the others (55.3%, LiGen, 64.1% AutoDock, and 75.3% Glide, respectively). The same trend of results can be observed if we consider the best scoring pose. However it is worth mentioning that the best pose corresponds to the best scoring pose in 26% of the experiments in the case of LiGenDock, in the 30% of the experiments performed with Glide but only in 12% in the case of AutoDock. Analysis of the poses with RMSD higher than 3 Å from the experimentally solved ligand revealed that in many cases they involve ligands with more than 10 rotatable bonds, as in the cases of 1b11 and 1gni (Figure 16A); in other cases LiGen failed in defining a good pharmacophore, especially for those proteins presenting a solvent exposed binding site, as for example in the case of 1nhu, 1tyr, 1v2o and 2g8r (Figure 16B); other badly predicted



**Figure 15** Parameters of experiment 28 of RSM allows a good prediction of ligand binding conformation with different types of ligand. Here we reported two examples, one (A) of a very small ligand (PDBcode:1pbd, crystallographic ligand pose in blue, cofactor FAD in gray) and one (B) of a ligand with drug-like size. (PDBcode:1eta,crystallographic ligand pose in blue)

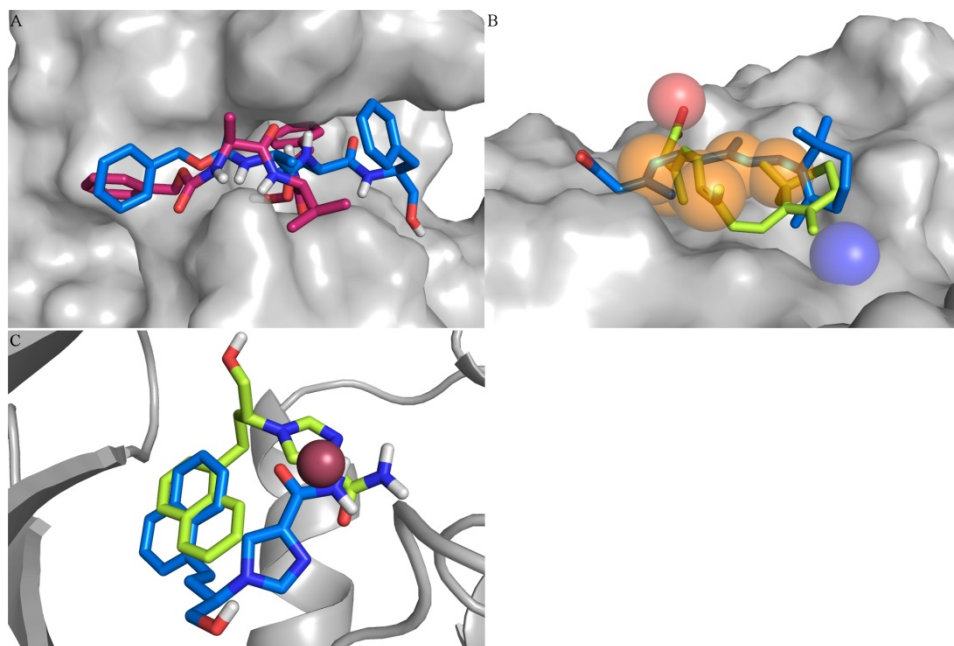
**Table 12.** Comparison of LiGen, AutoDock and Glide results using the PDBbind dataset

	best predicted pose			best scoring pose		
	LiGen	AutoDock	Glide	LiGen	AutoDock	Glide
<i>n</i> results	170	170	170	170	170	170
<i>n</i> RMSD <2 Å	95	109	128	56	69	98
<i>n</i> RMSD <3 Å	145	139	146	96	96	117
% rmds < 2	54.7	64.1	75.3	32.9	40.6	57.6
% rmds < 3	84.7	81.8	85.9	56.5	56.5	68.8
	LiGen	AutoDock	Glide			
% (number) best predicted pose = best scoring pose	26 (44)	% 12% (21)	30% (51)			
Pocket time (s)	5.46	9.83	161.42			
Dock time (s)	20.30	397.08	18.24			
Total time (s)	25.76	407.63	179.66			

poses are found for Zinc dependent metalloproteins, as 1ndy, 1zs0, 1zvx or 8cpa, indicating that some improvement must be introduced for scoring ligand-metal interactions (Figure 16C).

From the comparison reported in Table 12, it is clear that Glide performs better than LiGen and AutoDock; the LiGen performance is very similar to AutoDock in terms of poses predicted within an RMSD of 3Å. It should be noted, however, another interesting aspect coming out from Table 12, i.e. the difference on the average speed of the three programs. This is not particularly relevant for pose prediction, but can be for virtual





**Figure 16. Examples of problems found in LiGenDock validation using complexes from PDBbind. A) Ligand with more than 10 rotatable bonds (PDBcode:1b11, crystallographic ligand pose in blue). B) Solvent exposed binding site (PDBcode:1tyr, crystallographic ligand pose in blue). C) Interaction with Zn (PDBcode:1ndy, crystallographic ligand pose in blue)**

screening, and makes LiGenDock attractive for this kind of application. In particular, LiGen requires less than 30 seconds on average to produce ten poses, whereas Glide needs roughly 3.5 min on average and AutoDock about 7 min on average.

Small differences in times and RMSD values with respect to those reported elsewhere in the literature<sup>94, 99</sup> are due to different starting conformations in our test conditions with respect to the ones used by others. Given the good performance in predicting poses with RMSD less than 3Å in very short time, this parameters set used during self docking validation was also chosen as default setting for routine pose prediction experiments.

## 1.5.4 Virtual Screening Validation

To complete LiGenDock validation, we report in Table 13 the comparison of the VS performances between LiGen, Glide and DOCK6, using data published in other papers, appeared when this study was being performed, for the other two programs.<sup>94, 98</sup> As shown in Table 13, the global enrichment of the three programs is good, performing all better than random. Glide is the best one, having the average and the median AUC value of 0.80 and 0.82 respectively. LiGenDock, with average and median AUC values of 0.73 in both cases, performs slightly worse than Glide, but better than DOCK6, which average and median AUC are 0.60 and 0.56. The last two columns of Table 13 report the average percentage of ROC(1%) and ROC(2%). As it is evident from this table, Glide is the best performing program. LiGenDock performance is better than DOCK6 in the first 1% of the screened database (ROC(1%) value), and definitely better than random performance (random performance at 1% of the screened database is 0.5%), confirming the goodness of LiGen approach also for VS. Notably, even though also after the optimization the average early enrichment (ROC(1%)) is not outstanding, there was a significant improvement with respect to the original set of parameters, corroborating the application of the optimization protocol. A large enhancement was also registered for global enrichment (AUC) for ROC(5%). For us, values of the early enrichment metrics were not a surprising outcome. LiGen docking process is driven by the pharmacophore generated inside the binding site, so the suggested binding pose should not be seen as a results of an extensive conformational search for the global energy minimum, but as a result of a reduced/constrained conformational search to match the highest number of pharmacophoric features. In this sense a higher number of decoys can be recognized during the docking/virtual screening process. Moreover, in a real application of virtual screening this can be an advantage since it allows for the recognition of possibly new and diverse scaffolds. An example of poor benchmarking results but good outcomes in a real life application can be found in the work of Löwer et al. <sup>100</sup> The LiGen docking algorithm has been also developed for use in fragment docking in *de novo* design, so the recognition of new scaffolds is of primary importance, definitely more than high enrichment in recognition of already know ligands.

**Table 13. Comparison between LiGen, Glide and DOCK6 VS results. With ROC is indicated the AUC of the ROC curve, SD is the standard deviation, Max is the highest ROC value and Min is the lowest ROC value found.**

	<i>ROC</i>	<i>SD</i>	<i>Median</i>	<i>Max</i>	<i>Min</i>	<i>ROC(1%)</i>	<i>ROC(2%)</i>
LiGen	0.73	0.17	0.73	0.99	0.37	8.07	12.10
Glide	0.80	0.14	0.82	0.98	0.42	25.18	33.64
DOCK6	0.60	0.17	0.56	0.96	0.29	4.99	20.19

The high speed of LiGen's code allows Virtual Screening with the entire DUD database in less than 105 hours, that means about 3s for each ligand. This represents a very good results in term of times needed to screen large databases, since for example for Glide, are reported times of 10s per ligand.<sup>94</sup>

## 1.6 CONCLUSIONS

The primary goal of this project was the optimization of LiGen docking performance using a procedure based on experimental designs. LiGen is a *de novo* design suite of programs, presented by Beccari et al.,<sup>72</sup> consisting of a set of modules: LiGenPass for binding site recognition, LiGenPocket for binding site analysis and structure-based pharmacophore definition, LiGenDock for docking and virtual screening and LiGenBuilder for *de novo* design. In this study we focused on LiGenPocket and LiGenDock, which constitute the docking engine of the program. A number of parameters controlling the docking procedure were varied according to statistical experimental designs. First, the most influential parameters were identified through a fractional factorial design, yielding parameter sets that covered the selected interval of parameter values. The parameter sets thus designed were then applied in a docking study using a set of 100 protein-ligand complexes taken from the CCDC Astex dataset. The number of poses presenting an RMSD less than 2 Å between the best predicted docking poses and the corresponding crystallographic ligands was considered as response for fitting the design. A significant regression model between the docking runs using the designed parameter sets and the docking results (number of poses with RMSD less than 2 Å) was established, thus shedding light on the parameters with large influence on docking results. The most relevant parameters were the minimum distance between two pharmacophoric features (minimum feature distance), the maximum number of features identified in the binding site (maximum feature number), the degree of smoothing of the protein's Van der Waals volume (Van der Waals bumps), the grid spacing (grid accuracy), the tolerance in considering a ligand functional group superposed on the pharmacophoric feature (distance threshold) and the threshold indicating which pharmacophoric points should be considered during docking (neighbor threshold). Furthermore a response surface model was developed using these parameters to find the optimal parameters' set. As shown in Table 3, with the optimized set of parameters we obtained a number of poses with RMSD less than 2 Å almost

seven-times higher compared to the original set (41% of the optimized set with respect to the 6% of the original parameters' set). This gain in the accuracy of pose prediction was not followed by an increment in time consumed by the docking process, since the difference of the average time spent is less than 1 second. It should be noticed that the LiGenDock algorithm has been originally derived within a *de novo* design suite of programs, where the main objective is the identification of novel chemotypes able to interact with the partner macromolecule. Thus the LiGen's pharmacophore based approach partially suffers in terms of precision in exactly reproducing experimental binding poses, although a deep visual inspection of the results suggests that poses within 3Å from the co-crystallized ligands are still quite accurate. Moreover the comparison between LiGen docking results with AutoDock and Glide using a dataset extracted from the PDBbind database, confirms the quality of LiGen approach, even though Glide was the best performing program. As reported in Table 7, the number of LiGen's predicted poses within 3 Å from the co-crystallized ligand is similar to those predicted by Glide and a little better than by AutoDock (poses within 3Å from the crystallized ligands: LiGen 84.7%, AutoDock 81,%, Glide 85.9%).

Investigation of the influence of parameters on the VS results was also performed using experimental design, to find an optimal parameters set for virtual screening experiments. Global enrichment, represented by the mean ROC values of 0.73 after the optimization procedure, is consistent with values obtained with other software and reported in literature.<sup>57, 94, 98</sup> and also summarized in Table 11. The not particularly excellent performance in terms of early enrichment should not be considered as a negative result, since in other cases reported in literature the DUD dataset demonstrated to be very challenging.<sup>94, 98</sup> Moreover there are already examples in literature of pharmacophore-based virtual screening studies with modest benchmarking results but good outcomes in real life applications, as shown by the paper by Löwer and coworkers<sup>100</sup>. Furthermore the high speed reached, screening the entire DUD database in about 105 hours, makes LiGen very attractive for virtual screening applications. It should be commented that simultaneous optimization of both virtual screening and pose prediction performance would mean to carry out a two-properties optimization, possibly through the definition of a desirability function. This can be done, but we can anticipated that the results cannot be better than those described in the paper. The resulting set of parameters might possibly be seen as 'general purposes' parameters, but our data, clearly

indicate that the optimization of the two properties diverges, so that optimization of an 'averaged' desirability function must necessarily afford 'averaged' results.

Finally, the results obtained with the optimization of LiGen highlighted the usefulness of experimental designs for optimization purposes also in the field of computational drug discovery.

## BIBLIOGRAPHY

1. Dickson, M.; Gagnon, J. P., Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discovery* **2004**, *3*, 417-429.
2. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935-949.
3. Jain, A. N., Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput. Aided Mol. Des.* **2008**, *22*.
4. Brooijmans, N.; Kuntz, I. D., Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.
5. Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409-443.
6. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K., Gaussian docking functions. *Biopolymers* **2003**, *68*, 76-90.
7. Shoichet, B. K.; Kuntz, I. D., Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*.
8. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*.
9. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking: Current status and future challenges. *Proteins-Structure Function and Bioinformatics* **2006**, *65*, 15-26.
10. McMartin, C.; Bohacek, R. S., QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput. Aided Mol. Des.* **1997**, *11*, 333-344.
11. Stouten, P. F. W.; Kroemer, R. T., Core Concepts and Methods - Target Structure based - Docking and Scoring. In *Comprehensive Medicinal Chemistry II*, Taylor, J. B.; Triggle, D. J., Eds. Elsevier: Oxford, UK, 2006; Vol. 4, pp 255-281.
12. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
13. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.* **2003**, *17*, 755-763.
14. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
15. Jain, A. N., Surfex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499-511.
16. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P., eHITS: An innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006**, *7*, 421-435.
17. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM - a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*.
18. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.

19. Dias, R.; de Azevedo, W. F., Molecular Docking Algorithms. *Curr. Drug Targets* **2008**, *9*, 1040-1047.
20. Wang, R.; Liu, L.; Lai, L. H.; Tang, Y. Q., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **1998**, *4*.
21. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.
22. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScoreCSD-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296-6303.
23. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337-356.
24. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins: Structure, Function and Genetics* **2003**, *52*, 609-623.
25. Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912-5931.
26. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E., Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851-5855.
27. Tirado-Rives, J.; Jorgensen, W. L., Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880-5884.
28. Perola, E.; Walters, W. P.; Charifson, P. S., A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins-Structure Function and Bioinformatics* **2004**, *56*, 235-249.
29. Wang, J. M.; Hou, T.; Xu, X., Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Current Computer-Aided Drug Design* **2006**, *2*, 287-306.
30. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham Iii, T. E., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.
31. Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51*, 69-82.
32. Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D., Flexibases: A way to enhance the use of molecular docking methods. *J. Comput. Aided Mol. Des.* **1994**, *8*, 565-582.
33. Peng, H.; Huang, N.; Qi, J.; Xie, P.; Xu, C.; Wang, J.; Yang, C., Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3693-3699.
34. Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P., Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46*, 2656-2662.
35. Schneider, G.; Fechner, U., Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649-663.
36. Loving, K.; Alberts, I.; Sherman, W., Computational Approaches for Fragment-Based and *De novo* Design. *Curr. Top. Med. Chem.* **2010**, *10*, 14-32.
37. Bohm, H. J., The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **1992**, *6*, 61-78.
38. Honma, T.; Hayashi, K.; Aoyama, T.; Hashimoto, N.; Machida, T.; Fukasawa, K.; Iwama, T.; Ikeura, C.; Ikuta, M.; Suzuki-Takahashi, I.; Iwasawa, Y.; Hayama, T.; Nishimura, S.;



Morishima, H., Structure-based generation of a new class of potent Cdk4 inhibitors: new *de novo* design strategy and library design. *J. Med. Chem.* **2001**, *44*, 4615-27.

39. Tan, J. J.; Zhang, B.; Cong, X. J.; Yang, L. F.; Liu, B.; Kong, R.; Kui, Z. Y.; Wang, C. X.; Hu, L. M., Computer-aided design, synthesis, and biological activity evaluation of potent fusion inhibitors targeting HIV-1 gp41. *Med. Chem.* **2011**, *7*, 309-16.

40. Wang, R.; Gao, Y.; Lai, L. H., LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.* **2000**, *6*.

41. Yuan, Y.; Pei, J.; Lai, L., LigBuilder 2: A Practical *de novo* Drug Design Approach. *J. Chem. Inf. Model.* **2011**, *51*, 1083-1091.

42. Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P., SPROUT: recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Sci* **1994**, *34*, 207-17.

43. Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E., HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins* **1994**, *19*, 199-221.

44. Westhead, D. R.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B., PRO-LIGAND: an approach to *de novo* molecular design. 3. A genetic algorithm for structure refinement. *J. Comput. Aided Mol. Des.* **1995**, *9*, 139-48.

45. Waszkowycz, B.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Westhead, D. R., PRO\_LIGAND: an approach to *de novo* molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. *J. Med. Chem.* **1994**, *37*, 3994-4002.

46. Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R., PRO-LIGAND: an approach to *de novo* molecular design. 1. Application to the design of organic molecules. *J. Comput. Aided Mol. Des.* **1995**, *9*, 13-32.

47. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G., DOGS: reaction-driven *de novo* design of bioactive compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.

48. Ji, H., Fragment-Based Drug Design: Considerations for Good ADME Properties. In *ADMET for Medicinal Chemists: A Practical Guide*, first ed.; Tsaion, K.; Kates, S. A., Eds. John Wiley & Sons, Inc.: Hoboken, New Jersey, 2011; pp 417-485.

49. Kutchukian, P. S.; Shakhnovich, E. I., *De novo* design: balancing novelty and confined chemical space. *Expert Opinion on Drug Discovery* **2010**, *5*, 789-812.

50. Roe, D. C., Computer-Aided Molecular Design: *De novo* Design. In *Handbook of Chemoinformatics Algorithms*, first ed.; Faulon, J.-L.; Bender, A., Eds. Chapman and Hall/CRC Taylor & Francis Group: Boca Raton, Florida, 2010; pp 295-315.

51. Proschak, E.; Tanrikulu, Y.; Schneider, G., Fragment-based *De novo* Design of Drug-like Molecules. In *Chemoinformatics Approaches to Virtual Screening*, first ed.; Varnek, A.; Tropsha, A., Eds. The Royal Society of Chemistry: Cambridge, UK, 2008; Vol. 0, pp 217-239.

52. Jia, Y.; Chiu, T. L.; Amin, E. A.; Polunovsky, V.; Bitterman, P. B.; Wagner, C. R., Design, synthesis and evaluation of analogs of initiation factor 4E (eIF4E) cap-binding antagonist Bn7-GMP. *Eur. J. Med. Chem.* **2010**, *45*, 1304-13.

53. Hartenfeller, M.; Schneider, G., *De novo* drug design. *Methods in molecular biology (Clifton, N.J.)* **2011**, *672*, 299-323.

54. Wang, R.; Lai, L. H.; Wang, S. M., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* **2002**, *16*.

55. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A., SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765-73.

56. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S., Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558-565.

57. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C., Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455-1474.
58. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Comparing protein-ligand docking programs is difficult. *Proteins: Structure, Function and Genetics* **2005**, *60*, 325-332.
59. Jain, A. N.; Nicholls, A., Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.* **2008**, *22*.
60. Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K., Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*.
61. Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A., How to do an evaluation: Pitfalls and traps. *J. Comput. Aided Mol. Des.* **2008**, *22*, 179-190.
62. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*.
63. Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R., A new test set for validating predictions of protein-ligand interaction. *Proteins: Struct., Funct., Bioinf.* **2002**, *49*, 457-471.
64. Wang, R.; Fang, X. L., Y.; Yang, C.-Y.; Wang, S., The PDBbind Database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111-4119.
65. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T., Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes? *J. Comput. Aided Mol. Des.* **2008**, *22*, 213-228.
66. Bohari, M. H.; Sastry, G. N., FDA approved drugs complexed to their targets: evaluating pose prediction accuracy of docking protocols. *J. Mol. Model.* **2012**, *18*.
67. Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K., Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742-755.
68. Cleves, A. E.; Jain, A. N., Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput. Aided Mol. Des.* **2008**, *22*.
69. Truchon, J.-F.; Bayly, C. I., Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*.
70. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O., Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*.
71. Corbeil, C. R.; Williams, C. I.; Labute, P., Variability in docking success rates due to dataset preparation. *J. Comput. Aided Mol. Des.* **2012**, *26*.
72. Beccari, A. R.; Cavazzoni, C.; Beato, C.; Costantino, G., LiGen: A high performance workflow for chemistry driven *de novo* design. *J. Chem. Inf. Model.* **2013**, *53*, 1518-1527.
73. Brady, G. P., Jr.; Stouten, P. F., Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **2000**, *14*, 383-401.
74. Antes, I.; Merkwirth, C.; Lengauer, T., POEM: Parameter optimization using ensemble methods: Application to target specific scoring functions. *J. Chem. Inf. Model.* **2005**, *45*, 1291-1302.
75. Andersson, C. D.; Thysell, E.; Lindstrom, A.; Bylesjo, M.; Raubacher, F.; Linusson, A., A multivariate approach to investigate docking parameters' effects on docking performance. *J. Chem. Inf. Model.* **2007**, *47*, 1673-1687.
76. Andersson, C. D.; Chen, B. Y.; Linusson, A., Multivariate assessment of virtual screening experiments. *J. Chemom.* **2010**, *24*, 757-767.
77. Protein Data Bank (PDB). <http://www.rcsb.org/pdb/home/home.do>

78. Wang, R.; Fang, X. L., 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.
79. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*.
80. *Protein Preparation Wizard; Epik version 2.0; Impact version 5.5; Prime version 2.1.*, Schrödinger, LLC, New York, NY, 2009.
81. *Maestro 9.1*, Schrödinger, LLC; New York, NY, 2010.
82. Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C., ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534-546.
83. *ConfGen, version 2.1*, Schrödinger, LLC, New York, NY, 2009.
84. Process Improvement. In *NIST/SEMATECH e-Handbook of Statistical Methods*, National Institute of Standards and Technology (NIST).
85. Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nystrom, A.; Pettersen, J.; Bergman, R., Experimental design and optimization. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3-40.
86. Box, G. E. P.; Hunter, J. S.; Hunter, W. G., In *Statistics for Experimenters: Design, Innovation, and Discovery, Second edition*, Second ed.; John Wiley & Sons, Inc. : Hoboken, New Jersey, 2005.
87. Box, G. E. P.; Hunter, J. S.; Hunter, W. G., Factorial Designs at Two Levels: Advantages of Experimental Design; Fraction Factorial Designs: Economy in Experimentation. In *Statistics for Experimenters: Design, Innovation, and Discovery, Second edition*, John Wiley & Sons, Inc. : Hoboken, New Jersey, 2005; pp 173-279.
88. Box, G. E. P.; Hunter, J. S.; Hunter, W. G., Modelling Relationships, Sequential Assembly: Basics for Response Surface Methods. In *Statistics for Experimenters: Design, Innovation, and Discovery, Second edition*, John Wiley & Sons, Inc. : Hoboken, New Jersey, 2005; pp 437-487.
89. Bezerra, M. A.; Santelli, R. E.; Oliveira, E. P.; Villar, L. S.; Escaleira, L. A., Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta* **2008**, *76*.
90. Box, G. E. P.; Hunter, J. S.; Hunter, W. G., Basics: Probability, Parameters and Statistics. In *Statistics for Experimenters: Design, Innovation, and Discovery, Second edition*, Second ed.; John Wiley & Sons, Inc. : Hoboken, New Jersey, 2005; pp 17-65.
91. *MATLAB and Statistics Toolbox Release 2010*, MathWorks: Natick, Massachusetts, United States.
92. Onodera, K.; Satou, K.; Hirota, H., Evaluations of molecular docking programs for virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1609-1618.
93. Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R., Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J. Comput. Chem.* **2010**, *31*, 2109-2125.
94. Repasky, M. P.; Murphy, R. B.; Banks, J. L.; Greenwood, J. R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R. A., Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided Mol. Des.* **2012**, *26*.
95. Ewing, T.; Baber, J. C.; Feher, M., Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2423-2431.
96. Kirkman, T. W. Statistics to Use. <http://www.physics.csbsju.edu/stats/> (January 2012),
97. Massey, F. J., The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68-78.
98. Brozell, S. R.; Mukherjee, S.; Balius, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C., Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided Mol. Des.* **2012**, *26*.

CHAPTER 1 Development and optimization of LiGen, a new drug design software

99. Trott, O.; Olson, A. J., Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*.
100. Loewer, M.; Geppert, T.; Schneider, P.; Hoy, B.; Wessler, S.; Schneider, G., Inhibitors of Helicobacter pylori Protease HtrA Found by 'Virtual Ligand' Screening Combat Bacterial Invasion of Epithelia. *PLoS One* **2011**, *6*.

## **CHAPTER 2:**

# **Elucidation of the binding mode of a series of 3-HAO inhibitors**



## 2.1 INTRODUCTION

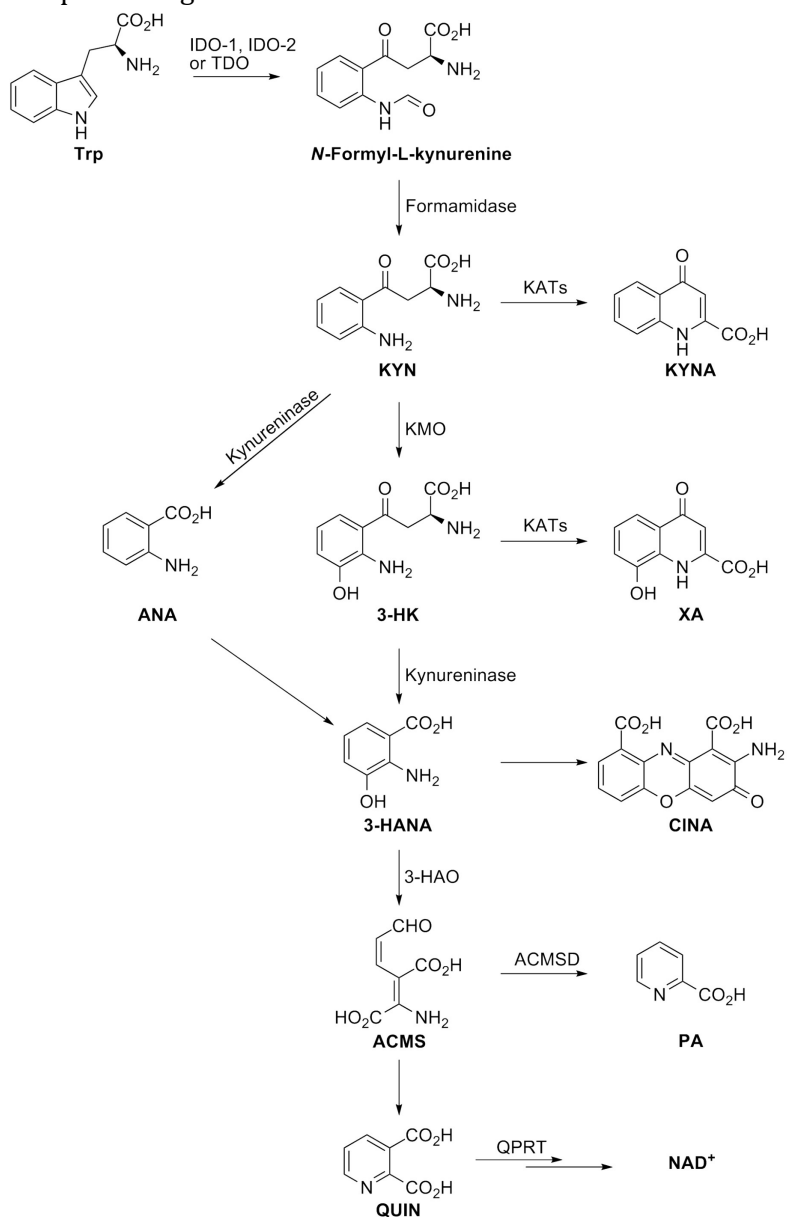
### 2.1.1 Kynurenine pathway

Tryptophan is one of the essential amino acid but it is also the precursor of serotonin and almost the 95% of it is metabolized through the kynurenine pathway, both in periphery and in the brain.<sup>1</sup> In the central nervous system (CNS), an imbalance of the intermediate metabolites of this pathway has been related to several neurodegenerative diseases,<sup>2</sup> as it will be discussed further in this chapter.

The first step of the kynurenine pathway (Figure 1) is the conversion of tryptophan to N-formyl-L-kynurenine by tryptophan 2,3-dioxygenase and indoleamine 2,3-dioxygenase 1 and 2 (IDO-1 and IDO-2). N-formyl-L-kynurenine is then hydrolyzed by formamidase to L-kynurenine (KYN). Kynurenine is a key metabolite of this pathway, because it can be metabolized through three distinct pathway: 1) it can be converted into anthranilic acid by kynureninase, 2) it can be metabolized by kynurenine aminotransferase I, II and III (KAT) into kynurenic acid (KYNA) or 3) KYN can be transformed into 3-hydroxykynurenine (3-HK) by kynurenine-3-hydroxylase (KMO). The three possible branches of the pathway also seem to have a different localization in the brain: KAT, that catalyzes the formation of KYNA is more present in astrocytes, whereas kynureninase and kynurenine-3-hydroxylase are prevalent in microglial cells.<sup>3</sup> Both their products, 3-HK and anthranilic acid, can be further metabolized into 3-hydroxyanthranilic acid, that is the substrate of 3-hydroxyanthranilic acid 3,4-dioxygenase (3-HAO), which catalyzes the oxidative ring-opening to form the reactive intermediate 2-amino-3-carboxymuconic-6-semialdehyde, that preferentially converts to quinolinic acid (QUIN) by a non-enzymatic cyclisation.

QUIN is the precursor of NAD<sup>+</sup> and NADP<sup>+</sup>, but it is also an important endogenous neurotoxin, that has been related to several neurological disorders such as Huntington's disease (HD),<sup>4</sup> Alzheimer's disease,<sup>5</sup> amyotrophic lateral sclerosis<sup>6</sup> and also major depressive disorders.<sup>7</sup> It acts

through different mechanism, but most importantly it is a weak but specific competitive agonist of the NMDA (N-methyl-D-aspartate) receptors.<sup>8, 9</sup> NMDA receptors are glutamate-



**Figure 1** The kynurenine pathway of tryptophan metabolism. Trp: L-tryptophan; IDO: indoleamine 2,3-dioxygenase; TDO: tryptophan 2,3-dioxygenase; KYN: L-kynurenine; KATs: kynurenine aminotransferases; KYNA: kynurenic acid; KMO: kynurenine 3-monooxygenase; ANA: anthranilic acid; 3-HK: 3-hydroxykynurenine; XA: xanthurenic acid; 3-HANA: 3-hydroxyanthranilic acid; CINA: cinnabarinic acid; 3-HAO: 3-hydroxyanthranilic acid 3,4-



dioxygenase; ACMS: 2-amino-3-carboxymuconic-6-semialdehyde; ACMSD: 2-Amino-3-carboxymuconic-6-semialdehyde decarboxylase; PA: picolinic acid; QUIN: quinolinic acid; QPRT: quinolinic acid phosphoribosyltransferase.

gated ion channels with a pivotal role in the regulation of synaptic function in the CNS.<sup>10</sup> In fact it is well known that the modulation of their function and signaling is crucial in neurodevelopment and synaptic plasticity.<sup>11</sup> NMDA receptors dysfunction and misregulation are related to a several number of neurodegenerative diseases, in particular they can promote neuronal death under excitotoxic pathological conditions, as such mediated by QUIN excess.<sup>12</sup> NMDA receptors are multimeric complexes of different subunits, NR1, NR2 and NR3, with different physiological and pharmacological properties and also distinct patterns of synaptic distribution.<sup>13</sup> NR2 subunit is particularly important for NMDA receptor functions, and determines many biophysical and pharmacological properties of the receptor.<sup>10</sup> There are four different types of NR2 subunit: NR2A, with lower affinity for glutamate, fast kinetics, greater open probability (the fraction of time a single channel remains open when activated) and prominent Ca<sup>2+</sup>-dependent desensitization; NR2B is less sensitive to Ca<sup>2+</sup>-dependent desensitization, which determines slow channel kinetics and a reduced open probability; NR2C and NR2D are both characterized by low conductance openings and reduced sensitivity to Mg<sup>2+</sup> block.<sup>11</sup> The other subunit NR1 is present in eight different subtypes whereas NR3 in two (A and B); in both NR1 and NR3 the agonist binding domain (ABD) binds glycine, whereas in the NR2 subunits ABD binds glutamate.<sup>14, 15</sup> In CNS most NMDA receptors are usually tetramers constituted of two NR1 and two NR2 subunits, and to activate the receptor, the simultaneous binding of the two different co-agonists is required. More recently, it has been observed that central NMDA receptors are mobile in the neuronal plasma-membrane and that they can laterally diffuse between synaptic and extrasynaptic sites.<sup>16</sup> The balance between synaptic and extrasynaptic NMDA receptors has been shown to be crucial for their activity and the two receptor populations can differentially signal to cell survival or apoptotic pathways; this has been suggested to possibly contribute to early cognitive dysfunction and onset of pathogenesis in neurodegenerative disorders such as Huntington's disease,<sup>17, 18</sup> which is associated with an increased NMDA receptor distribution from synaptic plasma-membrane to extrasynaptic sites.<sup>11, 19</sup> QUIN is a selective agonist for NMDA receptors subtypes containing NR2A and NR2B subunits.<sup>20</sup> Although

it is only one quarter as active as NMDA and approximately as active as glutamate in stimulating NMDA receptors,<sup>21</sup> QUIN excitotoxicity is increased by its rapidly saturated uptake system, that prolongs its stimulation in cases of increased QUIN production.<sup>22</sup> Quinolate phosphoribosyl transferase (QPRT) is the enzyme responsible for QUIN catabolism into NAD<sup>+</sup> and it has been noticed

that there are more cells containing 3-HAO than those containing QPRT.<sup>23</sup> Kinetic studies indicated that the two enzymes have similar  $K_m$  values, however 3-HAO reaction velocity is 80-fold higher than QPRT,<sup>24</sup> leading to QUIN accumulation under certain pathological conditions. Interestingly, QPRT has also a different brain localization compared to 3-HAO. For this reason QUIN, once produced by 3-HAO in microglial cells, must exit these cells to be metabolized by QPRT, present in astrocytes and neurons.<sup>21</sup>

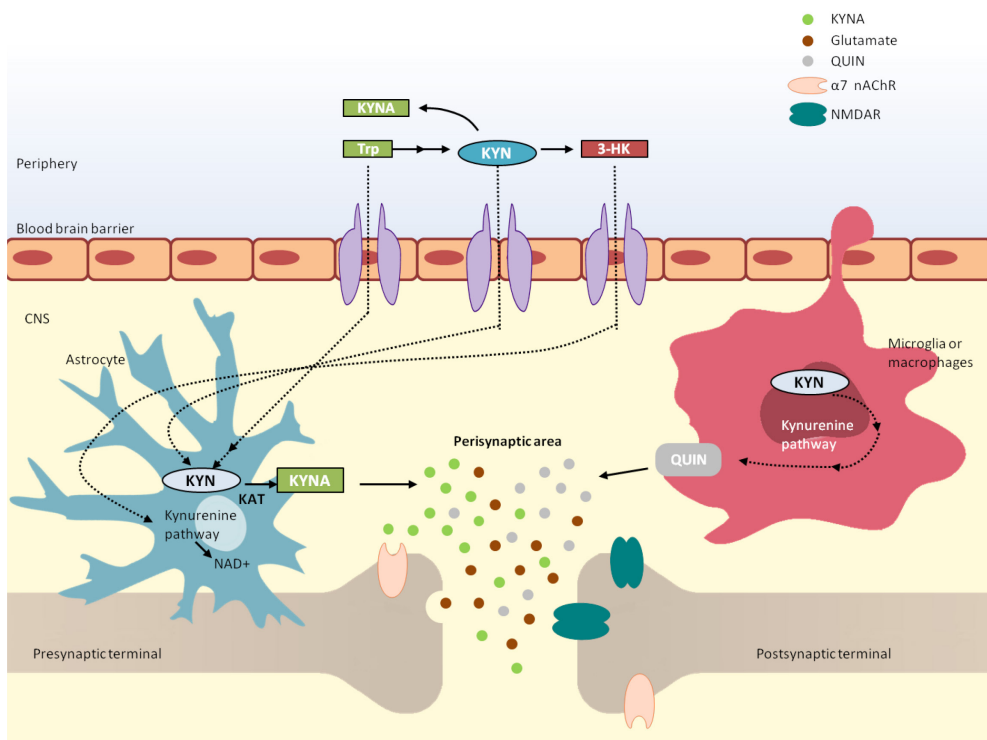
Apart from activating NMDA receptors, QUIN induces neurotoxicity through other different but equally important mechanisms: it can increase glutamate release and inhibit its reuptake by astrocytes;<sup>25</sup> QUIN induces also metabolic impairment through inhibition of B monoamine oxidase (MAO-B) in mitochondria, that leads to progressive mitochondrial dysfunction;<sup>26</sup> oxidative stress due to increased free radical generation has also been noticed;<sup>27, 28</sup> the increased radical production is in part dependent of its activity on NMDA receptors and in part independent, probably related to the stimulation of NOS activity in astrocytes and neurons.<sup>29</sup>

However QUIN is not the only neurotoxic kynurenine metabolite. Impairment in the balance between the production of KYNA in astrocytes, that has a neuroprotective role,<sup>30</sup> and 3-HK in microglia, that mediates neurotoxicity by increasing the levels of free radicals,<sup>31</sup> has been proposed to be involved in several neurological disorders (Table 1), together with QUIN-induced toxicity. KYNA is considered neuroprotective, because it acts as an antagonist on the strychnine-insensitive glycine-binding site and on the glutamate binding site in NMDA receptors,<sup>9</sup> and because of its action on  $\alpha 7$  nicotinic acetylcholine receptors (nAChRs), that suppresses the presynaptic release of glutamate.<sup>32</sup> Free radical-mediated 3-HK neurotoxicity is also increased by its metabolite 3-hydroxyanthranilic acid, which can undergo auto-oxidation generating superoxide anions.<sup>9</sup>

Alterations in tryptophan metabolism have been reported in several neurological diseases as reported in Table 1. The most relevant example of these is Huntington's disease. In fact a growing body of evidence demonstrates that the kynurenine pathway is altered in HD: elevated levels

of QUIN and 3-HK are registered in patients,<sup>33</sup> together with an increased activity of 3-HAO,<sup>34</sup> a decreased level of KYNA and also decreased activity of KATs; finally a recent study in animal models revealed also increased KMO and decreased kynureninase activity.<sup>30</sup> Moreover it has been shown in rats that QUIN is able to induce the expression of huntingtin gene.<sup>35</sup> Abnormal levels of kynurenine metabolites are also found in patient suffering from AIDS-dementia complex,

**Figure 2. Schematic representation of central aspects of kynurenine pathway in the**



**brain**

together with an increased IDO activity;<sup>36</sup> elevated QUIN concentrations in cerebrospinal fluid have been correlated with virus load and symptomatology, and with neuronal loss in several brain regions. After a cerebral insult, for example in case of ischaemia, glial cells are activated to produce and secrete cytokines and kynurenines, and elevated levels of all the enzymes of the kynurenine pathway except for KAT are registered in these patients.<sup>37</sup> In many other neurological pathologies such as schizophrenia, Parkinson's disease, and epilepsy among the others (see

Table 1), brain levels of kynurenines are altered but a direct correlation between kynurenines and the ethiology is still under investigation

**Table 1. Kynurenine metabolites alterations in neurological disorders**

<b>Huntington’s disease</b>
<ul style="list-style-type: none"> <li>• Decreased KYNA levels in the cortex, striatum and CSF<sup>38-40</sup></li> <li>• Decreased KAT activity in the striatum<sup>39</sup></li> <li>• Elevated 3-HAO activity in the brain<sup>33</sup></li> <li>• Increased 3-HK and QUIN levels in the brain<sup>41</sup></li> <li>• Decreased KYNA/QUIN and KYNA/3-HK ratios in the striatum at early stages of the disease<sup>34</sup></li> </ul>
<b>Alzheimer’s disease</b>
<ul style="list-style-type: none"> <li>• Elevated KYNA levels in the striatum and hippocampus and decreased KYNA levels in the blood and CSF<sup>42</sup></li> <li>• Elevated KAT activity in the striatum<sup>42</sup></li> <li>• Elevated IDO and QUIN immunoreactivity in the hippocampus in association with senile plaques<sup>5</sup></li> </ul>
<b>Cerebral ischaemia</b>
<ul style="list-style-type: none"> <li>• Elevated activity of IDO, kynureninase, KMO and 3-HAO but unaffected KAT activity<sup>43, 44</sup></li> </ul>
<b>Multiple sclerosis</b>
<ul style="list-style-type: none"> <li>• Elevated levels of KYNA in CSF during acute relapse but decreased KYNA levels in chronic remission<sup>45, 46</sup></li> </ul>
<b>Amyotrophic lateral sclerosis</b>
<ul style="list-style-type: none"> <li>• Elevated KYN and QUIN levels in the CSF and serum<sup>6</sup></li> <li>• Elevated microglia and neuronal IDO expression in the motor cortex and spinal cord<sup>6</sup></li> <li>• Elevated IDO activity in CSF<sup>6</sup></li> </ul>
<b>Parkinson’s disease</b>
<ul style="list-style-type: none"> <li>• Decreased KYNA levels in the frontal cortex, SNpc and putamen<sup>47</sup></li> <li>• Increased levels of 3-HK in the SNpc and putamen<sup>47</sup></li> </ul>
<b>Schizophrenia</b>
<ul style="list-style-type: none"> <li>• Increased KYNA levels in the CSF<sup>48</sup></li> </ul>
<b>Epilepsy</b>
<ul style="list-style-type: none"> <li>• Decreased levels of KYNA and KYN in CSF in infantile spasms and West syndrome<sup>49, 50</sup></li> <li>• Elevated 3-HK levels in CSF in infantile spasms<sup>49</sup></li> </ul>

## 2.1.2 Targeting the kynurenine pathway in the brain

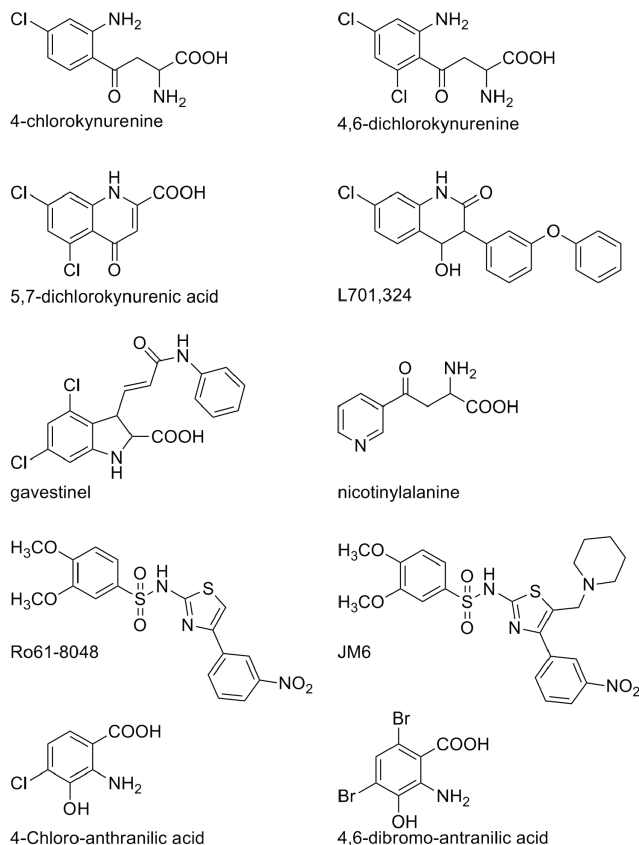
Kynurenine metabolism is altered in several neurological disorders and small changes in kynurenines concentration in CNS have a disproportionately large effect on neuronal function. Therefore, some strategies have been explored in the course of years to target the enzymes of this pathway.<sup>30</sup> The first strategy is to exploit the neuroprotective effect of KYNA, trying to increase its level in the CNS.<sup>51</sup> However direct administration is not a feasible route, due to its poor ability to penetrate the blood brain barrier (BBB). Some halogenated KYN derivatives have been studied as BBB-penetrant precursors of KYNA. Systemic administration in animal models of 4-chlorokynurenine or 4,6-dichlorokynurenine result in the production of halogenated KYNA in the brain, in particular 4,6-dichlorokynurenine (Figure 3), that is converted into 5,7-dichlorokynurenic acid, is the most active one, exhibiting a IC<sub>50</sub> of 80nM against strychnine-resistant glycine binding site; 4-chlorokynurenine also completed the Phase I clinical safety trial, but then failed in demonstrating activity as neuroprotective agent.<sup>30</sup>

Analogues of KYNA able to cross the BBB and exhibiting neuroprotective activity have also been developed.<sup>52</sup> Halogenated and thio-substituted KYNA derivatives showed increased selectivity and affinity for the strychnine-insensitive glycine-binding site in NMDA receptors compared to KYNA. Also other scaffolds, such as indole derivatives or benzazepindione compounds, demonstrated to be active as antagonist on the glycine binding site of NMDA receptors. One of these derivatives, the 3-[(E)-3-anilino-3-oxoprop-1-enyl]-4,6-dichloro-1H-indole-2-carboxylic acid, known as gavestinel (Figure 3), reached the Phase II of clinical trials in stroke but then failed in demonstrating efficacy.<sup>9, 53</sup>

Another strategy is to block the branch of the kynurenine pathway leading to the production of neurotoxic metabolites, promoting simultaneously the production of endogenous KYNA. In the previous years, some inhibitors targeting KMO, kynureninase or 3-HAO have been developed, and the most relevant ones are summarized in Figure 3.

