# Activity of diverse chalcones against several targets: Statistical analysis of a high-throughput virtual screen of a custom chalcone library.

A thesis submitted in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

IN CHEMISTRY

of

RHODES UNIVERSITY, SOUTH AFRICA

Department of Chemistry

Faculty of Science

By

**Arthur F. D. Sarron**

December 2018

# ABSTRACT:

Chalcone family molecules are well known to have therapeutic proprieties (anti-inflammatory, anti-microbial or anti-cancer, etc). However the mechanism of action in some cases is not well known. A virtual library of this family of compounds was constructed using custom scripts, based on the aldol condensation, and this library was modified further to analogues by expansion of the $\alpha,\beta$-unsaturated ketone linker. Acetophenone and benzaldehyde derivatives which are available and purchasable were used as a base to design the chalcone virtual library. 8063 chalcones were constructed and geometrically optimized with Gaussian 09. Their physicochemical characteristics linked to the Lipinski rules were analyzed with Knime and CDK. The entire library was after docked against several targets including HIV-1 integrase, MRSA pyruvate kinase, HSP90, COX-1, COX-2, $ALR_2$, MAOA, MAOB, acetylcholinesterase, butyrylcholinesterase and $PLA_2$. With the exception of MAOA, which does not have a crystal structure ligand, all dockings were validated by redocking the original ligand provided by the literature.

These targets are known in the literature to be inhibited by chalcone-derivatives. However, specificity of the particular known chalcone inhibitors to the particular targets is not known. To this end the performance of the generated chalcone library against the list of targets was of interest. The binding energy of ligand-protein complexes was generally good across the library. Statistical analysis including principal component analysis and hierarchical clustering analysis were made in order to investigate for any physical/chemical characteristics which might explain what chalcone features affect the binding energy of the ligand-protein complexes. The spherical polar coordinates defining the orientation of the binding poses were also calculated and used in the statistical analysis. The statistical analysis has allowed us to hypothesize the importance of these radial distances and the polar angles of key atoms in the chalcones in binding to the pyruvate kinase crystal structure. This was validated by the docking of another small library of compound models in which the $\alpha,\beta$-unsaturated ketone chain of the chalcone was replaced by incrementally longer conjugated chains.

Further studies on the chalcones themselves reveal rotameric systems in both *cis* and *trans*-configurations (which may impact binding), and also studied was the effect of Topliss-based modification and its impact of binding to HSP90. Molecular dynamics confirmed good binding of identified chalcone hits.

# ACKNOWLEDGMENTS

# Table of Contents

# Tables of Figures and Tables:

# List of Abbreviations:

| | |
|---|---|
| 4-C-CoAL | 4-coumarate-CoA ligase |
| 5-HT | serotonin |
| A | adenine |
| AA | amino acid |
| AChE | acetylcholinesterase |
| ADCS | aminodeoxychrorismate synthase |
| ADMET | absorption, distribution, metabolism, elimination and toxicity |
| ADP | adenosine diphosphosphate |
| ADT | arogenate dehydratase |
| agnes | agglomerative nesting hierachical clustering |
| AKT/PKB | steroid hormone receptor c-MET and the serine-threonine protein kinase |
| Ala | alanine |
| $ALR_2$ | aldose reductase |
| Arg | arginine |
| AS | anthranilate synthase |
| Asp | aspartic acid |
| ATP | adenosine triphosphate |
| B | asparagine or aspartic acid |
| BChE | butyrylcholinesterase |
| Bcl-2 | B-cell lymphoma 2 |
| BCR-ABL | fusion-genes Breakpoint Cluster Region-Abelson |
| BFGS | Broyden-Fletcher-Goldfarb-Shanno |
| BmimOTs | 1-butyl-3-methyl-1-H-imidazolium 4-methylbenzenesulfate |
| C | cytosine |
| CCD | catalytic core domain |
| CD4 | cluster of differentiation |
| CDC | Center for Disease Control and Prevention |
| CDER | Center for Drug Evaluation and Research |
| CHDD | Chinese Herbal Drug Database |
| CHPC | Center for High Performance Computing |
| CHS | naringenin-chalcone synthase |
| CM | chromate mutase |
| CM | chorismate mutase |
| CNPD | Chinese Natural Product Database |
| CNS | central nervous system |
| CoA | coenzyme A |
| contrib | contribution |
| cos2 | squared cosine |
| COX | cyclooxygenase |
| CS | chorismate synthase |
| $Cs_2CO_3$ | caesium carbonate |
| CTD | C-terminal domain |

| | |
|---|---|
| DA | dopamine |
| DAAs | direct acting viral |
| DAHP | 3-Deoxy-D-arabino-heptulosonate7-phosphate |
| DHQ | 3-dehydroquinic acid |
| DHS | 3-dehydroshikimic acid |
| diana | divisive analysis clustering |
| DIEA | N,N-diisopylethylamine |
| DNA | deoxyribonucleic acid |
| E. coli | Escherichia coli |
| E4P | D-erythrose-4-phosphate |
| EGFR | epidermal growth factor receptor |
| EPSP | 5-enolpyruvylshikimate 3-phosphate |
| EPSP synthase | 3-phosphoshikimate 1-carboxyvinyltransferase |
| FAD | flavin adenine dinucleotide |
| FBDD | fragment based drug design |
| FDA | Food and Drug Administration |
| G | guanine |
| G3P | glyceraldehyde 3-phosphate |
| Gln | glutamine |
| Glu | glutamic acid |
| Gly | glycine |
| GPCR | G protein-coupled receptors |
| H-bond | hydrogen-bond |
| HDF | high density lipoproteins |
| His | histidine |
| HIV | Human Immunodeficiency Virus |
| HOMO | highest occupied molecular orbital |
| HSP90 | heat-shock protein 90 |
| HTVS | high-throughput virtual screening |
| I or Ile | isoleucine |
| ICS | isochrismate |
| IDG | Illuminating the Druggable Genome |
| LDL | low density lipoproteins |
| Leu | leucine |
| logP | partition coefficient or octanol-water partition |
| LPC | lysophosphatidylcholines |
| Lp-PLA$_2$ or PAF-AH | lipoprotein-associated phospholipase A$_2$ |
| LUMO | lowest unoccupied molecular orbital |
| Lys | lysine |
| MAO | monoamine oxidase |
| Met | methionine |
| miRNA | microRNA |
| MlogP | Mannhold partition coefficient |
| MRSA | methicillin-resistant Staphylococcus aureus |

| | |
|---|---|
| MW | microwave |
| NADPH | nicotinamide adenine dinucleotide phosphate |
| NANPDB | Northern African Natural Products Database |
| NE | norepinephrine |
| NF-$\kappa$B | nuclear factors |
| NMR | nuclear magnetic resonance |
| NPT | constant number of particle |
| NSAID | non steroidal anti inflammatory drug |
| NTD | N-terminal domain |
| NVT | constant temperature, volume and number of particles |
| PAF | platelet activating factor |
| PAF-AH | platelet-activating factor acetylhydrolase |
| PAL | phenylalanine ammonia-lyase |
| PCA | principle component analysis |
| Pd(PPh$_3$)$_4$ | tetrakis(triphenylphosphine)palladium(0) |
| PDB | Protein Data Base |
| PdCl$_2$(PPH$_3$)$_2$ | bis(triphenylphosphine)palladium(II) dichloride |
| PDT | prephenate dehydratase |
| PEP | 2-phosphoenolpyruvate |
| PGE$_2$ | prostaglandin E$_2$ |
| PGH$_2$ | prostaglandin H$_2$ |
| Phe | phenylalanine |
| PIC | preintegration complex |
| PK | pyruvate kinase |
| PLA$_2$ | phospholipase A$_2$ |
| PLP | pyridoxal 5'-phosphate |
| PPA-AT | prephenate aminotransferase |
| PPDK | plastidic pyruvate orthophosphate dikinase |
| PPT | PEP/phosphate translocator |
| Pro | proline |
| QM | quantum mechanic |
| QSAR | quantitative structure activity relationships |
| r | radial distance |
| R&D | Research and Development |
| RCSB | Research Collaboratory for Structural Bioinformatics |
| RMSD | root-mean-squared deviation |
| ROC | Receiver Operating Characteristic |
| SANCDB | South African Natural Compound Database |
| SBDD | structure based drug design |
| SDH | shikimate dehydrogenase |
| Ser | serine |
| siRNA | small interfering RNA |
| SK | shikimate kinase |

| | |
|---|---|
| SNP | single-nucleotide polymorphisms |
| T | thymine |
| tC4M | *trans*-cinnamate 4-monooxygenase |
| TCM | Tradional Chinese Medicine |
| TDB | 1,5,7-trisazabicyclo[4,4,0]decene |
| TFA | trifluoroacetic acid |
| TGF β-2 | transforming growth factor beta-2 |
| THF | tetrahydrofuran |
| Thr | threonine |
| TOPO | topoisomerase |
| Trp | tryptophan |
| Tyr | tyrosine |
| V or Val | valine |
| VEGF | vascular endothelial growth factor |
| WHO | World Health Organization |
| ZINC | Zinc Is Not Commercial |
| θ or theta | polar angle |
| φ or psi | azimuthal angle |

# Chapter 1: Introduction

## 1. Strategy in drug design

The World Health Organization (WHO) has engaged many countries throughout the world in 2017 with the aim to scale up the fight against hepatitis B and C with a view to the total eradication of hepatitis by 2030. Although this aim is currently far from realization (in 2018), it is achievable following the WHO guidelines.[1] It is reported that greater than 86% of countries had identified milestones and 70% of these countries have initiated preparation of a national plan to reach these milestones in the efficient prevention, diagnosis and treatment of hepatitis.[2] In 2015, 325 million people were infected by a viral hepatitis (257 million for hepatitis B and 71 million for hepatitis C). During this time there were 1.34 million deaths from the disease. This number is worse than the number of deaths worldwide from the Human Immunodeficiency Virus (HIV).[2, 3]

The challenge for WHO with some countries (particularly Third World and emerging economies) is the provision of accessibility of essential drugs for hepatitis treatment; these are of a class of direct acting antiviral drugs (DAAs). Only 7% of those infected with hepatitis have access to these drugs. A compounding issue is the cost of these DAA's due to the limited presence of generics within this family of drugs. Unfortunately this limited presence of generics is not an isolated problem limited to this class of drugs, and it perpetuates as a result of several different factors. The drug-development process, including clinical trials will be discussed in this chapter in order to understand the multifactorial issues giving rise to these problems. Also discussed will be the different strategies followed in the production of potentially, new, therapeutic, small molecules.

### 1.1. Stagnation and other issues inherent in the development of new drugs

#### 1.1.1. The drug-development pipeline

The development of new drugs appropriate to a particular biological target is a long and expensive process. In many cases drug candidates proceed to a limited extent through the drug-discovery pipeline, ultimately failing to become an approved medicine. This drug-discovery pipeline has several stages, and this failure may occur at any one of these stages. Within the pipeline is rational drug design based on Emil Fisher's concept (from 1894) of a "lock and key" in which a drug will inhibit or activate an enzymatic target by a perfect match

or fit to a corresponding active site.[4, 5] This fit of drug to receptor is an interaction at a very specific receptor site, and this interaction forms a drug-receptor complex.

### 1.1.2. Targets for Drug Discovery

The identification of the therapeutic target or receptor is therefore the first step in this pipeline for the development of new drugs. The definition of a biological target is a receptor that is a protein, peptide or nucleic acid which has some biological purpose or activity. This biological activity can be increased, decreased or suppressed using 3 kinds of compounds: small molecules, larger biological compounds (such as antibodies) or recombinant proteins (e.g. insulin).[6]

The identification of the targets may come from a genomics approach with phenotype screening and validation by bioactivity assays.[6, 7] Typically, the first strategy in identifying a new target is the analysis of ribonucleic acid (RNA) and the protein expression in target tissues, or this analysis but in comparison between healthy and diseased tissue.[6] The highlighting or the suppression of particular genes with knockout mice (genetically modified to suppress or enhance that gene) helps to understand the role of the gene and the receptor that the gene encodes by extension.[8] This perturbation of genes combined with phenotypic readout has shown very good results.[6]  Observation of single-nucleotide polymorphisms (SNP) is used in gene mapping and to highlight mutations of diseases. In addition to this, wide association studies reveal the targets themselves inside of their genomes. Furthermore this technique may predict which drugs are suitable for the identified targets. For instance, the Loci for type II diabetes was identified by genome wide association. The researchers discovered that these particular Loci may be targeted by the glitazone drug class which was already in use for another disease.[7] The issue of this technique was the limitation of the technology and the restricted access to the database of genomes. In 2014, the National Institute of Health motivated to bridge these two gaps (in terms of both technology and access to data) with the successful launch of the Illuminating the Druggable Genome (IDG) programme. This programme has investigated genes and proteins responsible for off-target interactions of existing drugs, and has identified particular proteins that are not suitable as druggable targets. This programme furthermore defines the long-term expansion of the potential therapeutic space and will refine the research of new targets.[9]

Another strategy in the identification of new targets is to use the side-effects of released (approved) drugs. The side effect of a particular drug may well be linked to a particular separate receptor.[10]

Targets normally are protein in nature, enzymes with specific active sites, and these proteins can be regrouped as families following their natures. Some examples of families are the G protein-coupled receptors (GPCR), ligand-gated ion channels, nuclear receptors, phosphodiesterases, proteases, protein-kinases and voltage-gated ion channels, to name a few.[11] Some of these families have been targeted successfully by developed drugs, and the distribution of developed drug per family is illustrated in Figure 1.1.[12]

However targets are identified, they require validation. This validation is through *in vitro* and *in vivo* assays.
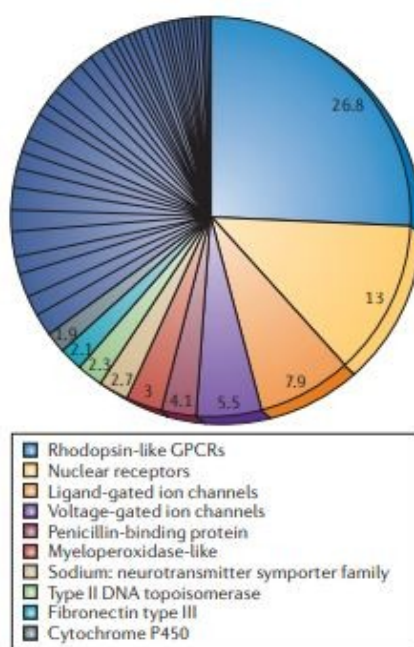


Figure 1.1: Gene-family distribution of current drugs per drug substance (with the permission of the publisher). [12]

### 1.1.3. Identification of Lead Compounds

A crucial task in the process of identifying possible drugs for a particular target is the identification of a lead compound. A lead compound is a molecule that possesses the appropriate biological activity, although it may only have weak activity and some side effects. For lead compounds, the binding interaction may be fragile, or they may form a ligand-receptor complex without long-term stability. Lead compounds are useful in that they provide a starting point for optimization towards a successful drug candidate.[5] The lead molecule may then be optimized by the addition, deletion or modification of chemical moieties that are

present in this original compound.[13] During the optimization/modification cycle the derived molecules are tested in preclinical assays, the purpose of which is to provide evidence for the effectiveness of these new compounds, and to also verify their side effects and pharmacokinetic proprieties. Further it is essential to identify and quantify the drug and its metabolites, and also to optimize its Galenic forms. Assays used in preclinical work for drug discovery may be *in vitro* or *in vivo*. [5]

### 1.1.4. Clinical Trials

Successful drug candidates from these assays are taken forward to clinical trials which are divided by 3 phases (Figure 1.2). Each of these three phases requires both a different type and a different number of patients to test the drug candidate. Phase I trials help to determine the maximal dose at which toxicity effects will appear and further to characterize the lifetime and identify the presence and concentrations of metabolites. Only 20 to 80 healthy volunteers are used in phase I trials (and these individuals will usually be men).[14] The phase II determines the optimal dose for a patient and accepts only about 100 volunteers for treatment or prevention of the research disease.[15] The last phase is used to verify in large scale (between 100 to 10000 sick volunteers) the efficiency and the safety of the drug. This last phase will compare with the reference treatment or placebo or both. Phase II and III are often repeated several times before final approval. All of these trials are compulsory and legally cannot be omitted. If the drug passes all three trials, application for authorization to sell the drug in the market is possible. A last non-obligatory phase (IV) may be used to check the side effects of the drug, to evaluate the increase of patient condition and the efficiency of the drug. Sometimes side-effects reveal new properties of the drug providing an opportunity for a new therapeutic target. During phase (IV) particular groups of individuals will be watched closely, such as pregnant women, children and the elderly. All of these clinical assays are parts of the clinical development. [5]
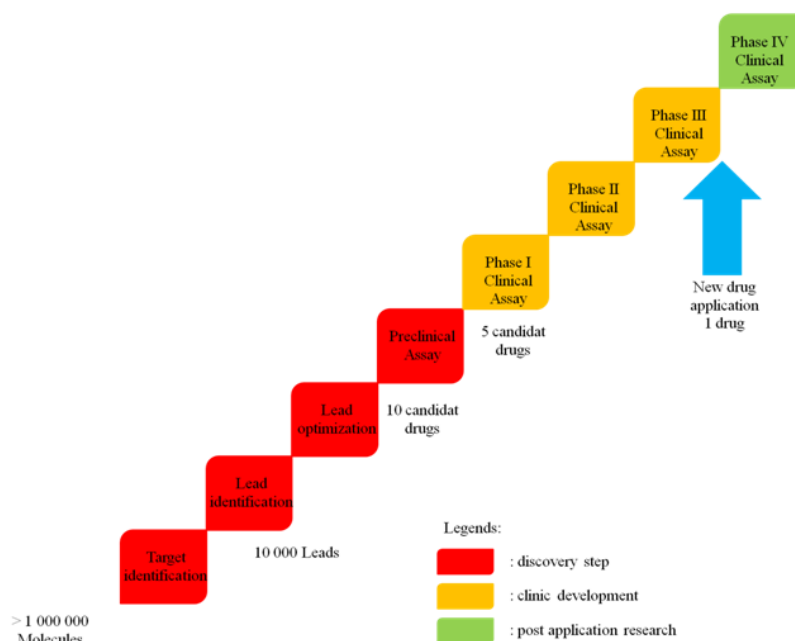
Figure 1.2: Life cycle of the drug. Inspired of the schemes inspired of[16, 17].

## 1.2. Issue of drug development

### 1.2.1. Economics of drug research and development

In 2017, only 46 drugs were approved by the Center for Drug Evaluation and Research (CDER). [18] This information is surprising given the marked difference in orders of magnitude of numbers of medications in development, compared with the numbers of those released on the market. This may be explained at various levels. In economic terms, the development of a drug can both be uncertain and also be very expensive especially if the project is abandoned.[19] From 1970 until present, some studies have highlighted the exponential increase in cost of research by year.[20] New companies working with new biotechnologies such as "large molecule" molecules may not have the capacity to follow through with a particular drug in terms of logistic and financial issues. Many of these smaller companies rely on larger pharmaceutical companies for clinical assays and commercialisation of the drugs. Taking these decisions are risky for both parties and one of the reasons why some pharmaceutical companies have very strict requirements for drugs proceeding to clinical trial.[20] Included in this decision-making are the use of statistics and probability tools to help in the prediction of the cost of success. This cost of success is compared against predicted prices and demand of the potential drug in developed countries. The predicted price in itself is dependent on several factors such as if the drug is innovative and the price of present drugs

available on the market. These statistical tools also take into account, for instance, where a market drug already exists in the therapeutics class (called "me-too" drugs), where the putative drug will therefore warrant a lower price than if none were already available. In this case a company would be discouraged to develop a new medicine in this therapeutic class.[21] Figure 1.3 represents the Trends in capitalized pre-human, clinical and total cost per approved new drugs in 2013.
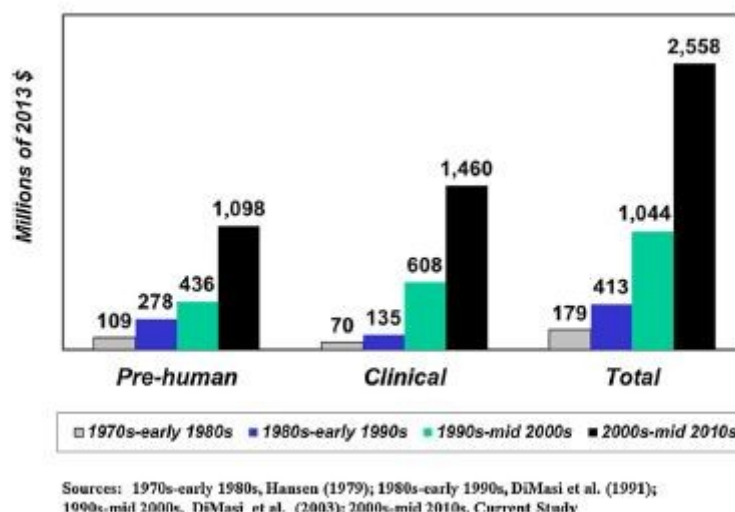


Figure 1.3: Trends in capitalized pre-human, clinical and total cost per approved new drugs (with the permission of the publisher.[22]

### 1.2.2. Similarity of the drug, side-effect and induction of resistance

That pharmaceutical companies are reluctant to develop "me-too" drugs as is previously described. However the situation is more complicated than it seems to be. Developing an entirely new drug (called a "pioneer") might also be a risky path to take. From the point of a pharmaceutics company, therapeutic areas do not always equate exactly to their interests. Some classes of disease do not affect enough of the population to counterbalance the cost of research into these therapeutics, including the possibility of failure of this research. This is more evident for companies in developed countries. One example that illustrates this is the research of new antiparasitic treatments. In 2004, only 2 antiparasitic treatments were released in comparison to 9 antiretroviral (specific to Human Immunodeficiency Virus) which were released in the same timeframe onto the market.[23] Oncologic treatments (including immunology modulators), central nervous system agents and systemic anti-infectives (in particular HIV) are on-trend and these represented 65% of the drug released market landscape

in 2008.[24] Figure 1.4 and 1.5 represent therapeutics area focus for specialty pharmaceuticals and new molecular entities disclosed in the R&D programs of the world's 15 largest pharmaceutical companies. Other issues faced in the research and development of new drugs relate to their potential side-effects. The origin of side-effects may be explained by the non-specificity of the drug. For instance, a particular drug may well dock against its target enzyme, but also dock to one or several other non-desired enzymes. One famous example of this is the drug Sindenafil. This compound was tested for angina (cardiovascular problem) and resulted in poor to average results with respect to the placebo. Researchers were surprised to note that some subjects wanted to keep their drugs after the study. An unintended consequence of the investigation was that this drug was shown to help men with erectile dysfunction. A new class of molecules was created by using this side effect of its original therapeutic class. In 2006 the action of the drug was also extended to pulmonary hypertension.[25] There is a correlation between the target binding similarity and side effect similarity with *in vitro* proteins. This may be used to help predict off-target binding.[10] With this in mind, parameters from these and from other studies have been used to create a network map of drugs predicted to have common protein targets, and the drug target relations between them represented in Figure 1.6.[10]
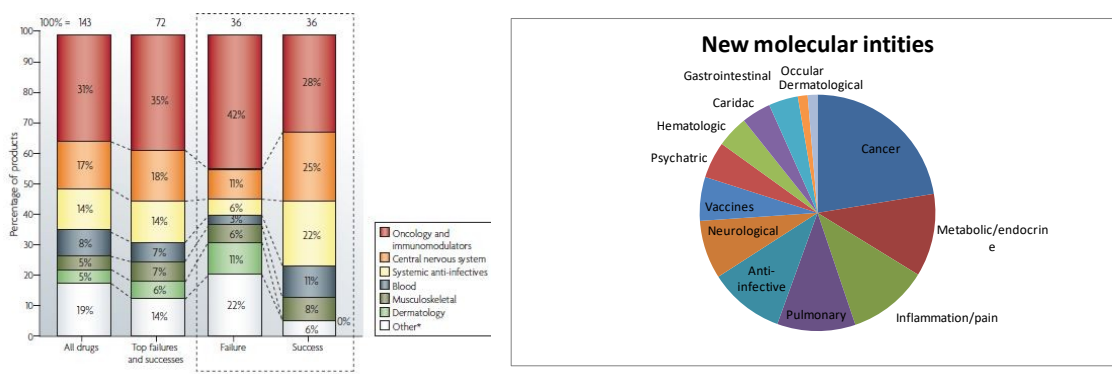


Figure 1.4 and 1.5: Therapeutics area focus for specialty pharmaceuticals and new molecular entities disclosed in the R&D programs of the world's 15 largest pharmaceutical companies. [24](with the permission of the publisher) and inspired of [23].

7
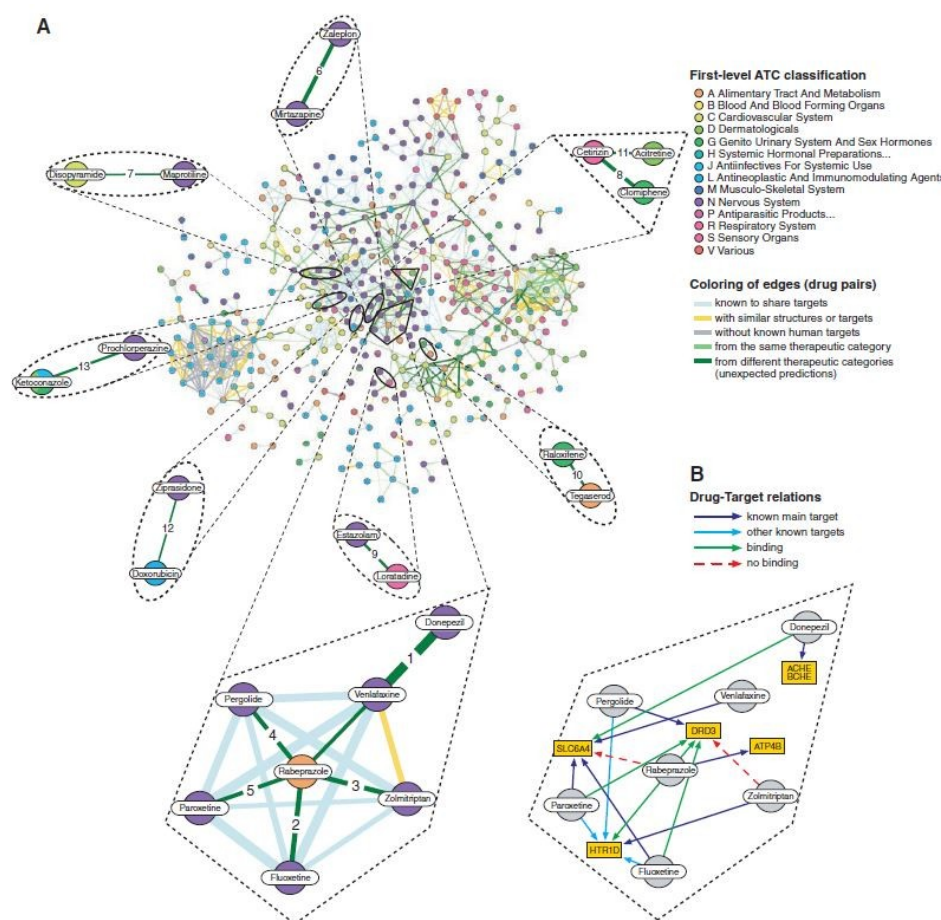
Figure 1.6: A network of drug predicted to have common protein targets and B sub network map of selected drug-target relations around rabeprazole (with the permission of the publisher).[10]

The last major issue of drug development is mutation of the target proteins. Amino acids changing during replication are due to errors appearing in the DNA responsible for the target protein. Mutation may provide a means for the exogenous to survive a particular treatment. This phenomenon gives rise to two types of mutation: minor and major. The consequence of mutation is defined as such depending upon the location of the mutated amino acid, whether internal or external to the active site. Major mutations in the active site pocket might well preserve the physical and chemical properties of the enzyme, with an inhibitory drug still working. However the modifications of an active site residue often lead to a change of conformation of the active site thus blocking access to the drug. Mutations in other locations may also be responsible for the reduction of the binding energy of inhibitory ligands. The best example of this is the mutation of the HIV-1 subtype C protease. Untreated patients produce $10^9$ to $10^{10}$ virions per day, while HIV-1 is only composed of $10^4$ nucleotides. Furthermore the virus mutates $3 \times 10^{-5}$ nucleotides per replication cycle due to transcription error from the reverse transcriptase of the virus. When a treatment is used, in many cases it prevents

replication of the native virus but not of the mutated one. After time, only the mutated virus remains. This action due to the drug is called "therapeutic pressure" and, in the worst case, results in failure of the medicine.[26] Polymorphism and drug mutations allow the protease to be flexible in response to inhibitors. [27]

## 1.3. Actual strategy proposed by industries and research group

### 1.3.1. Extraction from nature

These issues, particularly those relating to the economics of drug development and those of drug resistance drive industries and research teams to explore novel approaches to identify new potential compounds. Natural product molecules fitting these criteria can be found with the use of four different approaches. The first approach utilizes screening of a selection of plant extracts. This approach may be used for a specific class of compounds such as flavonoids or alkaloids family or from randomly selected plants for specific bioassays.[28] The ethnopharmacology approach is based on the observation, description and experimental approach of indigenous drug usage. The research links several fields such as botany, biochemistry, chemistry and pharmacology to human sciences like archaeology, history, linguistics or anthropology. This interdisciplinary approach started with incidents such as Serturner isolation in 1805 of morphine from the *Papaver somniferum* plant.[28] China and India are two major countries with an important legacy of traditional medicine spanning thousands of years. An anti-malarial drug was discovered using this approach: in 1972 artemisinin from *Artemesia alba* tree was isolated (*Artemesia alba* is a traditional Chinese medicine).[28] The last methodology is the zoo-pharmacognosy approach. The observation of animal interactions with botanical sources helps to identify a particular family of compounds as potential drugs. Some animals eat with intention specific botanical sources, which are not normally part of their diet and which present toxic characteristics for parasites. It has been recorded, for instance, that pregnant elephants in Kenya may eat particular leaves to induce birth.[29]

The principle of extraction of active compounds from natural sources is crucial both for the chemical identification of the compound, but also for its isolation from other compounds present within the natural source. [30] The first step of this long process is collection the plant which possesses the required therapeutic characteristics. Some measures have to be taken in order to preserve the active compound from degradation during the process of extraction. Organic solvents are used in the extraction and differences in solvent help to separate the

required molecules from the rest of the source, dependent on their physicochemical properties. An example of extraction is represented in Figure 1.7. Different analyses may be helpful in the verification of the presence of the desired molecule in one or more of the solvent fractions produced by chromatography. TLC and HPLC coupled to analysis techniques like UV or mass spectrometry provide this information prior to bioassay use of these fractions. Table 1 shows some therapeutic molecules extracted from natural sources and indication for treatment.
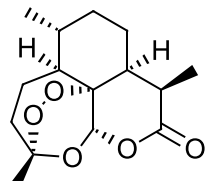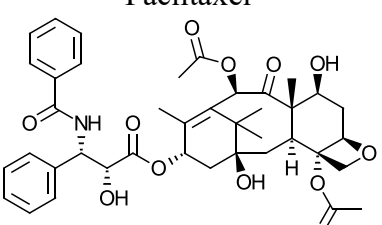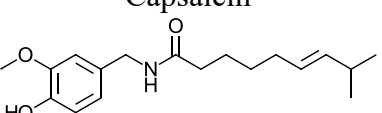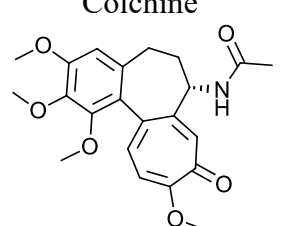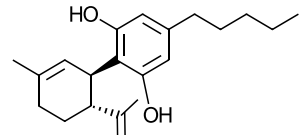
| Name and structure of the molecule | Plant species | Trade name and year of introduction | Indication |
|---|---|---|---|
| Artemisin  | *Artemisia annua L.* | Artemisin 1987 | Malaria treatment |
| Paclitaxel  | *Taxus brevifolia.* | Taxol 1993, Abraxane 2005, Nanoxel 2007 | Cancer chemotherapy |
| Capsaicin  | *Capsicum annum. L* or *C minium Mill.* | Qutenza 2010 | Postherpetic neuralgia |
| Colchine  | *Cholchicum spp.* | Colcrys 2009 | Gout |
| Cannabidol  | *Cannabis Sativa L.* | Sativex | Chronic neuropathic pain |

Table 1.1: Some natural compounds approved for therapeutic use. [31]
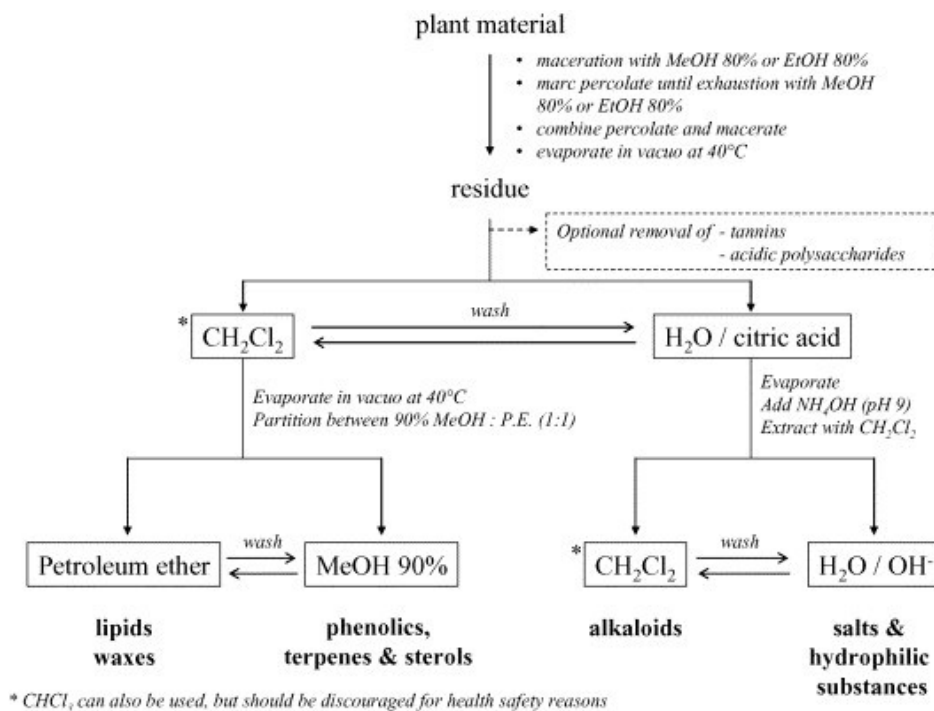
Figure 1.7: Standard scheme for extraction process (with the permission of the publisher).[30]

### 1.3.2. Semisynthesis and hemisynthesis

One of the two major problems encountered in the extraction of natural compounds is that some botanical sources are rare, under protection and are thus not available for collection. This may also be the result of local war, natural catastrophes or changing legal regulations.[31] The second problem relates to the biological activity of the natural compounds. Although active, the molecules isolated from a biological source may be too weak to present the full desired therapeutic effect, yet be enough to represent a hit within assays. As such the compound must be modified in order to increase its efficacy or to prevent its potential to induce drug resistance in an exo-organism. Artemisinin, for instance is a very poor antimalarial, but analogues of Artemisinin, such as an adamantane derived analogue (OZ277) and a dimeric analogue have been shown to have activity to the extent that a single dose of either destroys malaria in an infected mouse.[32] These two problems, relating to both accessibility and to activity (hence quantity required), could make pharmaceutical companies contribute to the extinction of species if they were to release particular natural compounds onto the market. In 2006, it was estimated more than 20% of the worlds medicinal and aromatic compounds are endangered.[31] One famous example is the chemotherapy drug: paclitaxel. This molecule is synthesized from a compound extracted from the bark of the

Pacific yew tree (*Taxus brevifolia*). Over the past 30 years, research has estimated that the decrease of the population of this tree is between 50% and 80%.[33] To avoid this type of situation, pharmaceutical companies try to recreate the therapeutic molecule by alternative biological methods or by chemical reactions with a minimum number of steps together with a comfortable yield.[34] An example of this is the sexual hormone progesterone which is used principally for hormone replacement therapy. It was isolated by Adolf Willard Allen in 1934, where, he used more than 181 kg fresh ovaries to extract only 1.03 mg of pure progesterone.[34, 35] Due to the very low yield of this extraction process the price of the drug was around 200 USD.$g^{-1}$. In 1940, Russel Marker synthesized progesterone from diosgenin (a sapogenin) present in the trillium *erectum* root. This synthesis has 5 steps with a good yield.[36] This semi-synthesis is represented in Figure 1.8. It was found later that by using the Mexican yam as primary base, it was possible to increase the concentration of diosgenin by an order of magnitude. The price of progesterone subsequently decreased to 1.5 USD.$g^{-1}$ and after several optimizations to 0.19 USD. $g^{-1}$.[34]
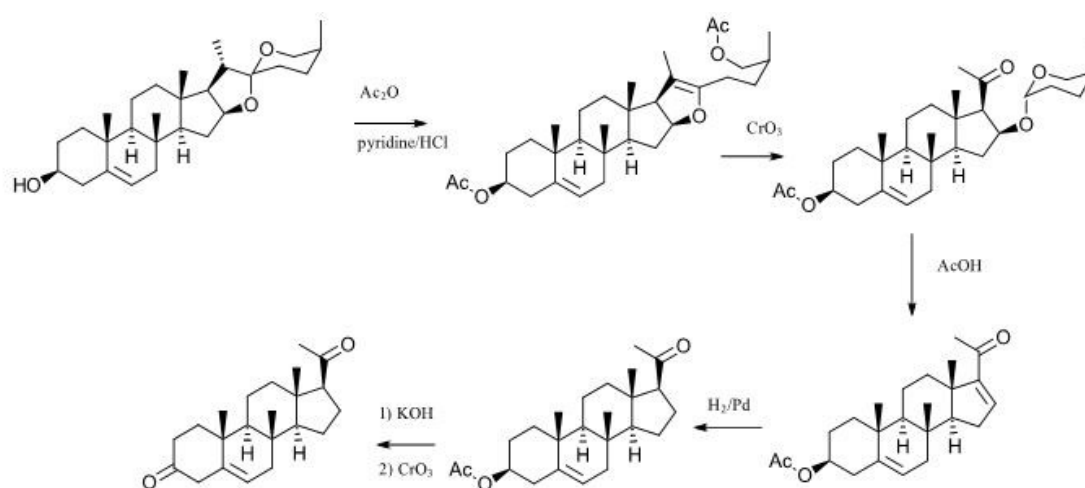


Figure 1.8: Semisynthesis of progesterone developed by Russel Marker.

Another example is with insulin, which is a protein hormone which is normally synthesized by the β cells inside the pancreatic islets within the pancreas. This hormone is responsible for the regulation of the metabolism of carbohydrates. A patient who is affected by type I diabetes (destruction of the pancreatic islet) or type II (insulin resistance) has trouble with insulin production.[37, 38] However, human insulin has a high similarity with the porcine one, with only one amino acid different (B30, Asparagine/Aspartic Acid at position 30).[39] The development of commercial insulin involved a mutation corresponding to this particular

amino acid. This action may be performed using either a triptic digestion or Edman degradation and iodotyrosylation using *Escherichia coli* (*E. coli*); this bioengineering resulted in the successful hemisynthesis of insulin. [39, 40]

### 1.3.3. Total Synthesis

Natural product isolation, or semi- and hemisynthesis are proven routes to natural bioactive compounds, and the synthetic routes become feasible where the natural product is rare. However, none of these address situations where the natural compound is poorly active. It may be desirable to modify the molecular structure in order to increase the efficiency of the drug.[41] This action may make the process more complex and expensive in order to obtain a highly active drug. A third course of action is the total synthesis of a particular drug, copying or mimicking the biological pathways in order to produce the desired molecule. Total synthesis may use organic synthesis and/or enzymatic reactions which produce secondary metabolites.[42] The first total synthesis of an "organic" compound was described in 1828 by Friedrich Wöhler, who converted the supposedly "inorganic" compound ammonium cyanate to the "organic" (natural) urea (resulting in the change of definitions in chemistry of the terms inorganic and organic).

Total synthesis has become increasingly more intricate, and successful synthesis of increasingly complex compounds is due to an increase in chemical understanding of a range of issues. These issues include the nature of chemical bonds, the role of orbitals and pericyclic reactions, the production and use of reactive species such as anions, cations, carbenes and radicals in forming new bonds, the development of catalysts, particularly those involving transition metals and, importantly, ways of controlling diastereoselectivity and the development of asymmetric synthesis. Asymmetric synthesis is of particular importance, due to the differential interaction of enantiomers (or diastereomers) with biological targets. Stereochemical strategies avoid the problem of creation of the wrong enantiomer (or diastereomer) with the use of stereocontrol. An Evan's reduction by example may be used in asymmetric synthesis to favour a single enantiomer with both good yield and good enantiomeric excess.[43, 44]

Total synthesis was revolutionized by E. J. Corey in 1961 with the introduction of retrosynthetic analysis, proven by many examples, but first by the synthesis of longifolene.[41, 45] Retrosynthetic analysis examines the final product and uses disconnections based on organic reaction mechanisms to identify precursors called synthons.

The process is repeated until the reagents identified correspond to commercially available compounds.[46] The recognition of key substructures as synthons helps to reveal essential bonds for disconnection. Functional groups in molecules are the key to successful disconnections. There are a plethora of strategies accessible to retrosynthesis, beyond the scope of this discussion.[46]

### 1.3.4. *In silico* approach

#### 1.3.4.1. Chemical Libraries

*In silico* drug design requires existing ligands whose 3D conformation in the active site of a target is known. There are 3 techniques which may be used to solve this problem. X-ray crystallography, nuclear magnetic resonance (NMR) data, and the use of protein and small molecule data banks in high-throughput virtual screening (HTVS).[47] One major problem with HTVS is the accessibility of databases of small molecules. Indeed in order to perform determination of the conformations of a large number of ligands bound in the active site of many proteins, X-ray diffraction and NMR are very limited techniques in terms of the extended time required per individual system. However, for HTVS, several libraries which have been built for different purposes and classification are available both in terms of protein and ligand. The most common justification of the creation of a virtual library or database of ligand molecules is the non-accessibility of structures from academic research including restricted-access publications, and from Masters and PhD theses.[48, 49] Natural products, as described previously play an important role *in silico* drug design. Over 30 years of studies (from 1981 to 2010), it has been shown the contribution of natural products to drug discovery (in terms of approved drugs reaching market) is approximately 64%. Given this, there are a wide variety of ligand databases, from databases of natural and of synthetic products through to libraries of theoretically attainable molecules.

#### 1.3.4.2. Limited Scope Databases

In terms of natural product databases, natural products extracted from traditional Chinese plants are represented by over 23 000 compounds isolated from 7 000 different natural sources. These compounds are curated within several Chinese databases such as the Traditional Chinese Medicine (TCM@Taiwan), [50] the Chinese Herbal Drug Database (CHDD) [51] and the Chinese Natural Product Database (CNPD) [52]. Several other national-

or region-based databases have been influenced by the development of these Chinese databases. One of several African databases is the Northern African Natural Products Database (NANPDB) [49] which includes only natural products which have been identified from organisms indigenous to the Northern parts of Africa (specifically Algeria, Egypt, Libya, Morocco, Sudan, South Sudan, Tunisia, Western Sahara and parts of Northern Mali). This database contains compounds that vary widely in terms of biological activities against a wide range of diseases, and this information is available from the database. In contrast there are other databases (again regional) that focus simply on one disease such as the database Afrocancer.[53] 20% of compounds in NANPDB have a known biological activity. It has also been shown that 57% of the molecules in this database follow the Lipinski rules in terms of being most drug-like without any Lipinski violations, while a further 32% have at most two Lipinski violations.[49] The database is widely accessible and is also hosted by OMICTOOLS.[54]

Other innovative databases include the South African Natural Compound Database (SANCDB).[48] Considering South Africa as a rich terrestrial and marine biodiversity, research has shown that more than 3 000 plants species specific to this country are actually used in treatments (this is mainly from traditional medicine use). SANCDB initially used as a base the Rhodes University thesis collection (206 references) and included compounds identified in these theses which were extracted from South African plant or marine life (166 sources).[48] SANCDB was designed with natural product chemists in mind, including the functionality for individual researcher to contribute to the expansion of this database by submitting compounds they had identified in from South African natural sources. In term of range of biological activity of compounds in this database, antimicrobial, anticancer, antidiabetic and antineurodisease proprieties are common on SANCDB. [48]

### 1.3.4.3. Broad Scope Databases

Other databases do not focus only on natural products. In 2006, the database DrugBank became available. It has an entirely different purpose to the natural product database, and that is to link bioinformatics and chemoinformatics datasets. It contains 4 distinguished categories of drugs: FDA-approved small molecule drugs, FDA-approved biotechnologic drugs (such as protein or peptides), vitamins, metabolites and experimental drugs (such as unapproved drugs, de-listed drugs and illicit drugs, enzyme inhibitors or potential toxins). These categories are updated frequently. The DrugBank also gives details for each molecule about its physico-

chemical, pharmaceutical and physiological properties and activities.[55] More properties are added to this database as it is under constant revision – this includes more recent additions of pharmacological, pharmacogenomic and molecular biological, pharmacokinetics, drug metabolism quantitative structure activity relationships (QSAR) and absorption, distribution, metabolism, elimination and toxicity (ADMET) sets of data. This database contains roughly 12 000 drugs entries including more than 2 700 FDA-approved drugs. In order to populate DrugBank with data, information from more than 27 500 publications from diverse journals were collected over the 12 past years.[56]

Another database of molecules, similar in some respects to DrugBank is PubChem. PubChem came online in 2004, released by the US National Institute of Health (NIH). PubChem contains diverse types of compounds such as small molecules, macromolecules, siRNAs (small interfering RNA), miRNAs (microRNA), carbohydrates, lipids and peptides. PubChem itself is actually a combination of 3 databases: PubChem Substance (e.g. plant extracts), PubChem BioAssay (bioactivity of particular substances or molecules) and PubChem Compound (molecule database). Contributors may deposit in one of these 3 databases depending on what information they have to hand about the compound to be submitted. In terms of intellectual property, PubChem includes options for owners or contributors of data to have the data held private for a specified period of time. In addition to the usual unpatented contributions, PubChem also possesses a large collection of data linked to the World Intellectual Property Organization including 6 million patents, 16 million unique chemical structures and 329 million substance patents. One popular feature of PubChem is the option for contributors to withhold public access to their submitted data until a specified date (for example, until a patent for the information is secured).[57] Some journals automatically submit information to PubChem about molecules and biological assays (e.g. Nature Publishing Group). Furthermore not only do publishers contribute to this database but chemical vendors, academic institutions, government agencies and pharmaceutical companies are also involved.[58] In 2017 PubChem contained over 2 million chemical structures, including over 200 million individual bioactivities. [59, 60]

A third example of a non-exclusive database is the Zinc Is Not Commercial (ZINC) database. ZINC, however has a different purpose from the other database sets. The catalogue of entries is mainly composed of purchasable compounds (and includes bioactivity data) and is updated monthly.[61] They have also updated their database to include natural products, metabolites, drugs and experimental drugs. The database contains information from 266 commercial suppliers.[62] By way of requirement to be listed as a purchasable drug, ZINC aims for 90%

of their compounds to be purchasable in 90 days. [63] The latest version of ZINC, ZINC15 contains 220 million molecules. ZINC also uses ChEMBL as base for the literature research.[62]

Table 1.2 illustrates the number of small molecules provided by the different databases cited previously. Note that the numbers in the table, are the latest information (2018) directly from the respective websites, rather than from literature (cited above) which may be months out of date.

| Name | Nature of the catalogue | Number of small molecule | Website link | Number of chalcone found | References |
|---|---|---|---|---|---|
| NANPDB | Natural Products | 4 500 | http://african-compounds.org/nanpdb/ | 7 | [49] |
| SANCDB | Natural Products | 716 | https://sancdb.rubi.ru.ac.za/ | 5 | [48] |
| DrugBank | Drugs (natural and synthetic) | 11 885 | https://www.drugbank.ca/ | 2 | [56] |
| PubChem | Drugs (natural and synthetic) | 96 532 203 | https://pubchem.ncbi.nlm.nih.gov/# | 1242 | [57] |
| ZINC$^{15}$ | Drugs (natural and synthetic) | 736 001 654 | http://zinc15.docking.org/ | 26 | [62] |

Table 1.2: Some examples of database of small molecule for high-throughput docking.

### 1.3.4.4. Library generation

### 1.3.4.1. Combinatorial approaches

All libraries described previously are limited by the number of molecules which they make available, even though they may be considered large libraries. In order to have a better idea of the capacity of a large library, it is important to understand the definition of chemical space. Chemical space corresponds to a multidimensional descriptor environment. It requires a metric, such as physicochemical properties (molecular weight, polar surface area etc.), mathematical (position in multidimensional space, similarity metrics, etc.) or even a QSAR

descriptor to enable comparison of molecules. Although, the dimensions of the chemical space are many, described by the many independent metrics, there is allowance for reduction in the number of molecules selected or used in a screening protocol, whilst maintaining the appropriate chemical diversity. The choice of descriptor is thus very important in order to improve the efficiency of database mining. Chemical space mining uses statistical tools to identify common denominators which may describe the chemical space of interest well. The most common tool used is that of principle component analysis (PCA). PCA is linear projection method which may transform from a large number of dimensions (L) a smaller one (S). The smaller dimension is also composed of uncorrelated vectors which are defined by eigenvectors of the matrix of variance–covariance from the original matrix.[64] These vectors are used as a new orthogonal axis.[65] The largest eigenvalues from this data explains most of the original value before the transformation of their dimensions. PCA may explain which parameters can mostly define a section of chemical space. [64]

This view of chemical space enables the generation of combinatorial libraries. Combinatorial libraries use large partial regions of chemical space with common denominators such as similar chemical functionality. However such generated libraries of compounds might not be as diverse as is required. One reason for this lack of diversity may be the presence of significant numbers of derivatives of specific molecules which might not be relevant and are therefore not included in the library. In order to use such libraries for medicinal-type virtual screening, these undesired compounds have to be removed from the combinatorial library. A first combinatorial library may give poor results, but from this library promising structures may be identified through the application of filters. Parameters such ADMET, QSAR or Lipinski rules may be applied to good performing structures in a combinatorial library in order to optimize these leads. If many leads are obtained from the library, they may be used themselves to form an optimized library.

There are two principal methodologies that may applied to the design of combinatorial libraries. A library may focus on a target family or a single biological target for instance. QSAR models of existing compounds binding to this target or targets is/are used to identify pharmacophores which are active against specific proteins. This methodology is also effective for enzymes that share similar proprieties.[65] Figure 1.9 represents a QSAR analysis of quinine as an antimalarial drug.
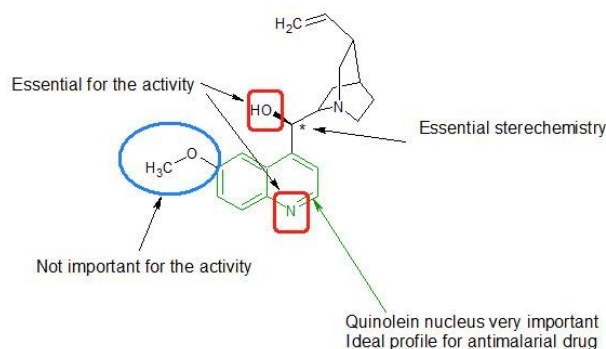
Figure 1.9: QSAR analysis of quinidine.