Nicotinylalanine increases the production of KYNA while inhibiting both kynureninase and KMO, and has been shown to prevent the induction of seizures.<sup>54</sup> The most potent KMO inhibitor developed so far is Ro61-8048, that shows a IC<sub>50</sub> of 37nM and is also orally active.<sup>55</sup> Ro61-8048 does not reduce QUIN acid in normal mice, but suppresses KMO activity and QUIN formation when IDO activity is induced by immune mediators.<sup>56</sup> With the

**Figure 3. Compounds active as kynurenine pathway inhibitors or competitive inhibitors of kynurenines action on NMDA receptors.**



goal of increasing KYNA levels, a peripherally active KMO pro-drug inhibitor (JM6 or 2-(3,4-dimethoxybenzenesulphonylamino)-4-(3-nitrophenyl)-5-(piperidin-1-yl)methylthiazole) was developed, which is supposed to act by increasing the blood levels of KYN. However recent findings questioned its activity as a prodrug for Ro61-8048.<sup>57</sup>

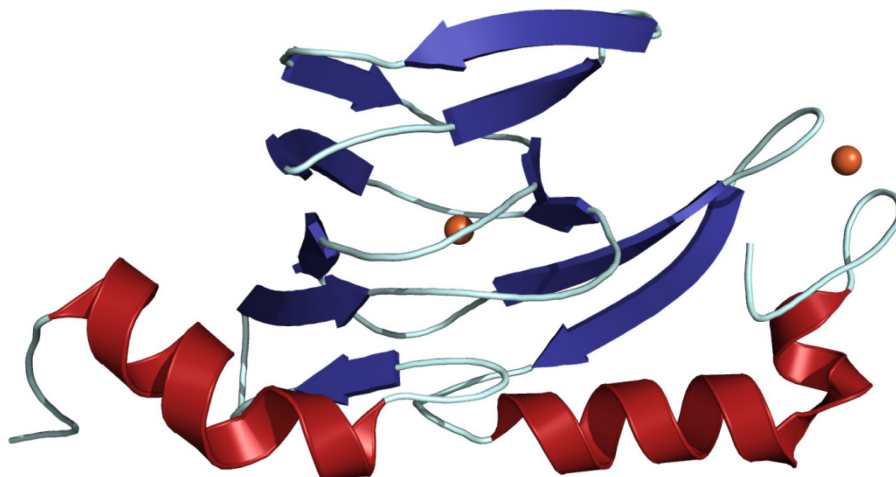
Another strategy to reduce QUIN formation and shifting the kynurenine pathway to the production of the neuroprotective KYNA, is to block the last step, catalyzed by 3-HAO. The first class of inhibitors is represented by the mono-, di-, and trisubstituted derivatives of the enzyme substrate 3-hydroxyanthranilic acid.<sup>58, 59</sup> Most of these analogs are extremely potent in vitro, the most potent one is the 4,6-dibromo-derivative, with IC<sub>50</sub> values in the low nanomolar range.<sup>60</sup> However the experimental use in vivo is severely limited by their instability under

physiological conditions, due to the tendency of the o-aminophenol nucleus to auto-oxidize and generate reactive radical species.<sup>61</sup>

### 2.1.3 3-HAO

One useful approach in trying to change the balance between QUIN and KYNA and thus to influence synaptic transmission and reduce excitotoxicity, is to inhibit 3-HAO, the last enzyme of the kynurenine pathway that catalyzes the transformation of 3-hydroxyanthranilic acid to 2-amino-3-carboxymuconic-6-semialdehyde, which spontaneously rearranges into QUIN. 3-HAO is a type III non-heme Fe<sup>2+</sup> dependent extradiol dioxygenase, that has been conserved during the evolutionary process from bacteria to higher species.<sup>62</sup> Dioxygenases are classified as intradiol or extradiol dioxygenases. Intradiol dioxygenases cleave the bond situated between the two hydroxyl groups, whereas extradiol dioxygenases cleave a bond adjacent to one of the two hydroxyl group.<sup>63</sup> Generally these enzymes both depend on mononuclear non-heme Fe(II), and although at a first sight they may seem similar, they have completely different structures and use a completely different catalytic mechanism.<sup>64</sup> Extradiol dioxygenases are divided into three evolutionary independent families: type I contains extradiol dioxygenases belonging to the vicinal oxygen superfamily, characterized by one or two domains (e.g. 2,3-dihydroxybiphenyl 1,2-dioxygenase II). To the second family (type II) belong multimeric extradiol dioxygenases (e.g. protococatechuate 4,5-dioxygenase). Type III extradiol dioxygenases are characterized by a cupin barrel fold, and thus belong to the cupin superfamily, that includes also enzymes such as the gentisate, homogentisate dioxygenase.<sup>65</sup> The name cupin derives from Latin *cupa*, that was used to indicate a small barrel, the shape resembled by the spatial disposition of the antiparallel  $\beta$ -sheets fragments. Catecholic extradiol dioxygenases utilize Fe<sup>2+</sup> and dioxygen to cleave the bond adjacent to one of the two ortho-hydroxyl groups, and residues forming the binding site are highly conserved among all the species.<sup>64</sup> 3-HAO presents the same conserved residues and shares the same mechanism of other extradiol dioxygenases to catalyze the cleavage of the bond adjacent to the hydroxyl group of 3-hydroxyanthranilic acid. Four prokaryotic 3-HAO crystal structures from *Ralstonia metallidurans* and three eukaryotic structures (yeast, bovine and human) have been crystallized, providing useful information to study this enzyme.

**Figure 4. Secondary structure of bacterial 3-HAO. As evident from the picture of the monomeric form of the enzyme, two iron site are present.**



### **2.1.3.1 Crystal structures**

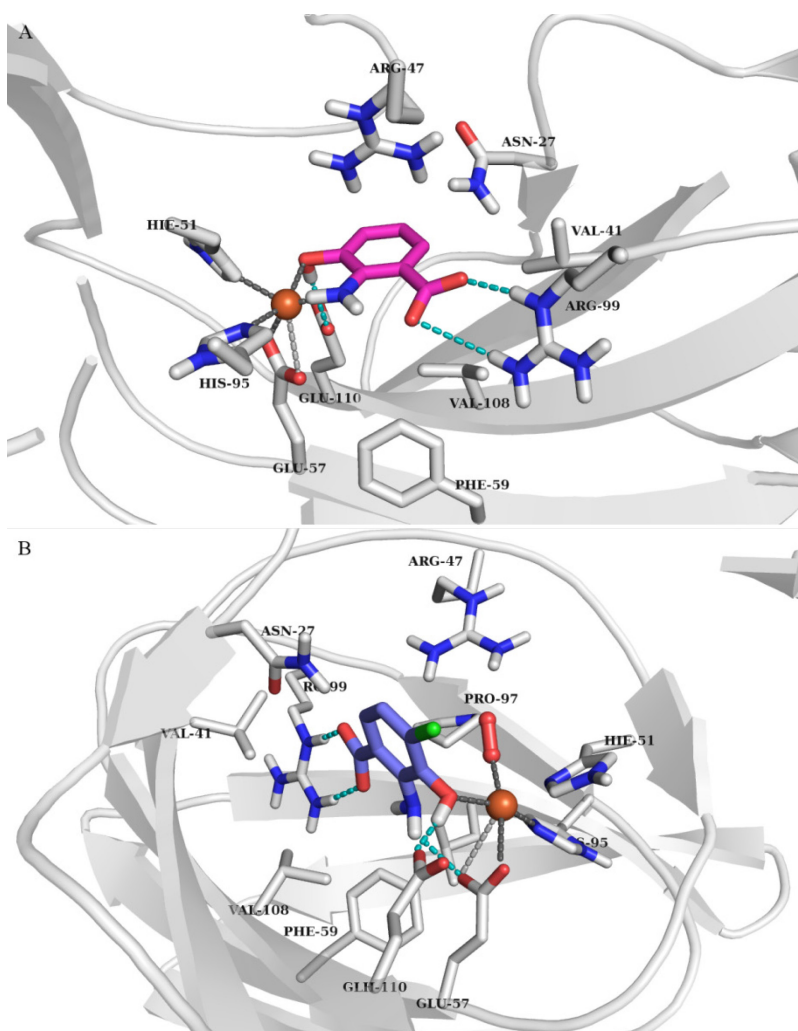
Bacterial 3-HAO was the first to be crystallized. It is available in its apo and holo forms, and in complex with the inhibitor 4-chloroanthranilic acid (4ClHAA) in two structures, one with dioxygen and the other with nitric oxide.<sup>62</sup> Crystal structures with substrate and inhibitor are of particular importance because they are the only bounded forms of 3-HAO available, revealing the binding modes and the changes occurring upon ligand binding. The prokaryotic form of the enzyme is a homodimer and each monomer, consisting of 174 residues, contains two iron binding sites: the catalytic site, buried inside the cupin barrel, and a rubredoxin-like site of unknown function, 24 Å away from the catalytic one. This second iron binding site is located in the C-terminal region and the iron ion is coordinated by four cysteine residues, forming a tetrahedral FeS<sub>4</sub> center.

The core motif of each monomer is the jellyroll β-barrel formed by two antiparallel β-sheets constituted of six strands. In addition to the cupin barrel, each monomer contains also a 3<sub>10</sub> helix and two α-helices. The catalytic site is formed by a cluster of buried hydrophilic residues, Arg47, His51, Glu57, His95, Arg99, Glu110, forming a hydrogen bonding network around the ferrous ion.



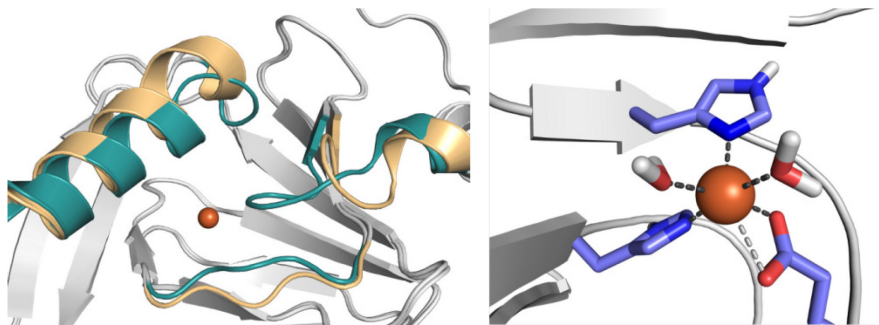
Fe(II) in the apo form of the enzyme presents a distorted octahedral geometry and is coordinated by two histidine residues, His51 and His95, a glutamate residue, Glu57, and two water molecules (Figure 6).

**Figure 5. A) Crystal complex of the substrate 3-hydroxyanthranilic acid and bacterial 3-HAO. B) Crystal complex of the inhibitor 4ClHAA and bacterial 3-HAO**



In the inhibitor-bound crystal structures (Figure 5B), the catalytic iron shows the same octahedral geometry and is coordinated to His51, bidentate Glu57, His95, 4ClHAA and O<sub>2</sub>/NO molecule. The crystal structures showed that the inhibitor acts as monodentate ligand, with the 2-amino

group involved in a hydrogen bond with Glu57. The carboxylate group of 4CIHAA forms an hydrogen bond with Arg99, whereas the 3-hydroxy group is involved both in the coordination of the iron atom and in a hydrogen bond with Glu110. The



**Figure 6. On the left the superposition of two crystal structure of bacterial 3-HAO; loops that move upon ligand binding are colored, in yellow is represented the open-apo form whereas in green-cyan the closed form**

hydrophobic portion of 4CIHAA is accommodate in the hydrophobic pocket formed by Val25, Phe121, Ile 142 and Leu 146.

The crystal structure with the substrate 3-hydroxyanthranilic acid (Figure 5A) was solved at lower resolution ( $3.2\text{\AA}$ ) if compared to the structures with the inhibitor. However a different binding mode of the substrate, compared with the inhibitor 4CIHAA, can be appreciated: 3-hydroxyanthranilic acid seems to bind as a bidentate ligand to the Fe(II), probably facilitating the subsequent dioxygen binding and the ring-opening reaction.

Comparing the crystal structures of 3-HAO in its bounded and unbounded form, some conformational changes are evident in the neighborhood of the binding site. The loop containing residues 21-27 moves towards the center of the  $\beta$ -barrel, bending on the top of the binding site as to close it, and also the  $\alpha 2$  helix unwinds one helical turn and moves towards the catalytic site. This latter step is favored by the presence of two consecutive proline residues, Pro147 and Pro148, in the  $\alpha 2$  helix, that disrupts the normal hydrogen bonding interactions within  $\alpha$  helix. Therefore to unwind one helical turn only one hydrogen bond, between Ile142 and Leu146, needs to be broken, that is compensated by the a newly formed hydrogen bond visible in the bounded structures between Ile142 and Asn27, diminishing the energetic barrier required by the conformational change. Another important conformational change happening upon ligand binding involves Arg47, that moves towards the

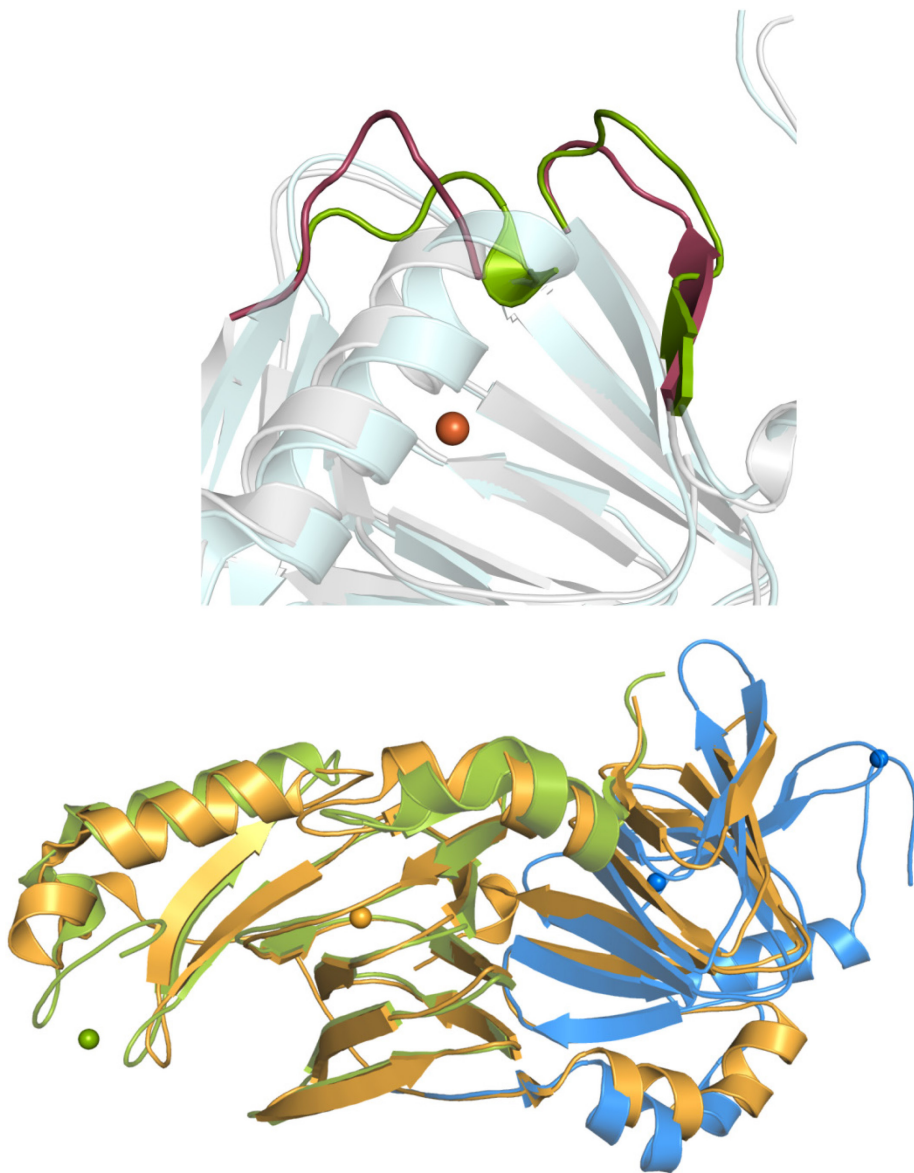
catalytic iron, making a hydrogen bond with the molecular oxygen and a cationic- $\pi$  interaction with the aromatic ring of the ligands.

The first crystal structure of eukaryotic 3-HAO was the one from *S. cerevisiae*.<sup>66</sup> The enzyme, consists of 176 residues and is functionally organized as a homodimer, as the prokaryotic one. The secondary structure and the overall fold are very similar to those of the bacterial 3-HAO. As the *Ralstonia metallidurans* enzyme, each monomer has two metal binding sites, the catalytic site and the rubredoxin-like site of unknown function. However the yeast structure contains nickel ions not iron, as a result of the purification protocol. The sequence identity between yeast 3-HAO and bacterial 3-HAO is 38%, with an higher degree of conservation for the amino acid residues lining the binding site. The only substitution in active site residues occurs at position 51, where an Asn residue replaces a Glu residue present in the prokaryotic structure. Other two structural differences are found in the intervening loop  $\beta 5$ - $\beta 6$  and in the loop  $\beta 11$ - $\alpha 2$ , that are longer in the yeast enzyme (Figure 7, at top). In particular the loop  $\beta 11$ - $\alpha 2$  is the one that the bacterial structures showed to undergo a conformational change, unwinding the first helical turn and moving towards the active site upon ligand binding. In yeast structure this loop is three residue longer and  $\alpha 2$  helix is two residue shorter, making it more loose than the bacterial one. The small loop  $\alpha 1$ - $\beta 1$  shows in the yeast structure in one monomer the open conformation as in the bacterial apo structure, whereas in the other monomer it adopts the closed conformation, suggesting that the two forms are in equilibrium in apo enzymes and the equilibrium is shifted towards the closed conformation upon ligand binding.

Two mammalian crystal structure of 3-HAO have been resolved, bovine<sup>67</sup> and human 3-HAO. The human structure does not contain  $Fe^{2+}$  but  $Ni^{2+}$  due to the purification protocol, and has not be accompanied by any paper, corroborating the crystal structure with further details on the experiments performed.

Conversely respect to bacterial and yeast structures, mammalian 3-HAO is no longer a homodimer but it has evolved to a bi-cupin monomer of 286 residues. Moreover the rubredoxin-like iron binding site, with four cystein residues coordinating the  $Fe^{2+}$  atom is no longer present. Mammalian structures are characterized by two  $\beta$ -barrel domains linked by a long stretch, but only one of them contains the catalytic site(Figure 7, at bottom). The domain containing the catalytic site is very similar and can be superposed very well with the  $\beta$ -barrels of the bacterial and yeast

**Figure 7.** At the top, the superposition of bacterial and yeast 3-HAO highlights the differences between the two structures (bacterial in bright green, yeast in dark red); on the bottom is reported the structural alignment between mammalian (pale orange) and bacterial 3-HAO in its dimeric form (green and lightblue represent chain A and B).

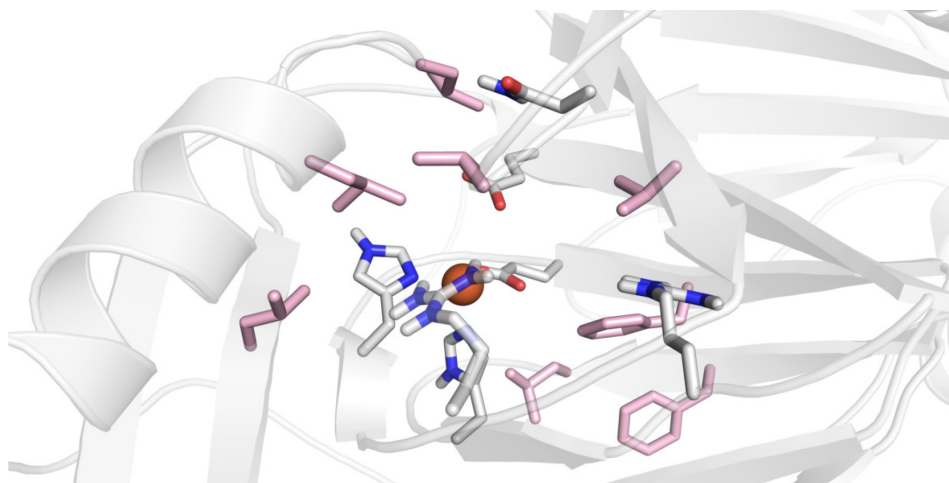


structures, whereas the other domain present in the mammalian structure is smaller. The spatial organization of the two domains resemble the one of the homodimers, with the two monomers related by a twofold axis, and the smaller domain almost superposed to the second monomer. The function of the second  $\beta$ -barrel domain is completely unknown and deserves further investigation. Human and bovine 3-HAO are very similar, presenting the 86.7% of identical amino acids. The main differences are related to the loop connecting  $\alpha 2$  ( $\alpha 1$  in bacterial and yeast structures, residues 18-24) and  $\beta 1$ , the one that acts as a lid on the top of the binding site, that is in open conformation in human structure but in closed conformation in the bovine one, even if both the structures were solved in their apo form. Another minor but relevant difference is present in loop  $\beta 11$ - $\alpha 3$  ( $\alpha 2$  in bacteria and yeast): in the human structure it assumes the conformation of the apo form, whereas in the bovine structure the first helical turn is unwind, as in the complexes with inhibitor and substrate. Despite the evolutionary change from a homodimer to a monomer, the catalytic mechanism of mammalian 3-HAO is supposed to be the same of the one proposed for *Ralstonia metallidurans*, presenting fully conserved catalytic residues and only minor changes in the amino acids surrounding the catalytic portion of the binding site.

### ***2.1.3.2 Binding site analysis***

The iron-containing binding site is highly conserved in all the organisms expressing 3-HAO.  $\text{Fe}^{2+}$  is bound deep inside the  $\beta$ -barrel and presents a distorted octahedral geometry; it is coordinated with two His residues (N $\delta 1$  of His51 and N $\epsilon 2$  of His95 in bacterial 3-HAO, N $\delta 1$  of His49 and N $\epsilon 2$  of His97 in yeast, N $\epsilon 2$  of both His47 and His91 in mammalian enzymes), a bidentate Glu residue (Glu57, Glu 55 and Glu53 in bacterial, yeast and mammal 3-HAO respectively), although one of the Fe-O bond is longer (2.8 Å) than typical iron-oxygen bonds, and with two water molecules, substrate or inhibitor and dioxygen. Upon ligand binding, the two water molecules are replaced either by the 3-hydroxyanthranilic acid, that coordinates iron in a bidentate manner, or by the inhibitor. 4ClHAA binds to the metal as a monodentate ligand, with the amino group forming an hydrogen bond with Glu instead of coordinating iron. The coordination sphere in 3-HAO complexed with inhibitor is completed by the molecular dioxygen or by the nitric oxide in the other available structure. In the binding site are present

also other important hydrophilic residues, highlighted by the comparison among bounded and unbounded structure from *Ralstonia metallidurans*: Arg99 (Arg101 in yeast, Arg95 in mammals) that interacts with the carboxylate group of the 3-hydroxyanthranilic acid; Glu 110 (Glu111 in yeast, Glu105 in mammals) that forms a strong hydrogen bond with the substrate, and experimental evidences suggested that the hydroxyl group of 3-hydroxyanthranilic acid binds to iron in its deprotonated form, corroborating the hypothesis that Glu is able to tear the hydrogen atom from the hydroxyl group, a crucial step in 3HAO mechanism of action. The importance of this interaction with Glu110 is also confirmed by mutagenesis studies performed with *Ralstonia metallidurans*,<sup>62</sup> showing a  $k_{cat}$  reduction of more than 2000-fold in the mutant E110A. Asn27 (Asn26 in yeast, Ans24 in mammals) is structurally and functionally important, forming a new hydrogen bond with Ile142, after the unwind of the first helical turn in the  $\alpha 2$  helix, stabilizing the closed conformation. Arg47 (Arg45 in yeast, Arg43 in mammals) is another key residue, forming a cationic- $\pi$  interaction with the aromatic ring of the substrate and a hydrogen bond with the molecular dioxygen. This residue undergoes a wide conformational change, moving 0.9Å towards the active site upon ligand



**Figure 8. Binding site residues in the bovine crystal structure. In pink are highlighted the hydrophobic residues.**

binding. The catalytic significance of this residue is also confirmed by the reduction of  $k_{cat}$  of more than 1000-fold in the mutant R47A of *Ralstonia*

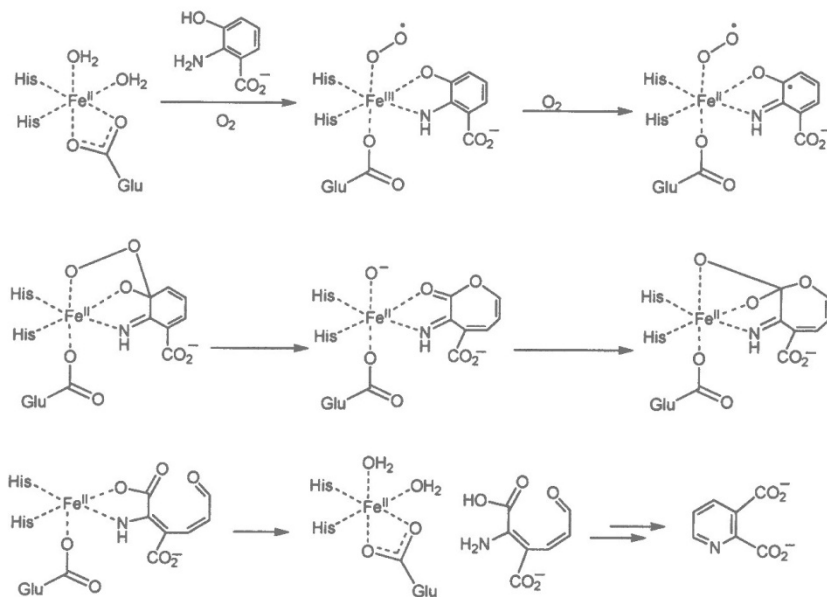
metallidurans. The conformational rearrangement is accompanied by the movement of Asp49 (Asp47 in yeast, Asp45 in mammals), that is involved in hydrogen bonds with Arg47, of about 0.7Å. The binding site is completed by some hydrophobic residues (yeast and mammalian in brackets, respectively): Val25(Val24, Val22), Val41 (Val39, Val37), Phe59 (Phe57, Phe55), Leu89 (Leu91, Leu85), Phe87 (Tyr89,Phe83),Phe121 (Ile122,Leu116), Ile142 (Leu143,Leu137), Leu146 (Val147,Leu141). Proline residues at the beginning of  $\alpha 2$  helix, residues creating the loop  $\alpha 2$ - $\beta 11$  and residues constituting the small loop  $\alpha 1$ - $\beta 1$  are also conserved among all species, highlighting the importance of their role in facilitating the conformational changes of the structure. Notably the second water molecule coordinating the iron atom in the apo form of bacterial 3-HAO showed an elongated electron density, as in the bovine structures, in which the authors preferred not to model this electron density as a water molecule, and suggested and that it could possibly represent the molecular dioxygen, although the oxygen binding is not predicted to happen before the ligand binding.<sup>62, 68</sup>

### ***2.1.3.3 Proposed mechanism of action and inhibition***

In vitro studies and the available crystal structures gave the opportunity to make some hypothesis on the mechanism of action and of inactivation of 3-HAO. Even if it belongs to the extradiol family, 3-HAO represents a peculiar subgroup, catalyzing the cleavage of an ortho-aminophenol compound. Other two examples of extradiol dioxygenases cleaving ortho-aminophenol are known: the 2-aminophenol 1,6-dioxygenase that belongs to type II extradiol dioxygenases<sup>69</sup> and the 4-amino-3-hydroxybenzoate 2,3-dioxygenase from *Bordetella* sp 10d.<sup>70</sup> This last enzyme has 30% and 24% identities with yeast and human 3-HAO respectively, presenting the same catalytic residues and a high degree of similarity in the surrounding amino acids, but curiously no identities with other extradiol dioxygenases; as 3-HAO, it exhibits high substrate specificity for its natural ligand, that differs from 3-hydroxyanthranilic acid only for the position of the amino group. Differences in residues lining the binding site should be involved in determining substrate specificity, differentiating the two enzymes and the reaction catalyzed. Only three residues are different: Arg99 (Arg101 in yeast, Arg95 in mammals)is replace by a phenylalanine in *Bordetella*, Phe59

(Phe57, Phe55) by a glutamine residue and finally Gln61 (Gln59, Gln57) by a serine, suggesting the importance of these residues in determining substrate specificity, in particular of the first two.

Zhang et al. in their detailed biochemical study proposed a catalytic mechanism for 3-HAO.<sup>62, 68</sup> The first step involves the displacement by the substrate of one water molecule and one of the bidentate Glu57; the substrate also loses an hydrogen atom from the hydroxyl group in favor of Glu110, to form a chelated monoanionic 3-hydroxyanthranilate; proton abstraction by Glu110 is facilitated by the complex hydrogen bond network formed by residues in proximity of the iron atom. Glu57 make an hydrogen bond with the 2-amino group of the substrate, probably pulling out an hydrogen atom from it. Substrate binding is also stabilized by the interaction of the carboxylic group with Arg99 and by hydrophobic interactions. The second step is the displacement of the second water molecule and the binding of the molecular oxygen to the vacant iron



**Figure 9. Schematic representation of the mechanism of action of 3-HAO proposed by Zhang et al.<sup>62</sup>**

coordination site; oxygen binding is stabilized by hydrogen bond with Arg47 and Asp49; at this stage the bound dioxygen acquires a negative charge due to electron transfer from the metal center, forming a diradical



intermediate imine-Fe<sup>2+</sup>-superoxide. At this point the activated dioxygen molecule attacks the radical C3 atom of the substrate, forming an alkenylperoxo intermediate, that undergoes a Criegee rearrangement with the cleavage of the peroxide bond to form an unsaturated seven-membered lactone intermediate; hydrolysis of the lactone gives then the product 2-amino-3-carboxymuconic-6-semialdehyde. In his study, Zhang proposed a transient oxidation of the Fe<sup>2+</sup> to Fe<sup>3+</sup>, but this hypothesis has been ruled out by following mechanistic studies on extradiol dioxygenases, that demonstrated that electron density is transferred from the aromatic ring of the substrate to the bound oxygen via the iron, thereby giving them both radical character and activating them for reaction with each other.<sup>71</sup> However some recent studies proposed again the formation of the oxidized iron atom, re-opening the debate on which species is formed during the reaction.<sup>72</sup>

In the same study of Zhang et al., two hypothesis were proposed to explain the mechanism of inhibition of 3-HAO by the halogen derivatives of the 3-hydroxyanthranilic acid.<sup>62, 68</sup> They excluded the possibility of a covalent adduct of 4ClHAA by using mass spectrometry, and proposed that the inhibition can be ascribed either to an unproductive positioning of the inhibitor in the active site or because the electron withdrawing chlorine substituent makes the electron transfer from the substrate to the oxidized iron atom more difficult to happen. The unfavorable binding of the 4ClHAA to the active site is suggested to be caused by a steric clash between the chlorine atom and the Ile142 residue, that moves toward the bonding site upon ligand binding, forcing the inhibitor to bind to the iron atom as a monodentate ligand; the distance between C3 atom and the superoxide is thus too long to allow the Criegee rearrangement, preventing the reaction to proceed.

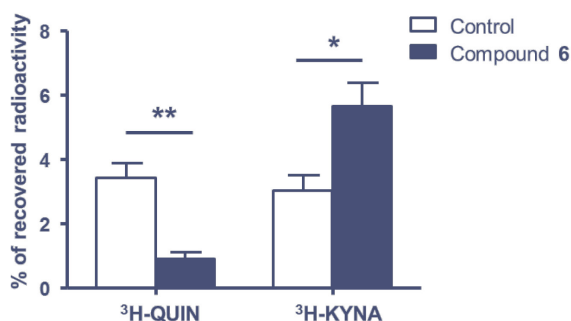
## 2.2 AIMS

3-HAO is the enzyme responsible for the production of QUIN, a neurotoxic metabolite of the kynurenine pathway. Elevated brain levels of QUIN are observed in several neurodegenerative diseases such as Huntington's disease, Alzheimer's disease, ichaemia and others. However before our synthetic laboratory discovered a new class of 3-HAO inhibitors, only a class of compounds, halogen derivatives of the substrate 3-hydroxyanthranilic acid, was available; unfortunately their experimental use is seriously limited by the lack of stability under physiological condition, due to the tendency of the o-aminophenol nucleus to undergo spontaneous auto-oxidation and to generate reactive radical species. Our new class of 3-HAO inhibitors is based on the 2-aminonicotinic 1-oxide nucleus and characterized by an increased stability, that could help to investigate the roles of QUIN in physiological and pathological conditions. To gain further insights into the mode of action of the 2-aminonicotinic 1-oxide derivatives, we decided to undertake a molecular modeling study. 3-HAO is an iron-dependent enzyme that catalyzes the ring opening of the 3-hydroxyanthranilic acid to form 2-amino-3-carboxymuconic-6-semialdehyde that then re-arranges into QUIN. Several crystal structures of 3-HAO are available in the PDB.<sup>73</sup> Analysis of these structures highlighted that the enzyme undergoes minor but important conformational changes upon ligand binding. Moreover the target under investigation deserves special cares, since in the binding site is present an iron atom, that cannot be appropriately treated with classical molecular mechanic approximations. Therefore, paying particular attention to two aforementioned aspects, we developed a protocol to perform molecular docking studies to provide a rational explanation for the activity and inactivity of the synthesized compounds.

## 2.3 MATERIALS AND METHODS

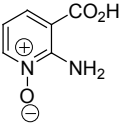
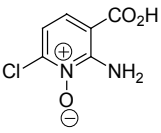
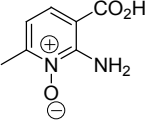
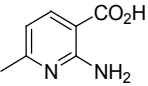
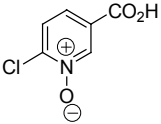
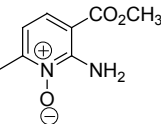
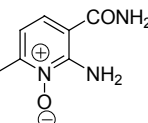
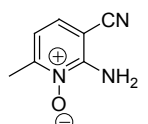
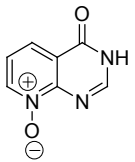
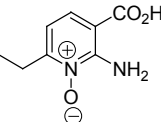
### 2.3.1 Compounds

In our synthetic laboratory a new class of 3-HAO inhibitor, based on the 2-aminonicotinic acid 1-oxide nucleus, has recently been identified. To escape from the unstable o-aminophenol moiety of reported inhibitors, while retaining most of its relevant structural and electronic features, the 2-aminonicotinic acid 1-oxide scaffold was proposed and several derivatives with different substituent were synthesized. The complete list of derivatives is reported in Table 1, whereas the complete synthetic procedure and the details of the *in vitro* experiments are available in the paper of Vallerini et al.<sup>61</sup> Briefly mono- and disubstituted 2-aminonicotinic acid 1-oxide derivatives were synthesized; substituents were generally small alkyl groups or halogens, except for the derivative **14** with a phenyl group attached to C4. Biological tests were performed in both rat and human brain homogenate, measuring the production of <sup>14</sup>C-QUIN after incubation with <sup>14</sup>C-3-HANA and different concentrations of test compounds. Activities reported in Table 1 are expressed as percentage of inhibition of 3-HAO, with standard errors of triplicate experiments. The IC<sub>50</sub> of the most potent compound (derivative **3**) was evaluated in rat and human brain homogenate, yielding 2.8 and 1.1  $\mu$ M respectively; compound **3** was also tested *in vivo*, in lesioned rat striatum, showing a reduction in QUIN production and an increase in KYNA levels (Figure 10).