Another basis for the generation of combinatorial libraries is structure based drug design (SBDD). This method allows for the focus on structure of the molecule instead of the target, where the target may be unclear. Where the target is known, SBDD may involve docking with a small library of compounds to identify and therefore select a molecular framework as a core for modulation or improvement. Three examples of software that allow for SBDD methodology are CombiDOCK, OptiDock and CombiGlide. There will generally selected, after docking, an optimal core (which could be unsubstituted and substituted). The software will make guided decisions on which substitutes may be incorporated into the core based on the algorithm followed. If the resultant molecule is judged to be satisfactory, this modified core is added to the library. CombiGlide is developed by Schrödinger available in Maestro and has the added advantage of being able to filter modified molecules by proprieties such as ADMET. [66]

CombiDOCK, OptiDock and CombiGlide processes are summarized in Figure 1.10.
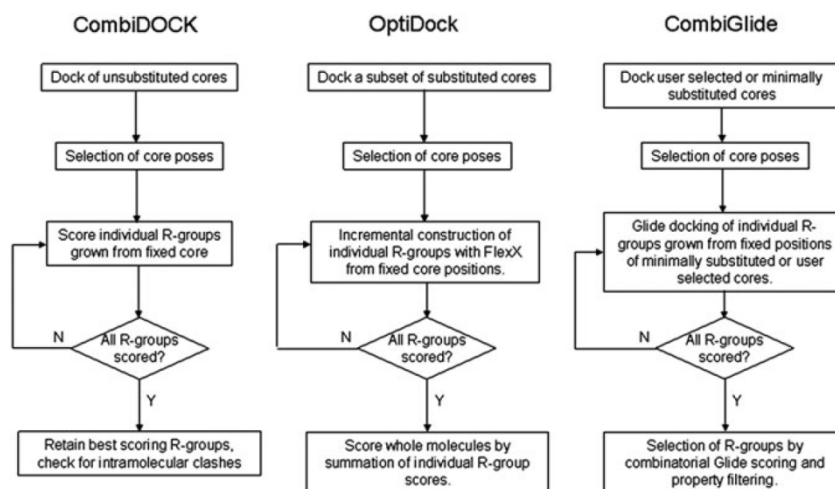


Figure 1.10: Process of library generation of SBDD of CombiDOCK, OptiDock and CombiGlide (with the permission of the publisher). [65]

Another quite different method of constructing libraries is based on essential fragments of active molecules. This method is called fragment based drug design (FBDD), and is a useful process in the optimization of leads where other methods present difficulties.[67] The library uses one lead as a base which is then modified by adding fragments. The fragments when isolated, present a low binding affinity for the target but the sum of the binding energy per atom, for all these fragments, is superior to the binding of a lead molecule. The aim is to create a molecule with linkers and fragments which have a high affinity to a specific region of the protein.[65] The Figure 1.11 represents the fragment based drug design process.
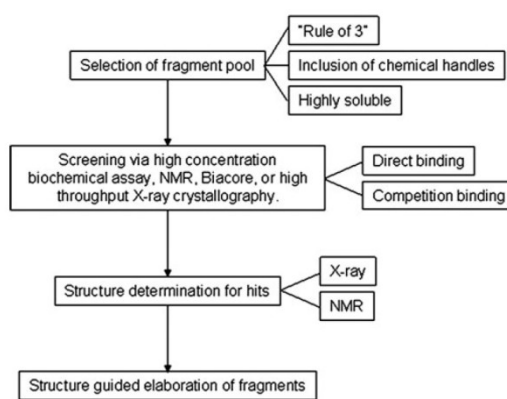


Figure 1.11: FBDD process (with the permission of the publisher). [65]

One example of library generation is the creation of the GDB libraries. These libraries use the general concept of chemical space as a basis, but are restricted by practical limitations. In 2004, GDB contained organic structures with a maximum of 11 atoms, but without three or four atoms ring, with some measure of stability and synthesizability. At this point the library contained less than 14 million structures.[68] The GDB, after updates now possesses 26.4 million molecules including 166.4 million stereoisomers.[69] The database uses molecule containing carbon, nitrogen, oxygen and halogen atoms following standard valence rules, synthetic pathways and measures of stability. Most of the molecules present in this new version of GDB, GDB17 are composed of nonaromatic heterocycles. The process of building GDB17 may be described in four steps. Hydrocarbon skeletons and carbon atoms were respectively considered as graphs and graph nodes. A single bond corresponds in this representation to a graph edge.[70] Using the program GENG, a full set of non-isomorphic connected graphs was created with up to 17 nodes with a maximum of four edges per node

(the maximum degree of all graphs was four; the maximum order of each graph was 17). Rings were chosen to be bigger than 4 atoms to remove steric and strain problems. After removing these graphs, the remaining acceptable graphs were translated into molecular format by substituting nodes with atoms and edges with bonds. A further selection of hydrocarbons was made by filtering out some graphs, such as those containing 14 carbons together with more than 3 bridgehead atoms (atoms shared by three or more rings) or graphs with 17 carbons but with small rings. From more than 114 million graphs, this further filtering reduced the value to 5.4 million structures. Next the hydrocarbon skeleton is transformed with the inclusion of double and triple bonds and filtering of these systems to remove highly unsubstituted systems like allenes or where unsaturation is within small rings. Duplicates (including tautomers) are removed from the resultant set. The molecular skeletons are then submitted to other filters to substitute carbon by oxygen and nitrogen, but at set positions and with a maximum number of these heteroatoms to avoid molecules that will easily polymerizer, or molecules that are likely to be toxic or explosive. The last part of the design is post-processing, diversifying the GDB such as by substitution of hydroxyl groups by halogen or oxygen atom by sulphur.[69, 70]  These steps lead to GDB-17 in its current form, containing now more than 166.4 million molecules. Compared to PubChem, GDB-17 contains many orders more molecules when size of molecule is taken into account, and this includes a large number if stereoisomers. However the entire library is too large to be use for high-throughput virtual screening, and for this purpose use of a subset is feasible.[69] In contrast to PubChem, the synthesis of compounds from GDB have not been investigated. Figure 1.12 shows the distribution (in terms of principal moments of inertia) of GDB-17 space together with a representation of the respective filters used to go from sets of graphs to actual molecular structures.
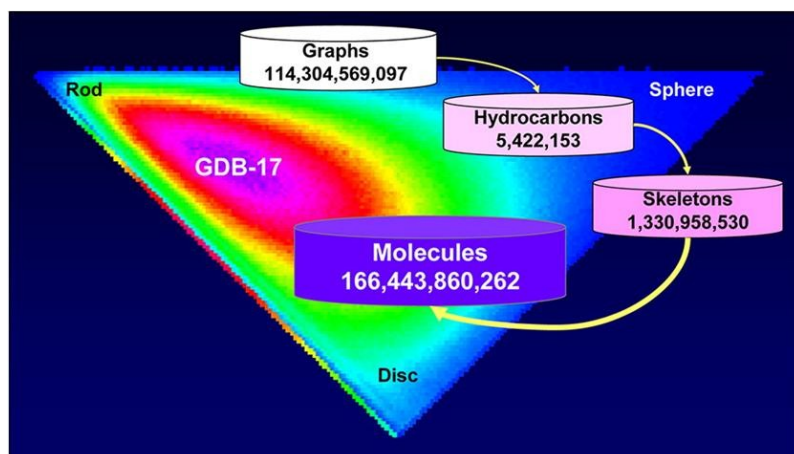
Other tools are available to the general researcher in order to facilitate the creation of virtual chemical libraries such as RDKit (which also has nodes available for use in KNIME). RDKit can be used to create libraries from fragments, or to calculate molecular properties (including fingerprints, for example).[71]. This software is one of several suitable for small molecular structure database creation.[72]

## 2. Chalcones: Powerful family with multiple therapeutic activities

### 2.1. Structures

In terms of natural products derived from plants, polyphenols represent a large family of molecules. These compounds are characterised by the presence at least one arene and free or engaged (ether, ester or heteroside) hydroxyl groups.[73] Polyphenols have their origin in the shikimate or acetate biosynthetic pathways (which we will address in the next section). Following the broad definition of what a polyphenol is, this encompasses a very large body of quite diverse molecules. The phenolic group possesses two important chemical/physical proprieties. The first one is the ease of cleavage of the hydroxyl bond (between the oxygen and the hydrogen). When this cleavage is heterolytic, we have an explanation for the acidic character of these molecules. On the other hand, if this cleavage is homolytic the result is that the phenol possesses reducing proprieties. This latter characteristic may be used as an antioxidant strategy.[74] Figure 1.13 presents the essential characteristics of phenols under reactive conditions.
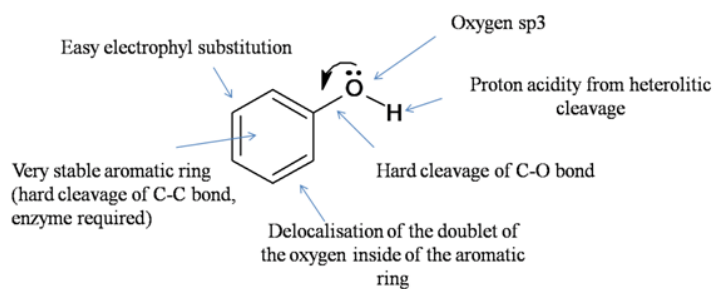
Figure 1.13: Strategic map of phenol reaction.

 The family of polyphenols may also be decomposed in structural subfamilies. These subfamilies include derivative phenols, and coumarin and flavonoid subcategories. Some chalcone derivatives may have the bridging oxy group that places them within the subfamily of flavonoids.

Chalcones include more than 700 identified natural products. They are characterised by the absence of a central heterocyclic ring (that flavonoids possess) which instead is replaced by a linker of three carbon including an α,β-unsaturated ketone moiety.[73] UV spectroscopy readily identifies the presence of this moiety.[75] Convention is that the aromatic or benzene ring closest to the ketone functionality is identified as ring A, while the other aromatic ring in proximity to the α,β- double bond is identified as ring B.[76] Figure 1.14 shows the simplest possible unsubstituted chalcone molecule.
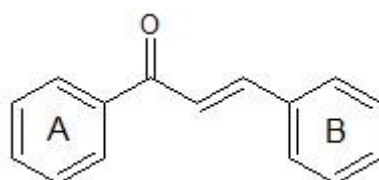


Figure 1.14: Chalcone structure.

## 2.2. Synthesis/Biosynthesis of chalcones

### 2.2.1. Natural production

#### 2.2.1.1. Shikimate pathways

##### 2.2.1.1.1. Common pathways

Exclusively produced by plants and micro-organisms, the biosynthesis of polyphenols follows two pathways. The first pathway, uses a precursor acetate to create only terpernophenolic and anthracenic molecules (*via* a Claisen condensation).[73] The second pathway, called the shikimate pathway is by far the predominant pathway in the production of natural polyphenol molecules.[77] Figure 1.15 represents the Shikimate pathway.
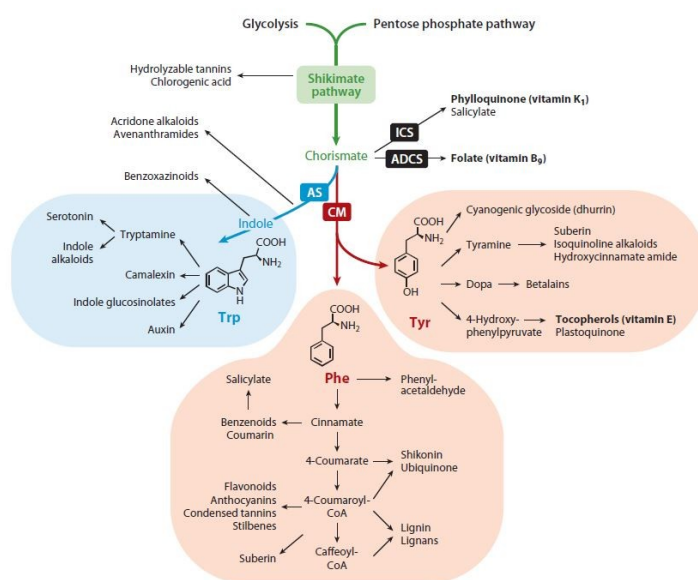


Figure 1.15: The aromatic amino acid pathway with compound. In green the shikimate pathways is represented which has as terminal compound the chorismate. The phenylalaline (Phe) and the tyrosine (Tyr) are developed in red with the chromate mutase (CM). The anthranilate synthase (AS) leads to the tryptophan pathways in blue. Molecules written in bold are corresponding to the essential nutriments in the human diet. Other abbreviation: ADCS, aminodeoxychrorismate synthase; CoA, coenzyme A; ICS, isochrismate (with the permission of the publisher). [77]

This process uses 2-phosphoenolpyruvate (PEP) and D-erythrose-4-phosphate (E4P) as precursor compounds. The pathway ultimately produces chorismic acid which may be transformed into three possible amino acids (tryptophan, tyrosine or phenylalanine). As such this pathway is crucial in the production of serotonin or vitamin E (Tocopherols), since both of these are produced from tryptophan and tyrosine. Phenylalanine from the shikimic acid

pathway may be used as precursor for the production of most polyphenols. In detail, this pathway takes as substrates PEP and E4P; an initial adol condensation produces 3-Deoxy-D-arabino-heptulosonate7-phosphate (DAHP) through DAHP synthetases.[78, 79] What follows is five consecutive steps in which DAHP is transformed into 3-dehydroquinic acid by the suppression of the phosphoester.[80] The first of these steps is alcohol reduction and β-elimination of the phosphate monoester group of the DAHP hemiketal form.[80] The reduction of this alcohol (which is in proximity to the ether group) leads to the ring opening of the reduced molecule. The final step is an intramolecular aldol condensation which closes the ring and produces 3-dehydroquinic acid. These steps are represented in the Figure 1.16.
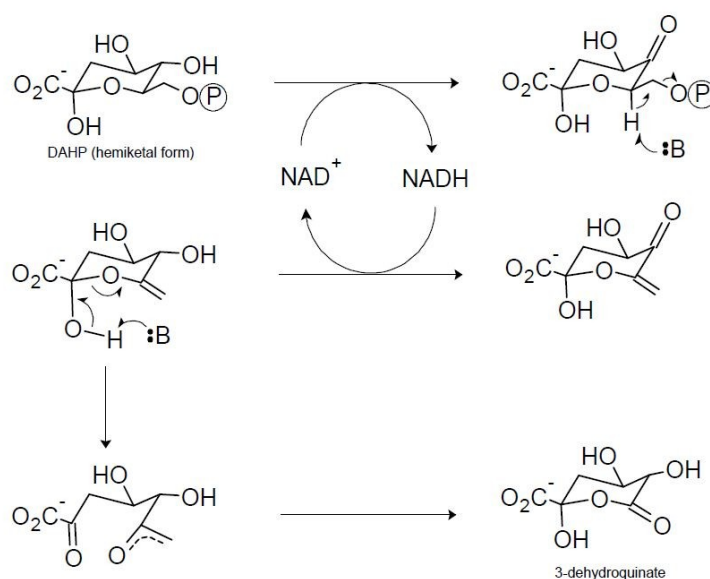


Figure 1.16: Reaction of formation for the 3-dehydroquinate from DAHP.

The next step in the shikimic acid pathway is dehydration of the 3-dehydroquinic acid (DHQ) in order to give the intermediate 3-dehydroshikimic acid (DHS) mediated by DHQ dehydrate. DHQ dehydrate is an enzyme with two forms: type I and type II with entirely different sequences and mechanisms, but identical function as a result of convergent evolution.[80] Type I follows a *syn* elimination reaction with a Schiff base and was discovered in *E. coli*. Studies of *Aspergillus nidulans*, containing the type II enzyme have been shown to follow an *anti* elimination.[81] Shikimate dehydrogenase uses NADPH to reduce the DHS into shikimic acid. The hydroxyl of the shikimic acid is phosphorylated into shikimic acid 3-phosphate through the shikimate kinase with adenosine triphosphate (ATP) as cosubstrate.[77] The 3-phoshoshikimate 1-carboxyvinyltransferase (also called EPSP synthase) performs the condensation of the previous molecule via addition-elimination to afford the enol ether 5-

enolpyruvylshikimate 3-phosphate (EPSP). The last reaction before the differentiation of the amino acid production is the 1,4-*anti*-elimination of the phosphoric acid group of the EPSP the chorismate synthetase to form chorismic acid. This reaction may only proceed with the cofactor (reduced flavin) donating one electron to the substrate. This pathway is represented by Figure 1.17.
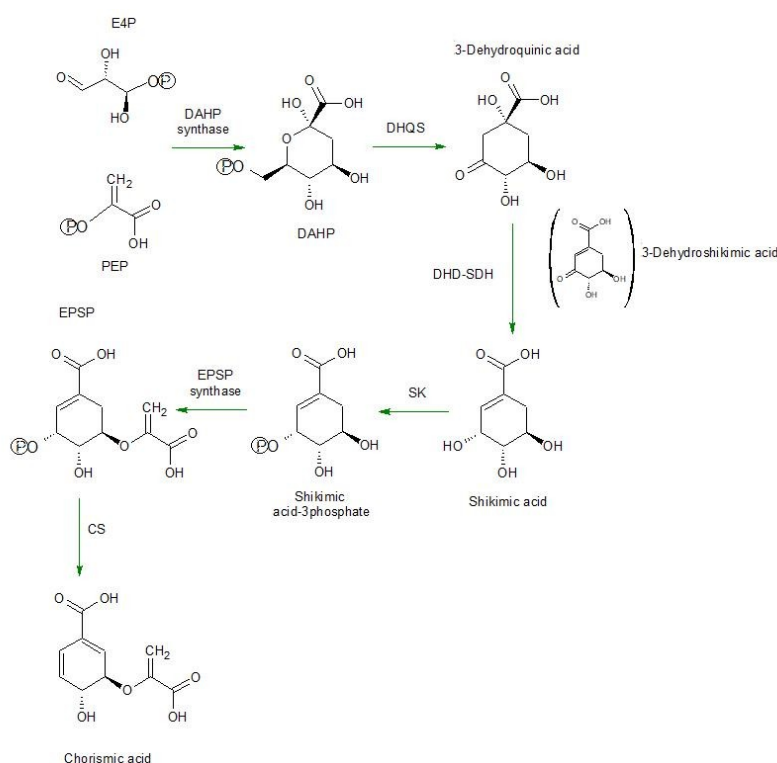


Figure 1.17: Natural pathways for chorismic acid. E4P: D-erythrose-4-phosphate, PEP: 2-phosphoenolpyruvate, DAHP: 3-Deoxy-D-arabino-heptulosonate7-phosphate, DHQS: dehydration of the 3-dehydroquinic acid synthetase, DHD: 3-dehydroquinate dehydratase, SDH: shikimate dehydrogenase, SK: shikimate kinase, EPSP: 5-*enol*pyruvylshikimate 3-phosphate, CS: chorismate synthase.

### 2.2.1.1.2. Phenylalanine pathway

Since the focus of this work is on chalcone derivatives, only the phenylalanine route in the shikimic acid pathway will be discussed. This pathway has recently been elucidated from gene encoding analysis of the responsible enzyme of this pathway.[77] In order to produce phenylalanine, chorismic acid from the shikimate pathway is converted to prephenic acid *via* a pericyclic Claisen rearrangement mediated by chorismate mutase (CM). At this stage the prephenic acid may be transformed by one of three different pathways. These three pathways produce respectively the amino acids tyrosine, phenylalanine, and arogenic acid. Arogenic

acid may also be an intermediate molecule in the production of tyrosine and phenylalanine. The longer pathway in the production of phenylalanine starts by the conversion of the prephenic acid into arogenic acid. This is mediated by prephenate aminotransferase (PPA-AT) using PLP as a cofactor. This conversion uses a reversible transamination between the substrate and L-glutamate and L-aspartate which act as amino group donors. It is important to notice the presence of the lysine amino acid in the active site of this enzyme which helps to recognise the dicarboxylic acid functionality of the substrate. As mentioned before, this arogenic acid can be transformed into phenylalanine or tyrosine. The route from arogenic acid to phenylalanine is *via* a decarboxylation and dehydratation involving arogenate dehydratase (ADT). This reaction is spontaneous under acidic conditions.[78]

However, the path to produce phenylalanine from the prephenic acid uses prephenate dehydratase (PDT) to form phenylpyruvic acid. This enzyme shares the same mechanism of reaction and the same conditions of action as ADT. The final step is a reversible transamination with PLP as cofactor. The transamination is only possible with an amine group donor and a keto acid acceptor from glutamic acid. Another enzyme that possesses the capability of converting phenylpyruvic acid into phenylalanine is the branched-chain acid aminotransferase called IlvE.

The regulation of all described syntheses is possible at different levels. The first of these is transcriptional regulation which depends on the species of the biological source. In most cases the expression of the gene of DAPH synthase will regulate the biosynthesis of these amino acids. The second way to regulate the production of these amino acids is during post-transcription. AS and CM have the same substrate (chorismate). As a result of this, they enter into competition with each other. Tryptophan, when its concentration is too high, will activate CM in order to produce tyrosine. Similarly, tyrosine will initiate the synthesis of phenylalanine *via* activating the ADT enzyme. The final means of control of production of these amino acids is the regulation of the supply of the E4P and PEP. Inside the Calvin cycle of photosynthesis, the transketolase transforms G3P and fructose-6-phosphate into xylose-5-phosphate and E4P. It has been shown that inhibition of this transketolase enzyme limits formation of E4P and that this inhibition is greater than other methods of regulation.[77] The origin of PEP, on the other hand is from the conversion within the scheme of plastidic glycolysis, mediated by the phosphoglyceromutase and enolase enzymes. PEP may also enter the plastid from the cytosol thanks to the PEP/phosphate translocator (PPT). PEP may also be formed in the plastid from pyruvic acid, mediated by plastidic pyruvate orthophosphate dikinase (PPDK). However these enzymes from PEP regulation are also involved in other

pathways such as the synthesis of fatty acids. PEP inhibition should not be considered as the primary method of regulation of the synthesis of these amino acids.[77]

All reactions described in the section are summarized in Figure 1.18.
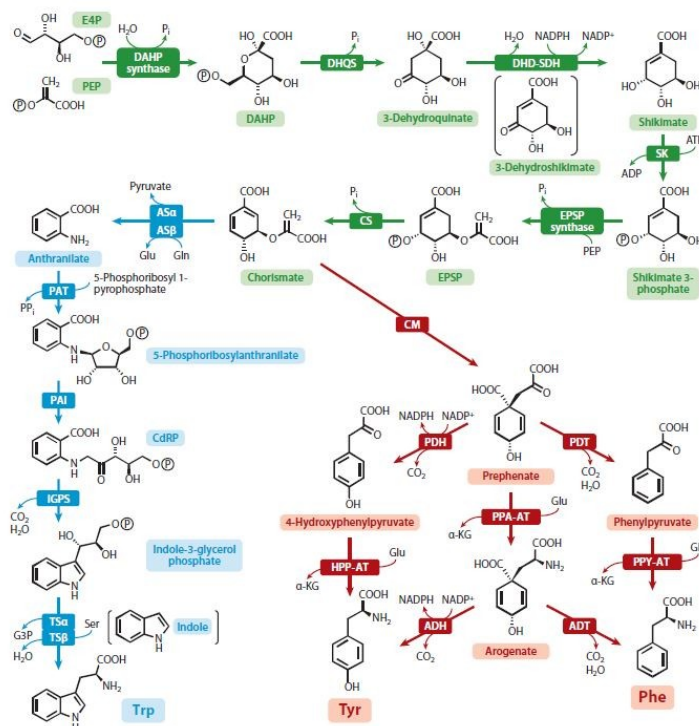


Figure 1.18: The full aromatic amino acid pathways in plants. The green path correspond to the shikimate way. (With the permission of the publisher). [77]

### 2.2.2. Biosynthesis of chalcones

The biosynthesis of chalcones uses the precursor L-phenylalanine; the biosynthesis of L-phenylalanine is fully described in the previous section. The L-phenylalanine will be de-aminated to form *trans*-cinnamic acid, mediated by the action of phenylalanine ammonia-lyase (PAL). This step is important because it will lead to two aromatic rings linked by 3 carbon atoms (a C6-C3-C6 structure) (Figure 18).[82] The *trans*-cinnamate 4-monooxygenase (tC4M) adds a hydroxyl group at the *para* position of the phenyl group, using cofactor NADPH, in order to form *p*-coumaric acid.[83] Coenzyme A (CoA) is transferred to the *p*-coumarate reductively resulting in the formation of 4-coumaroyl-CoA. The process is catalyzed by the enzyme 4-coumarate-CoA ligase (4-C-CoAL) involving the cofactor adenosine triphosphate (ATP). The next step uses naringenin-chalcone synthase (CHS) to perform a condensation of one molecule of 4-coumaroyl-CoA with 3 molecules of malonyl-CoA.[79, 81] This reaction is initiated by the transfer of the coumaroyl moieties of the 4-coumaroyl-CoA to the active site. The malonyl-CoA molecules are decarboxylated by the

28

enzyme forming carbanions derived from the acetyl-CoA and produced is carbon dioxide.[79] This final step of the reaction allows the extension of the alkyl chain of the diketide molecule. At the end it will form a tetraketide. The second ring of the chalcone is created by following a regiospecific intramolecular Claisen condensation.[84] Regulation of chalcone synthase is either through the products of this biosynthetic pathway such as the CoA ester, the resultant chalcone or naringenin in a noncompetitive way. The induction of this enzyme can be initiated in certain conditions such as UV light. The control of the CHS may also be achieved by the expression of sense and antisense genes.[84] Figure 1.19 and Table 1.3 represent the chalcone biosynthetic pathway and examples of natural chalcones.
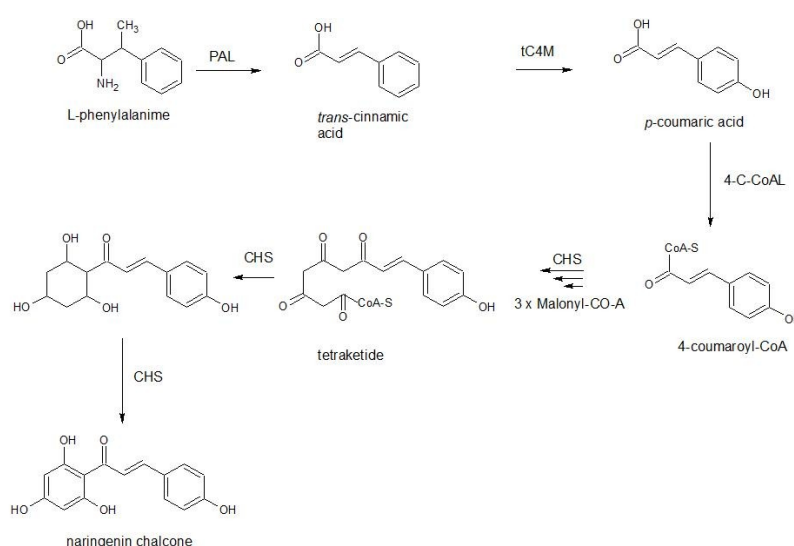


Figure 1.19: Chalcone natural synthesis. PAL: phenylalanine amino-lyase, tC4M: trans-cinnamate 4-monooxygenase, 4C-CoAL: 4-coumarate-CoA ligase, CHS: naringenin-chalcone synthase.
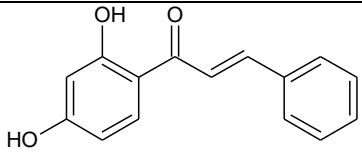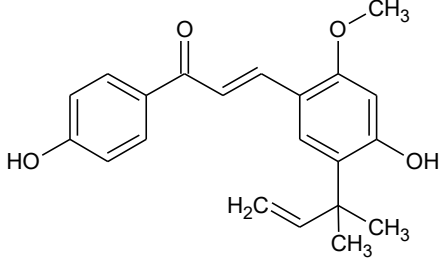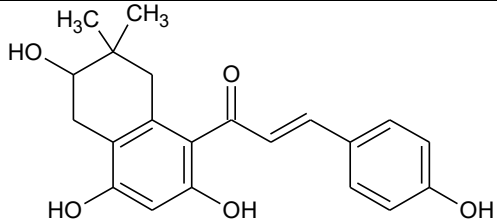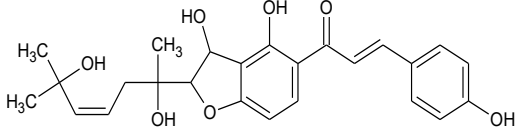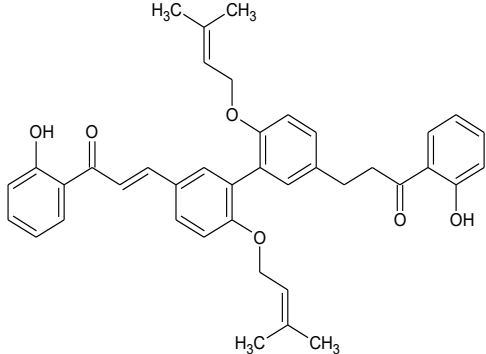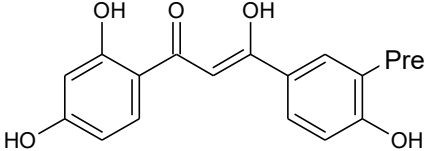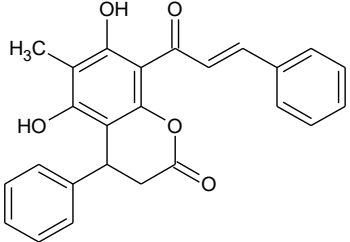
| Structure | Name | Plant | Activity |
|---|---|---|---|
|  | N.A | *Z.punctata* | antifungal |
| | | *M.calabura* | antibacterial |
|  | licochalcone A | *G.glabra* | antimicrobial, antioxidant, cell cycle modulator, antimalarial |
| | | *G.inflata* | antibacterial, anti-inflammatory, topoisomerase I inhibitor |
|  | bavachromanol | *P.corylifolia* | antiplasmodial |
|  | xanthokeismin | *A.keiskei* | superoxide scavenger |
|  | N.A | *G.lutea* | MAO inhibitor |
|  | glyinflanin A | *G.inflata* | HIV inhibitor |
| | glycyrdione A | | |
|  | Parasiticin A | *C.parasiticus* | cytotoxic |

Table 1.3: Examples of natural chalcones.[85]

### 2.2.3. Total synthesis

The synthesis of Chalcone derivatives total syntheses is often simple, requiring fewer reagents than many other natural product syntheses. Key to synthesis is the condensation of two aromatic reagents, one providing a nucleophilic group while the other provides an electrophilic group. [86] The most important synthetic pathways to chalcones are highlighted in this subsection.

### 2.2.3.1. Claisen-Schmidt reaction

The popular reaction for the formation of chalcones is based on the condensation between a benzaldehyde and an acetophenone. The A ring and B ring will be provided respectively by the acetophenone and the benzaldehyde.[87] Depending on the catalyst used, the reaction may be a normal condensation, a solid phase reaction or a solvent free one. The classic reaction conditions use with the benzaldehyde and acetophenone aqueous sodium hydroxide, potassium hydroxide or ethanolic sodium ethoxide at 50°C for several hours.[88, 89] The solid product is washed with 50% alcohol and finally dried. This process generally produces greater than 80% yield.[90] It is represented by the reaction 1 in the Figure 19.

The solid state reaction is slightly different. The acetophenone derivative used has to have a hydroxyl substitution which will be used to attach the acetophenone to the resin (Trityl chloride resin). Several solvents may be used such as dichloromethane or pyridine together with catalysts such as N,N-diisopylethylamine (DIEA). After attachment to the resin, the acetophenone derivative reacts with the benzaldehyde derivative in the classic manner. The chalcones is finally cleaved from the resin with trifluoroacetic acid.[87, 91]

The last reaction is related to the solvent-free Claisen-Schmidt reaction. Solvents normally used in non solvent-free reactions, as described above (including dichloromethane) are toxic and non eco-friendly.[92] Microwave irradiation may be used in a green chemistry approach avoiding the use of solvents. As an example, a mix of 2'-hydroxyacetophenone, benzaldehyde and anhydrous potassium carbonate is submitted to microwave irradiation for 3 to 5 min, yielding 2'-hydroxychalcones with yields between 80% to 90%.[88, 93] Two similar reactions have also been used. Instead of anhydrous potassium carbonate as catalyst, diiodine-aluminium has also been successfully used, reducing synthesis time to 80 seconds at 60°C. Perhaps this approach is less green, since ethyl acetate is required, and is added to the mixture prior to filtration, and washing of the product with sodium thiosulfate and water. Even after recrystallization from hot ethanol the yield is still in the range of 80% to 95%. [92] Another

variant is with the use 1,5,7-trisazabicyclo[4,4,0]decene (TBD) or pyridine or diisopropyl ethylamine and 2-chlorotrytilhychloride resin.[87, 91] The general reactions are represented in Figure 1.20.
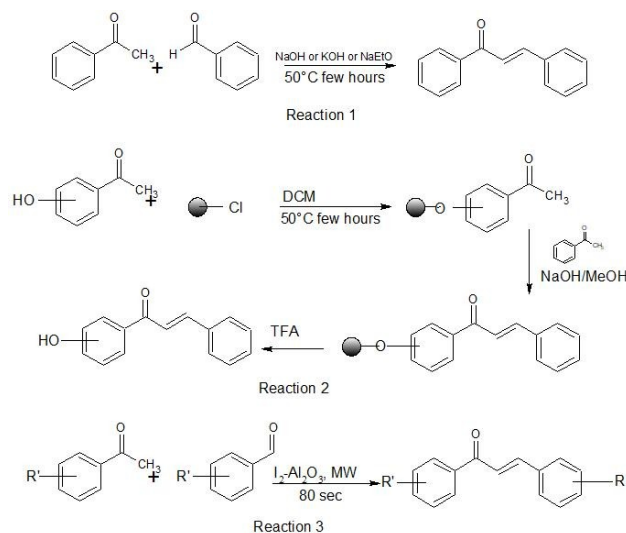


Figure 1.20: Claisen-Schmidt reactions. Reaction 1: the classic reaction. The second one is the solid phase and the last one represents solvent-free microwave synthesis.

### 2.2.3.2. Carbonylative Heck coupling reaction

The carbonylative Heck coupling reaction is an alternative to using acetophenone and a benzaldehyde as reagents; rather it uses an aryl halide together with styrene, catalyzed by palladium. The reaction requires the presence of carbon monoxide under pressure. It has been shown that aryl iodides and bromides lead to moderate to good yields (57-89%). Examples of conditions used include a pressure of carbon monoxide of 10 bar and a reaction temperature of 120°C.[94] Figure 1.21 represents this process.
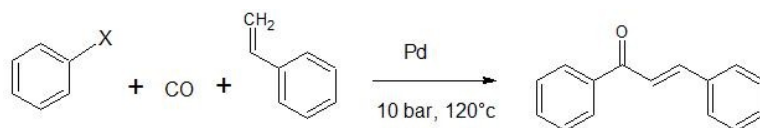


Figure 1.21: Carbonylative Heck coupling reaction.

## 2.2.3.3. Other coupling reactions

Another way to form chalcones *via* coupling is to use benzaldehyde and phenylacetylene derivatives within an ionic liquid environment, such as the use of the ionic liquid 1-butyl-3-methyl-1-H-imidazolium 4-methylbenzenesulfate (BmimOTs). Reaction times typically are 12 hours at 100°C.[86, 87] It is interesting that this coupling reactions leads to the *E* stereoisomer. Water added at the end of the reaction allows for the precipitation of products in generally high yield. Benzaldehyde derivatives with electron-donating (i.e.: chloro, bromo or nitro substituents) increases also the production of *E* stereoisomers. Some poor yields may be explained by the oxidation aliphatic aldehydes to carboxylic acids.[95] The coupling reaction in ionic liquid is represented in Figure 1.22.
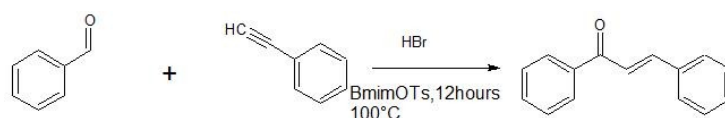


Figure 1.22: Coupling reaction.

## 2.2.3.4. Suzuki-Miyaura coupling reaction

This reaction involves the coupling of a benzoyl chloride derivative with a styrylboronic acid (pathway A), or between an arylboronic acid and a cinnamoyl chloride (pathway B).[86, 87] The second pathways presents a poor to moderate yield (between 41% and 51%) requiring the use of catalysts such as anhydrous toluene, $Pd(PPh_3)_4$ and $Cs_2CO_3$. Pathway B is more efficient with yields between 68 and 93%.[96] These pathways are shown in Figure 1.23.
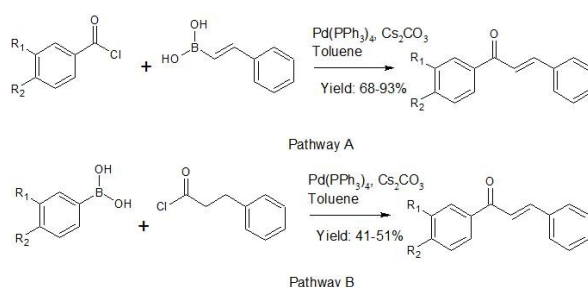


Figure 1.23: Suzuki-Miyaura coupling reactions.

## 2.2.3.5. Continuous-flow deuteration reaction

Deuterium, an isotope of hydrogen is commonly used in the pharmacological context. Isotopic labelling helps to track drugs and their metabolites inside the body. A chalcone,

labelled with deuterium has been synthesized from benzoyl chloride and phenylacetylene using PdCl$_2$(PPH$_3$)$_2$ and THF as catalyst and solvent. This reaction creates an ynone molecule, after which the reaction mixture is saturated with deuterium gas. Using a high pressure of deuterium and a temperature of 25°C good yields were obtained.[86, 97] This reaction is represented in Figure 1.24.
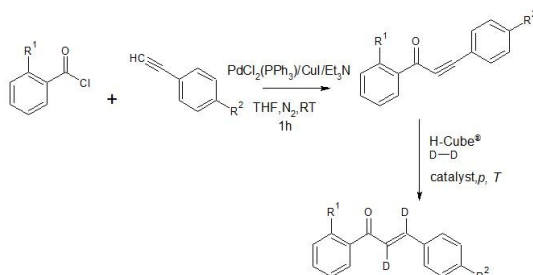


Figure 1.24: Continuous-flow deuteration reaction.

### 2.2.3.6. Sonogashira isomerisation coupling

Yet another coupling reaction uses the reaction of a phenyl halide with a substituted propargyl alcohol. This is catalyzed by PdCl$_2$(PPh$_3$)$_2$ under microwave conditions, using THF as solvent.[86, 87] Figure 1.25 represents this reaction.
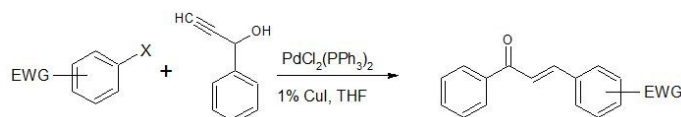


Figure 1.25: Sonogashira isomerisation coupling reaction.

### 2.2.3.7. One-pot synthesis

Researchers have reported a one pot synthesis of chalcones, consisting of successive reactions in a single reactor. This avoids the need to purify intermediate compounds as the reaction progresses.[86] Chromium trioxide was used in this case to oxidize phenylmethanol in benzaldehyde, which, in turn condensed with the acetophenone forming the chalcone.[87] This reaction is presented in Figure 1.26.
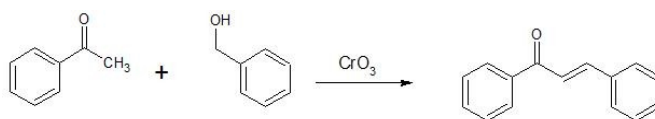


Figure 1.26 One-pot synthesis.

A final example uses a solid acid catalyst; unconventionally an ion-exchange resin was used as the heterogeneous catalyst.[86, 87, 98] Under microwave conditions, a chalcone was formed, as represented in Figure 1.27.
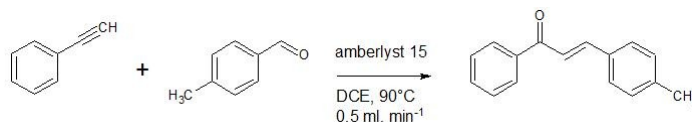


Figure 1.27: Solid acid catalyst mediated synthesis.

## 2.3. Pharmacological properties

Chalcones are well-known for their multiple and diverse therapeutic proprieties. However, in many cases their mechanism of action is not fully understood.[83] The range of properties is extensive. This includes *in vitro* and/or *in vivo* analgesic, anti-oxidant, anti-fungal, anti-bacterial, anti-protozoan, anti-mutative, anti-cancerous, anti-inflammatory and gastric protector properties. Some chalcones have a direct effect on certain pathogenic pathways. The actions of chalcones against specific enzymes in particular diseases will be discussed in the following section.

### 2.3.1. Anti-tumoral proprieties

Chalcone derivatives are known to work well against certain types of cancer (breast, lungs, colon leukaemia, melanoma, ovarian, renal prostate, Central Nervous System). [76, 99] They act on many different levels. These include effects on multi-drug resistant channels (efflux), or through inhibition of the vascular endothelial growth factor (VEGF) which is an angiogenesis stimulation signal protein. Some chalcones also inhibit important microtubule and topoisomerase (TOPO) proteins relating to mitosis, meiosis and the regulation and replication of DNA.[100]

### 2.3.2. Anti-inflammatory characteristics

Inflammation is a normal response to the presence of exogene, when this process is controlled by the body. Inflammation may also be as a response to various conditions such as neurodiseases or diabetes. Major actors of this inflammation are free radical species (oxides of nitrogen or oxygen) or enzymes.[76, 101] It has been demonstrated that some chalcone

derivatives inhibit cyclooxygenase (COX). This enzyme is responsible of the production prostaglandin and and thromboxane with arachidonic acid as substrate. [100, 102]

### 2.3.3. Anti-bacterial agents

Meningitis, tuberculosis, listeria, methicillin-resistant *Staphylococcus aureus* (MRSA) are examples of bacterial infections. At best patients with these diseases are invalids, at worst they can easily lead to death if they are not taken seriously in good time. Some studies have shown that some chalcones inhibit different enzymes responsible for the development and the proliferation of these bacteria. For example the tuberculosis uses the protein tyrosine phosphatase A and B to attenuate the host immune system. It has been demonstrated that inhibition by a chalcone leads to the death of the bacterial cells, aided by macrophages and interferon-γ.[87, 100] Other bacterial proteins are also inhibited by chalcones, including pyruvate kinase (aerobic condition) and lactate dehydrogenase (anaerobic condition). Essential to the bacteria cells, these proteins provide ATP. It has been reported that some chalcones prevent this activity in *S. aureus* (including MRSA), *Listeria monocytogen*, and *Salmonella typhimurium*.[76, 87] Finally, DNA gyrase is an enzyme responsible of the negative supercoiling or relieving strain in DNA. This enzyme is absent in humans. This makes it an ideal target for anti-bacterial drugs. Quinoline-derived chalcones show particular potential against this protein.

### 2.3.4. Anti-viral proprieties

A virus is an infectious agent which uses another organism's cell as a host to replicate. Human immunodeficiency virus (HIV), the groups of rhinovirus, the dengue virus and the many influenza viruses are all members of this family. HIV, for example, uses several proteins to infect the host cell and to use the host cell to aid in replication. It has been shown with *in-vitro* tests that there are chalcone derivatives that have anti-integrase activity for the HIV protein integrase. This enzyme is responsible to the integration of viral DNA into the DNA of the host nucleus. Furthermore other natural chalcones are good inhibitors of the HIV-protease enzyme which is responsible for cleavage of the *gag-pol* gene segment, and thus is essential for HIV replication.[76, 87, 103]

### 2.3.5. Anti-malarial agents

Malaria is a parasitic disease caused by *Plasmodium* species (including the *falciparum* and *vivax* strains) which is transmitted to humans by female *Anophele* mosquitoes. Test on some

chalcone derivatives against the more dangerous *P. falciparum* have shown that chalcones work very well in preventing degradation of the heme which is toxic for the parasite.[87] Furthermore it has been discovered that some chalcones are efficient against chloroquin resistant *P. falciparum* (Dd2 and W2) and the chloriquin-sensitive (D10) strains.[76, 87] Chalcone derivatives are considered as promising leads for anti-malarial treatment.

## 3. Paralytic enzymes for which chalcone inhibitors are known.

### 3.1. HIV-1 integrase

According to WHO reports, about 37 million people are currently infected by HIV with about 1 million deaths per year from this disease.[104] Despite a reduction in the number of infections (1.8 million people per year), it is still one of the most deadly and challenging diseases facing humanity. HIV is a retrovirus which can be divided in two strains, and the most prevalent of these is HIV-1. HIV-1 is constituted of 4 groups M, N, O, P, and is closely related to the Simian Immunity Virus.[105] Group M constitutes the most prevalent of infection cases. From this group, there are 9 subtypes confirmed (A, B, C, D, F, G, H, J, K) together with recombinant forms.

The HIV is a retrovirus which infects host cluster of differentiation (CD4) cells. The HIV life cycle is dependent on various enzymes. Firstly, reverse transcriptase (p66-p51) produces firstly a single-stranded $_v$DNA (viral) from the substrate $_v$RNA (viral). This genetic material is integrated into host DNA mediated by the integrase enzyme (p32). HIV protease is responsible for cleaving the viral polyprotein. [38]

Integrase is one of the targets in the treatment of HIV. Inhibition of integration of the viral $_v$DNA into the host DNA prevents the replication of the virus causing loss of activity.[106]

As mentioned before, the 32kDa integrase protein is responsible for holding the $_v$DNA and inserting it to host cell DNA. The mechanism of this process is divided into two actions: 3' processing and strand transfer.[107] The first action is to create a preintegration complex (PIC) with the viral $_v$DNA in the cytoplasm of the host cell. The integrase, aided by a divalent metal atom cofactor such as Manganese or Magnesium ($Mn^{2+}$ or $Mg^{2+}$) hydrolyses a pair of guanine and thymine nucleotides (GT) at the 3' end of the $_v$DNA. The divalent metal cation contributes to the stability of the PIC. By removing GT dinucleotide, it will uncover the end of the corresponding $CA_{OH}3$'(cytosine and adenine) of the $_v$DNA which will be incorporated into the DNA of the host.[108] The second action involves the actual $_v$DNA transfer into the host chromatid. Two mechanisms of transfer are proposed. The first involves

transesterification from viral to host DNA which creates phosphodiester bonds. The second involves a phosphoryl transfer *via* nucleophilic substitution ($S_N2$) pathway which inverts the chirality.[107]

Integrase is composed of three domains: the N-terminal domain (NTD) (residue 1 to 46), the catalytic core domain (CCD) (residue 56 to 186) and the C-terminal domain (CTD) (residue 195 to 288). A chain links NTD to CDT (residue 47 to 55) and a flexible loop connect CCD and CTD (residue 187 to 194).[108] The NTD and CCD are responsible of the activity of 3' processing and DNA strand transfer.[107] Inside NTD, pairs of histidine and cysteine residues are stabilized by $Zn^{2+}$ cations. The catalytic site of the CCD includes aspartic acid at positions 64 and 116 and glutamic acid 152 which link to the divalent metal ($Mg^{2+}$ or $Mn^{2+}$) cofactors. Mutation of these amino acids results in preventing this enzyme from fulfilling its function. The mutation of leucines 241 and 242 to alanine alters the structure CTD domain, preventing proper action. An allosteric site which may be targeted to inhibit integrase activities has been identified in CTD. [109] Figure 1.28 and Figure 1.29 represents respectively PDB structure of the integrase [4NYF] and the integrase domains.[110]
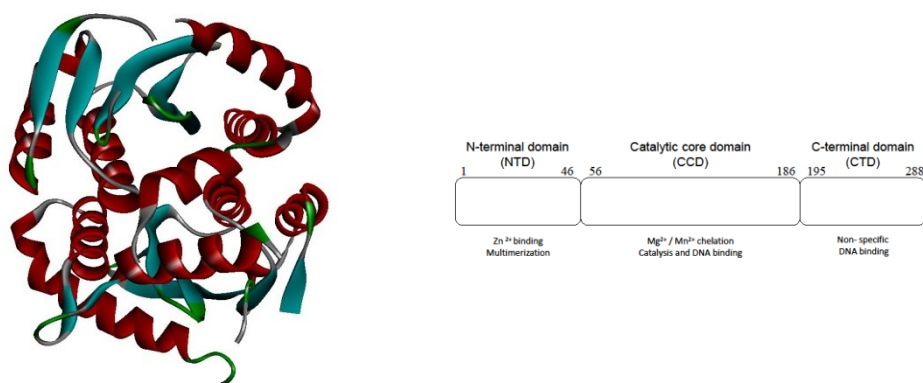


Figure 1.28 and 1.29: Integrase crystal structure [PDB ID: 4NYF] and domains.[110]

Chalcone derivatives have shown promising therapeutic effects for HIV-1, particularly against the integrase enzyme. Deng *et al* proposed a pharmacophore model from their substructure analysis and inhibition assays.[111] Vasu Babu *et al* have proposed potential inhibition of HIV-integrase by C-methyl chalcones derivatives through the use of docking experiments. In this particular case, the study was extended to include the synthesis of these compounds.[112] Another quantitative structure-activity relationship (QSAR) study was performed by Sharma *et al*, who synthesized and evaluated the inhibition activity of 3-keto salicylic acid chalcone derivatives.[113] Chalcone derivatives and pharmacophore examples from these studies are presented in Figure 1.30.

## 3.2. MRSA pyruvate kinase

Methicillin-resistance *Staphylococcus aureus* (MRSA) is a gram positive bacterial strain which is characterized by resistance to large spectrum antibiotics. These microorganisms are encountered hospitals and clinics and are spread by direct contact. If the MRSA-infection is not treated well and quickly, the disease may lead to sepsis and death. Considered as a major threat by the Center for Disease Control and Prevention (CDC), they particularly track MRSA infection epidemic outbreaks.[114] This strain resists not only all beta-lactam antibiotics (penicillins, carbapenems and cephalosporins), but also other classes of antibiotics (aminoglycoside, chloramphenicol, clindamycin, fluoroquinolones and macrolides).[115]

In 2011, Zoraghi *et al* presented as potential targets for MRSA treatment, the MRSA pyruvate kinase (PK) enzyme. This protein is essential to the organism for the production of pyruvate for the Krebs cycle. This enzyme catalyses irreversibly the phosphoenolpyruvate to pyruvate transformation, together with phosphorylation of adenosine diphosphosphate (ADP) to ATP. Inhibition of pyruvate kinase leads to interruption of the regulation of the carbohydrate metabolism and a reduction in available ATP, halting bacterial metabolism and killing the bacteria.[87, 116]

The pyruvate kinase is a homo-tetramer, with subunits being composed of N-terminal, A, B and C domains each with separate roles. The active site is located in the A and B domains while there is an allosteric site in the C domain. It appears that the C domain also contributes to the rigidity of the enzyme through salt-bridges. It is interesting to note that MRSA and human pyruvate kinases differ particularly in domain C.[116, 117] Figure 1.31 represents MRSA pyruvate kinase.
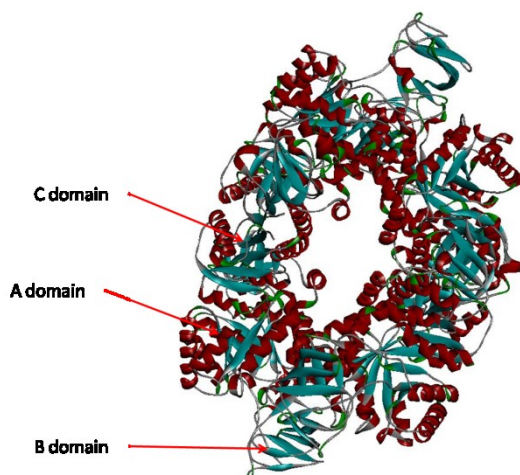
Figure 1.31: MRSA Pyruvate kinase with A,B and C domains [PDB ID: 3t05].[116]

Some studies have highlighted the anti-bacterial proprieties of chalcones especially against MRSA pyruvate kinase. Cushunie *et al*, through a structure-activity relationship, showed that the α-β unsaturated bond of chalcones is critical in its activity against the bacteria. It was also shown that a hydroxyl group in the ring A is also essential for this activity, and particularly good activity is observed when this is a *para*-hydroxyl group. Some *para*-substituents on ring B improve the activity. Moving to the *meta-* and to the *ortho-* positions, trifluoromethyl or bromo groups on ring B steadily increase the antibacterial activity. Chalcones may also act against MRSA by reducing the fluidity of the cytoplasmic membrane, or inhibiting DNA synthesis *via* an inhibition of the topoisomerase enzyme. Finally chalcones are known to inhibit energy catabolism, by inhibiting the NADH-cytochrome c reductase enzyme.[118, 119]

Osoriò *et al* tested chalcones derivatives against 14 different strains of MRSA. They proved these chalcones to be inactive MRSA.[115] Labrière *et al* prepared chalcone derivatives from the adol condensation between 2, 3, 4, 9–tetrahydro-1*H*-carbazol-1-one and some aldehydes derivatives. *In vitro* assays of these compounds showed inhibitory activity against pyruvate kinase from several MRSA strains.[120] Some of these chalcones are presented in Figure 1.32.



Chalcones with the best activity from Osòrio et al [115]
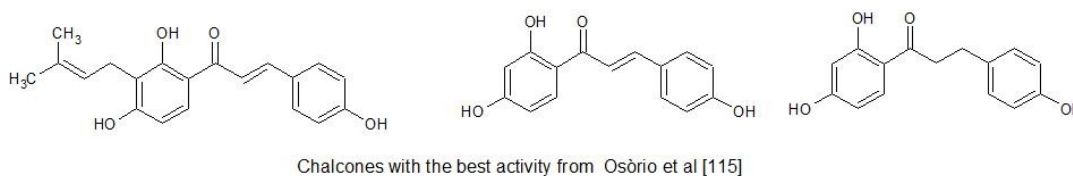
Figure 1.32: Examples of chalcones with an anti-PK activity.[115]

## 3.3. Heat-shock protein 90 (HSP90)

Following the International Agency for the Research on Cancer, cancers were responsible for 8.3 million deaths in 2012.[121] Cancers, by their nature and diversity may be highly complex and extremely difficult to treat. There are several targets which may be used for potential treatments.

A promising enzyme which is responsible for cancer proliferation is the heat-shock protein 90 (HSP90). HSP90 is a 90 kDa, responsible for the protection of cells against cellular stress (heat, radiation, etc.).[122] There are 20 enzymes known to work with HSP90 as co-chaperons.[123] HSP90 interacts with over 200 client proteins assisting with their folding and subsequent stabilization. These 200 client proteins include the following pro-cancerous proteins: mutant proto-oncogenes cKit, Her2, mutant epidermal growth factor receptor (EGFR), fusion-genes Breakpoint Cluster Region-Abelson (BCR-ABL), steroid hormone receptor c-MET and the serine-threonine protein kinase (AKT/PKB). The HSP90 therefore supports the onco-protein functions such as angiogenesis, anti-apoptosis and metastasis activities.[124] The protein HSP90 is overexpressed in malignant cells.

The HSP90 is a homodimer protein, divided into three domains. The role of its N-terminal is to bind to ATP. It also regulates the client substrates through hydrolysis.[124] The second domain is the M-domain (middle) and is responsible for the recognition of substrates and binding, and is also responsible for hydrolysis of ATP and the regulation of co-chaperones. The C-terminal is involved in the formation of the homodimer, and also includes an ATP binding site.[123, 124]

The mechanism of action is reliant on the changing conformation of HSP90. HSP90 has an open conformation ("V position") when it is inactive. The ATP binding sets off a change of conformation from open conformation to closed one which hydrolyzes the ATP (and the change is through several intermediate conformations).[122] The N-terminal is then responsible for removal of ADP, returning the homodimer to its original conformation.[123] Figure 1.33 shows a monomer of the HSP90 crystal structure and the open and closed conformations of the HSP90 enzyme.
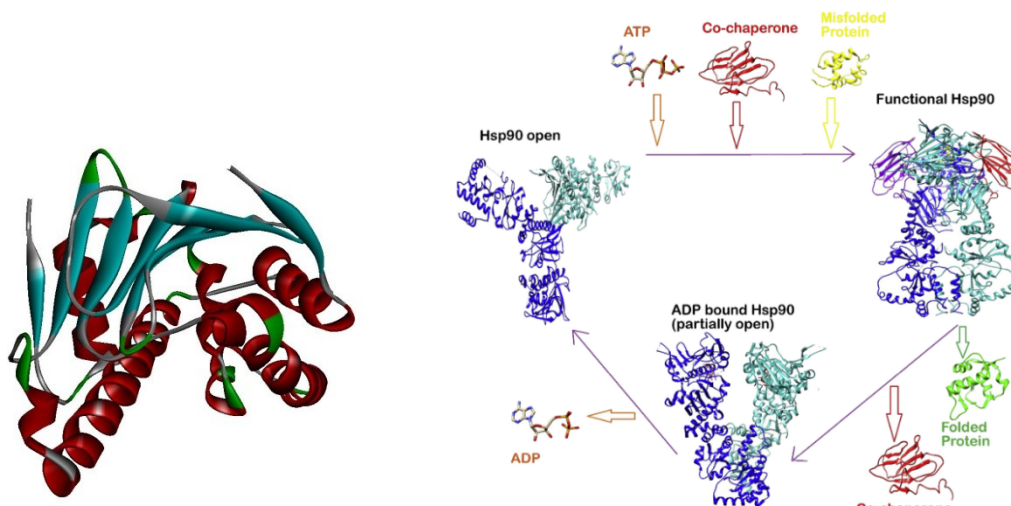
The chalcone family shows interesting inhibitory activity against the HSP90. A high-throughput screen using a natural product library was performed by Davenport *et al* in 2014.[126] This library included chalcones which inhibitory activity against the HSP90 targets. Curcumin is an example from this set that inhibits this protein.

Another study presents natural chalcones as indirect HSP90 inhibitors. Flavokawain B induces the degradation of HSP90 client proteins such as EGFR, Her2 or Met and disrupts the action of the enzyme as a consequence. Licochalcone A acts as an inhibitor in the same way; this compound is present in liquorice root. Docking experiments show that the chalcone licochalcone binds to the ATP site of the enzyme.[127]

(E)-1-(2-methoxyphenyl)-3-(2-nitrophenyl)prop-2-en-1-one was tested *in vitro* against prostate cancer cell line LNCaP and was shown to inhibit HSP90 activity by blocking the androgen receptor-HSP90 complex.[128]

Jeong *et al* designed and tested, through both *in silico* docking and *in vitro* assays synthetic chalcones against HSP90 from gefitinib (anti-cancerous drug) resistant non-small cell lung cancer cell line (H1975). One of their compounds **1g** (Figure 1.34) presents a high potency with respect to stopping proliferation of the cancerous cells. This molecule blocks ATP access to the active site on the N-terminal of HSP90.[129]  In 2017 **1f** (Figure 1.34) was shown to have  moderate inhibitory activity against HSP90 in H1975 line cells.[130] Figure 1.34 shows all of these chalcones mentioned which present HSP90 inhibitory activity.

Curcumin [126]  Flavokawain [127]  Licochalcone A [127]

(E)-1-(2-methoxyphenyl)-3-(2-nitropheny l)prop-2-en-1-one [128]  1g [128]  1f [129]
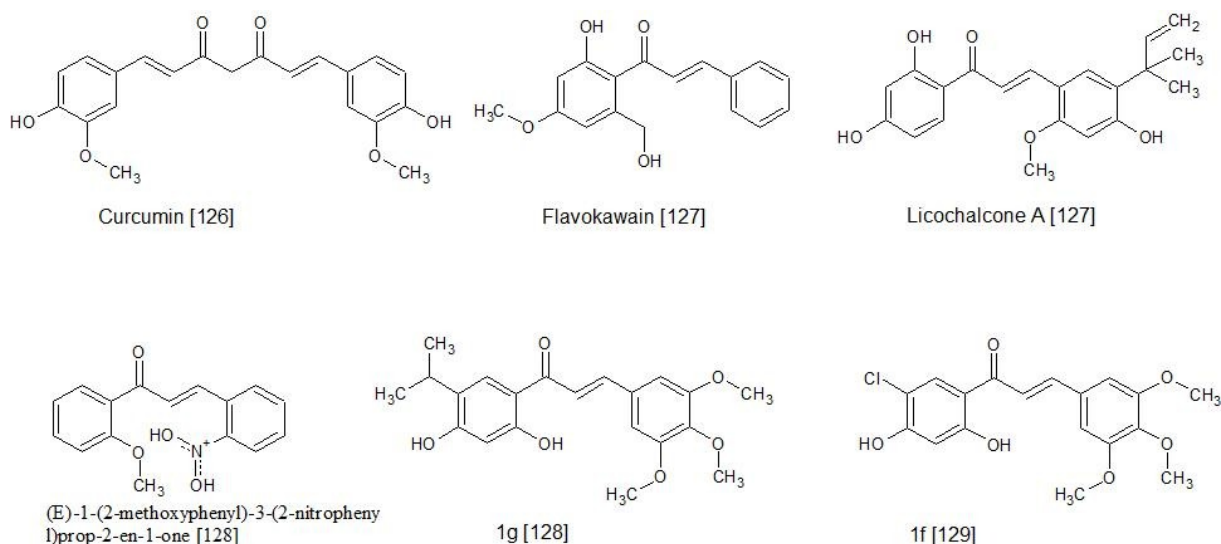
Figure 1.34: Chalcone which presents an inhibition activity against HSP90.[126-129]

## 3.4. Cyclooxygenases (COX)

Other proteins are responsible for proliferation of cancer. It has been shown that cyclooxygenase-2 is overexpressed in malignant cells in lung (90%), colon (71%) and breast cancer (38%).[131] Cyclooxygenases are proteins which convert arachidonic acid into prostaglandin $H_2$ (PGH$_2$) through oxidation, and through a cascade prostaglandin $E_2$ (PGE$_2$). PGE$_2$ (depending on the situation) either promotes or inhibits inflammation and homeostatic action which may lead to cardiovascular disease or cancer. [132-134] COXs are 70-kDa homodimers which possess a heme molecule. Their structures are divided into three subunits. The first domain is the epidermal growth factor, the second is a membrane-binding one and the last one is the catalytic domain. [135]

COXs are divided in two isoenzymes: COX-1 and COX-2. COX-1 shares 63% of homology with COX-2. The catalytic subunit is mostly conserved and the differences are concentrated in the N- and C terminal domains.[136, 137] Within the active site, there is a single amino acid that is different (V509I, COX-1 has V, COX-2 has I). COX-1 is expressed in mammalian tissues, particularly in stomach and kidney cells. The function of COX-1 is physioprotective and includes cytoprotection of the gastric epithelium, regulation of renal blood flow and control of platelet aggregation (through production of thromboxane).[131, 138]

COX-2 expression is induced by bacterial lipopolysaccharides, growth factors, cytokines phorbol ester and NF-κB nuclear factors.[132, 138] The overexpression of COX-2 prevents apoptosis of abnormal cells and encourages them to mutate, and hence promote metastatic properties. It has been shown that the level of transforming growth factor beta-2 (TGF β-2)

43

receptor and E-cadherin decrease, while the level of Bcl-2 expression increases due to the increased expression of COX-2. The second effect of this upregulation is an increased prevalence of angiogenesis and metastasis. The inflammatory factor NF-κB plays a critical role in the previous steps. There are two regions located on the 5' promoter which may bind to NF-κB. Through this binding action, there is positive activation of NF-κB. Nitric oxide and peroxynitrite (when their levels are low) are also involved in the regulation of the COX-2. [138]

Figure 1.35 presents a 3D representation of COX-1 and COX-2.
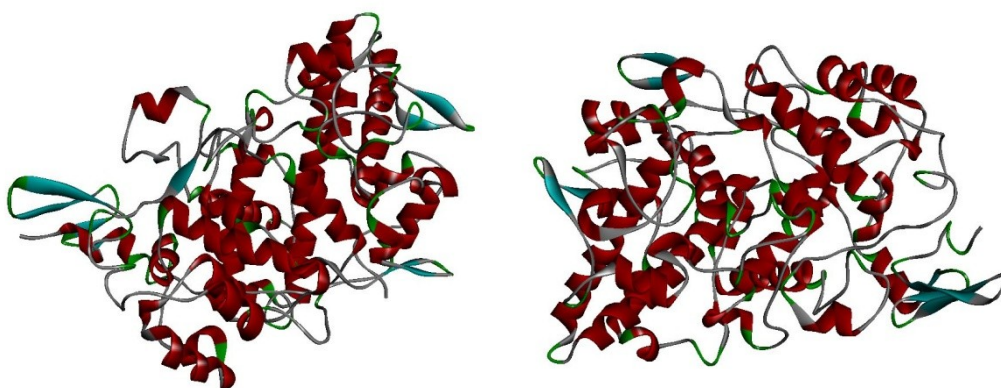


Figure 1.35: COX-1 (left)[PDB ID: 3N8Z] and COX-2 (right) [PDB ID: 5F19] single crystal structure.[134, 139]

The non-selective inhibition of the cyclooxygenases is problematic because it leads to side effects, for instance undesirable inhibition of COX-1 could lead to stomach haemorrhaging. A classic example of a class of drugs which do not make this distinction are the non steroidal anti inflammatory drugs (NSAIDs). The administration of these drugs may lead to stomach haemorrhage if overused.

Herencia *et al* have shown *in vivo* and *in vitro* anti-inflammatory activity of four chalcones derivatives which have in common a 2-choloroquinolinyl moiety in the ring B of the chalcone. Furthermore these molecules are selective to COX-2 and do not inhibit COX-1.[140] Herencia, as a result of this designed three new chalcone derivatives using the previous described moiety in two of these, with the final chalcone being entirely original. After *in vivo* and *in vitro* assays, researchers discovered that these molecules inhibited the generation of $PGE_2$ but not directly from inhibition of cyclooxygenases.[141]

Bandgar *et al* synthesized and evaluated the inhibitory activity of chalcone derivatives containing nitrogen atoms, including action against COX-1 and COX-2. They showed selectivity of inhibition of COX-2 relative to COX-1 for six systems.[142]

It has also been shown from some chalcone derivatives that certain moieties such as *para*-MeSO₂NH and NH₃ contribute to the activity against the cyclooxygenase enzyme. These molecules were synthesized using a Claisen condensation and *in silico* and *in vitro* tests successfully presented inhibition of COX-2.[143] Both of these studies provide compounds that are inhibitors of the COX-2 enzyme. [142-143]

Examples of these mentioned chalcone derivatives, that are active against COX-2 are represented in Figure 1.36.



Molecules 4 and 6 [140]

Molecules 7b and 3b [142]

Molecules 7b and 7f [143]
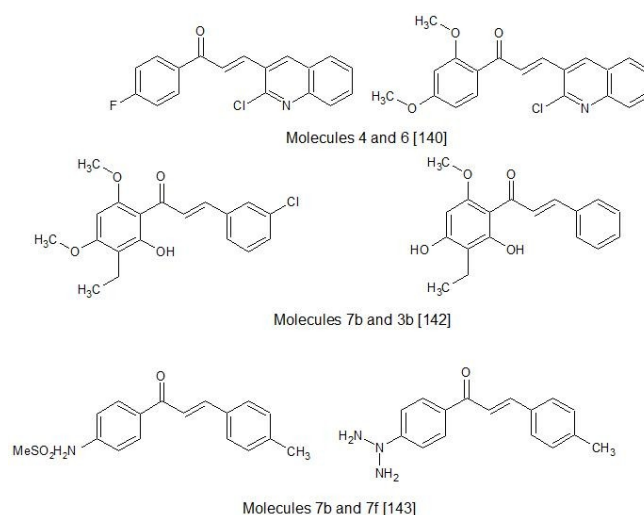
Figure 1.36: active chalcones against COX-2.[140, 142, 143]

## 3.5. Monoamine oxidases (MAOs)

Monoamine oxidases are flavoproteins located in the mitochondrial membrane. They possess flavin adenine dinucleotide (FAD) cofactor which is covalently bonded to a cysteine residue. MAOs encompass two isoenzymes MAOA and MAOB, which have 70% sequence similarity.[144] They coexist in most tissues except for the placenta where MAOA is predominant, and platelets and lymphocytes where MOAB predominates. In the brain MAOA is highly expressed in catecholaminergic neurons, while MAOB is expressed in astrocytes and serotonergic neurons. [145] Each of the different MAOs has specific substrates. The substrates of MAOA are epinephrine, norepinephrine (NE), dopamine (DA) and serotonin (5-HT).[145, 146] Phenylethylamine and benzylamine are degraded by MAOB. Either MAO catalyses transformation of these substrates *via* an oxidative deamination process. This may lead to a undesirable products such as 1-methyl-4phenyl-pyridinium (a Parkinson-producing neurotoxin from 1-methyl-4-phenyl-1,2,3,6-tetrapyridine) and hydrogen peroxide (which is one of the origins of harmful free radicals).[144, 146]

Depending on the enzyme, dysfunction leads to separate diseases. Aggressive behaviour has been observed in rats when the level of NE, DA and 5-HT is abnormally high due to the dysfunction of MAOA.[146] Dysfunction in MOAB, on the other hand is related to neurodiseases such as schizophrenia, alcoholism, affective disorder, depression, Alzheimer's disease and Parkinson's disease.[145, 147]

Both MAOA and MAOB exist as dimers. MAOs may be divided into three domains: the transmembrane region, the FAD-binding domain and the substrate-binding domain. The transmembrane domain is described as a helical structure and represents the C-terminal of the enzyme which passes the mitochondrial outer-membrane. The reason for binding to the mitochondria is still not well understood.[144] The transmembrane helix has apolar proprieties which may explain its ability penetrate the membrane of the mitochondria.[148] The MAOA transmembrane domain is longer than the MAOB one. The second domain presented is the FAD-binding site region. Critical to the structure of the MAO, the FAD cofactor has to bind within each protein in a dedicated region surrounded by four parallel β sheets, an α–helix and three antiparallel β-sheets.[146] FAD is always covalently bound to a cysteine residue. The active site is different for the MAOA and MAOB. The active site of MAOA is a single substrate hydrophobic cavity for which the volume corresponds to about 500 $Å^3$. The MAOB active site is linked to FAD-binding site which explains the larger active site volumes (approximately 700 $Å^3$).[149] In most cases, ligands or substrate binds to the FAD (through covalent bonds or non-covalent interactions). FAD inside MAOB adopts a non-planar conformation. [148]

Crystal structures of MAOA and MAOB are presented in Figure 1.37.
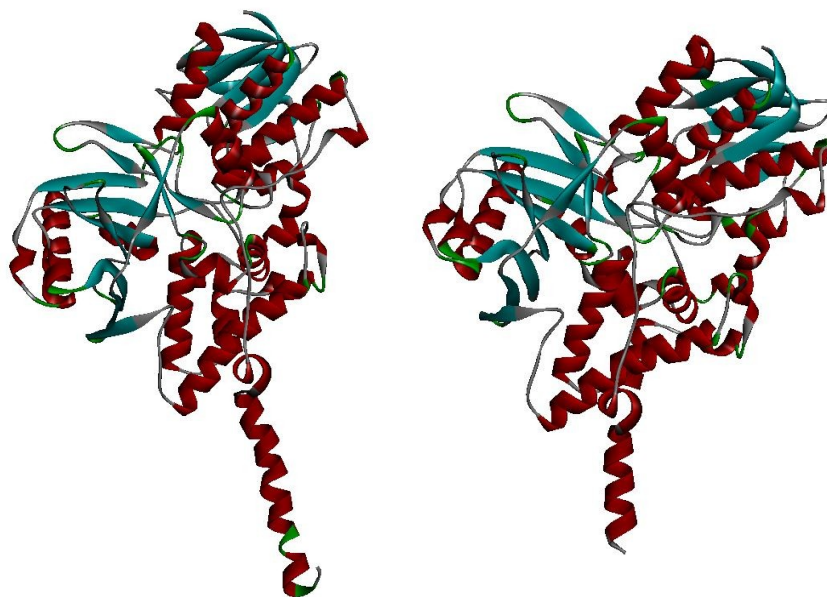
Some studies have identified chalcones as inhibitors of MAOs. Chimenti *et al*, after measuring the production of hydrogen peroxide by human MAOs, found that two naturally occurring chalcones are able to selectively and irreversibly inhibit human MAOB. They also identified that the presence of the hydroxy or methoxy in 2' and 4- position in ring B of the chalcone (which is close to the FAD molecule) and a chloro group in 4-position in ring A is necessary for potency. [151]

Morales-Camilo *et al* performed a study with rat-MAOB, in which they synthesised several chalcone derivatives before evaluation using an *in silico* assay, a QSAR study and some *in vitro* testing. The result was that the chalcones synthesized inhibited selectively while having no activity against MAOA. In this study the presence of a methoxy group in the ring A appeared to be important.[152]

Chalcones derivatives which inhibit MAOB are presented in Figure 1.38.

## 3.6. Hydrolases

### 3.6.1. Acetylcholinesterase (AChE)

Acetylcholinesterase is an enzyme which hydrolyses the cationic neurotransmitter, acetylcholine into acetate and choline. Mainly found in the brain, it is closely associated with the membrane of the synapse and regulates the transmission of acetylcholine to the acetylcholine receptors. Non-regulation of acetylcholine may lead to diverse range of diseases such as Alzheimer's disease. The first inhibitors of acetylcholinesterase were not developed for therapeutic purposes but rather to incapacitate or kill people. Sarin and soman are examples of such inhibitors, used in chemical warfare.[153] Their mechanism of inhibition is the creation a covalent bond within the catalytic site.[154]   On the other hand, AChEAChE is also used as a therapeutic target for the treatment of glaucoma, *myasthenia gravi*s and neuron degeneration.[153]

AChEAChE possesses multistructural forms of catalytic subunits: monomer, dimer and tetramer. The tetramer (20 kDa) is predominant in the Central Nervous System (CNS).[155] Two pairs of subunits, linked by a disulfide bridge near to the C-terminus form a subdomain, and are also anchored to the membrane. The AChE contains an active site which is divided in two subsites. Of these, the esteratic subsite is where hydrolysis of acetylcholine occurs. The acetyl portion of acetylcholine binds to this position. The second subsite is a charged anionic subsite. This is the location where the quaternium nitrogen from the choline portion of acetylcholine is bound. There is, in addition, an aromatic cation site in proximity to the subsites which binds the hydrophobic alkyl chain of the substrate. There is also a peripheral anionic subunit which is not directly connected to the active site. [156] Figure 1.39 shows the crystal structure of an AChE protein.
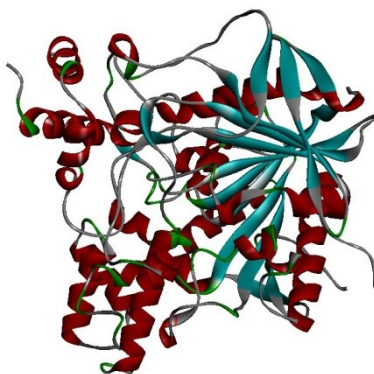
Liu *et al* used a chalcone natural product, flavokawain B as a base for synthesis of derivatives using a Mannich base. *In vitro* inhibition, kinetic and docking studies shows the potential of this chalcone family in inhibiting the AChE enzyme. Docking of the **compound 8** (Figure 1.40) shows that it may straddle both the peripheral anionic subsite and the aromatic cation site. This compound also appears to be able to bind both the unoccupied and occupied enzyme. [157]

Liu *et al* did another study using different chalcone derivatives of chalcone, using assays together with logP study (using the shake flask method). Their intent was also to observe the potential liposolubility of these chalcones in terms of then potential to cross the blood-brain barrier. All chalcones used in this study were able to do so. They also reported the selectivity of their library for the acetylcholinesterase over butyrylcholinesterase.[158] With further modifications of chalcones, with halogen substituents they were able to improve the action of the chalcones as AChE inhibitors. One compound (3i, Figure 1.40) shows both a strong inhibition and a high selectivity against AChE.[159]

Sukumaran *et al* synthesized 2'-hydroxychalcone derivatives and also tested them kinetic, *in vitro* and *in silico* assays against AChE. They report that chalcones which include a methoxy group on ring A and a halogen atom on ring B are the most efficient inhibitors of AChE from their library.[160]

Figure 1.40 represents few examples of chalcones derivate previously described.
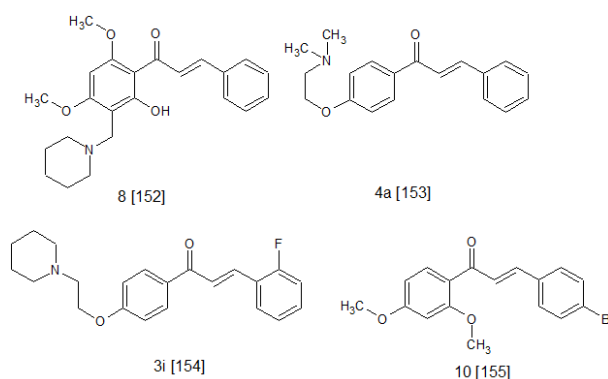
### 3.6.2. Butyrylcholinesterase (BChE)

The second hydrolase enzyme discussed is butyrylcholinesterase. This protein may be found in large quantity in plasma.[161] Although it is similar to AChE (four subunits linked at the end of the C-terminal), there are some differences present in BChE. The tetramer is composed of four subunits which surround a polyproline-rich peptide. Some amino acid residues inside of the acyl-binding site are different and the conformation of the active site is not the same.[161]

Its purpose is detoxification of eaten or inhaled exogens. Toxins are hydrolyzed in order to be inactive. Thus BChE may be used to protect against neurotoxic agents (such as sarin). The enzyme able to react fast to the presence of organophosphorus compounds by creating a covalent (stable) bond to them. In another example of protection, the enzyme converts the aspirin to salicylic acid or heroin to morphine. This may mediate cocaine intoxication and furthermore its addiction to some extent. BChE is also responsible of the acetylcholine hydrolysis.[162] Figure 1.41 represents the crystal structure of BChE.



Figure 1.41: Crystal structure of butyrylcholinesterase [PDB ID: 5DYW].[163]

50

It is interesting to note that, as described in the discussion of AChE, that the potential of chalcones appears to be selective in AChE over BChE.

## 3.7. Aldose reductase (ALR$_2$)

Aldose reductase is a member of the enzyme superfamily of aldo-keto reductases.[164] This enzyme catalyzes the conversion of aldose (and other different types of aldehydes) to their corresponding alcohols which is dependent on NADPH.[165] This enzyme is found in high concentration in the liver and kidneys. [166] By converting glucose to sorbitol (another sugar which hyperosmotic) ARL$_2$ is responsible for the proliferation of diabetic complications such as diabetic retinopathy, nephropathy and neuropathy.[164, 165] The enzyme is composed of 315 residues and includes a $\beta_8/\alpha_8$ barrel fold. Its mechanism of action is characterized by the linking to and removing of NADPH from ALR$_2$, occurring at the onset and right at the end of bioreaction respectively.[167] The cofactor NADPH is received into a V-shaped site and forced into a specific conformation to allow the enzyme to open and close its active site.[168] The active site is located close to the C-terminal of the $\beta$-barrel.[165, 169] Figure 1.42 represents a crystal structure of the aldose reductase enzyme.
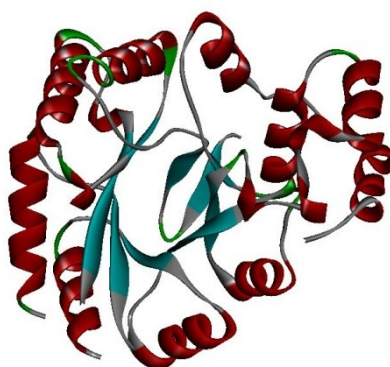


Figure 1.42: Aldose reductase crystal structure [PDB ID: 5JH1]. [170]

Severi *et al* have synthesized chalcone derivatives and have performed *in vitro* assays after filtering out compounds presenting a high liposolubility characteristic. From this study, they identified the importance two hydroxy moieties in 2' and 4' positions on the A ring of the chalcone for good inhibition, or, alternately, the presence of a carboxylic group on this ring.[171]

Iwata *et al* isolated some natural chalcones such as Licochalcone A and echnatin and synthesized other chalcone derivates. They confirmed that hydroxyl moieties in the position 2'and 4' in the ring A are crucial to good binding. To this, they further observed that substituents in position 2 on ring B also have a large effect.[172]

A review by Mahapatra *et al* compiled information known about chalcones as inhibitors of ALR$_2$. They were able to create a structure-activity relationship in response to this information. This further confirmed the importance of hydroxyl groups on ring A. They also showed that a thioglycolic acid substituent further increased the inhibition of the aldose reductase. On the other hand they observed that modification of the unsaturated bond, or replacement of the rings A and/or B by heterocycles decreased the activity of molecules against ALR$_2$. [173]

Two of these chalcone derivative inhibitors of ALR$_2$ are presented in Figure 1.43.



24 [171]        2', 4'-dihydroxy-4-methoxychalcone [172]
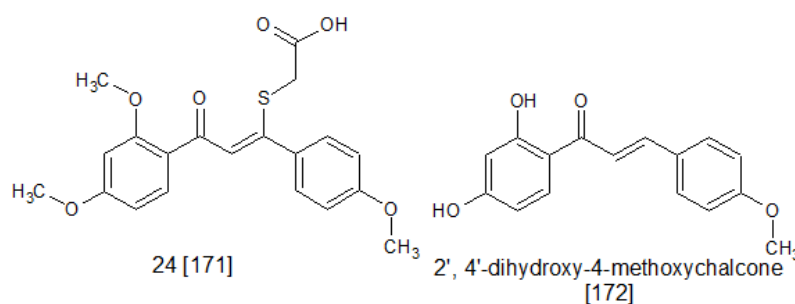
Figure 1.43: Examples of chalcones derivatives, inhibitor of ALR$_2$.[171, 172]

## 3.9. Lipoprotein-associated phospholipase A$_2$ (Lp-PLA$_2$ or PAF-AH)

The family of phospholipases is a superfamily which is composed 15 different groups and many more subgroups of enzymes.[174] To be classed in this family, the enzyme has to meet the following requirements before division into the appropriate subgroup. The enzyme has to present hydrolysis activity against the fatty acids (arachidonic acid or lyso LP) from the *sn*-2 ester bond of the phospholipid (shown in Figure 32). The enzyme may then have secondary activities. The second criterion is that the full sequence of the protein has to be known before this categorization. Homologous enzymes are separated with a subgroup letter (for example: IA from cobra or krait venom and IB from mammal pancreas). The same enzymes from different species share the same letter (for example: IIB shares rat and mouse testis sources). The last criterion is if the same PLA$_2$ gene from the same subgroup produces a splice variant enzyme, this is indicated by an Arabic number added to the end of the classification (e.g.

VIA-1 and VIA-2).[175, 176] PLA$_2$ enzymes are also divided in three main clusters. The first cluster uses a cation Ca$^{2+}$ and possesses an active site of a histidine and aspartate pair. The second one uses a catalytic serine and does not need the presence of Ca$^{2+}$ to be functional.[175] Through its large diversity, the superfamily of phospholipase A$_2$ is responsible for many diseases including asthma, arthritis or cancer.[177] Figure 1.44 represents the mechanism of action of the PLA$_2$ on a fatty acid substrate.
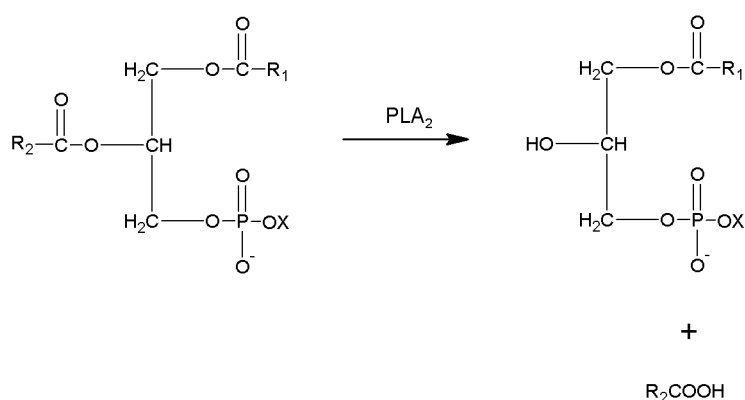


Figure 1.44: Reaction mechanism of PLA$_2$.

Two groups (VII and VIII) from the catalytic serine PLA$_2$ have the capacity to hydrolyse the sn2- position of the platelet activating factor (PAF) into acetate in addition to hydrolysing fatty acid chains. The enzymes of these groups bind also high density lipoproteins (HDL) and low density lipoproteins (LDL).[174] One of these enzymes, called lipoprotein-associated PLA$_2$ [Lp-PLA$_2$](or called GVIIA PLA$_2$ or platelet-activating factor acetylhydrolase [PAF-AH]) is present in plasma.[176] This 45 kDa enzyme is able to hydrolyze the platelet-activating factor (PAF) in two molecules: acetate and lysoPAF. This enzyme does not require Ca$^{2+}$ ion to be functional.[177] The enzyme also works in a water-soluble glycerophospholipids environment. [176] The regulation of this protein may be done during the processing of the differentiation of cells from the immune system. The PAF-AH is presented in high concentrations in the liver, lung, thymus, spleen and kidney.[177]

The structure of GVIIA PLA2 includes a lipase α/β hydrolase fold with a catalytic pocket. Two hydrophobic pockets serves as the active site for the lipoprotein and an acid pocket may use to bind the HDL and LDL.[177, 178] A crystal structure of this protein is presented in Figure 1.45.

Pl-PLA$_2$ is responsible for cardiovascular diseases such as coronary disease or atherosclerosis by releasing pro-inflammatory mediators (LPC and oxidized non-esterified fatty acids).[176, 177] It has also been shown that blocking Lp-PLA$_2$ may prevent asthma and protect against septicaemia. [177] The crystal structure of Lp-PLA$_2$ is shown in Figure 1.45.
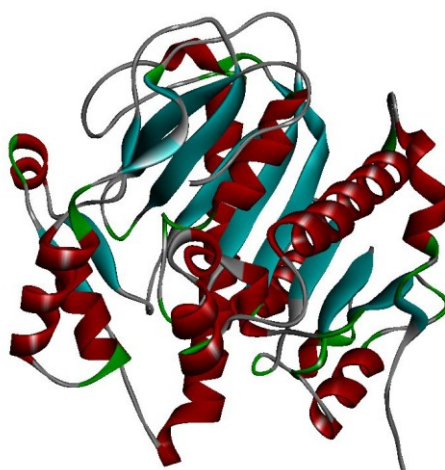


Figure 1.45: Lp-PLA$_2$ crystal structure [PDB ID: 5LYY]. [178]

At present, no study related to the chalcone family as inhibitor was found for this specific lipoprotein-associated phospholipase A$_2$. Most of the current literature investigations have focussed on group II or V of PLA$_2$. [140, 179-181]

The phospholipase A$_2$ is also one of targets of the venom of *Bothrops* snakes.[182] Subject to a previous study in this group, it was of interest to include this in this work.

## Aims of the current Thesis

Since the chalcone family presents a large promising therapeutic spectrum including anti-inflammatory, anti-microbial or anti-cancer proprieties, where specificity is not always clear, the aims of this thesis are, in a virtual screening approach:

1. To generate a library of chalcones based on the Claisen-Schmidt reaction.
2. To screen this library against a wide variety of biological targets, known to be susceptible to inhibition by chalcones including validation of the screening and molecular dynamics.
3. To analyse with statistical tools, and investigate relationships, if any, of the physicochemical characteristics of chalcones and the binding energy of the ligand-target complexes.
4. To improve to the binding energy through one or more modification strategies, of good binding ligands identified in the library.

# REFERENCES

1.      WHO. Close to 3 million people access hepatitis C cure.  2017  [cited 2018 30 July]; Available from:    http://www.who.int/news-room/detail/31-10-2017-close-to-3-million-people-access-hepatitis-c-cure.

2.      WHO. Eliminate hepatitis: WHO.  2017   [cited 2018 30 July]; Available from: http://www.who.int/news-room/detail/27-07-2017-eliminate-hepatitis-who.

3.      WHO, Global Hepatitis Report, 2017. 2017, WHO.

4.      A.Koutsoukas, Simms, J. Kirchmair, P.J. Bond, A.V. Whitmore, S. Zimmer, M.P. Young, J.L. Jenkins, M. Glick, R.C. Glen, and A. Bender, From in silico target prediction to multi-target drug design: Current databases, methods and applications. J Proteomics, 2011. **74**: p. 2554-2574.

5.      Chavatte, P., Médicaments, in Master Conception du Médicament. 2012.

6.      I. Gashaw, P. Ellinghaus, A. Sommer, and K. Asadullah, What makes a good drug target? Drug Discov Today, 2011. **175**(23-24): p. 1037-1043.

7.      Chris Finan, Anna Gaulton, Felix A. Kruger, R. Thomas Lumbers, Tina Shah, Jorgen Engmann, Luana Galver, Ryan Kelley, Anneli Karisson, Rita Santos, John P. Overington, Aroon D. Hingorani, and J.P. Casas, The druggable genome and support for target identification and validation in drug development. Sci Transl Med, 2017. **9**(383).

8.      Gibbs, J.B., Mechanism-based target identification and drug discovery in cancer research. Science, 2000. **287**(5460): p. 1969-1973.

9.      G. Rodger, C. Austin, J. Andersan, A. Pawlyk, C. Colvis, R. Margolis, and J. Baker, Glimmes in illuminating the druggable genome. Nat Rev Drugs Discov, 2018. **17**(5): p. 301-302.

10.     M. Campilllos, M. Kuhn, A-C. Gavin, L.J. Jensen, and P. Bork, Drug Target  Identification Using Side Effect Similarity. Science, 2008. **321**(5886): p. 263-266.

11.     A. L. Hopkins and C.R. Groom, The druggable genome. Nat Rev Drugs Discov, 2002. **1**(9): p. 727-730.

12.     J. P. Overington, B. Al-Lazikani, and A.L. Hopkins, How many drugs targets are there? Nat Rev Drugs Discov, 2006. **5**: p. 993-996.

13.     Chavatte, P. Relative Quantitative Structure à Activité. in Master Conception du Médicament. 2014.

14.     Lynch, S.S. Drug Development.  2016 Novmeber 2016 [cited 2018 05 February]; Available from:    http://www.merckmanuals.com/en-ca/professional/clinical-pharmacology/concepts-in-pharmacotherapy/drug-development.

15.     R. Miller, W. Ewy, B.W. Corrigan, D. Ouellet, D. Hermann, K. G. Kowalski, P. Lockwood, J. R. Koup, S. Donevan, A. El-Kattan, C.SW. Li, J. L. Werth, D. E. Feltner, and R.L. Lalonde, How Modeling and Simulation Have Enhanced Decision Making in New Drug Development. J Pharmacokinet Phar, 2005. **32**(2): p. 185-197.

16.     R. L. Lalonde, K.G. Kowalski, M. M. Hutmacher, W. Ewy, D.J. Nichols, P.A. Milligan, B.W. Corrigan, P.A. Lockwood, S.A. Marshall, L.J. Benincosa, T.G. Tensfeldt, K. PArivar, M. Amantea, P. Glue, H. Koide, and R. Miller, Model-based Drug Development. Clin Pharmacol Ther, 2007. **82**(1): p. 21-32.

17.     J.K. Willmann, N. van Bruggen, L. M. Dinkelborg, and S.S. Gambhir, Molecular imagin in drug development. Nat Rev Drugs Discov, 2008. **7**: p. 591-607.

18.     Research, C.F.D.E.a., Advancing HealthThrough Innovation 2017 New Drug Therapy Approval, ed. U.S.F.D. Administration. 2017. 36.

19.     R. R. Camejo, C. McGrath, and R. Herings, A dynamic perspective on pharmaceutical competition, drug development and cost effectiveness. Health Policy, 2011. **100**: p. 18-24.

20.     W.S. Comanor and F.M. Scherer, Mergers and innovation in the pharmaceutical industry. J Health Econ, 2013. **32**: p. 106-113.

21.    *D. Bardey, A. Bommier, and B. Jullien, Retail price regulation and innovation: Reference pricing in the pharmaceutical industry. J Health Econ, 2010. **29**: p. 303-316.*

22.    *J.A. DiMasi, H.G. Grabowski, and R.W. Hansen, Innovation in the pharmaceutical industry: New estimates of R&Dcosts. J Health Econ, 2016. **47**: p. 20-33.*

23.    *B. Spellberg, J.H. Powers, E.P. Brass, L.G. Miller, and J. J.E. Edwards, Trends in Antimicrobial Drug Development: Implications for the Future. Clin Infect Dis, 2004. **38**: p. 1279-1286.*

24.    *M. Gudiksen, E. Fleming, L. Furstenthal, and P. Ma, What drives success for specialty pharmaceuticals? Nat Rev Drugs Discov, 2008. **7**: p. 563-567.*

25.    *H.A. Ghofrani , I. H. Osterloh, and F. Grimminger, Sidenafil: from agina to erectil dysfunction to pulmonary hypertension and beyond. Nat Rev Drugs Discov, 2006. **5**(8): p. 689-702.*

26.    *Onywera, H., Influence of non-Synonymous Sequence Mutations on the Architecture of HIV-1 Clade C Protease Receptor Site: Docking and Molecular Dynamics Studies. 2013, Rhodes University: Grahamstown. p. 138.*

27.    *Mathu, A.M.N., Structural analysis of effects of mutations on HIV-1 subtype C protease active site, in Science. 2012, Rhodes University: Grahamstown. p. 91.*

28.    *C. Katiyar, A. Gupta, Satyajyoti Kanjilal, and S. Katiyar, Drug discovery from plants sources: An integrated approach. Ayu, 2012. **33**(1): p. 10-19.*

29.    *Shurkin, J., New Feature: Animals that self-medicate. PNAS, 2014. **111**(49): p. 17339-17341.*

30.    *P. Cos, A. J. Vlictinck, D. V. Berghe, and L. Maes, Anti-infective potential of natural products: How to develop a stronger in vitro 'proof-of-concept'. J Ethnopharmacol, 2006. **106**: p. 209-302.*

31.    *A. G. Atanasov, B. Waltenberg, Eva-Maria Pferschy-Wenzig, Thomas Linder, Christoph Wawrosch, Pavel Uhrin, Veronika Temml, Limei Wang, Stefan Schwaiger, Elice H. Heiss, Judith M. Rollinger, Daniela Schuster, Johannes M . Breuss, Valery Bochkov, Marko D. Mihovilovic, Brigitte Kopp, Rudolf Bauer, Verena M. Dirsch, and H. Stuppner, Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv, 2015. **33**: p. 1582-1614.*

32.    *G. M. Cragg and D.J. Newman, Natural products: a continuing source of novel drug leads. Biochim Biophys Acta, 2013. **1830**(6): p. 3670-3695.*

33.    *Mayor, S., Tree that provides paclitaxel is put on the list of endangered species. BMJ, 2011. **343**: p. 1.*

34.    *Lallemand, J.-Y. La place de l'hémisynthèse dans la production des médicaments. in NOUVEAUX CONCEPTS/NOUVELLES QUESTIONS Aspect cognitifs et sociaux des problématiques et des dispositifs de recherches aujourd'hui. 2005. Centre Lambert, Université Paris-Sud.*

35.    *Willard M. Allen and C. Goetsch, A simplified Method for the preparation of crystalline progesterone from pig ovaries. J Biol Chem, 1936. **116**(653).*

36.    *R. E. Marker and J. Kueger, Sterols. CXII. Sapogenins. XLI. The Preparation of Trillin and its Conversion to Progesterone. J Am Chem Soc, 1940. **62**(12): p. 3349-3350.*

37.    *Wilkins, L.W., ed. Foye's Principles of Medicinal Chemistry. 7th edition ed. 2013.*

38.    *T. L. Lemke and D.A. Williams, Foye's principles of medicinal chemistry. 7th edition ed. 2002, Philadelphia: Lippincott Williams & Wilkins.*

39.    *D. Balogh, W. J.Begley, D. Bremner, M. J. Wyvratt, and L.A. Paquette, Enzyme-assisted semisynthesis of human insulin. J Am Chem Soc, 1979. **101**(3): p. 751-752.*

40.    *Richard K. Assoiant and H. S.Tager, [($^{125}$I)Iodotyrosyl$^{B1}$]insulin semisynthesis,receptor binding and cell-mediated degradation of a B chain-labeled. J Biol Chem, 1981. **256**(8): p. 4042-4049.*

41.    *K. C. Nicolaou, D. Vourloumis, N. Winssinger, and P.S. Baran, The Art and Science of Total Synthesis at the Dawn of the Twenty-First Century. Angew Chem Int Ed Engl, 2000. **39**(1): p. 44-122.*

42.    *M. C. de la Torre and M.A. Sierra, Comments on Recent Achievements in Biomimetic Organic Synthesis. Angew Chem Int Ed Engl, 2003. **43**(2): p. 160-181.*

43. *A. K. Saksena and P. Mangiaracina, Recent Studies on Veratrum alkaloids: a new reaction of sodium triacetoxyborohydrique [NaBH(OAc)$_3$]. Tetrahedron Lett, 1983.* **24***(3): p. 273-276.*

44. *D. A. Evans, K.T.Chapman, and E.M. Carreira, Directed Reduction of ß-Hydroxy Ketones Employing Tetramethylammonium Triacetoxyborohydride. J Am Chem Soc, 1988.* **110***(11): p. 3560-3578.*

45. *E. J. Corey, M. Ohno, P. A. Vatakencherry, and R.B. Mitra, TOTAL SYNTHESIS OF d,l-LONGIFOLENE. J Am Chem Soc, 1961.* **83***(5): p. 1251-1253.*

46. *Corey, E.J., Retrosynthetic Thinking--Essentials and Examples. Chem Soc Rev, 1988.* **17***: p. 111-133.*

47. *D-L. Ma, D. S-H. Chan, and C.-H. Leung, Molecular docking for virtual screening of natural product databases. Chem Sci, 2011.* **2***(9): p. 1656-1665.*

48. *R. Hatherley, D. K. Brown, T. M. Musyoka, D. L. Penkler, N. Faya, K. A. Lobb, and Ö.T. Bishop, SANCDB: a South Africa natural compound database. J Cheminformatics, 2015.* **7***(29): p. 1-9.*

49. *F. Ntie-Kang, K. K. Telukunta, K. Döring, C. V. Simoben, A. F. A. Moumbock, Y. I. Malange, L. E. Njume, J. N. Yong, W. Sippl, and S. Günther, NANPDB: A Resource for Natural Products from Northern African Sources. J Nat Prod, 2017.* **80***(7): p. 2067-2076.*

50. *Y-C.Chen, C., TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. PLoS One, 2011.* **6***(1): p. 1-5.*

51. *X. Qiao, T. Hou, W. Zhang, S.L. Guo, and X. Xu, A 3D Structure Database of Components from Chinese Traditional Medicinal Herbs. J Chem Inf Model, 2002.* **42***(3): p. 481-489.*

52. *J. Shen, X. Xu, F. Cheng, H. Liu, X. Luo, J. Shen, K. Chen, W. Zhao, X. Shen, and H. Jiang, Virtual Screening on Natural Products for Discovering Active Compounds and Target Information. Curr Med Chem, 2003.* **10***(21): p. 2327-2342.*

53. *F. Ntie-Kang, J. N. Nwodo, A. Ibezim, C. V. Simoben, B. Karaman, V. F. Ngwa, W. Sippl, M. U. Adiwu, and L.M. Mbaze, Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. J Chem Inf Model, 2014.* **54***(9): p. 2433-2450.*

54. *V. J. Henry, A. E. Bandrowski, A-S. Pepin, B. J. Gonzalez, and A. Desfeux, OMICtools: an informative directory for multi-omic data analysis. Database (Oxford), 2014: p. 1-5.*

55. *D. S. Wishard, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Xhang, and J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res, 2006.* **34***(Issue suppl_1, 1 January 2006): p. D668–D672.*

56. *D. S. Wishard, Y. D. Feunang, A. C. Guo, A.M. E. J. Lo, T.S. J. R. Grant, D. Johnson, Z.S. C. Li, N. Assempour, I. Iynkkaran, Yifeng, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res, 2017.* **46***(D1): p. D1074–D1082.*

57. *S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu. A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J . Zhang, and S.H. Bryant, PubChem Substance and Compound databases. Nucleic Acids Res, 2016.* **44***(Database issue): p. D1202-D1213.*

58. *S. Kim, P. A. Thiessen, T. Chen, B. Yu, B. A. Shoemaker, J. Wang, E. E. Bolton, Y. Wang, and S.H. Bryant, Literature information in PubChem: associations between PubChem records and scientific articles. J Cheminformatics, 2016.* **8***(32): p. 1-15.*

59. *T. Cheng, M. Hao, T. Takeda, S. H. Bryant, and Y. Wang, Large-Scale Prediction of Drug-Target Interaction: a Data-Centric Review. AAPS J, 2017.* **19***(5): p. 1264-1275.*

60. *Y. Wang, T. Cheng, and S.H. Bryant, PubChem BioAssay: A Decade's Development toward Open High-Throughput Screening Data Sharing. SLAS Discov, 2017.* **22***(6): p. 655-666.*

61. *J. J. Irwin and B.K. Shoichet, ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. J Chem Inf Model, 2005.* **45***(1): p. 177-182.*