**Figure 10.** Acute conversion of <sup>3</sup>H-kynurenine to <sup>3</sup>H-QUIN and <sup>3</sup>H-KYNA in the excitotoxically lesioned rat striatum *in vivo*. Data are the mean  $\pm$  SEM of 4 (control) and 5 (compound 6) rats, respectively. \*\* $p < 0.01$ , \* $p < 0.05$  vs. controls (unpaired Student's t-test). See text for experimental details.

**Table 1. Results of the *in vitro* 3-HAO inhibition test for all the synthesized compounds**

Cpd	Structure	Rat Brain % Inhibition			Human Brain % Inhibition		
		10 $\mu$ M	100 $\mu$ M	1 mM	10 $\mu$ M	100 $\mu$ M	1 mM
1		47.6 $\pm$ 8.4	84.9 $\pm$ 6.1	92.4 $\pm$ 4.0	74.0 $\pm$ 7.3	88.8 $\pm$ 5.5	94.0 $\pm$ 2.6
2		67.3 $\pm$ 16.2	84.0 $\pm$ 4.5	95.2 $\pm$ 4.2	72.9 $\pm$ 13.8	89.6 $\pm$ 3.4	95.2 $\pm$ 4.2
3		76.9 $\pm$ 11.0	94.1 $\pm$ 5.2	96.0 $\pm$ 3.6	85.5 $\pm$ 6.3	92.3 $\pm$ 4.8	94.4 $\pm$ 5.1
4		-2.1 $\pm$ 2.2	5.1 $\pm$ 4.5	42.5 $\pm$ 20.7	-3.6 $\pm$ 4.6	0.7 $\pm$ 4.4	24.8 $\pm$ 5.7
5		2.9 $\pm$ 2.9	11.4 $\pm$ 12.7	37.0 $\pm$ 9.5	-5.4 $\pm$ 12.3	-3.0 $\pm$ 11.9	30.4 $\pm$ 17.2
6		1.0 $\pm$ 2.7	1.2 $\pm$ 5.0	9.3 $\pm$ 9.0	-7.9 $\pm$ 7.5	-5.8 $\pm$ 6.4	6.0 $\pm$ 3.4
7		-1.1 $\pm$ 3.5	-2.0 $\pm$ 3.0	11.5 $\pm$ 3.0	1.7 $\pm$ 0.9	1.3 $\pm$ 0.9	20.1 $\pm$ 2.0
8		-2.4 $\pm$ 4.1	-0.6 $\pm$ 6.9	2.4 $\pm$ 6.7	-2.9 $\pm$ 3.4	-0.5 $\pm$ 7.8	5.6 $\pm$ 6.5
9		0.6 $\pm$ 3.6	6.4 $\pm$ 1.4	46.9 $\pm$ 8.6	5.5 $\pm$ 1.7	13.0 $\pm$ 1.6	44.3 $\pm$ 4.0
10		69.9 $\pm$ 1.8	90.7 $\pm$ 0.9	94.2 $\pm$ 1.8	76.6 $\pm$ 3.7	88.4 $\pm$ 2.8	90.4 $\pm$ 2.5

CHAPTER 2 – Elucidation of the binding mode of a series of 3-HAO inhibitors

11		$-0.8 \pm 3.9$	$4.3 \pm 5.4$	$46.6 \pm 1.9$	$0.7 \pm 4.0$	$16.3 \pm 2.5$	$69.5 \pm 2.6$
12		$2.8 \pm 3.7$	$3.1 \pm 0.6$	$23.0 \pm 2.3$	$8.5 \pm 6.7$	$10.9 \pm 4.7$	$20.0 \pm 9.6$
13		$1.1 \pm 1.6$	$2.2 \pm 3.0$	$25.7 \pm 0.7$	$4.0 \pm 10.7$	$10.2 \pm 11.0$	$24.1 \pm 15.9$
14		$2.4 \pm 1.9$	$5.5 \pm 2.5$	$14.8 \pm 3.1$	$-2.8 \pm 2.2$	$4.1 \pm 3.9$	$7.4 \pm 7.0$
15		$1.2 \pm 2.4$	$2.9 \pm 3.0$	$14.6 \pm 1.6$	$-5.4 \pm 2.3$	$1.5 \pm 7.3$	$10.6 \pm 2.2$
16		$17.6 \pm 2.7$	$61.6 \pm 3.8$	$85.8 \pm 2.8$	$31.3 \pm 9.7$	$76.3 \pm 3.4$	$89.1 \pm 3.0$
17		$8.7 \pm 2.6$	$30.7 \pm 9.5$	$67.6 \pm 15.0$	$16.3 \pm 9.5$	$52.6 \pm 11.3$	$80.1 \pm 8.6$
18		$9.1 \pm 6.8$	$42.6 \pm 6.7$	$74.4 \pm 6.9$	$24.8 \pm 2.5$	$68.6 \pm 3.9$	$85.9 \pm 4.9$
19		$14.5 \pm 5.7$	$55.2 \pm 6.2$	$83.7 \pm 3.3$	$22.9 \pm 17.2$	$67.2 \pm 17.0$	$89.1 \pm 6.7$

### 2.3.2 Docking to Metalloproteins

Molecular docking is one of the most important techniques in drug discovery, as described in the previous chapter. Metalloproteins play an important role in physiological processes, however, unfortunately, docking studies involving metalloproteins pose a serious challenge because the ligand interaction with the transition metal can be treated appropriately only at the quantum mechanical (QM) level.<sup>74</sup> Molecular mechanics approximation commonly used for docking and molecular dynamic simulations (MD), do not handle properly polarization and electron transfer. Actually, it is well known that force field charges for metal ions are not optimized as real partial charges, and usually the formal charge values are also used for partial charges, without any modification. Instead QM approach accurately estimates the atomic partial charges. However QM calculations for extended number of atoms in protein-ligand complexes require big computational efforts in terms of computing hours, and are generally not suitable for docking purposes. Therefore quantum mechanical/molecular mechanics (QM/MM) technique has been developed, that treats the catalytic site of the protein at QM level, and the remaining part of the system at molecular mechanics level.<sup>75, 76</sup> The equation for the energy in a QM/MM treated system is

$$E_{QM/MM} = E_{MM}(o) + E_{QM}(i) + E_{QM-MM}(i+o)$$

where

$$E_{QM-MM}(i+o) = E^b_{QM/MM} + E^{vdW}_{QM/MM} + E^{el}_{QM/MM}$$

Some approaches have been proposed in the past years to efficiently deal with docking using a metalloprotein as target, combining QM, QM/MM, molecular dynamic simulations and docking. Attempts have been made to improve the docking accuracy for zinc-dependent enzymes, by reparametrization of the the metal ion force field. Sternberg published a study on zinc proteins in which he used a fluctuating atomic charge model of force field, parametrized by semi-empirical QM method.<sup>77</sup> Khandekwal et al proposed a method for predicting binding affinities of metalloproteins by combining a series of QM/MM and force field based MD calculations.<sup>78</sup> Cho and Rinaldo made the first attempt to combine QM/MM and docking,

treating at QM levels only the ligand docked into the active site and the metal atom.<sup>74</sup> In a following study they extended this approach to include in the QM region also the atoms surrounding the binding site, along with the metal ion and the ligand atoms already considered. This determined an increase in computing time of almost 20-fold, but consistently improved the results obtained. A useful example from their study, to suggest a protocol for the treatment of 3-HAO, is represented by the prediction of the binding mode of 4-hydroxybenzoate to protocatechuate 3,4-dioxygenase (PDB ID: 2BUR): with traditional docking approach and also considering only ligand and metal atoms in the QM region, the binding pose was wrongly predicted, with the carboxylic group oriented towards the iron atom. Including also the surrounding protein atoms in the QM region, they were able to predict correctly the binding mode with the following docking run. The authors also pointed out that, because the transfer of charge from the metal is predominantly to surrounding protein atoms, it is conceivable to devise a protocol in which such calculation is performed before the docking process itself.

### 2.3.3 Workflow

Accordingly with the previously reported literature example, we applied the following protocol to elucidate the binding mode of the 2-aminonicotinic acid 1-oxide derivatives to 3-HAO.

- Preparation of the protein structure. The Protein Preparation Workflow<sup>79</sup> contained in the Maestro program suite was used to add hydrogen atoms, assign bond orders, create zero-order bonds to the iron ion, optimize the hydrogen bonding network and, finally, to refine the protein structure with a maximum RMSD of tolerance of 0.3 Å.

- Assignment of partial charges using QM/MM approach. Partial charges were assigned with the single-point energy calculation of Qsite to ensure a more accurate charge assignment to the iron atom and the surrounding binding site residues. We used DFT-B3LYP, lacvp++\*\* as basis set and continuum solvation model for the QM region. In the QM region were included the Fe<sup>2+</sup> atom and atoms belonging to residues Arg43, His47, Glu53, His91, Arg95, and Glu105 (mammalian

structure numeration). For the MM treated region we used OPLS2005 as forcefield.

- Preparation of the ligands. Molecular models of compounds were built using Maestro; tautomerization and protonation state at  $\text{pH } 7.0 \pm 1$  were assigned with LigPrep. Ligands geometry optimization and partial charges were calculated with Jaguar, using DFT-B3LYP, lacvp++\*\* as basis set, a maximum number of 100 steps, and using the Poisson–Boltzmann solver as solvent model for water.

- Ligand docking. Docking simulations were performed using Glide5.6. The binding site grid box was centered on the centroid of the following residues: Fe<sup>2+</sup>, Arg43, His47, Glu53, Phe55, His91, Arg95, and Glu105 (mammalian numeration). During grid generation, the previously calculated protein partial charges were retained. Standard precision (SP) mode was applied in the docking process. Two constraints were defined to avoid the generation of poses not in agreement with experimental data, the iron–ligand interaction and the hydrogen bond between ligand and Arg95 residue, and at least one of them was imposed to be satisfied during docking. Ten poses for each ligand were required.

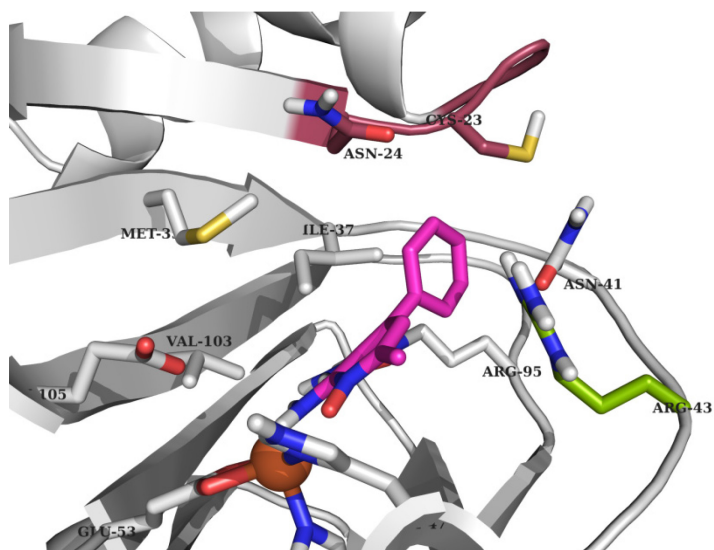
At first to elucidate the binding mode of the 2-aminonicotinic acid 1-oxide derivatives to 3-HAO, we decided to use the human crystal structure. Preliminary docking studies using the human crystal structure and the aforementioned protocol, were entirely unable to explain the lack of activity showed in the experimental assays by some of the synthesized compounds compared to the active ones. Therefore we decided to build a comparative model of the human structure using the bovine one as template to adopt the closed conformation showed by the bounded bacterial and by bovine crystal structure. Human and bovine 3-HAO share the same secondary structure, and residues are almost fully conserved (86% and 93% of identical and similar residues, respectively), making the building of the model an easy task. The model was built using Prime software available in Maestro suite. The docking protocol was then applied to the human model.



## 2.4 RESULTS AND DISCUSSION

A new series of 3-HAO inhibitors has been developed by our synthetic laboratory, and to gain further insight into the mode of action of the 2-aminonicotinic acid 1-oxide derivatives we decided to perform some molecular modeling studies. In vitro experimental assays were performed on rat and human brain homogenates.

Human 3-HAO crystal structure was the initial choice to perform docking studies. However this study did not provide the expected results, being the human 3-HAO crystal structure entirely unable to discriminate between active and inactive compounds (Figure 11). The reason of this failure was easily identified in the open conformation assumed by the human 3-HAO crystal structure, whereas the three bacterial structures with 4ClHAA and 3-hydroxyanthranilic acid presented a closed conformation. Moreover the human crystal structure has so far not been corroborated by further experimentation (for example by restoring activity after purification) and the number of co-crystallized water molecules is just above the value considered the cut-off between a good and a bad crystal structure.<sup>80</sup> Human and bovine 3-HAO share the same secondary structure, and residues are almost fully conserved (86% and 93% of identical and similar residues, respectively). Therefore we decided to use the bovine crystal structure, that presents a closed conformation despite the fact it is in its apo-form, as a template to build a comparative model of the human enzyme. The alignment between the human and the bovine protein sequences, together with the *Ralstonia metallidurans* one, is reported in Figure 12. Within the binding region around the co-crystallized 4ClHAA in the bacterial 3-HAO, only three residues are not fully conserved between the bacterial and the mammalian structures: Thr39 and has been replaced in the mammalian structures by Met35, Ile142 is changed in mammalian structures into Leu137 and finally Val41 is conserved in the bovine structure as Val37 but substituted by Ile37 in the human structure. The last two mutations preserve the characteristics of the bacterial residues, maintaining the hydrophobic character of this part of the pocket.



**Figure 11.** Pose of an inactive compound (compound 14) in the binding site of human crystal structure. As it can be appreciated from the picture, the open conformation of the small loop (in dark red) together with the conformation of Arg43 (in bright green) that points towards the top of the cavity instead of pointing towards the center, increases the volume of the pocket, allowing the binding of bigger compounds.

Conversely the substitution of the threonine residue with a methionine one not only impact the shape of the pocket, being the last one more bulky and elongating towards the center of the binding pocket, but also changing the electronic properties from hydrophilic to hydrophobic. Considering a wider binding site region of 6Å around the co-crystallized inhibitor, no other differences can be appreciated between human and bovine sequences, while other differences between mammalian and bacterial sequences occur: Asp53 is mutated into Glu49, conserving the same electronic properties; Gly26 and Leu139 are substituted by two cysteine residues, Cys23 and Cys134; Ala66 and Val143 are changed into Met62 and Gly138 respectively, maintaining almost the same hydrophobic features; Phe121 and Val137 are mutated respectively into Leu116 and Phe132. This substitution impacts the shape of the pocket, as can be seen in Figure 9: the spatial arrangement of the two mutated residues in mammalian structures seems to open a small sub-cavity, compared to bacterial structure; however the small pocket seems difficult to be exploited for designing compounds targeting also this portion of the binding site, because it is located just behind and slightly above the iron atom.

**Figure 12. Sequence alignment between bovine, bacterial and human 3-HAO structures. Residues within 4Å from the binding site are highlighted in green, while in cyan are those within 6Å. Amino acid differences are represented in bold.**

```

3FE5    ----ERPVRVKAWVEENRGSFLPPVCNKLLHQKQ-
LKIMFVGGPNTRKDYHEEGEEVFY
1YFW
MLTYGAPFNFPRWIDEHAHLLKPPVGNRQVWQDSDFIVTVVGGNHRTDYHDDPLEEEFY
2QNK    ---SERRLGVRAWVKENRGSFQPPVCNKLMHQEQ-
LKVMFVGGPNTRKDYHEEGEEVFY

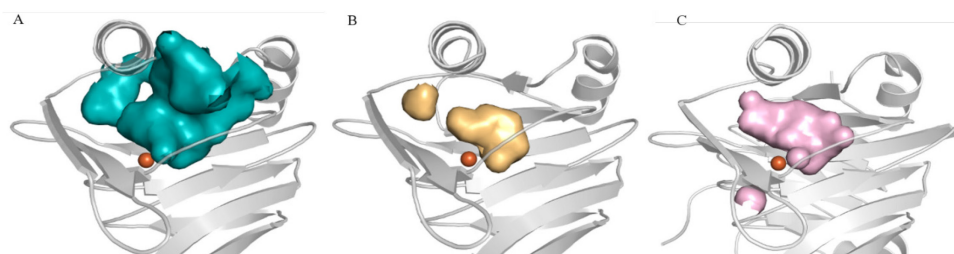
3FE5    QLEGDMLRVLERGKHRDVIRQGEIFLLPAGVPHSEQR-
FANTVGLVIRRRLKTELDG
1YFW
QLRGNAYLNLWVDGRRERADLKEGDIFLLPPHVRHSEQRPEAGSACLVIERQRPAGMLDG
2QNK    QLEGDMLRVLEQGKHRDVIRQGEIFLLPARVPHSEQR-
FANTVGLVIRRRLETELDG

3FE5
LRYYVGDTTDVLFEKWFYCEDLGTQLAPIIQEFFSSEQYRTGKPNPDQLLKEPPFLSTR
1YFW    FEWYCDACGHLVHRVEVLKSTVTDLPPLFESFYASEDKRRCPHCGVHPGRA----
--
2QNK
LRYYVGDTMDVLFEKWFYCKDLGTQLAPIIQEFFSSEQYRTGKPIPDQLLKEPPFLSTR

3FE5
SVMEPMCLEAWLDGHRKELQAGTPLSLFGDTYESQVMVHGQSSEGLRRDVDVWLWQLEG
1YFW    -----
--
2QNK
SIMEPMSLDAWLDSHRELQAGTPLSLFGDTYETQVIAYGQSSEGLRQNVDVWLWQLEG

3FE5    SSVVTMEGQRLSLTLDDSLLVPAGTLYGWERGQSVALSVTQDPACKKS--
1YFW    -----
2QNK    SSVVTMGRRLSLAPDDSLLVLAGTSYAWERTQGSVALSVVTQDPACKKPLG

```



**Figure 13 . Comparison volumes binding sites. A) Bacterial apo 3-HAO crystal structure. B) Bacterial closed 3-HAO conformation. C) Mammalian (bovine) crystal structure**

Once analyzed the protein 3D structures and the human model just built, we proceeded with the docking of the synthesized compounds using the human model, trying to find an explanation for the behavior of the compounds in the experimental assays. Before applying the protocol illustrated in the experimental section we made an attempt to dock the compounds treating both small molecules and protein using only molecular mechanics approximations. However as expected the results obtained were not satisfactory, showing distorted geometries in iron coordination or not taking into account the coordination. The correct assignment of atomic partial charges is indeed crucial for metalloproteins docking and cannot be fulfilled using molecular mechanics. At this point we applied the aforementioned protocol.

QM methods take into account polarization and charge transfer that are overlooked by classical MM methods. A comparison between binding site partial charges assigned using molecular mechanics and quantum mechanics is reported in Figure 14 and makes immediately clear the reason for using QM assigned partial charges for the binding site. Iron partial charge with molecular mechanics is considered equal to the formal charge +2.000 whereas QM takes into account the influence of the surrounding atoms and assigns a partial charge of +1.384 to the iron atom. Minor changes involves residues atoms directly coordinating the metal ion. A major difference concerns the partial charges of the carboxylic group of Glu105, the residue that, according to mechanistic studies, is responsible for the deprotonation of the hydroxyl group of 3-hydroxyanthranilic acid. MM approach treats all the carboxylic groups at the same manner, thus assigns partial charges of -0.800 to the two oxygen atoms and +0.700 to the carbon atom. Treating the region from a QM point of view, allows to recognize the peculiarity of this binding site region. In fact QM approach assigns a partial charge of -1.118 to the oxygen atom of Glu105 that takes the hydrogen atom of the substrate, to the other carboxylic oxygen atom a partial negative charge of -0.892 and to the carbon atom a partial charge of +1.017. Therefore the higher nucleophilic character of the Glu105 that is able to remove the hydrogen atom from the substrate can be modeled by using QM partial charges. Moreover the QM partial charges of the two arginine residues present in the binding site are substantially different from the MM ones. Accordingly with MM approximation in the two arginine residues N $\eta$ 1 and N $\eta$ 2 have a partial charge of -0.800 and N $\epsilon$  of -0.700. Conversely QM method assign to nitrogen atoms of Arg43, the residue coordinating the molecular oxygen, a value of -1.068, -0.783 and -0.245 to N $\eta$ 1, N $\eta$ 2 and N $\epsilon$

respectively, whereas to nitrogen atoms of Arg95, the one coordinating the carboxylic group of the substrate, a value of -1.209, -1.095 and -1.310 to N $\eta$ 1, N $\eta$ 2 and N $\epsilon$  respectively, and an increased positive partial charge to the hydrogen atoms bounded to N $\eta$ 1.

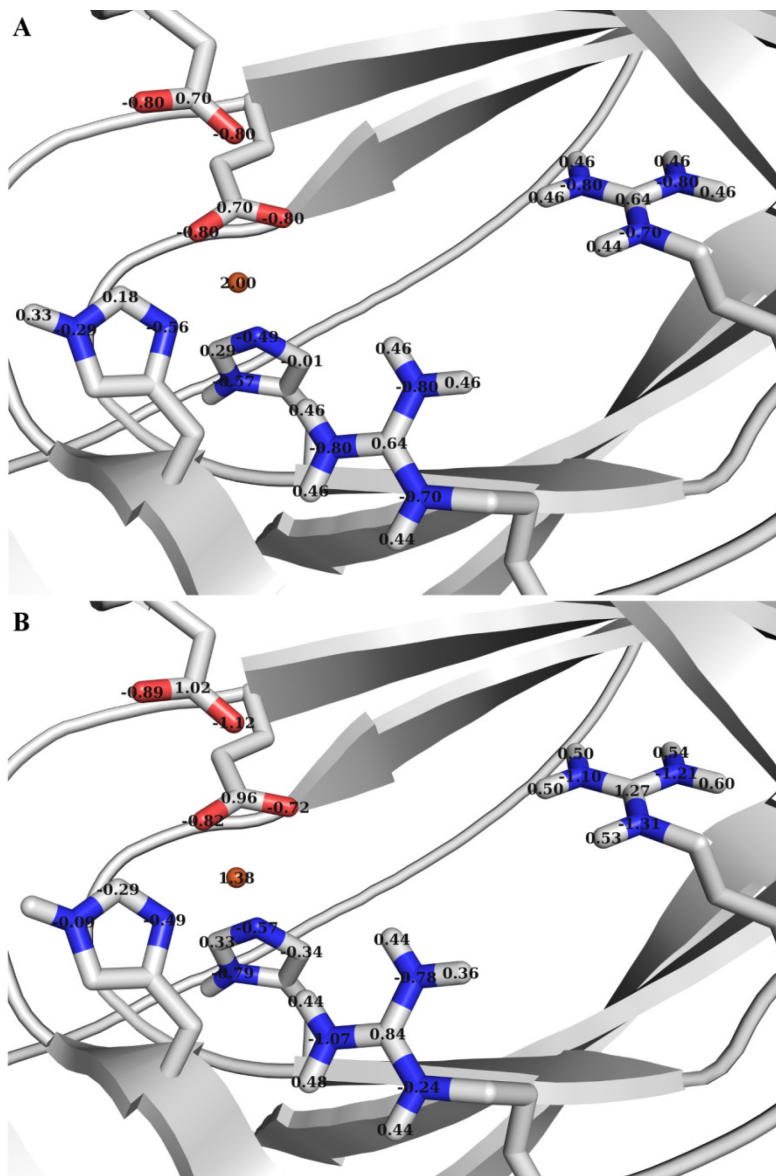
QM charges have also been assigned to the compounds to be docked. Some minor differences compared to the MM charges are present: the two carboxylic oxygen atoms and the secondary amino nitrogen present an increased negative charge and in most of the compounds the slightly positive partial charge of pyridine nitrogen is almost halved compared to the MM assigned atomic charges.

After having assigned the correct charges to protein and compounds, we performed some docking studies using Glide SP. Previous studies showed that molecular dioxygen binding happens only after ligand binding, thus it was not considered in our docking experiments. Results suggested a binding mode of compound **3**, the most potent one, consistent with the crystallographic binding mode of 4ClHAA to bacterial 3-HAO (Figure 5B and 15A), i.e. the 1-oxide portion coordinates the Fe<sup>2+</sup>, and the carboxylic acid moiety at position C<sub>3</sub> interacts with Arg95. The 2-amino portion is not involved in iron coordination, as for 4ClHAA, but interacts with the free oxygen of Glu53, that upon ligand binding, acts as monodentate ligand in coordinating the iron atom. The 2-aminonicotinic acid 1-oxide scaffold is well accommodated in the 3-HAO binding site, making further hydrophobic interactions with residues shaping the binding cavity, Met35, Iso37, Phe55, Val103.

Since we kept the receptor rigid during docking, Arg43 is posed as it was solved in the bovine 3-HAO crystal structure, pointing towards the active site. This arginine conformation is the same of the one assumed in bacterial crystal complexes with 3-hydroxyanthranilic acid and 4ClHAA, where Arg43 was coordinating the molecular oxygen and may have formed a  $\pi$ -cationic interaction with the ligand ring. In this case, however, it is pointing towards the electron deficient pyridine N-oxide ring, thus questioning whether the proposed interaction is actually significant in stabilizing the complex.

Therefore we calculated the Glide Score per residue, to have an idea of the interaction energy between Arg43 and compound **3**, and comparing the results to those obtain by docking 4ClHAA, and 3-hydroxyanthranilic acid, to better evaluate how this interaction is considered. As can be seen in Figure 16, the relative interaction of compound **3** with Arg43 is slightly smaller (less favorable) than for the other two compounds. However, the

**Figure 14. A) Partial charges assigned using MM approximation. B) Partial charges assigned using QM method**



Glide score per residue only provides a qualitative trend, and in our opinion cannot be used to draw quantitative conclusions. Moreover it represents a score of the force field based scoring function, that cannot represent accurately this type of ligand-protein interaction that includes iron chelation.

Docking poses show that small substituents at position C<sub>6</sub> of the aromatic ring (compounds **2**, **3** and **10**) are generally tolerated since they can be accommodated in the small sub-pocket delineated by Arg108, Cys134, Leu137, Leu141 and His147. These results are in agreement with the activity displayed by the compounds in the experimental assays.

The smallest active compound, compound **1**, do not have any substituent. Molecules from **4** to **9** were synthesized to investigate the essential of the inhibitory activity of 2-amino nicotinic acid derivatives. Removal of the amino group or of the N-oxide moiety resulted in completely inactive compounds; modification of the carboxylic group into an ester, an amide or into a cyano group produced inactive compounds as well. In compound **9** the amino and carboxylic moieties were fused together into a bi-cyclic lactam derivatives but also in this case the compound resulted completely inactive in the experimental assays.

Compounds presenting di-substitution at positions C<sub>4</sub> and C<sub>6</sub> of the pyridine ring and compounds mono-substituted at position C<sub>5</sub> or di-substituted at positions C<sub>5</sub> and C<sub>6</sub> resulted inactive or weakly active as 3-HAO inhibitors. From the analysis of the docking studies it can be proposed that the main reason of this lack of activity could be steric clashes with binding site residues. In fact, like the bacterial enzyme, mammalian 3-HAO has on one side a small flexible loop that acts as a lid, moving from an open to a closed conformation above the binding site especially upon ligand binding, together with the unwinding of the first helical turn of the  $\alpha$ 3 helix that partially covers the binding site on the other side of the cavity. In our docking studies we used an almost closed conformation, as in the bovine 3-HAO crystal structure, because dockings using the open conformation of the human 3-HAO structure were not in agreement with the experimental data.

Figure 15. A) Pose of compound 3. B) Pose of compound 2.

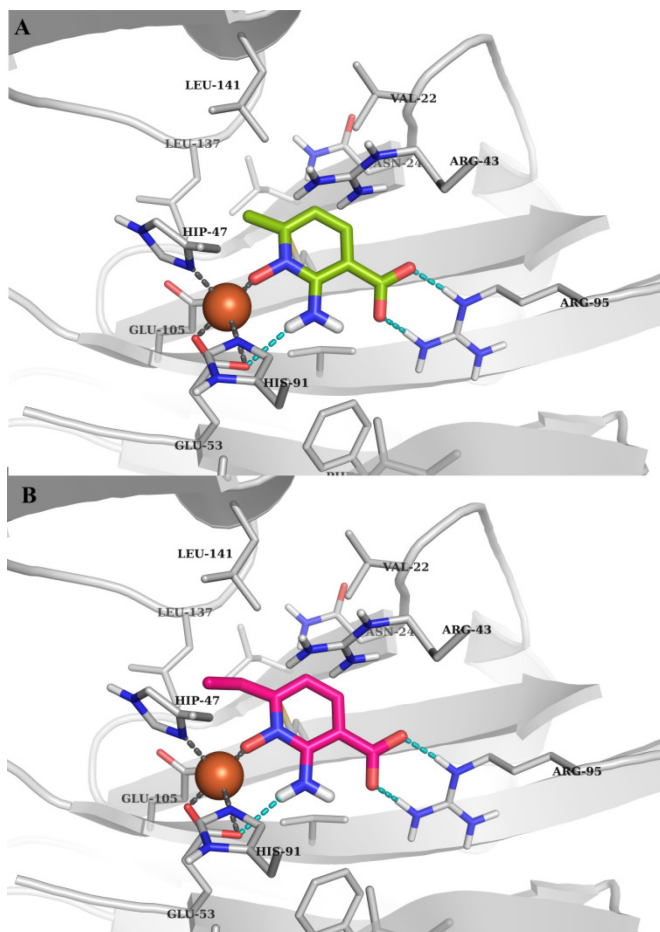
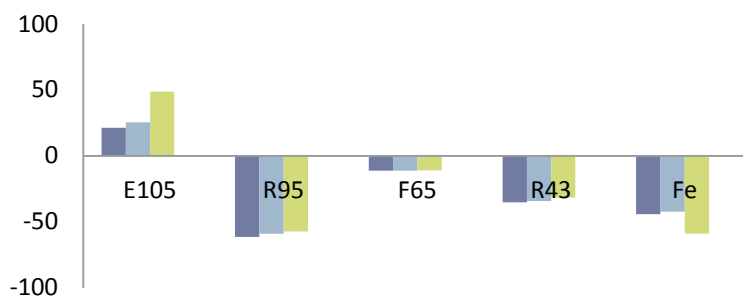


Figure 16. Graph showing the different contributions to Glide Score of the interaction of 3-hydroxyanthranilic acid, 4CIHAA and compound 3 with the main residues of the human model of 3-HAO. In purple-blue are reported the contribution of the interaction between 3-hydroxyanthranilic acid and 3-HAO, in light blue and in green those of 4CIHAA and compound 3, respectively. Residue numeration is relative to mammalian 3-HAO





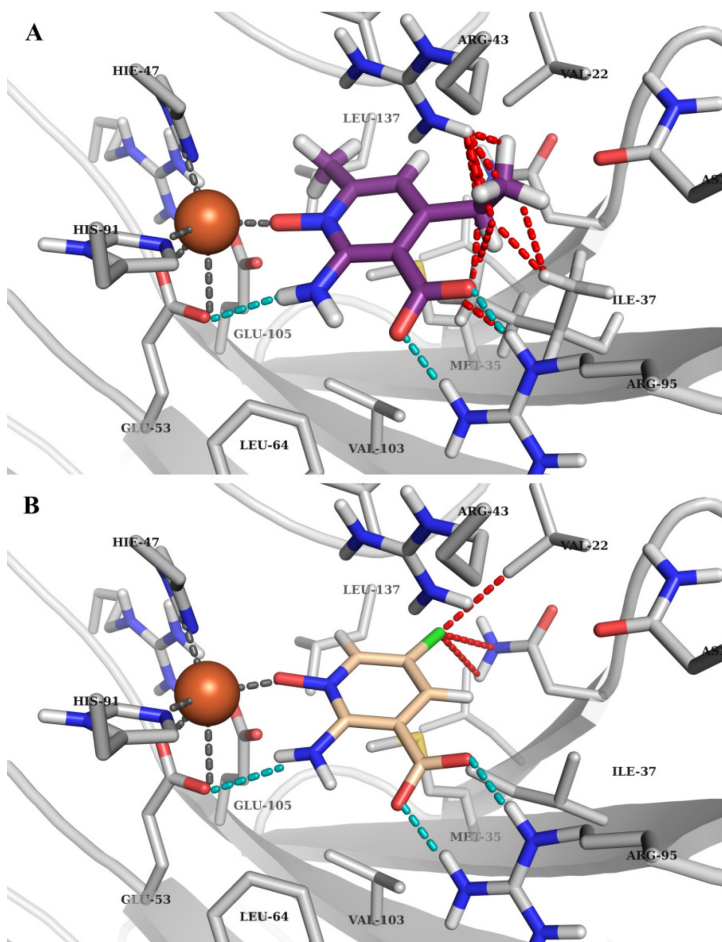
Docking poses generated for compounds with di-substitution at positions C<sub>4</sub> and C<sub>6</sub> of the pyridine ring (compounds **11-15**) present steric clashes with residues surrounding substituents at position C<sub>4</sub>, Val22 and Arg43, as shown in Figure 17.

Compounds **16-19** (substituted at position C<sub>5</sub> or di-substituted at positions C<sub>5</sub> and C<sub>6</sub>) displayed low inhibitory activity at 10 μM, whereas at 100 μM the inhibition was generally higher than 60%. A possible explanation for the poor activity may be related to the unfavorable contact of the compounds in the binding site that generate some small steric clashes.

These results, although in agreement with the small volume of the binding pocket, revealed a different behavior of the 2-amino nicotinic acid derivatives compared with the previously reported anthranilic derivatives. In fact, Anthranilic inhibitors with small alkyl substituents or halogens at positions C5 and C6 of the *o*-aminophenol ring,<sup>15</sup> equivalent to positions C<sub>4</sub> and C<sub>5</sub> of the 2-aminonicotinic acid 1-oxide nucleus, were reported to be active. This different behavior of the two series of compounds may be ascribed to a different mechanism of inhibition, although sharing the same binding mode.

On the basis of the mechanism of dioxygenation proposed by Zhang et al., one can speculate that the introduction of electron-withdrawing substituents in the anthranilic acid nucleus has the effect of blocking the single electron transfer from the electron-rich substrate to the catalytic iron, resulting in the production of a radical at position C<sub>3</sub> and in the restoration of the oxidation state of the metal atom, thus trapping the system in an unproductive oxidized state of the iron ion. Halogen substituents, in particular, can lower the HOMO energy of the *o*-aminophenol moiety, thus blocking this oxidation reaction. Moreover halogenated anthranilic derivatives are very reactive molecules, and may chelate the catalytic iron with high efficiency, allowing the accommodation of bulkier substituents by

**Figure 17. A) Pose of compound 14 and B) pose of compound 21. Steric clashes are evident for compound 14, whereas probably also an electronic component determines the unfavorable interaction of compound 21, which presents less evident steric clashes with the surrounding residues**



forcing the equilibrium towards the open or to a partially open conformation of the enzyme.

In contrast, our 2-aminonicotinate 1-oxide derivatives are inherently unsusceptible to oxidation at the pyridine nitrogen (formally equivalent to carbon 3 of the *o*-aminophenol moiety), and thus less sensitive to electronic effects. Therefore the driving force that shifts the equilibrium towards a more open conformation of the enzyme, able to accommodate bulkier substituents, is missing in our case, resulting in experimental inactivity for compounds presenting substituents at positions C<sub>4</sub> and C<sub>5</sub> of the 2-aminonicotinic acid 1-oxide nucleus.

## 2.5 CONCLUSIONS

In this study, we provided a rational explanation for the activity of a series of 2-aminonicotinic acid 1-oxide derivatives. Several compounds showed activity in human and rat brain homogenates, the most potent one, compound **3**, was also tested in a rat model of a neurodegenerative condition, where it was able to acutely shift the balance between neurotoxic (QUIN) and neuroprotective (KYNA) kynurenine metabolites towards the latter.

The adopted protocol that applies classical docking using QM assigned partial charges for compounds and binding sites, revealed to be efficient. Using QM/MM to assign QM approximated atomic charges to binding site and treating the rest of the protein with classical MM approximation, allowed us to have a more realistic picture of the actual charge distribution in the binding site. All the experimentally active compounds share the same binding mode, whereas part of the bulkiest derivative, do not fit into the binding site or fail to produce the correct binding pose, with the carboxylic moiety and the N-oxide group pointing towards the arginine residue and the iron atom respectively. The equilibrium between the open and the close conformation of the small loop on the top of the binding site is shifted towards the closed form upon ligand binding. The strong metal chelation exerted by the anthranilic derivatives is enough to force the equilibrium towards the open form to accommodate some relatively bulky substituents. However, in the case of the 2-aminonicotinic acid derivatives, the metal chelation is probably weaker than for the anthranilic derivatives, is not sufficient to ensure the accommodation of substituents at positions C5 and C6 of the pyridine ring, preventing these compounds to efficiently bind to the enzyme. At a visual inspection, the other inactive molecules that are correctly docked into the binding site, show steric clashes with the surrounding residues that probably are the cause of the loss of activity compared to smaller derivatives.

Analysis of the activity profile of the compounds can only be qualitative, due to the available activity data. IC50 data was only available

for the most active compound whereas for all the other compounds only the percentage of inhibition at 1mM, 100µM and 10 µM was available. Moreover, inhibitory activity has been measured with brain homogenate and not with the purified protein, so direct correlation would have been imprecise.

Clearly, several issues must be carefully considered for the further development of the described compounds described. Given the low molecular weight, the presence of the *N*-oxide, and the high value of the polar surface area, it can be expected that blood-brain barrier penetrability, as well as metabolic liability, will need extensive rounds of optimization. Tolerated substituents at position C6 of the pyridine ring can be further explored to ensure another point of anchorage to the enzyme, for example interacting with Cys134 or Arg108, and possibly stabilizing a more open conformation of the active site, allowing the substitution at the other positions of the pyridine ring.

## BIBLIOGRAPHY

1. Wolf, H., The effect of hormones and vitamin B6 on urinary excretion of metabolites of the kynurenine pathway. *Scandinavian journal of clinical and laboratory investigation. Supplementum* **1974**, *136*, 1-186.
2. Costantino, G., New promises for manipulation of kynurenine pathway in cancer and neurological diseases. *Expert Opinion on Therapeutic Targets* **2009**, *13*, 247-258.
3. Amori, L.; Guidetti, P.; Pellicciari, R.; Kajii, Y.; Schwarcz, R., On the relationship between the two branches of the kynurenine pathway in the rat brain in vivo. *J. Neurochem.* **2009**, *109*, 316-325.
4. Giorgini, F.; Guidetti, P.; Nguyen, Q. V.; Bennett, S. C.; Muchowski, P. J., A genomic screen in yeast implicates kynurenine 3-monooxygenase as a therapeutic target for Huntington disease. *Nature Genetics* **2005**, *37*, 526-531.
5. Guillemin, G. J.; Brew, B. J.; Noonan, C. E.; Takikawa, O.; Cullen, K. M., Indoleamine 2,3 dioxygenase and quinolinic acid immunoreactivity in Alzheimer's disease hippocampus. *Neuropathology and Applied Neurobiology* **2005**, *31*, 395-404.
6. Chen, Y.; Stankovic, R.; Cullen, K. M.; Meininger, V.; Garner, B.; Coggan, S.; Grant, R.; Brew, B. J.; Guillemin, G. J., The Kynurenine Pathway and Inflammation in Amyotrophic Lateral Sclerosis. *Neurotoxicity Research* **2010**, *18*, 132-142.
7. Maes, M.; Verkerk, R.; Bonaccorso, S.; Ombelet, W.; Bosmans, E.; Scharpé, S., Depressive and anxiety symptoms in the early puerperium are related to increased degradation of tryptophan into kynurenine, a phenomenon which is related to immune activation. *Life Sci.* **2002**, *71*, 1837-1848.
8. Stone, T. W.; Perkins, M. N., Quinolinic acid: A potent endogenous excitant at amino acid receptors in CNS. *Eur. J. Pharmacol.* **1981**, *72*, 411-412.
9. Stone, T. W.; Darlington, L. G., Endogenous kynurenines as targets for drug discovery and development. *Nat. Rev. Drug Discovery* **2002**, *1*, 609-620.
10. Cull-Candy, S. G.; Leszkiewicz, D. N., Role of distinct NMDA receptor subtypes at central synapses. *Science's STKE [electronic resource] : signal transduction knowledge environment* **2004**, *2004*.
11. Lau, C. G.; Zukin, R. S., NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. *Nature Reviews Neuroscience* **2007**, *8*, 413-426.

12. Paoletti, P.; Neyton, J., NMDA receptor subunits: function and pharmacology. *Curr. Opin. Pharmacol.* **2007**, *7*, 39-47.
13. Carroll, R. C.; Zukin, R. S., NMDA-receptor trafficking and targeting: Implications for synaptic transmission and plasticity. *Trends in Neurosciences* **2002**, *25*, 571-577.
14. Furukawa, H.; Singh, S. K.; Mancusso, R.; Gouaux, E., Subunit arrangement and function in NMDA receptors. *Nature* **2005**, *438*, 185-192.
15. Yao, Y.; Mayer, M. L., Characterization of a soluble ligand binding domain of the NMDA receptor regulatory subunit NR3A. *J. Neurosci.* **2006**, *26*, 4559-4566.
16. Groc, L.; Heine, M.; Cousins, S. L.; Stephenson, F. A.; Lounis, B.; Cognet, L.; Choquet, D., NMDA receptor surface mobility depends on NR2A-2B subunits. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 18769-18774.
17. Levine, M. S.; Cepeda, C.; André, V. M., Location, Location, Location: Contrasting Roles of Synaptic and Extrasynaptic NMDA Receptors in Huntington's Disease. *Neuron* **2010**, *65*, 145-147.
18. Milnerwood, A. J.; Gladding, C. M.; Pouladi, M. A.; Kaufman, A. M.; Hines, R. M.; Boyd, J. D.; Ko, R. W. Y.; Vasuta, O. C.; Graham, R. K.; Hayden, M. R.; Murphy, T. H.; Raymond, L. A., Early Increase in Extrasynaptic NMDA Receptor Signaling and Expression Contributes to Phenotype Onset in Huntington's Disease Mice. *Neuron* **2010**, *65*, 178-190.
19. Kudryashova, I. V., Synaptic and extrasynaptic NMDA receptors: Problems and prospects. *Neurochemical Journal* **2007**, *1*, 275-280.
20. Guillemain, G. J., Quinolinic acid, the inescapable neurotoxin. *Febs Journal* **2012**, *279*, 1356-1365.
21. Lugo-Huitrón, R.; Ugalde Muñiz, P.; Pineda, B.; Pedraza-Chaverrí, J.; Ríos, C.; Pérez-De La Cruz, V., Quinolinic acid: An endogenous neurotoxin with multiple targets. *Oxidative Medicine and Cellular Longevity* **2013**.
22. Rahman, A.; Ting, K.; Cullen, K. M.; Braid, N.; Brew, B. J.; Guillemain, G. J., The Excitotoxin Quinolinic Acid Induces Tau Phosphorylation in Human Neurons. *PLoS One* **2009**, *4*.
23. Kohler, C.; Eriksson, L. G.; Flood, P. R.; Hardie, J. A.; Okuno, E.; Schwarcz, R., Quinolinic acid metabolism in the rat brain. Immunohistochemical identification of 3-hydroxyanthranilic acid oxygenase and quinolinic acid phosphoribosyltransferase in the hippocampal region. *J. Neurosci.* **1988**, *8*, 975-987.
24. Foster, A. C.; Okuno, E.; Brougher, D. S.; Schwarcz, R., A radioenzymatic assay for quinolinic acid. *Anal. Biochem.* **1986**, *158*, 98-103.
25. Tavares, R. G.; Tasca, C. I.; Santos, C. E. S.; Alves, L. B.; Porciúncula, L. O.; Emanuelli, T.; Souza, D. O., Quinolinic acid stimulates synaptosomal glutamate

release and inhibits glutamate uptake into astrocytes. *Neurochem. Int.* **2002**, *40*, 621-627.

26. Naoi, M.; Ishiki, R.; Nomura, Y.; Hasegawa, S.; Nagatsu, T., Quinolinic acid: an endogenous inhibitor specific for type B monoamine oxidase in human brain synaptosomes. *Neuroscience Letters* **1987**, *74*, 232-236.

27. Santamaría, A.; Jiménez-Capdeville, M. E.; Camacho, A.; Rodríguez-Martínez, E.; Flores, A.; Galván-Arzate, S., In vivo hydroxyl radical formation after quinolinic acid infusion into rat corpus striatum. *NeuroReport* **2001**, *12*, 2693-2696.

28. Behan, W. M. H.; McDonald, M.; Darlington, L. G.; Stone, T. W., Oxidative stress as a mechanism for quinolinic acid-induced hippocampal damage: Protection by melatonin and deprenyl. *British Journal of Pharmacology* **1999**, *128*, 1754-1760.

29. Braidy, N.; Grant, R.; Adams, S.; Guillemin, G. J., Neuroprotective effects of naturally occurring polyphenols on quinolinic acid-induced excitotoxicity in human neurons. *FEBS Journal* **2010**, *277*, 368-382.

30. Vécsei, L.; Szalárdy, L.; Fülöp, F.; Toldi, J., Kynurenines in the CNS: Recent advances and new questions. *Nat. Rev. Drug Discovery* **2013**, *12*, 64-82.

31. Szalárdy, L.; Klivenyi, P.; Zadori, D.; Fuellep, F.; Toldi, J.; Vecsei, L., Mitochondrial Disturbances, Tryptophan Metabolites and Neurodegeneration: Medicinal Chemistry Aspects. *Current Medicinal Chemistry* **2012**, *19*, 1899-1920.

32. Hilmas, C.; Pereira, E. F. R.; Alkondon, M.; Rassoulpour, A.; Schwarcz, R.; Albuquerque, E. X., The brain metabolite kynurenic acid inhibits  $\alpha 7$  nicotinic receptor activity and increases non- $\alpha 7$  nicotinic receptor expression: Physiopathological implications. *J. Neurosci.* **2001**, *21*, 7463-7473.

33. Schwarcz, R.; Okuno, E.; White, R. J.; Bird, E. D.; Whetsell Jr, W. O., 3-Hydroxyanthranilate oxygenase activity is increased in the brains of Huntington disease victims. *Proceedings of the National Academy of Sciences of the United States of America* **1988**, *85*, 4079-4081.

34. Guidetti, P.; Luthi-Carter, R. E.; Augood, S. J.; Schwarcz, R., Neostriatal and cortical quinolinate levels are increased in early grade Huntington's disease. *Neurobiology of Disease* **2004**, *17*, 455-461.

35. Carlock, L.; Walker, P. D.; Shan, Y.; Gutridge, K., Transcription of the Huntington disease gene during the quinolinic acid excitotoxic cascade. *NeuroReport* **1995**, *6*, 1121-1124.

36. Heyes, M. P.; Rubinow, D.; Lance, C.; Markey, S. P., Cerebrospinal fluid quinolinic acid concentrations are increased in acquired immune deficiency syndrome. *Annals of Neurology* **1989**, *26*, 275-277.

37. Heyes, M. P.; Nowak Jr, T. S., Delayed increases in regional brain quinolinic acid follow transient ischemia in the gerbil. *Journal of Cerebral Blood Flow and Metabolism* **1990**, *10*, 660-667.

38. Beal, M. F.; Matson, W. R.; Storey, E.; Milbury, P.; Ryan, E. A.; Ogawa, T.; Bird, E. D., KYNURENIC ACID CONCENTRATIONS ARE REDUCED IN HUNTINGTONS-DISEASE CEREBRAL-CORTEX. *Journal of the Neurological Sciences* **1992**, *108*, 80-87.
39. Jauch, D.; Urbańska, E. M.; Guidetti, P.; Bird, E. D.; Vonsattel, J. P. G.; Whetsell Jr, W. O.; Schwarcz, R., Dysfunction of brain kynurenic acid metabolism in Huntington's disease: Focus on kynurenine aminotransferases. *Journal of the Neurological Sciences* **1995**, *130*, 39-47.
40. Heyes, M. P.; Saito, K.; Crowley, J. S.; Davis, L. E.; Demitrack, M. A.; Der, M.; Dilling, L. A.; Elia, J.; Kruesi, M. J. P.; Lackner, A.; Larsen, S. A.; Lee, K.; Leonard, H. L.; Markey, S. P.; Martin, A.; Milstein, S.; Mouradian, M. M.; Pranzatelli, M. R.; Quearry, B. J., Quinolinic acid and kynurenine pathway metabolism in inflammatory and non-inflammatory neurological disease. *Brain* **1992**, *115*, 1249-1273.
41. Guidetti, P.; Reddy, P. H.; Tagle, D. A.; Schwarcz, R., Early kynurenergic impairment in Huntington's Disease and in a transgenic animal model. *Neuroscience Letters* **2000**, *283*, 233-235.
42. Baran, H.; Jellinger, K.; Deecke, L., Kynurenine metabolism in Alzheimer's disease. *Journal of Neural Transmission* **1999**, *106*, 165-181.
43. Gold, A. B.; Herrmann, N.; Swardfager, W.; Black, S. E.; Aviv, R. I.; Tennen, G.; Kiss, A.; Lanctot, K. L., The relationship between indoleamine 2,3-dioxygenase activity and post-stroke cognitive impairment. *Journal of Neuroinflammation* **2011**, *8*.
44. Saito, K.; Nowak, T. S.; Markey, S. P.; Heyes, M. P., MECHANISM OF DELAYED INCREASES IN KYNURENINE PATHWAY METABOLISM IN DAMAGED BRAIN-REGIONS FOLLOWING TRANSIENT CEREBRAL-ISCHEMIA. *J. Neurochem.* **1993**, *60*, 180-192.
45. Rejdak, K.; Petzold, A.; Kocki, T.; Kurzepa, J.; Grieb, P.; Turski, W. A.; Stelmasiak, Z., Astrocytic activation in relation to inflammatory markers during clinical exacerbation of relapsing-remitting multiple sclerosis. *Journal of Neural Transmission* **2007**, *114*, 1011-1015.
46. Rejdak, K.; Bartosik-Psujek, H.; Dobosz, B.; Kocki, T.; Grieb, P.; Giovannoni, G.; Turski, W. A.; Stelmasiak, Z., Decreased level of kynurenic acid in cerebrospinal fluid of relapsing-onset multiple sclerosis patients. *Neuroscience Letters* **2002**, *331*, 63-65.
47. Ogawa, T.; Matson, W. R.; Beal, M. F.; Myers, R. H.; Bird, E. D.; Milbury, P.; Saso, S., KYNURENINE PATHWAY ABNORMALITIES IN PARKINSONS-DISEASE. *Neurology* **1992**, *42*, 1702-1706.
48. Erhardt, S.; Blennow, K.; Nordin, C.; Skogh, E.; Lindström, L. H.; Engberg, G., Kynurenic acid levels are elevated in the cerebrospinal fluid of patients with schizophrenia. *Neuroscience Letters* **2001**, *313*, 96-98.



49. Yamamoto, H.; Shindo, I.; Egawa, B.; Horiguchi, K., KYNURENIC ACID IS DECREASED IN CEREBROSPINAL-FLUID OF PATIENTS WITH INFANTILE SPASMS. *Pediatric Neurology* **1994**, *10*, 9-12.
50. Yamamoto, H.; Murakami, H.; Horiguchi, K.; Egawa, B., STUDIES ON CEREBROSPINAL-FLUID KYNURENIC ACID CONCENTRATIONS IN EPILEPTIC CHILDREN. *Brain & Development* **1995**, *17*, 327-329.
51. Nozaki, K.; Beal, M. F., NEUROPROTECTIVE EFFECTS OF L-KYNURENINE ON HYPOXIA ISCHEMIA AND NMDA LESIONS IN NEONATAL RATS. *Journal of Cerebral Blood Flow and Metabolism* **1992**, *12*, 400-407.
52. Stone, T. W., Development and therapeutic potential of kynurenic acid and kynurenine derivatives for neuroprotection. *Trends in Pharmacological Sciences* **2000**, *21*, 149-154.
53. Cugola, A.; Gavraghi, G., Indole antagonists of excitatory amino acids. **1993**.
54. Connick, J. H.; Heywood, G. C.; Sills, G. J.; Thompson, G. G.; Brodie, M. J.; Stone, T. W., Nicotynilalanine increases cerebral kynurenic acid content and has anticonvulsant activity. *General Pharmacology* **1992**, *23*, 235-239.
55. Rover, S.; Cesura, A. M.; Huguenin, P.; Kettler, R.; Szenté, A., Synthesis and biochemical evaluation of N-(4-phenylthiazol-2-yl)benzenesulfonamides as high-affinity inhibitors of kynurenine 3-hydroxylase. *J. Med. Chem.* **1997**, *40*, 4378-4385.
56. Chiarugi, A.; Moroni, F., Quinolinic acid formation in immune-activated mice: Studies with (m-nitrobenzoyl)-alanine (mNBA) and 3,4-dimethoxy-[N-4-(3-nitrophenyl) thiazol-2yl]-benzenesulfonamide (Ro 61-8048), two potent and selective inhibitors of kynurenine hydroxylase. *Neuropharmacology* **1999**, *38*, 1225-1233.
57. Beconi, M. G.; Yates, D.; Lyons, K.; Matthews, K.; Clifton, S.; Mead, T.; Prime, M.; Winkler, D.; O'Connell, C.; Walter, D.; Toledo-Sherman, L.; Munoz-Sanjuan, I.; Dominguez, C., Metabolism and pharmacokinetics of JM6 in mice: JM6 is not a prodrug for Ro-61-8048. *Drug Metabolism and Disposition* **2012**, *40*, 2297-2306.
58. Walsh, J. L.; Todd, W. P.; Carpenter, B. K.; Schwarcz, R., 4-halo-3-hydroxyanthranilic acids: Potent competitive inhibitors of 3-hydroxy-anthranilic acid oxygenase in vitro. *Biochem. Pharmacol.* **1991**, *42*, 985-990.
59. Linderberg, M.; Hellberg, S.; Bjork, S.; Gotthammar, B.; Hogberg, T.; Persson, K.; Schwarcz, R.; Luthman, J.; Johansson, R., Synthesis and QSAR of substituted 3-hydroxyanthranilic acid derivatives as inhibitors of 3-hydroxyanthranilic acid dioxygenase (3-HAO). *Eur. J. Med. Chem.* **1999**, *34*, 729-744.
60. Schwarcz, R.; Bruno, J. P.; Muchowski, P. J.; Wu, H.-Q., Kynurenines in the mammalian brain: when physiology meets pathology. *Nature Reviews Neuroscience* **2012**, *13*, 465-477.

61. Vallerini, G. P.; Amori, L.; Beato, C.; Tararina, M.; Wang, X. D.; Schwarcz, R.; Costantino, G., 2-Aminonicotinic acid 1-oxides are chemically stable inhibitors of quinolinic acid synthesis in the mammalian brain: A step toward new antiexcitotoxic agents. *J. Med. Chem.* **2013**, *56*, 9482-9495.
62. Zhang, Y.; Colabroy, K. L.; Begley, T. P.; Ealick, S. E., Structural studies on 3-hydroxyanthranilate-3,4-dioxygenase: The catalytic mechanism of a complex oxidation involved in NAD biosynthesis. *Biochemistry* **2005**, *44*, 7632-7643.
63. Siegbahn, P. E. M.; Haeffner, F., Mechanism for catechol ring-cleavage by non-heme iron extradiol dioxygenases. *J. Am. Chem. Soc.* **2004**, *126*, 8919-8932.
64. Bugg, T. D.; Ramaswamy, S., Non-heme iron-dependent dioxygenases: unravelling catalytic mechanisms for complex enzymatic oxidations. *Curr. Opin. Chem. Biol.* **2008**, *12*, 134-140.
65. Vaillancourt, F. H.; Bolin, J. T.; Eltis, L. D., The ins and outs of ring-cleaving dioxygenases. *Crit. Rev. Biochem. Mol. Biol.* **2006**, *41*, 241-267.
66. Li, X.; Guo, M.; Fan, J.; Tang, W.; Wang, D.; Ge, H.; Rong, H.; Teng, M.; Niu, L.; Liu, Q.; Hao, Q., Crystal structure of 3-hydroxyanthranilic acid 3,4-dioxygenase from *Saccharomyces cerevisiae*: A special subgroup of the type III extradiol dioxygenases. *Protein Sci.* **2006**, *15*, 761-773.
67. Dilović, I.; Gliubich, F.; Malpeli, G.; Zanotti, G.; Matković-Čalogović, D., Crystal structure of bovine 3-hydroxyanthranilate 3,4-dioxygenase. *Biopolymers - Peptide Science Section* **2009**, *91*, 1189-1195.
68. Colabroy, K. L.; Zhai, H. L.; Li, T. F.; Ge, Y.; Zhang, Y.; Liu, A. M.; Ealick, S. E.; McLafferty, F. W.; Begley, T. P., The mechanism of inactivation of 3-hydroxyanthranilate-3,4-dioxygenase by 4-chloro-3-hydroxyanthranilate. *Biochemistry* **2005**, *44*, 7623-7631.
69. Lendenmann, U.; Spain, J. C., 2-Aminophenol 1,6-dioxygenase: A novel aromatic ring cleavage enzyme purified from *Pseudomonas pseudoalcaligenes* JS45. *J. Bacteriol.* **1996**, *178*, 6227-6232.
70. Murakami, S.; Sawami, Y.; Takenaka, S.; Aoki, K., Cloning of a gene encoding 4-amino-3-hydroxybenzoate 2,3-dioxygenase from *Bordetella* sp. 10d. *Biochem. Biophys. Res. Commun.* **2004**, *314*, 489-494.
71. Kovaleva, E. G.; Lipscomb, J. D., Crystal structures of Fe<sup>2+</sup> dioxygenase superoxo, alkylperoxo, and bound product intermediates. *Science* **2007**, *316*, 453-457.
72. Mbughuni, M. M.; Chakrabarti, M.; Hayden, J. A.; Bominaar, E. L.; Hendrich, M. P.; Münck, E.; Lipscomb, J. D., Trapping and spectroscopic characterization of an FeIII-superoxo intermediate from a nonheme mononuclear iron-containing enzyme. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 16788-16793.
73. Protein Data Bank (PDB). <http://www.rcsb.org/pdb/home/home.do>

74. Cho, A. E.; Rinaldo, D., Extension of QM/MM docking and its applications to metalloproteins. *J. Comput. Chem.* **2009**, *30*, 2609-2616.
75. Burger, S. K.; Thompson, D. C.; Ayers, P. W., Quantum mechanics/molecular mechanics strategies for docking pose refinement: Distinguishing between binders and decoys in cytochrome c peroxidase. *J. Chem. Inf. Model.* **2011**, *51*, 93-101.
76. Hayik, S. A.; Dunbrack, R.; Merz, K. M., Mixed quantum mechanics/molecular mechanics scoring function to predict protein-ligand binding affinity. *Journal of Chemical Theory and Computation* **2010**, *6*, 3079-3091.
77. Sternberg, U.; Koch, F. T.; Brauer, M.; Kunert, M.; Anders, E., Molecular mechanics for zinc complexes with fluctuating atomic charges. *J. Mol. Model.* **2001**, *7*, 54-64.
78. Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S., A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* **2005**, *48*, 5437-5447.
79. *Protein Preparation Wizard; Epik version 2.0; Impact version 5.5; Prime version 2.1.*, Schrödinger, LLC, New York, NY, 2009.
80. Kleywegt, G. J., Validation of protein crystal structures. *Acta Crystallographica Section D-Biological Crystallography* **2000**, *56*, 249-265.



## **CHAPTER 3:**

### **Creation of a new database of Structural Alerts**



## 3.1 INTRODUCTION

An important issue in the very early stages of drug discovery is the identification of potentially reactive and toxic compounds.<sup>1</sup> Highlighting safety risks only in later stages of drug discovery is an incredible waste of money throughout the preceding steps.<sup>2</sup> Moreover Idiosyncratic Adverse Drug Reactions (IADRs), rare and bizarre side effects, probably induced by multiple different mechanisms,<sup>3, 4</sup> are usually noticed only after the drug has been introduced into the market and reached a wider number of people, representing an even bigger problem, both for patients and pharmaceutical industries.<sup>2</sup> Of the 454 new chemical entities approved in US and Canada between 1992 and 2011, the 30.7% received a black box warning or was either withdrawn from the US market by the FDA due to adverse reactions.<sup>5</sup> Given these data, it is clear that assessing the potential toxicity of a drug candidate as early as possible along the drug discovery/development pipeline, has several advantages.

Nevertheless understanding the cause of toxicity of a chemical entity to improve its safety profile is anything but easy. Adverse Drug Reactions (ADRs) are often associated with the primary pharmacology of the drug; these type of reactions can easily be detected with pharmacology-toxicology studies and usually exhibit a dose-response relationship.<sup>6</sup> More difficult to notice are instead those ADRs, and especially IADRs, related to drug metabolites, produced by drug biotransformations.<sup>7</sup> In a study of 2007, Guengerich and MacDonald reported a statistical analysis of Bristol-Myers Squibb on the causes of toxicity in animal models, underlying that in almost the 30% of the cases, the cause was target-related, but the same number of cases was found also for biotransformation-related toxicity. They also reported that channel inhibition counted for the 18% of the toxicity reports and immune-mediated toxicity in another 7% of the cases.<sup>2</sup>

Nowadays there is a huge interest in predicting drug metabolism, to avoid metabolic liabilities in new drug candidates reducing the percentage of failures due to drug bioactivation, on the grounds that reactive

metabolites can be involved in drug-drug interactions, genotoxicity, hepatotoxicity and immune-mediated adverse drug reactions.<sup>8</sup>

The main route of toxicity of these reactive metabolites is unspecific covalent binding to proteins, nucleic acids and other macromolecules. In particular the covalent binding to GSH(glutathione), while being a detoxifying mechanism to reduce the reactivity of the metabolites, can lead to the depletion of the GSH levels in cells, leading to oxidative stress toxicity.<sup>9-11</sup> DNA modifications due to covalent binding have been established being the cause of toxicity for some drugs, for example in acetaminophen induced IADRs.<sup>12, 13</sup> Also intrinsically electrophilic compounds may react with tissue nucleophiles, leading to toxicity<sup>14</sup>. However, the events following protein covalent modifications and leading to the relevant toxic biological events are difficult to delineate in many cases. Another aspect highlighting the difficulties in understanding the causes of toxicity, is that, despite several attempts made, none of the studies on covalent binding have revealed a target that can explain the biology underlying the toxic events.<sup>15</sup>

Even if a lot of efforts have been undertaken in the field of predicting the toxicological profile of new drug candidates, we are still far from the goal. One of the main reasons is that toxicity is a complex physiological and pathological event and *in silico* tools developed so far cannot deal with all the possible aspects, just with one or few of them.<sup>16, 17</sup>

One approach that has been widely used in past fifteen years is represented by quantitative structure-activity relationship (QSAR) models, that try to correlate structural properties of the compounds with their toxicity. However two main issues, the applicability domain and the interpretability of the models, have limited the use of QSAR.<sup>17</sup> The first drawback is intrinsically related to the QSAR approach, because we can only try to predict the toxicological profile of chemicals that are similar to the compounds used to build our model, for which we already have toxicity data. The second limitation can be viewed from two sides: one related to the type of descriptors used to built the model, sometimes not easy to be understood, the other related to the confidence in using our model, tightly connected to the type of data that are available to build the models. In fact, most of the available data are results of *in vitro* experiments, whereas a smaller amount of data is about *in vivo* (animal) toxicity tests, and even less data are available on human toxicity. However the aim of the models is to predict human toxicity, so predictions are not always straightforward.<sup>18</sup>



In the attempt to predict the ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) profile of drug candidates also protein 3D structures have been used.<sup>19</sup> Several crystal structures of cytochromes and other proteins related with drug metabolism have become available in the past ten years, increasing the number of structure based studies in this field.<sup>20</sup> Cytochromes P450 (CYPs) are the most studied metabolizing enzymes, however these protein are difficult to investigate, in part because of the presence of large and flexible binding sites, sometimes also able to accommodate more than one ligand simultaneously. Previous studies suggested that modeling small molecules binding to CYPs without considering protein flexibility, can lead to artifacts and erroneous predictions. The same problems occur in predicting ligand binding to P-glycoprotein (P-gp), the most studied ATP-binding cassette (ABC) transporters; moreover P-gp can recognize a broad spectrum of ligands, charged, neutral, linear, cyclic and aromatic compounds, and seems to have up to seven binding sites.<sup>19</sup> Given the complexity of the problem, prediction using 3D structures of the proteins is challenging and time consuming, therefore limiting this kind of approach to the analysis of a small number of compounds.

Other types of computational approaches have been used together or instead of the previous ones, in trying to give the best picture of the possible drug biotransformation processes. Reactivity-based techniques have been developed to predict compound liabilities using descriptors derived from the electronic structure of the molecule, applying semi-empirical quantum mechanical methods.<sup>16</sup> Molecular interaction fields (MIFs), that encode the variation of the interaction energies between a target molecule and chemical probes in 3D space, can be used to build 3D-QSAR models; this is the approached followed by MetaSite, however it predicts only cytochrome-mediated metabolism.<sup>21</sup>

The structural alerts (SAs) or “toxicophores” are another kind of approach that is used to address potential reactivity and toxicity of compounds. SAs directly associate molecular patterns of the compounds with toxicity, giving us qualitative information about the intrinsic chemical reactivity or the tendency to form chemically reactive metabolites upon biotransformation.<sup>7</sup> The efficiency of this approach is confirmed by some literature examples, even though SAs cannot be consider alone as predictors of toxicity, because they cannot by themselves predict the type and the frequency of ADRs that may arise, but should be used as complementary technique with other approaches, for example with QSAR models previously

described. However toxicophores are a valuable tool to screen databases of compounds. For example, it is a well established practice to apply reactivity filters before running virtual screening experiments and SAs can also be used to reduce the number of molecules to test before HTS assays. Functional groups susceptible to bioactivation reactions have been reviewed for the first time by S.D. Nelson in 1994<sup>22</sup> and it is the basis of all SAs regarding bioactivation that have been published later.

Recently, as High Throughput Screening (HTS) established as the main discipline in early drug discovery in pharmaceutical industries, new SAs have been proposed. These structural alerts identify substructural features, not related to toxicity and therefore not recognized by common filters, that have been connected to false positive results in vitro assays.<sup>23</sup> A compound may act as a false positive for several reasons, through aggregate formation,<sup>24</sup> interference with detection methods<sup>25</sup> (e.g. fluorescent compounds) or with the assay media, or by reacting in an unspecific manner with the target protein or with multiple proteins<sup>26, 27</sup>. While aggregation can be prevented or minimized by adding a surfactant to the assay media, unspecific protein binding, operated by the so called “frequent hitters” due to the high frequency with which they are found as hits in HTS, is more difficult to understand and to prevent. In 2010 Baell J.B. and Holloway G.A. published new substructure filters to reject compounds displaying this type of assay- interfering behavior, basing their study on the analysis of the large screening campaign results started in 2003 at their institution.<sup>27</sup> Among compounds displaying unspecific protein binding, are also included chemicals establishing strong non-covalent interactions, such as metal chelation mediated by hydroxamate derivatives in metalloproteinase assays.<sup>28</sup> Of course this class of SAs has little meaning in case of other type of experimental assays compared to the important role they can play in HTS to avoid or identify false positive results.

On the basis of experimental evidences, it is clear that SAs are a useful tool in early drug discovery to prioritize compounds, when the number of molecules considered is huge and more accurate approaches as QSAR or structure-based techniques cannot be applied. However it is important to underline that information conveyed by SAs should not be over-interpreted: for example some authors report phenyl ring as being a substructure prone to be bioactivated into an epoxide metabolite and therefore should be excluded<sup>1</sup>; even if this information is in its own right, almost all of the drug on the market present this feature, but only some of them present severe ADRs related to this metabolite. Thus, SAs information should be used

wisely and not to exaggerate the safety hazard associated with a compound presenting a potentially reactive substructure feature.

## 3.2 AIMS

SAs are very important in early drug discovery projects to identify intrinsically reactive or compounds prone to bioactivation. Several authors in the past years published SAs for different endpoints, for example for genotoxicity<sup>29</sup>, hepatotoxicity,<sup>30, 31</sup> or simple list of reactive groups<sup>32</sup> etc. The aim of the project was to build a SAs database to be included in ICM software<sup>33</sup> and to be also freely available as a webpage on the laboratory website, to filter or simply to flag compounds presenting a functional group recognized by one of the SAs. Our idea was to include in the SAs database all the meaningful SAs previously published, evaluating their relevance on the basis of the data available in the literature, excluding those too stringent or not supported with significant examples in literature. We decided to group the database entries in three different categories, according to the three types of SAs we collected: functional groups of intrinsically reactive molecules, structural features of chemicals susceptible to bioactivation and finally substructures related to assay interfering compounds. For each entry in the database, an in-depth analysis is reported on the webpage. Explanation of the reason of the alert creation, references to papers and ToxAlerts database, will be available on the webpage and are provided here in Appendix B.

## 3.3 MATERIALS AND METHODS

### 3.3.1. Source of Information

Scientific literature on structural alerts and toxicity endpoints has been examined together with two publicly available databases of SAs. The first one is the internet database ToxAlerts.<sup>1</sup> This database is very comprehensive, but with lot of redundant information and with several alerts not supported by experimental data but referring to filters of vendors catalogs. The second database is a collection of filtering rules used at Ely-Lilly and published in 2012.<sup>34</sup> These rules, even if containing several useful SAs, contain also other entries to exclude compounds without rings or with more than three rings, that are beyond the scope of this database. These filtering rules are associated with a penalty score, and compound are excluded only after a certain score, limiting in this way the influence of less important SAs. Penalty scores are also assigned to compounds according to their molecular weight, or to not-neutral compounds, preferring drug-like compounds. Conversely, our database is meant to simply consider the presence of liabilities in molecular structures possibly leading to toxicity or to HTS assays interference, without taking into account the drug-like concept that is nowadays called into question, contesting its suitability especially in the early stages of the drug discovery process<sup>35</sup>.

### 3.3.2 Structural Alerts definition using SMARTS

Smiles Arbitrary Target Specification (SMARTS) strings were used to represent the structural alerts contained in the database. SMARTS strings are a common representation for molecular substructures, because they can easily be applied in database screening and filtering. SMARTS strings are based on SMILES (Simplified Molecular-Input Line-Entry System)<sup>36</sup>, a line notation used to represent molecular structures using ASCII strings, but are specifically developed to represent molecular substructures, adding to

SMILES codification wildcards, connectivity descriptors and logical operators. For the creation of the SAs database the SMARTS notations available in ICM software suite have been used (available in table 1), which along with canonical SMARTS present some extensions:

- $\wedge n$ : that indicates the hybridization of the atom, e.g.: [C;  $\wedge 2$ ] indicates an sp<sup>2</sup> hybridized carbon atom.;
- $yn$ : that represents the number of the ring in the molecular substructure representation the atom belongs to, e.g.: [C;y1] indicates a carbon atom belonging to the first ring;
- $Yn$ : is used to indicate the minimum number of hydrogen atoms bonded to the considered atom, e.g.: [N,Y2] represents a nitrogen atom with at least two hydrogen atoms attached.

**Table 1. SMILES and SMARTS symbol used in the database**

Symbol	Description	Examples
*	any atom	*
a	Aromatic atom	aN(=O)O
A	Aliphatic atom	AAA
C	aliphatic carbon	
c	aromatic carbon	
[#n]	Atomic number	[#6] any carbon atom
Dn	the number of heavy neighbors	[*;D2] any atom with two non-H connections
Hn	Precise number of attached hydrogens	[*;H2]
Rn	the number of rings the atom belongs to	[#6;R2] any carbon in two rings
rn	the size of smallest ring the atom belongs to	[*;r6]
vn	valence, sum of bond orders of all neighbors	
Xn	Total number of neighbors including heavy atoms and hydrogens	
-n	negative charge	[--], [-2]
+n	positive charge	[++], [+2]
$\wedge n$	sp <sup>1</sup> ,sp <sup>2</sup> ,sp <sup>3</sup> hybridization	[C; $\wedge 2$ ] sp <sup>2</sup> carbon
yn	ring number in SSSR	[*;y1] any atom which belongs to the first ring
Yn	number of at least attached hydrogens	[*;Y2] atom with two or more hydrogens
@	anticlockwise chirality	C[C@H](F)O
@@	clockwise chirality	
~	any bond	C~C
:	aromatic bond	c:c
-,=#	single, double and triple bonds	C#C
=&!@	bond SMART notation for double, not in ring	acC=&!@Cca
!primitive	negation	[!C] non-aliphatic carbon, [*;!R] any atom not in a ring
expr1&expr2	logical and (high precedence)	[c,n&H1] any arom carbon OR H-pyrrole nitrogen
expr1,expr2	logical or	[C,N,O] C or N or O
expr1;expr2	logical and (low precedence)	[c,n;H1] arom carbon OR nitrogen with one hydrogen

### 3.3.3 Database Structure

Database entries are divided into three categories according to the type of structural alert they represent, and each group is assigned a different level of hazard (“Rank”):

- the first group contains intrinsically reactive functional groups, mainly electrophiles that can react with protein nucleophiles giving covalent adducts; this group contains the most dangerous SAs and are collected in “Rank 3” group;
- the second category contains substructure susceptible to metabolic activation; in this group are present those substructure that can potentially evolve in a reactive metabolite but are not toxic per se (“Rank 2” group);
- the less dangerous group is the one called “Rank 1” that put together features of promiscuous compounds and other SAs with contrasting or ambiguous reports.

Generally when a structural alert can be ascribed to more than one category, it is reported in the one related to the higher level of risk.


Each entry in the database is characterized by the following information, corresponding to different fields of the database:

- ID: a unique identification string for each entry, constituted by four characters;
- NA: the name of the functional group of substructure;
- SM: the SMARTS string representing the structure;
- RK: the Rank, i.e. the subgroup belonging to (Rank 1, 2 or 3);
- DE: the description of the structural alert, that is the type of problem it represents, i.e. covalent binding, reactive metabolites formation etc. and, if available, also an explanation of the mechanism of action of the functional group;
- DR: database reference, the reference to ToxAlerts ID code and to Wikipedia;
- RF: reference to papers reporting the alert;
- RE: for Rank 3 alerts, the amino acidic residues usually involved in the covalent modification;
- KW: keywords useful for searching through the database;

The database will be made available as an HTML webpage.

STRUCTURE ALERT
UC San Diego

Search
Structure Alert
genotox
x
Q



**β-haloamines; haloethylamine, N-mustard**

C(!@C[Cl.Br.I])!@N

Rank: 3

beta-haloamine also called N-mustard are non-specific DNA alkylating agents, that have been used as anticancer agents; they act undergoing first an intramolecular cyclization that forms an aziridinium intermediate, and then reacting with nucleophilic centre on guanine base in DNA strands; they can also react with cysteine (and lysine) residues through an SN2 reaction to form a covalent adducts; some haloethylamine derivatives have been also reported to be involve in acute aquatic toxicity;

ToxAlerts.TA414; ToxAlerts.TA624; ToxAlerts.TA344; ToxAlerts.TA362; ToxAlerts.TA435; ToxAlerts.TA687; ToxAlerts.TA810;

wiki: Nitrogen\_mustard;

- 🔗 [Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology.](#)
- 🔗 [A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity.](#)
- 🔗 [Derivation and validation of toxicophores for mutagenicity prediction.](#)
- 🔗 [Electrophiles and acute toxicity to fish.](#)

genotox; carcinogen;

*haloamines; haloethylamine; N-mustard; genotox (N-mustard)*

Figure 1. Snapshot of the webpage of database (still under construction).



## 3.4 RESULTS AND DISCUSSION

SAs represent a valuable and well established tool in early drug discovery. After the analysis of the available literature, the newly created database contains 144 entries, organized into three different categories corresponding to the different level of danger represented by the molecular liabilities: into the “Rank 3” subgroup are included 42 molecular patterns responsible for covalent protein binding, “Rank 2” contains 28 SAs referring to substructures that can potentially be bioactivated into reactive metabolites, resulting in covalent adducts with proteins or in oxidative stress due to depletion of GSH levels in cell. Finally “Rank 1” grouped 74 molecular substructures related with assays interference and SAs that do not present a completely clear correlation with a toxicity event. Appendix B contains a printed version of the database, before its elaboration into a webpage, listing all the entries and reporting the information in the fields previously described. A more in depth description of the three categories of the database is reported in the following paragraphs, in particular analyzing those entries providing a high number of matches when profiled against DrugBank (version 3.0) database<sup>37</sup>. DrugBank is a public database containing comprehensive information about approved, experimental, withdrawn drugs. Version 3.0 collects 6515 entries of experimental and approved small molecule drug entries that we used to profile our database.

### 3.4.1 Rank 3

Almost all the entries of this subgroup of the database are electrophiles exerting their toxicity through protein covalent binding. DrugBank vs.3 contains 429 entries that are recognized by these class of structural alerts. The highest number of DrugBank entries is recognized by the aldehyde, thiol, michael’s acceptor and epoxide structural alerts (Table 2 and Figure 1). For all the entries of this category, several reports relating these functional groups with toxicity endpoints, are available in literature and are

listed in the RF field of the database. Aldehydes can react with active site serine or cysteine residues forming hemiacetals, and, even if the reaction is usually reversible, the time scale can vary greatly.<sup>14</sup> Aliphatic aldehydes and a small number of aromatic aldehydes may also interact with protein amino groups, for example with lysine residues, forming a Schiff-base adducts through a nucleophilic addition reaction.<sup>38,39</sup>

Thiols have the potential to undergo a  $S_N2$  reaction with cysteine residues in proteins, creating a disulfide bridge. If the shape and size of thiol-compound are compatible with the catalytic site of metalloenzymes (generally in case of small thiol-containing derivatives), it may also nonspecifically interact with the catalytic metal ion, forming a coordination complex that block the enzyme activity.<sup>14,28</sup>

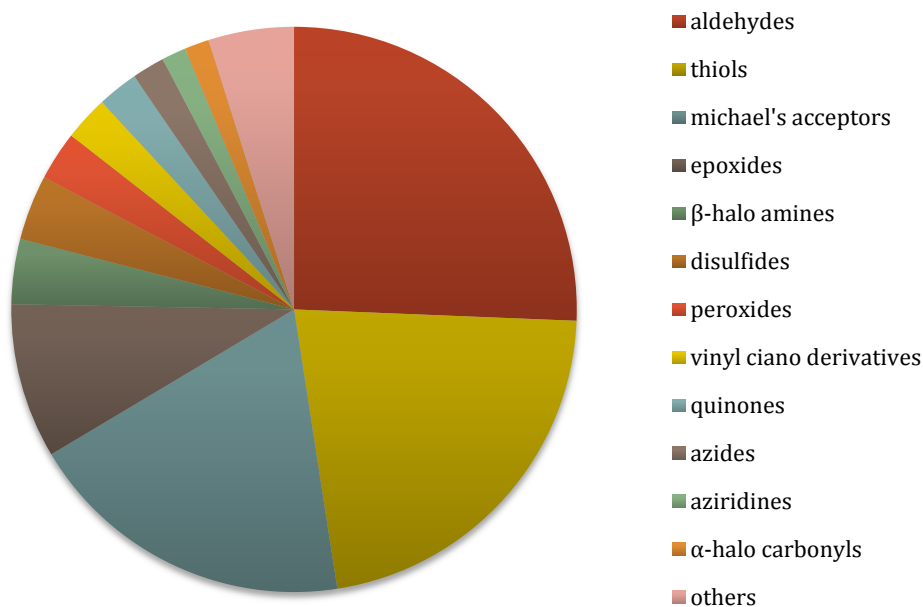
Michael's acceptors are  $\alpha$ - $\beta$  unsaturated carbonyl compounds. In this category are group together  $\alpha$ - $\beta$  unsaturated ketones, acrylates and acrylamides given the similarity in reactions and endpoints. All these functional groups can behave as acylating agents, reacting with biological nucleophiles, in proteins and/or nucleic acids, upon the electron deficient  $\beta$ -carbon atom.<sup>38</sup> The lower is the electron density at the  $\beta$ -carbon, the greater is the likelihood of undergoing nucleophilic attack. An example of Micheal's acceptors reaction is represented by  $\alpha$ - $\beta$  unsaturated ketones reacting with active serine or cysteine of serine or cysteine hydrolases respectively, irreversibly inhibiting the enzyme or with slow recovery rate.<sup>40, 41</sup> In literature are also reported some evidences that some polarized alkenes do not act directly as acylating agents, but are metabolically activated by conversion to their epoxides, and then the epoxide derivatives form the protein or DNA adducts.

Epoxides are unstable strained heterocycles, highly reactive and prone to ring opening at C-O bond, forming covalent adducts with biological nucleophiles through a  $S_N2$  reaction;<sup>12, 14, 41</sup> due to their high reactivity epoxides have been shown to act as alkylating agents and often adducts formed by reaction with epoxides are more stable than those formed by Michael's acceptors, thus with less chances of being hydrolyzed to recover the functional form of the enzyme.