62. *Sterling, T., ZINC 15 – Ligand Discovery for Everyone. J Chem Inf Model, 2015.* **55***(11): p. 2324-2337.*

63. *J. J. Irwin, T. Sterling, M. M Mysinger, E. S. Bolstad, and R.G. Coleman, ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model, 2012.* **52***(7): p. 1757-1768.*

64. *V. Le Guilloux, L. Colliandre, S. Bourg, G. Guénegou, J. Dubois-Chevalier, and L. Morin-Allory, Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. J Chem Inf Model, 2011. **51**(8): p. 1762-1774.*

65. *D.M. Schnur, B. R. Beno, A. J. Tebben, and C. Cavallaro, Methods for Combinatorial and Parallel Library Design. Methods Mol Biol, 2011. **672**: p. 387-434.*

66. *Schrödinger, Schrödinger Release 2018-4: CombiGlide. 2018: New York, NY.*

67. *Zhou, J.Z., Structure-directed combinatorial library design. Curr Opin Chem Biol, 2008. **12**(3): p. 379-385.*

68. *T. Fink, H. Bruggesser, and J.-L. Reymond, Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. Angew Chem Int Ed Engl, 2005. **44**(10): p. 1504-1508.*

69. *L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. J Chem Inf Model, 2012. **52**(11): p. 2864-2875.*

70. *T. Fink and J.-L. Reymond, Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. J Chem Inf Model, 2007. **47**(2): p. 342-353.*

71. *A. Vulpetti and C. Dalvit, Design and Generation of Highly Diverse Fluorinated Fragment Libraries and their Efficient Screening with Improved 19F NMR Methodology. ChemMedChem, 2013. **8**(12): p. 2057-2069.*

72. *S. Saubern, R. Guha, and J.B. Baell, KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. Mol Inform, 2011. **30**(10): p. 847-850.*

73. *Bruneton, J., Pharmacogonsie Phytochimie Plantes Médicinales. 4ᵗʰ Edition ed, ed. Lavoisier. 2009.*

74. *Y. Chen, H. Xian, J. Zheng, and G. Liang, Structure-Thermodynamics-Antioxidant Activity Relationships of Selected Natural Phenolic Acids and Derivatives: An Experimental and Theoretical Evaluation. PLoS One, 2015. **10**(3).*

75. *D-M Song, K-H. Jung, J-H. Moon, and D.-M. Shin, Photochemistry of chalcone and the application of chalcone-derivatives in photo-alignment layer of liquid crystal display. Opt Mater, 2003. **21**(1-3): p. 667-671.*

76. *P. Singh, A. Anand, and V. Kumar, Recent developments in biological activities of chalcones: A mini review. Euro J Med Chem, 2014. **85**: p. 758-777.*

77. *H. Maeda and N. Dudareva, The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. Annu Rev Plant Biol, 2012. **63**: p. 73-105.*

78. *V. Tzin and G. Galili, The Biosynthetic Pathways for shikimate and Aromatic Amino Acids in Arabidopsis thaliana, in The Arabidopsis Book, A.S.o.P. Biologists, Editor. 2010.*

79. *Knaggs, A.R., The Biosynthesis of Shikimate Metabolites. Nat Prod Rep, 2001. **18**: p. 334-355.*

80. *K. M. Herrmann and L.M. Weaver, The shikimate pathway. Annu Rev Plant Physiol Plant Mol Biol, 1999. **50**: p. 473-503.*

81. *Dewick, P.M., The Biosynthesis of Shikimate Metabolites. Nat Prod Rep, 1998. **15**(1): p. 17-58.*

82. *Tsao, R., Chemistry and Biochemistry of Dietary Polyphenols. Nutrients, 2010. **2**: p. 1231-1246.*

83. *V. R. Yadav, S. Prasad, B. Sung, and B.B. Aggarwal, The role of chalcones in suppression of NF-κB-mediated inflammation and cancer. Int Immunopharmacol, 2011. **11**: p. 295-309.*

84. *T. T. H. Dao, H. J. M. Linthorst, and R. Verpoorte, Chalcone synthase and its functions in plant resistance. Phytochem Rev, 2011. **10**: p. 397-412.*

85. *Z. Rozmer and P. Perjési, Naturally occurring chalcones and their biological activities. Phytochem Rev, 2016. **15**(1): p. 87-120.*

86. *M. N. Gomes, E. N. Muratov, M. Pereira, J. C. Peixoto, L. P. Rosseto, P.V. L. Cravo, C. H. Andrade, and B.J. Neves, Chalcone Derivatives: Promising Starting Points for Drug Design. Molecules, 2017. **22**(8).*

87. *D. K. Mahapatra, S. K. Bharti, and V. Asati, Chalcones scaffolds as anti-infective agents: Structural and molecular target perspectives. Euro J Med Chem, 2015. **101**: p. 496-524.*

88. *S. N. A. Bukhari, M. Jasamai, and I. Janta, Synthesis and Biological Evaluation of Chalcone Derivatives (Mini Review). Mini-Rev Med Chem, 2012. **12**(13): p. 1394-403.*

89. *C-T. Hiesh, T-J. Hsieh, M. El-Shazly, D-W. Chuang, C.-T.Y. Y-H. Tsai, S-F. Wu, Y-C Wu, and F.-R. Chang, Synthesis of chalcone derivatives as potential anti-diabetic agents. Bioorg Med Chem Lett, 2012. **22**(12): p. 3912-3915.*

90. *P. M. Sivakumar, S. P. Seenivasan, V.Kumar, and M. Doble, Synthesis antimycobacterial activity evaluation, and QSAR studies of chalcone derivatives. Bioorg Med Chem Lett, 2007. **17**(6): p. 1695-1700.*

91. *M. S. Cheng, R. S. Li, and G. Kenyon, A Solid Phase Synthesis of Chalcones by Claisen-Schmidt Condensations. Chinese Chem Lett, 2000. **11**(10): p. 851-854.*

92. *D. Kakati and J.C. Sarma, Microwave assisted solvent free synthesis of 1,3-diphenylpropenones. Chemistry Central Journal, 2001. **5**(8).*

93. *Srivastava, Y.K., Ecofriendly microwave assisted synthesis ofsome chalcones. Rasayan J Chem, 2008. **1**(4): p. 884-886.*

94. *X-F Wu, H.Neumann, A. Spanneberg, T. Schulz, H.Jiao, and M. Beller, Development of a General Palladium-Catalyzed carbonylative Heck Reaction of Aryl Halides. J Am Chem Soc, 2010. **132**(41): p. 14596-14602.*

95. *L-W. Xu, L. Li, C-G. Xia, and P.-Q. Zhao, Efficient Coupling Reactions of Arylalkynes and Aldehydes Leading to the Synthesis of Enones. Helv Chim Acta, 2004. **87**(12): p. 3080-384.*

96. *M. A. Selepe and F.R.v. Heerden, Application of the Suzuki-Miyaura Reaction in the Synthesis of Flavonoids. Molecules, 2013. **18**(4): p. 4739-4765.*

97. *S. B. Ötvös, C-T. Hsieh, Y-C. Wu, J-H. Li, F-R. Chang, and F. Fülöp, Continuous-Flow Synthesis of Deuterium-Labeled Antidiabetic Chalcones: Studies towards the Selective Deuteration of the Alkynone Core. Molecules, 2016. **21**(3).*

98. *M. Rueping, T. Bootwicha, H. Baars, and E. Sugiono, Continuous-flow hydration–condensation reaction: Synthesis of α,β-unsaturated ketones from alkynes and aldehydes by using a heterogeneous solid acid catalyst. Beilstein J Org Chem, 2011. **7**: p. 1680-1687.*

99. *D. K. Mahapatra, S. K. Bharti, and V. Asati, Anti-cancer chalcones: Structural and molecular target perspectives. Euro J Med Chem, 2015. **98**: p. 69-114.*

100. *B. Zhou and C. Xing, Diverse Molecular Targets for Chalcones with Varied Bioactivities. Med Chem, 2015. **8**: p. 388-404.*

101. *H-H. Ko, L-T. Tsao, K-L. Yu, C-T. Liu, J-P. Wang, and C.-N. Lin, Structure–Activity Relationship Studies on Chalcone Derivatives: The Potent Inhibition of Chemical Mediators Release. Bioorg Med Chem, 2003. **11**(1): p. 105-111.*

102. *D. K. Mahapatra and S.K. Bharti, Therapeutic potential of chalcones as cardiovascular agents. Life Sci, 2016. **148**: p. 154-172.*

103. *S. Cheenpracha, C. Karalai, C. Ponglimanont, S. Subhadhirasakul, and S. Tewtrakul, Anti-HIV61 protease activity of compounds from Boesenbergia pandurata. Bioorg Med Chem, 2006. **14**(6): p. 1710-1714.*

104. *WHO, WHO HIV update. 2018, WHO.*

105. *A. Rambaut, D. Posada, K. A. Crandall, and E.C. Holmes, The causes and consequences of HIV evolution. Nat Rev Genet, 2004. **5**(1): p. 52-61.*

106. *M. Marlowitz, J. O. Morales-Ramirez, B-Y. Nguyen, C. M. Kovacs, R. Isaacs, L. R. Gilde, L. Wenning, J. Zhao, and H. Teppler, Antiretroviral Activity Pharmacokinetics and Tolerance of MK-0518, a Novel Inhibitor of HIV-1 Integrase, Dosed As Monotherapy for 10 Days in Treatment-Naive HIV-1-Infected Individuals. J Acq Immun Def Synd, 2006. **43**(5): p. 509-515.*

107. *A. Engelman and P. Cherepanov, Retroviral Integrase Structure and DNA Recombination Mechanism. Microbiol Spectr, 2014. **2**(6): p. 1-22.*

108. *J.J. Kessl, C. J. McKee, J. O. Eidahl, N. Shkriabai, A. Katz, and M. Kvaratskhelia, HIV-1 Integrase-DNA Recognition Mechanisms. Viruses, 2009. **1**(3): p. 713-736.*

109.   K. Gupta, V. Turkki, S. Sherrill-Mix, Y. Hwang, G. Eilers, L. Taylor, C. McDanal, P. Wang, D. Temelkoff, R. T. Nolte, E. Velthuisen, J. Jeffrey, G. D. Van Duyne, and F.D. Bushman, *Structural Basis for Inhibitor-Induced Aggregation of HIV Integrase*. PloS Bio, 2016. **14**(12).

110.   L. D. Fader, E. Malenfant, M. Parisien, R. Carson, F. Bilodeau, S. Landry, M. Pesant, C. Brochu, S. Morin, C. Chabot, T. Halmos, Y. Bousquet, M. D. Bailey, S. H. Kawai, R. Coulombe, S. LaPlante, A. Jakalian, P. K. Bhardwaj, D. Wernic, P. Schroeder, M. Amad, P. Edwards, M. Garneau, J. Duan, M. Cordingley, R. Bethell, S. W. Mason, M. Bös, P. Bonneau, M-A. Poupart, A_M. Faucher, B. Simoneau, C. Fenwick, C. Yoakim, and Y. Tsantrizos, *Discovery of BI 224436, a Noncatalytic Site Integrase Inhibitor (NCINI) of HIV-1*. ACS Med. Chem. Lett., 2014. **5**(4): p. 422-427.

111.   J. Deng, T. Shanchez, L. Q. Al-Mawsawi, R. Dayam, R. A .Yunes, A. Garofalo, M. B. Bolger, and N. Neamati, *Discovery of structurally diverse HIV-1 integrase inhibitors based on a chalcone pharmacophore*. Bioorg Med Chem, 2007. **15**(14): p. 4985-5002.

112.   A. Vasu Babu, R. Navudu, and G. Trimurtulu, *Synthesis of C-methyl chalcones as HIV-integrase inhibitors—computational approach*. Med Chem Res, 2014. **23**(2): p. 877-881.

113.   H. Sharma, S. Patil, T. W. Sanchez, N. Neamati, R. F. Schinazi, and J.K. Buolamwini, *Synthesis, Biological Evaluation and 3D-QSAR Studies of 3-Keto Salicylic Acid Chalcones and Related Amides as Novel HIV-1 Integrase Inhibitors*. Bioorg Med Chem, 2011. **19**(6).

114.   CDC. *General Information About MRSA in Healthcare Settings*. 2018 [cited 2018 12/09/2018]; Available from: https://www.cdc.gov/mrsa/healthcare/index.html#q3.

115.   T. M. Osòrio, F. D. Monache, L. D. Chiaradia, A. Mascarello, T. R. Stumpf, C. R. Zanetti, D. B. Silveira, C. R. M. Barardi, E. d F. A. Smânia, A. Viancelli, L. A. T. Garcia, R. A. Yunes, R. J. Nunes, and A.S. Jr, *Antibacterial activity of chalcones, hydrazones and oxadiazoles against methicillin-resistant Staphylococcus aureus*. Bioorg Med Chem Lett, 2012. **22**(1): p. 225-230.

116.   R. Zoraghi, L. Worral, R. H. See, W. L. Popplewell, H. Gong, T. Samaai, R. D. Swayze, S. Kaur, M. Vuckovic, B. B. Finlay, R. C. Bruham, W. R. McMaster, M. T. Davies-Coleman, N. C. Strynadka, R. J. Andersen, and N.E. Reiner, *Methicillin-resistant Staphylococcus aureus (MRSA) Pyruvate Kinase as a Target for Bis-indole Alkaloids with Antibacterial Activities*. J Biol Chem, 2011. **286**(52): p. 44716-44725.

117.   N. S. Kumar, E. M. Dullaghan, B. B. Finlay, H. Gong, N. E. Reiner, J. J. P. Selvam, L. M. Thorson, S. Campbell, N. Vitko, A. R. Richardson, R. Zoraghi, and R.N. Young, *Discovery and optimization of a new class of pyruvate kinase inhibtors as potential therapeutics for the treatment of methicillin-resistant Staphylococcus aureus infections*. Bioorg Med Chem, 2014. **22**(5): p. 1708-1725.

118.   T. P. T. Cushnie and A.J. Lamb, *Recents advances in understanding the antibacterial properties of flavonoids*. Int J Antimicrob Ag, 2011. **38**: p. 99-107.

119.   L. E. Alcaràz, S. E. Blanco, O. N. Puig, F. Tomàs, and F.H. Ferretti, *Antibacterial Activity of Flavonoids Against Methicillin-resistant Staphylococcus aureus strains*. J Theor Biol, 2000. **205**(205): p. 231-240.

120.   C. Labrière, H. Gong, B. B. Finlay, N. E. Reiner, and R.N. Young, *Further investigastion of inhibitors of MRSA pyruvate kinase: Towards the conception of novel antimicrobial agents*. Euro J Med Chem, 2017. **125**: p. 1-13.

121.   J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, and F.B. D. Forman, *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012*. Int J Cancer, 2015. **136**(5): p. E359-386.

122.   J. Li and J. Buchner, *Structure, Function and Regulation of the Hsp90 Machinery*. Biomed J, 2013. **36**(3): p. 106-117.

123.   S. Verma, S. Goyal, S. Jamal, A. Singh, and A. Grover, *Hsp90: Friends, clients and natural foes*. Biochimie, 2016. **127**: p. 227-240.

124.   D. S. Hong, U. Banerji, B. Tavana, G. C. George, J. Aaron, and R. Kurzrock, *Targeting the molecular chaperone heat shock protein 90 (HSP90): Lessons learned and future directions*. Cancer Treat Rev, 2013. **39**(4): p. 375-387.

125. *S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N . Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, and X. Barril, Dynamic undocking and the quasi-bound state as tools for drug discovery. Nat Chem, 2017. **9**(3): p. 201-206.*

126. *J. Davenport, M. Balch, L. Galam, A. Girgis, J. Hall, B. S. J. Blagg, and R.L. Matts, High-throughput screen of natural product libraries for hsp90 inhibitors. Biology (Basel), 2014. **3**(1): p. 101-138.*

127. *F. D. Piaz, S. Terracciano, N. De Tommasi, and A. Braca, Hsp90 Activity Modulation by Plant Secondary Metabolites. Planta Med, 2015. **81**(14): p. 1223-1239.*

128. *Y. S. Kim, V. Kumar, S. Lee, A. Iwai, L. Neckers, and S.V. Malhotra, Methoxychalcone inhibitors of androgen receptor translocation and function. Bioorg Med Chem Lett, 2012. **22**(5): p. 2105-2109.*

129. *C-H. Jeong, H. B. Park, W. J. Jang, S. H. Jung, and Y.H. Seo, Discovery of hybrid Hsp90 inhibitors and their anti-neoplastic effects against gefitinib-resistant non-small cell lung cancer (NSCLC). Bioorg Med Chem Lett, 2014. **24**(1): p. 224-227.*

130. *J. H. Jeong, Y. J. Oh, T. K. Kwon, and Y.H. Seo, Chalcone-templated Hsp90 inhibitors and their effects on gefitinib resistance in non-small cell lung cancer (NSCLC). Arch Pharm Res, 2017. **40**(1): p. 96-105.*

131. *R. A. Soslow, A. J. Danneberg, D. Rush, B. M. Woener, K. N. Khan, J. Masferrer, and A.T. Koki, COX-2 Is Expressed in Human Pulmonary, Colonic, and Mammary Tumors. Cancer, 2000. **89**(12): p. 2637-2645.*

132. *B. F. McAdam, F. Catella-Lawson, I. A. Mardini, S. Kapoor, J. A. Lawson, and G.A. FitzGerald, Systemic biosynthesis of prostacyclin by cyclooxygenase (COX)-2: The human pharmacology of a selective inhibitor of COX-2. PNAS, 1999. **96**(1): p. 272-277.*

133. *J. Y. Park, M. H. Pillinger, and S.B. Abramson, Prostaglandin E2 synthesis and secretion: The role of PGE2 synthases. Clin Immunol, 2006. **119**(3): p. 229-240.*

134. *M. J. Lucido, B. J. Orlando, A. J. Vecchio, and M.G. Malkowski, Crystal Structure of Aspirin Acetyl Human Cyclooxygenase-2: Insight Into the Formation of Products with Reversed Stereochemistry. Biochemistry, 2016. **55**(8): p. 1226-1238.*

135. *L. J. Marnett, S. W. Rowlinson, D. C. Goodwin, A. S. Kalgutkar, and C.A. Lanzo, Arachidonic Acid Oxygenation by COX-1 and COX-2 J Biol Chem, 1999. **274**(33): p. 22903-22906.*

136. *J.k. Gierse, J. J. McDonald, S. D. Hauser, S. H. Rangwala, C. M. Koboldt, and K. Seibert, A Single Amino Acid Difference between Cyclooxygenase-1 (COX-1) and -2 (COX-2) Reverses the Selectivity of COX-2 Specific Inhibitors. J Biol Chem, 1996. **271**(26): p. 15810-15814.*

137. *S. B. Appleby, A. Ristimäki, K. Narko, and T. Hla, Structure of the human cyclo-oxygenase-2 gene. Biochem J, 1994. **302**(Pt 3): p. 723-727.*

138. *Y-J.Surh, K-S. Chun, H-H. Cha, S. S. Han, Y-S. Keum, K-K. Park, and S.S. Lee, Molecular mechanisms underlying chemopreventive activities of anti-inflammatory phytochemicals: down regulation of COX-2 and iNOS through suppression of NF-κB activation. Mutat Res, 2001. **480-481**: p. 2413-268.*

139. *R. S. Sidhu, J. Y. Lee, C. Yuan, C. Yuan, and W.L. Smith, Comparison of Cyclooxygenase-1 Crystal Structures: Cross-Talk between Monomers Comprising Cyclooxygenase-1 Homodimers. Biochemistry, 2010. **49**(33): p. 7069-7079.*

140. *F. Herencia, M. L. Ferràndiz, Amalia Ubeda, J. N. Domìnguez, J. E. Charris, G. M. Lobo, and M.J. Alcaraz, Synthesis and anti-inflammatory activity of chalcone derivatives. Bioorg Med Chem Lett, 1998. **8**(10): p. 1169-1174.*

141. *F. Herencia, M. L. Ferràndiz, Amalia Ubeda, I. Guillén, J. N. Domìnguez, J. E. Charris, G. M. Lobo, and M.J. Alcaraz, Novel anti-infammatory chalcone derivatives inhibit the induction of nitric oxide synthase and cyclooxygenase-2 in mouse peritoneal macrophages. FEBS Lett, 1999. **453**(1-2): p. 129-134.*

142. *B. P. Bandgar, S. A. Patil, and B.L. R. N. Gacche, Synthesis and biological evaluation of nitrogen-containing chalcones as possible anti-inflammatory and antioxidant agents. Bioorg Med Chem Lett, 2010. **20**(2): p. 730-733.*

143. *A. Zarghi, T. Zebardast, F. Hakimion, F. H. Shirazi, P. N. Praveen Rao, and E.E. Knaus, Synthesis and biological evaluation of 1,3-diphenylprop-2-en-1-ones possessing a methansulfonamido or an azido pharmacophore as cyclooxygenase-1/-2 inhibitors. Bioorg Med Chem, 2006. **14**(20): p. 7044-7050.*

144. *S-Y. Son, J. Ma, Y. Kondou, M. Yoshimura, E. Yamashita, and T. Tsukihara, Structure of human monoamine oxidase A at 2.2-Å resolution: The control of opening the entry for substrates/inhibitors. PNAS, 2008. **105**(15): p. 5739-5744.*

145. *J. C. Shih, J. Grimsby, K. Chen, and Q.-S. Zhu, Structure and Promoter Organization of the Human Monoamine Oxidase A and B Genes. J Psychiatr Neurosci, 1993. **18**(1): p. 25-32.*

146. *J. Ma, M. Yoshimura, E. Yamashita, A. Nakagawa, A. Ito, and T. Tsukihara, Structure of Rat Monoamine Oxidase A and Its Specific Recognitions for Substrates and Inhibitors. J Mol Biol, 2004. **338**(1): p. 103-114.*

147. *Z-Y. Chen, G. S. Hotamisligil, J-K. Huang, D.E. L. W, N. Aydin-Muderrisoglu, J. F. Powell, R. H. Huang, X. O. Breakefield, I. Craig, and Y.-P.P. Hsu, Structure of the human gene for monoamine oxidase type A. Nucleic Acids Res, 1991. **19**(16): p. 4537-4541.*

148. *D. E. Edmondson, A. Mattevi, C. Binda, M. Li, and F. Hubálek, Structure and Mechanism of Monoamine Oxidase. Curr Med Chem, 2004. **11**(15): p. 1983-1993.*

149. *L. de Colibus, M. Li, C. Binda, A. Lustig, D. E. Edmondson, and A. Mattevi, Three-dimensional structure of human monoamine oxidase A (MAO A): Relation to the structures of rat MAO A and human MAO B. PNAS, 2005. **102**(36): p. 12684-12689.*

150. *P. de Deurwaerdère, C. Binda, R. Corne, C. Leone, A. Valeri, M. Valoti, R. R. Ramsay, Y. Fall, and J. Marco-Contelles, Comparative Analysis of the Neurochemical Profile and MAO Inhibition Properties of N-(Furan-2-ylmethyl)-N methylprop-2-yn-1-amine. ACS Chem Neurosci, 2017. **8**(5): p. 1026-1035.*

151. *F. Chimenti, R. Fioravanti, A. Bolasco, P. Chimenti, D. Secci, F. Rossi, M. Yànez, F. Orallo, F. Ortuso, and S. Alcaro, Chalcones: A Valid Scaffold for Monoamine Oxidases Inhibitors. J Med Chem, 2009. **52**(9): p. 2818-2824.*

152. *N. Morales-Camilo, C. O. Salas, C. Sanhueza, C. Espinosa-Bustos, S. Sepùlveda-Boza, M. Reyes-Parada, F. Gonzalez-Nilo, M. Caroli-Rezende, and A. Fierro, Synthesis, Biological Evaluation, and Molecular Simulation of Chalcones and Aurones as Selective MAO-B Inhibitors. Chem Biol Drug Des, 2015. **85**(6): p. 685-695.*

153. *Quinn, D.M., Acetylcholinesterase: Enzyme Structure, Reaction Dynamics, and Virtual Transition States. Chem Rev, 1987. **87**(5): p. 955-979.*

154. *J. Cheung, E. N. Gary, K. Shiomi, and T.L. Rosenberry, Structures of Human Acetylcholinesterase Bound to Dihydrotanshinone I and Territrem B Show Peripheral Site Flexibility. ACS Med. Chem. Lett., 2013. **4**(11): p. 1091-1096.*

155. *H. L. Fernandez, R. D. Moreno, and N.C. Inestrosa, Tetrameric (G₄) Acetylcholinesterase: Structure, Localization, and Physiological Regulation. J Neurochem, 1996. **66**(4): p. 1335-1346.*

156. *H. Dvir, I. Silma, M. Harel, T.L. Rosenberry, and J.L. Sussman, Acetylcholinesterase: From 3D structure to functionChemico-Biological Interactions. Chem Biol Interact, 2010. **187**(1-3): p. 10-22.*

157. *H-R. Liu, X-Q. Huang, D-H. Lou, X-J. Liu, W-K. Liu, and Q.-A. Wang, Synthesis and acetylcholinesterase inhibitory activity of Mannich base derivatives flavokawain B. Bioorg Med Chem Lett, 2014. **24**(19): p. 4749-4753.*

158. *H-R. Liu, X-J. Liu, H-Q Fan, J-J Tang, X-H. Gao, and W.-K. Liu, Design, synthesis and pharmacological evaluation of chalcone derivatives as acetylcholinesterase inhibitors. Bioorg Med Chem, 2014. **22**(21): p. 6124-6133.*

159. *H-R. Liu, C. Zhou, H-Q Fan, J-J. Tang, L-B. Liu, X-Hui. Gao, Q-A. Wang, and W.-K. Liu, Novel Potent and Selective Acetylcholinesterase Inhibitors as Potential Drugs for the Treatment of Alzheimer's Disease: Synthesis, Pharmacological Evaluation, and Molecular Modeling of*

*Amino-Alkyl-Substituted Fluoro-Chalcones Derivatives. Chem Biol Drug Des, 2015.* **86***(4): p. 517-522.*

160. *S. D. Sukumaran, C. F. Chee, G. Viswanathan, M. J. C. Buckle, R. Othman, N. A. Rahman, and L.Y. Chung, Synthesis, Biological Evaluation and Molecular Modelling of 21-Hydroxychalcones as Acetylcholinesterase Inhibitors. Molecules, 2016.* **21***(7): p. 1-10.*

161. *Y. Nicolet, O. Lockridge, P. Masson, J. C. Fontecilla-Camps, and F. Nachon, Crystal Structure of Human Butyrylcholinesterase and of Its Complexes with Substrate and Products. J Biol Chem, 2003.* **278***(42): p. 41141-41147.*

162. *Lockridge, O., Review of human butyrylcholinesterase structure, function, genetic variants, history of use in the clinic, and potential therapeutic uses. Pharmacol Therapeut, 2015.* **148***: p. 34-46.*

163. *U. Košak, B. Brus, D. Knez, R. Šink, S. Žakelj, J. Trontelj, A. Pišlar, J. Šlenc, M. Gobec, M. Živin, L. Tratnjek, M. Perše, K. Sałat, A. Podkowa, B. Filipek, F. Nachon, X. Brazzolotto, A. Więckowska, B. Malawska, J. Stojan, I. M. Raščan, J. Kos, N. Coquelle, J-P. Colletier, and S. Gobec, Development of an in-vivo active reversible butyrylcholinesterase inhibitor. Sci Rep, 2016.* **6***: p. 1-16.*

164. *J. M. Jez, T. G. FLynn, and T.M. Penning, A New Nomenclature for the Aldo-Keto Reductase Superfamily. Biochem Pharmacol, 1997.* **54***(6): p. 639-647.*

165. *D. K. Wilson, I. Tarle, J. M. Petrash, and F.A. Quiocho, Refined 1.8 A structure of human aldose reductase complexed with the potent inhibitor zopolrestat. PNAS, 1993.* **90***(21): p. 9847-9851.*

166. *O. A. Barski, K. H. Gabbay, and K.M. Bohren, Characterization of the Human Aldehyde Reductase Gene and Promoter. Genomics, 1999.* **60***(188-198).*

167. *Penning, T.M., The aldo-keto reductases (AKRs): Overview. Chem Biol Interact, 2015.* **234***: p. 236-246.*

168. *D. W. Borhani, T. M. Harter, and J.M. Petrash, The Crystal Structure of the Aldose Reductase NADPH Binary Complex. J Biol Chem, 1992.* **297***(34): p. 24841-24847.*

169. *J. M. Rondeau, F. Tête-Favier, A. Podjarny, J-M. Reymann, P. barth, J-F. Biellmann, and D. Moras, Novel NADPH-binding domain revealed by the crystal structure of aldose reductase. Nature, 1992.* **355***: p. 469-472.*

170. *P. O. de Giuseppe, M. L. dos Santos, S. M. de Sousa, K. E. Koch, J. A. Yunes, R. Aparicio, and M.T. Murakami, A comparative structural analysis reveals distinctive features of cofactor binding and substrate specificity in plant aldo-keto reductases. Biochem Biophys Res Commun, 2016.* **474***(4): p. 696-701.*

171. *F. Severi, S. Benvenuti, L. Costantino, G. Vampa, M. Melegari, and L. Antolini, Synthesis and activity of a new series of chalcones as aldose reductase inhibitors. Euro J Med Chem, 1998.* **33***(11): p. 859-866.*

172. *S. Iwata, N. Nagata, A. Omae, S. Yamaguchi, Y. Okada, S. Shibata, and T. Okuyama, Inhibitory Effect of Chalcone Derivatives on Recombinant Human Aldose Reductase. Bio Pharm Bull, 1999.* **22***(3): p. 323-325.*

173. *D. K. Mahapatra, V. Asati, and S.K. Bharti, Chalcones and their therapeutic targets for the management of diabetes: Structural and pharmacological perspectives. Euro J Med Chem, 2015.* **92***: p. 839-865.*

174. *J. E. Burke and E.A. Dennis, Phospholipase A$_2$ Biochemistry. Cardiovasc Drugs Ther, 2009.* **23***(1): p. 49-59.*

175. *D. A. Six and E.A. Dennis, The expanding superfamily of phospholipase A$_2$ enzymes: classication and characterization. Biochimica et Biophysica Acta, 2000.* **1488***(1-2): p. 1-19.*

176. *R. H. Schaloske and E.A. Dennis, The phospholipase A$_2$ superfamily and its group numbering system. Biochimica et Biophysica Acta, 2006.* **1761***(11): p. 1246-1259.*

177. *M. Murakami, Y. Taaketomi, Y. Miki, H. Sato, T. Hirabayashi, and K. Yamamoto, Recent progress in phospholipase A2 research: From cells to animals to humans. Prog Lipid Res, 2011.* **50***(2): p. 152-192.*

178. A. J-A. Woolford, P. J. Day, V. Bénéton, V. Berdini, J. E. Coyle, Y. Dudit, P. Grondin, P. Huet, L. Y. W. Lee, E. S. Manas, R. L. McMenamin, C. W. Murray, L. W. Page, V. K. Patel, F. Potvain, S. J. Rich, Y. sang, D. O. Somers, L. Trottet, Z. Wan, and X. Zhang, *Fragment-Based Approach to the Development of an Orally Biovailable Lactam Inhibitor of Lipoprotein-Associated Phospholipase (Lp-PLA$_2$). J Med Chem, 2016.* **59**(2): p. 10738-10749.

179. H. P. Kim, K. H. Son, H. W. Chang, and S.S. Kanga, *Anti-inflammatory Plant Flavonoids and Cellular Action Mechanisms. J Pharmacol Sci, 2004.* **96**(3): p. 229-245.

180. J. F. Ballesteros, M. J. Sanz, A. Ubeda, M. A. Miranda, S. Iborra, M. Paya, and M.J. Alcaraz, *Synthesis and Pharmacological Evaluation of 2'-Hydroxychalcones and Flavones as Inhibitors of Inflammatory Mediators Generation. J Med Chem, 1995.* **38**(14): p. 2794-2797.

181. I. Jantan, S. N. A. Bukhari, and O.A. Adekoya, *Studies of synthetic chalcone derivatives as potential inhibitors of secretory phospholipase A$_2$, cyclooxygenases, lipoxygenase and pro-inflammatory cytokines. Drug Des Devel Ther, 2014.* **8**: p. 1405-1418.

182. Chippaux, J.-P., *Venin de Serpents et Envenimations, ed. I. Editions. 2002. 288.*

# Chapter 2: Library design and characteristic analysis

In the introduction the rationale for exploring chalcones is presented and shows the importance of this class of molecules in the therapeutic area. Furthermore this chalcone family have known activities against diverse proteins. Derivatives of chalcones (bischalcones) have shown an inhibitory activity against the α-glucosidase protein which contributes to diabetes. [1] Idris *et al* have also shown the inhibitor potential of several chalcones scaffolds against COX-2 and microsomal prostaglandin E synthase-1 proteins.[2] These researchers also showed that certain moeity modifications of the chalcones improve their binding energy against MRSA. [3]  All of the examples presented in the previous chapter illustrate the importance of chalcone structure in potential drugs. Furthermore in literature this family of molecules are studied mainly in biological assays or high-throughput virtual screening research. However most of these studies show a limited interest in the contribution of common physical or chemistry denominator (H-bond donor, etc) towards activity. This avenue of research may explain the score of binding energies, at least, of chalcones against proteins. It was decided to use this family then as the basis for the design of our library. There are several ways in which chemical libraries may be generated, for example using the MS Combi (Schrödinger) [4], or by reading SMARTS reactions or MDL rxn files as a reaction, and enumerating the reaction possibilities within RDKit. [5] In this study, however, given that a python script framework was already in place in the project, the enumeration of a virtual chemical library was through the use of nested loops within this python framework.

## 1.  Library design

### 1.1.  Primary compound selection

The generation of the library of chalcone derivatives was based on the knowledge of the Claisen-Schmidt reaction. The corresponding chemical reaction scheme for this is represented in Figure 2.1. Phase I correspond to the Claisen-Schmidt reaction. The reactivity of the Claisen-Schmidt reaction will change relative to which substituents are present, but we assume that the Claisen-Schmidt reaction will at least provide access to adducts, even if the yield is poor. It was decided to extend the library by replacing the α,β unsaturated bond by an alkane chain and by adding moieties to this bond. Steps II and III are corresponding to this electrophilic addition of halogens to the α,β alkene chain of the chalcones It was realized later in the conception of the library that in phase III of the reaction scheme, molecules $III_c$ and $III_d$

would be difficult to prepare due to the steric bulk of the (R, R) and (S, S) configurations (Figure 2.1). This was cleare from the superposition of all precursor fragments (acetophenones, benzaldehydes and alkyl bromide) within Discovery Studio Visualizer. Based on this, only products $III_a$ and $III_b$ from this phase would therefore be realisable and this informed the choice to include only these in the virtual library. Given collaborations with an experimental group, the library necessitated a collection of accessible acetophenone, benzaldehydes and alkyl bromide derivates. These compounds were selected as those that are available to purchase from the commercial website of Acros Organic (https://www.acros.com). In order to further restrict the number of molecules from this site informing the library generation, two primary criteria were imposed to the selection. Further to the molecules having to be commercially available from the Acros website, a second requirement relating to the alkyl side chain was also used. This second criterion is the alkyl chain of the halogen derivate has to be short in order to avoid any steric blockage between the chalcones and different protein active sites. For example, if the alkyl chain ($R^1$ and $R^2$) from the alkyl bromide part is too long, the chain may collapse in molecules III to the acetophenone or/and benzaldehyde derivatives.
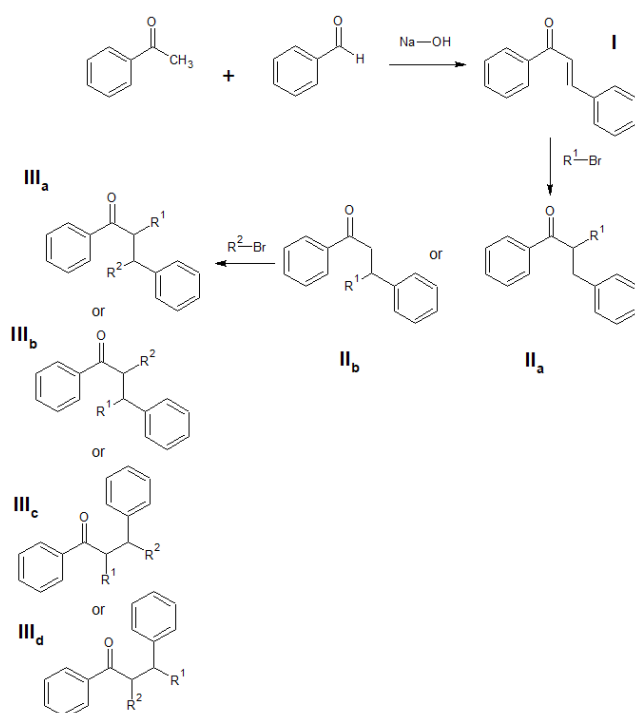


Figure 2.1: Claisen-Schmidt reaction and extensions process. After selection of the molecules, compounds $III_c$ and $III_d$ are impossible to be designed from the steric bulk of the (R,R) and (S,S) configuration.

Based on these two criteria, 84 acetophenones, 96 benzaldehydes and 19 alkyl bromides were used to generate a library of chalcones. Other criteria were imposed in the process of generation of structures, such as the nature of the atoms linked to the acetophenone and benzaldehyde derivatives. It was decided to privilege heteroatoms such as nitrogen or oxygen and avoid sulphur atom as benzene ring substituents of the acetophenones and benzaldehydes, in order to not allow the possibility of disulfides forming for compounds ultimately at the experimental stage. Given the purpose in mind for this library a criterion for the limitation of number of cycles (e.g.: cyclohexane or cyclohexene) as polycyclic substituents was further introduced. Even with these restrictions in place, the acetophenones and benzaldehydes used as the basis for this library possessed a wide range of functionalities including hydroxy groups, alkyl chains, and ether or amine moieties. The majority of alkyl bromides compounds used as this substituent also consisted of a conjugated bond. It was realized later in this work (after the analysis of the results of the high-throughput virtual screening), that duplicates of starting structures were present in the benzaldehyde set of structures. Filtering out of duplicates was effected later in the library generation process. The Figures 2.2 to 2.4, 2.5 to 2.6 and 2.7 represent respectively the 84 acetophenones, the 96 benzaldehyde derivates and alkyl bromides selected for the library.
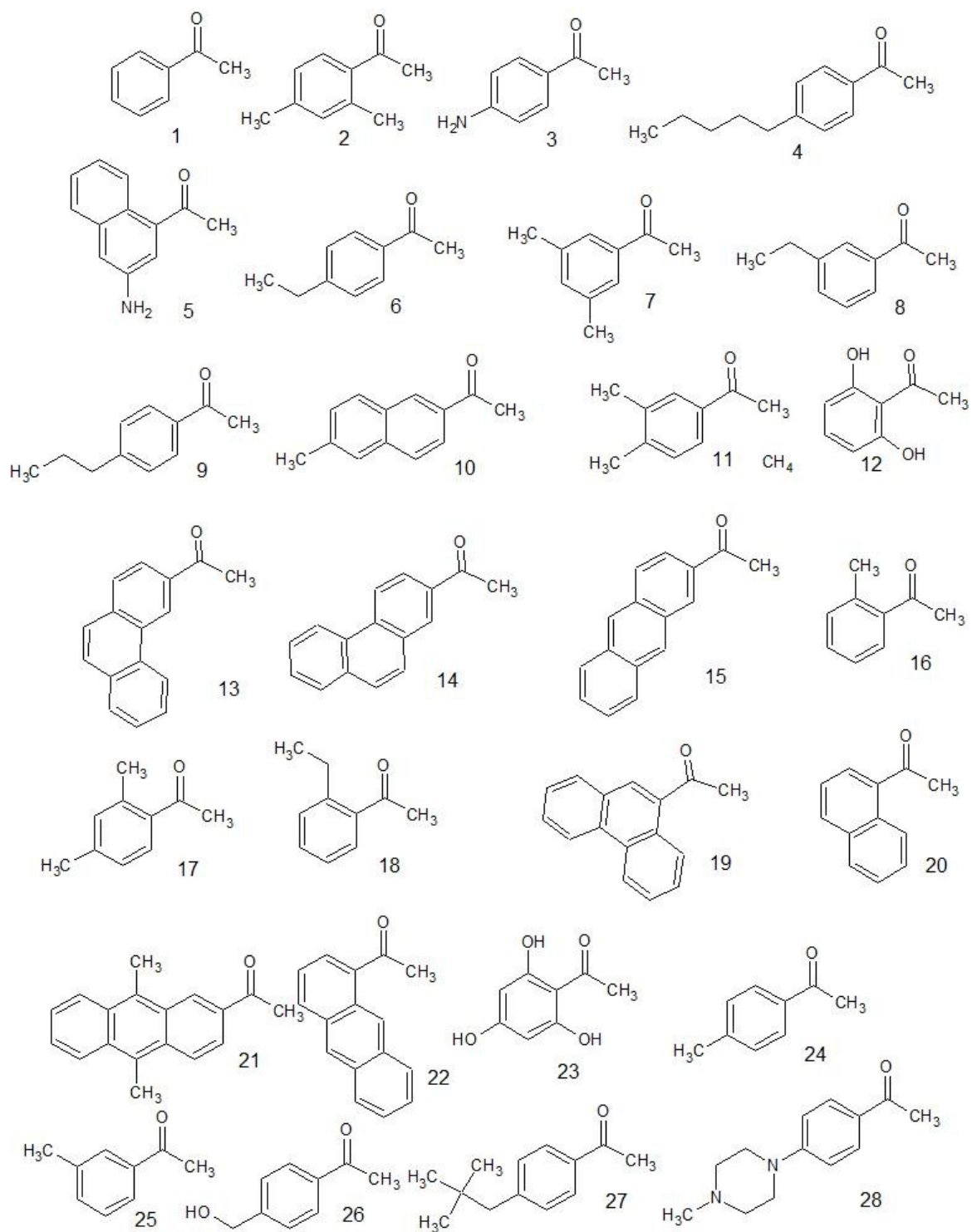
Figure 2.2: Acetophenones used as a basis for virtual library generation according to Figure 2.1.  A_1 to A_28 are presented in this Figure (1st parts).
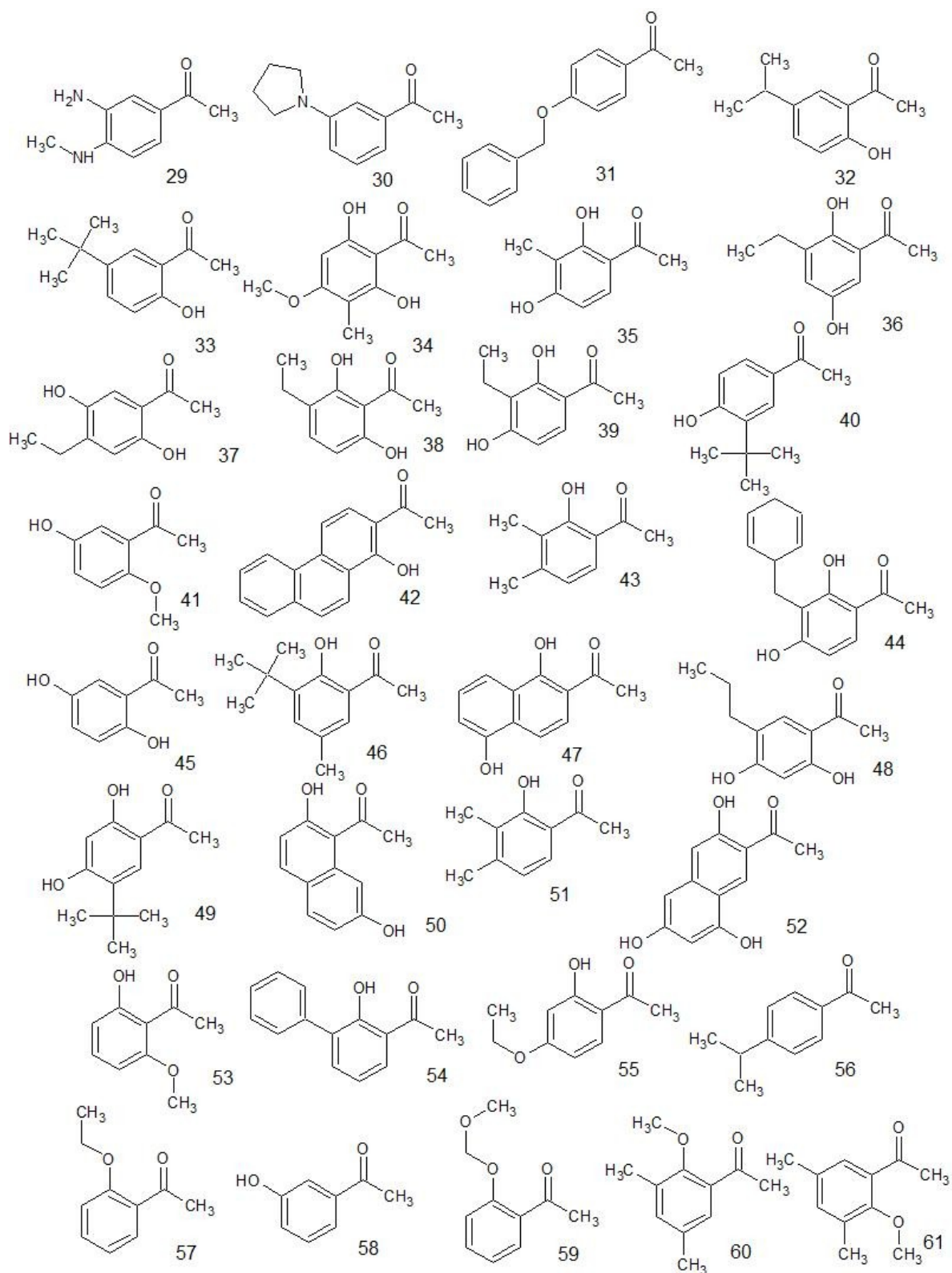
Figure 2.3: Acetophenones used as a basis for virtual library generation according to Figure 2.1. A_29 to A_61 are presented in this Figure (2<sup>nd</sup> parts).
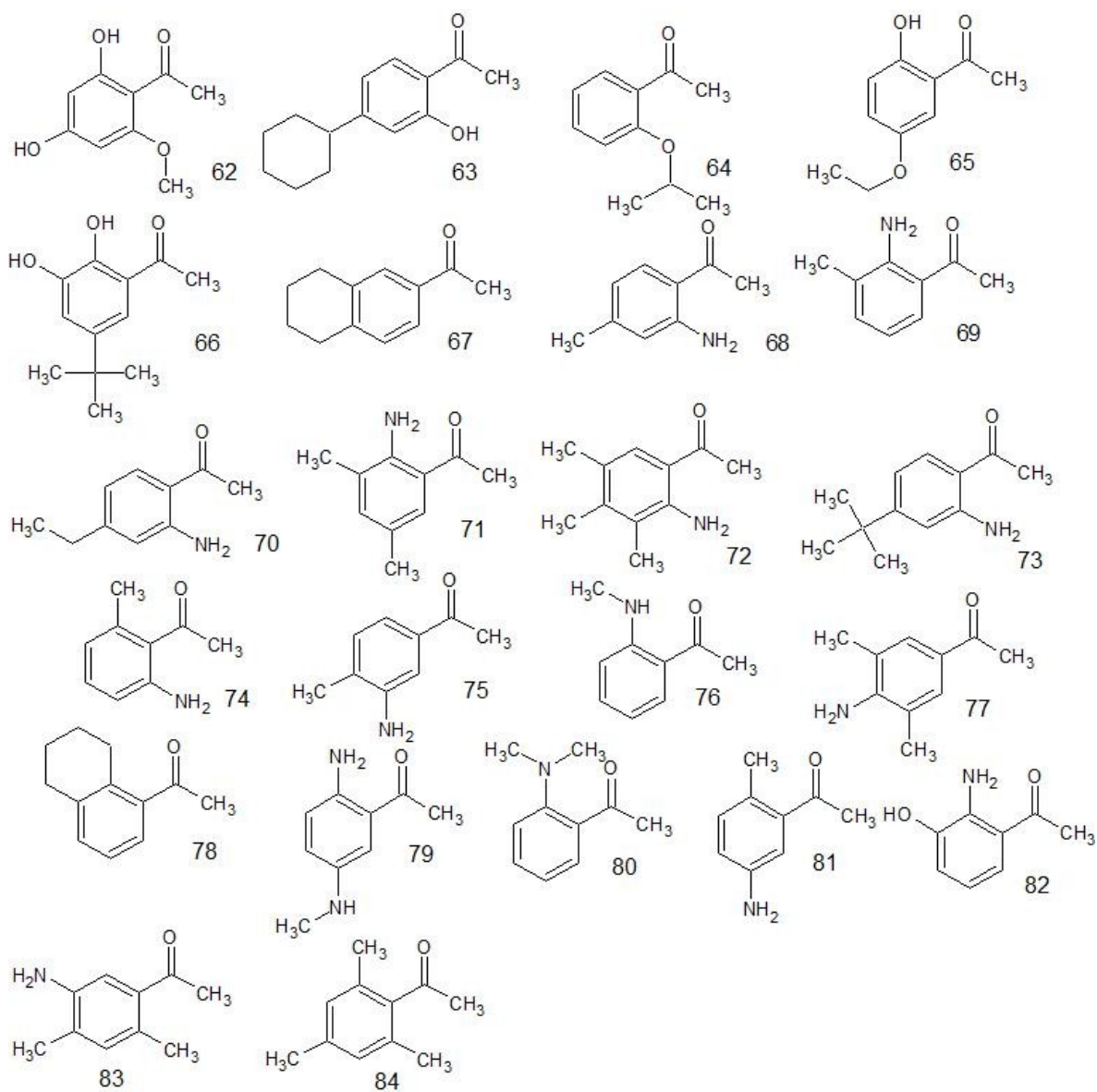
Figure 2.4: Acetophenones used as a basis for virtual library generation according to Figure 2.1.  A_62 to A_84 are presented in this Figure (3rd parts).
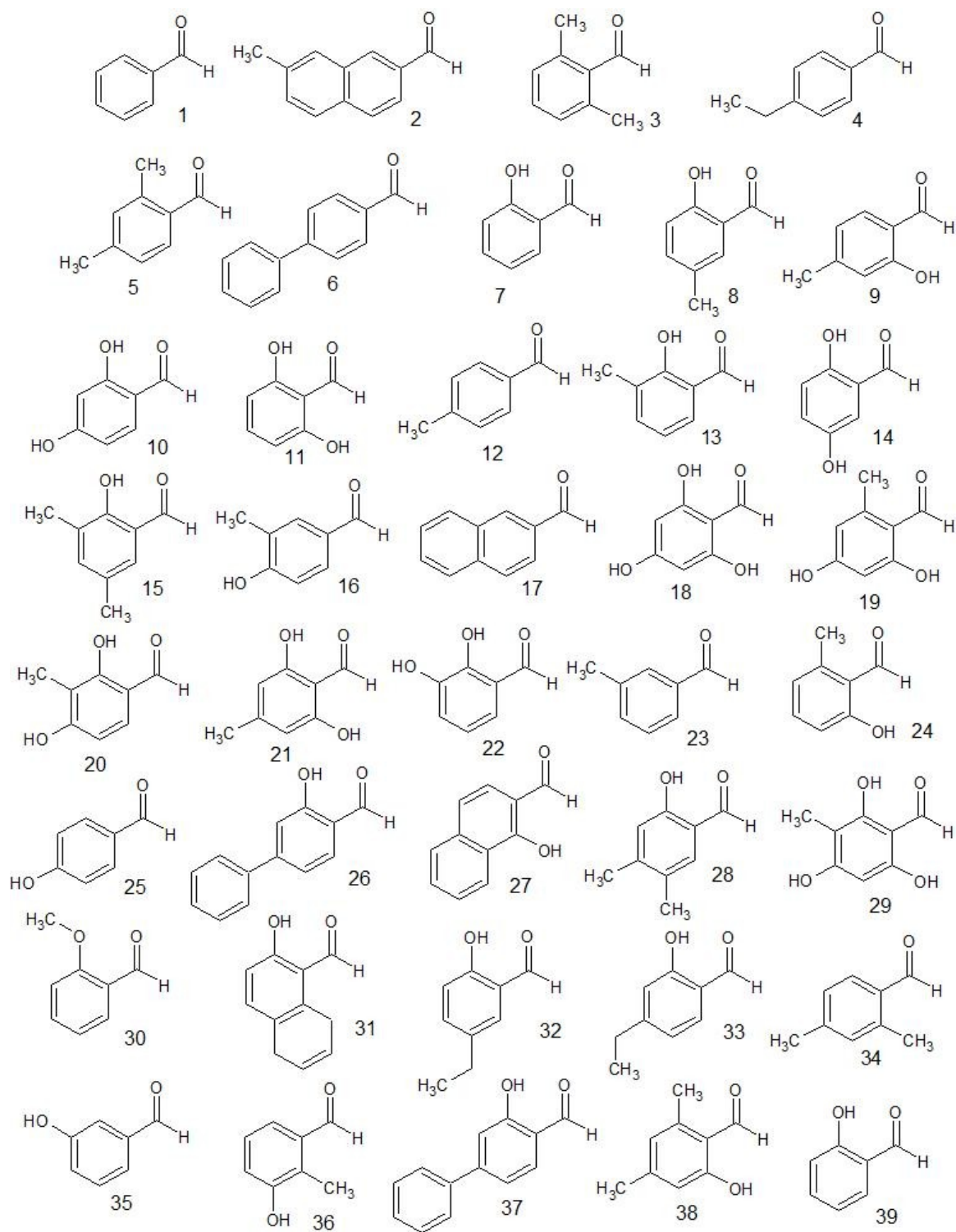
Figure 2.5: Benzaldehydes derivated used as a basis for virtual library generation according to Figure 2.1.  B_1 to B_39 are presented in this Figure (1st part).
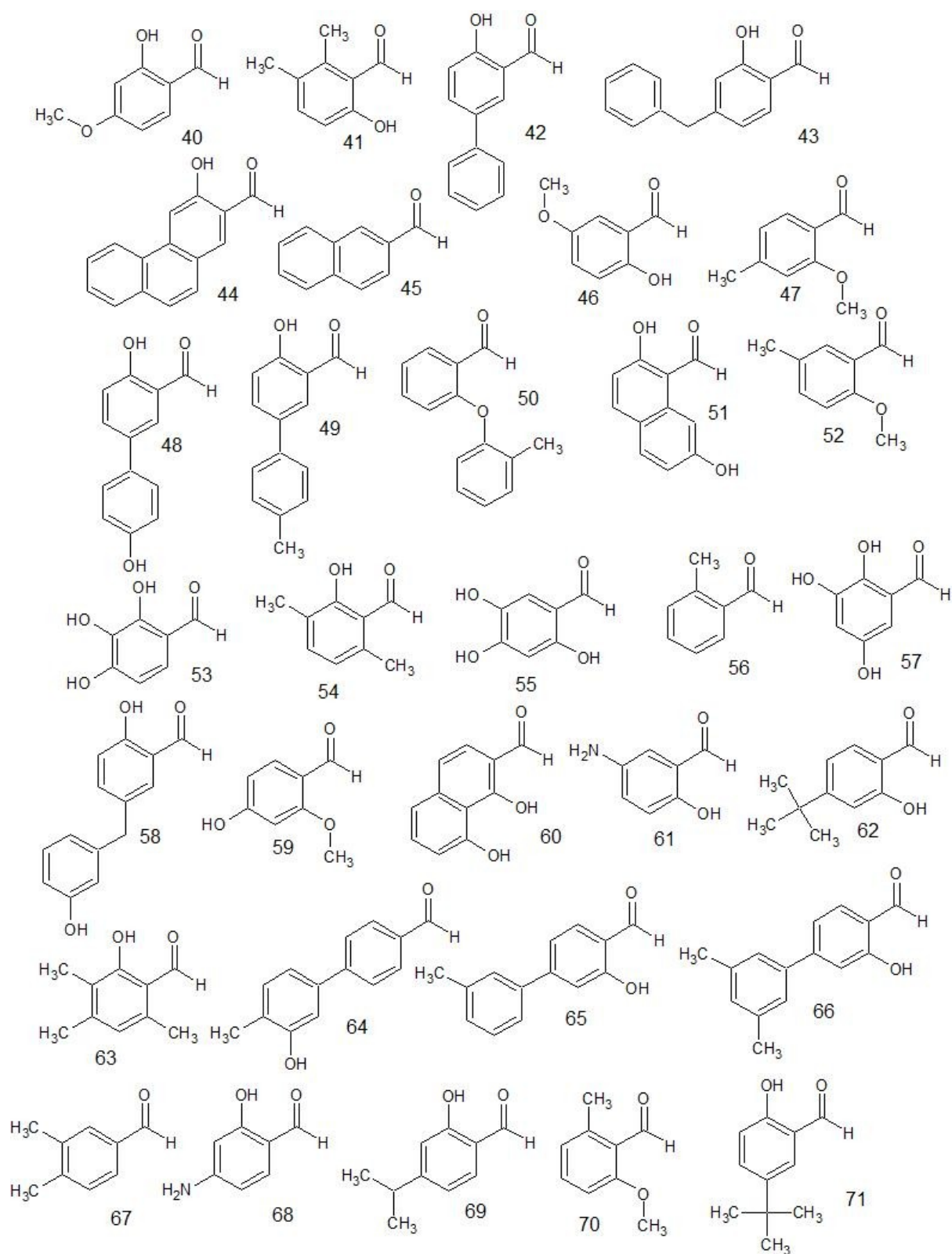
Figure 2.6: Benzaldehydes derivated used as a basis for virtual library generation according to Figure 2.1. B_40 to B_71 are presented in this Figure (2nd part).
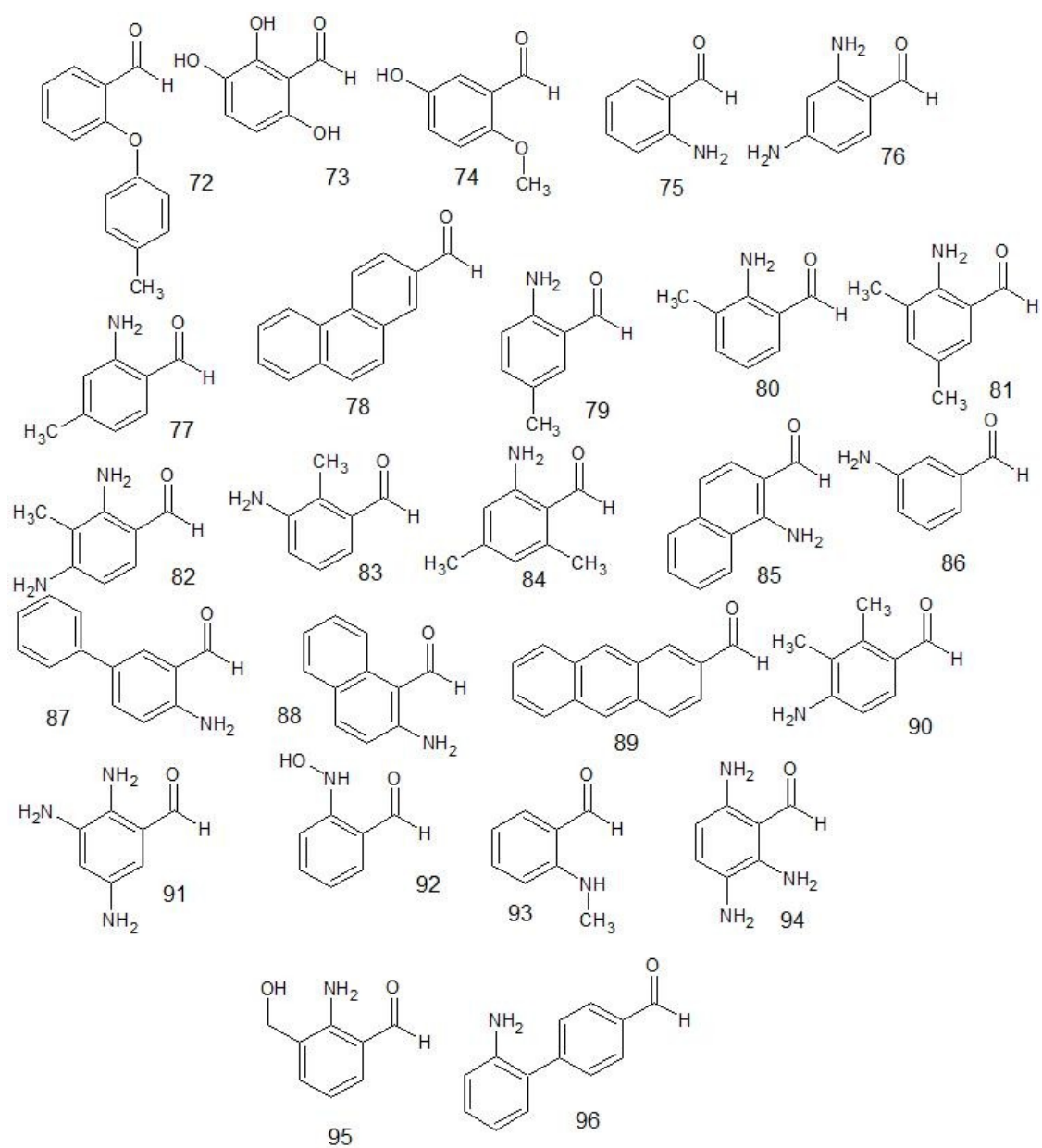
Figure 2.7: Benzaldehydes derivated used as a basis for virtual library generation according to Figure 2.1. B_72 to B_96 are presented in this Figure (3rd part).
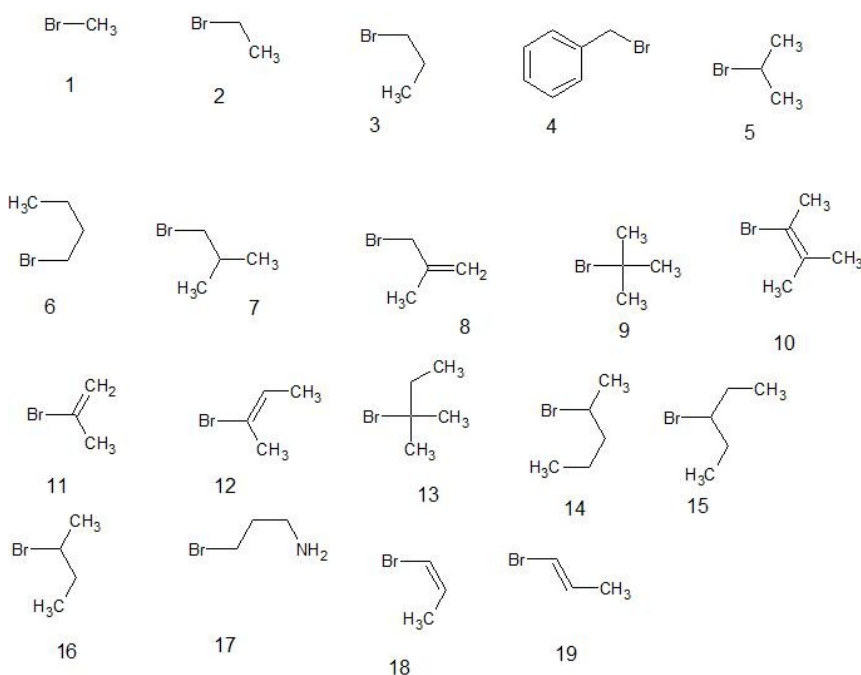
Figure 2.8: Alkyl bromide derivates used as a basis for virtual library generation according to Figure 2.1.