The fifth category is represented by  $\beta$ -halo amines, a class of derivatives used as anticancer agents, in particular di(haloethyl)amines, also known as N-mustard.<sup>13</sup> Nitrogen mustards first undergo an intramolecular cyclization forming an aziridinium intermediate, that then react with nucleophilic centre on guanine bases in DNA strands;<sup>12</sup> they can also

react with cysteine (and lysine) residues through an SN2 reaction to form a covalent adducts.<sup>38</sup>

Another group is the one constituted by disulfide-containing derivatives. The disulfide bridge in physiological environment can break, forming two thiol-derivatives that can undergo the same reactions highlighted for thiol-containing compounds.<sup>14, 42</sup> With this alert we decided to address only linear



**Figure 2. Pie chart representing the distribution of the recognized functional groups.**

**Table 2 Number of recognized entries in DrugBank by Rank3 subgroup.**

RANK 3	n of matches
Aldehydes	110
thiols	94
michael's acceptors	81
Epoxides	38
β-halo amines	16
Disulfides	16
Peroxides	12
Vinyl-cyano derivatives	11
quinones	10

Azides	8
Aziridines	6
$\alpha$ -halo carbonyls	6
Others	21

disulfide bridges, given the higher degree of stability of those embedded in a ring system.

Peroxides are generally unstable, and peroxide-containing molecules are susceptible to form oxygen radicals, that can eventually cause protein, lipid and DNA oxidation<sup>38</sup>. Even if examples of stable peroxide bridges are available,<sup>43</sup> to avoid the possibility of radical-mediated toxicity, they are commonly unwanted in lead candidate compounds, therefore we decided to consider them in our SAs.

Vinyl-cyano derivatives, or acrylonitriles, are a very reactive class of Michael's acceptors, that can undergo Michael addition due to the attack of a biological nucleophile upon the electron deficient beta-carbon.

Quinones are another important class of SAs, because they can undergo a Michael's addition reaction with biological nucleophiles. The possibility of being metabolized, forming potential reactive epoxide-containing intermediates, increases the potential toxicity of quinones.<sup>34, 42</sup>

Also azide, aziridine-containing and  $\alpha$ -halo carbonyl compounds are common SAs in drug discovery. Some azide derivatives are mutagens and are listed as carcinogenic compounds.<sup>41</sup> Aziridines are strained three-membered rings, that behave like epoxides, reacting as alkylating agents after ring opening.<sup>12, 14</sup> Finally, in  $\alpha$ -halo carbonyls the reactivity of the sp<sup>3</sup> carbon bound to the halogen atom is increased by the carbonyl moiety in  $\alpha$  position. Thus, the sp<sup>3</sup> carbon atom undergoes a nucleophilic attack by an endogenous nucleophile, forming a new covalent bond through a SN<sub>2</sub> reaction.

### 3.4.2 Rank 2

Entries in this category are related with the possibility of reactive metabolites formation due to bioactivation. Drug metabolism is a crucial step, that can determine the length of drug action, prodrug activation, drug-

drug interactions and can limit or increase drug-related toxicity. In fact, several drugs have been withdrawn from market or acquired a black box warning due to ADRs or IADRs mediated by one, or more than one, reactive metabolite.

As previously discuss, we decided to not include phenyl rings as structural alerts, even if the metabolism of the phenyl ring proceeds through the formation of an epoxide intermediate, with the idea to not exaggerate the hazard possibility related to drug bioactivation.

The two subgroups that recognize almost two thirds of all the compounds matched by this type of SAs in DrugBank are thiophene/furan and aniline.

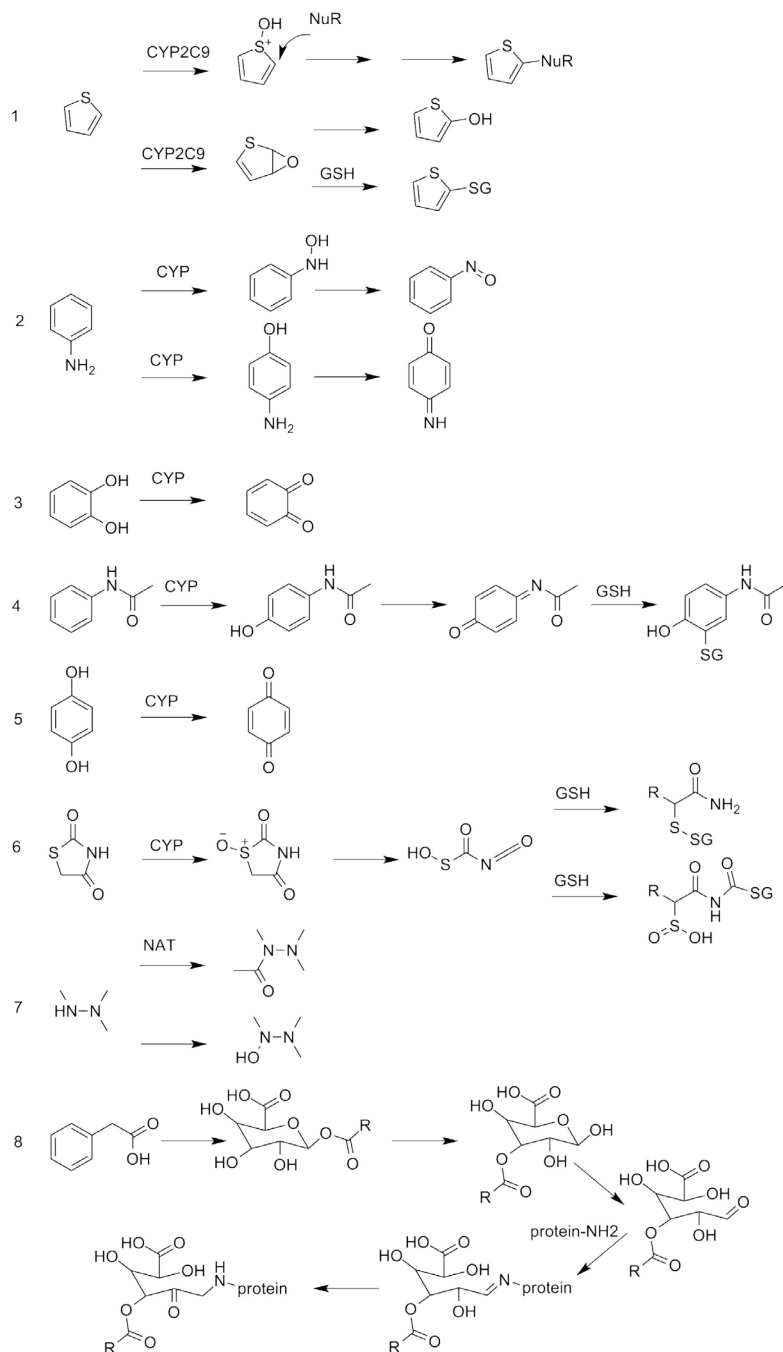
Five-membered heteroaromatic rings such as thiophenes, furans, pyrroles and their benzo-fused derivatives are widely used in medicinal chemistry. However they are susceptible to biotransformation into reactive metabolites involved in drug toxicity. For the same reason we excluded phenyl rings from our alerts, we decided to limit the SA for five-membered heterocycles to furans and thiophens, which have been clearly related to toxicity endpoints, excluding pyrroles and all their benzo-fused derivatives. Thiophene rings are bioactivated by CYP2C9, and can either form an S-oxide or an epoxide intermediate (as shown in Figure 3). S-oxide intermediates are unstable and react easily with biological nucleophiles, often with nucleophilic residues present in CYP2C9 catalytic site, leading to inactivation of the cytochrome. The epoxide intermediate can follow two different routes: or an immediate reaction with glutathione, or the epoxide can hydrolyze forming an hydroxyderivative, that in cases of electron-deficient thiophene rings can lead to ring opening, forming the corresponding  $\alpha$ - $\beta$ -unsaturated aldehyde.<sup>44</sup> To this category belongs Tienilic acid, a diuretic drug that was withdrawn from the market due to several cases of hepatotoxicity, caused by CYP2C9-mediated bioactivation of the thiophene ring.<sup>45</sup>

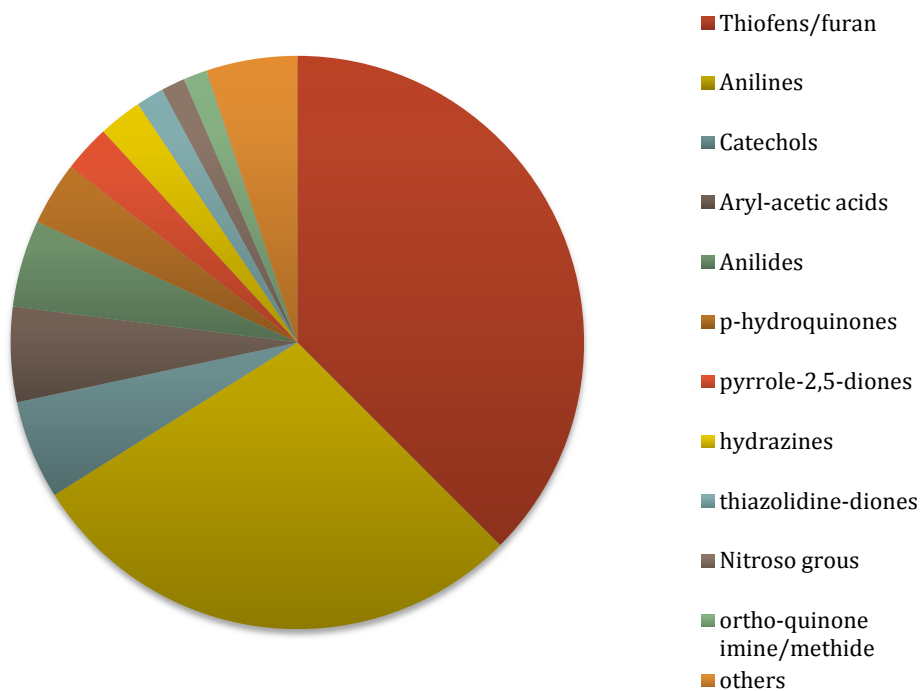
The second group is constituted by aniline derivatives, that, although commonly used as functional groups in drugs and drug candidates, have long been associated with chemical carcinogenesis. Their possible genotoxic effect is mediated by a reactive metabolite originated by CYP-mediated biotransformation. Primary arylamines can either undergo N-hydroxylation or ortho para hydroxylation, both operated by cytochromes. The unstable N-hydroxylamine can immediately be conjugated to sulphate or acetate, leaving groups that can later lead to the formation of a nitrenium ion, or be further oxidized to a nitroso intermediate. The ortho or para hydroxylated

derivative can be subsequently oxidized to a reactive quinone-imine metabolite. Both nitroso and quinone-imine derivatives can be trapped by GSH, but it is also possible a reaction with protein nucleophiles, leading to hepatotoxicity.<sup>46</sup>

Catechols, p-hydroquinones and o/p-quinone imine or methide are all quinone-related SAs. Quinones, quinone-methides or quinone-imines are reactive compounds that may react with biological nucleophiles. When they are formed in liver during drug biotransformation they react with GSH, forming stable adducts, that can lead to depletion of GH level in hepatocytes. Compounds presenting a catechol moiety can be conjugated with glucuronic acid or oxidized by CYP to o-quinone derivatives, whereas p-hydroquinones can be activated to p-quinone metabolites. Ortho or para quinone-imines or quinone-methides can also react with GSH and protein nucleophiles; their precursors, after being oxidized by CYP, undergo the same type of reactions.<sup>47</sup>

**Figure 3. Representation of the functional groups undergoing biotransformation in human liver. 1) thiophene; 2) aniline; 3) o-hydroxyphenol 4) anilide; 5) p-hydroxyphenol; 6) thiazolidinedione; 7) hydrazine; 8) aryl-acetic acid.**





**Figure 4 . Pie chart representing the distribution of the recognized functional groups by alerts collected in Rank2 subgroup.**

**Table 3 Number of recognized entries in DrugBank by Rank2 subgroup.**

RANK 2	n of matches
Thiofens/furan	168
Anilines	128
Catechols	25
Aryl-acetic acids	24
Anilides	22
p-hydroquinones	16
pyrrole-2,5-diones	12
Hydrazines and hydrazide	19
thiazolidine-diones	7
Nitroso grous	6
ortho-quinone imine/methide	6
Others	23



Aryl-acetic acid derivatives are widely used in NSAIDs (Non Steroidal Anti Inflammatory Drugs). However the aryl-acetic portion of these molecules have been related to the hepatotoxicity potential of these drugs, and some of them have also been withdrawn from the market. The hepatotoxic effect is believed to be immune-mediated and ascribed to the  $\beta$ -1-O-acyl glucuronide metabolites,<sup>48, 49</sup> as shown in Figure 3, that is the main metabolites of many NSAIDs, also of the non-toxic ones. The difference between toxic and non-toxic derivatives has been shown to be related to the substitution on the  $\alpha$ -carbon: aryl-acetic derivatives with an alkyl substituent on the  $\alpha$ -carbon (for example aryl-propionic acid derivatives) are less reactive against protein nucleophiles than the unsubstituted ones, suggesting both an electronic and steric effect of the substituent. The unsubstituted glucuronide derivatives have been shown to modify proteins through a simple transacylation reaction, forming a covalent adduct.<sup>50</sup> Therefore, given the safety profile of 2-substituted aryl-acetic acid derivatives as ibuprofen, this entry of SAs database clearly refers only to the unsubstituted moieties.

Other two entries of the database that have been highlighted to possibly be transformed into reactive metabolites are thiazolidinediones and hydrazine and hydrazides. The first ones can be oxidized by CYP3A4 on the sulfur atom, causing an immediate opening of the thiazolidine ring to form S-hydroxyl-isocyanate metabolites, that can react with GSH either on the sulfur atom or on the isocyanate moiety.<sup>51</sup> The latter conjugate also leads to the oxidation of the sulfur atom into a sulfate derivative. This type of derivatives is, the main metabolite of the thiazolidinedione-containing anti-diabetic drugs (rosiglitazone, pioglitazone and troglitazone) and has been shown to be able to inhibit the ATP-binding cassette transporter bile salt export pump (BSEP) for all these drugs.<sup>52</sup> In the specific case of troglitazone, that has been withdrawn from the USA market by FDA in 2000 three years later having been introduced, there are also other metabolites of the o-alkylphenyl portion that contribute to the hepatotoxic effect, together with an higher daily dosage, if compared with the other two compounds of the same class.<sup>53</sup> Hydrazines and hydrazides are metabolized by CYP and then further conjugated with acetyls. However CYP-mediated oxidation may lead to the production of radicals, especially in case of not fully-substituted nitrogen atoms.<sup>54, 55</sup> In this case an hydrazone ion can be formed, that can easily react with proteins to form covalent adducts.<sup>56</sup> Since the safety concerns are mainly related to the partially unsubstituted hydrazines and hydrazides, fully substituted derivatives are not considered by the SA entry.

### 3.4.3 Rank 1

In this category containing 74 entries are listed all those entries related with minor intrinsically or bioactivation-related safety problems and with assay interference. The highest number of entries is retrieved by the undesired atoms alert, filtering atoms such as silicon or boron, that sometimes are used as carbon bioisosteric replacement but are generally unwanted in lead compounds.

The second largest group is represented by ammonium quaternary salts. The elevated number of matches is explained by their widespread use in anticancer drugs, due to their ability to directly interact with DNA, and particularly aryl compounds with quaternary nitrogen are often involved with DNA intercalation.<sup>34</sup> Moreover quaternary amines usually do not cross the gastro intestinal barrier, therefore are unsuitable for oral administration of systemic drugs.

Four-membered lacton and lactam rings or their thio-analogs are very common in antibacterial drugs, constituting the active part of penicillin and cephalosporine antibiotics. They contain a strained ring system that can easily undergo ring opening reactions, leading to instability problems.<sup>34</sup> Furthermore even if  $\beta$ -lactamic antibacterial drugs are generally well tolerated, they are frequently associated with drug allergies and anaphylaxis reactions.<sup>57, 58</sup>

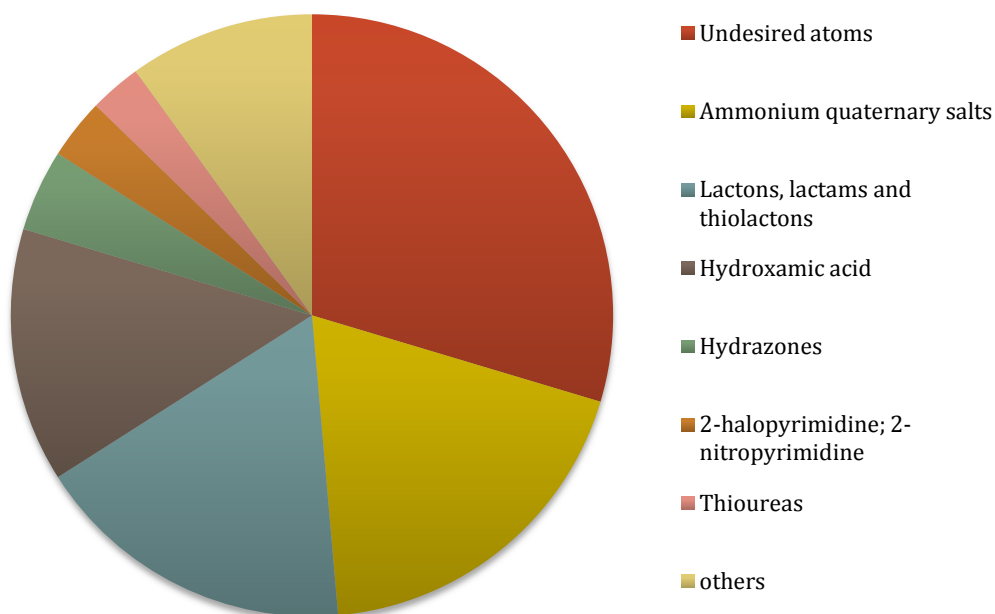
Another category of SAs is hydroxamic acids, that are connected with problems in metal ions chelation as well as reaction with biological thiols; metal ion chelation also represents a common artifact in biological screening, therefore in vitro assays of hydroxamic acids should properly handle this possibility.<sup>28</sup>

Hydrazones are electrophilic functional groups that can bind to proteins in an unspecific way, especially forming aggregates in in-vitro assays consequently giving false positives results;<sup>27, 34, 38</sup> some hydrazones derivatives gave positives results in AMES assay for genotoxicity, however many others did not, making it difficult to directly relate the genotoxic outcomes to the hydrazone functionality.<sup>59, 60</sup>

Polyhalogenated aromatic and heteroaromatic rings have been related to different toxicity outcomes and also been linked to interference in HTS assays. Aromatic rings with two or more activating groups as halogens or nitrogroups, are prone to undergo  $S_NAr$  with protein nucleophiles.<sup>14, 34, 38</sup>

Compounds containing thioureas have been reported by some authors to undergo metabolic bioactivation into reactive metabolites, but this point

of view is not well established and some authors do not consider thioureas as source of safety concerns.<sup>2, 61</sup> An example of thioureas undergoing metabolic bioactivation is Propylthiouracil, an antithyroid drug used in treatment of Graves's disease, that is associated with several reactive metabolites, probably generated from the oxidation of the thiourea. However the connection between toxic effect and reactive metabolites formation has not been clearly assessed yet, and deserves further investigations.<sup>13, 62, 63</sup>



**Figure 5.** Pie chart representing the distribution of the recognized functional groups recognized by SA in DrugBank

**Table 4.** Number of recognized entries in DrugBank by each entry.

RANK 1	n of matches
Undesired atoms	128
Ammonium quaternary salts	82
Lactons, lactams and thiolactons	75
Hydroxamic acid	59
Hydrazones	19
2-halopyrimidine; 2-nitropyrimidine	14
Thioureas	12
others	43

### 3.4.4 Profiling Of Best Selling Drugs

Out of a set of 136 drugs taken from the list of the top 200 best selling drugs in US in 2012(excluding duplicates and biological drugs)<sup>64</sup> a total of 19 drugs failed to pass the SAs rules, representing the 15.44 % of the marked small molecule drugs. The complete list of drugs failing the rules is available in table 4.  $\beta$ -lactamic antibiotics matched the SA for lactons and lactams. Aniline is the reason of three rejections, even though no reactive metabolites have been found for Darunavir and Mesalazine, while the hepatic metabolism of Lenalidomide has not been studied. Tiotropium is recognized by three SAs: epoxide, quaternary nitrogen and thiophene; since it is mainly used topically by inhalation, the absorbed and metabolized fraction is very low, thus avoiding the possible formation of reactive metabolites. Colesevelam is not absorbed but acts directly in the gut, and the SA quaternary nitrogen is useful in preventing the drug from being absorbed. The same reasoning is valid also for ipratropium, but applied to lungs. Also sevelamer, with an epoxide SA, is not absorbed, thus do not present any problem of protein alkylation due to ring opening. It is interesting to notice that atorvastatin presents an alkyl-pyrrole moiety, that has been reported to cause HTS assay interference, even if the mechanism of this interfering action has not been clarified yet.

**Table 5 results of the profiling of the best selling drugs in US using the SA database.**

<b>Drug</b>	<b>Indication</b>	<b>SA</b>
<i>Bortezomib</i>	multiple mieloma	undesired atoms
<i>Darunavir</i>	anti HIV	Aniline
<i>Lenalidomide</i>	Anemia	Aniline
<i>mesalazine</i>	Antinflammamtory	Aniline
<i>acetaminphen</i>	antipyretic and analgesic	Anilide
<i>sevelamer</i>	Hyperphosphatemia	Epoxide
<i>tiotropium</i>	Bronchodilator	epoxide; quaternary alkyl nitrogen; thiophene
<i>pioglitazone</i>	Antidiabetic	Thiazolidinedione
<i>atazanavir</i>	Antiretroviral	hydrazine
<i>colesevelam</i>	antihypercholesterolemia	quaternary alkyl nitrogen
<i>ipratropium bromide</i>	Bronchodilator	quaternary alkyl nitrogen
<i>mometasone</i>	Asthma	$\alpha$ -halo carbonyl
<i>rosuvastatin</i>	antihypercholesterolemia	4-vinylpyrimidine
<i>duloxetine</i>	Antidepressive	thiophene
<i>ezetimibe</i>	antihypercholesterolemia	$\beta$ -lactam
<i>piperacillin</i>	Antibiotic	$\beta$ -lactam
<i>tazobactam</i>	Antibiotic	$\beta$ -lactam
<i>atorvastatin</i>	antihypercholesterolemia	alkyl-pyrrole
<i>Bendamustine</i>	Antineoplastic	$\beta$ -halo amine
<i>Rilpivirine</i>	Anti HIV	Vinyl ciano

**Table 6. comparison of our SA, ElyLilly Rules and ToxAlert DB.**

<b>SAs DATABASE</b>	<b>Number of Structures of Top 200 Selling Drugs</b>	<b>% Matches</b>
<i>Our database</i>	136	15.44
<i>Ely Lilly Rules (data from ref 34)</i>	123	30.08
<i>ToxAlerts*</i>	136	98.53

\*Data from ToxAlerts were obtained using all the available entries in the website

## 3.5 CONCLUSIONS

SAs are a widely used approach in drug discovery to avoid reactive chemicals or compounds susceptible to bioactivation. Different sets of SAs are available in literature or from various websites, handling different levels of structure liabilities. Some collections of different types of SAs are available, sometimes containing redundant information or in other cases SAs without any clear explanation for all of them. Our intent was to assemble the various available SAs in one database, considering only the SAs with a well-established bodies of evidences, while rejecting the ambiguous or not fully verified ones. In fact a common risk is to overestimate the danger linked with SAs, creating a set of filters that really narrows the chemical space for drug discovery without any concrete reason. Our final database contains 142 entries, divided in three categories representing three different degrees of alert for chemicals: SAs classified as Rank3 are those meant to match intrinsically reactive portion of molecules, Rank2 collects SAs related to metabolic instability and Rank1, SAs related with interferences in chemical assays and known SAs with a not-yet fully clarified mechanism of toxicity. The final database will be available on line and integrated in ICM software suite. Number of SAs within the top 200 marketed drugs in line with numbers previously reported by authors who did a similar work, in comparison with the really high and exaggerated number of SAs found using redundant and not completely verified collection of SAs (Table 5). We find that the application of our rules can be used to identify drug candidate presenting structural liabilities, without exaggerating the risk connected to reactivity and biotransformation. The so-flagged compound should then be analyzed whether to introduce structural modification to reduce the reactivity or to mask the metabolically-unstable site, to discard the molecules or to ignore the possible liability.

.

## 3.6 Bibliography

1. Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V., ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310-2316.
2. Guengerich, F. P.; MacDonald, J. S., Applying mechanisms of chemical toxicity to predict drug safety. *Chem. Res. Toxicol.* **2007**, *20*, 344-369.
3. Kalgutkar, A. S.; Soglia, J. R., Minimising the potential for metabolic activation in drug discovery. *Expert Opinion on Drug Metabolism and Toxicology* **2005**, *1*, 91-142.
4. Uetrecht, J., Idiosyncratic drug reactions: Current understanding. In *Annu. Rev. Pharmacol. Toxicol.*, 2007; Vol. 47, pp 513-539.
5. Rawson, N. S. B., New drug approval times and safety warnings in the United States and Canada, 1992-2011. *Journal of population therapeutics and clinical pharmacology = Journal de la therapeutique des populations et de la pharmacologie clinique* **2013**, *20*, e67-81.
6. Kalgutkar, A. S.; Didiuk, M. T., Structural alerts, reactive metabolites, and protein covalent binding: How reliable are these attributes as predictors of drug toxicity? *Chemistry and Biodiversity* **2009**, *6*, 2115-2137.
7. Park, B. K.; Boobis, A.; Clarke, S.; Goldring, C. E. P.; Jones, D.; Kenna, J. G.; Lambert, C.; Lavery, H. G.; Naisbitt, D. J.; Nelson, S.; Nicoll-Griffith, D. A.; Obach, R. S.; Routledge, P.; Smith, D. A.; Tweedie, D. J.; Vermeulen, N.; Williams, D. P.; Wilson, I. D.; Baillie, T. A., Managing the challenge of chemically reactive metabolites in drug development. *Nat. Rev. Drug Discovery* **2011**, *10*, 292-306.
8. Kalgutkar, A. S., Handling reactive metabolite positives in drug discovery: What has retrospective structure-toxicity analyses taught us? *Chem. Biol. Interact.* **2011**, *192*, 46-55.
9. Casini, A.; Giorli, M.; Hyland, R. J.; Serroni, A.; Gilfor, D.; Farber, J. L., MECHANISMS OF CELL INJURY IN THE KILLING OF CULTURED-HEPATOCYTES BY BROMOBENZENE. *J. Biol. Chem.* **1982**, *257*, 6721-6728.
10. Smith, M. T.; Thor, H.; Orrenius, S., THE ROLE OF LIPID-PEROXIDATION IN THE TOXICITY OF FOREIGN COMPOUNDS TO LIVER-CELLS. *Biochem. Pharmacol.* **1983**, *32*, 763-764.
11. Gibson, J. D.; Pumford, N. R.; Samokyszyn, V. M.; Hinson, J. A., Mechanism of acetaminophen-induced hepatotoxicity: Covalent binding versus oxidative stress. *Chem. Res. Toxicol.* **1996**, *9*, 580-585.
12. Enoch, S. J.; Cronin, M. T. D., A review of the electrophilic reaction chemistry involved in covalent DNA binding. *Crit. Rev. Toxicol.* **2010**, *40*, 728-748.
13. Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D., Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A

Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24*, 1345-1410.

14. Enoch, S. J.; Ellison, C. M.; Schultz, T. W.; Cronin, M. T. D., A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol.* **2011**, *41*, 783-802.

15. Orrenius, S.; Zhivotovsky, B., The future of toxicology - Does it matter how cells die? *Chem. Res. Toxicol.* **2006**, *19*, 729-733.

16. Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C., Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617-648.

17. Sushko, I.; Novotarskyi, S.; Koerner, R.; Pandey, A. K.; Cherkasov, A.; Lo, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Mueller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Oberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V., Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094-2111.

18. Modi, S.; Hughes, M.; Garrow, A.; White, A., The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug Discovery Today* **2012**, *17*, 135-142.

19. Moroy, G.; Martiny, V. Y.; Vayer, P.; Villoutreix, B. O.; Miteva, M. A., Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discovery Today* **2012**, *17*, 44-55.

20. Sun, H.; Scott, D. O., Structure-based drug metabolism predictions for drug design. *Chemical Biology and Drug Design* **2010**, *75*, 3-17.

21. Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R., MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970-6979.

22. Hinson, J. A.; Pumford, N. R.; Nelson, S. D., THE ROLE OF METABOLIC-ACTIVATION IN DRUG TOXICITY. *Drug Metabolism Reviews* **1994**, *26*, 395-412.

23. Šink, R.; Gobec, S.; Pečar, S.; Zega, A., False positives in the early stages of drug discovery. *Current Medicinal Chemistry* **2010**, *17*, 4231-4255.

24. Shoichet, B. K., Screening in a spirit haunted world. *Drug Discovery Today* **2006**, *11*, 607-615.

25. Thorne, N.; Auld, D. S.; Inglese, J., Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **2010**, *14*, 315-324.

26. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K., A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712-1722.

27. Baell, J. B.; Holloway, G. A., New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719-2740.

28. Rishton, G. M., Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86-96.

29. Kazius, J.; McGuire, R.; Bursi, R., Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312-320.



30. Hakimelahi, G. H.; Khodarahmi, G. A., The identification of toxicophores for the prediction of mutagenicity, hepatotoxicity and cardiotoxicity. *Journal of the Iranian Chemical Society* **2005**, *2*, 244-267.
31. Park, B. K.; Kitteringham, N. R.; Maggs, J. L.; Pirmohamed, M.; Williams, D. P., The role of metabolic activation in drug-induced hepatotoxicity. In *Annu. Rev. Pharmacool. Toxicol.*, 2005; Vol. 45, pp 177-202.
32. Liebler, D. C., Protein damage by reactive electrophiles: Targets and consequences. *Chem. Res. Toxicol.* **2008**, *21*, 117-128.
33. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM - a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*.
34. Bruns, R. F.; Watson, I. A., Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55*, 9763-9772.
35. Hann, M. M.; Oprea, T. I., Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255-263.
36. Weininger, D., SMILES, A CHEMICAL LANGUAGE AND INFORMATION-SYSTEM .1. INTRODUCTION TO METHODOLOGY AND ENCODING RULES. *J. Chem. Inf. Comput. Sci* **1988**, *28*, 31-36.
37. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S., DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035-D1041.
38. Schwoebel, J. A. H.; Koleva, Y. K.; Enoch, S. J.; Bajot, F.; Hewitt, M.; Madden, J. C.; Roberts, D. W.; Schutz, T. W.; Cronin, M. T. D., Measurement and Estimation of Electrophilic Reactivity for Predictive Toxicology. *Chem. Rev.* **2011**, *111*, 2562-2596.
39. Enoch, S. J.; Madden, J. C.; Cronin, M. T. D., Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *Sar and Qsar in Environmental Research* **2008**, *19*, 555-578.
40. Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A., An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060-1068.
41. Benigni, R.; Bossa, C., Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutation Research-Reviews in Mutation Research* **2008**, *659*, 248-261.
42. Evans, D. C.; Baillie, T. A., Minimizing the potential for metabolic activation as an integral part of drug design. *Current Opinion in Drug Discovery & Development* **2005**, *8*, 44-50.
43. Rishton, G. M., Molecular diversity in the context of leadlikeness: compound properties that enable effective biochemical screening. *Curr. Opin. Chem. Biol.* **2008**, *12*, 340-351.
44. Dansette, P. M.; Amar, C.; Smith, C.; Pons, C.; Mansuy, D., OXIDATIVE ACTIVATION OF THE THIOPHENE RING BY HEPATIC-ENZYMES - HYDROXYLATION AND FORMATION OF ELECTROPHILIC METABOLITES DURING METABOLISM OF TIENILIC ACID AND ITS ISOMER BY RAT-LIVER MICROSOMES. *Biochem. Pharmacol.* **1990**, *39*, 911-918.
45. Dansette, P. M.; Amar, C.; Valadon, P.; Pons, C.; Beaune, P. H.; Mansuy, D., HYDROXYLATION AND FORMATION OF ELECTROPHILIC METABOLITES OF TIENILIC ACID AND ITS ISOMER BY HUMAN LIVER-MICROSOMES - CATALYSIS BY

A CYTOCHROME-P450 IIC DIFFERENT FROM THAT RESPONSIBLE FOR MEPHENYTOIN HYDROXYLATION. *Biochem. Pharmacol.* **1991**, *41*, 553-560.

46. Yu, J.; Mathisen, D. E.; Burdette, D.; Brown, D. G.; Becker, C.; Aharony, D., Identification of Multiple Glutathione Conjugates of 8-Amino-2-methyl-4-phenyl-1,2,3,4-tetrahydroisoquinoline maleate (Nomifensine) in Liver Microsomes and Hepatocyte Preparations: Evidence of the Bioactivation of Nomifensine. *Drug Metabolism and Disposition* **2010**, *38*, 46-60.

47. Smith, K. S.; Smith, P. L.; Heady, T. N.; Trugman, J. M.; Harman, W. D.; Macdonald, T. L., In vitro metabolism of tolcapone to reactive intermediates: Relevance to tolcapone liver toxicity. *Chem. Res. Toxicol.* **2003**, *16*, 123-128.

48. Dong, J. Q.; Liu, J. H.; Smith, P. C., Role of benoxaprofen and flunoxaprofen acyl glucuronides in covalent binding to rat plasma and liver proteins in vivo. *Biochem. Pharmacol.* **2005**, *70*, 937-948.

49. Smith, P. C.; Benet, L. Z.; McDonagh, A. F., COVALENT BINDING OF ZOMEPIRAC GLUCURONIDE TO PROTEINS - EVIDENCE FOR A SCHIFF-BASE MECHANISM. *Drug Metabolism and Disposition* **1990**, *18*, 639-644.

50. Baba, A.; Yoshioka, T., Structure-Activity Relationships for the Degradation Reaction of 1-beta-O-Acyl Glucuronides. Part 3: Electronic and Steric Descriptors Predicting the Reactivity of Aralkyl Carboxylic Acid 1-beta-O-Acyl Glucuronides. *Chem. Res. Toxicol.* **2009**, *22*, 1998-2008.

51. Kassahun, K.; Pearson, P. G.; Tang, W.; McIntosh, L.; Leung, K.; Elmore, C.; Dean, D.; Wang, R.; Doss, G.; Baillie, T. A., Studies on the metabolism of troglitazone to reactive intermediates in vitro and in vivo. Evidence for novel biotransformation pathways involving quinone methide formation and thiazolidinedione ring scission. *Chem. Res. Toxicol.* **2001**, *14*, 62-70.

52. Alvarez-Sanchez, R.; Montavon, F.; Hartung, T.; Paehler, A., Thiazolidinedione bioactivation: A comparison of the bioactivation potentials of troglitazone, rosiglitazone, and pioglitazone using stable isotope-labeled analogues and liquid chromatography tandem mass spectrometry. *Chem. Res. Toxicol.* **2006**, *19*, 1106-1116.

53. Masubuchi, Y.; Kano, S.; Horie, T., Mitochondrial permeability transition as a potential determinant of hepatotoxicity of antidiabetic thiazolidinediones. *Toxicology* **2006**, *222*, 233-239.

54. Lauterburg, B. H.; Smith, C. V.; Todd, E. L.; Mitchell, J. R., OXIDATION OF HYDRAZINE METABOLITES FORMED FROM ISONIAZID. *Clinical Pharmacology & Therapeutics* **1985**, *38*, 566-571.

55. Nelson, S. D.; Mitchell, J. R.; Timbrell, J. A.; Snodgrass, W. R.; Corcoran, G. B., ISONIAZID AND IPRONIAZID - ACTIVATION OF METABOLITES TO TOXIC INTERMEDIATES IN MAN AND RAT. *Science* **1976**, *193*, 901-903.

56. Preziosi, P., Isoniazid: Metabolic aspects and toxicological correlates. *Current Drug Metabolism* **2007**, *8*, 839-851.

57. Fontana, R. J.; Shakil, A. O.; Greenson, J. K.; Boyd, I.; Lee, W. M., Acute liver failure due to amoxicillin and amoxicillin/clavulanate. *Digestive Diseases and Sciences* **2005**, *50*, 1785-1790.

58. Connor, S. C.; Everett, J. R.; Jennings, K. R.; Nicholson, J. K.; Woodnutt, G., HIGH-RESOLUTION H-1-NMR SPECTROSCOPIC STUDIES OF THE METABOLISM AND EXCRETION OF AMPICILLIN IN RATS AND AMOXICILLIN IN RATS AND MAN. *J. Pharm. Pharmacol.* **1994**, *46*, 128-134.

59. Ballantyne, B.; Slesinski, R. S.; Myers, R. C., THE ACUTE TOXICITY AND MUTAGENIC POTENTIAL OF 3-METHYL-2-BENZOTHIAZOLINONE HYDRAZONE. *Toxicology and Industrial Health* **1988**, *4*, 23-37.
60. Sterba, M.; Simunek, T.; Mazurova, Y.; Adamcova, M.; Popelova, O.; Kaplanova, J.; Ponka, P.; Gersl, V., Safety and tolerability of repeated administration of pyridoxal 2-chlorobenzoyl hydrazone in rabbits. *Human & Experimental Toxicology* **2005**, *24*, 581-589.
61. Nelson, S. D., Structure toxicity relationships - How useful are they in predicting toxicities of new drugs? In *Biological Reactive Intermediates Vi: Chemical and Biological Mechanisms in Susceptibility to and Prevention of Environmental Diseases*, Dansette, P. M.; Snyder, R.; Delaforge, M.; Gibson, G. G.; Greim, H.; Jollow, D. J.; Monks, T. J.; Sipes, I. G., Eds. 2001; Vol. 500, pp 33-43.
62. Waldhauser, L.; Uetrecht, J., OXIDATION OF PROPYLTHIOURACIL TO REACTIVE METABOLITES BY ACTIVATED NEUTROPHILS - IMPLICATIONS FOR AGRANULOCYTOSIS. *Drug Metabolism and Disposition* **1991**, *19*, 354-359.
63. Taurog, A.; Dorris, M. L., PROPYLTHIOURACIL AND METHIMAZOLE DISPLAY CONTRASTING PATHWAYS OF PERIPHERAL METABOLISM IN BOTH RAT AND HUMAN. *Endocrinology* **1988**, *122*, 592-601.
64. Edon Vitaku, E. A. I., Jón T. Njarðarson Top 200 Pharmaceutical Products by US Retail Sales in 2012. <http://cbc.arizona.edu/njardarson/group/homepage> (accessed in October 2013),

# Appendix A

**Table I:** Molecular weight and number of rotatable bonds for complexes used for the optimization process.

**Table II.** Experimental design 1, fractional factorial design.

**Table III.** Response Surface Model

**Table IV.** LiGen, AutoDock and Glide docking results with the PDBbind complexes .

**Table V:** Parameters and results of the full factorial design performed for the optimization of the VS protocol.

**Table VI.** Comparison of VS results considering only the “own decoys” subset of DUD, with the original set of parameters (on the left side of the table) and the optimized ones (on the right).

**Table VII.** Parameters of the experiment number 28; this set of parameters has been chosen as default set for LiGenDock.