## 1.2. Preparation of the primary compound

After the selection of the starting compound structures, each of these materials was given a unique code composed of a letter and number for easy identification. For example the 15th acetophenone in the list of acetophenone starting materials was assigned the code A_15.

In order to generate the products of the chemical reactions according to the scheme, each family of starting compounds was optimized, with appropriate conformations chosen. The structures in each family of compounds were aligned spatially using the first compound in each series, A_1 and B_1 as templates for this superposition. Discovery Studio Visualizer [6] was used to effect this superposition, with five atoms in each pair of molecules used as reference points to superimpose firstly all of the acetophenone and secondly all of the benzaldehyde precursors. The results of the superposition of the three types of molecules gave planar aligned molecules which might be interpreted by software as a single molecule (chalcone) if the elemental rules of chemistry are respected. For superposition of the alkyl bromide species, due to the small number of atoms, only the bromide atom and its adjacent carbon were taken as a reference for this superposition. Each set of geometrically aligned compounds were saved in a unique file of sdf format. The result of this was three sdf files: for the acetophenone; the benzaldehyde; and the alkyl bromide moieties. However the electrophilic addition of the bromide derivatives to the $\alpha,\beta$ unsaturated bond of the chalcone

may be done to the α position or β position or both in the same time. In order to respect this reaction, the process was repeated twice for the alkyl bromide for R and S configurations. The fixed points are shown in Figure 2.9. Ultimately the alkylation was not included in this library (but the system is in place for this extension).
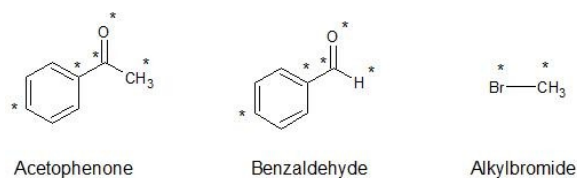


Figure 2.9: Atoms used as fixed point. They are represented by an asterisk.

The python script developed to effect the generation of chalcone (and derivative) product libraries took as input these 5 sdf files (acetophenone library, benzaldehyde library and alkyl halide library), modifying structures as appropriate to account for the chemical change. The script handled appropriate file formatting and generation as a product sdf file. The script was designed to appropriately merge the acetophenone and benzaldehyde unique IDs to create a further unique id for each element of the product library, by recognizing keywords within the sdf file each time the program read a specific keyword such as "acetophenone" or "benzaldehyde". The product structures ID therefore is named "acetophenone_[number]_benzaldehyde_[number]". Chemical change was handled by appropriate dropping or repurposing of atoms to the new chemical functionality. For example, the appropriate two hydrogens atoms of the acetophenone and one oxygen atoms of the benzaldehyde (corresponding to water elimination) were automatically removed from each product file structure. After this process, 8 063 molecules for phase I of the reaction scheme were generated. However, after screening for duplicate structures, the library was found to present 7 728 unique molecules. The completion of generation of phase II and III molecules to extend this library was left as future work to be explored. Figure 2.10 represents a simplified version of the workflow followed for the generation of the library.
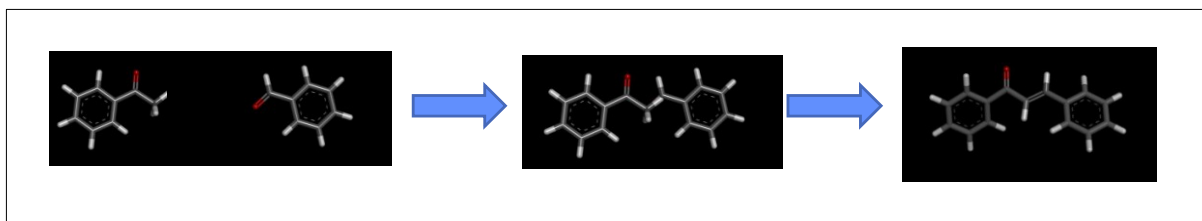
The entire library was further submitted to geometrical optimization using the semi-empirical AM1 method, within Gaussian 09 [7]. AM1 is not a particularly accurate method to use; however given the size of the library, this method was chosen due to the ability then to obtain complete results in reasonable time. In order to effect this, the pdb files were converted to Gaussian ".com" input format using the OpenBabel software [8]. A script was written to effect this optimization of all library compounds using an OpenPBS batch queuing system, with appropriate calls to OpenBabel where chemical file format conversion was necessary.

## 1.3. Characteristic of the library

### 1.3.1. Chemical/Physical proprieties

The physicochemical proprieties (MlogP, hydrogen bond acceptor and donor, topological polar surface area, molecular weight and heavy atoms count) of the chalcone library were calculated using RDKit. The workflow in which these values were obtained is described in chapter 4. The molecular weight fluctuates between 200 and 450 Daltons, the Mannhold LogP between 2.5 and 5, and the heavy atom count between 15 and 35 atoms. The topological polar surface area of the chalcone library is also in the range of 20 to 160. Naturally the chalcone library did not possess any molecules with no hydrogen bond acceptor (due to the presence of a ketone). The number of hydrogen bond acceptors, however, fluctuated between 1 and 4 with the exception of the chalcones which derived from benzaldehyde **91** (which has many hydroxyl groups present, the number of hydrogen bond acceptor reached 5 or 7). There was a range of 0 to 6 hydrogen bond donors present per molecule, although the majority of molecules had between 0 and 4 hydrogen donors. All chemical/physical proprieties are presented in form of diagram in Figure 2.11.
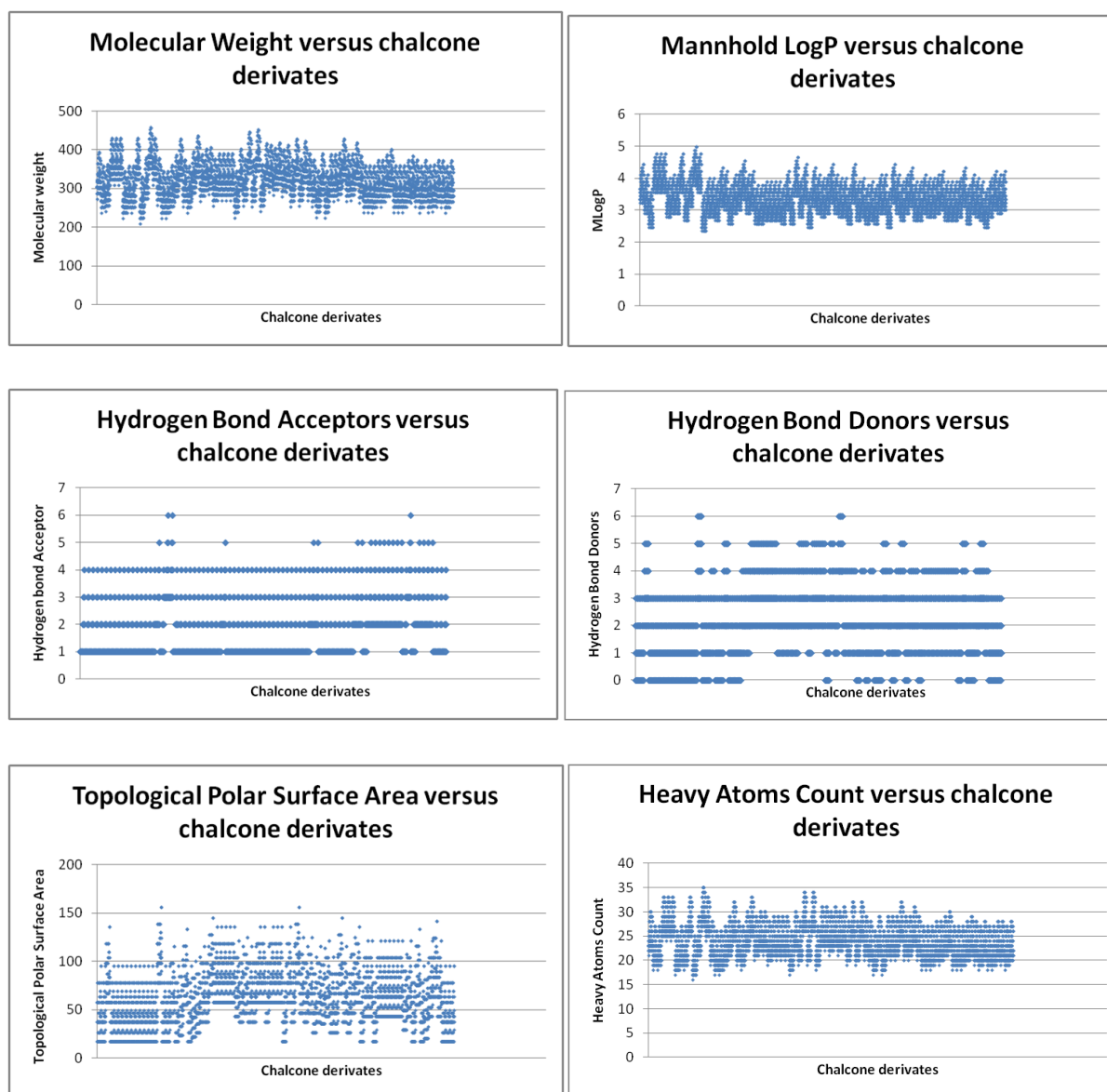
Figure 2.11: Graphs of different chemical/physical proprieties versus the chalcone derivates. *From the top left to the down right*: molecular weight, MlogP, hydrogen bond acceptor, hydrogen bond donors, topological polar surface area and heavy atoms count. The arbitrary x axis represents each chalcone within the library.

In terms of the Lipinski rule of five, the library only presented violation of the rules regarding hydrogen bond donors and Mannhold logP.[9] The molecular weight of all molecules is under 500 Daltons and the hydrogen bond acceptor count is less than 10 for all molecules of the chalcone library. Under Lipinski's rules, there should be less than 5 hydrogen bond donors per molecule. There were a few compounds, those derived from acetophenones_23 and 52 and from the benzaldehydes 18, 29, 53, 55, 57, 73, 91, 95, which possessed 6 hydrogen bonds, but the remainder were within limits. Further, the Mannhold octanol-water partition index is

over 4.15 (instead of 5 due to this specific method of calculation) for only 320 molecules within the library, which corresponds to another violation of the Lipinski rule of five. Since there are fewer molecules in the library that represent violations of Lipinski's rule of five, than those that don't, some of the violators are highlighted here. Acetophenone_28_benzaldehyde_91, acetophenone_29_benzaldehyde_94 and acetophenone 79_benzaldehyde_91 which possess 6 hydrogen bond acceptors are represented in Figures 2.12, 2.13 and 2.14 respectively. It is observed that violations are due to the presence of large numbers of nitrogen atoms.
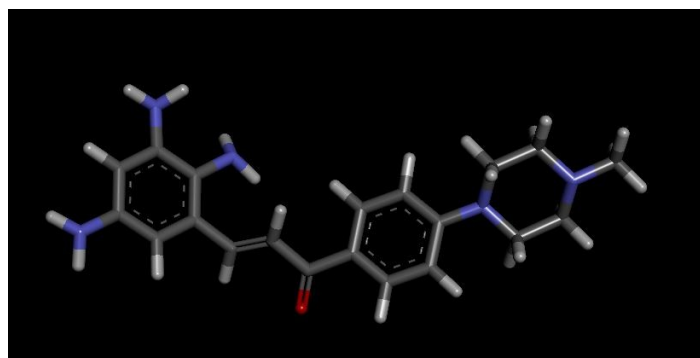


Figure 2.12: Acetophenone_28_benzaldehyde_91.
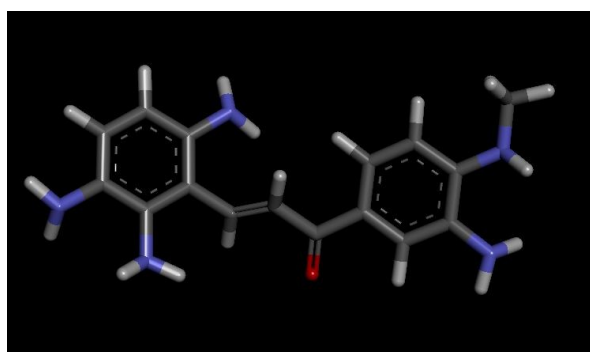


Figure 2.13:Acetophenone_29_benzaldehyde_94.

Violation of hydrogen bond donor numbers is not only due to nitrogen atoms. 6 hydrogen bond donors are also present for a number of oxygen containing compounds. Figures 2.15, 2.16 and 2.17 show respectively acetophenone_23_benzaldehyde_29, acetophenone_23_benzaldehyde_73 and acetophenone_52_benzaldehyde_29 which possess numbers of phenolic OH groups.
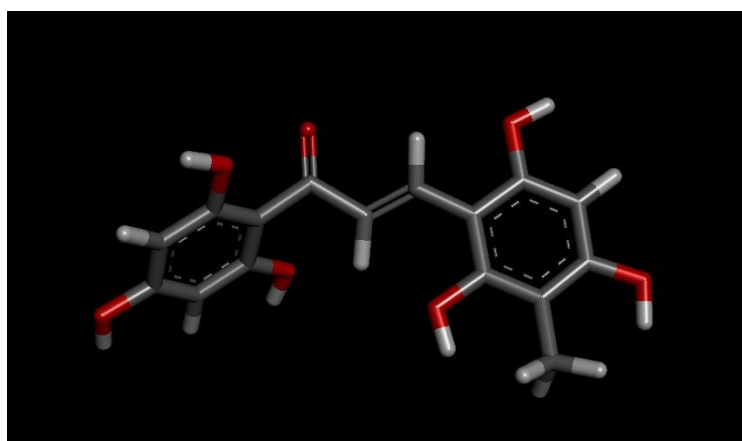


Figure 2.15: Acetophenone_23_benzaldehyde_29
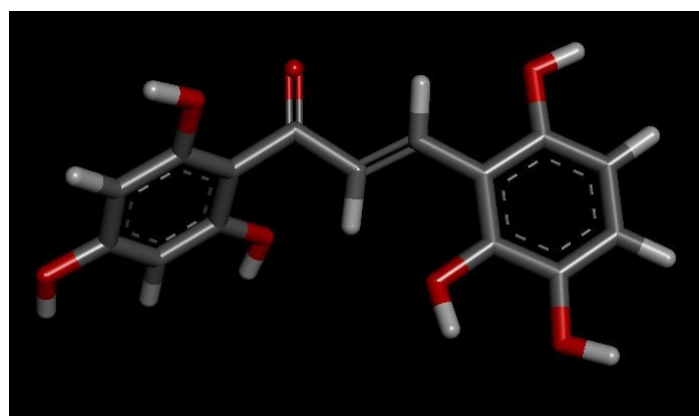


Figure 2.16: Acetophenone_23_benzaldehyde_73.
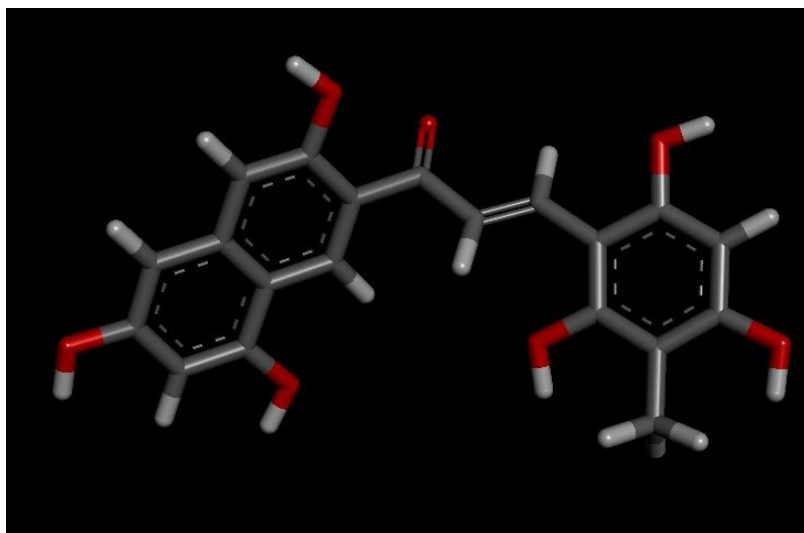
Figure 2.17: Acetophenone_52_benzaldehyde_29.

The last example of molecules violating Lipinski's rule of 5 are those in which the Mannhold LogP value is greater than 4.15. This characteristic is brought about in molecules with very few heteroatomic substituents, such as the molecules acetophenone_21_benzaldehyde_64, acetophenone_22_benzaldehyde_89 and acetophenone_22_benzaldehyde_78, which are represented in Figures 2.18, 2.19 and 2.20 respectively.
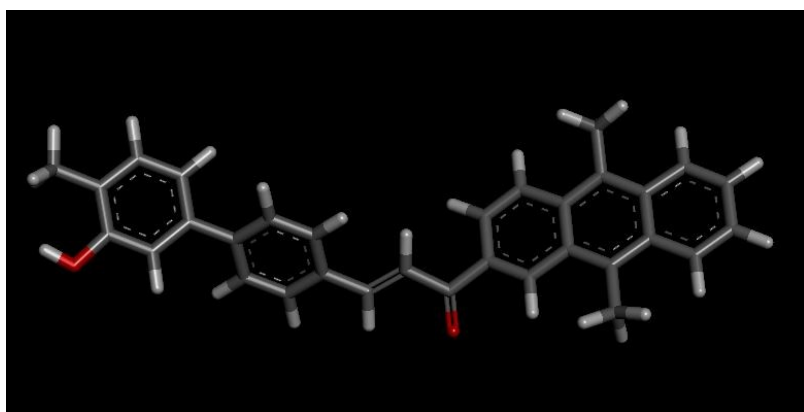


Figure 2.18: Acetophenone_21_benzaldehyde_64 (MlogP: 4.76).
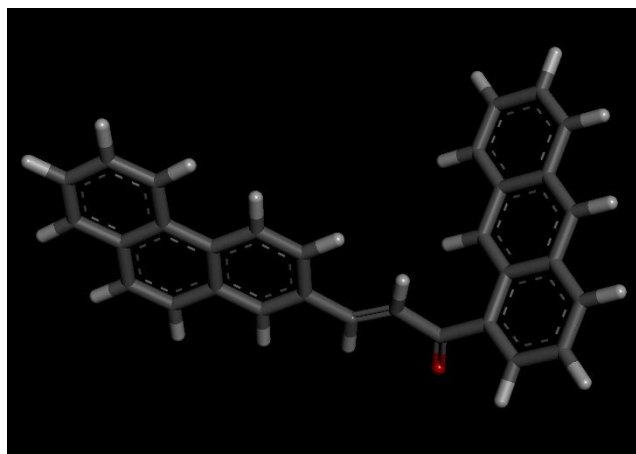
Figure 2.19: Acetophenone_22_benzaldehyde_89 (MlogP: 4.76).
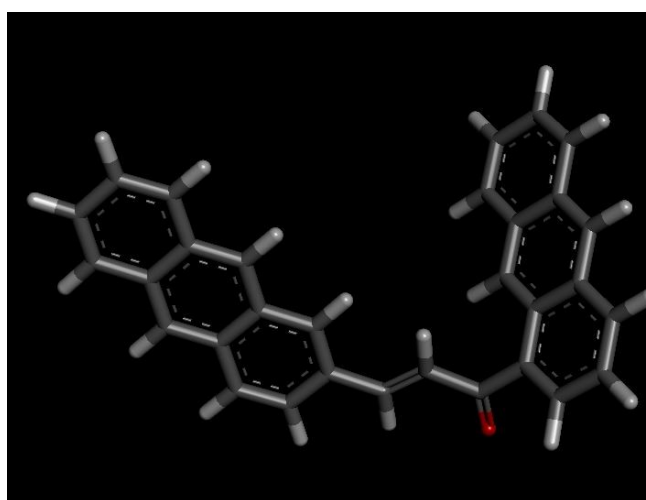


Figure 2.20: Acetophenone_22_benzaldehyde_78 (MlogP: 4.76).

Chemical space was also investigated. Given the very focussed nature of this library, naturally not a wide area within all accessible chemical space is covered. However, the variations within the library were of interest. The Mannhold LogP and topological polar surface provide a general metric of the shape of the chalcones. In order to visualize the diversity within the library, the Mannhold LogP versus the topological polar surface area of the library was plotted and this is represented in Figure 2.21. The MlogP decreases in general when the topological polar surface area increases. This may be explained in terms of the decrease in hydrophobicity of the chalcone as more and more polar substituents are added.

**MLogP versus topological polar surface area**

*MLopgP* (y-axis)

*Topological polar surface area* (x-axis)
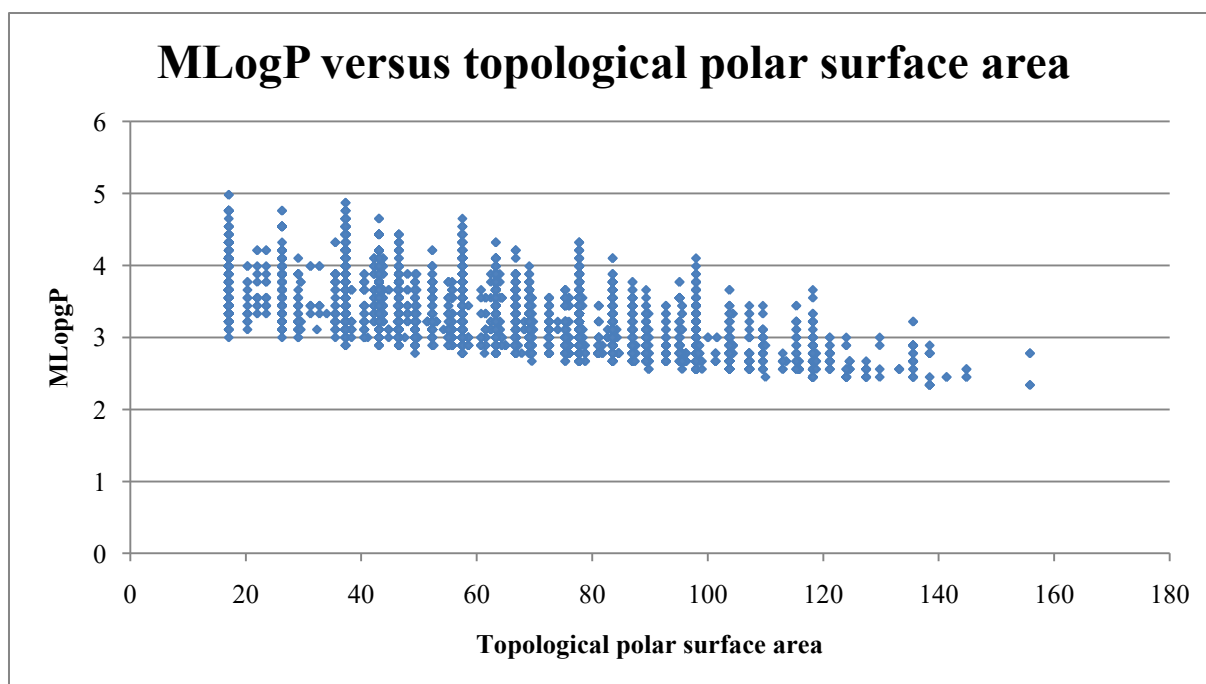
Figure 2.21: Representation of the chemical space by the graph of Mannhold LogP versus the topological polar surface area.

### 1.3.2. Stability and reactivity

It was also of interest (in terms of further reactivity of the chalcones, or in the formation of covalent bonds to substrates) to investigate the molecular orbitals of compounds within the library. This purpose, the HOMO and LUMO energy values were extracted from the final step of geometry optimization. From these values, the band gap was calculated and plotted in the graph together with the HOMO and LUMO energy values. Figure 2.22 represents molecular orbital energy values and the band gap against the chalcone derivatives.

It is interesting to note 373 compounds such as acetophone_38_benzaldehyde_79 or acetophenone_62_benzaldehyde_83 possess a positive energy value for the LUMO (these 373 compounds have a LUMO energy varying between +0.03122 and +0.21908 Hartree). The band gaps are very small and may be influenced by the length of the chalcone molecules, which may allow greater freedom for electrons to circulate from one side of the molecule to the other. The reproducibility of orbital energy values obtained was anticipated, and anticipated to be due to the common α,β-unsaturated linker.
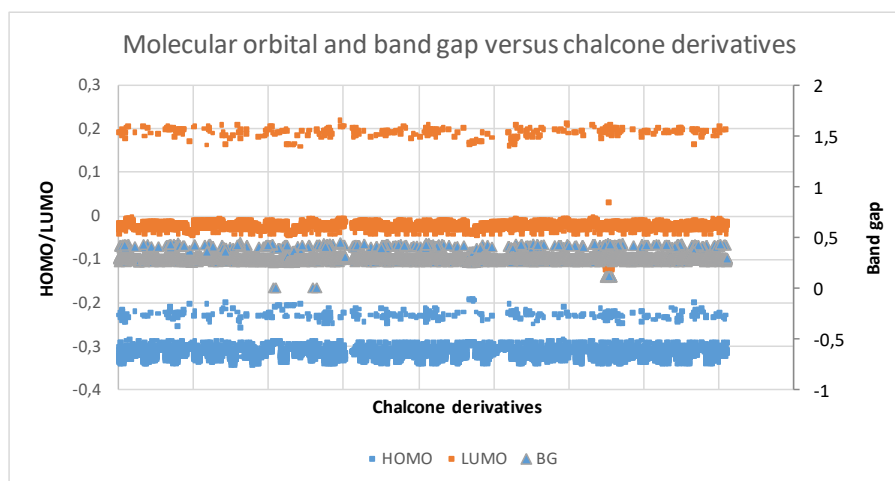
**Figure 2.22: Molecular orbital and band gap versus chalcone derivatives.**

In order to confirm this, recalculation and visualization of the HOMO/LUMO of a chalcone ( in this case one which possesses a positive LUMO energy value was chosen) was performed using Orca software.[10] The molecule acetophone_1_benzaldehyde_78 which had a LUMO energy value of +0.17756 Hartree was selected. To this end, the pdb format was translated into an Orca inp file, together with the appropriate directives. The further optimization was at the density-functional theory level using the pure functional BP using the RI implementation and the def2-TZVP basis set. Orbital pictures were generated using orca_plot. The values of the LUMO and HOMO obtained from this calculation were -0.208818 Hartree and -0.119524 Hartree. The HOMO calculated from Orca is close to be the same calculated previously (-0.22875 Hartree). However the Orca LUMO value was lower than that calculated previously (+0.17756 Hartree). Figures 2.23 and 2.24 represent respectively the HOMO and LUMO orbitals for acetophenone_1_benzaldehyde_78. Both frontier orbitals are, as expected, concentrated around the α,β-unsaturated system with the LUMO over the double bond (corresponding to a site for conjugate addition) and the HOMO centred on the lone pairs of the ketone oxygen.
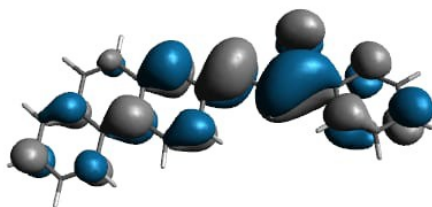
Figure 2.23: Representation of the LUMO orbital of the acetophenone_1_benzaldehyde_78. Code color. Blue, positive orbital; grey, negative orbital.
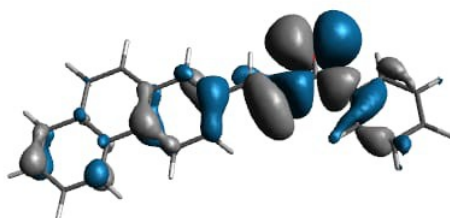


Figure 2.24: Representation of the HOMO orbital of the acetophenone_1_benzaldehyde_78. Code color. Blue, positive orbital; grey, negative orbital.

## Conclusion

Generation of a chemical library using custom python scripts, optimization of the resultant structures at the QM level and characterization of the library in terms of physicochemical properties and reactivity was successfully completed. For the focussed nature of the library, there is still some diversity in terms of chemical functionality, and most of the library possesses favourable drug-like properties.

# REFERENCES

1.	Cai, C.-Y., L. Rao, Y. Rao, J.-X. Guo, Z.-Z. Xiao, J.-Y. Cao, Z.-S. Huang, and B. Wang, *Analogues of xanthones Chalcones and bis-chalcones as $\alpha$−glucosidase inhibitors and anti-diabetes candidates.* Euro J Med Chem, 2017. **130**: p. 51-59.
2.	Idris, M.H.M., S.N.M. Amin, M. Selvaraj, H. Jamari, T.L. Kek, and M.Z. Salleh, *High-Throughput Structure-Based Drug Design of Chalcones Scaffolds as Dual Inhibitor of Cyclooxygenase-2 and Microsomal Prostaglandin E Synthase-1.* J Pharm Sci Emerg Drugs, 2018. **6**(1): p. 1-14.
3.	Cushnie, T.P.T. and A.J. Lamb, *Recent advances in understanding the antibacterial properties of flavonoids.* Int J Antimicrob Ag, 2011. **38**(2): p. 99-107.
4.	Schrödinger, *Schrödinger Release 2018-4: MS Combi*. 2018: New York.
5.	Landrum, G. *RDKit: Open-source cheminformatics*.
6.	BIOVIA, D.S., *Discovery Studio Modeling Environment*. 2016, San Diego: Dassault Systèmes.
7.	Frisch, M.J., G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, M.C. X. Li, A. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F.L. F. Ding, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery, Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, and D.J. Fox, *Gaussian 09*. 2016, Gaussian, Inc.: Wallingford CT.
8.	O'Boyle, N.M., M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison, *Open Babel: An open chemical toolbox.* J Cheminformatics, 2011. **3**(33): p. 1-14.
9.	Lipinski, C.A., F. Lombardo, B.W. Dominy, and P.J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv REv, 2001. **46**(1-3): p. 3-26.
10.	Neese, F., *Software update: the ORCA program system, version 4.0.* Wires Comput Mol Sci, 2018. **8**(1): p. 1-6.

# Chapter 3: High-throughput virtual screening and molecular dynamics

Following the generation of the virtual library, questions around suitability and specificity of this library against a range of targets were to be explored. To this end high throughput virtual screening of the library of chalcones was performed against the range of targets described, using molecular docking mediated by AutoDock Vina. A validation of docking was made with the original drug provided by the literature and compared to the top five best (lowest energy) compounds. Some of the best docked complexes (with the HSP90 protein) were taken further and studied using molecular dynamics.

## 1. High-throughput virtual screening

### 1.1. AutoDock Vina

AutoDock Vina, is open source software, which is able to make use of multi-core computer systems, and as such was appropriate for the high-throughput virtual screening in this study. Using a scoring function, this program simulates the interaction between a ligand and a target (calculating and exploring ligand, and sometimes target conformational space) and calculates the binding energy of the resultant ligand-protein complex. The scoring function is based on a hybrid function which is composed of both empirical and knowledge-based functions. The binding energy is calculated from the sum of the different interactions of atom pairs, dependent on surface distance (taking into account both the interatomic distance and the atomic radii of both of the pair of atoms). [1] The scoring function may be expressed in short in Equation 3.1.

$$E_{binding} = \sum e_{pair}(d) \quad (3.1)$$

The sum of the atom pair interactions may be decomposed into different types of interactions. These interactions are comprised of two separate Gaussian functions (describing the attractive term for the binding energy), and a third term describing the repulsions. The third steric term is equal to the square of the distance if the overlap of atoms is larger than zero. These equations are respectively represented in Equations 3.2, 3.3 and 3.4. [1]

$$gauss_1(d) = e^{-\left(d/0.5\right)^2} \quad (3.2)$$

$$gauss_2(d) = e^{-\left((d-3)/2\right)^2} \quad (3.3)$$

$$repulsion(d) = \begin{cases} d^2, if \ d < 0 \\ 0, if \ d \geq 0 \end{cases} \qquad (3.4)$$

Also included are hydrogen bond (acceptor and donor), hydrophobic interactions and a number of torsions between four atoms of rotatable bond. These values also account for variation of entropy.[2] The complete expression for free energy of binding is represented in Equation 3.5:

$$\Delta G_{binding} = \Delta G_{gauss} + \Delta G_{repulsion} + \Delta G_{hbond} + \Delta G_{hydrop \ hobic} + \Delta G_{torsion} \quad (3.5)$$

The search algorithm is optimized both in terms of acceleration of the calculation and also in terms of the generation of accurate poses. It combines two methods: a mutation-based method combined with a local Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. Both of these methods are an implementation of the quasi-Newton method for finding the minima of the scoring function. This combination helps in general to solve the non-linear optimization problem. In the case of AutoDock Vina, the scoring function and its gradients are used to this purpose. The BFGS takes in consideration the derivatives of the position, orientation and torsions to calculate the required rotation of the torsions. [3]

The search for the best (lowest energy) conformation is performed using a stochastic Monte Carlo method. The binding energy at each point is calculated and compared to the energy of the previous position. If the new value is lower than the old one, the new position is accepted. If the result is higher from the old one, a Boltzmann-Statistical criterion is applied. If $e^{(-\Delta E/kbT)}$ is well above zero, the position is retained. Otherwise the new position is rejected. The process is repeated unless the maximum of retained or rejected positions is reached. The calculation stops when the number of cycles end. [1]

The software uses the multithreading and allows calculations to be performed in parallel.[1]

## 1.2. Preparation of the targets

The targets used for the high-throughput virtual screening were downloaded from the website of RCSB PBD (https://www.rcsb.org) in the pdb format.[4] The selection was made following one main criterion: the absence of mutation (with the exception of the drug resistant HIV-1 integrase). From the eleven targets, five targets (COX-1, COX-2, MAOB, acetylcholinesterase A and B) possess two co-crystal structures inside of their files. It was found that the structure of the pyruvate kinase is complex with 4 subunits. After a literature

search, it was decided the protein would be separated in two structures: the AB domain (which retains responsibility of the activity of pyruvate kinase) and the C domain (which contains an allosteric site). All of the structures had water, ions and crystal structure ligands removed before being converted to pdbqt format. All proteins with their PDB ID are represented in Table 3.1.

| Enzyme | PDB ID | Reference |
|---|---|---|
| Integrase | 5HOT | [5] |
| Pyruvate kinase | 3T05 | [6] |
| HSP90 | 5FNC | [7] |
| COX-1 | 3N8Z | [8] |
| COX-2 | 5F19 | [9] |
| ALR$_2$ | 5JH1 | [10] |
| MAOA | 2Z5X | [11] |
| MAOB | 5MRL | [12] |
| AChE | 4M0E | [13] |
| BChE | 5DYW | [14] |
| PLA$_2$ | 5LYY | [15] |

Table 3.1: Proteins used for the high-throughput virtual screening.

## 1.3. Parameters

For successful molecular docking during the screening process, sensible parameters for use by Vina were determined. The first parameter, the center of the search area within the target was determined by localisation of the active site for each of crystal structures. This location was informed by literature pertaining to each of the crystal structures. Determination of this location was aided in many cased by the presence of crystal structure ligand (a known inhibitor). In two of the systems with no crystal structure ligand, the acetylcholine esterase (AChE both chains A and B) and the aldose reductase (ALR$_2$) structures two active site residues were used with a simple center point between these was used for the center of search (Figure 3.1 a); however, in general four active site residues were used to define the center of search (Figure 3.2 b). For the integrase structure, the inhibitor binds to an allosteric site, and this site was used for molecular docking. In the case of the integrase, N-4 in the crystal structure ligand 6-chloro-4-N-[(4-methylphenyl)methyl]pyridine-2,4-diamine provided the coordinates for the center of search. With the pyruvate kinase, two domains of the homotetramer were treated with blind docking. The center of each of these domains was used as center of the grid box and use the maximal size of the box allowed (126/126/126). Figures 3.1 and 3.2 respectively represent a diagram, formula to determine the active site center and the coordinates of the active site of each protein.
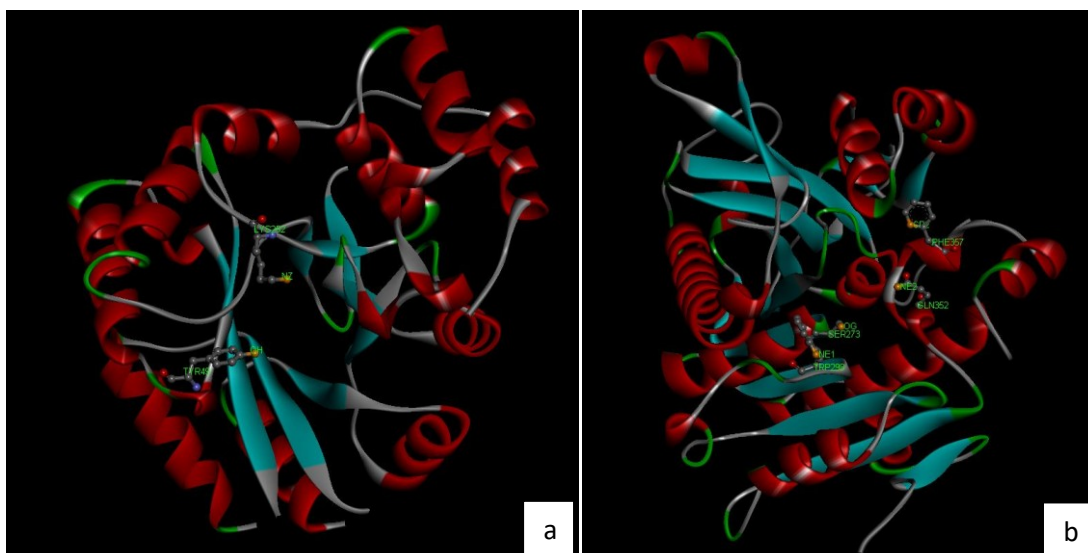
Figure 3.1: Methodology of the search of the center point of the active site of the protein. a; protein $ALR_2$ using two active site residues, b; protein $sLAP_2$ using four active site residues.



$$\overrightarrow{OASC} = \frac{(\overrightarrow{OAA_1} + \overrightarrow{OAA_2})}{2}$$

or

$$\overrightarrow{OASC'} = \frac{\sum \overrightarrow{OAA'}}{4}$$

(applied to x,y and z axis)
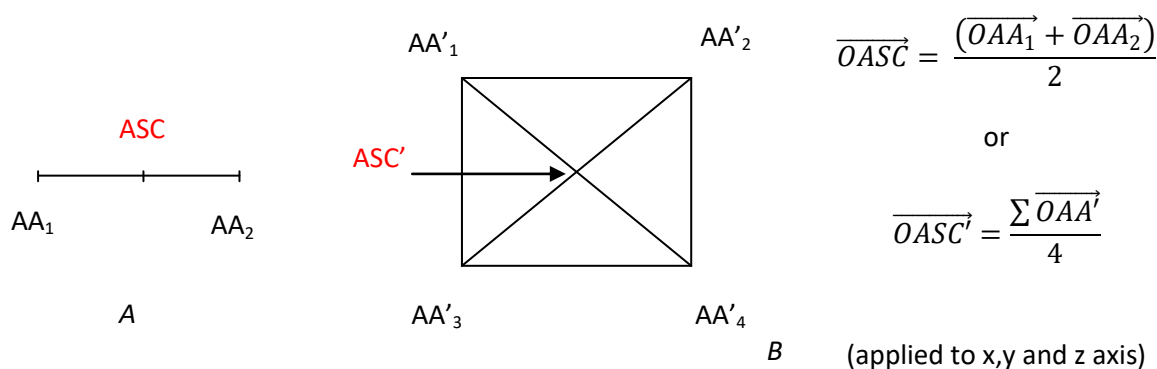
Figure 3.2: A represents the method used for the AChE (structure A and B) and $ALR_2$ and B represent the square method used for the rest of the proteins. AA: amino acid residue; ASC: Active Site Center.