**Table I.** Molecular weight and number of rotatable bonds for complexes used for the optimization process.

<i>PDB code</i>	<i>Rotatable bonds</i>	<i>Molecular weight</i>	<i>PDB code</i>	<i>Rotatable bonds</i>	<i>Molecular weight</i>	<i>PDB code</i>	<i>Rotatable bonds</i>	<i>Molecular weight</i>
1a4g	7	332.31	1fkg	9	449.59	1tph	6	171.05
1a9u	1	377.44	1frp	13	340.12	1tpp	4	212.25
1acj	1	200.29	1ghb	4	246.27	1trk	12	428.34
1acm	7	253.11	1glp	14	357.34	1tyl	2	157.21
1apu	16	485.66	1gpy	9	260.14	1ukz	8	347.22
1aqw	11	306.32	1hdc	6	568.76	1ulb	1	151.13
1ase	8	265.16	1hfc	10	349.43	1ydr	2	291.37
1b59	7	313.39	1imb	9	260.14	1yee	11	360.26
1bgo	12	481.61	1ivb	4	240.17	2ak3	8	347.22
1bl7	1	338.39	1ivq	16	613.84	2cht	3	228.2
1blh	6	245.17	1ldm	1	89.05	2cmd	6	192.13
1bmq	11	546.64	1mld	6	189.1	2cpp	0	152.24
1byb	24	666.58	1mmq	7	456.58	2dbl	5	417.57
1byg	2	485.7	1okl	3	250.32	2fox	12	456.35
1cbs	9	300.44	1pbd	2	137.14	2h4n	3	222.24

---

1cdg	12	342.3	1pdz	5	156.03	2phh	2	138.12
1cil	4	324.43	1pgp	13	276.14	2qwk	7	284.36
1cle	24	649.11	1phd	0	144.18	2r07	7	326.39
1coy	1	288.43	1phg	3	226.28	2tsc	12	477.48
1cqp	7	404.55	1ppi	28	808.77	2yhx	7	297.31
1cvu	14	304.47	1pso	24	684.9	3cla	7	323.13
1d4p	5	375.58	1qbr	14	758.91	3cpa	7	238.24
1dd7	4	479.49	1rbp	9	286.46	3ert	10	387.52
1dhf	10	441.4	1rds	12	590.42	3hvt	0	266.3
1die	5	163.17	1rob	8	323.2	4aah	3	330.21
1dy9	16	525.51	1rt2	7	364.44	4cox	4	357.79
1ejn	6	348.52	1slt	6	383.36	4cts	3	132.07
1elc	13	506.57	1snc	9	402.19	4er2	24	685.9
1eta	7	776.87	1tdb	6	326.18	4fab	3	334.33
1ets	11	538.78	1tka	8	425.34	4phv	15	618.77
1ett	8	441.66	1tmn	13	479.58	4tpi	6	216.28
1f0s	5	427.5	1tng	2	113.2	5abp	6	180.16
1fen	7	270.46	1tni	5	149.24	7tim	6	171.05
1fgi	5	296.33						

---

**Table II.** Experimental design 1, fractional factorial design.

Exp num.	Min F Dist	Max F Num	Dist Co	Prot Vdw B	Grid Acc	Lig Neig Thr	Score Dist Thr	Grid Dist Thr	Hyd Thr	Dist Thr	Pose Over	Ag Delta	Conf Vdw B	Neig Thr	Conf Ag Delta	% poses RMSD < 2Å
1	1	5	4	0.6	1	0.5	1	1	0.1	1	0.5	10	0.5	50	3	0
2	1	5	4	0.6	1	0.5	3	3	0.3	3	1	50	1	150	10	0
3	1	5	4	0.6	1	3	1	3	0.3	3	1	50	1	150	10	0
4	1	5	4	0.6	1	3	3	1	0.1	1	0.5	10	0.5	50	3	0
5	1	5	4	0.6	4	0.5	1	3	0.3	3	1	50	1	150	10	14
6	1	5	4	0.6	4	0.5	3	1	0.1	1	0.5	10	0.5	50	3	9
7	1	5	4	0.6	4	3	1	1	0.1	1	0.5	10	0.5	50	3	9
8	1	5	4	0.6	4	3	3	3	0.3	3	1	50	1	150	10	14
9	1	5	4	1	1	0.5	1	3	0.3	3	1	10	0.5	50	3	0
10	1	5	4	1	1	0.5	3	1	0.1	1	0.5	50	1	150	10	0
11	1	5	4	1	1	3	1	1	0.1	1	0.5	50	1	150	10	0
12	1	5	4	1	1	3	3	3	0.3	3	1	10	0.5	50	3	0
13	1	5	4	1	4	0.5	1	1	0.1	1	0.5	50	1	150	10	1
14	1	5	4	1	4	0.5	3	3	0.3	3	1	10	0.5	50	3	9
15	1	5	4	1	4	3	1	3	0.3	3	1	10	0.5	50	3	9
16	1	5	4	1	4	3	3	1	0.1	1	0.5	50	1	150	10	1
17	1	5	6	0.6	1	0.5	1	3	0.3	1	0.5	50	1	50	3	7
18	1	5	6	0.6	1	0.5	3	1	0.1	3	1	10	0.5	150	10	0



19	1	5	6	0.6	1	3	1	1	0.1	3	1	10	0.5	150	10	0
20	1	5	6	0.6	1	3	3	3	0.3	1	0.5	50	1	50	3	7
21	1	5	6	0.6	4	0.5	1	1	0.1	3	1	10	0.5	150	10	11
22	1	5	6	0.6	4	0.5	3	3	0.3	1	0.5	50	1	50	3	4
23	1	5	6	0.6	4	3	1	3	0.3	1	0.5	50	1	50	3	4
24	1	5	6	0.6	4	3	3	1	0.1	3	1	10	0.5	150	10	11
25	1	5	6	1	1	0.5	1	1	0.1	3	1	50	1	50	3	0
26	1	5	6	1	1	0.5	3	3	0.3	1	0.5	10	0.5	150	10	0
27	1	5	6	1	1	3	1	3	0.3	1	0.5	10	0.5	150	10	0
28	1	5	6	1	1	3	3	1	0.1	3	1	50	1	50	3	0
29	1	5	6	1	4	0.5	1	3	0.3	1	0.5	10	0.5	150	10	5
30	1	5	6	1	4	0.5	3	1	0.1	3	1	50	1	50	3	4
31	1	5	6	1	4	3	1	1	0.1	3	1	50	1	50	3	4
32	1	5	6	1	4	3	3	3	0.3	1	0.5	10	0.5	150	10	5
33	1	15	4	0.6	1	0.5	1	3	0.1	3	0.5	50	0.5	150	3	0
34	1	15	4	0.6	1	0.5	3	1	0.3	1	1	10	1	50	10	0
35	1	15	4	0.6	1	3	1	1	0.3	1	1	10	1	50	10	0
36	1	15	4	0.6	1	3	3	3	0.1	3	0.5	50	0.5	150	3	0
37	1	15	4	0.6	4	0.5	1	1	0.3	1	1	10	1	50	10	23
38	1	15	4	0.6	4	0.5	3	3	0.1	3	0.5	50	0.5	150	3	25
39	1	15	4	0.6	4	3	1	3	0.1	3	0.5	50	0.5	150	3	25
40	1	15	4	0.6	4	3	3	1	0.3	1	1	10	1	50	10	23
41	1	15	4	1	1	0.5	1	1	0.3	1	1	50	0.5	150	3	0
42	1	15	4	1	1	0.5	3	3	0.1	3	0.5	10	1	50	10	0

43	1	15	4	1	1	3	1	3	0.1	3	0.5	10	1	50	10	0
44	1	15	4	1	1	3	3	1	0.3	1	1	50	0.5	150	3	0
45	1	15	4	1	4	0.5	1	3	0.1	3	0.5	10	1	50	10	25
46	1	15	4	1	4	0.5	3	1	0.3	1	1	50	0.5	150	3	15
47	1	15	4	1	4	3	1	1	0.3	1	1	50	0.5	150	3	15
48	1	15	4	1	4	3	3	3	0.1	3	0.5	10	1	50	10	25
49	1	15	6	0.6	1	0.5	1	1	0.3	3	0.5	10	1	150	3	0
50	1	15	6	0.6	1	0.5	3	3	0.1	1	1	50	0.5	50	10	8
51	1	15	6	0.6	1	3	1	3	0.1	1	1	50	0.5	50	10	8
52	1	15	6	0.6	1	3	3	1	0.3	3	0.5	10	1	150	3	0
53	1	15	6	0.6	4	0.5	1	3	0.1	1	1	50	0.5	50	10	16
54	1	15	6	0.6	4	0.5	3	1	0.3	3	0.5	10	1	150	3	28
55	1	15	6	0.6	4	3	1	1	0.3	3	0.5	10	1	150	3	28
56	1	15	6	0.6	4	3	3	3	0.1	1	1	50	0.5	50	10	16
57	1	15	6	1	1	0.5	1	3	0.1	1	1	10	1	150	3	0
58	1	15	6	1	1	0.5	3	1	0.3	3	0.5	50	0.5	50	10	0
59	1	15	6	1	1	3	1	1	0.3	3	0.5	50	0.5	50	10	0
60	1	15	6	1	1	3	3	3	0.1	1	1	10	1	150	3	0
61	1	15	6	1	4	0.5	1	1	0.3	3	0.5	50	0.5	50	10	15
62	1	15	6	1	4	0.5	3	3	0.1	1	1	10	1	150	3	10
63	1	15	6	1	4	3	1	3	0.1	1	1	10	1	150	3	10
64	1	15	6	1	4	3	3	1	0.3	3	0.5	50	0.5	50	10	15
65	3	5	4	0.6	1	0.5	1	3	0.1	1	1	10	1	150	3	0
66	3	5	4	0.6	1	0.5	3	1	0.3	3	0.5	50	0.5	50	10	0

67	3	5	4	0.6	1	3	1	1	0.3	3	0.5	50	0.5	50	10	0
68	3	5	4	0.6	1	3	3	3	0.1	1	1	10	1	150	3	0
69	3	5	4	0.6	4	0.5	1	1	0.3	3	0.5	50	0.5	50	10	8
70	3	5	4	0.6	4	0.5	3	3	0.1	1	1	10	1	150	3	11
71	3	5	4	0.6	4	3	1	3	0.1	1	1	10	1	150	3	10
72	3	5	4	0.6	4	3	3	1	0.3	3	0.5	50	0.5	50	10	8
73	3	5	4	1	1	0.5	1	1	0.3	3	0.5	10	1	150	3	0
74	3	5	4	1	1	0.5	3	3	0.1	1	1	50	0.5	50	10	0
75	3	5	4	1	1	3	1	3	0.1	1	1	50	0.5	50	10	0
76	3	5	4	1	1	3	3	1	0.3	3	0.5	10	1	150	3	0
77	3	5	4	1	4	0.5	1	3	0.1	1	1	50	0.5	50	10	8
78	3	5	4	1	4	0.5	3	1	0.3	3	0.5	10	1	150	3	3
79	3	5	4	1	4	3	1	1	0.3	3	0.5	10	1	150	3	3
80	3	5	4	1	4	3	3	3	0.1	1	1	50	0.5	50	10	8
81	3	5	6	0.6	1	0.5	1	1	0.3	1	1	50	0.5	150	3	0
82	3	5	6	0.6	1	0.5	3	3	0.1	3	0.5	10	1	50	10	1
83	3	5	6	0.6	1	3	1	3	0.1	3	0.5	10	1	50	10	1
84	3	5	6	0.6	1	3	3	1	0.3	1	1	50	0.5	150	3	0
85	3	5	6	0.6	4	0.5	1	3	0.1	3	0.5	10	1	50	10	10
86	3	5	6	0.6	4	0.5	3	1	0.3	1	1	50	0.5	150	3	4
87	3	5	6	0.6	4	3	1	1	0.3	1	1	50	0.5	150	3	4
88	3	5	6	0.6	4	3	3	3	0.1	3	0.5	10	1	50	10	10
89	3	5	6	1	1	0.5	1	3	0.1	3	0.5	50	0.5	150	3	0
90	3	5	6	1	1	0.5	3	1	0.3	1	1	10	1	50	10	0

91	3	5	6	1	1	3	1	1	0.3	1	1	10	1	50	10	0
92	3	5	6	1	1	3	3	3	0.1	3	0.5	50	0.5	150	3	0
93	3	5	6	1	4	0.5	1	1	0.3	1	1	10	1	50	10	4
94	3	5	6	1	4	0.5	3	3	0.1	3	0.5	50	0.5	150	3	7
95	3	5	6	1	4	3	1	3	0.1	3	0.5	50	0.5	150	3	7
96	3	5	6	1	4	3	3	1	0.3	1	1	10	1	50	10	4
97	3	15	4	0.6	1	0.5	1	1	0.1	3	1	50	1	50	3	0
98	3	15	4	0.6	1	0.5	3	3	0.3	1	0.5	10	0.5	150	10	0
99	3	15	4	0.6	1	3	1	3	0.3	1	0.5	10	0.5	150	10	0
100	3	15	4	0.6	1	3	3	1	0.1	3	1	50	1	50	3	0
101	3	15	4	0.6	4	0.5	1	3	0.3	1	0.5	10	0.5	150	10	10
102	3	15	4	0.6	4	0.5	3	1	0.1	3	1	50	1	50	3	21
103	3	15	4	0.6	4	3	1	1	0.1	3	1	50	1	50	3	20
104	3	15	4	0.6	4	3	3	3	0.3	1	0.5	10	0.5	150	10	11
105	3	15	4	1	1	0.5	1	3	0.3	1	0.5	50	1	50	3	0
106	3	15	4	1	1	0.5	3	1	0.1	3	1	10	0.5	150	10	0
107	3	15	4	1	1	3	1	1	0.1	3	1	10	0.5	150	10	0
108	3	15	4	1	1	3	3	3	0.3	1	0.5	50	1	50	3	0
109	3	15	4	1	4	0.5	1	1	0.1	3	1	10	0.5	150	10	6
110	3	15	4	1	4	0.5	3	3	0.3	1	0.5	50	1	50	3	6
111	3	15	4	1	4	3	1	3	0.3	1	0.5	50	1	50	3	6
112	3	15	4	1	4	3	3	1	0.1	3	1	10	0.5	150	10	6
113	3	15	6	0.6	1	0.5	1	3	0.3	3	1	10	0.5	50	3	3
114	3	15	6	0.6	1	0.5	3	1	0.1	1	0.5	50	1	150	10	0

115	3	15	6	0.6	1	3	1	1	0.1	1	0.5	50	1	150	10	0
116	3	15	6	0.6	1	3	3	3	0.3	3	1	10	0.5	50	3	3
117	3	15	6	0.6	4	0.5	1	1	0.1	1	0.5	50	1	150	10	9
118	3	15	6	0.6	4	0.5	3	3	0.3	3	1	10	0.5	50	3	19
119	3	15	6	0.6	4	3	1	3	0.3	3	1	10	0.5	50	3	19
120	3	15	6	0.6	4	3	3	1	0.1	1	0.5	50	1	150	10	9
121	3	15	6	1	1	0.5	1	1	0.1	1	0.5	10	0.5	50	3	0
122	3	15	6	1	1	0.5	3	3	0.3	3	1	50	1	150	10	0
123	3	15	6	1	1	3	1	3	0.3	3	1	50	1	150	10	0
124	3	15	6	1	1	3	3	1	0.1	1	0.5	10	0.5	50	3	0
125	3	15	6	1	4	0.5	1	3	0.3	3	1	50	1	150	10	6
126	3	15	6	1	4	0.5	3	1	0.1	1	0.5	10	0.5	50	3	11
127	3	15	6	1	4	3	1	1	0.1	1	0.5	10	0.5	50	3	11
128	3	15	6	1	4	3	3	3	0.3	3	1	50	1	150	10	6
129	2	10	5	0.8	2.5	1.75	2	2	0.2	2	0.75	30	0.75	100	6.5	20

Parameters set of all the tests of the screening design. In the last column is reported the percentage of predicted poses with an RMSD than 2Å respect to the co-crystallized ligands.

**Table III.** Response Surface Model.

	<b>Min F Dist</b>	<b>Max Num</b>	<b>F</b>	<b>Prot B</b>	<b>Vdw</b>	<b>Grid Acc</b>	<b>Dist Thr</b>	<b>Neig Thr</b>	<b>Number of results</b>	<b>% &lt;2Å</b>	<b>RMSD Å</b>	<b>% RMSD &lt;3 Å</b>	<b>best scoring poses: %</b>
exp01	1.5	12	0.6	2	2	25	95	32	57	17			
exp02	1.5	12	0.6	2	3	75	85	22	45	15			
exp03	1.5	12	0.6	4	2	75	95	34	58	18			
exp04	1.5	12	0.6	4	3	25	96	38	60	20			
exp05	1.5	12	0.8	2	2	75	65	17	35	11			
exp06	1.5	12	0.8	2	3	25	92	33	51	19			
exp07	1.5	12	0.8	4	2	25	91	25	46	13			
exp08	1.5	12	0.8	4	3	75	93	33	57	21			
exp09	1.5	20	0.6	2	2	75	81	25	48	12			
exp10	1.5	20	0.6	2	3	25	99	41	70	21			
exp11	1.5	20	0.6	4	2	25	99	39	66	23			
exp12	1.5	20	0.6	4	3	75	99	43	68	26			
exp13	1.5	20	0.8	2	2	25	96	42	67	22			
exp14	1.5	20	0.8	2	3	75	70	20	36	11			
exp15	1.5	20	0.8	4	2	75	95	39	68	19			
exp16	1.5	20	0.8	4	3	25	95	38	66	24			
exp17	2	12	0.6	2	2	75	81	19	38	12			

exp18	2	12	0.6	2	3	25	98	41	60	25
exp19	2	12	0.6	4	2	25	99	36	58	18
exp20	2	12	0.6	4	3	75	99	42	66	23
exp21	2	12	0.8	2	2	25	93	32	56	23
exp22	2	12	0.8	2	3	75	68	19	36	10
exp23	2	12	0.8	4	2	75	95	36	59	22
exp24	2	12	0.8	4	3	25	99	31	58	18
exp25	2	20	0.6	2	2	25	98	41	69	18
exp26	2	20	0.6	2	3	75	82	22	40	11
exp27	2	20	0.6	4	2	75	100	38	67	22
exp28	2	20	0.6	4	3	25	100	40	70	21
exp29	2	20	0.8	2	2	75	64	12	26	8
exp30	2	20	0.8	2	3	25	96	38	62	21
exp31	2	20	0.8	4	2	25	99	38	61	19
exp32	2	20	0.8	4	3	75	99	34	63	17
exp33	1.5	16	0.7	3	2.5	50	96	37	64	18
exp34	2	16	0.7	3	2.5	50	99	43	69	18
exp35	1.75	12	0.7	3	2.5	50	95	29	62	16
exp36	1.75	20	0.7	3	2.5	50	99	42	69	21
exp37	1.75	16	0.6	3	2.5	50	98	43	66	22
exp38	1.75	16	0.8	3	2.5	50	95	41	62	27
exp39	1.75	16	0.7	2	2.5	50	93	34	57	20

exp40	1.75	16	0.7	4	2.5	50	98	36	67	20
exp41	1.75	16	0.7	3	2	50	97	38	63	19
exp42	1.75	16	0.7	3	3	50	99	41	63	25
exp43	1.75	16	0.7	3	2.5	25	98	41	67	21
exp44	1.75	16	0.7	3	2.5	75	97	41	67	21
exp45	1.75	16	0.7	3	2.5	50	98	41	67	21
exp46	1.75	16	0.7	3	2.5	50	98	41	67	21
exp47	1.75	16	0.7	3	2.5	50	98	41	67	21
exp48	1.75	16	0.7	3	2.5	50	98	41	67	21
exp49	1.75	16	0.7	3	2.5	50	98	41	67	21
exp50	1.75	16	0.7	3	2.5	50	98	41	67	21
exp51	1.75	16	0.7	3	2.5	50	98	41	67	21
exp52	1.75	16	0.7	3	2.5	50	98	41	67	21
exp53	1.75	16	0.7	3	2.5	50	98	41	67	21
exp54	1.75	16	0.7	3	2.5	50	98	41	67	21
exp55	1.75	16	0.7	3	2.5	50	98	41	67	21
exp56	1.75	16	0.7	3	2.5	50	98	41	67	12
exp57	1.75	16	0.7	3	2.5	50	98	41	67	21
exp58	1.75	16	0.7	3	2.5	50	98	41	67	21
exp59	1.75	16	0.7	3	2.5	50	96	41	67	26



Parameters set used in the RSM design together with the number of results obtained for each parameter set out of the 100 protein tested, the percentage of predicted poses with RMSD less than 2 and 3 Å, the percentage of the best scoring poses with RMSD less than 2Å.

**Table IV.** LiGen, AutoDock and Glide docking results with the PDBbind complexes . In table A) are reported the RMSD of the best predicted pose respect to the co-crystallized ligand for LiGen, AutoDock and Glide respectively. In table B) are reported the RMSD of the best scoring pose respect to the co-crystallized ligand for LiGen, AutoDock and Glide respectively.

A)

<i>PDB code</i>	<i>LiGen</i>	<i>AutoDock</i>	<i>Glide</i>	<i>PDB code</i>	<i>LiGen</i>	<i>AutoDock</i>	<i>Glide</i>	<i>PDB code</i>	<i>LiGen</i>	<i>AutoDock</i>	<i>Glide</i>
	<i>n</i>	<i>k</i>	<i>e</i>		<i>n</i>	<i>k</i>	<i>e</i>		<i>n</i>	<i>k</i>	<i>e</i>
<b>1e66</b>	0.99	0.51	0.46	<b>1l2s</b>	1.76	1.02	0.45	<b>1vzq</b>	2.86	0.7	0.29
<b>10gs</b>	2.82	3.8	1.35	<b>1l83</b>	0.36	1.15	0.35	<b>1x1z</b>	1.39	0.45	0.3
<b>1a69</b>	1.11	2.19	0.97	<b>1li3</b>	2.22	0.99	0.43	<b>1xgj</b>	2.43	2.22	5.89
<b>1abf</b>	0.82	0.29	0.32	<b>1li6</b>	2.53	3.35	1.69	<b>1y1m</b>	0.82	0.51	0.33
<b>1ai5</b>	1.1	1.35	0.92	<b>1lol</b>	2.06	0.53	2.47	<b>1y6q</b>	6.16	0.69	0.64
<b>1ajp</b>	2.83	4.47	0.63	<b>1loq</b>	1.65	0.67	0.45	<b>1ydt</b>	2.85	1.04	0.61
<b>1ajq</b>	1.18	1.02	0.34	<b>1m0n</b>	1.95	0.81	0.72	<b>1zc9</b>	2.36	2.87	3.04
<b>1avn</b>	2.64	5.01	2.43	<b>1m0q</b>	2.1	0.68	0.81	<b>1zoe</b>	1.28	5.55	2.73
<b>1ax0</b>	2.49	0.59	0.47	<b>1m2q</b>	3.7	3.71	1.31	<b>1zs0</b>	3.99	1.32	0.5
<b>1axz</b>	1.1	0.61	0.57	<b>1n2v</b>	0.92	2.74	0.47	<b>1zvx</b>	5.91	0.94	0.55
<b>1b11</b>	5.31	5.13	1.14	<b>1nc1</b>	2.1	0.93	0.73	<b>2aou</b>	3.21	1.13	2.3
<b>1b7h</b>	2.89	1.86	0.51	<b>1ndw</b>	1.59	1.19	1.27	<b>2b1v</b>	0.95	0.8	0.63
<b>1b8o</b>	2.77	2.33	1.08	<b>1ndy</b>	4.16	1.55	2.42	<b>2baj</b>	3.27	0.64	0.43
<b>1b9j</b>	1.9	1.63	0.64	<b>1ndz</b>	3.31	2.29	1.8	<b>2bok</b>	1.53	0.53	0.52
<b>1bcu</b>	1.13	2.28	0.32	<b>1nfy</b>	1.75	0.69	1.42	<b>2brb</b>	1.35	1.04	0.71

<b>1bra</b>	1.53	0.47	0.44	<b>1nhu</b>	3.24	4.91	5.22	<b>2brm</b>	1.69	1.09	4.39
<b>1c84</b>	1.62	0.26	0.32	<b>1nja</b>	1.29	1.62	1.62	<b>2bz6</b>	2.07	0.62	0.81
<b>1d09</b>	1.31	0.45	0.41	<b>1nje</b>	2.07	1.66	4.32	<b>2c02</b>	3.37	2.58	2.81
<b>1d7j</b>	1.74	1.37	1.78	<b>1nvq</b>	0.75	3.31	0.56	<b>2ceq</b>	0.7	3.16	1.44
<b>1det</b>	2.73	0.74	0.75	<b>1nwl</b>	2.78	2.76	3.24	<b>2cer</b>	2.08	3.19	3.43
<b>1df8</b>	1.05	0.39	0.18	<b>1o0h</b>	1.72	1.11	2.33	<b>2cet</b>	0.81	3.31	3.66
<b>1dhi</b>	2.24	1.3	2.78	<b>1o3f</b>	1.44	0.64	0.63	<b>2cgr</b>	1.1	1.62	0.74
<b>1e1v</b>	1.46	8.66	4.99	<b>1o3p</b>	1.25	1.26	1.66	<b>2ctc</b>	0.75	0.53	0.35
<b>1e5a</b>	0.65	0.84	0.53	<b>1ols</b>	2.93	1.53	0.39	<b>2d0k</b>	2.33	1.01	2.39
<b>1ela</b>	1.94	2.04	2.36	<b>1olu</b>	2.81	0.88	2.43	<b>2d1o</b>	2.42	1.8	0.75
<b>1elb</b>	2.94	2.62	4	<b>1om1</b>	0.95	1.08	0.35	<b>2d3u</b>	1.09	0.78	0.56
<b>1f4e</b>	0.78	2.31	1.66	<b>1p1q</b>	2.59	1.46	0.68	<b>2d3z</b>	1.26	0.72	0.38
<b>1f4f</b>	1.78	2.02	2.91	<b>1pb9</b>	0.37	0.22	0.27	<b>2drc</b>	2.46	1.29	2.86
<b>1f4g</b>	2.86	1.59	1.07	<b>1pbq</b>	0.64	0.41	0.3	<b>2f01</b>	1.22	0.34	0.29
<b>1f5k</b>	0.61	0.24	0.47	<b>1pr5</b>	2.26	2.14	1.72	<b>2fai</b>	1.14	0.55	0.32
<b>1fcx</b>	0.57	0.52	0.25	<b>1pxo</b>	1.9	2.34	2.39	<b>2flb</b>	1.15	1.46	0.72
<b>1fcz</b>	1.28	0.48	0.25	<b>1q7a</b>	0.83	2.26	1.13	<b>2fzc</b>	2.37	3.79	4.16
<b>1fd0</b>	0.8	0.62	0.27	<b>1q8t</b>	1.11	1.38	0.88	<b>2g5u</b>	0.59	0.5	0.51
<b>1fh7</b>	1.43	0.48	0.37	<b>1rdi</b>	1.17	2.37		<b>2g8r</b>	3.42	6.68	6.28
<b>1fh8</b>	2.28	1.76	0.27	<b>1rdj</b>	2.86	1	13.7	<b>2gss</b>	2.91	2.04	2.63
<b>1fh9</b>	2.33	2.04	0.71	<b>1rdl</b>	1.23	3.24	0.49	<b>2h3e</b>	1.03	0.62	0.34
<b>1fki</b>	2.08	4.63	1.1	<b>1re8</b>	2.84	2.63	0.87	<b>2hdq</b>	0.51	1.2	0.66

<b>1flr</b>	0.5	0.26	0.31	<b>1rnt</b>	2.82	2.92	3.11	<b>2j77</b>	0.87	0.29	0.8
<b>1ftm</b>	0.91	0.81	0.52	<b>1s39</b>	0.39	0.32	0.38	<b>2j78</b>	1.81	3.35	3.37
<b>1gni</b>	3.24	1.65	1.85	<b>1sqa</b>	1.29	1.3	0.97	<b>2qwb</b>	4.2	2.89	1.8
<b>1gpk</b>	1.48	4.86	0.33	<b>1sv3</b>	1.33	4.74	1.45	<b>2qwd</b>	1.67	0.64	0.4
<b>1gz9</b>	1.37	2.94	0.3	<b>1syh</b>	1.66	0.22	0.14	<b>2qwe</b>	1.2	0.74	0.29
<b>1h23</b>	2.15	2.22	3.4	<b>1tmn</b>	2.27	3.69	1.78	<b>2rkm</b>	2.36	1	0.33
<b>1ha2</b>	1.01	0.86	0.68	<b>1toi</b>	0.71	1.41	0.63	<b>2std</b>	2.38	0.75	0.55
<b>1hi4</b>	4.11	1.82	3.69	<b>1toj</b>	0.92	0.71	0.27	<b>2usn</b>	4.3	2.7	1.84
<b>1if7</b>	2.04	3.53	1.13	<b>1tok</b>	0.82	0.34	0.25	<b>3pce</b>	2.81	4.72	4.3
<b>1j16</b>	0.35	0.44	0.21	<b>1trd</b>	1.42	0.66	0.61	<b>3pch</b>	0.7	0.85	0.51
<b>1j17</b>	5.52	1.22	1.98	<b>1tsy</b>	1.58	3.85	1.28	<b>3pcj</b>	2.46	1.88	0.38
<b>1jaq</b>	3.04	3.42	5	<b>1ttm</b>	2.7	1.58	1.27	<b>3std</b>	1.26	6.03	0.64
<b>1jqd</b>	3.1	1.75	1.02	<b>1tyr</b>	3.77	3.69	3.39	<b>4tim</b>	1.55	0.89	0.63
<b>1jqe</b>	5.03	0.93	0.89	<b>1u2y</b>	1.52	3.79	3.63	<b>4tln</b>	1.45	2.46	1.8
<b>1jys</b>	0.66	0.7	0.79	<b>1utp</b>	3.25	2.13	1.42	<b>5abp</b>	0.72	0.27	0.35
<b>1k4g</b>	1.96	1.33	0.98	<b>1uwt</b>	1.3	3.35	3.44	<b>6fiv</b>	5.86	2.5	1.4
<b>1k9s</b>	1.49	1.07	0.7	<b>1v16</b>	2.62	2.29	2.1	<b>6rnt</b>	2.25	4.4	5.2
<b>1kv1</b>	1.49	0.6	0.57	<b>1v2o</b>	4.09	7.84	2.83	<b>6std</b>	2.62	0.9	0.86
<b>1kv5</b>	1.01	0.62	0.4	<b>1v48</b>	1.91	0.79	0.46	<b>8abp</b>	0.77	0.12	0.22
				<b>1vfn</b>	2.64	2.43	0.54	<b>8cpa</b>	3.65	1.61	2.98

B)

<i>PDBcode</i>	<i>LiGe</i>	<i>AutoDoc</i>	<i>Glid</i>	<i>PDBcode</i>	<i>LiGe</i>	<i>AutoDoc</i>	<i>Glid</i>	<i>PDBcode</i>	<i>LiGe</i>	<i>AutoDoc</i>	<i>Glid</i>
<i>e</i>	<i>n</i>	<i>k</i>	<i>e</i>	<i>e</i>	<i>n</i>	<i>k</i>	<i>e</i>	<i>e</i>	<i>n</i>	<i>k</i>	<i>e</i>
<i>1e66</i>	<i>1.03</i>	<i>0.53</i>	<i>3.91</i>	<i>1l2s</i>	<i>1.76</i>	<i>1.02</i>	<i>0.45</i>	<i>1vzq</i>	<i>8.05</i>	<i>0.75</i>	<i>0.29</i>
<i>10gs</i>	<i>3.87</i>	<i>7.91</i>	<i>2.43</i>	<i>1l83</i>	<i>1.06</i>	<i>1.15</i>	<i>0.97</i>	<i>1x1z</i>	<i>1.88</i>	<i>0.66</i>	<i>0.52</i>
<i>1a69</i>	<i>1.11</i>	<i>2.36</i>	<i>1.95</i>	<i>1li3</i>	<i>3.27</i>	<i>0.99</i>	<i>0.43</i>	<i>1xgj</i>	<i>3.13</i>	<i>7.9</i>	<i>7.81</i>
<i>1abf</i>	<i>1.52</i>	<i>0.32</i>	<i>0.32</i>	<i>1li6</i>	<i>2.77</i>	<i>3.35</i>	<i>2.38</i>	<i>1y1m</i>	<i>1.17</i>	<i>0.52</i>	<i>0.47</i>
<i>1ai5</i>	<i>4.54</i>	<i>1.37</i>	<i>0.92</i>	<i>1lol</i>	<i>2.09</i>	<i>0.53</i>	<i>2.54</i>	<i>1y6q</i>	<i>6.64</i>	<i>0.85</i>	<i>0.72</i>
<i>1ajp</i>	<i>4.23</i>	<i>4.47</i>	<i>4.11</i>	<i>1loq</i>	<i>2.26</i>	<i>0.67</i>	<i>0.6</i>	<i>1ydt</i>	<i>2.85</i>	<i>2.66</i>	<i>0.88</i>
<i>1ajq</i>	<i>1.82</i>	<i>1.68</i>	<i>1.46</i>	<i>1m0n</i>	<i>1.99</i>	<i>0.81</i>	<i>0.73</i>	<i>1zc9</i>	<i>2.92</i>	<i>3.21</i>	<i>3.3</i>
<i>1avn</i>	<i>4.65</i>	<i>6.03</i>	<i>3.62</i>	<i>1m0q</i>	<i>4.71</i>	<i>0.68</i>	<i>5.51</i>	<i>1zoe</i>	<i>4.57</i>	<i>5.58</i>	<i>5.06</i>
<i>1ax0</i>	<i>4.74</i>	<i>5.27</i>	<i>0.47</i>	<i>1m2q</i>	<i>4.77</i>	<i>3.71</i>	<i>1.31</i>	<i>1zs0</i>	<i>3.99</i>	<i>4.21</i>	<i>1.63</i>
<i>1axz</i>	<i>4</i>	<i>4.15</i>	<i>0.57</i>	<i>1n2v</i>	<i>4.08</i>	<i>2.74</i>	<i>0.62</i>	<i>1zvx</i>	<i>6.8</i>	<i>4.63</i>	<i>0.58</i>
<i>1b11</i>	<i>8.66</i>	<i>5.51</i>	<i>1.29</i>	<i>1nc1</i>	<i>5.65</i>	<i>0.93</i>	<i>0.77</i>	<i>2aou</i>	<i>8.17</i>	<i>1.83</i>	<i>2.6</i>
<i>1b7h</i>	<i>6.66</i>	<i>6.3</i>	<i>0.51</i>	<i>1ndw</i>	<i>4.07</i>	<i>1.19</i>	<i>7.33</i>	<i>2b1v</i>	<i>2.52</i>	<i>0.81</i>	<i>0.63</i>
<i>1b8o</i>	<i>2.77</i>	<i>2.33</i>	<i>2.51</i>	<i>1ndy</i>	<i>6.51</i>	<i>1.94</i>	<i>6.79</i>	<i>2baj</i>	<i>4.26</i>	<i>0.68</i>	<i>1.66</i>
<i>1b9j</i>	<i>7.97</i>	<i>7.4</i>	<i>0.65</i>	<i>1ndz</i>	<i>3.31</i>	<i>7.69</i>	<i>7.66</i>	<i>2bok</i>	<i>8.24</i>	<i>0.56</i>	<i>0.52</i>
<i>1bcu</i>	<i>1.13</i>	<i>2.88</i>	<i>0.32</i>	<i>1nfy</i>	<i>8.19</i>	<i>0.74</i>	<i>2.74</i>	<i>2brb</i>	<i>5.67</i>	<i>1.41</i>	<i>0.91</i>
<i>1bra</i>	<i>2.51</i>	<i>1.72</i>	<i>0.59</i>	<i>1nhu</i>	<i>6.61</i>	<i>7.29</i>	<i>6.51</i>	<i>2brm</i>	<i>1.99</i>	<i>6.92</i>	<i>4.43</i>
<i>1c84</i>	<i>1.78</i>	<i>0.89</i>	<i>1.01</i>	<i>1nja</i>	<i>2.22</i>	<i>3.52</i>	<i>1.62</i>	<i>2bz6</i>	<i>2.07</i>	<i>1.33</i>	<i>1.04</i>

<i>1d09</i>	1.87	0.82	0.54	<i>1nje</i>	2.21	1.82	6.41	<i>2c02</i>	5.08	2.98	3.72
<i>1d7j</i>	1.89	2.3	2.03	<i>1nvq</i>	0.83	3.33	0.56	<i>2ceq</i>	0.7	3.32	8.03
<i>1det</i>	4.59	2.61	1.34	<i>1nwl</i>	7.5	7.5	7.08	<i>2cer</i>	2.08	8.32	6.48
<i>1df8</i>	1.05	0.46	0.18	<i>1o0h</i>	1.77	3.55	9.07	<i>2cet</i>	0.81	3.93	7.6
<i>1dhi</i>	7.72	5.78	4.83	<i>1o3f</i>	1.44	1.7	1.23	<i>2cgr</i>	1.1	2.41	3.06
<i>1e1v</i>	2.87	9.22	7.22	<i>1o3p</i>	2.07	2	1.77	<i>2ctc</i>	0.76	0.53	0.35
<i>1e5a</i>	1.09	5.62	5.69	<i>1ols</i>	2.93	4.47	0.39	<i>2d0k</i>	2.33	1.1	3.22
<i>1ela</i>	1.98	5.26	2.36	<i>1olu</i>	2.81	0.88	2.43	<i>2d1o</i>	2.42	4.06	1.26
<i>1elb</i>	2.94	2.62	4.41	<i>1om1</i>	1.23	1.22	0.35	<i>2d3u</i>	1.11	0.8	1.44
<i>1f4e</i>	1.99	2.39	1.66	<i>1p1q</i>	4.73	2.96	1.74	<i>2d3z</i>	3.13	0.72	0.42
<i>1f4f</i>	9.01	3.42	3.39	<i>1pb9</i>	0.37	0.25	0.27	<i>2drc</i>	3.64	1.3	4.68
<i>1f4g</i>	9.69	3.84	1.07	<i>1pbq</i>	0.64	0.61	0.3	<i>2f01</i>	2.39	0.34	0.67
<i>1f5k</i>	0.61	0.31	0.47	<i>1pr5</i>	3.21	2.22	6.41	<i>2fai</i>	2.73	2.21	1.44
<i>1fcx</i>	0.57	0.55	0.31	<i>1pxo</i>	1.9	2.59	2.83	<i>2flb</i>	2.52	1.63	0.72
<i>1fcz</i>	1.28	0.5	0.25	<i>1q7a</i>	0.94	5.56	4.42	<i>2fzc</i>	4.33	4.77	4.52
<i>1fd0</i>	0.8	0.62	0.27	<i>1q8t</i>	1.47	1.71	0.88	<i>2g5u</i>	4.2	0.52	0.51
<i>1fh7</i>	1.43	1.96	0.42	<i>1rdi</i>	2.85	2.37	6.2	<i>2g8r</i>	7.44	7.18	6.28
<i>1fh8</i>	5.65	2.6	0.27	<i>1rdj</i>	4.24	5.18	13.7	<i>2gss</i>	3.18	7.49	2.8
<i>1fh9</i>	6.36	2.82	0.71	<i>1rdl</i>	4.07	3.32	3.74	<i>2h3e</i>	1.55	0.62	0.65
<i>1fki</i>	4.02	4.64	1.1	<i>1re8</i>	2.84	3.84	0.87	<i>2hdq</i>	2.18	1.23	1.55
<i>1flr</i>	2.49	0.26	0.82	<i>1rnt</i>	3.57	3.33	3.64	<i>2j77</i>	1.9	2.46	0.8
<i>1ftm</i>	0.91	2.21	0.52	<i>1s39</i>	0.52	0.33	0.38	<i>2j78</i>	3.65	3.82	3.85

<i>1gni</i>	5.96	5.52	1.88	<i>1sqa</i>	1.29	1.52	0.97	<i>2qwb</i>	5.04	5.02	1.8
<i>1gpk</i>	4.95	4.87	4.84	<i>1sv3</i>	2	4.74	1.76	<i>2qwd</i>	2.8	1.34	0.4
<i>1gz9</i>	1.37	2.94	0.74	<i>1syh</i>	1.66	0.22	0.14	<i>2qwe</i>	1.2	1.76	0.29
<i>1h23</i>	5.9	4.6	4.24	<i>1tmn</i>	2.27	8.02	2.67	<i>2rkm</i>	2.36	1	0.74
<i>1ha2</i>	1.01	0.86	0.76	<i>1toi</i>	1.35	4.36	1.34	<i>2std</i>	2.48	0.75	2.51
<i>1hi4</i>	4.67	3.66	4.72	<i>1toj</i>	0.92	4.38	0.27	<i>2usn</i>	4.71	6.64	9.84
<i>1if7</i>	2.39	9.57	2.01	<i>1tok</i>	1.48	0.4	0.25	<i>3pce</i>	4.41	5.04	4.68
<i>1j16</i>	0.35	3.08	0.53	<i>1trd</i>	1.42	1.07	1.09	<i>3pch</i>	4.07	1	4.2
<i>1j17</i>	8.15	1.22	2.46	<i>1tsy</i>	2.76	4.63	1.35	<i>3pcj</i>	3.24	2.51	4.05
<i>1jaq</i>	6.23	5.24	5.39	<i>1ttm</i>	3.27	3.88	2.96	<i>3std</i>	8.5	6.03	0.87
<i>1jqd</i>	3.53	8.6	1.02	<i>1tyr</i>	7.93	5.68	4.19	<i>4tim</i>	4.14	0.95	1.11
<i>1jqe</i>	5.19	5.24	1.62	<i>1u2y</i>	2.2	3.85	3.63	<i>4tln</i>	2.41	3.33	2.29
<i>1jys</i>	0.75	2.5	3.14	<i>1utp</i>	3.25	3.71	2.16	<i>5abp</i>	0.75	0.36	0.5
<i>1k4g</i>	4.24	1.37	1.39	<i>1uwt</i>	3.7	3.45	4.14	<i>6fiv</i>	8.08	4.19	5.36
<i>1k9s</i>	2.36	2.4	0.7	<i>1v16</i>	2.62	2.29	2.1	<i>6rnt</i>	6.86	4.4	6.67
<i>1kv1</i>	7.36	0.65	0.59	<i>1v2o</i>	4.89	8.73	3.35	<i>6std</i>	2.88	2.56	0.86
<i>1kv5</i>	1.44	0.62	0.43	<i>1v48</i>	2.71	0.89	1.33	<i>8abp</i>	0.94	3.95	0.51
				<i>1vfn</i>	2.87	2.43	4.63	<i>8cpa</i>	3.65	4.21	3.72





Mean	12.90	11.00	8.17	6.12	12.06	9.01	8.30	7.29	9.89	10.14	4.80	0.00	11.81	12.38	5.80	5.43
Median	1.05	1.90	3.50	3.65	0.85	1.25	2.80	5.90	2.65	0.35	0.00	0.00	3.70	0.85	0.00	1.50
SD	23.64	19.79	10.55	8.55	19.63	16.52	11.62	8.34	17.49	18.86	8.83	0.00	19.72	20.86	9.54	7.37
ROC (5%)																
Mean	21.27	19.37	16.44	12.65	17.69	16.63	12.86	13.04	20.61	19.08	13.71	0.00	16.89	18.13	11.76	10.56
Median	8.15	4.25	9.20	5.10	9.10	2.75	8.15	11.70	10.85	5.40	8.55	0.00	6.55	7.50	5.80	4.50
SD	29.44	29.55	17.11	12.88	23.28	24.23	14.80	11.51	25.45	25.43	15.23	0.00	23.28	24.54	13.95	13.26
ROC (20%)																
Mean	38.62	37.36	33.91	38.89	39.18	40.24	34.45	41.74	41.12	38.32	34.58	7.23	39.46	38.64	29.36	29.50
Median	23.65	24.45	25.70	25	30.10	29.70	34.05	36.50	27.25	29.80	27.30	0.00	28.95	27.25	25.30	24.85
SD	31.67	32.67	25.27	32.17	32.29	34.08	25.97	29.99	29.78	26.50	29.23	20.84	32.80	32.98	26.93	24.08
BEDR OC $\alpha=20.0$																
Mean	0.18	0.17	0.14	0.13	0.17	0.15	0.13	0.14	0.18	0.15	0.11	0.00	0.17	0.17	0.11	0.10
Median	0.07	0.05	0.08	0.08	0.10	0.06	0.09	0.11	0.10	0.05	0.06	0.00	0.10	0.07	0.06	0.06
SD	0.24	0.23	0.14	0.12	0.20	0.19	0.13	0.11	0.20	0.19	0.12	0.01	0.19	0.21	0.12	0.11

**Table VI.** Comparison of VS results considering only the “own decoys” subset of DUD, with the original set of parameters (on the left side of the table) and the optimized ones (on the right).