Calculations were performed at the Center for High Performance Computing (CHPC, Cape Town). Docking experiments were set to use 24 cores (the number of cores per host). The energy range for discrimination was set to 4. For blind docking the exhaustiveness was set to 240 in comparison to other runs in which the exhaustiveness was set to 120 (the increase was due to the increase in search space for blind docking). All of the parameters for molecular docking are represented in Table 3.3.

A script was written to collate the parameters of the docking experiments and the values related to the grid box (Cartesians coordinates and sizes of the boxes) and write them to separate AutoDock Vina input files. The script detected the presence of the prepared crystal structure targets and used this name (together with individual ligand names) to uniquely

91

identify and name all vina input files. This crystal structure name enabled automation of entering of the grid box parameters. A portion of the pseudo code is presented in Figure 3.3.

```
for protein in protein_list:
        for ligand in ligand_directory:
                create_vinafie
                write_protein_name_and_path
                write_ligand_name_and_path
                write_specific_output_name
                if prot name == "AchE_4m0e_A_apo":
                    write_protein_specific_center
                    write_protein_specific_search_size
                    write_protein_specific_exhaustiveness
                elif prot name == "AchE_4m0e_B_apo":
                    write_protein_specific_center
                    write_protein_specific_search_size
                    write_protein_specific_exhaustiveness
        …etc
```

Figure 3.3: Pseudo-code of generation of vina files.

| Name | Crystal structures | AA$_1$/Atom Used | AA$_2$/Atom used | AA$_3$/Atom used | AA$_4$/Atom used | Cartesian Coordinates proposed | | | Size of the grid box | Type of docking |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | X | Y | Z | | |
| AChE | A | PHE295/N | TYR124/OH | N.A | N.A | -13.683 | -42.076 | 25.519 | 70/50/54 | Normal |
| | B | PHE295/N | TYR124/OH | N.A | N.A | 9.262 | -54.442 | -22.451 | 70/66/54 | Normal |
| ALR$_2$ | A | LYS252/N | TYR49/OH | N.A | N.A | 47.209 | 30.348 | 22.641 | 66/66/44 | Normal |
| BChE | A | ASP70/OD2 | SER198/OG | VAL288/CG2 | TRP82/CZ2 | -3.168 | 8.297 | -13.162 | 40/60/40 | Normal |
| | B | ASP70/OD2 | SER198/OG | VAL288/CG2 | TRP82/CZ2 | 17.839 | -27.038 | -41.616 | 40/72/40 | Normal |
| COX-1 | A | ARG120/NH1 | TYR385/OH | ALA527/CB | VAL349/CG1 | -21.795 | 50.203 | 9.357 | 40/70/54 | Normal |
| | B | ARG120/NH1 | TYR385/OH | ALA527/CB | VAL349/CG1 | -31.614 | 44.657 | -31.252 | 40/66/40 | Normal |
| COX-2 | A | ARG120/NH2 | TYR385/OH | TYR385/OH | VAL434/CG1 | 16.277 | 43.857 | 64.83 | 76/46/40 | Normal |
| | B | ARG120/NH2 | TYR385/OH | TYR385/OH | VAL434/CG1 | 17.561 | 49.511 | 15.668 | 40/56/78 | Normal |
| HSP90 | A | N.A | N.A | N.A | N.A | 2.269 | 10.687 | 24.493 | 48/44/46 | Normal |
| Integrase | A and B | B_HIS171/NE2 | A_ALA124/O | B_GLU170/OE2 | A_ALA124/O | 20.195 | 64.106 | 7.992 | 40/40/60 | Normal |
| MAOA | A | ARG51/NE | THR435/N | VAL244/O | TYR69/OH | 25.888 | 33.387 | -18.386 | 40/40/40 | Normal |
| MAOB | A | ARG42/NH1 | MET436/N | TYR60/OH | VAL235/N | 55.734 | 143.978 | 18.443 | 54/66/40 | Normal |
| | B | ARG42/NH1 | MET436/N | TYR60/OH | VAL235/N | 28.293 | 125.366 | 11.785 | 70/40/62 | Normal |
| Pyruvate kinase | AB | N.A | N.A | N.A | N.A | 0.579 | 33.8 | 23.22 | 126/126/126 | Blind |
| | C | N.A | N.A | N.A | N.A | 31.494 | 11.371 | 31.974 | 126/126/126 | Blind |
| sLAP$_2$ | A | TRP298/NE1 | PHE357/CD2 | SER273/OG | GLN352/NE2 | 25.211 | 4.677 | -0.923 | 56/56/56 | Normal |

Table 3.2: Summarize of the Amino acids used for the square method, Cartesian coordinates and size of the grid box and type of docking. AA: amino acid, N.A: Non-applicable.

| | Normal docking | Blind docking |
|---|---|---|
| Energy range | 4 | 4 |
| Exhaustiveness | 120 | 240 |
| CPU | 24 | 24 |

Table 3.3: Parameters used for normal and blind docking.

## 1.4. Results

The results were collated with script designed to extract only the first (lowest binding energy) pose from each docking experiment. In order to visualize the results, a heatmap was generated (using python and matlab scripts) which is presented in Figure 3.5. For this heatmap (shown in Figure 3.4) a gradient color code was used to identify the docked energy, with better binding (lower energy) represented in purple, and higher energy (weaker binding) of ligand-protein complexes represented in yellow. The ligands were ordered on the chart from the lowest to the highest binding energy, averaged across their binding to all targets. Each of the rows represents a single chalcone derivative from the virtual library, ordered thus, while the columns match each of the crystal structures of the proteins (11 proteins). Not all binding was negative. 29 chalcone derivatives presented a positive binding energy for the MAOA protein (i.e. energy input is required to force binding to occur). To exclude these unwanted cases from "washing out" the heat map, the script generating the heatmap set an upper limit of color to -4 kcal.mol$^{-1}$. -4 kcal.mol$^{-1}$ can hardly be considered good binding in this context, so there is little chance for confusion in the interpretation of the heatmap.
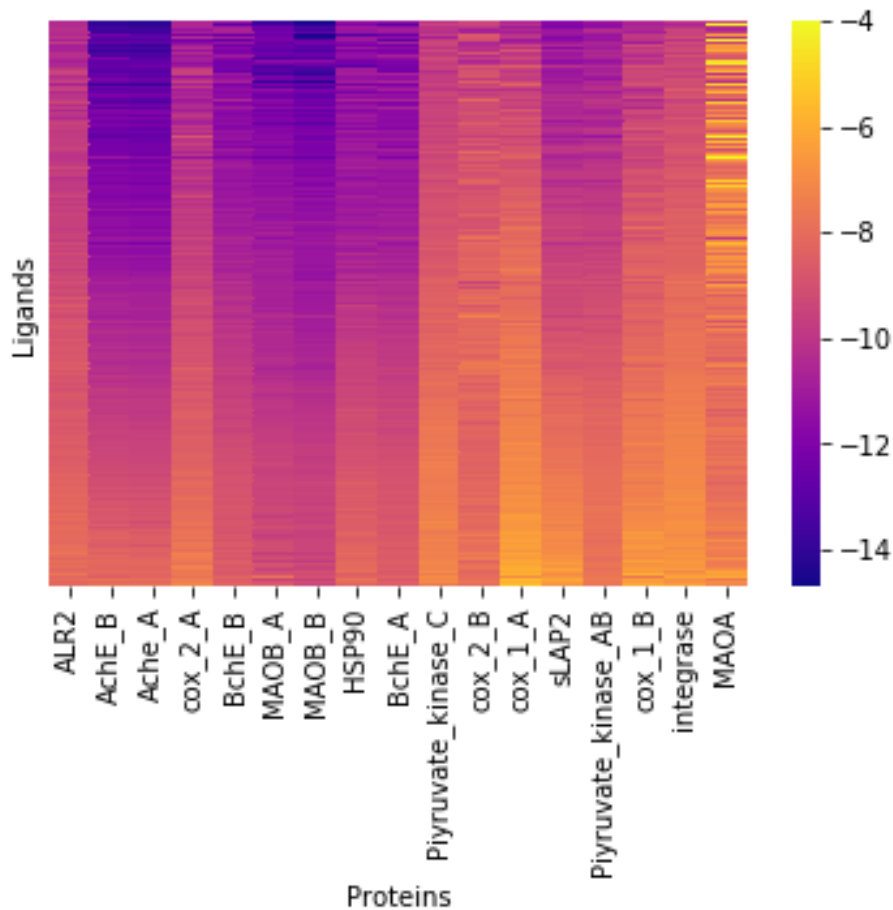
Figure 3.4: Heatmap of the high-throughput virtual screening.

```python
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.ticker as ticker
import matplotlib.pyplot as plt

Cpath='C:/Users\General/GoogleD work/Coding/Heatmap/'
AD = pd.read_csv((Cpath+'new_files_2018.csv'),delimiter=';',decimal=",")
labels=pd.read_csv((Cpath+'Label_ heatmap.csv'),delimiter=',',decimal=".")
Names=labels.iloc[:,0]
nms=Names.tolist()
a=labels.iloc[:,1]
a=np.arange(np.size(nms))

Energy=AD.iloc[:,0]
E=Energy.tolist()
Protein=AD.iloc[:,1]
P=Protein.tolist()
Mean=AD.iloc[:,2]
M=Mean.tolist()
data = pd.DataFrame({'X': E, 'Ligands': M, 'Proteins': P)
data_pivoted = data.pivot_table("X", "Ligands", "Proteins")
print(data_pivoted)
fig= sns.heatmap(data_pivoted, cmap="plasma",xticklabels = True, yticklabels= False)
fig.set_xticklabels(nms, minor=False,ha="center")
fig.spines['bottom'].set_visible(False)
plt.show()
```

Figure 3.5: Code of generation of the heatmap.

## 1.5. Validation of Docking

In order to assess the docking procedure followed for the high-throughput virtual screening, known chalcone inhibitors were docked to their respective targets using exactly the same parameters. Models of chalcone inhibitors of these targets (28 molecules in total), identified in the literature and presented in Chapter 1 (Figures 1.30, 1.32, 1.34, 1.36, 1.38, 1.40 and 1.43) [16-35] were constructed and submitted to geometric optimization following the same protocol as the chalcone library at a QM level using Gaussians09. The Figure 3.6 shows these molecules with their code names.
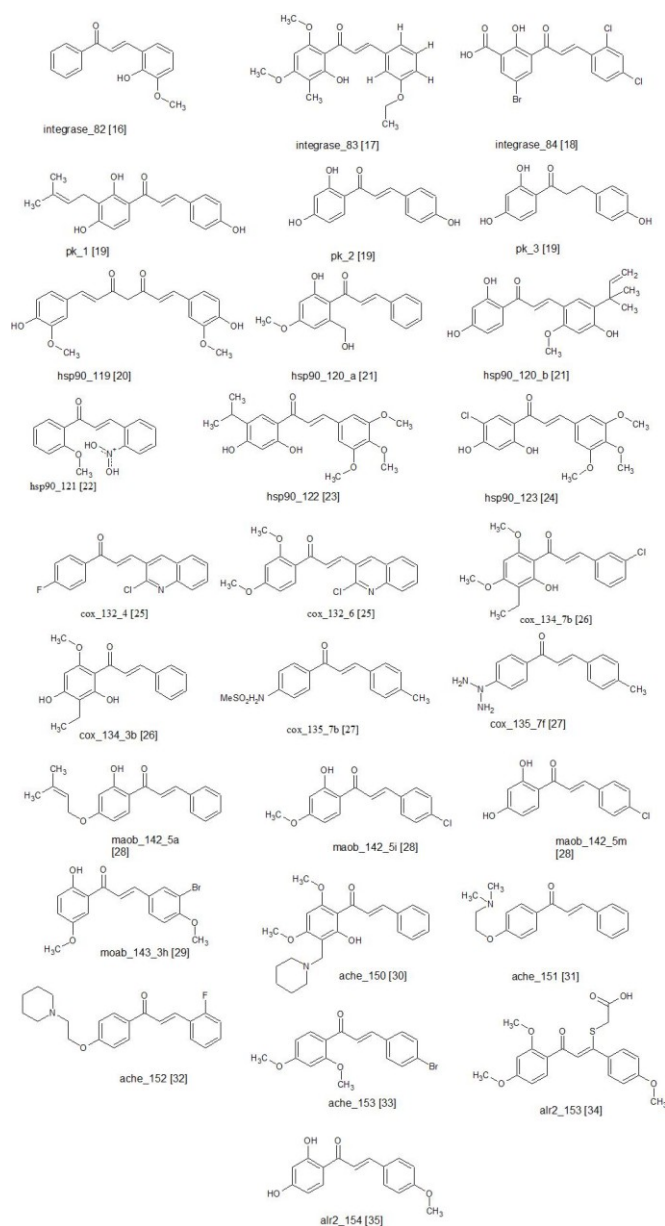


Figure 3.6: Chalcone inhibitors used for the validation docking. [16-35]

Subsequently, they were docked against the same targets following the same protocol. Several comparisons were possible: binding energy of known chalcone inhibitors vs. that of library inhibitors; position of docked known chalcone inhibitors relative to crystal structure ligands; and the position of known chalcone inhibitors vs. position of library inhibitors in the active site (particularly for blind docking evaluation). In terms of binding energy the comparison between chalcone library binding energy and known chalcone inhibitor binding energy was performed for structures AChE A and B substructures, $ALR_2$, COX-1 and COX-2 with A and B structures, HSP90, integrase, MAOB A and B structures and Pyruvate kinase AB and C co-crystal structures.

The binding energy of the library chalcones are lower than the best performing of the known inhibitor chalcones. It is interesting that library chalcone inhibitors improve on binding of known inhibitors by up to 1.6 kcal.mol$^{-1}$ for the COX-2_B and up to 3.4 kcal.mol$^{-1}$ for the $ALR_2$ protein. Also of interest is that in some cases inhibitors known for other targets have a better binding energy (affinity) than the inhibitors reported for the target itself. Figures 3.7 and 3.8 show the binding energy of the inhibitor ligands and the top five best (lowest energy) chalcones per protein.

The positioning of the chalcone library ligand docking relative to the active site of the original ligands from the literature and also relative to where known chalcone inhibitors bind to the diverse targets are presented in Figures 3.9 to 3.14. Except for the pyruvate kinase AB model, all positions of the ligands are located in the active site. Generally, the chalcones from the chalcone library are closely aligned, upon docking, to the original ligands. There are minor variations in terms of individual ligand orientation. The positioning of ligands for the MAOA structure shows some translation of the chalcone library molecules from the known inhibitor library. In the case of the pyruvate kinase, it is important to reiterate that the study was undertaken using blind docking. The pyruvate kinase AB co-crystal structure, chalcone library inhibitors docked to different locations within the protein structure to where known chalcone inhibitors were placed. However, for the pyruvate kinase C model, all orientations of the chalcones (known inhibitor and library chalcones) were found to occur at the same position.

In light of the above and considering that the original known inhibitor ligands were constructed *de novo* and optimized identically to the chalcone library ligands, positioning after docking matched crystal structure ligands, and considering that positioning of library chalcones matched that of known inhibitors we may make positive conclusions both regarding

the validity of the docking and the performance of the chalcone library in providing positive inhibitors to these targets. This success is across all targets, with the exception of the pyruvate kinase AB model.

**AchE_A**

| Name | Binding energy |
| --- | --- |
| acetophenone_15_benzaldehyde_78 | -14,2 |
| acetophenone_13_benzaldehyde_78 | -14,1 |
| acetophenone_13_benzaldehyde_27 | -13,9 |
| acetophenone_13_benzaldehyde_60 | -13,9 |
| acetophenone_14_benzaldehyde_78 | -13,9 |
| AchE_4m0e_A_apo_cox_132_4_opt1 | -10,8 |
| AchE_4m0e_A_apo_cox_132_6_opt1 | -10,6 |
| AchE_4m0e_A_apo_integrase_84_opt1 | -10 |
| AchE_4m0e_A_apo_ache_152_opt1 *(orange)* | -9,9 |
| AchE_4m0e_A_apo_pk_1_opt1 | -9,8 |
| AchE_4m0e_A_apo_cox_135_7f_opt1 | -9,3 |
| AchE_4m0e_A_apo_hsp90_120_b_opt1 | -9,3 |
| AchE_4m0e_A_apo_integrase_82_opt1 | -9,2 |
| AchE_4m0e_A_apo_hsp90_119_opt1 | -9,2 |
| AchE_4m0e_A_apo_maob_142_5a_opt1 | -9 |
| AchE_4m0e_A_apo_maob_143_3h_opt1 | -9 |
| AchE_4m0e_A_apo_cox_135_7b_opt1 | -9 |
| AchE_4m0e_A_apo_ache_153_opt1 *(orange)* | -9 |
| AchE_4m0e_A_apo_hsp90_122_opt1 | -8,9 |
| AchE_4m0e_A_apo_ache_150_opt1 *(orange)* | -8,8 |
| AchE_4m0e_A_apo_maob_142_5i_opt1 | -8,8 |
| AchE_4m0e_A_apo_pk_2_opt1 | -8,7 |
| AchE_4m0e_A_apo_maob_142_5m_opt1 | -8,6 |
| AchE_4m0e_A_apo_arl2_164_opt1 | -8,6 |
| AchE_4m0e_A_apo_integrase_83_opt1 | -8,5 |
| AchE_4m0e_A_apo_pk_3_opt1 | -8,5 |
| AchE_4m0e_A_apo_ache_151_opt1 *(orange)* | -8,5 |
| AchE_4m0e_A_apo_hsp90_123_opt1 | -8,4 |
| AchE_4m0e_A_apo_hsp90_121_opt1 | -8,3 |
| AchE_4m0e_A_apo_hsp90_120_a_opt1 | -8,1 |
| AchE_4m0e_A_apo_cox_134_3b_opt1 | -8 |
| AchE_4m0e_A_apo_cox_134_7b_opt1 | -7,8 |
| AchE_4m0e_A_apo_arl2_163_opt1 | -7,5 |

**AchE_B**

| Name | Binding energy |
| --- | --- |
| acetophenone_13_benzaldehyde_78 | -14,1 |
| acetophenone_15_benzaldehyde_78 | -14,1 |
| acetophenone_13_benzaldehyde_17 | -13,8 |
| acetophenone_13_benzaldehyde_27 | -13,8 |
| acetophenone_13_benzaldehyde_45 | -13,8 |
| AchE_4m0e_B_apo_cox_132_4_opt1 | -10,6 |
| AchE_4m0e_B_apo_cox_132_6_opt1 | -10,4 |
| AchE_4m0e_B_apo_integrase_84_opt1 | -9,8 |
| AchE_4m0e_B_apo_ache_152_opt1 *(orange)* | -9,6 |
| AchE_4m0e_B_apo_hsp90_119_opt1 | -9,3 |
| AchE_4m0e_B_apo_hsp90_120_b_opt1 | -9,2 |
| AchE_4m0e_B_apo_cox_135_7f_opt1 | -9,2 |
| AchE_4m0e_B_apo_pk_1_opt1 | -9,2 |
| AchE_4m0e_B_apo_cox_135_7b_opt1 | -9,1 |
| AchE_4m0e_B_apo_integrase_82_opt1 | -9,1 |
| AchE_4m0e_B_apo_ache_153_opt1 *(orange)* | -9,1 |
| AchE_4m0e_B_apo_maob_143_3h_opt1 | -9 |
| AchE_4m0e_B_apo_maob_142_5a_opt1 | -8,9 |
| AchE_4m0e_B_apo_ache_150_opt1 *(orange)* | -8,9 |
| AchE_4m0e_B_apo_hsp90_122_opt1 | -8,8 |
| AchE_4m0e_B_apo_maob_142_5i_opt1 | -8,7 |
| AchE_4m0e_B_apo_pk_3_opt1 | -8,6 |
| AchE_4m0e_B_apo_pk_2_opt1 | -8,6 |
| AchE_4m0e_B_apo_arl2_164_opt1 | -8,6 |
| AchE_4m0e_B_apo_maob_142_5m_opt1 | -8,6 |
| AchE_4m0e_B_apo_integrase_83_opt1 | -8,4 |
| AchE_4m0e_B_apo_ache_151_opt1 *(orange)* | -8,4 |
| AchE_4m0e_B_apo_hsp90_123_opt1 | -8,3 |
| AchE_4m0e_B_apo_cox_134_3b_opt1 | -8,2 |
| AchE_4m0e_B_apo_hsp90_120_a_opt1 | -8,1 |
| AchE_4m0e_B_apo_hsp90_121_opt1 | -7,9 |
| AchE_4m0e_B_apo_cox_134_7b_opt1 | -7,9 |
| AchE_4m0e_B_apo_arl2_163_opt1 | -7,4 |

**ALR2**

| Name | Binding energy |
| --- | --- |
| acetophenone_19_benzaldehyde_2 | -11,3 |
| acetophenone_19_benzaldehyde_17 | -11,1 |
| acetophenone_19_benzaldehyde_45 | -11,1 |
| acetophenone_21_benzaldehyde_89 | -11,1 |
| acetophenone_21_benzaldehyde_88 | -11 |
| ALR2_5jh1_apo_integrase_84_opt1 | -9,2 |
| ALR2_5jh1_apo_cox_132_4_opt1 | -8,9 |
| ALR2_5jh1_apo_cox_132_6_opt1 | -8,7 |
| ALR2_5jh1_apo_cox_135_7f_opt1 | -8,4 |
| ALR2_5jh1_apo_integrase_82_opt1 | -8,3 |
| ALR2_5jh1_apo_hsp90_120_a_opt1 | -8,2 |
| ALR2_5jh1_apo_pk_1_opt1 | -8,1 |
| ALR2_5jh1_apo_maob_142_5i_opt1 | -8 |
| ALR2_5jh1_apo_maob_142_5a_opt1 | -8 |
| ALR2_5jh1_apo_pk_2_opt1 | -7,9 |
| ALR2_5jh1_apo_maob_142_5m_opt1 | -7,9 |
| ALR2_5jh1_apo_ache_152_opt1 | -7,9 |
| ALR2_5jh1_apo_cox_134_7b_opt1 | -7,8 |
| ALR2_5jh1_apo_cox_134_3b_opt1 | -7,7 |
| ALR2_5jh1_apo_integrase_83_opt1 | -7,7 |
| ALR2_5jh1_apo_cox_135_7b_opt1 | -7,7 |
| ALR2_5jh1_apo_maob_143_3h_opt1 | -7,7 |
| ALR2_5jh1_apo_hsp90_121_opt1 | -7,7 |
| ALR2_5jh1_apo_ache_150_opt1 | -7,7 |
| ALR2_5jh1_apo_pk_3_opt1 | -7,6 |
| ALR2_5jh1_apo_arl2_164_opt1 *(orange)* | -7,6 |
| ALR2_5jh1_apo_hsp90_120_b_opt1 | -7,6 |
| ALR2_5jh1_apo_ache_151_opt1 | -7,5 |
| ALR2_5jh1_apo_ache_153_opt1 | -7,5 |
| ALR2_5jh1_apo_arl2_163_opt1 *(orange)* | -7,2 |
| ALR2_5jh1_apo_hsp90_119_opt1 | -7,1 |
| ALR2_5jh1_apo_hsp90_122_opt1 | -6,9 |
| ALR2_5jh1_apo_hsp90_123_opt1 | -6,7 |

**COX-1_A**

| Name | Binding energy |
| --- | --- |
| acetophenone_19_benzaldehyde_89 | -11,1 |
| acetophenone_21_benzaldehyde_89 | -11,1 |
| acetophenone_14_benzaldehyde_89 | -10,9 |
| acetophenone_42_benzaldehyde_89 | -10,9 |
| acetophenone_10_benzaldehyde_89 | -10,8 |
| cox_1_3n8z_A_apo_cox_132_4_opt1 *(orange)* | -8,1 |
| cox_1_3n8z_A_apo_hsp90_119_opt1 | -7,7 |
| cox_1_3n8z_A_apo_ache_152_opt1 | -7,6 |
| cox_1_3n8z_A_apo_integrase_82_opt1 | -7,5 |
| cox_1_3n8z_A_apo_maob_142_5m_opt1 | -7,3 |
| cox_1_3n8z_A_apo_maob_142_5a_opt1 | -7,2 |
| cox_1_3n8z_A_apo_cox_135_7f_opt1 *(orange)* | -7,1 |
| cox_1_3n8z_A_apo_pk_1_opt1 | -7 |
| cox_1_3n8z_A_apo_pk_2_opt1 | -7 |
| cox_1_3n8z_A_apo_arl2_164_opt1 | -7 |
| cox_1_3n8z_A_apo_ache_151_opt1 | -6,9 |
| cox_1_3n8z_A_apo_pk_3_opt1 | -6,9 |
| cox_1_3n8z_A_apo_cox_135_7b_opt1 *(orange)* | -6,9 |
| cox_1_3n8z_A_apo_integrase_83_opt1 | -6,9 |
| cox_1_3n8z_A_apo_ache_150_opt1 | -6,9 |
| cox_1_3n8z_A_apo_cox_132_6_opt1 *(orange)* | -6,8 |
| cox_1_3n8z_A_apo_integrase_84_opt1 | -6,7 |
| cox_1_3n8z_A_apo_hsp90_120_b_opt1 | -6,5 |
| cox_1_3n8z_A_apo_maob_142_5i_opt1 | -6,4 |
| cox_1_3n8z_A_apo_cox_134_3b_opt1 *(orange)* | -6,4 |
| cox_1_3n8z_A_apo_maob_143_3h_opt1 | -6,4 |
| cox_1_3n8z_A_apo_cox_134_7b_opt1 *(orange)* | -6,4 |
| cox_1_3n8z_A_apo_hsp90_120_a_opt1 | -6,4 |
| cox_1_3n8z_A_apo_ache_153_opt1 | -6,2 |
| cox_1_3n8z_A_apo_hsp90_122_opt1 | -6,1 |
| cox_1_3n8z_A_apo_hsp90_121_opt1 | -6 |
| cox_1_3n8z_A_apo_hsp90_123_opt1 | -5,9 |
| cox_1_3n8z_A_apo_arl2_163_opt1 | -5,7 |

**COX-1_B**

| Name | Binding Energy |
| --- | --- |
| acetophenone_14_benzaldehyde_2 | -10,9 |
| acetophenone_54_benzaldehyde_89 | -10,9 |
| acetophenone_15_benzaldehyde_2 | -10,8 |
| acetophenone_15_benzaldehyde_32 | -10,8 |
| acetophenone_54_benzaldehyde_2 | -10,8 |
| cox_1_3n8z_B_apo_ache_152_opt1 | -8,5 |
| cox_1_3n8z_B_apo_hsp90_119_opt1 | -8,5 |
| cox_1_3n8z_B_apo_maob_142_5a_opt1 | -8,4 |
| cox_1_3n8z_B_apo_cox_132_4_opt1 *(orange)* | -8,3 |
| cox_1_3n8z_B_apo_integrase_82_opt1 | -8 |
| cox_1_3n8z_B_apo_ache_151_opt1 | -7,8 |
| cox_1_3n8z_B_apo_maob_143_3h_opt1 | -7,7 |
| cox_1_3n8z_B_apo_pk_1_opt1 | -7,7 |
| cox_1_3n8z_B_apo_integrase_84_opt1 | -7,6 |
| cox_1_3n8z_B_apo_maob_142_5m_opt1 | -7,4 |
| cox_1_3n8z_B_apo_cox_135_7b_opt1 *(orange)* | -7,3 |
| cox_1_3n8z_B_apo_hsp90_121_opt1 | -7,3 |
| cox_1_3n8z_B_apo_arl2_164_opt1 | -7,2 |
| cox_1_3n8z_B_apo_pk_2_opt1 | -7,1 |
| cox_1_3n8z_B_apo_pk_3_opt1 | -7,1 |
| cox_1_3n8z_B_apo_cox_134_7b_opt1 *(orange)* | -7 |
| cox_1_3n8z_B_apo_maob_142_5i_opt1 | -7 |
| cox_1_3n8z_B_apo_cox_135_7f_opt1 *(orange)* | -7 |
| cox_1_3n8z_B_apo_hsp90_120_a_opt1 | -6,9 |
| cox_1_3n8z_B_apo_cox_134_3b_opt1 *(orange)* | -6,8 |
| cox_1_3n8z_B_apo_integrase_83_opt1 | -6,7 |
| cox_1_3n8z_B_apo_arl2_163_opt1 | -6,6 |
| cox_1_3n8z_B_apo_hsp90_122_opt1 | -6,6 |
| cox_1_3n8z_B_apo_cox_132_6_opt1 *(orange)* | -6,5 |
| cox_1_3n8z_B_apo_ache_150_opt1 | -6,4 |
| cox_1_3n8z_B_apo_hsp90_120_b_opt1 | -6,4 |
| cox_1_3n8z_B_apo_hsp90_123_opt1 | -6,1 |
| cox_1_3n8z_B_apo_ache_153_opt1 | -6 |

**COX-2_A**

| Name | Binding Energy |
| --- | --- |
| acetophenone_19_benzaldehyde_89 | -12,2 |
| acetophenone_22_benzaldehyde_78 | -12,1 |
| acetophenone_19_benzaldehyde_78 | -12 |
| acetophenone_22_benzaldehyde_89 | -12 |
| acetophenone_14_benzaldehyde_2 | -11,7 |
| cox_2_5f19_A_apo_cox_132_4_opt1 *(orange)* | -9,1 |
| cox_2_5f19_A_apo_ache_152_opt1 | -8,7 |
| cox_2_5f19_A_apo_integrase_82_opt1 | -8,5 |
| cox_2_5f19_A_apo_integrase_84_opt1 | -8,4 |
| cox_2_5f19_A_apo_cox_135_7f_opt1 *(orange)* | -8,3 |
| cox_2_5f19_A_apo_maob_142_5i_opt1 | -8,2 |
| cox_2_5f19_A_apo_pk_1_opt1 | -8,1 |
| cox_2_5f19_A_apo_maob_142_5a_opt1 | -8,1 |
| cox_2_5f19_A_apo_arl2_164_opt1 | -8,1 |
| cox_2_5f19_A_apo_pk_3_opt1 | -8,1 |
| cox_2_5f19_A_apo_maob_142_5m_opt1 | -8,1 |
| cox_2_5f19_A_apo_pk_2_opt1 | -8 |
| cox_2_5f19_A_apo_hsp90_119_opt1 | -8 |
| cox_2_5f19_A_apo_hsp90_121_opt1 | -8 |
| cox_2_5f19_A_apo_maob_143_3h_opt1 | -7,9 |
| cox_2_5f19_A_apo_cox_132_6_opt1 *(orange)* | -7,9 |
| cox_2_5f19_A_apo_hsp90_123_opt1 | -7,7 |
| cox_2_5f19_A_apo_ache_151_opt1 | -7,7 |
| cox_2_5f19_A_apo_hsp90_120_a_opt1 | -7,7 |
| cox_2_5f19_A_apo_integrase_83_opt1 | -7,7 |
| cox_2_5f19_A_apo_cox_135_7b_opt1 *(orange)* | -7,6 |
| cox_2_5f19_A_apo_ache_153_opt1 | -7,5 |
| cox_2_5f19_A_apo_cox_134_3b_opt1 *(orange)* | -7,5 |
| cox_2_5f19_A_apo_hsp90_122_opt1 | -7,4 |
| cox_2_5f19_A_apo_hsp90_120_b_opt1 | -7,2 |
| cox_2_5f19_A_apo_cox_134_7b_opt1 *(orange)* | -7 |
| cox_2_5f19_A_apo_arl2_163_opt1 | -6,7 |
| cox_2_5f19_A_apo_ache_150_opt1 | -6,6 |

Figure 3.7: Binding energy of the drugs and the five top lowest binding energy of the chalcone library against AChE_A, AChE_B, ALR$_2$, COX-1_A, COX-1_B, COX-2_A. Orange entry is corresponding to the original drug for protein.

| Name | Binding energy |
|---|---|
| acetophenone_15_benzaldehyde_45 | -10,8 |
| acetophenone_15_benzaldehyde_17 | -10,7 |
| acetophenone_15_benzaldehyde_2 | -10,7 |
| acetophenone_14_benzaldehyde_2 | -10,6 |
| acetophenone_21_benzaldehyde_2 | -10,6 |
| cox_2_5f19_B_apo_cox_132_4_opt1 *(orange)* | -9 |
| cox_2_5f19_B_apo_ache_152_opt1 | -8,6 |
| cox_2_5f19_B_apo_cox_135_7f_opt1 *(orange)* | -8,5 |
| cox_2_5f19_B_apo_integrase_82_opt1 | -8,5 |
| cox_2_5f19_B_apo_pk_3_opt1 | -8,3 |
| cox_2_5f19_B_apo_maob_142_5a_opt1 | -8,2 |
| cox_2_5f19_B_apo_hsp90_121_opt1 | -8,1 |
| cox_2_5f19_B_apo_pk_2_opt1 | -8 |
| cox_2_5f19_B_apo_hsp90_120_b_opt1 | -8 |
| cox_2_5f19_B_apo_pk_1_opt1 | -7,8 |
| cox_2_5f19_B_apo_arl2_164_opt1 | -7,7 |
| cox_2_5f19_B_apo_hsp90_119_opt1 | -7,6 |
| cox_2_5f19_B_apo_integrase_84_opt1 | -7,6 |
| cox_2_5f19_B_apo_maob_142_5m_opt1 | -7,6 |
| cox_2_5f19_B_apo_ache_151_opt1 | -7,5 |
| cox_2_5f19_B_apo_cox_134_3b_opt1 *(orange)* | -7,3 |
| cox_2_5f19_B_apo_maob_142_5i_opt1 | -7,3 |
| cox_2_5f19_B_apo_maob_143_3h_opt1 | -7,1 |
| cox_2_5f19_B_apo_hsp90_120_a_opt1 | -6,8 |
| cox_2_5f19_B_apo_cox_135_7b_opt1 *(orange)* | -6,8 |
| cox_2_5f19_B_apo_hsp90_122_opt1 | -6,8 |
| cox_2_5f19_B_apo_cox_132_6_opt1 *(orange)* | -6,7 |
| cox_2_5f19_B_apo_integrase_83_opt1 | -6,6 |
| cox_2_5f19_B_apo_ache_153_opt1 | -6,4 |
| cox_2_5f19_B_apo_hsp90_123_opt1 | -6,3 |
| cox_2_5f19_B_apo_arl2_163_opt1 | -6,1 |
| cox_2_5f19_B_apo_ache_150_opt1 | -5,9 |
| cox_2_5f19_B_apo_cox_134_7b_opt1 *(orange)* | -5,6 |

COX-2_B

| Name | Binding energy |
|---|---|
| acetophenone_22_benzaldehyde_78 | -13,5 |
| acetophenone_22_benzaldehyde_58 | -13,4 |
| acetophenone_22_benzaldehyde_44 | -13,2 |
| acetophenone_22_benzaldehyde_72 | -13,2 |
| acetophenone_22_benzaldehyde_89 | -13,2 |
| HSP90_5fnc_apo_hsp90_120_b_opt1 *(orange)* | -9,8 |
| HSP90_5fnc_apo_ache_152_opt1 | -9,4 |
| HSP90_5fnc_apo_cox_132_4_opt1 | -9,4 |
| HSP90_5fnc_apo_maob_142_5a_opt1 | -9,2 |
| HSP90_5fnc_apo_cox_132_6_opt1 | -9,1 |
| HSP90_5fnc_apo_pk_1_opt1 | -9 |
| HSP90_5fnc_apo_integrase_84_opt1 | -9 |
| HSP90_5fnc_apo_maob_143_3h_opt1 | -8,9 |
| HSP90_5fnc_apo_maob_142_5m_opt1 | -8,8 |
| HSP90_5fnc_apo_cox_135_7b_opt1 | -8,8 |
| HSP90_5fnc_apo_arl2_164_opt1 | -8,8 |
| HSP90_5fnc_apo_hsp90_121_opt1 *(orange)* | -8,8 |
| HSP90_5fnc_apo_pk_2_opt1 | -8,7 |
| HSP90_5fnc_apo_integrase_83_opt1 | -8,7 |
| HSP90_5fnc_apo_cox_135_7f_opt1 | -8,6 |
| HSP90_5fnc_apo_maob_142_5i_opt1 | -8,4 |
| HSP90_5fnc_apo_hsp90_119_opt1 *(orange)* | -8,4 |
| HSP90_5fnc_apo_ache_151_opt1 | -8,4 |
| HSP90_5fnc_apo_hsp90_122_opt1 *(orange)* | -8,3 |
| HSP90_5fnc_apo_integrase_82_opt1 | -8,3 |
| HSP90_5fnc_apo_cox_134_3b_opt1 | -8,3 |
| HSP90_5fnc_apo_ache_150_opt1 | -8,3 |
| HSP90_5fnc_apo_ache_153_opt1 | -8,2 |
| HSP90_5fnc_apo_hsp90_120_a_opt1 *(orange)* | -8,2 |
| HSP90_5fnc_apo_arl2_163_opt1 | -8,1 |
| HSP90_5fnc_apo_hsp90_123_opt1 *(orange)* | -8,1 |
| HSP90_5fnc_apo_cox_134_7b_opt1 | -8,1 |

HSP90

| Name | Binding energy |
|---|---|
| acetophenone_13_benzaldehyde_78 | -10,6 |
| acetophenone_13_benzaldehyde_89 | -10,6 |
| acetophenone_19_benzaldehyde_78 | -10,5 |
| acetophenone_21_benzaldehyde_78 | -10,3 |
| acetophenone_22_benzaldehyde_78 | -10,3 |
| integrase_4nyf_apo_cox_132_4_opt1 | -8,1 |
| integrase_4nyf_apo_cox_132_6_opt1 | -7,6 |
| integrase_4nyf_apo_integrase_84_opt1 *(orange)* | -7,4 |
| integrase_4nyf_apo_cox_135_7b_opt1 | -7,3 |
| integrase_4nyf_apo_ache_152_opt1 | -7,2 |
| integrase_4nyf_apo_hsp90_120_b_opt1 | -7,1 |
| integrase_4nyf_apo_pk_1_opt1 | -7 |
| integrase_4nyf_apo_maob_142_5a_opt1 | -7 |
| integrase_4nyf_apo_cox_135_7f_opt1 | -6,9 |
| integrase_4nyf_apo_hsp90_120_a_opt1 | -6,8 |
| integrase_4nyf_apo_maob_142_5m_opt1 | -6,8 |
| integrase_4nyf_apo_integrase_82_opt1 *(orange)* | -6,8 |
| integrase_4nyf_apo_hsp90_121_opt1 | -6,8 |
| integrase_4nyf_apo_ache_153_opt1 | -6,7 |
| integrase_4nyf_apo_cox_134_3b_opt1 | -6,7 |
| integrase_4nyf_apo_pk_3_opt1 | -6,7 |
| integrase_4nyf_apo_maob_142_5i_opt1 | -6,7 |
| integrase_4nyf_apo_arl2_164_opt1 | -6,7 |
| integrase_4nyf_apo_arl2_163_opt1 | -6,7 |
| integrase_4nyf_apo_maob_143_3h_opt1 | -6,6 |
| integrase_4nyf_apo_pk_2_opt1 | -6,6 |
| integrase_4nyf_apo_integrase_83_opt1 *(orange)* | -6,6 |
| integrase_4nyf_apo_ache_151_opt1 | -6,5 |
| integrase_4nyf_apo_ache_150_opt1 | -6,5 |
| integrase_4nyf_apo_hsp90_119_opt1 | -6,5 |
| integrase_4nyf_apo_cox_134_7b_opt1 | -6,5 |
| integrase_4nyf_apo_hsp90_122_opt1 | -6,3 |
| integrase_4nyf_apo_hsp90_123_opt1 | -6,1 |

integrase

| Name | Binding energy |
|---|---|
| acetophenone_15_benzaldehyde_78 | -13,9 |
| acetophenone_15_benzaldehyde_26 | -13,8 |
| acetophenone_15_benzaldehyde_37 | -13,8 |
| acetophenone_13_benzaldehyde_78 | -13,6 |
| acetophenone_22_benzaldehyde_64 | -13,6 |
| MAOB_5mrl_A_apo_ache_152_opt1 | -10,8 |
| MAOB_5mrl_A_apo_maob_142_5a_opt1 *(orange)* | -10,4 |
| MAOB_5mrl_A_apo_cox_135_7b_opt1 | -10,2 |
| MAOB_5mrl_A_apo_cox_132_4_opt1 | -10,1 |
| MAOB_5mrl_A_apo_pk_1_opt1 | -9,8 |
| MAOB_5mrl_A_apo_cox_135_7f_opt1 | -9,7 |
| MAOB_5mrl_A_apo_maob_142_5i_opt1 *(orange)* | -9,7 |
| MAOB_5mrl_A_apo_maob_142_5m_opt1 *(orange)* | -9,6 |
| MAOB_5mrl_A_apo_hsp90_120_a_opt1 | -9,5 |
| MAOB_5mrl_A_apo_cox_132_6_opt1 | -9,5 |
| MAOB_5mrl_A_apo_hsp90_119_opt1 | -9,4 |
| MAOB_5mrl_A_apo_integrase_82_opt1 | -9,4 |
| MAOB_5mrl_A_apo_pk_2_opt1 | -9,4 |
| MAOB_5mrl_A_apo_ache_150_opt1 | -9,4 |
| MAOB_5mrl_A_apo_ache_151_opt1 | -9,3 |
| MAOB_5mrl_A_apo_integrase_84_opt1 | -9,2 |
| MAOB_5mrl_A_apo_arl2_164_opt1 | -8,9 |
| MAOB_5mrl_A_apo_cox_134_3b_opt1 | -8,9 |
| MAOB_5mrl_A_apo_hsp90_122_opt1 | -8,7 |
| MAOB_5mrl_A_apo_pk_3_opt1 | -8,7 |
| MAOB_5mrl_A_apo_ache_153_opt1 | -8,6 |
| MAOB_5mrl_A_apo_hsp90_121_opt1 | -8,6 |
| MAOB_5mrl_A_apo_maob_143_3h_opt1 *(orange)* | -8,5 |
| MAOB_5mrl_A_apo_hsp90_123_opt1 | -8,5 |
| MAOB_5mrl_A_apo_hsp90_120_b_opt1 | -8,5 |
| MAOB_5mrl_A_apo_integrase_83_opt1 | -8,4 |
| MAOB_5mrl_A_apo_cox_134_7b_opt1 | -8,3 |
| MAOB_5mrl_A_apo_arl2_163_opt1 | -8,1 |

MAOB_A

| Name | Binding energy |
|---|---|
| acetophenone_15_benzaldehyde_78 | -14,7 |
| acetophenone_22_benzaldehyde_89 | -14,5 |
| acetophenone_15_benzaldehyde_44 | -14,4 |
| acetophenone_13_benzaldehyde_78 | -14,2 |
| acetophenone_15_benzaldehyde_26 | -14,2 |
| MAOB_5mrl_B_apo_ache_152_opt1 | -10,8 |
| MAOB_5mrl_B_apo_maob_142_5a_opt1 *(orange)* | -10,4 |
| MAOB_5mrl_B_apo_cox_132_6_opt1 | -10,3 |
| MAOB_5mrl_B_apo_cox_132_4_opt1 | -10,3 |
| MAOB_5mrl_B_apo_cox_135_7b_opt1 | -10,1 |
| MAOB_5mrl_B_apo_integrase_84_opt1 | -10,1 |
| MAOB_5mrl_B_apo_pk_1_opt1 | -9,8 |
| MAOB_5mrl_B_apo_hsp90_119_opt1 | -9,7 |
| MAOB_5mrl_B_apo_cox_135_7f_opt1 | -9,3 |
| MAOB_5mrl_B_apo_hsp90_120_a_opt1 | -9,3 |
| MAOB_5mrl_B_apo_integrase_82_opt1 | -9,3 |
| MAOB_5mrl_B_apo_ache_153_opt1 | -9,2 |
| MAOB_5mrl_B_apo_ache_151_opt1 | -9,2 |
| MAOB_5mrl_B_apo_hsp90_121_opt1 | -9,2 |
| MAOB_5mrl_B_apo_arl2_164_opt1 | -9,1 |
| MAOB_5mrl_B_apo_maob_143_3h_opt1 *(orange)* | -9,1 |
| MAOB_5mrl_B_apo_maob_142_5m_opt1 *(orange)* | -9 |
| MAOB_5mrl_B_apo_maob_142_5i_opt1 *(orange)* | -9 |
| MAOB_5mrl_B_apo_cox_134_3b_opt1 | -9 |
| MAOB_5mrl_B_apo_pk_3_opt1 | -8,8 |
| MAOB_5mrl_B_apo_hsp90_122_opt1 | -8,8 |
| MAOB_5mrl_B_apo_pk_2_opt1 | -8,8 |
| MAOB_5mrl_B_apo_ache_150_opt1 | -8,8 |
| MAOB_5mrl_B_apo_integrase_83_opt1 | -8,7 |
| MAOB_5mrl_B_apo_arl2_163_opt1 | -8,7 |
| MAOB_5mrl_B_apo_cox_134_7b_opt1 | -8,7 |
| MAOB_5mrl_B_apo_hsp90_123_opt1 | -8,5 |
| MAOB_5mrl_B_apo_hsp90_120_b_opt1 | -8,4 |

MAOB_B

| Name | Binding energy |
|---|---|
| acetophenone_21_benzaldehyde_49 | -12 |
| acetophenone_21_benzaldehyde_87 | -11,8 |
| acetophenone_21_benzaldehyde_78 | -11,7 |
| acetophenone_21_benzaldehyde_89 | -11,6 |
| acetophenone_13_benzaldehyde_78 | -11,5 |
| piyruvte_kinase_3t05_AB_apo_cox_132_4_opt1 | -8,5 |
| piyruvte_kinase_3t05_AB_apo_cox_132_6_opt1 | -8,4 |
| piyruvte_kinase_3t05_AB_apo_pk_1_opt1 *(orange)* | -8,2 |
| piyruvte_kinase_3t05_AB_apo_ache_150_opt1 | -8,2 |
| piyruvte_kinase_3t05_AB_apo_integrase_84_opt1 | -8,1 |
| piyruvte_kinase_3t05_AB_apo_hsp90_119_opt1 | -8 |
| piyruvte_kinase_3t05_AB_apo_maob_142_5a_opt1 | -7,9 |
| piyruvte_kinase_3t05_AB_apo_maob_142_5m_opt1 | -7,9 |
| piyruvte_kinase_3t05_AB_apo_cox_135_7b_opt1 | -7,9 |
| piyruvte_kinase_3t05_AB_apo_maob_142_5i_opt1 | -7,9 |
| piyruvte_kinase_3t05_AB_apo_cox_135_7f_opt1 | -7,9 |
| piyruvte_kinase_3t05_AB_apo_ache_152_opt1 | -7,8 |
| piyruvte_kinase_3t05_AB_apo_pk_2_opt1 *(orange)* | -7,8 |
| piyruvte_kinase_3t05_AB_apo_hsp90_120_a_opt1 | -7,7 |
| piyruvte_kinase_3t05_AB_apo_pk_3_opt1 *(orange)* | -7,7 |
| piyruvte_kinase_3t05_AB_apo_cox_134_3b_opt1 | -7,7 |
| piyruvte_kinase_3t05_AB_apo_hsp90_120_b_opt1 | -7,7 |
| piyruvte_kinase_3t05_AB_apo_integrase_82_opt1 | -7,6 |
| piyruvte_kinase_3t05_AB_apo_hsp90_122_opt1 | -7,6 |
| piyruvte_kinase_3t05_AB_apo_integrase_83_opt1 | -7,5 |
| piyruvte_kinase_3t05_AB_apo_maob_143_3h_opt1 | -7,5 |
| piyruvte_kinase_3t05_AB_apo_hsp90_121_opt1 | -7,5 |
| piyruvte_kinase_3t05_AB_apo_cox_134_7b_opt1 | -7,5 |
| piyruvte_kinase_3t05_AB_apo_arl2_163_opt1 | -7,4 |
| piyruvte_kinase_3t05_AB_apo_arl2_164_opt1 | -7,3 |
| piyruvte_kinase_3t05_AB_apo_hsp90_123_opt1 | -7,3 |
| piyruvte_kinase_3t05_AB_apo_ache_153_opt1 | -7,3 |
| piyruvte_kinase_3t05_AB_apo_ache_151_opt1 | -7,3 |

PK_AB

| Name | Binding energy |
|---|---|
| acetophenone_15_benzaldehyde_89 | -11,2 |
| acetophenone_15_benzaldehyde_44 | -10,8 |
| acetophenone_15_benzaldehyde_64 | -10,6 |
| acetophenone_42_benzaldehyde_89 | -10,6 |
| acetophenone_10_benzaldehyde_89 | -10,5 |
| piyruvte_kinase_3t05_C_apo_cox_132_4_opt1 | -8,5 |
| piyruvte_kinase_3t05_C_apo_cox_132_6_opt1 | -8,4 |
| piyruvte_kinase_3t05_C_apo_integrase_84_opt1 | -8,3 |
| piyruvte_kinase_3t05_C_apo_pk_1_opt1 *(orange)* | -8,1 |
| piyruvte_kinase_3t05_C_apo_ache_150_opt1 | -8 |
| piyruvte_kinase_3t05_C_apo_pk_2_opt1 *(orange)* | -7,9 |
| piyruvte_kinase_3t05_C_apo_cox_135_7b_opt1 | -7,9 |
| piyruvte_kinase_3t05_C_apo_cox_134_3b_opt1 | -7,9 |
| piyruvte_kinase_3t05_C_apo_maob_142_5m_opt1 | -7,9 |
| piyruvte_kinase_3t05_C_apo_maob_142_5a_opt1 | -7,8 |
| piyruvte_kinase_3t05_C_apo_hsp90_120_a_opt1 | -7,8 |
| piyruvte_kinase_3t05_C_apo_maob_142_5i_opt1 | -7,7 |
| piyruvte_kinase_3t05_C_apo_cox_135_7f_opt1 | -7,6 |
| piyruvte_kinase_3t05_C_apo_ache_152_opt1 | -7,6 |
| piyruvte_kinase_3t05_C_apo_integrase_82_opt1 | -7,6 |
| piyruvte_kinase_3t05_C_apo_pk_3_opt1 *(orange)* | -7,5 |
| piyruvte_kinase_3t05_C_apo_hsp90_121_opt1 | -7,5 |
| piyruvte_kinase_3t05_C_apo_hsp90_120_b_opt1 | -7,4 |
| piyruvte_kinase_3t05_C_apo_arl2_163_opt1 | -7,2 |
| piyruvte_kinase_3t05_C_apo_hsp90_122_opt1 | -7,2 |
| piyruvte_kinase_3t05_C_apo_arl2_164_opt1 | -7,1 |
| piyruvte_kinase_3t05_C_apo_ache_153_opt1 | -7 |
| piyruvte_kinase_3t05_C_apo_hsp90_123_opt1 | -7 |
| piyruvte_kinase_3t05_C_apo_ache_151_opt1 | -6,9 |
| piyruvte_kinase_3t05_C_apo_cox_134_7b_opt1 | -6,6 |
| piyruvte_kinase_3t05_C_apo_maob_143_3h_opt1 | -6,3 |
| piyruvte_kinase_3t05_C_apo_integrase_83_opt1 | -6,3 |
| piyruvte_kinase_3t05_C_apo_hsp90_119_opt1 | -6 |

PK_C

Figure 3.8: Binding energy of the drugs and the five top lowest binding energy of the chalcone library against COX-2_B, HSP90, integrase, MAOB_A, MAOB_B, PK_AB, PK_C. Orange entry is corresponding to the original drug for protein.
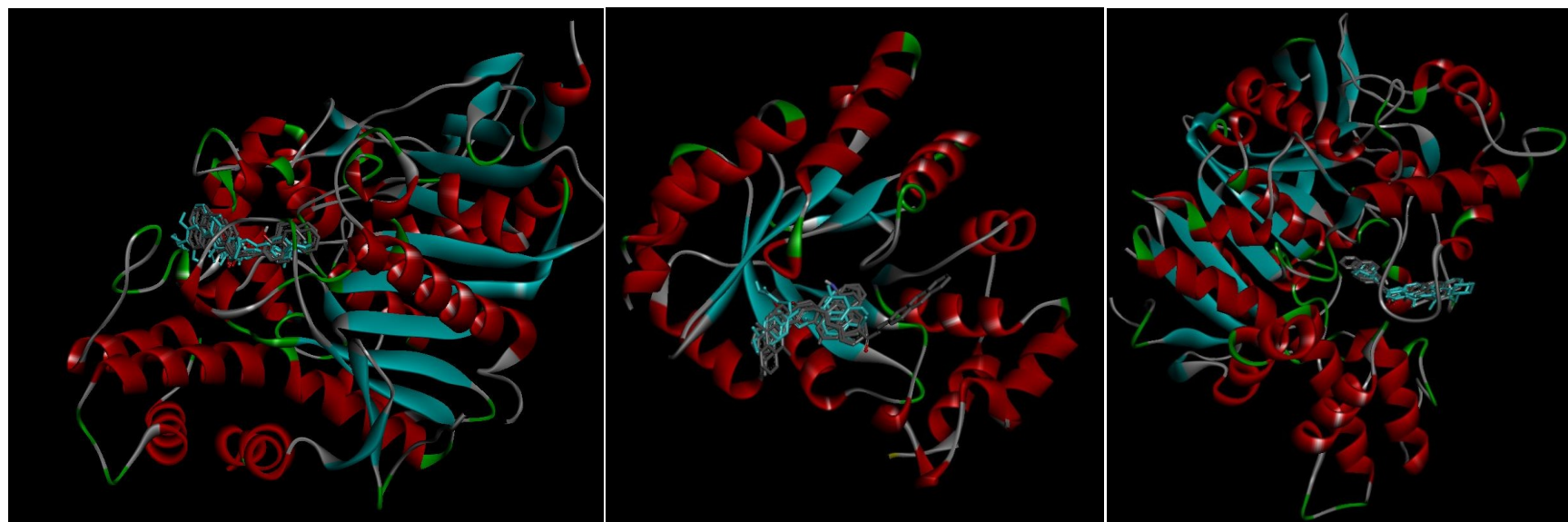
Figure 3.9: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour) for the AChE_A (*left*), AChE_B (*middle*) and ALR$_2$ (*right.*).
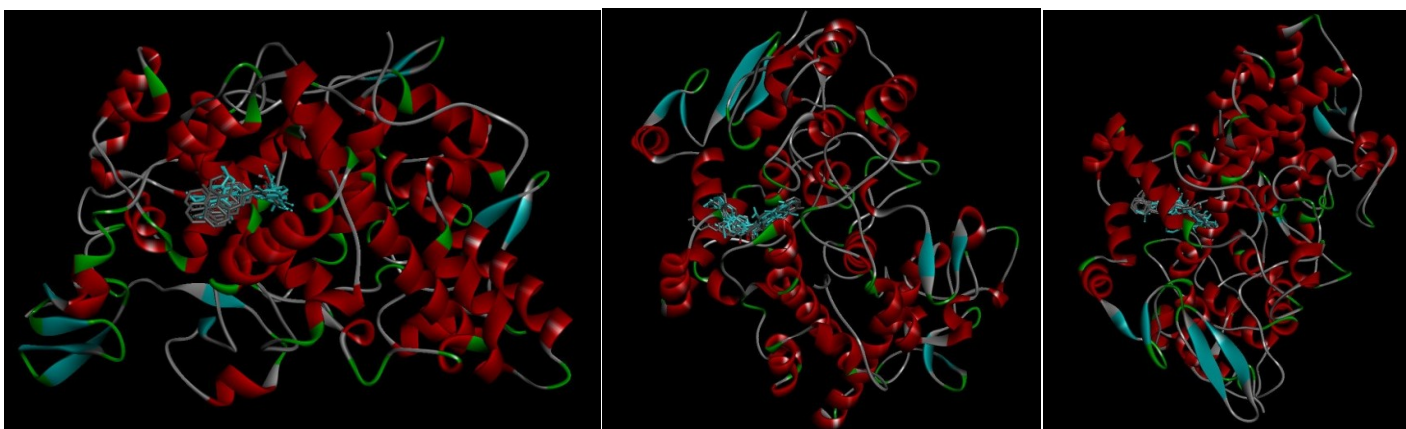
Figure 3.10: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour) for the COX-1_A (*left*), COX-1_B (*middle*) and COX-2_A (*right*).
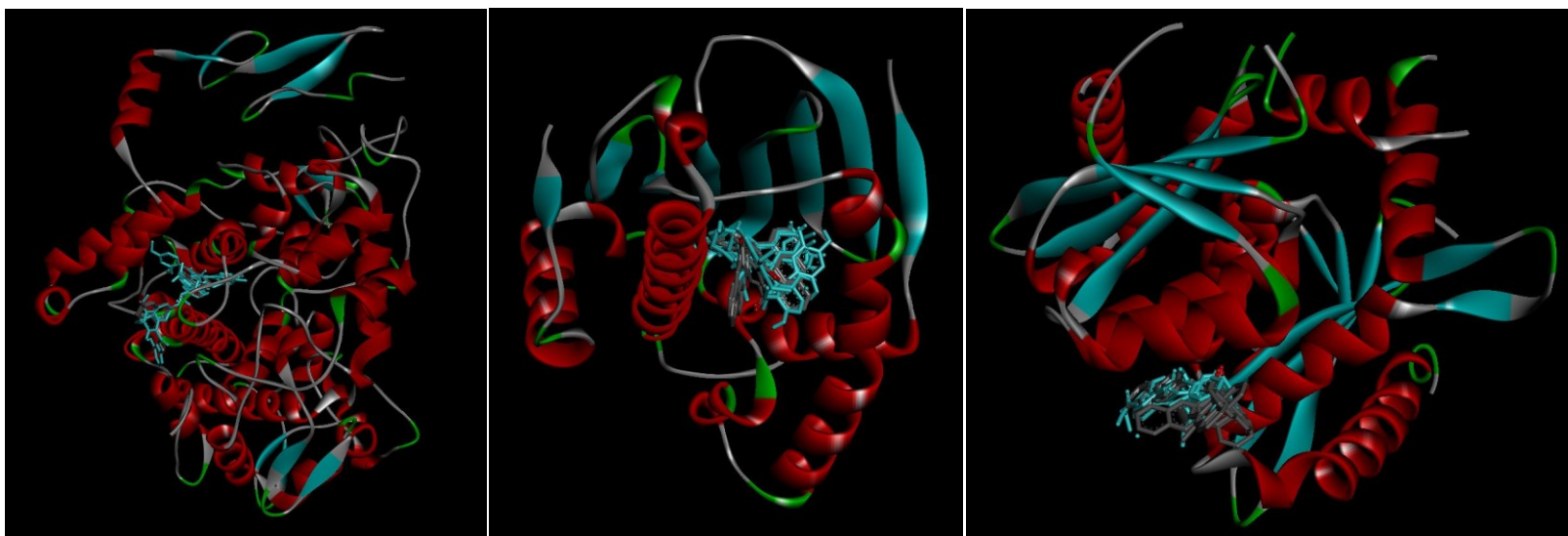
Figure 3.11: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour)  for the COX-2_B (*left*), HSP90 (*middle*) and integrase (*right*).
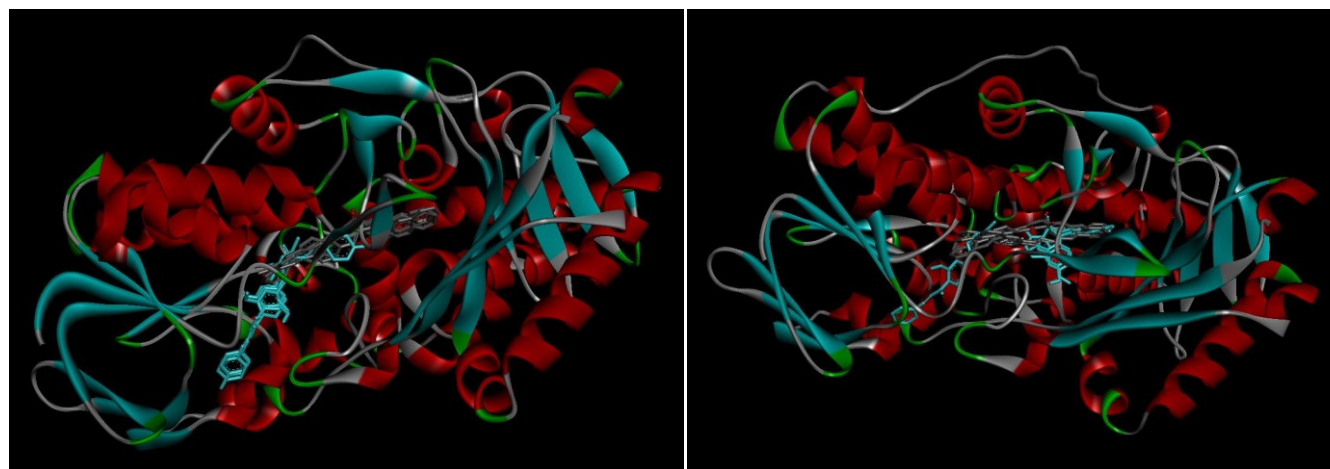


Figure 3.12: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour)  for the MAOB_A (*left*) and MAOB_B (*right*).

102
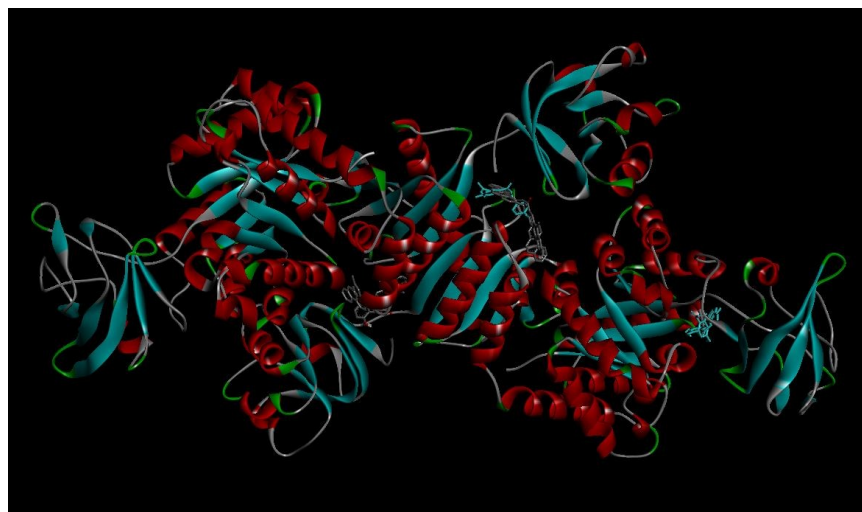
Figure 3.13: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour) for the pyruvate kinase AB model.
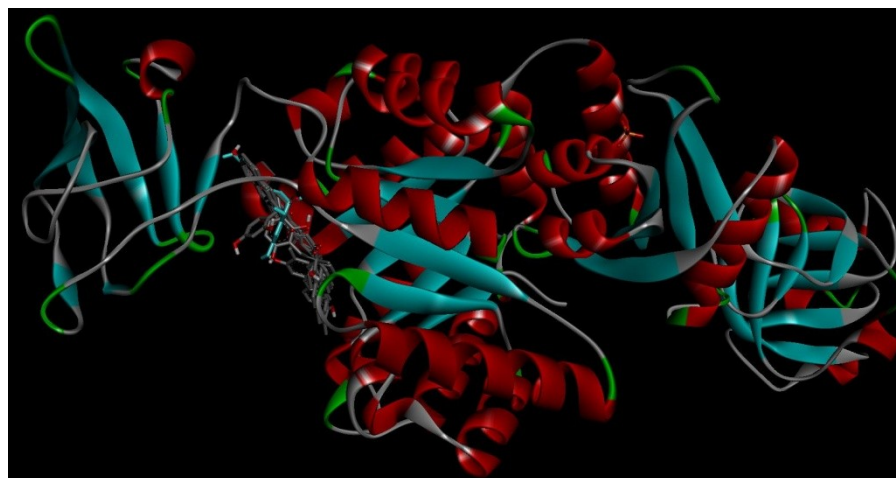


Figure 3.14: Poses of the corresponding inhibitor ligands (blue) and the top five best (lowest energy) of the chalcone library (standard colour) for the pyruvate kinase C co-crystal structure.

103

## 1.6. Binding to MAOA

The results of high throughput virtual screening of the chalcone library provides for values of binding energy of the ligand-protein complexes that, in general, are good across all protein targets. The exception is with MAOA (refer Figure 3.4). Although across all targets the binding energies reach as low as -14 kcal.mol$^{-1}$, MAOA consistently provides for poor ligand binding. Table 3.4 shows chalcone library ligands where the calculated binding to MAOA is actually positive in value (indicating very poor binding indeed). It was decided to verify the interactions within these complexes with Discovery Studio Visualizer. Figure 3.15 represents the poses of these poor-binding chalcones in MAOA. Although these ligands are very poor binders to MAOA, they bind well to other targets. For example, the acetophenone_22_benzaldehyde_66 which presents a binding energy for the MAOA of 8.1 kcal.mol$^{-1}$ has also respectively a binding energy of -12.3, -11.7 and -11.4 kcal.mol$^{-1}$ for the acetylcholine esterase model A, COX-2 A and MAOB A structure. The positive binding is not an issue with the individual ligands, therefore, and it is clear that it is the MAOB-ligand complex that is problematic. The ligands are thus not inhibitors of MAOB even though the observed binding energy is encouraging.
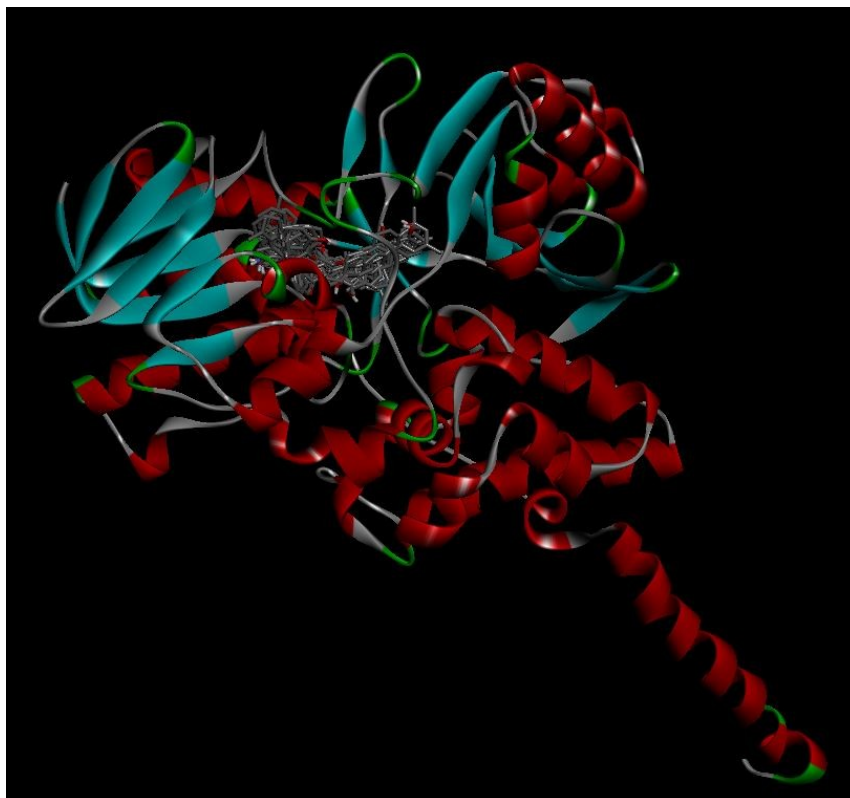


Figure 3.15: Complex MAOA-positive binding energy chalcones.

Figure 3.16 presents the interactions between MAOA and the ligand acetophenone_22_benzaldehyde_78 (binding energy with MAOA: 0.9 kcal.mol$^{-1}$). The analysis of the interaction complex shows different kinds of interaction: amine-π stacked, π-alkyl and hydrogen bonds which are used for the recognition of the protein, the affinity and the orientation of the ligand. Despite the presences of these interactions, the favourable interactions are overwhelmed by the three bump interactions which severe steric interactions with highly unfavourable energy cost. This explains the positive values of the binding energy for some compounds.

| Name | Protein | Energy |
|---|---|---|
| acetophenone_14_benzaldehyde_50 | MAOA | 0,5 |
| acetophenone_19_benzaldehyde_31 | MAOA | 0,3 |
| acetophenone_19_benzaldehyde_50 | MAOA | 1,2 |
| acetophenone_19_benzaldehyde_51 | MAOA | 0,4 |
| acetophenone_19_benzaldehyde_64 | MAOA | 3,4 |
| acetophenone_19_benzaldehyde_66 | MAOA | 2,4 |
| acetophenone_19_benzaldehyde_88 | MAOA | 0,5 |
| acetophenone_21_benzaldehyde_50 | MAOA | 2,1 |
| acetophenone_21_benzaldehyde_51 | MAOA | 1 |
| acetophenone_21_benzaldehyde_66 | MAOA | 4,1 |
| acetophenone_21_benzaldehyde_72 | MAOA | 1,5 |
| acetophenone_21_benzaldehyde_88 | MAOA | 0,1 |
| acetophenone_22_benzaldehyde_44 | MAOA | 2,6 |
| acetophenone_22_benzaldehyde_64 | MAOA | 0 |
| acetophenone_22_benzaldehyde_66 | MAOA | 8,1 |
| acetophenone_22_benzaldehyde_78 | MAOA | 0,9 |
| acetophenone_42_benzaldehyde_66 | MAOA | 0,2 |

Table 3.4: Ligand-protein complexes which have a positive binding energy.
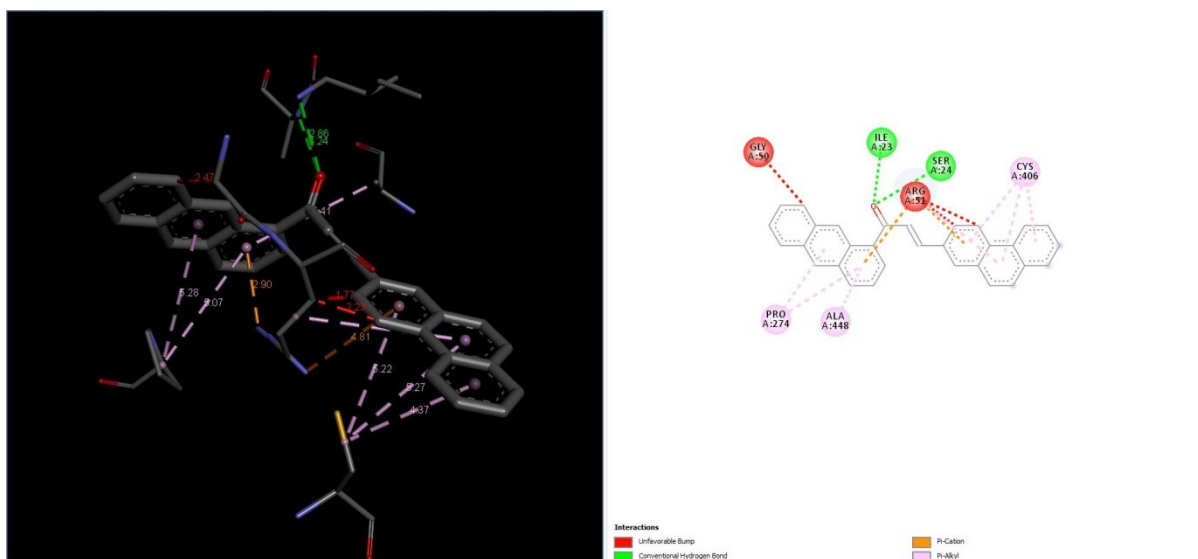
Figure 3.16: Interaction between acetophenone_22_benzaldehyde_78 and the protein MAOA in 3D (*left*) and 2D map (*right*).

Table 3.4 is currently a small proportion of the docking results for the MAOA. It is interesting to study the interaction between the MAOA and the best lowest energy ligands (acetophenone_31_benzaldehyde_89) for this protein. Figure 3.17 represents the interaction between this ligand (binding energy: -12 kcal.mol$^{-1}$) and the protein MAOA. The complex presents five types of interaction ($\pi$-alkyl, amide-$\pi$ stacked, $\pi$-cation, $\pi$-sulfur, carbon hydrogen bond and van der Walls). The benzaldehyde part possesses rich dipole-dipole interactions with the protein. The oxygen of the acetophenone group creates a carbon hydrogen bond with the threonine 435. These interactions (electrostatic, hydrogen bond and hydrophobic) are responsible for specificity and the recognition of the ligand by the protein. In comparison to the interaction of MAOA and the acetophenone_22_benzaldehyde_78, the acetophenone_31_benzaldehyde_89 does not have bump interactions. It also possesses more different types of interactions.

106
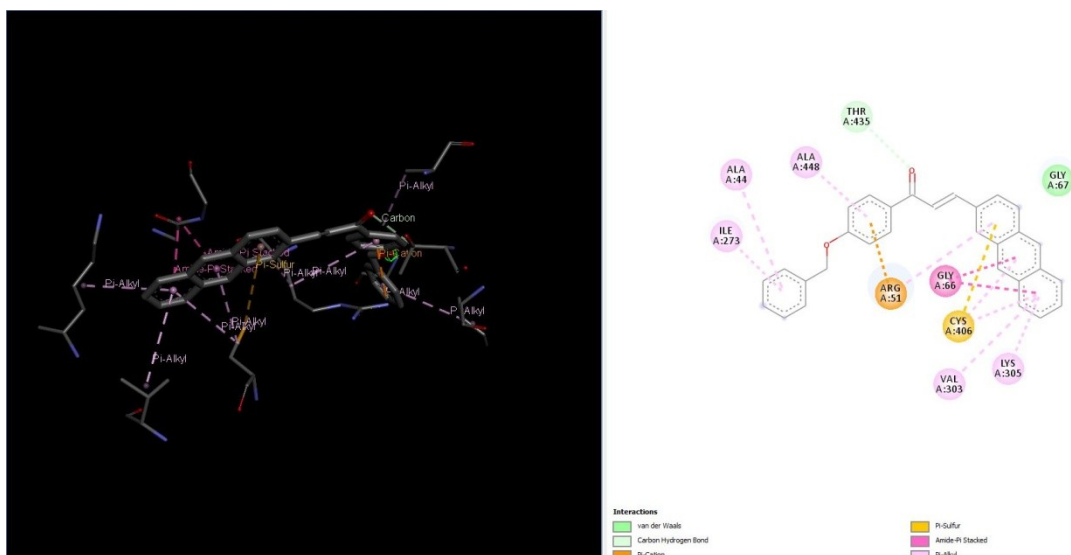
Figure 3.17: Interaction between acetophenone_31_benzaldehyde_89 and the protein MAOA in 3D (*left*) and 2D map (*right*).

With the exception of the MAOA binding, the binding energy of all ligands within the library to the other targets was below -4 kcal.mol⁻¹. COX-1 (both A and B structures), COX-2_B and pyruvate kinase (both AB and C) provided for the best binding overall across all ligands in the library. By comparison, the binding to AChE and MAOB structures was not considered good. In term of interactions, we investigated the library compound acetophenone_15_benzaldehyde_78 binding to AChE_A (-14.2 kcal.mol⁻¹) and MAOB_A (-13.9 kcal.mol⁻¹) and B (-14.7 kcal.mol⁻¹) co crystal. The presence of π-π interactions was a feature of the best binding complexes. Of these three complexes, the binding of this particular ligand to MAOB _B presents the most diverse set of interactions (hydrogen bond, π-cation, π-π, etc ...); by comparison the AChE-A complex presents exclusively π-π interactions. These three complexes, highlighting the interactions in both 3D and 2D maps are represented in from Figure 3.18 to 3.20.
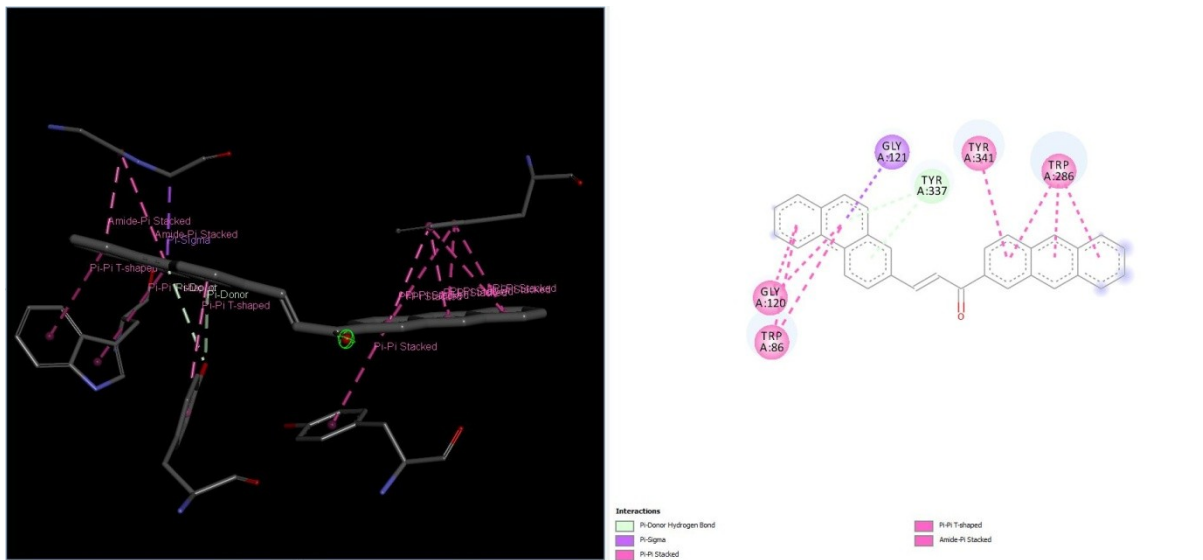
Figure 3.18: Interaction between acetophenone_15_benzaldehyde_78 and the protein AChE_A in 3D (*left*) and 2D map (*right*).
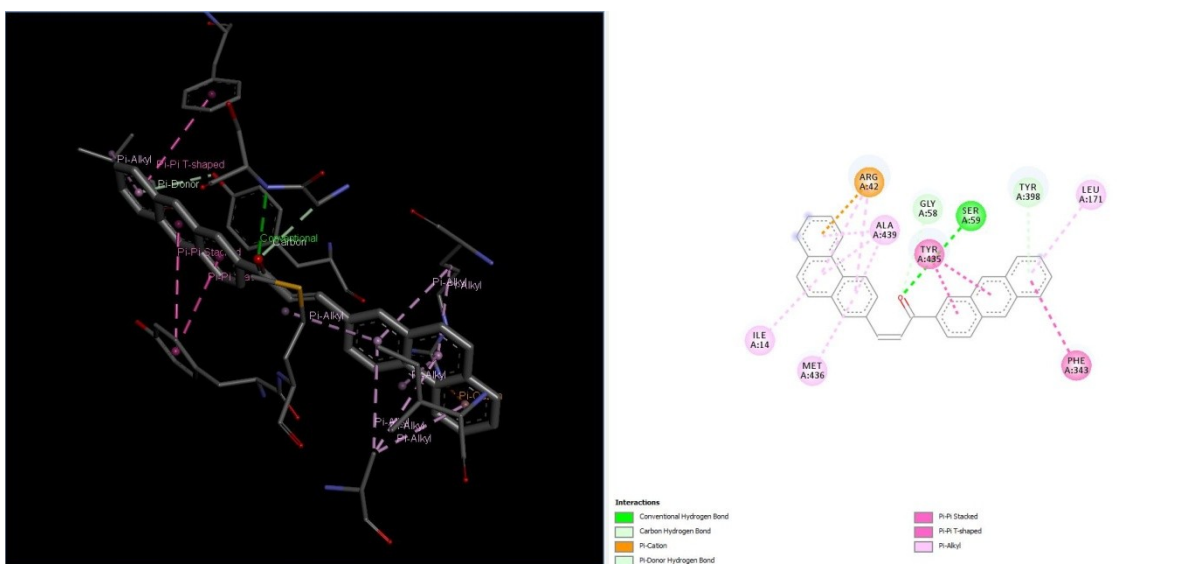


Figure 3.19: Interaction between acetophenone_15_benzaldehyde_78 and the protein MAOB_A in 3D (*left*) and 2D map (*right*).
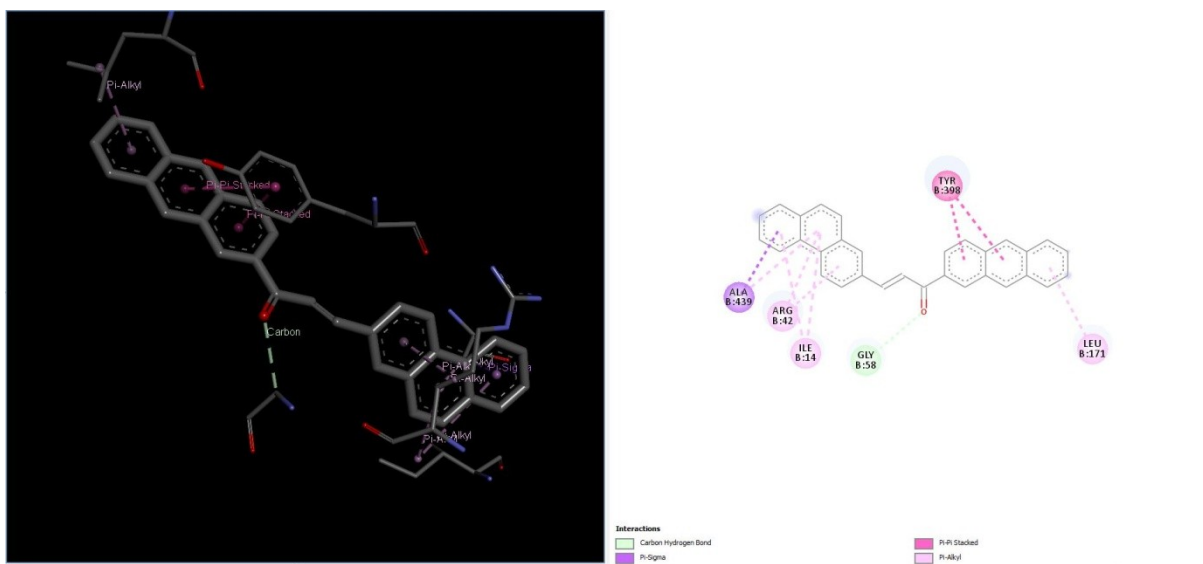
Figure 3.20: Interaction between acetophenone_15_benzaldehyde_78 and the protein AChE_A in 3D (*left*) and 2D map (*right*).

The five best complexes in terms of binding energy for each of the targets are represented in Table 3.5. It is interesting to notice that the common denominator of the acetophenone_13, acetophenone_15-, acetophenone_21-, acetophenone_22-, benzaldehyde_78- and benzaldehyde_89- moieties in the best performing ligands. In terms of specificity, some targets appear to respond favourably to a specific denominator such as AChE_B with the acetophenone_13, HSP90 with the acetophenone_22, the pyruvate kinase AB with acetophenone_21, COX-1_A with benzaldehyde_89 and MAOA with the benzaldehyde_89 moiety. These moieties are represented in Figure 3.21.
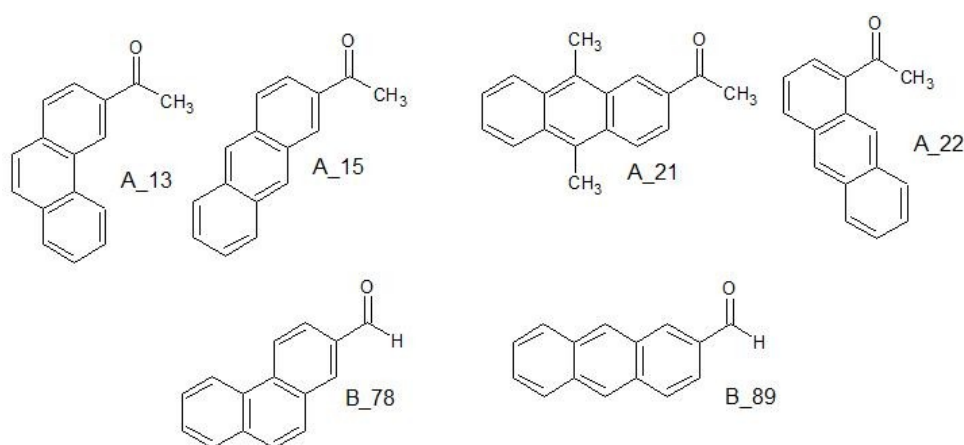


Figure 3.21: Common denominator inside of the chalcone derivatives library. A acetophenone; B: benzaldehyde.

**AchE_A**

| | |
|---|---|
| acetophenone_15_benzaldehyde_78 | -14,2 |
| acetophenone_13_benzaldehyde_78 | -14,1 |
| acetophenone_13_benzaldehyde_27 | -13,9 |
| acetophenone_13_benzaldehyde_60 | -13,9 |
| acetophenone_14_benzaldehyde_78 | -13,9 |

**AchE_B**

| | |
|---|---|
| acetophenone_13_benzaldehyde_78 | -14,1 |
| acetophenone_15_benzaldehyde_78 | -14,1 |
| acetophenone_13_benzaldehyde_17 | -13,8 |
| acetophenone_13_benzaldehyde_27 | -13,8 |
| acetophenone_13_benzaldehyde_45 | -13,8 |

**ALR2**

| | |
|---|---|
| acetophenone_19_benzaldehyde_2 | -11,3 |
| acetophenone_19_benzaldehyde_17 | -11,1 |
| acetophenone_19_benzaldehyde_45 | -11,1 |
| acetophenone_21_benzaldehyde_89 | -11,1 |
| acetophenone_21_benzaldehyde_88 | -11 |

**BchE_A**

| | |
|---|---|
| acetophenone_21_benzaldehyde_72 | -13,1 |
| acetophenone_22_benzaldehyde_72 | -13,1 |
| acetophenone_13_benzaldehyde_78 | -12,9 |
| acetophenone_19_benzaldehyde_44 | -12,9 |
| acetophenone_21_benzaldehyde_50 | -12,9 |

**BchE_B**

| | |
|---|---|
| acetophenone_13_benzaldehyde_78 | -12,9 |
| acetophenone_21_benzaldehyde_50 | -12,9 |
| acetophenone_21_benzaldehyde_58 | -12,9 |
| acetophenone_22_benzaldehyde_50 | -12,9 |
| acetophenone_22_benzaldehyde_72 | -12,9 |

**COX-1_A**

| | |
|---|---|
| acetophenone_19_benzaldehyde_89 | -11,1 |
| acetophenone_21_benzaldehyde_89 | -11,1 |
| acetophenone_14_benzaldehyde_89 | -10,9 |
| acetophenone_42_benzaldehyde_89 | -10,9 |
| acetophenone_10_benzaldehyde_89 | -10,8 |

**COX-1_B**

| | |
|---|---|
| acetophenone_14_benzaldehyde_2 | -10,9 |
| acetophenone_54_benzaldehyde_89 | -10,9 |
| acetophenone_15_benzaldehyde_2 | -10,8 |
| acetophenone_15_benzaldehyde_32 | -10,8 |
| acetophenone_54_benzaldehyde_2 | -10,8 |

**COX-2_A**

| | |
|---|---|
| acetophenone_19_benzaldehyde_89 | -12,2 |
| acetophenone_22_benzaldehyde_78 | -12,1 |
| acetophenone_19_benzaldehyde_78 | -12 |
| acetophenone_22_benzaldehyde_89 | -12 |
| acetophenone_14_benzaldehyde_2 | -11,7 |

**COX-2_B**

| | |
|---|---|
| acetophenone_15_benzaldehyde_45 | -10,8 |
| acetophenone_15_benzaldehyde_17 | -10,7 |
| acetophenone_15_benzaldehyde_2 | -10,7 |
| acetophenone_14_benzaldehyde_2 | -10,6 |
| acetophenone_21_benzaldehyde_2 | -10,6 |

**HSP90**

| | |
|---|---|
| acetophenone_22_benzaldehyde_78 | -13,5 |
| acetophenone_22_benzaldehyde_58 | -13,4 |
| acetophenone_22_benzaldehyde_44 | -13,2 |
| acetophenone_22_benzaldehyde_72 | -13,2 |
| acetophenone_22_benzaldehyde_89 | -13,2 |

**Integrase**

| | |
|---|---|
| acetophenone_13_benzaldehyde_78 | -10,6 |
| acetophenone_13_benzaldehyde_89 | -10,6 |
| acetophenone_19_benzaldehyde_78 | -10,5 |
| acetophenone_21_benzaldehyde_78 | -10,3 |
| acetophenone_22_benzaldehyde_78 | -10,3 |

**MAOA**

| | |
|---|---|
| acetophenone_31_benzaldehyde_89 | -12 |
| acetophenone_63_benzaldehyde_89 | -11,8 |
| acetophenone_15_benzaldehyde_89 | -11,6 |
| acetophenone_28_benzaldehyde_89 | -11,5 |
| acetophenone_10_benzaldehyde_89 | -11,1 |

**MAOB_A**

| | |
|---|---|
| acetophenone_15_benzaldehyde_78 | -13,9 |
| acetophenone_15_benzaldehyde_26 | -13,8 |
| acetophenone_15_benzaldehyde_37 | -13,8 |
| acetophenone_13_benzaldehyde_78 | -13,6 |
| acetophenone_22_benzaldehyde_64 | -13,6 |

**MAOB_B**

| | |
|---|---|
| acetophenone_15_benzaldehyde_78 | -14,7 |
| acetophenone_22_benzaldehyde_89 | -14,5 |
| acetophenone_15_benzaldehyde_44 | -14,4 |
| acetophenone_13_benzaldehyde_78 | -14,2 |
| acetophenone_15_benzaldehyde_26 | -14,2 |

**Piyruvate_kinase_AB**

| | |
|---|---|
| acetophenone_21_benzaldehyde_49 | -12 |
| acetophenone_21_benzaldehyde_87 | -11,8 |
| acetophenone_21_benzaldehyde_78 | -11,7 |
| acetophenone_21_benzaldehyde_89 | -11,6 |
| acetophenone_13_benzaldehyde_78 | -11,5 |

**Piyruvate_kinase_C**

| | |
|---|---|
| acetophenone_15_benzaldehyde_89 | -11,2 |
| acetophenone_15_benzaldehyde_44 | -10,8 |
| acetophenone_15_benzaldehyde_64 | -10,6 |
| acetophenone_42_benzaldehyde_89 | -10,6 |
| acetophenone_10_benzaldehyde_89 | -10,5 |

**sLAP2**

| | |
|---|---|
| acetophenone_21_benzaldehyde_44 | -12,4 |
| acetophenone_21_benzaldehyde_89 | -12,3 |
| acetophenone_21_benzaldehyde_66 | -12,2 |
| acetophenone_21_benzaldehyde_78 | -12,1 |
| acetophenone_42_benzaldehyde_65 | -12 |

Table 3.5: Best binding energy compounds for each co-crystal structures (expressed in kcal.mol$^{-1}$).

## 2. Stability experiment with dynamics

### 2.1. Parameters

Three molecules were used for this study. It was decided to used the original ligand from the pdb structures of the HSP90 [ID code: 5fnc], one of the best lowest energy ligand discovered from the high-throughput virtual screening against the same protein, acetophenone_22_benzaldehyde_17 and its modified version including four chlorine atoms (this modified system is important in Chapter 5).[7] Figure 3.22 represents these three molecules. Conversion from the pdb format to the GROMACS format, together with generation of ligand topologies (using the AMBER03 force field) was effected using the PRODRG server.[36]
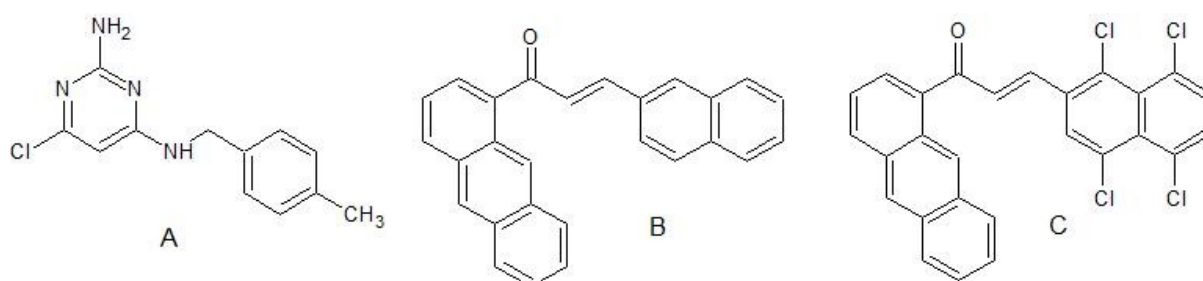


Figure 3.22: Molecules used for the molecular dynamic studies. A, Molecule from the HSP90 pdb structure; B, acetophenone_22_benzaldehyde_17; C, acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl.

Preparation of the protein involved several steps. The water, the ions and ligands were discarded from the protein pdb file. Some correction due to missing atoms had to be effected. Missing atoms for the terminal lysine (Lys224), glutamic acid 16 and glutamine 123 were provided. The **pdb2gmx** command was used to generate protein topologies, GROMACS file formats (for the HSP90 protein). The force field used for this study was the AMBER03 force field.[37] Protein and ligands were combined, such that four models of the HSP90 proteins, alone and with each of the three different docked ligands were inserted in a cubic box with at least 3nm of space between the macromolecule or complex and the edge of the box. The box was filled next with solvent. A specific equilibrated 3D point water solvent model, spc216.gro was applied to solvate the box. Due to an excess of sulphate ions, the protein had a charge of -**7** and was neutralized. Neutralization was carried out using sodium chloride (sodium and chloride ions) with an excess of 7 of sodium ions to neutralize the proteins. The ions replaced some water molecules provided for by the solvation. This process was repeated for the protein alone and with three ligands presented previously. Input files were written by reference to

coordinate and topologies using the GROMACS ***grompp*** module. Two examples of solvation and neutralization of the HSP90 protein alone and with the original ligand provided from the pdb file were presented in Figure 3.23.
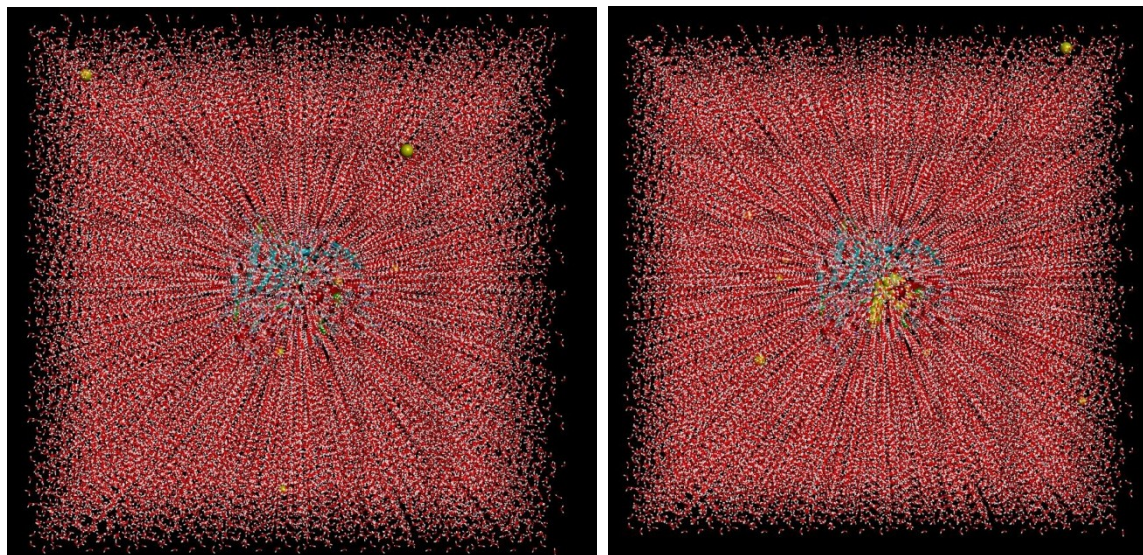


Figure 3.23: Solvation and neutralization of the HSP90 protein alone (*left*) and complex of the HSP90 protein and the original ligand (*right*). The ions and the ligand are highlighted in yellow.

The system was assembled and prepared with the module ***grompp*** before submitting to energy minimization process with ***mdrun*** command.

Index groups were generated with ***make_ndx***. The ***genrestr*** command was used to generate restraints for equilibration, particularly the ligand to prevent untoward excessive movement in the initial stages of equilibrations. The restraints were in terms of force constants at 1000 kJ.mol$^{-1}$.nm$^{-2}$. Equilibrium dynamics was firstly conducted under constant temperature, volume and number of particles (*NVT*) and file preparation and running used the ***grompp*** and ***mdrun*** commands. The second part of the system equilibration was focussed on the stabilization of the pressure (*NPT*). From this point, due to heavy computational demand, parallel processing at the Center for High Performance Computing was used for dynamics simulation.

After the *NPT* equilibration, the system was fully prepared for production molecular dynamics calculations. The position restraints were fully removed for production dynamics.

The analysis was divided in three distinct parts. The trajectories of the system were corrected in terms of re-centering on the protein. From these corrected trajectories, ligand and protein movement during dynamics was appropriately analysed in terms of RMSD, and Radius of

Gyration. The dynamics with the *apo* protein (in the absence of ligand) was also analysed in the same way. The analysis was completed by Principle Component Analysis of the protein covariance during dynamics with the commands ***covar*** and ***anaeig***.

## 2.2. Results

The RMSD of the HSP90 protein during all four molecular dynamics simulations (Figure 3.24) reveals that the ligands stabilise the HSP90 – all complexes exhibit a lower RMSD over time than the protein alone. The chalcone ligands (labelled 22 and Cl) appear to stabilise the HSP90 in terms of RMSD more than the original crystal structure ligand (labelled CS). Of the two chalcones, the undecorated acetophenone_22_benzaldehyde_17 ligand appears to stabilise the most, although this increase in stabilisation appears slight. The amplitude of the oscillations of the RMSD seems weak.
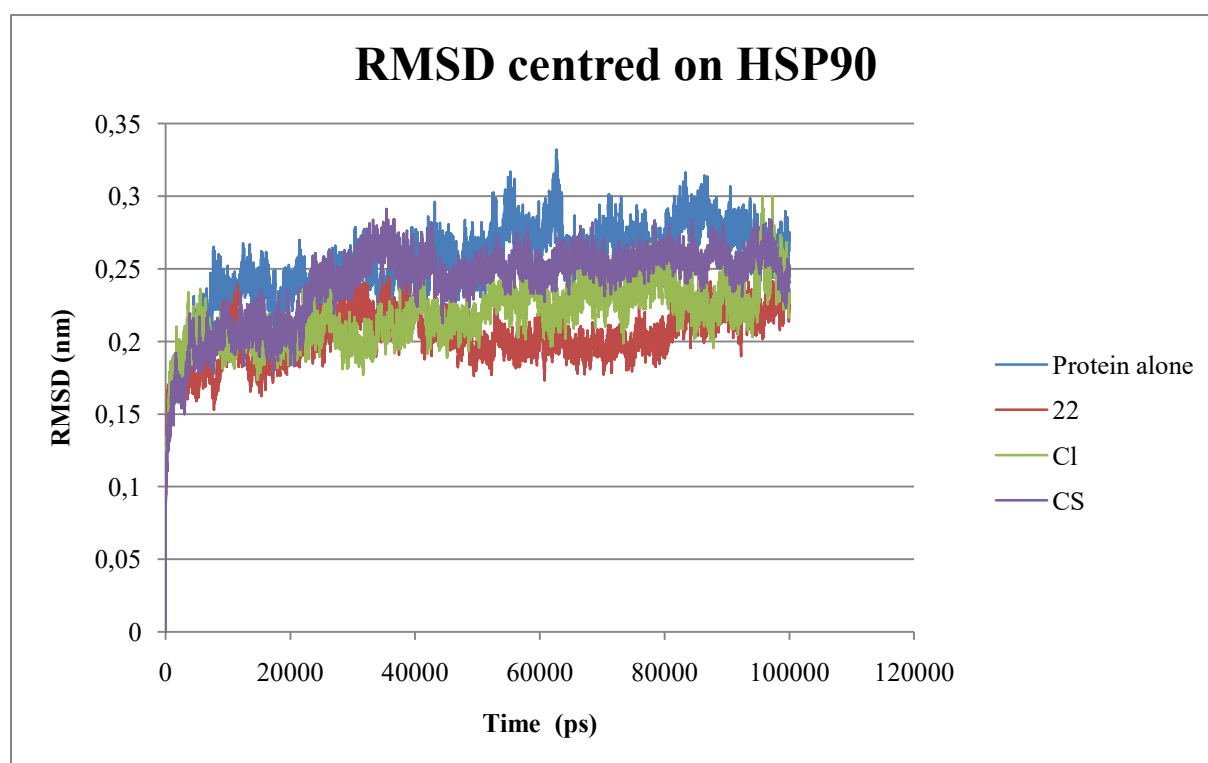


Figure 3.24: RMSD centred on the HPS90 protein systems. 22, complex HSP90-acetophenone_22_benzaldehyde_17; Cl, complex HSP90-acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl; CS, original complex from the pdb file.