Original parameters – “own decoys” subset									Optimized parameters - “own decoys” subset					
	PDB code	ROC C	ROC (1%)	ROC (2%)	ROC (5%)	ROC (20%)	BEDROC $\alpha=20.0$	ROC C	ROC (1%)	ROC (2%)	ROC (5%)	ROC (20%)	BEDROC $\alpha=20.0$	
serine proteasi														
FXa	1f0r	0.64	1.4	2.8	3.5	38.7	0.09	0.54	1.4	2.1	2.1	17.6	0.04	
thrombin	1ba8	0.63	0	3.1	12.3	50.8	0.06	0.75	1.5	1.5	4.6	56.9	0.12	
trypsin	1bjv	0.47	0	0	2.3	22.7	0.02	1.01	9.1	9.1	11.4	11.4	0.12	
kinase														
FGFr1	1agw	0.59	0	3.4	6.8	15.3	0.07	0.85	0.8	2.5	10.2	47.5	0.26	
CDK2	1ckp	0.83	0	0	6	12	0.06	0.37	0	2	2	8	0.03	
EGFr	1m17	0.63	0.5	0.5	2.3	32.9	0.04	0.95	2.3	3.4	32.9	87.4	0.33	
HSP90	1uy6	0.45	0	0	0	33.3	0.02	0.55	0	0	0	12.5	0.05	
SRC	2src	0.93	1.9	3.2	9.7	53.5	0.11	0.67	1.3	3.2	8.4	40	0.11	

TK	1kim	0.31	0	0	0	27.3	0.05	0.65	0	0	4.5	27.3	0.05
p38	1kv2	0.8	2	3.9	10.9	23.8	0.13	0.63	2.3	7.8	15.6	38.3	0.17
metalloenzyme													
ACE	1o86	0.9	2	4.1	8.2	28.6	0.05	0.66	2	2	2	22.4	0.05
COMT	1h1d	0.2	0	0	9.1	9.1	0.02	0.69	0	0	0	45.5	0.07
PDE5	1xp0	0.56	0	5.9	5.9	27.5	0.06	0.65	3.9	3.9	13.7	35.3	0.15
nuclear hormone receptor													
ERagonist	1l2i	0.89	0	0	3	35.8	0.1	0.88	3	14.9	38.8	76.1	0.37
ERantagonist	3ert	0.52	0	0	0	23.1	0.06	0.59	5.1	5.1	7.7	12.8	0.06
GR	1m2z	0.3	1.3	1.3	1.3	1.3	0.02	0.91	41	52.6	57.7	79.5	0.38
MR	2aa2	0.86	0	0	0	40	0.36	0.95	0	0	0	33.3	0.05
PPARg	1fm9	0.63	0	0	1.2	6.2	0.01	0.79	0	0	2.5	13.6	0.04
PR	1sr7	0	0	0	0	0	0	0.69	3.7	7.4	11.1	29.6	0.16
RXR	1mvc	0	0	0	0	0	0	0.99	35	60	70	75	0.63
folate enzyme													
DHFR	3dfr	0.26	0	0	0	6	0	0.54	0.5	1.5	6	21.9	0.07

GART	1c2t	0.1 3	0	0	0	9.5	0	0.5 5	0	0	4.8	14.3	0.07
other enzyme													
COX-2	1cx2	0.7 7	0.9	1.1	1.1	1.1	0.01	0.8 4	21	33.6	46	68.4	0.45
PARP	1efy	0.5 9	3	3	3	18.2	0.07	0.5 8	12.1	18.2	21.2	48.5	0.27
AChE	1eve	0.6 3	1	1	1.9	22.9	0.02	0.5 8	1	2.9	6.7	29.5	0.08
HIVPR	1hpx	0.6 6	5.7	9.4	20.8	54.7	0.13	0.6 9	1.9	3.8	11.3	35.8	0.13
HMGR	1hw8	0.7 4	2.9	2.9	8.6	48.6	0.12	0.9 9	0	0	8.6	20	0.13
InhA	1p44	0.6 2	0	2.4	7.1	34.1	0.09	0.7 3	3.5	5.9	9.4	35.3	0.12
COX-1	1p4g	0.6 6	0	4	4	4	0.03	0.7 4	4	4	4	20	0.07
HIVRT	1rt1	0.5 4	0	0	5	22.5	0.02	0.6 6	0	2.5	7.5	32.5	0.09
AmpC	1xgj	0.4 7	0	0	0	0	0	0.3 6	0	0	0	0	0
SAHH	1a7a	0.3	3	3	3	3	0.04	0.8 9	45.5	66.7	93.9	97	0.36
GPB	1a8i	0.5 7	0	0	5.8	11.5	0.1	0.7 1	0	1.9	3.8	23.1	0.07
ALR2	1ah3	0.5 8	3.8	3.8	15.4	30.8	0.03	0.6 4	3.8	3.8	7.7	34.6	0.13
PNP	1b8o	0.6 7	4	8	16	36	0.16	0.5 6	0	0	28	72	0.02
NA	1a4g	0.9 8	10.2	10.2	18.4	38.8	0.14	0.8 7	8.2	22.4	49	83.7	0.44

mean	0.56	0.92	1.85	4.8	22.38	0.06	0.71	5.94	9.58	16.75	39.07	0.16		
mediana	0.61	0	0.75	3	22.8	0.05	0.69	1.7	3.05	8.05	33.95	0.12		
sd	0.24	1.45	2.36	5.15	16.84	0.07	0.16	11.27	16.73	21.6	25.18	0.15		

# Appendix B

In the following pages is reported a printable version of the database discussed in CHAPTER 3. Every row corresponds to one database entry and columns are database fields.

- ID: unique identification code of the entry;
- NA: common name of the entry;
- SM: SMARTS notification that represents the structural alert
- RK: rank
- DE: description of the entry
- RF: literature references
- DR: references to other databases or websites
- KW: keywords to search through the database
- RE: residues target of the covalently binding SAs.

ID	NA	SM	RK	DE	RF	DR	KW	RE
<b>HL</b> <b>AC</b>	Acyl Halide	CC([F,Cl,Br,I])=O	3	Acyl halides are electrophilic agents that can covalently and unspecifically bind to biological nucleophiles in proteins and DNA; acyl halide derivatives are considered genotoxic carcinogens as they give positive results in AMES test, skin sensitizers as they can also covalently bind to skin proteins. Acyl halide derivatives have also been reported as possibly causing acute aquatic toxicity due to their electrophilic nature. They act	PMID:23061697; PMID:20693071; PMID:15656773; PMID:8144710; PMID:18621573; PMID:15935536; PMID:2269228; PMID:21809939; PMID:12565011;	ToxAlerts:TA358;ToxAlerts:TA577; ToxAlerts:TA704; wiki:Acyl_halide	acid halide; acylating ; skin; genotox; aquatic toxicity;	CYS; NH2; LYS
<b>AC</b> <b>IZ</b>	Acylimidazoles; N-acetylimidazole	N(C=NC1)(C=1)!@C(!@C)=O	2	N-acetylimidazoles (NAI) are tyrosine-selective acylating agents used in protein chemical modification to clarify structure-function relationships of this type of residue. Hence the presence of this type of moiety in a molecule could lead to selective modification of tyrosine residues, in particular solvent exposed ones, in the target protein; acylation of threonine and serine residues is also possible, even if less common.	PMID:23061697; PMID:5870321; SARF:2; PMID:20131845;	ToxAlerts:TA1736;	acylating ; acylimidazole;	TYR; SER; THR; NH2
<b>AT</b> <b>RZ</b>	Acyltriazoles; N-acyltriazole	c1nnn(C(=O))c1.c1n(C(=O))cnn1	2	Acyltriazole are acylating agent; they have been reported to interfere in HTS assays, forming unspecific covalent bond with proteins.	PMID:20131845; PMID:16711725;	ToxAlerts:TA1761; ToxAlerts:TA1799; ToxAlerts:TA2239;		TYR; SER; THR; NH2
<b>AN</b> <b>HY</b>	Anhydrides	CC([O,S]C(C)=[O,S])=[O,S]	3	Acyclic anhydrides and cyclic acid anhydride are electrophilic acylating agent; they can react with biological	PMID:15656773; PMID:23061697; PMID:20693071;	ToxAlerts:TA226; ToxAlerts:TA267; ToxAlerts:TA286;	acylating ; anhydrid	NH2, CYS

				nucleophiles present in proteins, forming covalent adducts that may result in skin sensitization reaction in humans and animal models, HTS assay interference and acute aquatic toxicity;	PMID:8144710; PMID:2269228; PMID:21809939; PMID:18853302; PMID:12565011; PMID:16711725;	ToxAlerts:TA476; ToxAlerts:TA580; ToxAlerts:TA585; ToxAlerts:TA607; ToxAlerts:TA672; ToxAlerts:TA706; ToxAlerts:TA715; wiki:Acid_anhydride; ToxAlerts:TA1810;	e;	
<b>SIS N</b>	Si (silicon) or Sn (tin)	[Si,Sn,B,Se]	1	unstable, form toxic phospho-organics, nonspecific binder to -N-H, -O-H or similar group. Not allowed atom (RF 1)	PMID:23061697; PMID:20131845;	ToxAlerts:TA1871; ToxAlerts:TA1733; ToxAlerts:TA1780; ToxAlerts:TA701	Si; silicon; silicium	CYS; NH2; LYS;
<b>SU HA</b>	Sulfonylhalides; sulphonyl halide; sulphonyl halide; sulfonyl halide	S([F,Cl,Br,I])(C)(=O)=O	3	Sulfonyl halides are considered acylating agent; they can react with endogenous nucleophiles, that can attack the sulfonyl halide at the sulfur atom, forming a tetrahedral intermediate and causing the release of the halogen leaving group; sulfonyl halide are considered akin sensitizer agents as they can react with protein nucleophiles producing chemically modified proteins that can lead to T-cell-mediated allergic response;	PMID:23061697; PMID:20693071; PMID:8144710; PMID:21809939; PMID:12565011; PMID:21401043;	ToxAlerts:TA224; ToxAlerts:TA708; ToxAlerts:TA1860; wiki:Sulfonyl_halide;	sulfonyl halide; sulphonyl halide; acylating	CYS; NH2
<b>SU AH</b>	Sulfonylanhydrides	S(OS(=O)(=O)C)(C)(=O)=O	2	sulfonylanhydride may act as acylating agent by itself or after hydrolysis; it can bind to nucleophilic protein groups as sulfonyl halide do, but there are no reports of causing skin sensitization, only evidences of	PMID:21809939; PMID:20131845; PMID:16711725;	ToxAlerts:TA1783;	sulphonylanhydride; sulfonylanhydride;	CYS; A; G;



<b>AK SN</b>	Alkylsulfonates	<chem>S(=O)(=O)C(=O)C</chem>	3	interfering with HTS assays; Alkylsulfonate have been reported to act as alkylating agent undergoing a SN2 reaction on the sp3 carbon atom; if the carbon atom directly bound to the sulfur atom is involved in a carbon-carbon double or triple bond, so in case of alpha-beta unsaturated sulphonate they can also undergo a Michael addition reaction to form a covalent bond with protein nucleophilic residues; alkylsulfonated are reported to be skin sensitizers; several	PMID:20693071; PMID:21809939; PMID:8144710;	wiki:Alkyl_sulfonate; ToxAlerts:TA234; ToxAlerts:TA281; ToxAlerts:TA727; ToxAlerts:TA775;	alkylsulfonate; alkylsulfonate;	CYS;
<b>AL SU</b>	Alkylsulfates	<chem>S(=O)(=O)OC(=O)C</chem>	1	Alkylsulfates can in principle be considered alkylating agent reacting via SN2 mechanism because they have a reactive carbon sp3 atom; however skin sensitization has been reported only at high concentrations only for some derivatives; no mutagenic, genotoxic or carcinogenic effect has been related to alkylsulfate use;	PMID:23061697; PMID:20693071; PMID:21809939; PMID:8144710;	ToxAlerts:TA234; ToxAlerts:TA282; ToxAlerts:TA774; wiki:Organosulfate;	alkylsulfate; alkylating;	CYS; NH2; LYS; A; G;
<b>ICY N</b>	Isocyanates	<chem>N=C=C=O</chem>	3	isocyanates are electrophilic functional groups that behaves as acylating agent; isocyanate can react with nucleophilic site on protein and nucleic acid, causing skin sensitization reaction, DNA and RNA mutation leading to genotoxicity; acute aquatic toxicity has also been reported;	PMID:23061697; PMID:20693071; PMID:8144710; PMID:2269228; PMID:21809939; PMID:18621573; PMID:15935536;	wiki:Isocyanate; ToxAlerts:TA229; ToxAlerts:TA315; ToxAlerts:TA372; ToxAlerts:TA423; ToxAlerts:TA609; ToxAlerts:TA678; ToxAlerts:TA716;	isocyanate; acylating; genotoxic;	CYS; NH2; LYS; A; G;
<b>IT CN</b>	Isothiocyanates	<chem>N=C=S</chem>	3	Isothiocyanate are strong electrophilic groups reacting with nucleophilic	PMID:20693071; PMID:8144710;	wiki:Isothiocyanate;	isothiocyanate;	CYS; NH2;

				group in protein residues, DNA and RNA; they act through acylating reaction mechanism that can be also considered a Schiff base formation reaction; isothiocyanate derivatives are reported to cause skin sensitization and acute aquatic toxicity; even though several papers reported genotoxic effects for thiocyanate derivatives, their genotoxicity is controversi	PMID:2269228; PMID:21809939; PMID:17317210; PMID:8159717; PMID:19402170; PMID:21461351; PMID:21241062; PMID:21401043; PMID:18621573; PMID:15935536;	ToxAlerts:TA230; ToxAlerts:TA316; ToxAlerts:TA372; ToxAlerts:TA424; ToxAlerts:TA610; ToxAlerts:TA680; ToxAlerts:TA717;	acylating ; genotox	LYS;
<b>CB DI</b>	Carbodiimides	<chem>N(C)=C=NC</chem>	1	Carbodiimides are electrophilic group reacting with nucleophilic group in protein residues, through acylation; they hydrolyze to form ureas;	PMID:21809939; PMID:20131845;	ToxAlerts:TA718; wiki:Carbodiimide; ToxAlerts:TA1865;	carbodiimide; acylating	CYS;
<b>TH CY</b>	Thiocyanates	<chem>CSC#N</chem>	2	Thiocyanate are alkylating agent that may undergo SN2 reaction with cysteine residues, forming a covalent disulfide bridge; thiocyanate derivatives have been reported to give positives results in acute aquatic toxicity tests;	PMID:23061697; PMID:2269228; PMID:21809939;	wiki:Thiocyanate; ToxAlerts:TA653; ToxAlerts:TA679; ToxAlerts:TA799;	thiocyanate; alkylation; cysteine; disulfur bridge	CYS;
<b>HY DA</b>	Hydroxamic acid	<chem>C!@C(=O)!@[N;H][O;H]</chem>	1	hydroxamate functional group has been shown to act as metal chelator; they may chelate metal ions present in metalloproteinases as well as react with thiols; metal ion chelation represents a common artifacts in biological screening; several compounds targeting histone deacetylase carrying hydroxamate functional group are under	PMID:23061697; PMID:22837956; PMID:23701657; PMID:19239360; PMID:21139608; PMID:12565011;	wiki:Hydroxamic acid; ToxAlerts:TA1723;	hydroxamic acid; hydroxamate;	CYS;

				development as anticancer and anti-HIV drugs;			
<b>ET CA</b>	Ethylcarbamates	<chem>[C;H3][C;H2]OC(=O)NC</chem>	1	carbamates are useful functional groups in drug discovery, they are used in cholinesterase inhibitors such as rivastigmine, that covalently modifies S200 in acetylcholinesterase, in the anti-epileptic drug felbamate and in other approved drugs; however some derivatives have been related to cancer; in particular ethylcarbamate, used in the past as antineoplastic drug, has been added by IARC to group 2A carcinogen, that means it is	PMID:20205516; PMID:20529350; PMID:18621573;	wiki:Ethyl_carbamate; wiki:Carbamate; ToxAlerts:TA373; ToxAlerts:TA440; ToxAlerts:TA690;	ethylcarbamate; carbamate;
<b>AZ ID</b>	Azides	<chem>[N-]=[N+]=NC</chem>	3	Azides are classified as genotoxic carcinogens, mutagens, and skin sensitizers; azide can be bioactivated to produce highly electrophilic compound, that may react with nucleophilic site in DNA, RNA and proteins;	PMID:20693071; PMID:18621573; PMID:16922655; PMID:16711725;	wiki:Azide; ToxAlerts:TA223; ToxAlerts:TA225; ToxAlerts:TA336; ToxAlerts:TA379; ToxAlerts:TA825; ToxAlerts:TA1775;	azide; Idiosyncratic toxicity; genotoxicity
<b>TH NE</b>	Thione esters	<chem>C([\$(C=C),\$(C#C),\$(a)])OC(=S)([H,C])</chem>	1	Thione esters of aromatic derivatives or deriving from unsaturated alcohol have been suggested to act as skin sensitizers;	PMID:18853302; PMID:23061697;	ToxAlerts:TA462;	thione ester;
<b>DT HN</b>	dithione	<chem>CC(=S)C(=S)C</chem>	2	Dithiones have been reported to induce skin sensitization that can react with protein hard nucleophiles through a Schiff base formation reaction;	PMID:23061697; PMID:18853302; PMID:20131845;	ToxAlerts:TA454; ToxAlerts:TA455; ToxAlerts:TA1958;	dithione
<b>PE RO</b>	Peroxides	<chem>OO</chem>	3	peroxides are functional group with an redox-unstable oxygen-oxygen single	PMID:20693071; PMID:8144710;	wiki:Peroxide; ToxAlerts:TA227;	peroxide ; radical;

				bond, that can split into reactive radicals; radicals just formed may causes toxicity via protein, lipids and DNA oxidation; peroxides and peroxide-radicals are related to skin sensitization reaction, genotoxicity and artifacts in biological and biochemical assaysd;	PMID:12565011; PMID:16711725;	Toxalerts:TA303; Toxalerts:TA304; Toxalerts:TA305; Toxalerts:TA307;		
<b>AC DS</b>	Acyclic disulfides	[S;R0][S;R0]	3	acyclic disulfide can be hydrolized into thiols, and subsequently undergoing SN2 reaction with cysteines exposed on protein surface to form covalent disulfide bridges; compounds containing disulfide bridges have been related to skin sensitization, acute aquatic toxicity and to idiosyncratic toxicity; disulfide derivatives may also cause assay interferences, reaction aspecifically with proteins and with sulfur-containing media comp	PMID:23061697; PMID:8144710; PMID:2269228; PMID:21809939; PMID:16922655; PMID:12565011; PMID:20131845;	ToxAlert:TA301; ToxAlert:TA648; ToxAlert:TA649; ToxAlert:TA651; ToxAlert:TA795; ToxAlert:TA797; ToxAlert:TA839; wiki:Disulfide; ToxAlert:TA1870;	disulfide ; sulfur bridges;	CYS;
<b>NI TS</b>	Nitroso	O=[N;D2]C	2	alkyl and aryl nitroso derivatives can react with a broad spectrum of biological nucleophiles through Schiff base formation reaction; nitroso derivatives have been related to genotoxic carcinogenicity, mutagenicity; the degree of hazard related to nitroso derivatives depends on substituents, that can enhance or mitigate the electrophilic potential of the nitroso group, making it more prone to react with nucleophilic site;	PMID:18621573; PMID:15935536; PMID:15634026;	wiki:Nitroso; Toxalerts:TA324; ToxAlerts:TA331; ToxAlerts:TA382; ToxAlerts:TA397; ToxAlerts:TA456;	nitroso;	

<b>AN IL</b>	Aniline	c1ccccc1[N;H2]	2	Aniline derivatives are reported as structural alerts for bioactivation: the phenil ring might be oxidated to produce a reactive quinone-imine metabolite. Several marketed drug containing an alanine moiety received a black box warning from FDA due to idiosyncratic adverse drug reactions;	PMID:16922655; PMID:21702456; PMID:17302443; PMID:21455238;	wiki:Aniline	aniline;
<b>AN LD</b>	Anilides	c1ccccc1[N;H]C(=O)[C;H3]	2	Anilide derivatives are reported as structural alerts for bioactivation: the phenil ring might be oxidated to produce a reactive quinone-imine metabolite. Several marketed drug containing an alanine moiety received a black box warning from FDA due to idiosyncratic adverse drug reactions;	PMID:16922655; PMID:21702456; PMID:17302443; PMID:21455238;	wiki:Aniline	anilide;
<b>NN IT</b>	N-nitro	O=[N+](N)[O-]	3	Compounds presenting N-nitro groups are metabolic unstable; due to their high electophilicity they can react with highly nucleophilic site on DNA, RNA and proteins; they give positive results in AMES assay to assess genotoxicity, in particular aliphatic N-nitro derivatives;	PMID:23061697; PMID:18621573;	ToxAlerts:TA350; ToxAlerts:TA380;	N-nitro A; G; CYS; LYS;
<b>NN IS</b>	N-nitroso, nitrosamine	O=[N;D2]N	3	nitrosamines, especially nitrosoureas and nitrosoguanidine, may react via SN2 mechanism to add a nitroso group to cysteine (and/or lysine) residues exposed on protein surface, in the so called "nitrosation reaction"; due to their ability of alkylate proteins they have been related to genotoxic	PMID:23061697; PMID:21809939; PMID:18621573; wiki:Nitrosamine;	ToxAlerts:TA378; ToxAlerts:TA399; ToxAlerts:TA400; ToxAlerts:TA401; ToxAlerts:TA402; ToxAlerts:TA403; ToxAlerts:TA404; ToxAlerts:TA464;	N-nitroso; nitrosamine;

				carcinogenicity, mutagenicity; n-nitroso compounds are also metabolic unstable;		ToxAlerts:TA790; ToxAlerts:TA971	
<b>AZ OX</b>	Azo, Diazos compounds	C/[N;D2]!@[N;D 2]/C	3	the stability of the azo-compound depend on the substituents attached to the nitrogen atoms; the most stable are those bearing two aryl groups, whereas alkyl derivatives are less stable; however under reductive conditions azo-compounds bearing aromatic rings as substituents, break down into two aromatic amines; azo-groups with aromatic substituents are frequent in dyes and pigments and some derivatives, used as dye, have been found	PMID:23061697; PMID:21809939; PMID:15634026; PMID:12565011;	wiki:Azo_compoun d; ToxAlerts:TA258; ToxAlerts:TA262; ToxAlerts:TA326; ToxAlerts:TA337; ToxAlerts:TA339; ToxAlerts:TA351; ToxAlerts:TA371; ToxAlerts:TA388; ToxAlerts:TA425; ToxAlerts:TA428;	azo; diazos genotox
<b>AZ OX ;</b>	Azoxy	C/N=[N+](/!@C)O	3	Azoxy compounds are oxidate azo derivatives also called N-oxide; azoxy compounds give positive results in AMES genotoxicity assay;	PMID:18621573; PMID:15935536;	wiki:Azoxy; ToxAlerts:TA371; ToxAlerts:TA406;	azoxy;
<b>NI TR</b>	Nitroso dimers	C/[N+](=[N+](/C) O)O	3	unstable, nonspecific binder, breaks by N=N and forms mutagenic and cancerogenic nitroso-like compounds			
<b>HL AM</b>	Haloamine s	CN([F,Cl,Br])[C;R]	3	Haloamines are functional groups containing an halogen atom directly bound to a nitrogen atom; linear haloamines are quite reactive, particularly prone to oxidation, and they have been shown to be genotoxic in in vitro assays; haloimides (RC(=O)N(X)(C=O)R) have been reported to be genotoxic as well, and	PMID:23061697; PMID:20693071; PMID:21809939; PMID:15935536;	ToxAlerts:TA417 (haloamine); ToxAlerts:TA250; ToxAlerts:TA804;	haloami ne; haloimid e; oxidizer; genotox

				to induce skin sensitization; there are no clear evidences of the mechanism, however an hypothetical mechanism involving a SN2 re				
<b>SF EH</b>	Sulfenyl halides	[F,Cl,Br,I][S;D2]C	2	Sulfenyl halides have a halogen atom directly bound to a sulfur atom; as haloamine, they are unstable and can form RCS+, that act as oxidizers; due to the electrophilic nature, sulfenyl halides can react with cysteine residues, via SN2 reaction mechanism, forming a covalent disulfide bridge and leading to the final alkylated protein adduct; compounds carrying this functional group gave positive results in acute aquatic toxicity t	PMID:23061697; PMID:2269228; PMID:21809939;	wiki:Sulfenyl_chlor ide; ToxAlerts:TA654; ToxAlerts:TA798;	sulfenyl halides; oxidizer; disulfide bridge	
<b>SF H</b>	Sulfinyl halides	[F,Cl,Br,I][S;D3](C )=O	3	unstable, nonspecific binder to hydroxy- or amino-group				
<b>C2 C2</b>	1,2 dienes (allenes)	CC=C=C	2	Allenes are strong acylating agents, that can react via Michael type addition reaction to form a covalent adduct with the protein; due to the two double bond they are more reactive than simple alkenes; moreover the reactivity can be further increased by the presence of electrowithdrawing substituents like NO2, C=O, C#N, S=O;	PMID:23061697; PMID:18853302;	wiki:Allene; ToxAlerts:TA471	allene; diens	
<b>R3 O</b>	Oxiranes; epoxide	C(O1)C1	3	Epoxides are unstable strained heterocycles, highly reactive and prone to ring opening at C-O bond, forming covalent adducts with	PMID:23061697; PMID:21809939; PMID:18621573; PMID:15935536;	wiki:Oxirane; ToxAlerts:TA409; ToxAlerts:TA254; ToxAlerts:TA271;	oxirane; alkylatin g; epoxide;	CYS; NH2; LYS;

				biological nucleophiles through a SN2 reaction; due to their high reactivity epoxides have been shown to act as alkylating agents, skin sensitizers and genotoxic carcinogens and to induce acute aquatic toxicity;	PMID:2269228; PMID:20693071; PMID:15656773; PMID:8144710; PMID:15634026; PMID:12565011;	ToxAlerts:TA283; ToxAlerts:TA334; ToxAlerts:TA364; ToxAlerts:TA442; ToxAlerts:TA613; ToxAlerts:TA666; ToxAlerts:TA785;		
<b>R3 S</b>	Thiiranes; thiorane	C(S1)C1	3	Thiirane are strained sulfur-containing ring systems; they can undergo a ring opening reaction at C-S bond, give SN2 reaction with biological nucleophiles, in particular S-containing ones, forming an alkylated adduct via SS bridge;	PMID:23061697; PMID:21809939;	ToxAlerts:TA787; wiki:Thiirane;	thiorane; thiirane	CYS;
<b>R3 N</b>	Aziridines	C(N1)C1	3	Aziridines are three-membered ring systems, highly unstable; due to their strong electrophilic nature they are subject to be attacked by endogenous nucleophiles that causes the opening of the ring system and the formation of a covalent adduct with the nucleophiles via SN2 reaction mechanism; given their high reactivity aziridines react with several nucleophiles, present in DNA, RNA and proteins, causing DNA and RNA mutations that	PMID:23061697; PMID:2269228; PMID:21809939; PMID:18621573; PMID:12565011;	wiki:Aziridine; ToxAlerts:TA335; ToxAlerts:TA364; ToxAlerts:TA409; ToxAlerts:TA614; ToxAlerts:TA786;	aziridine s; genotox;	CYS; A; G; NH2; LYS;
<b>TH AD</b>	Thiazolidi nedione	C1SC(=O)NC1(=O)	2	The thiazolidinedione moiety present in some drugs withdrawn from the market or with black box warning is considered a possible cause of the IADR; CYP3A can catalyze the opening of the thiazolidinedione ring,	PMID:16922655; PMID:21702456; PMID:17302443; PMID:21455238;	ToxAlerts:TA849;	thiazolid inedione ;	



				producing a sulfur-containing metabolite that, in case of troglitazone, pioglitazone and rosiglitazone, all containing the thiazolidinedione moiety, has been related to BSEP inhibition;				
<b>TE RT</b>	tertiary alkyl halides	CC([C;R0])([C;R0])[Cl,Br,I]	3	alkyl halides are potential electrophilic agents that can undergo SN2 reaction with biological nucleophiles producing protein alkylation. Some studies report genotoxicity (RF11), acute aquatic toxicity (RF 6), skin sensitization (RF 2,3,4) and the possibility of idiosyncratic toxicity due to formation of reactive metabolites (RF12)	PMID:20693071; PMID:2269228; PMID:21809939; PMID:15634026; PMID:16922655; PMID:12565011;	wiki:Haloalkane; ToxAlerts:TA439; ToxAlerts:TA634; ToxAlerts:TA635; ToxAlerts:TA327; ToxAlerts:TA342; ToxAlerts:TA655; ToxAlerts:TA856	tertiary alkyl halides; alkylating	
<b>SE CO</b>	secondary alkyl halides	[C;^3]@[C;H][C;H2][Cl,Br,I]	2	alkyl halides are potential electrophilic agents that can undergo SN2 reaction with biological nucleophiles resulting in protein alkylation; some studies report genotoxicity, acute aquatic toxicity, skin sensitization reaction and the possibility of idiosyncratic toxicity due to the formation of reactive metabolites;	PMID:20693071; PMID:2269228; PMID:21809939; PMID:15634026; PMID:15935536; PMID:16922655; PMID:12565011;	wiki:Haloalkane; ToxAlerts:TA772; ToxAlerts:TA439; ToxAlerts:TA634; ToxAlerts:TA635; ToxAlerts:TA327; ToxAlerts:TA342; ToxAlerts:TA407; ToxAlerts:TA655;	secondary alkyl halide; alkylating	CYS; NH@;
<b>AL DE</b>	aldehydes	[c;C][C;H](=O)	3	aldehyde are quite reactive functional groups that can act as non-specific binders forming covalent bonds with serine or cysteine residues on protein surfaces, resulting in hemiacetal derivatives; they can also react with free -NH2 (lys and arg + protein terminal NH2) via Schiff base reaction	PMID:23061697; PMID:20693071; PMID:2269228; PMID:21809939; PMID:18621573; PMID:18853302; PMID:12565011; PMID:21401043;	ToxAlerts:TA244; ToxAlerts:TA264; ToxAlerts:TA290; ToxAlerts:TA368; ToxAlerts:TA432; ToxAlerts:TA449; ToxAlerts:TA450; ToxAlerts:TA451;	aldehyde ;	CYS; SER; NH2; LYS; ARG;

				to form imines; endpoints reported as consequences of these chemical reactions are skin sensitization, genotoxic carcinogenicity, skin sensi		ToxAlerts:TA611; ToxAlerts:TA677; ToxAlerts:TA762; wiki:Aldehyde;		
<b>AR AC</b>	Arylacetic,	<chem>c1ccccc1[C;H2]C(=O)[O;H]</chem>	2	arylacetic derivatives are a common scaffold in NSAIDs, however some marketed drug have been withdrawn from the market or received a black box warning due to IADRs; these idiosyncratic ADRs are believed to be immune-mediated responses linked to the beta-1-O-acyl glucuronide metabolites, that can covalently modify proteins via transacylation reaction or through acyl migration within the beta-O-glucuronide to a reactive alpha-hydro	PMID:16922655; PMID:21702456; PMID:17302443; PMID:21455238;	wiki:Non-steroidal_anti-inflammatory_grug	arylacetic acid, nsaid	LYS; NH2;
<b>HY ZI</b>	hydrazines and hydrazides	<chem>[N;D3](C)(C)!@N!@[C]</chem>	2	hydrazines are nucleophiles derivatives and, if the nitrogen atoms are not fully substituted, they can be highly unstable, leading to breakage of the N=N bond and to the formation of hydrazonium ion, that easily react with proteins, forming covalent adducts and breaking peptide bonds; hydrazine derivatives have been shown to cause skin sensitization, genotoxic carcinogenicity, and are also source of promiscuity in experimental ass	PMID:23061697; PMID:15656773; PMID:18621573; PMID:21809939; PMID:16922655;	wiki:Hydrazine; ToxAlerts:TA262; ToxAlerts:TA318; ToxAlerts:TA354; ToxAlerts:TA370; ToxAlerts:TA405; ToxAlerts:TA431; ToxAlerts:TA668; ToxAlerts:TA825; ToxAlerts:TA1794;	genotox; hydrazine; e; hydrazides	
<b>HY ZO</b>	hydrazone	<chem>N(!@N)=C</chem>	1	hydrazones are electrophilic functional group that can bind to	PMID:23061697; PMID:20131845;	wiki:Hydrazone; ToxAlerts:1905;	hydrazone;	

				proteins in an unspecific way, especially forming aggregates in in-vitro assays consequently giving false positives results; some hydrazones derivatives gave positives results in AMES assay for genotoxicity, however many others did not, making it difficult to directly relate the genotoxic outcomes to the hydrazone functionality alone;	PMID:21401043;	ToxAlerts:1911; ToxAlerts:1934; ToxAlerts:1935;	
<b>HY AM</b>	hydroxylamines	c[N;D2][O;H]	3	hydroxylamines are nucleophiles functional groups, whose nucleophilicity is increased by lone pair electrons on the alpha atom; they are highly reactive, mostly through the oxygen atom and especially towards phosphorus atoms; due to their reactivity they can act as unspecific binders: and induce random mutations in DNA and RNA; several hydroxylamine derivatives bind to oxyhemoglobin leading to the formation of hematotoxic radical	PMID:23061697; PMID:21809939; PMID:15935536; PMID:15634026; PMID:12565011; PMID:16984161; PMID:10087986; PMID:9821018;	wiki:Hydroxylamine; ToxAlerts:TA341; ToxAlerts:TA398; ToxAlerts:TA356; ToxAlerts:TA404;	hydroxylamine; alkylating; genotox;
<b>NO XD</b>	N-oxides	[O-][N+]1=CC=CC=C1	1	aromatic N-oxides are both more electrophilic and nucleophilic compared to pyridine; they can undergo both electrophilic or nucleophilic attack at the oxygen atom and at the 2- or 4-position; some already approved or under development drugs contain the pyridine n-oxide moiety; some	PMID:18621573; PMID:3277047;	wiki:Amine_oxide; ToxAlerts:TA383; ToxAlerts:TA429;	genotox; pyridine n-oxide; n-oxide;

				pyridine n-oxide derivatives were shown to be genotoxic;			
<b>AM MO</b>	Ammonium alkyl quaternary salts	[N+](C)(C)(C)C	1	alkyl quaternary ammonium salts are reported being skin sensitizers and to provoke acute aquatic toxicity; some of them cause toxicity due to DNA intercalation, as ethidium bromide; ammonium salts with reactive substituent, like epoxide, have been reported to be genotoxic; moreover the positive charge makes more difficult to cross the intestinal wall in case of the development of oral drugs;	PMID:20693071; PMID:2269228; PMID:16084005; PMID:23061697; PMID:5859041;	wiki:Ammonium_salt; ToxAlerts:TA259; ToxAlerts:TA642;	ammonium quaternary salt; ammonium salt; quaternary salt;
<b>TH IO</b>	thiol	[C;Y1][S;H]	3	Thiols can easily react with solvent exposed cysteine residues on protein surface and with reducing agent DTT often used in biochemical assay media; moreover thiols may chelate metal ions during in-vitro assays and interact with endogenous metalloproteinases or others metallo-dependent enzymes; thiol derivatives have been related to skin sensitization reaction and production of reactive metabolites leading to idiosyncratic toxicity	PMID:16711725; PMID:8144710; PMID:21809939; PMID:12565011;	ToxAlerts:TA300; ToxAlerts:TA794; ToxAlerts:TA838; ToxAlerts:TA1724;	thiol; CYS
<b>SU LF</b>	Sulfonium salts	[S+](C)(C)C	1	sulfonium salts have been reported to cause acute aquatic toxicity and to cause assay interference, probably due to protein alkylation;	PMID:2269228; PMID:23061697; PMID:16711725; PMID:21401043;	wiki:Sulfonium_salt; ;ToxAlerts:TA643; ToxAlerts:TA1817;	
<b>AH LC</b>	a-halocarbo	C([C;R0][Cl,Br,I])(=O)C	3	In a-halocarbonyl the reactivity of the sp <sup>3</sup> carbon bound to the halogen atom	PMID:21809939; PMID:12565011;	Toxalerts:TA273; Toxalerts:TA416;	a-halocarb

	nyl			is increased by the carbonyl moiety in alpha position. The sp <sup>3</sup> carbon atom undergo a nucleophilic attack, forming a new covalent bond with the endogenous nucleophilic group through a SN <sub>2</sub> reaction; alpha-halocarbonyls have been related to genotoxicity and skin sensitization reaction; alpha-(poly)halogenated carbonyls have been shown to possibly cause inter	PMID:16711725; PMID:15935536; PMID:15656773;	Toxalerts:TA779; wiki:Haloketone; ToxAlerts:TA1746;	onyl; a-haloketone; alpha-halocarbonyl;
<b>BH</b> <b>LA</b>	b-haloamine s; haloethylamine, N-mustard	[C;Y1](!@[C;H2][Cl,Br,I])!@N	3	beta-haloamine also called N-mustard are non-specific DNA alkylating agents, that have been used as anticancer agents; they act undergoing first an intramolecular cyclization that forms an aziridinium intermediate, and then reacting with nucleophilic centre on guanine base in DNA strands; they can also react with cysteine (and lysine) residues through an SN <sub>2</sub> reaction to form a covalent adducts; some haloethylamine derivatives have	PMID:18621573; PMID:21809939; PMID:15634026; PMID:2269228;	ToxAlerts:TA414; ToxAlerts:TA624; ToxAlerts:TA344; ToxAlerts:TA362; ToxAlerts:TA435; ToxAlerts:TA687; ToxAlerts:TA810; wiki:Nitrogen_mustard;	haloamines; haloethylamine; N-mustard, genotoxin (N-mustard)
<b>BH</b> <b>LS</b>	b-halosulfide s, S-mustard	C(!@C[Cl,Br,I])!@S	2	beta-halosulfides or S-mustard are highly reactive and cytotoxic agents; they undergo an intramolecular cyclization, forming an episulfonium intermediate that with guanidine in DNA strand to form a covalent adduct; the damage in DNA strand can lead to cell death as well as to uncontrolled cell replication, leading to cancer	PMID:21809939; PMID:18621573; PMID:15634026; PMID:2269228;	ToxAlerts:TA344; ToxAlerts:TA362; ToxAlerts:TA435; ToxAlerts:TA623; ToxAlerts:TA688; ToxAlerts:TA810; wiki:Sulfur_mustard;	genotoxin (S-mustard)

				development; beta-halosulfide derivatives are strongly lipophilic and can be easily absorbed by skin; they				
<b>AC</b> <b>RL</b>	acrylates; alpha-beta unsaturated carboxylic esters	<chem>C(=O)C=C([H])O</chem>	2	carboxylic acid and esters alpha-beta unsaturated are considered Michael acceptors because they can undergo Micheal addition reaction after the attack of a nucleophile upon the electron deficient beta-carbon; di-substitution at the beta-carbon may prevent Michael addition due to sterical hindrance; acrylate derivatives have been reported to cause acute aquatic toxicity, skin sensitization reaction; given their Michael acceptor nature	PMID:2269228; PMID:21809939; PMID:16711725; PMID:20693071; PMID:18621573; PMID:15935536;	Toxalerts:TA240; Toxalerts:TA630; Toxalerts:TA631; ToxAlerts:TA292; ToxAlerts:TA723; Toxalerts:TA467; Toxalerts:TA468; Toxalerts:TA1815; Toalerts:TA1821; wiki:Acrylate; ToxAlerts:TA367; ToxAlerts:TA418	acrylate; acrylate ester; alpha- beta unsatura ted carbonyl ;	CYS; NH2;
<b>AC</b> <b>RA</b>	acrylamide s; alpha- beta unsaturated amide	<chem>C(=O)C=C([H])N</chem>	3	acrylamide esters alpha-beta unsaturated are considered Michael acceptors because they can undergo Micheal addition reaction after the attack of a nucleophile upon the electron deficient beta-carbon; di-substitution at the beta-carbon may prevent Michael addition due to sterical hindrance; acrylate derivatives have been reported to cause acute aquatic toxicity, skin sensitization reaction; given their Michael acceptor nature they	PMID:21809939; PMID:20693071; PMID:18621573; PMID:15935536; PMID:2269228;	wiki:Acrylamide; ToxAlerts:TA367; ToxAlerts:TA240; ToxAlerts:TA625; ToxAlerts:TA242; ToxAlerts:TA292; ToxAlerts:TA724; ToxAlerts:625; ToxAlerts:TA367; ToxAlerts:TA418;	acrylami de; alpha- beta unsatura ted carbonyl ;	CYS; NH2;
<b>VI</b> <b>NK</b>	vinylketon es; alpha- beta	<chem>C(=O)C=C([H])C</chem>	3	alpha-beta unsaturated carbonyl derivatives are Micheal acceptors, hence are prone to Micheal addition by	PMID:21809939; PMID:12565011; PMID:20693071;	ToxAlerts:TA240; ToxAlerts:TA242; ToxAlerts:TA292;	vinylket one; alpha-	CYS; NH2;

	unsaturated ketone			reacting with endogenous nucleophiles, forming a covalent adduct after attacking the electron deficient beta carbon; like all other Michael acceptors, double substitution on the beta carbon atom can prevent or even avoid the reaction; vinyl ketone derivatives were found to be skin sensitizers and irritants, especially for the	PMID:18853302; PMID:15935536; PMID:2269228;	ToxAlerts:TA367; ToxAlerts:TA453; ToxAlerts:TA722; ToxAlerts:TA418; ToxAlerts:TA632; wiki:Methyl_vinyl_ketone;	beta unsaturated ketone;
<b>VI NS</b>	vinylsulfones; alpha-beta unsaturated sulfone	<chem>S(=O)(C!@=C([H]))(=O)[H]</chem>	3	vinylsulfones are alkenes presenting an electron withdrawing group that makes the carbon atom in beta position with respect to the substituent very electrophilic, and subsequently reacting with endogenous nucleophiles on protein residues, causing skin sensitization reactions and acute aquatic toxicity; as well as for other Michael acceptors, the Michael addition reaction can be prevented by double bulky substituents on the beta carbon	PMID:21809939; PMID:23061697; PMID:2269228; PMID:18853302;	ToxAlerts:TA629; ToxAlerts:TA728;	vinylsulfone; alpha-beta unsaturated sulfone; alpha-beta unsaturated sulfone; CYS; NH2;
<b>NT RV</b>	nitrovinyl; alpha-beta unsaturated nitro compounds	<chem>N(=O)(C!@=C([H]))=O</chem>	3	compounds presenting a nitrovinyl moiety are Michael's acceptor, that react with biological nucleophiles forming covalent adducts; they have been shown to give positive results in acute aquatic toxicity tests; di-substitution at the beta-carbon can sterically hinder the Michael addition; formation of reactive metabolites has	PMID:23061697; PMID:21809939; PMID:2269228; PMID:16922655;	ToxAlerts:TA628; ToxAlerts:TA725;	nitrovinyl; alpha-beta unsaturated nitro compounds;

				also been pointed out for compounds carrying this functional group;				
<b>AC RN</b>	acrylonitriles; vinyl ciano compound s;	<chem>C(#N)C!@=C</chem>	3	acrylonitriles are highly reactive Michael's acceptor that can undergo Michael addition due to the attack of a biological nucleophile upon the electron deficient beta-carbon; acrylonitrile derivatives are reported to be skin sensitizers and to induce acute aquatic toxicity; acrylonitrile itself is listed in IARC classification among compounds possibly carcinogenic for humans;	PMID:21809939; PMID:23061697; PMID:18853302; PMID:2269228; PMID:16922655;	wiki:Acrylonitrile; ToxAlerts:TA726; ToxAlerts:TA743; ToxAlerts:TA726;	acrylonitrile; vinyl ciano compound;	CYS; NH2;
<b>PD A1</b>	pyrimidine acrylates	<chem>N(=C(N=CC1)C!@=C([H]))C=1</chem>	3	pyrimidine acrylate derivatives can behave as Michael's acceptor, hence reacting with biological nucleophiles upon the electron deficient beta-carbon to form a covalent adduct; as all Micheal's acceptors di-substitution at the beta-carbon can sterically hinders the Michael addition reaction; pyrimidine acrylate derivatives have been reported to be skin sensitizer agents;	PMID:21809939; PMID:18853302;	ToxAlerts:TA733; ToxAlerts:TA467; ToxAlerts:TA468;	pyrimidine acrylate;	CYS; NH2;
<b>PD A2</b>	pyrimidine acrylates	<chem>N(=CN=C(C1)C!@=C([H]))C=1</chem>	3	pyrimidine acrylate derivatives can behave as Michael's acceptor, hence reacting with biological nucleophiles upon the electron deficient beta-carbon to form a covalent adduct; as all Micheal's acceptors di-substitution at the beta-carbon can sterically hinders the Michael addition reaction;	PMID:21809939; PMID:18853302;	ToxAlerts:TA733; ToxAlerts:TA467; ToxAlerts:TA468;	pyrimidine acrylate;	CYS; NH2;



				pyrimidine acrylate derivatives have been reported to be skin sensitizer agents;				
<b>PD A3</b>	pyrimidine acrylates	<chem>N(=CN=CC1C!@=C([H]))C=1</chem>	3	pyrimidine acrylate derivatives can behave as Michael's acceptor, hence reacting with biological nucleophiles upon the electron deficient beta-carbon to form a covalent adduct; as all Micheal's acceptors di-substitution at the beta-carbon can sterically hinders the Michael addition reaction; pyrimidine acrylate derivatives have been reported to be skin sensitizer agents;	PMID:21809939; PMID:18853302;	ToxAlerts:TA733; ToxAlerts:TA467; ToxAlerts:TA468;	pyrimidine acrylate;	CYS; NH2;
<b>PY RR</b>	pyrrole-2,5-dione	<chem>C(C(NC1=O)=O)=C1</chem>	2	pyrrole-2,5-dione also called acid imide, is an alkylating agent that acts via Micheal type addition reactions with biological nucleophiles; acid imide derivatives have been reported to cause skin sensitization reaction and they also gave positives results in acute aquatic toxicity tests; as being cyclic alpha-beta insaturated carbonyls, they have been also reported among genotoxic carcinogen compounds and among promiscuous deirvati	PMID:15656773; PMID:21809939; PMID:8144710; PMID:15935536; PMID:18853302; PMID:2269228; PMID:16711725;	ToxAlerts:TA272; ToxAlerts:TA761; ToxAlerts:TA292; ToxAlerts:TA418; ToxAlerts:TA467; ToxAlerts:TA625; ToxAlerts:TA632; ToxAlerts:TA1855;	pyrrole-2,5-dione; acid imide;	CYS; NH2;
<b>KE TE</b>	Ketenes	<chem>CC=C=O</chem>	3	Ketenes are reactive electrophilic derivatives that can react with protein nucleophiles to form acylated adducts; this potentially reactive specie has been shown to induce acute aquatic toxicity;	PMID:23061697; PMID:21809939;	wiki:Ketene; ToxAlerts:TA605; ToxAlerts:TA676; ToxAlerts:TA720;	ketene;	CYS; NH2;

<b>OP HN</b>	o,p-dinitrohaloarenes ; o,p-dinitrohalobenzen	<chem>c1([Cl,Br,F,I])c([N]([O])=O)cc([N]([O])=O)cc1</chem>	3	o,p-dinitrohalobenzenes as well as o,o,p-trinitrohalobenzenes have been reported to be able to undergo a SNAr reaction, forming protein adducts; they have also been reported to be skin sensitizer agents, possibly mutagenic and causing aquatic toxicity;	PMID:21809939; PMID:20693071; PMID:18853302; PMID:16084005; PMID:2269228; PMID:15656773;	ToxAlerts:TA236; ToxAlerts:TA443; ToxAlerts:TA639; ToxAlerts:TA664; ToxAlerts:TA265;	o,p-dinitrohaloarenes; o,p-dinitrohalobenzenes	CYS; NH2; G;
<b>OO HN</b>	o,o-dinitrohaloarenes ; o,o-dinitrohalobenzenes	<chem>c1c(N([O])=O)c([Cl,Br,F,I])c(N([O])=O)cc1</chem>	3	o,o-dinitrohalobenzenes as well as o,o,p-trinitrohalobenzenes have been reported to be able to undergo a SNAr reaction, forming protein adducts; they have also been reported to be skin sensitizer agents, possibly mutagenic and causing aquatic toxicity;	PMID:21809939; PMID:20693071; PMID:18853302; PMID:16084005; PMID:2269228; PMID:15656773;	ToxAlerts:TA236; ToxAlerts:TA443; ToxAlerts:TA639; ToxAlerts:TA664; ToxAlerts:TA265;	o,p-dinitrohaloarenes; o,p-dinitrohalobenzenes	CYS; NH2; G;
<b>AH P1</b>	activated 2,5-halopyridine	<chem>n1c([Cl,F,Br,I],[N+](=O)O])ccc([Cl,F,Br,I],[N+](=O)O])c1</chem>	3	activated 2,5-halo-nitropyridine have been reported to be able to undergo a SNAr reaction, forming protein adducts; they have also been reported to be skin sensitizer agents;	PMID:21809939; PMID:18853302;	ToxAlerts:TA444; ToxAlerts:TA818;	2,5-halopyridine	CYS; NH2;
<b>AP H2</b>	activated 2,3-halopyridine	<chem>n1c([Cl,F,Br,I],[N+](=O)O])c([Cl,F,Br,I],[N+](=O)O])ccc1</chem>	3	activated 2,3-halo-nitropyridine have been reported to be able to undergo a SNAr reaction, forming protein adducts; they have also been reported to be skin sensitizer agents;	PMID:21809939; PMID:18853302;	ToxAlerts:TA444; ToxAlerts:TA818;	2,3-halopyridine;	CYS; NH2;
<b>BZ HL</b>	Benzylhalogenides; alpha-halobenzyls ;	<chem>C(=CC=C(C1)[C]^3[Cl,Br,I])C=1</chem>	3	alpha-halobenzyls present an reactive sp3 carbon atom that can undergo a SN2 reaction with biological nucleophiles; benzylic halides have shown in experimental assays genotoxicity, acute aquatic toxicity and	PMID:21809939; PMID:2269228; PMID:15935536; PMID:3277047;	ToxAlerts:TA408; ToxAlerts:TA638; ToxAlerts:TA783; ToxAlerts:TA439; ToxAlerts:TA700;	Benzylhalogenide; alpha-halobenzyl; benzylic	CYS; NH2;

<b>H ME O</b>	halomethyl ethers; alpha-haloethers	[C;^3]O!@C([F,Cl,Br,I])	3	skin sensitization reactions; alpha-haloethers present a reactive sp3 carbon atom that can undergo a SN2 reaction with biological nucleophiles; positive results have been reported for acute aquatic toxicity tests with halomethyl ether derivatives and the possibility of idiosyncratic toxic reaction has also been pointed out;	PMID:21809939; PMID:2269228; PMID:16922655;	ToxAlerts:TA622; ToxAlerts:TA784; ToxAlerts:TA856;	halide; alpha-haloether; halomethyl ethers;	CYS; NH2;
<b>TH FU</b>	Thiophene, Furan	[s,o]1[c;R1][c;R1][c;R1]c1	2	thiophene and furan heterocycles have been reported to possibly induce idiosyncratic adverse drug reaction; some thiophene derivatives are reported to act as interfering compounds in vitro biochemical tests;	PMID:16922655; PMID:21702456; PMID:17302443; PMID:21455238;	ToxAlerts:TA847; ToxAlerts:TA1955; ToxAlerts:TA2216; ToxAlerts:TA846;	thiophene, furan	
<b>4H PM</b>	4-halopyrimidine; 4-nitropyrimidine	c1([F,Cl,Br,I,\$(N(=O)O)])ncnc1	1	4-halopyrimidine and 4-nitropyrimidine are potential electrophilic agents that can react with biological nucleophiles; they have also been reported to be skin sensitizers and to induce acute aquatic toxicity;	PMID:21809939; PMID:18853302; PMID:2269228;	ToxAlerts:TA446; ToxAlerts:636;	4-halopyrimidine; 4-nitropyrimidine;	
<b>2H PM</b>	2-halopyrimidine; 2-nitropyrimidine	n1c([F,Cl,Br,I,\$(N(=O)O)])ncnc1	1	2-halopyrimidine and 2-nitropyrimidine are potential electrophilic agents that can react with biological nucleophiles; they have also been reported to be skin sensitizers and to induce acute aquatic toxicity;	PMID:21809939; PMID:18853302; PMID:12565011; PMID:2269228;	ToxAlerts:TA445; ToxAlerts:636;	2-halopyrimidine; 2-nitropyrimidine;	
<b>CY AN</b>	cyanohydrines; alpha-hydroxy	CC([O;H])C#N	1	cyanohydrines have been reported as agents causing aquatic toxicity and false positive results in HTS assays.	PMID:23061697; PMID:16711725;	ToxAlerts:TA761; wiki:Cyanohydrin; ToxAlerts:TA1750;		

		nitriles					
<b>LT</b>	Linear	N!@C(=S)!@N	1	Thioureas were reported to induce the formation of reactive metabolites and causing idiosyncratic toxicity; they have been also listed among derivatives able to induce non-genotoxic cancer in experimental assays; thioureas with electron-rich substituents have also been reported to possibly cause HTS assay interference;	PMID:16922655; PMID:18621573; PMID:20131845;	ToxAlerts:TA836; wiki:Thiourea;	thioureas;
<b>HU</b>	thioureas						
<b>LD</b>	Linear	N!@C(=S)!@SC	1	Dithiocarbamates are possible acylating agent and skin sensitizers (RF 4); some thiocarbamate derivatives are listed as genotoxic compounds and aquatic toxicants	PMID:23061697; PMID:8144710; PMID:18621573;	wiki:Dithiocarbamates; ToxAlerts:TA317; ToxAlerts:TA373; ToxAlerts:TA690;	dithiocarbamate; NH <sub>2</sub> ; CYS;
<b>TC</b>	dithiocarbamates;						
<b>PH</b>	polyhalogenated compound	*([Cl,Br,I])([Cl,Br,I])([Cl,Br,I])	1	compounds containing too many halogens, the general idea is more than 3 or 4 halogen atom per molecule, have been pointed out to be involved in biochemical assay interference, in particular aromatic derivatives; but also in promoting toxicity: cycloalkanes with three or more halogen atoms have been related to non genotoxic carcinogenicity,	PMID:18621573; PMID:23061697; PMID:16711725;	wiki:Polyhalogenated_compound; ToxAlerts:TA1779; ToxAlerts:1787; ToxAlerts:1852;	polyhalogenated;
<b>C3</b>							
<b>DT</b>	Dithioester	CC(=S)SC	1	dithioester derivatives have been reported as skin sensitizers and as promiscuous agents in HTS assays, probably due to hydrolysis of the thioester bond releasing a thiol derivative that can interact with sulfur	PMID:23061697; PMID:20693071; PMID:21809939;	wiki:Thioester; ToxAlerts:TA232; ToxAlerts:TA289; ToxAlerts:TA1828;	thioester; ; dithioester;
<b>HE</b>							

				containing biological nucleophiles or sulfur containing nucleophiles in the assay media;				
<b>OQ UI</b>	o- Quinones	<chem>C(C=CC(C1=O)=O)=C1</chem>	3	o-quinones derivatives may induce toxic reactions in two different ways: they can undergo a Micheal addition reaction: a biological nucleophile may attack the beta C respect to one carbonilic group, causing the loss of an hydrogen atom and the formation of a covalent adduct; but quinones can also give induce unwanted responses by intercalating into double stranded DNA; independently of the mechanism, skin sensitization, idiosyncra	PMID:23061697; PMID:16922655; PMID:21809939; PMID:18621573; PMID:16922655; PMID:20693071; PMID:15656773; PMID:20131845; PMID:8144710; PMID:16711725;	ToxAlerts:TA237; ToxAlerts:TA291; ToxAlerts:TA369; ToxAlerts:TA752; ToxAlerts:TA276; ToxAlerts:TA2286; ToxAlerts:TA841; ToxAlerts:TA1781; ToxAlerts:TA2286; wiki:Quinone;	ortho- quinone;	
<b>PQ UI</b>	p- Quinones	<chem>C(C(C=CC1=O)=O)=C1</chem>	3	o-quinones derivatives may induce toxic reactions in two different ways: they can undergo a Micheal addition reaction: a biological nucleophile may attack the beta C respect to one carbonilic group, causing the loss of an hydrogen atom and the formation of a covalent adduct; but quinones can also give induce unwanted responses by intercalating into double stranded DNA; independently of the mechanism, skin sensitization, idiosyncra	PMID:15656773; PMID:16922655; PMID:18621573; PMID:16922655; PMID:20131845; PMID:16711725; PMID:20693071; PMID:8144710; PMID:2269228;	ToxAlerts:TA237; ToxAlerts:TA291; ToxAlerts:TA369; ToxAlerts:TA752; wiki:Quinone; ToxAlerts:TA2286; ToxAlerts:TA1906; ToxAlerts:TA1869; ToxAlerts:TA1825; ToxAlerts:TA633; ToxAlerts:TA841;	para- quinone;	
<b>AL AC</b>	allyl acetate	<chem>CC(=O)O!@C!@C=[C;H2]</chem>	1	allyl acetate group presents a reactive sp3 carbon atom that can undergo a SN2 reaction with biological nucleophiles (rf 7); allyl acetate and their thio analogs have been listed	PMID:21809939; PMID: 16084005; PMID:18853302;	ToxAlerts:TA462; ToxAlerts:TA778;	allyl acetate; alkylatin g agent;	CYS; NH2;

				among skin sensitizers agents; genotoxicity has been reported for allyl acetate derivatives, however it is still not well-established;				
<b>LA CT</b>	lactones; thiolactones; lactams	C1C(=O)[O,S,N]C1	1	lactones and lactams and their thio-analogs are four terms strained ring systems, that can undergo a ring opening acylation reaction involving nucleophilic attack at the carbonyl group; beta propiolactone, lactams and thio derivatives have been reported to be genotoxic, to induce skin sensitization in experimental assays, and to induce acute aquatic toxicity; beta-lactams are a common functional group in antibiotic drugs but it is	PMID:21809939; PMID:15634026; PMID:18621573; PMID:15935536; PMID:18853302; PMID:2269228; PMID:16711725;	ToxAlerts:TA352; ToxAlerts:TA363; ToxAlerts:TA410; ToxAlerts:TA437; ToxAlerts:TA461; ToxAlerts:TA582; ToxAlerts:TA612; ToxAlerts:TA673; ToxAlerts:TA712; ToxAlerts:TA788; ToxAlerts:TA1844; wiki:Lactone;	lactone; thiolactone ; lactam	CYS; NH2;
<b>CY PP</b>	Cyclopropenone	C1=CC1(=O)	1	Cyclopropenones are three terms strained ring system, that can undergo a ring opening acylation reaction involving nucleophilic attack at the carbonyl group. However, due to their instability they are not common as functional group in compounds collections;	PMID:21809939;	ToxAlerts:TA710		CYS; NH2;
<b>AZ LC</b>	Az lactone	C1C(=O)OC=N1	1	Az lactones are five-members strained ring system, that can undergo a ring opening acylation reaction involving nucleophilic attack at the carbonyl group. They are not common functional groups due to the instability of the ring that tends to open quite easily;	PMID:21809939;	ToxAlerts:TA714		CYS; NH2

<b>HO QU</b>	ortho- hydroquin one; catechol	<chem>c1c(O[H])c(O[H])c cc1</chem>	2	orthohydroquinones or catechols are orthoquinone precursors that can be activated into quinones by epatic metabolism. In particular after bioactivation they can undergo Michael addition reactions; ortho-hydroquinones are classified as skin sensitizers and inducers of idiosyncratic reaction due to reactive metabolites generation; due to their reactivity also false positive results un HTS assays has been highlighted;	PMID:21809939; PMID:20693071; PMID:15656773; PMID:16922655; PMID:8144710; PMID:20131845;	ToxAlerts:TA277; ToxAlerts:TA278; ToxAlerts:TA755; ToxAlerts:TA840; ToxAlerts:TA239; ToxAlerts:TA294; ToxAlerts:TA298; wiki:Catechol; ToxAlerts:TA1918; ToxAlerts:TA2375;	o- hydroqui none; catechol;	CYS; NH2;
<b>HP QU</b>	para- hydroquin one	<chem>c1c(O[H])ccc(O[H]) c1</chem>	2	parahydroquinones are p-quinone precursors that can be activated into quinone by epatic metabolism. AS all quinones, they can undergo Micheal addition reaction; p-hydroquinones are reported to be skin sensitizers and inducers of idiosyncratic reaction due to reactive metabolites generation; due to their reactivity also false positive results un HTS assays has been highlighted;	PMID:21809939; PMID:20693071; PMID:15656773; PMID:8144710; PMID:16922655;	ToxAlerts:TA238; ToxAlerts:TA277; ToxAlerts:TA278; ToxAlerts:TA293; ToxAlerts:TA296; ToxAlerts:TA755; ToxAlerts:TA840; ToxAlerts:TA1791; ToxAlerts:TA1854	p- hydroqui none;	CYS; NH2;
<b>OQ IM</b>	ortho- quinone imine; ortho- quinone methide	<chem>C1=CC(=O)C(=[N, C])C=C1</chem>	2	ortho-quinone imine and ortho-quinone methide can be metabolized into o-quinones, inducing the same type of toxicity as o-quinones;	PMID:21809939; PMID:16922655;	ToxAlerts:TA753; ToxAlerts:TA843; ToxAlerts:TA758;		CYS; NH2;
<b>PQ IM</b>	para- quinone imine;	<chem>C1=CC(=O)C=CC1( =[N,C])</chem>	2	para-quinone imine and para-quinone methide can be metabolized into p-quinones, and so inducing the same	PMID:21809939; PMID:16922655;	ToxAlerts:TA753; ToxAlerts:TA843; ToxAlerts:TA758;		CYS; NH2;

				type of toxicity;				
<b>OA</b>	para-quinone methide							
<b>PH</b>	ortho-aminophenol	<chem>c1c(O[H])c([N;H2])ccc1</chem>	2	ortho-aminophenol can be metabolized into o-quinones by hepatic metabolism. It has also been classified as a skin sensitizer agent. Several ortho-aminophenol derivatives are reported to be genotoxic;	PMID:15634026; PMID:21809939; PMID:18853302;	ToxAlerts:TA589; ToxAlerts:TA756;	ortho-aminophenol; o-aminophenol;	CYS; NH2;
<b>OQ</b>	ortho-quinone diimine	<chem>C1=CC(=N)C(=[N,C])C=C1</chem>	2	ortho-quinone diimine can be metabolized into o-quinones, determining the same type of Michael addition reactions;	PMID:21809939; MID:16711725;	ToxAlerts:TA754; ToxAlerts:TA1781;		CYS; NH2;
<b>PQ</b>	para-quinone diimine	<chem>C1=CC(=N!@C)C=CC1(=[N,C;R0])</chem>	2	para-quinone diimine can be metabolized into p-quinones and so giving the same Michael addition reaction, especially in experimental assays;	PMID:21809939; PMID:16711725;	ToxAlerts:TA754; ToxAlerts:TA1788; ToxAlerts:TA1886;	para-quinone diimine	CYS; NH2;
<b>PY</b>	pyranone	<chem>C1=COC=CC1(=O)</chem>	1	pyranones can react with nucleophiles and give Michael addition reaction	PMID:21809939;	ToxAlerts:TA759; wiki:Pyrone;	pyranone	
<b>PO</b>	polyaromatic hydrocarbons	<chem>c1c2cccc2cc3cccc13</chem>	2	linear (not fused ring with indentation) polyaromatic hydrocarbons can be metabolically activated by CYP450 into quinones and after that undergo Michael addition reactions, with DNA or proteins, or they can intercalate as they are into double strand DNA; for this reason several studies report these compounds as genotoxic; accounts for skin sensitization reactions and acute aquatic toxicity	PMID:21809939; PMID:18621573; PMID:15935536; PMID:15634026; PMID:18853302;	ToxAlerts:TA375; ToxAlerts:TA422; ToxAlerts:TA760; ToxAlerts:TA465; ToxAlerts:TA656; ToxAlerts:TA328;	polyaromatic; hydrocarbon;	CYS; NH2;



				(due to excessive lipophilicity) have als				
<b>IT AZ</b>	isothiazol-3-one	C1=CSNC1(=O)	2	isothiazol-3-ones have been shown to react with sulfur atom of cysteine residues, undergoing a SN2 reaction that finally produce the formation of the protein adduct after ring opening; some studies classify them as skin sensitizers.	PMID:21809939; PMID:15656773; PMID:18853302;	ToxAlerts:TA269; ToxAlerts:TA575; ToxAlerts:TA801; ToxAlerts:TA802;	isothiazol-3-one;	CYS;
<b>HL AK</b>	polarized haloalkene	C([H,C])([C;R0,H,F,Cl,Br,I])=!@C([Br,Cl,F,I])([C;R0,H,F,Cl,Br,I])	2	Alkene with halogen substituents are susceptible to a SN2 reaction at the sp2 carbon atom. Monohaloalkenes (as the SMARTS definition) have been reported to cause genotoxic effects whereas geminal dihaloalkenes have been shown to cause skin sensitization	PMID:21809939; PMID:18621573; PMID:18853302;	ToxAlerts:TA805; ToxAlerts:TA361; ToxAlerts:TA434; ToxAlerts:TA472;	haloalkene;	CYS; NH2;
<b>IN CI</b>	1,3-Bis(hydroxymethyl)-5,5-dimethylimidazolidine-2,4-dione; DMDM hydantoin	C1(=O)N(CO[H])C(=O)N(CO[H])C1([C;^3])([C;^3])	1	DMDM hydantoin is a preservative used in skin formulations, able to release formaldehyde, a known skin sensitizer, that react with biological NH2 groups through a Schiff base formation reaction, resulting in cross-link between protein chains; some studies suggested it is able to induce skin allergies; particular attention should be paid in the development of derivatives of this scaffold, especially in case of drug for local administration	PMID:21809939; PMID:20693071; PMID:8144710;	ToxAlerts:TA820;	DMDM hydantoin;	NH2;
<b>CR W 1</b>	crown ether	C10[C;H2]O[C;H2][C;H2]OC1	1	crown ethers can chelate metal ions present in human body but also with ions present in the media during experimental assays;	PMID:23061697;	wiki:Crown_ether; ToxAlerts:TA1758;	crown ether;	

<b>CRW2</b>	crown ether	C1N[C;H2][C;H2]N[C;H2][C;H2]N[C;H2][C;H2]NC1	1	crown ethers can chelate metal ions present in human body but also with ions present in the media during experimental assays;	PMID:23061697;	wiki:Crown_ether; ToxAlerts:TA1758;	crown ether;
<b>CRW3</b>	crown ether	C1N[C;H2][C;H2][C;H2]N[C;H2][C;H2]N[C;H2]C1	1	crown ethers can chelate metal ions present in human body but also with ions present in the media during experimental assays;	PMID:23061697;	wiki:Crown_ether; ToxAlerts:TA1758	crown ether;
<b>PHC1</b>	polyhalogenated compound	*([Cl,Br,I])*([Cl,Br,I])([Cl,Br,I])	1	compounds containing too many halogens, the general idea is more than 3 or 4 halogen atom per molecule, have been pointed out to be involved in biochemical assay interference, in particular aromatic derivatives; but also in promoting toxicity: cycloalkanes with three or more halogen atoms have been related to non genotoxic carcinogenicity,	PMID:18621573; PMID:23061697; PMID:16711725;	wiki:Polyhalogenated_compound; ToxAlerts:TA1779; ToxAlerts:1787; ToxAlerts:1852;	polyhalogenated;
<b>PHC2</b>	polyhalogenated compound	*([Cl,I,Br])*([Cl,Br,I])*([Cl,Br,I])([Cl,Br,I],n)	1	aromatic compounds containing too many halogens, the well-established idea is more than 3 or 4 halogen atom per molecule, have been pointed out to be involved in biochemical assay interference ;	PMID:18621573; PMID:23061697; PMID:16711725;	wiki:Polyhalogenated_compound; ToxAlerts:TA377; ToxAlerts:1852;	polyhalogenated;
<b>PLFC</b>	activated polyfluorinated compound	*(F)*(F)*(F)*([N+](=O)(O)),([N](=O)=O),\$(C(=O)),\$(C(=O)[CH3]),\$(C#N),\$(OC(=O)))	1	compounds containing five or more fluorine atoms;	PMID:21809939; PMID:16711725; PMID:23061697;	ToxAlerts:TA816; ToxAlerts:TA817;	polyfluorinated
<b>RHD1</b>	rhodanine	C1(=O)NC(=S)SC1(=C)	1	Rhodanines and rhodanine-like compounds have been reported to be a recurrent scaffold coming out from	PMID:23061697; PMID:20131845; PMID:16711725;	ToxAlerts:TA1774; ToxAlerts:TA1910; ToxAlerts:TA2139;	rhodanine;

				HTS assay campaigns, suggesting that these derivatives are probably false positives and/or unspecific binders; rhodanines can interfere with photometric detection methods where light absorption is in the range of 570-620 nm and some derivatives are also highly colored, interfering with colorimetric detection method; there are also		wiki:Rhodanine;
<b>RH D2</b>	rhodanine-like	<chem>C1(=O)N=C(N[H])SC1(=C)</chem>	1	Possible equilibrium form for rhodanine-like compounds presenting a double bond with nitrogen instead of sulfur; rhodanine-like compounds have been reported to be a recurrent scaffold coming out from HTS assay campaigns, suggesting that these derivatives are probably false positives and/or unspecific binders; rhodanines can interfere with photometric detection methods where light absorption is in the range of 570-620 nm and some d	PMID:23061697; PMID:20131845; PMID:16711725;	wiki:Rhodanine; ToxAlerts:TA2139; ToxAlerts:TA2140;
<b>AD C1</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=O)[N,O,S]N=C1</chem>	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this kind of compound present an alpha-beta unsaturated carbonyl group, together	PMID:20131845;	ToxAlerts:TA1903;

				with another unsaturated bond in a heterocycle ring, that makes molecule more reactive and the beta carbon on first double bond ca		
<b>AD C2</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=O)NC=S</chem> 1	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this k	PMID:20131845;	ToxAlerts:TA1921;
<b>AD C3</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=[O,N,S])[O,N,S]C=C1</chem>	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this k	PMID:20131845;	ToxAlerts:TA1922;
<b>AD C4</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=O)NNC1(=O)</chem>	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this k	PMID:20131845;	ToxAlerts:TA1923;
<b>AC D5</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=O)[N,O,S]C=N1</chem>	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this k	PMID:20131845;	ToxAlerts:TA1924;

<b>AC D6</b>	alkylidene substituted with 5 membered heterocycle ring	<chem>C1(=C)C(=S)NN=C1</chem>	1	this type of 5 membered heterocycle ring has been reported being an interfering compound in experimental assays; the mechanism is unknown, however it is probably due to the ability of reacting with protein nucleophiles, because this k	PMID:20131845;	ToxAlerts:TA2128;
<b>FT HQ</b>	fused tetrahydro quinoline	<chem>C1Nc2ccccc2C3C=CCC13</chem>	1	they have been reported as interfering compounds in library screening, however the mechanism of interference is unknown;	PMID:20131845; PMID:23061697;	ToxAlerts:TA1946;
<b>AA P1</b>	1,2,3-aralkyl pyrrole	<chem>[c;h]1c(C)n(c2cccc2)c(C)[c;h]1</chem>	1	aralkyl pyrroles have been pointed out as possible interfering in HTS assays, but the mechanism of interference was not clarified;	PMID:20131845;	ToxAlerts:TA1916;
<b>AA P2</b>	1,2,3-aralkyl pyrrole	<chem>[c;h]1c(C)n(C)c(c2ccccc2)[c;h]1</chem>	1	aralkyl pyrroles have been pointed out as possible interfering in HTS assays, but the mechanism of interference was not clarified;	PMID:20131845;	ToxAlerts:TA2269;
<b>BZ FU</b>	benzofuran sulfonamide	<chem>c1ccc(S(=O)(=O)N(c3ccccc3))c2n[o,s]nc12</chem>	1	this type of aromatic compounds are quite reactive and can react with protein nucleophiles and nucleophiles in the assay media, giving false positive results;	PMID:20131845;	
<b>AC T1</b>	2-amino-3-carbonyl-thiophene	<chem>c1([C,N,O,Cl,Br,I,F,S,P])c([C,N,O,Cl,Br,F,S,I,P])sc(N([H])[H])c1(C(=O))</chem>	1	2-amino-3-carbonyl thiophene derivatives have been pointed out as possible assay interfering compounds	PMID:20131845;	ToxAlerts:TA1953;
<b>AC T2</b>	2-amide-3-carbonyl-thiophene	<chem>c1([a,C,H])csc(N(H)C(=O)C)c1(C(=O))</chem>	1	2-amide-3-carbonyl thiophene derivatives have been pointed out as possible assay interfering compounds	PMID:20131845;	ToxAlerts:TA1954;
<b>PR</b>		<chem>c1(c2ccc(Cl)cc2)n</chem>	1	compound highlighted as promiscuous	PMID:12565011;	

<b>01</b>	<chem>c(c3ccc(S(c4ccc(Cl)cc4))cc3)cc(C(=O)O)c1</chem>		in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	
<b>PR 02</b>	<chem>c1(c2cc(Cl)c(Cl)cc2)nc(c3ccc(C(C)(C)C)cc3)cc(C(=O)O)c1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 03</b>	<chem>c1(c2ccc(Cl)cc2)nc(c3ccc(c4cccc4)cc3)cc(C(=O)O)c1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 04</b>	<chem>c1(c2ccc(Cl)cc2)nc(c3ccc(CCC)cc3)cc(C(=O)O)c1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 05</b>	<chem>c1(c2ccc(Cl)cc2)nc(c3ccc(Cl)cc3)cc(C(=O)O(c4ccc(OC(C(=O)O)CC(C)C)cc4))c1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 06</b>	<chem>c1(c3nc(c4ccc(C)c4)cc(C(=O)O)c3)cccc2cccc12</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 07</b>	<chem>C1Cc2cc(OC)ccc2C3CCC4(C)[C@H](NCCN(c5ncc(N(=O)~O)cc5))CCC4C13</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 08</b>	<chem>C1Cc2cc(OC)ccc2C3CCC4(C)[C@H](NCCN(c5ccccc5))CC4C13</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;

<b>PR 09</b>	<chem>[H][C@@]1(CC(O)C)CC[C@@]1(C2[C@@H](C[C@@]34C)O)C(C)CCC2C3CC[C@@H]4NCCN</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 10</b>	<chem>O=C1NC2=CC=CC=C2/C1=C/C3=CC=C(C4=CC=CS4)S3</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 11</b>	<chem>CC(C)(C)C1=C(O)C(C(C)(C)C)=CC(/C=C2C(C=CC=N3)=C3NC\2=O)=C1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 12</b>	<chem>O=C1NC2=C(C=CC=C2)/C1=C\C3=C(C=C(N(C)C)C=C3</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 13</b>	<chem>O=C1NC2=C(C=CC=C2)/C1=C\C3=C(C=C(CCCC4)C4=C3O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 14</b>	<chem>O=C(C1=CC=CS1)NC2=NN=C(S2)SC3=[SH]C4=CC=C(C(C)C=C4N3</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 15</b>	<chem>OC(C=C1)=CC=C1/N=C/C2=CC=CC(Br)=C2</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 16</b>	<chem>O=C1C2=C(C)C=C(Cl)C=C2S/C1=C3C(C(C)C)=CC(Cl)=C</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific	PMID:12565011;

<b>PR 17</b>	<chem>4)=C4S/3)=O BrC1=CC(S(=O)([O-])=O)=C(/N=N/C2=C(S(=O)([O-])=O)C=C(C(S(=O)([O-])=O)C(/N=N/C3=CC=C(Br)C=C3S(=O)([O-])=O)=C4O)C4=C2O)C=C1</chem>	1	noncompetitive enzyme inhibition; compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 18</b>	<chem>OC1=C(/N=N/C2=CC=C(OCC(O)=O)C=C2)C=CC3=C1C=CC(Br)=C3</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 19</b>	<chem>O=S(C(C=C1)=CC=C1OCC2=CC=CC=C2)(C3=CC=C(OC(CCCCC)C(C)=O)C=C3)=O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 20</b>	<chem>C1(C(C=CC=C2)=C2/C3=C/C4CNCCC4)=C3C=CC=C1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:12565011;
<b>PR 21</b>	<chem>C(=C1c2ccccc2C(=O)O1)Nc1ccc(cc1)Oc1ccccc1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 22</b>	<chem>C(c1ccccc1O)=NNC(C(NN=Cc1ccccc1O)=O)=O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific	PMID:11931626;



<b>PR 23</b>	<chem>c1ccc2c(c1)ccc(c2N=Nc1c(cc(c2cccc12)S([O-])(=O)=O)O)O</chem>	1	noncompetitive enzyme inhibition; compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 24</b>	<chem>c1ccc2c(c1)ccc(c2O)N=Nc1c(cc(c2ccc12)S([O-])(=O)=O)O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 25</b>	<chem>C1=[C-](C(C(=CC1=C(c1c c(c(c1)I)[O-])I)c1cccc1C([O-])=O)I)=O)I</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 26</b>	<chem>c1[c-]2C(c3cc(c(c3Oc2c(c1)I)[O-])I)I[O-])I)c1c(C(O)=O)c(c(c1[Cl])[Cl])[Cl][Cl]</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 27</b>	<chem>CC(C)c1ccc(C=C2C(Nc3cccc23)=O)c c1</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 28</b>	<chem>c1cc(c(cc1[N+])([O-])=O)[N+]( [O-])=O)NN1C([C@H]2[C+](C1=O)[C@@]1(C=C([C@]2([C@@]1([Cl])[Cl])[Cl])[Cl])[Cl])[Cl]</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;

	=0			
<b>PR 29</b>	<chem>c1cc(c(cc1c1csc(Nc2ccc(cc2F)F)n1)O)O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 30</b>	<chem>c1ccc2c(c(cc(c2c1)S([O-])(=O)=O)[N-])=Nc1ccc(cc1)c1cc(c(cc1)N=Nc1cc(c2cccc2c1N)S([O-])(=O)=O)N</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;
<b>PR 31</b>	<chem>c1cc(c(cc1C1=C(C(c2c(cc(cc2O1)O)O)=O)O)O)O</chem>	1	compound highlighted as promiscuous in screening assays, due to aggregation and maybe also unspecific noncompetitive enzyme inhibition;	PMID:11931626;