The RMSD of the three ligands during dynamics (Figure 3.25), on the other hand, shows the stability of the original crystal structure ligand (labelled CS) relative to both chalcones. The RMSD for these chalcones show strong amplitudes of oscillation. The tetrachloro derivative (Cl) appears to presents three separate positions around 20 000, between 40000 and 60 000

and 80 000 and 90 000 ps. It would appear that the acetophenone_22_benzaldehyde_17 ligand is moving away from its original position over the time.
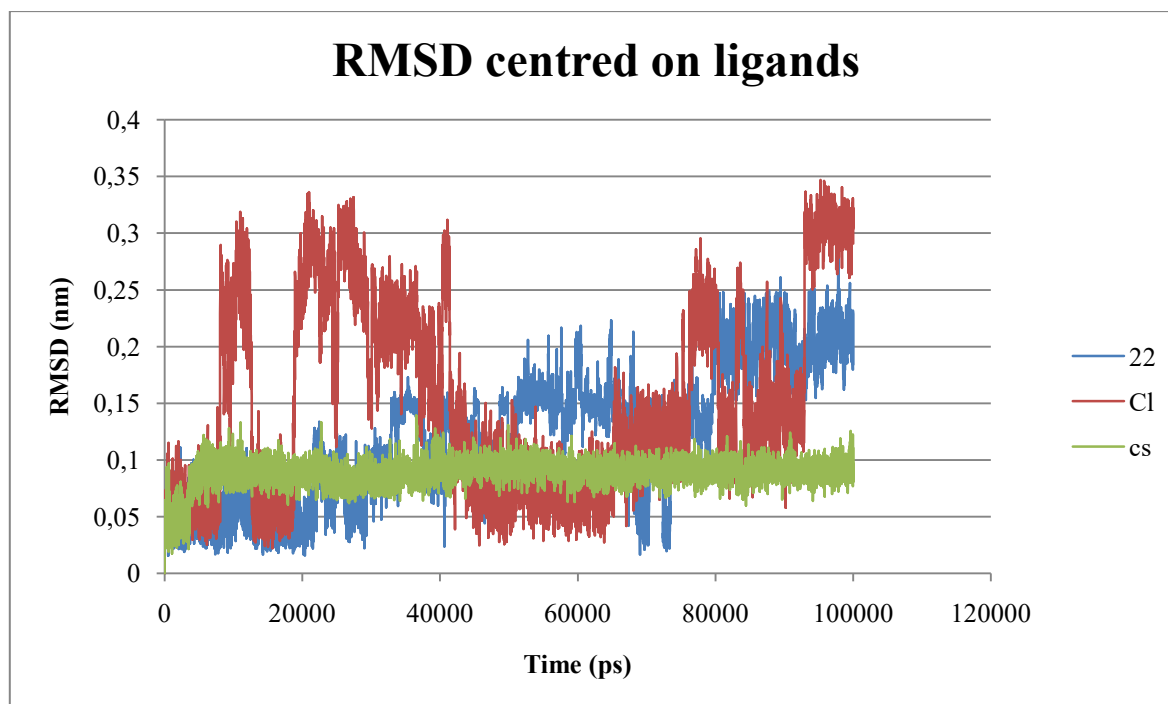


Figure 3.25: RMSD centred on ligands. 22, complex HSP90-acetophenone_22_benzaldehyde_17; Cl, complex HSP90-acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl; CS, original complex from the pdb file.

The radius of gyration of the protein HSP90 oscillates between 1.68 and approximately 1.735 nm over the time. The complex cs-HSP90 appears to be the most stable system with a weak oscillation in terms of compactness (between 1.686 and 1.735 nm over the time). By contrast, the greatest variation of the compactness is with modified chalcone (containing four chlorine atoms) (between 1.67 and 1.74 nm over the time). However all complexes appear to be relatively stable in terms of compactness over time. Figure 3.26 represents the radius of gyration of the protein HSP90 and its different complexes.
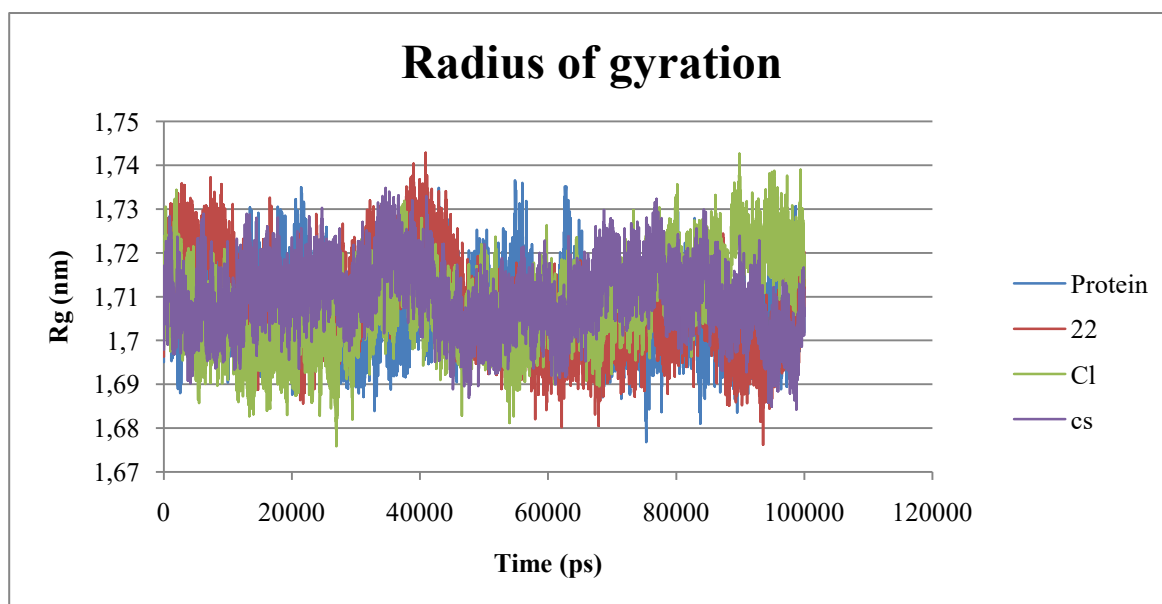
**Radius of gyration**

*Figure 3.26: Radius of gyration of the HSP90 protein and its complexes. 22, complex HSP90-acetophenone_22_benzaldehyde_17; Cl, complex HSP90-acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl; CS, original complex from the pdb file.*

PCA analyses of protein structure during dynamics (in the presence and absence of ligands) were plotted. Arbitrary color codes were setup in order to follow the evolution of the dynamics of the HSP90 protein over the time. In all cases the PCA plot is represented by green for the dynamics timeframe between 0 and 2500 ps, grey for structures between 2500 to 5000ps, yellow from 5000 to 7500 ps and blue dots for the final stages of dynamics between 7500 and 10000ps.

The Figure 3.27 represents the PCA score plot of the HSP90 protein in the absence of any ligands. There is a clear progression of the structure over time. Bear in mind that dimension 1 carries most of the structural variation over time and is representative of changes with time with this protein. This may be due to some minor conformational changes of the protein itself.

The Figure 3.28 represents the scores plot of the protein HSP90 in the presence of acetophenone_22_benzaldehyde_17 as ligand. The behaviour is clearly different when time progression is taken into account. This dimension 1 shows a large variation away from initial structures, before a return back towards the initial values.

Figure 3.29 shows the PCA score plot of the HSP90 protein in the presence of the acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl ligand. This PCA plot again shows a variation in timeframe as compared to Figure 3.27 (where the protein is alone). The motion towards the required dynamics structure is delayed until about the 2500 to 5000 ps timeframe, from which there appears to be stabilization away from the initial structure.
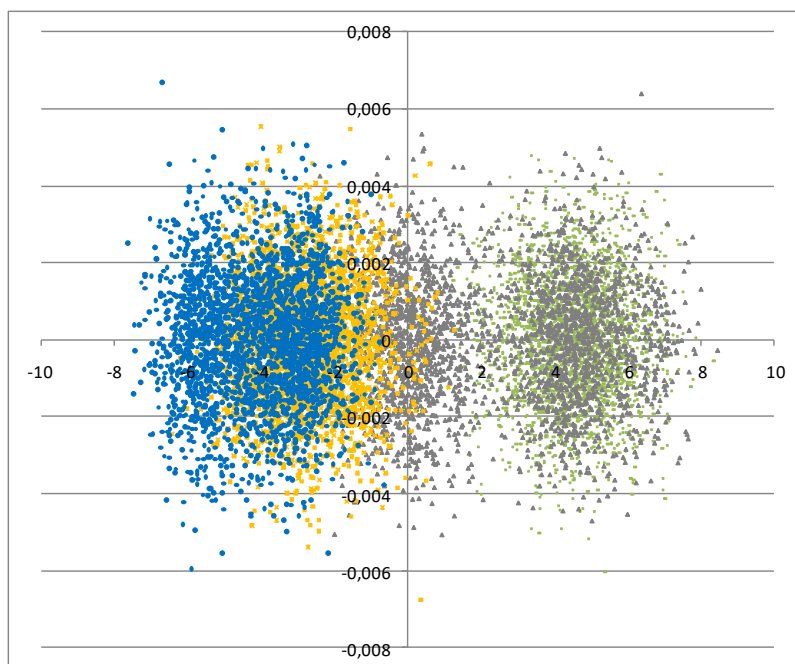


Figure 3.29: Scores plot of the HSP90 protein with the acetophenone_22_benzaldehyde_17 with 4 chlorines atoms as ligand. Green time from 0 to 2500 ps; grey, time from 2500 to 5000 ps; orange, time from 5000 to 7500 ps; blue, time from 7500 to 10000 ps.

Figure 3.30 represents the scores plot of the HSP90 protein with its original crystal structure ligand present. The plot presents fours different clusters over times which might correspond to four different conformations. It might also be an indication of instability. A different explanation may be that this is the result of induced fit effects. Each cluster appears to be compact and is nearly fully represented by dimension 1 (x-axis).
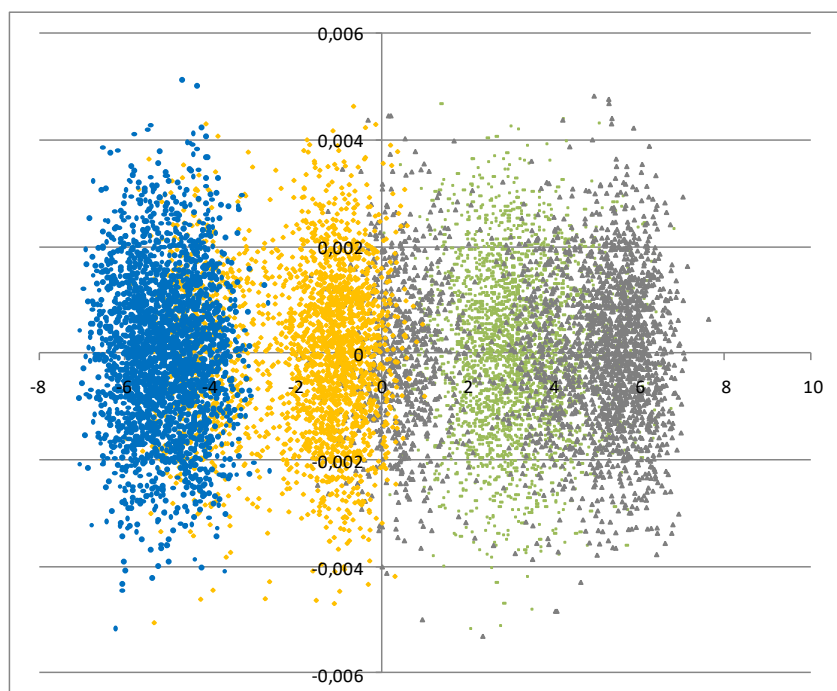
Figure 3.30: Scores plot of the HSP90 protein including its original molecule as ligand. Green s time from 0 to 2500 ps; grey, time from 2500 to 5000 ps; orange, time from 5000 to 7500 ps; blue, time from 7500 to 10000 ps.

# Conclusion

The high-throughput virtual screening has provided for a plethora of good binding ligands to the range of targets, where the binding energy appears to outperform binding by known chalcone inhibitors. Moieties that appear to induce specificity in binding have been identified. Also interactions between chalcones and targets have been identified and this may be used as the start of strategy to improve the molecules within the library in terms of binding. In the test case for molecular dynamics (involving HSP90), however, good binding did not always translate into long term stability within the active site of the target especially for the protein including the original ligand. Modifications (in terms of the TOPLISS decision tree) are an attempt to mediate this in terms of providing for better drug-likeness but also providing for more centers for protein-ligand binding and hence longer term stability within the active site of an inhibitor.

# REFERENCES

1. O. Trot and A.J. Olson, *Autodock Vina: Improving the speed and accuracy if docking with a new scoring function, efficient optimization and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-461.

2. Quiroga, R. and M.A. Villareal, *Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening.* PLoS One, 2016. **11**(5).

3. Handoko, S.D., X. Oyang, C.T.T. Su, C.K. Kwoh, and Y.S. Ong, *QuickVina: Accelerating AutoDock Vina Using Gradient-Based Heuristics for Global Optimization.* IEEE/ACM Trans Comput Biol Bioinform, 2012. **9**(5).

4. H. M. Berman, J.W., Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-242.

5. L. D. Fader, E. Malenfant, M. Parisien, R. Carson, F. Bilodeau, S. Landry, M. Pesant, C. Brochu, S. Morin, C. Chabot, T. Halmos, Y. Bousquet, M. D. Bailey, S. H. Kawai, R. Coulombe, S. LaPlante, A. Jakalian, P. K. Bhardwaj, D. Wernic, P. Schroeder, M. Amad, P. Edwards, M. Garneau, J. Duan, M. Cordingley, R. Bethell, S. W. Mason, M. Bös, P. Bonneau, M-A. Poupart, A_M. Faucher, B. Simoneau, C. Fenwick, C. Yoakim, and Y. Tsantrizos, *Discovery of BI 224436, a Noncatalytic Site Integrase Inhibitor (NCINI) of HIV-1.* ACS Med. Chem. Lett., 2014. **5**(4): p. 422-427.

6. R. Zoraghi, L. Worral, R. H. See, W. L. Popplewell, H. Gong, T. Samaai, R. D. Swayze, S. Kaur, M. Vuckovic, B. B. Finlay, R. C. Bruham, W. R. McMaster, M. T. Davies-Coleman, N. C. Strynadka, R. J. Andersen, and N.E. Reiner, *Methicillin-resistant Staphylococcus aureus (MRSA) Pyruvate Kinase as a Target for Bis-indole Alkaloids with Antibacterial Activities.* J Biol Chem, 2011. **286**(52): p. 44716-44725.

7. S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N . Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, and X. Barril, *Dynamic undocking and the quasi-bound state as tools for drug discovery.* Nat Chem, 2017. **9**(3): p. 201-206.

8. M. J. Lucido, B. J. Orlando, A. J. Vecchio, and M.G. Malkowski, *Crystal Structure of Aspirin Acetyl Human Cyclooxygenase-2: Insight Into the Formation of Products with Reversed Stereochemistry.* Biochemistry, 2016. **55**(8): p. 1226-1238.

9. R. S. Sidhu, J. Y. Lee, C. Yuan, C. Yuan, and W.L. Smith, *Comparison of Cyclooxygenase-1 Crystal Structures: Cross-Talk between Monomers Comprising Cyclooxygenase-1 Homodimers.* Biochemistry, 2010. **49**(33): p. 7069-7079.

10. P. O. de Giuseppe, M. L. dos Santos, S. M. de Sousa, K. E. Koch, J. A. Yunes, R. Aparicio, and M.T. Murakami, *A comparative structural analysis reveals distinctive features of cofactor binding and substrate specificity in plant aldo-keto reductases.* Biochem Biophys Res Commun, 2016. **474**(4): p. 696-701.

11. S-Y. Son, J. Ma, Y. Kondou, M. Yoshimura, E. Yamashita, and T. Tsukihara, *Structure of human monoamine oxidase A at 2.2-Å resolution: The control of opening the entry for substrates/inhibitors.* PNAS, 2008. **105**(15): p. 5739-5744.

12. P. de Deurwaerdère, C. Binda, R. Corne, C. Leone, A. Valeri, M. Valoti, R. R. Ramsay, Y. Fall, and J. Marco-Contelles, *Comparative Analysis of the Neurochemical Profile and MAO Inhibition Properties of N-(Furan-2-ylmethyl)-N methylprop-2-yn-1-amine.* ACS Chem Neurosci, 2017. **8**(5): p. 1026-1035.

13. J. Cheung, E. N. Gary, K. Shiomi, and T.L. Rosenberry, *Structures of Human Acetylcholinesterase Bound to Dihydrotanshinone I and Territrem B Show Peripheral Site Flexibility.* ACS Med. Chem. Lett., 2013. **4**(11): p. 1091-1096.

14. U. Košak, B. Brus, D. Knez, R. Šink, S. Žakelj, J. Trontelj, A. Pišlar, J. Šlenc, M. Gobec, M. Živin, L. Tratnjek, M. Perše, K. Sałat, A. Podkowa, B. Filipek, F. Nachon, X. Brazzolotto, A. Więckowska, B. Malawska, J. Stojan, I. M. Raščan, J. Kos, N. Coquelle, J-P. Colletier, and S.

Gobec, *Development of an in-vivo active reversible butyrylcholinesterase inhibitor.* Sci Rep, 2016. **6**: p. 1-16.

15. A. J-A. Woolford, P. J. Day, V. Bénéton, V. Berdini, J. E. Coyle, Y. Dudit, P. Grondin, P. Huet, L. Y. W. Lee, E. S. Manas, R. L. McMenamin, C. W. Murray, L. W. Page, V. K. Patel, F. Potvain, S. J. Rich, Y. sang, D. O. Somers, L. Trottet, Z. Wan, and X. Zhang, *Fragment-Based Approach to the Development of an Orally Biovailable Lactam Inhibitor of Lipoprotein-Associated Phospholipase (Lp-PLA$_2$).* J Med Chem, 2016. **59**(2): p. 10738-10749.

16. J. Deng, T. Shanchez, L. Q. Al-Mawsawi, R. Dayam, R. A .Yunes, A. Garofalo, M. B. Bolger, and N. Neamati, *Discovery of structurally diverse HIV-1 integrase inhibitors based on a chalcone pharmacophore.* Bioorg Med Chem, 2007. **15**(14): p. 4985-5002.

17. A. Vasu Babu, R. Navudu, and G. Trimurtulu, *Synthesis of C-methyl chalcones as HIV-integrase inhibitors—computational approach.* Med Chem Res, 2014. **23**(2): p. 877-881.

18. H. Sharma, S. Patil, T. W. Sanchez, N. Neamati, R. F. Schinazi, and J.K. Buolamwini, *Synthesis, Biological Evaluation and 3D-QSAR Studies of 3-Keto Salicylic Acid Chalcones and Related Amides as Novel HIV-1 Integrase Inhibitors.* Bioorg Med Chem, 2011. **19**(6).

19. T. M. Osòrio, F. D. Monache, L. D. Chiaradia, A. Mascarello, T. R. Stumpf, C. R. Zanetti, D. B. Silveira, C. R. M. Barardi, E. d F. A. Smânia, A. Viancelli, L. A. T. Garcia, R. A. Yunes, R. J. Nunes, and A.S. Jr, *Antibacterial activity of chalcones, hydrazones and oxadiazoles against methicillin-resistant Staphylococcus aureus.* Bioorg Med Chem Lett, 2012. **22**(1): p. 225-230.

20. J. Davenport, M. Balch, L. Galam, A. Girgis, J. Hall, B. S. J. Blagg, and R.L. Matts, *High-throughput screen of natural product libraries for hsp90 inhibitors.* Biology (Basel), 2014. **3**(1): p. 101-138.

21. F. D. Piaz, S. Terracciano, N. De Tommasi, and A. Braca, *Hsp90 Activity Modulation by Plant Secondary Metabolites.* Planta Med, 2015. **81**(14): p. 1223-1239.

22. Y. S. Kim, V. Kumar, S. Lee, A. Iwai, L. Neckers, and S.V. Malhotra, *Methoxychalcone inhibitors of androgen receptor translocation and function.* Bioorg Med Chem Lett, 2012. **22**(5): p. 2105-2109.

23. C-H. Jeong, H. B. Park, W. J. Jang, S. H. Jung, and Y.H. Seo, *Discovery of hybrid Hsp90 inhibitors and their anti-neoplastic effects against gefitinib-resistant non-small cell lung cancer (NSCLC).* Bioorg Med Chem Lett, 2014. **24**(1): p. 224-227.

24. J. H. Jeong, Y. J. Oh, T. K. Kwon, and Y.H. Seo, *Chalcone-templated Hsp90 inhibitors and their effects on gefitinib resistance in non-small cell lung cancer (NSCLC).* Arch Pharm Res, 2017. **40**(1): p. 96-105.

25. F. Herencia, M. L. Ferràndiz, Amalia Ubeda, J. N. Domìnguez, J. E. Charris, G. M. Lobo, and M.J. Alcaraz, *Synthesis and anti-inflammatory activity of chalcone derivatives.* Bioorg Med Chem Lett, 1998. **8**(10): p. 1169-1174.

26. B. P. Bandgar, S. A. Patil, and B.L. R. N. Gacche, *Synthesis and biological evaluation of nitrogen-containing chalcones as possible anti-inflammatory and antioxidant agents.* Bioorg Med Chem Lett, 2010. **20**(2): p. 730-733.

27. A. Zarghi, T. Zebardast, F. Hakimion, F. H. Shirazi, P. N. Praveen Rao, and E.E. Knaus, *Synthesis and biological evaluation of 1,3-diphenylprop-2-en-1-ones possessing a methansulfonamido or an azido pharmacophore as cyclooxygenase-1/-2 inhibitors.* Bioorg Med Chem, 2006. **14**(20): p. 7044-7050.

28. F. Chimenti, R. Fioravanti, A. Bolasco, P. Chimenti, D. Secci, F. Rossi, M. Yànez, F. Orallo, F. Ortuso, and S. Alcaro, *Chalcones: A Valid Scaffold for Monoamine Oxidases Inhibitors.* J Med Chem, 2009. **52**(9): p. 2818-2824.

29. N. Morales-Camilo, C. O. Salas, C. Sanhueza, C. Espinosa-Bustos, S. Sepùlveda-Boza, M. Reyes-Parada, F. Gonzalez-Nilo, M. Caroli-Rezende, and A. Fierro, *Synthesis, Biological Evaluation, and Molecular Simulation of Chalcones and Aurones as Selective MAO-B Inhibitors.* Chem Biol Drug Des, 2015. **85**(6): p. 685-695.

30.    H-R. Liu, X-Q. Huang, D-H. Lou, X-J. Liu, W-K. Liu, and Q.-A. Wang, *Synthesis and acetylcholinesterase inhibitory activity of Mannich base derivatives flavokawain B.* Bioorg Med Chem Lett, 2014. **24**(19): p. 4749-4753.

31.    H-R. Liu, X-J. Liu, H-Q Fan, J-J Tang, X-H. Gao, and W.-K. Liu, *Design, synthesis and pharmacological evaluation of chalcone derivatives as acetylcholinesterase inhibitors.* Bioorg Med Chem, 2014. **22**(21): p. 6124-6133.

32.    H-R. Liu, C. Zhou, H-Q Fan, J-J. Tang, L-B. Liu, X-Hui. Gao, Q-A. Wang, and W.-K. Liu, *Novel Potent and Selective Acetylcholinesterase Inhibitors as Potential Drugs for the Treatment of Alzheimer's Disease: Synthesis, Pharmacological Evaluation, and Molecular Modeling of Amino-Alkyl-Substituted Fluoro-Chalcones Derivatives.* Chem Biol Drug Des, 2015. **86**(4): p. 517-522.

33.    S. D. Sukumaran, C. F. Chee, G. Viswanathan, M. J. C. Buckle, R. Othman, N. A. Rahman, and L.Y. Chung, *Synthesis, Biological Evaluation and Molecular Modelling of 21-Hydroxychalcones as Acetylcholinesterase Inhibitors.* Molecules, 2016. **21**(7): p. 1-10.

34.    F. Severi, S. Benvenuti, L. Costantino, G. Vampa, M. Melegari, and L. Antolini, *Synthesis and activity of a new series of chalcones as aldose reductase inhibitors.* Euro J Med Chem, 1998. **33**(11): p. 859-866.

35.    S. Iwata, N. Nagata, A. Omae, S. Yamaguchi, Y. Okada, S. Shibata, and T. Okuyama, *Inhibitory Effect of Chalcone Derivatives on Recombinant Human Aldose Reductase.* Bio Pharm Bull, 1999. **22**(3): p. 323-325.

36.    A. W Schüttelkopf and D.M.F.v. Aalten, *PRODRG: a tool for high-throughput crystallography of protein-ligand complexes.* Acta Cryst. D, 2004. **D60**: p. 1355-1363.

37.    Y. Duan, C. Wu, Shibasish Chowdhury, Mathew. Lee, Gouoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, C. James, Junmei Wang, and P. Kollman, *A point-charge force field for molecular mechanics simulations of protein based on condensed-phase quantum mechanical calculations.* J Comput Chem, 2003. **24**(16): p. 1999-2012.

# Chapter 4: Statistical analysis of the results of the high-throughput virtual screening

Research undertaken to ascertain the presence and identity of some common denominator which could explain the binding energies observed, was severely limited owing to the large number of structures present in the library which were not able to be studied individually. In order to complete this study, it was therefore decided that a statistical approach was appropriate, and this was conducted with a diverse set of tools.

## 1. Statistical analysis

### 1.1. Tools used for statistical analysis

#### 1.1.1. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical approach, which helps to reduce redundant data and build a clearer representation of the data as a whole. A PCA study is conducted with the primary aim to identify useful, correlated variables. Two-dimensional representations of two well-chosen principal components allows for a simple visualization of cluster presence within the data and may be orientated by variable parameters.[1] Using a ($n$ x $p$) matrix in which $p$ variables are observed $n$ times, the PCA converts these variables into new, uncorrelated variables. The conversion is from the covariance matrix *via* an Eigenvalue decomposition. These new variables are called principal components. The first principal component is the principal component defined by having the largest explained variance in total. This first principal component is used as the first orthogonal axis for the remaining principal components, which are ranked from the highest variance to the lowest one. A good contribution of the percentage of explained variance for the first principal component is around 80% to 90%[1] . The projection and the expression of the variance of the correlation of the variable (PCA loading diagram) describe the amount of correlation, if one exists. When two vectors, which represent two sets of variables, show similar patterns, a positive correlation between the variables is shown. By contrast, when these two PCA loading vectors are in opposite quadrants, the correlation is negative [1, 2]. In addition, in a PCA plot, a vector is also more significant in describing variation in the data when its distance from the origin is large. For example, Figure 4.1 shows an example of the loadings plot of the first two components of a PCA of the concentration of pollution particles in European cities. The

contribution of the concentration of ozone is more significant, seen by the size of the vector in comparison with the concentration of the other pollutants.[3] Furthermore, the angles between each of the vectors show the level of correlation between the vectors; the smaller the angle, the more highly correlated the variable. Thus, from Figure 4.1, the concentrations of $NO_2$ and $O_3$ are highly uncorrelated, whereas there is a good correlation between $PM_{10}$ and NO.
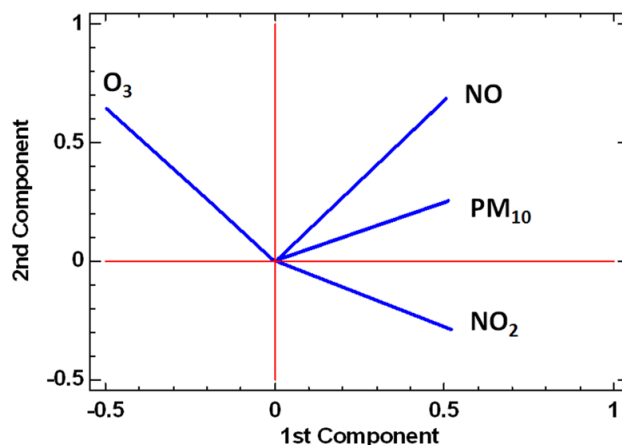


Figure 4.1: Loadings plot of a PCA of the concentration of the different pollutant particles in European cities. (with the permission of the publisher).[3]

In this study, PCA can be used to determine which of several parameters may be used to explain the variation observed in the binding energy of the different complexes. Even if only one variable changes for a different population at a time, the plotting with two principal components will consider all variables and identify all populations distincly.[1] PCA and factor analysis respectively give details about diagonal and off-diagonal elements which are respectively the variance and the covariance between variables.[1] RStudio is the software package used to perform PCA and draw representations from results in this study.

### 1.1.2. Cluster analysis: dendrogram

Further investigation of the relationship between all factors in binding, including the binding pose and interaction energy, may be performed using a hierarchical clustering analysis. This type of analysis may help to show the connection between the PCA and the meaning of the resulting major principal components from PCA. The aim of cluster analysis is to divide groups of variables into clusters using their similarities.[1] This discrimination method helps to identify parameters with poor to no information. Several clustering models are available depending on the purpose of the study. Examples of two options used to separate variables

into different clusters are as follows.[4] The first method incorporates agglomerative clustering. It uses a bottom-up approach to merge two or more objects, which possess the most similarity. The process continues with clusters merged into greater clusters, until all clusters are merged into one single cluster. The opposite method is a divisive (top-down) approach. From a large cluster containing all objects, two clusters are created such that they contain elements with similar properties. The process continues until all clusters are separated.

Cluster analysis has a metric, and this metric helps to measure the distance between two clusters. Three different metrics are proposed. The first one is the Euclidean distance which measures, geometrically, the distance between objects in multidimensional space. The second method uses the squared Euclidean distance.[5] Based on the Euclidean distance, now squared, this gives greater significance to larger distances.[6] The last metric presented is the Manhattan metric. It calculates the average distance between two points in all dimensions. This metric is useful to reduce the effect of external noise when the coordinates are not separated.[7]

The last parameter that may be used to build the dendrogram during cluster analysis is the applied algorithm used to separate two clusters. Five methods are proposed.[8] The complete linkage clustering calculates dissimilarities between values of cluster one and cluster two, takes the maximum of these values and then uses it as a distance between these two clusters. An opposite approach, called single linkage, is also possible. In this case, it uses the smallest value of dissimilarities between two clusters. The average clustering takes the average of the dissimilarities as the distance between the two clusters. The centroid linkage finds the dissimilarities between the centroid of the two clusters. The last method is the Ward's minimum variance. It uses the variance to calculate the distance between two clusters. RStudio may be used to draw the dendrograms with all of the parameters cited previously.

## 1.2. Preparation of data sets and code building

### 1.2.1. Research into suitable parameters from chemical/physical characteristics

The binding energies were collected from the high-throughput virtual screening which is described in the previous chapter. The chemical and physical characteristics of each chalcone molecule were calculated using the CDK node in Knime.[9, 10] The properties selected for the studies were chosen according to the Lipinski rules which include: Mannhold logP,

hydrogen bond acceptors, molecular weight. heavy atom count and topological polar surface area were also included. Chemical/physical properties were merged with the binding energy results to create tables with all available information. The processes and nodes used, are represented in Figure 4.2.
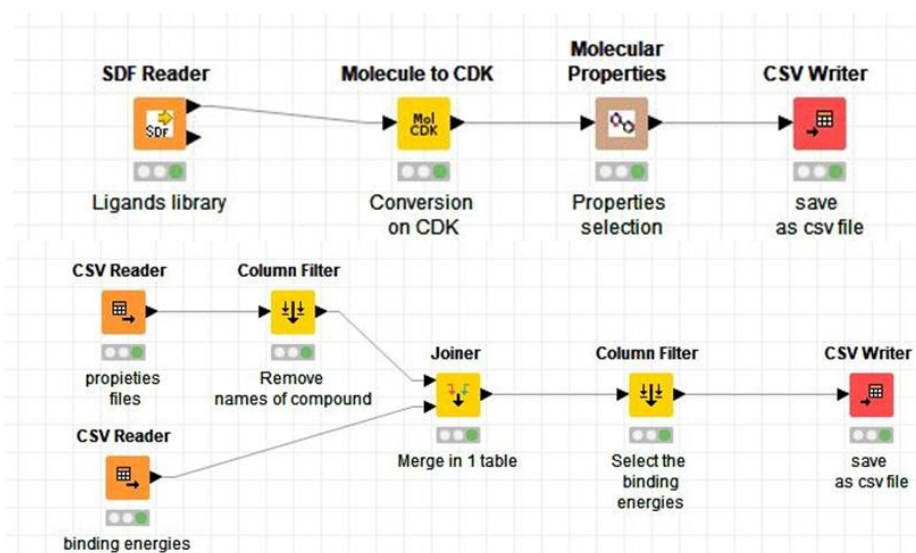


Figure 4.2: Nodes and processes of Knime used to get the properties of chalcone molecules and data manipulation. The name is provided from the binding energies CSV files.

### 1.2.2. Determination of the spherical polar coordinates of specific atoms in chalcone molecules

The search for a common denominator using only physical and chemical characteristics as parameters did not provide satisfactory results. As such the information was supplemented with information connecting the binding position and orientation with respect to both the acetophenone-derived, and the benzaldehyde-derived portions of the set of chalcones. In order to supplement with this information, a conversion of the Cartesian coordinates of a) the carbon adjacent to the ketone functionality and b) the carbon next to the $\alpha,\beta$ unsaturated bond to spherical polar coordinates was performed. The origin of the polar coordinate reference system is the active site centre used for the high-throughput virtual screening provided with Autodock Vina (Chapter 3). Figure 4.3 represents a chalcone molecule and highlights the carbon atoms selected for the studies. The choice of this system was made in order to address the limitation of the Cartesian coordinates, which can only give the absolute positions of the molecule in 3D space with no information regarding the position of the active site. The spherical polar coordinates not only provide the distance of parts of the molecule from the active site but also their relative positioning using polar and azimuthal angles. This coordinate

system fully allows for the situating and orientation of the molecules within the active site space.
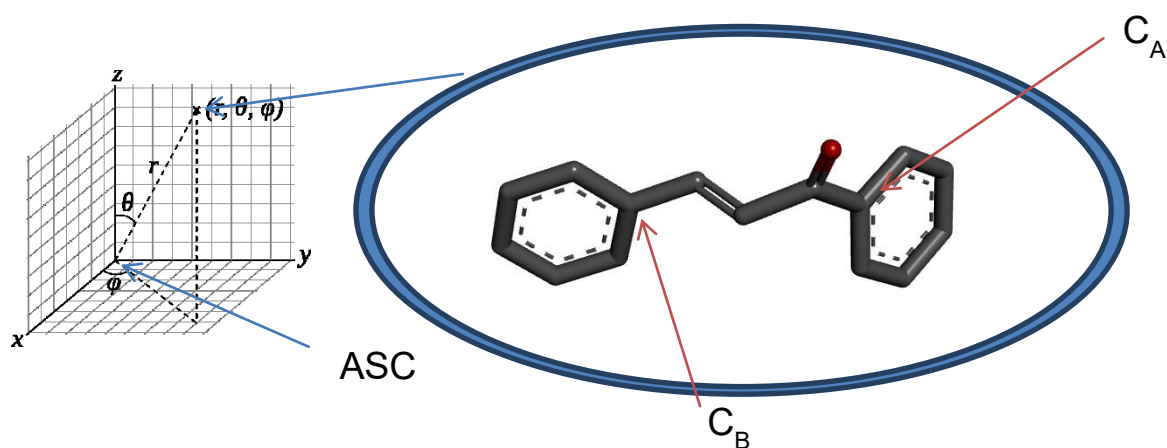
A Perl script was written in order to identify the carbons $C_A$ and $C_B$ close to the ketone moieties and to the unsaturated bond. Each set of Cartesian coordinates from the chalcone molecules were extracted. In order to automate the identification of key centres in the chalcone, a script, based on bond distances, was constructed for this purpose. The script calculated the vector between all oxygen atoms and all carbon atoms within the molecule. Since the only carbonyl functionality on any model was the α,β-unsaturated system, if the distance was lower than 1.29 Å, the functionality was the carbonyl and both the carbonyl carbon and oxygen were identified. The two carbon atoms, adjacent to the carbonyl carbon atom, were now examined and the number of carbons bound to it, counted. If two carbons were found linked to it, $C_A$ was successfully identified. If only one further attached carbon was found, then this carbon atom was identified as the α carbon of the α,β-unsaturated system. Using this atom, the second carbon, the β-atom of the alkene bond was easily located. $C_B$ was taken as the carbon atom adjacent to this β-atom in the alkene functional group. The coordinates of $C_A$ and $C_B$ were kept in memory. Using trigonometric formulae, these were then converted to spherical polar coordinates. The formulas used, are presented in Equations 4.1 to 4.3. All of the spherical polar coordinates were saved in one file per protein.

$$r = \sqrt{x^2 + y^2 + z^2} \quad (4.1)$$

$$\theta = \cos^{-1} \frac{z}{r} \quad (4.2)$$

$$\psi = \tan^{-1} \frac{y}{x} \quad (4.3)$$

These data were then merged with their appropriate binding energies according to the same process used in the last section with Knime.[9] The molecule name column was removed from the coordinate file and the binding energy column related to a specific protein was selected and retained. These two parts of the data were merged in one file and verified before being saved in CSV files. The Knime process used is presented in Figure 4.4 below.
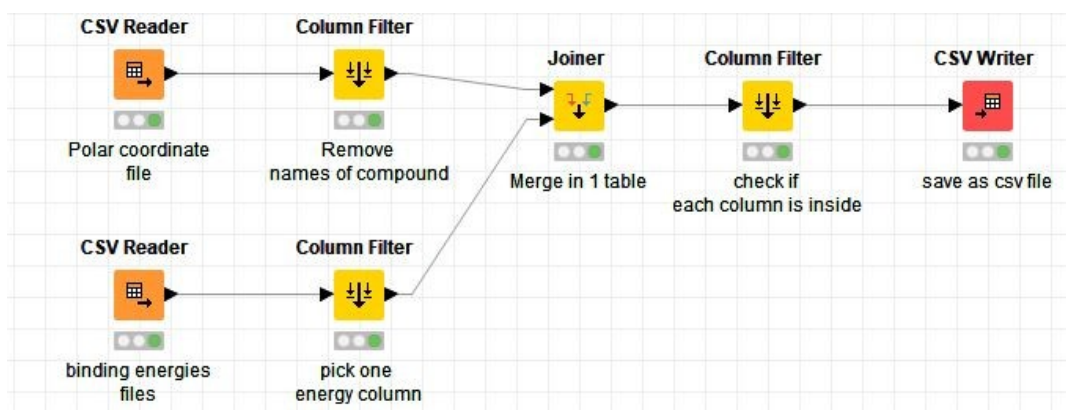


Figure 4.4: Knime process used to obtain the properties of the chalcone molecules and data manipulation.

### 1.2.3. Correlation, PCA and cluster analysis codes in RStudio

The binding energies per protein were merged into a single file (Table 4.1). Initially, Excel was used for the correlation analysis, and the default commands within RStudio were used for both the principal component analysis and the hierarchical cluster analysis. Later, however, data was analyzed using the more powerful package FactomineR in RStudio.[11, 12] The correlation of the chemical/physical characteristics (or the polar spherical coordinates) and the binding energy utilized FactomineR was done in order to prepare data for PCA and cluster analysis.

| row ID | Ar | Atheta | Apsi | Br | Btheta | Bpsi | sLAP2 binding energy (kcal.mol-1) |
|---|---|---|---|---|---|---|---|
| acetophenone_1_benzaldehyde_1 | 5.32028044373602 | 2.09697103770699 | -0.811837265586008 | 9.28209706908951 | 2.07952425455229 | -0.391593362395018 | -7.9 |
| acetophenone_1_benzaldehyde_10 | 9.28914883075947 | 2.06022127475246 | -0.30142234922622 | 5.55146404113365 | 2.15806342725071 | -0.836678430364751 | -7.8 |
| acetophenone_1_benzaldehyde_11 | 4.29851788410843 | 2.13483408588502 | -1.09188272824731 | 1.72553556903357 | 2.34791981411153 | -1.22410879821559 | -6.9 |
| acetophenone_1_benzaldehyde_12 | 10.2117176322106 | 2.13817325988986 | -0.283882361442274 | 6.6099189858878 | 2.19235949991715 | -0.671195270032588 | -7.5 |
| acetophenone_1_benzaldehyde_13 | 9.2980191976571 | 2.12130658936227 | -0.352522967088773 | 5.44989476962629 | 2.11422551100817 | -0.784337075422174 | -8.1 |
| acetophenone_1_benzaldehyde_14 | 9.2661598302641 | 2.06080920624736 | -0.301907130472293 | 5.53999720216536 | 2.15445824194266 | -0.839878511859403 | -7.6 |
| acetophenone_1_benzaldehyde_15 | 9.29952665462065 | 2.13057040168669 | -0.338124111093203 | 5.24232209998584 | 2.13267758979536 | -0.750006249397747 | -7.8 |
| acetophenone_1_benzaldehyde_16 | 10.11246626694 | 2.15140250079352 | -0.365110475420452 | 6.34605018889703 | 2.19151708770951 | -0.773343952752488 | -8.2 |
| acetophenone_1_benzaldehyde_17 | 2.38886772342045 | 2.3349962547413 | -1.22410597016396 | 6.87744640109976 | 2.09847591548506 | -0.601819294995548 | -9.1 |
| acetophenone_1_benzaldehyde_18 | 9.57864760809166 | 2.07553140711808 | -0.312170280214921 | 6.00661651847361 | 2.16102729170843 | -0.844802990881265 | -7.0 |
| acetophenone_1_benzaldehyde_19 | 10.153519389847 | 2.11122241650071 | -0.308311479763316 | 6.28676275995842 | 2.20655399131074 | -0.648680565877028 | -7.1 |
| acetophenone_1_benzaldehyde_2 | 10.1943274913061 | 2.15200054384207 | -0.367342359887607 | 6.3970200875095 | 2.20200196190777 | -0.749928732433864 | -8.9 |
| acetophenone_1_benzaldehyde_20 | 9.24074358479879 | 2.12397046252422 | -0.302323415941194 | 4.99518097770241 | 2.20902559555654 | -0.591182682699685 | -7.5 |
| acetophenone_1_benzaldehyde_21 | 4.09149862519835 | 2.15922342301857 | -1.10057851010192 | 1.93164230643253 | 2.26310616591881 | -0.948279625558322 | -7.2 |

Table 4.1: Example of CSV files of the polar spherical coordinates with binding energy of the sLAP$_2$ complex. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

## 1.2.4. Preliminary analysis in RStudio

CSV files of the complete data were imported into RStudio in order to perform cluster analysis and PCA calculations. An R script was written to calculate the PCA components and to produce the graphs corresponding to the analysis. We used a ***prcomp*** default principal component analysis command in an attempt to perform the PCA and draw up the results. Figure 4.5 presents the result of this simple analysis. However, the large amount of individual values obscures the biplot of the loadings and scores plots. In addition, the quality and the contribution of the different values were not easily deduced. From these unsatisfactory representations of PCA, it was decided to follow a more involved method using specific packages.
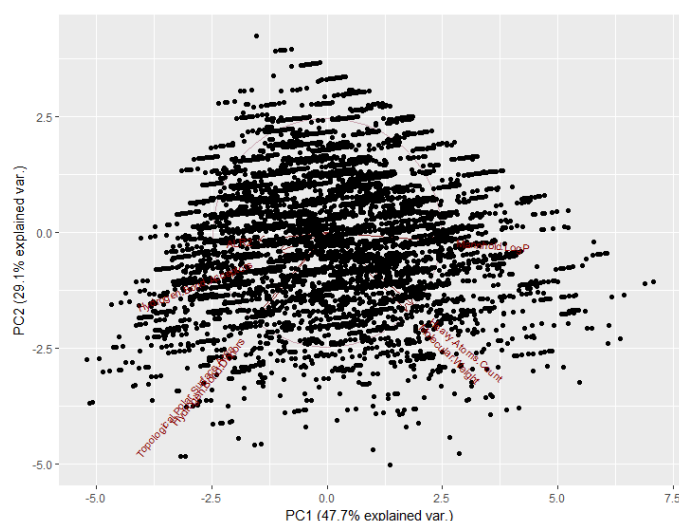


Figure 4 5: Attempt to draw properties and binding energy of ALR$_2$ PCA including loadings and scores plots at the same time.

As a result the package FactomineR, as implemented in RStudio, was used for the PCA calculation and Factoextra was used to visualize the resultant data [11]. Clustersim was also used in order to perform normalization of values. After importing the data, modification of the data table was performed to remove columns containing names and thus exclude non-numeric data from the statistical analysis. Next, the numeric values were normalized following the standardization method (x-mean/standard deviation). The normalized data was submitted to the PCA process, and the results were used in a plot of the number of dimensions versus the percentage of variance explained. Successful PCA calculations and visualizations were thus effected.

The hierarchical cluster analysis was also undertaken using RStudio and the R library Cluster was installed within R and included in the scripting. It was decided to test the agglomerative and divisive hierarchical clustering processes on the data before a final decision was implemented in the analysis of data. Figure 4.6 shows the agglomerative and the divisive dendrogram of the binding energy of $ALR_2$ and the chemical/physical properties of the chalcones binding to $ALR_2$.
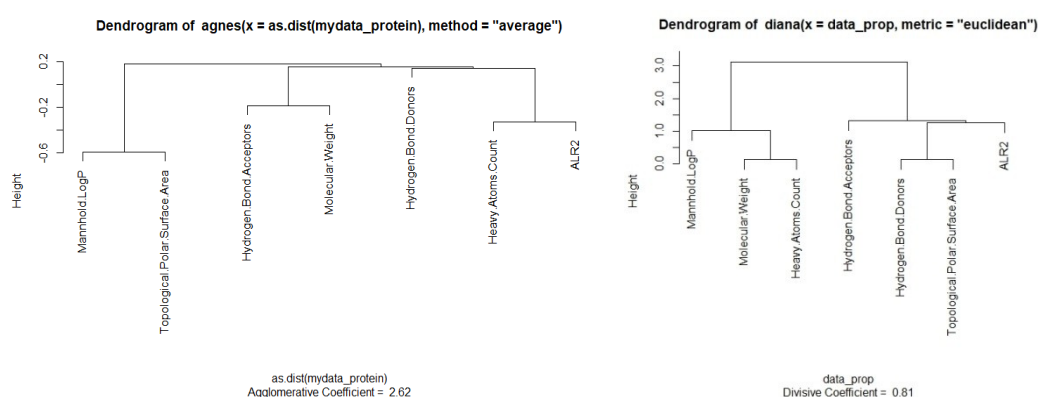


Figure 4.6: Agglomerative hierarchical clustering with average metric (*left*) and divisive hierarchical clustering with Euclidean metric (*right*) of the binding energy of the $ALR_2$ protein and the chemical/physical proprieties of the chalcones.

The reason for the choice of the divisive hierarchical clustering is that we wanted to exclude the possibility of orienting the analysis by using one cluster for the base of the dendrogram. A divisive coefficient was found to be better than the agglomerative one (since the divisive coefficient is supposed to be close to unity). The Euclidean method was used to calculate the distance between different clusters for agglomerative and divisive hierarchical clustering because this is provided as default in RStudio. The method option was only implemented in agglomerative hierarchical clustering. It was decided to use the average method to plot the dendrogram as compromise between the complete and the single linkage methods.

These methodologies were applied for the binding energy of each complex, chemical/physical (molecular weight, heavy atom count, Mannhold LogP, topological polar surface area, hydrogen bond donor and acceptor) characteristics of the chalcones and the spherical polar coordinates of the carbons next to the ketone function and the conjugated bond. Figure 4.7 represents the code which was used in RStudio for the PCA and for the cluster analysis.

```
################# INSTALLING PACKAGE FOR PCA  #################################################
install.packages("FactoMineR")
install.packages("factoextra")
install.packages("devtools")
install.packages("clusterSim")

################ Activate the library #########################################################
library(FactoMineR)
library(factoextra)
library(cluster)
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
library(clusterSim)

################ Read the table from previously ###############################################

setwd("C:/path_Results_stats")
donne <- read.csv("table_for_R_binding_energy.csv",sep = ";", dec = ",", header = T)
donne_PK_C <- read.csv("PK_C_Results_coord.csv",sep = ",", dec = ".", header = T)
################ Prepare data for protein analysis ############################################
donne_protein <- donne[,-1]
ligand_name <- donne[,1]
donne_PK_C <- donne_PK_C[,-1]

################ Correlation analisys for protein #############################################
mydata_protein <-cor(donne_protein)
mydata_PK_C <-cor(donne_PK_C)

################ Normalization of the data ####################################################
donne_PK_C_NORM <- data.Normalization(donne_PK_C,   type = "n1",   normalization = "column") #####   n1   corresponding   to
normalization cf help
donne_prot_NORM<- data.Normalization(donne_protein, type = "n1", normalization = "column")

################# Score plot, indiv PCA and var in test #######################################
PK_C_NORM.pca <- prcomp(donne_PK_C_NORM, scale. = FALSE)
fviz_eig(PK_C_NORM.pca)
fviz_pca_ind(PK_C_NORM.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE, label ="none")
fviz_pca_var(PK_C_NORM.pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))

Prot_NORM.pca <- prcomp(donne_prot_NORM, scale.= FALSE)
fviz_eig(Prot_NORM.pca)
fviz_pca_ind(Prot_NORM.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE, label ="none")
fviz_pca_var(Prot_NORM.pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))

################# Cluster analysis for specific data ##########################################
plot(diana(mydata_protein,metric = "euclidean"))
plot(diana(mydata_PK_C,metric = "euclidean"))
```

Figure 4.7: Example of code built to use the different statistical analysis tools and draw their graphs.

## 1.3. Results

### 1.3.1. Correlation between the binding energy and targets across the full dataset

Given that chalcones are known to bind to each of the targets, questions arise firstly as to whether there are individual ligands that show selectivity to specific targets, and secondly, how the library as a whole is (or is not) selective towards specific targets. To this end, the binding energy results from the MAOA protein (with consistently high binding energies indicative of poor interaction) were of particular interest. A simple correlation analysis within Excel was chosen as a means to verify differences in performance of the chalcone library

against the various targets, with the expectation that the MAOA results would be also highlighted by this analysis. Table 4.2 presents the correlation results of the set of binding energies over all crystal structure targets. The correlation results from MAOA indeed showed that it was independent from all other proteins with a weak correlation score. However, although this is interesting, a more useful indication would be the use of proteins with weak correlations to other targets, but with strong binding to the particular protein. For MAOA, the range of the coefficients is close to zero or slightly negative. The second target presenting some independence was the COX-2 co-crystal structure B with a correlation score ranging from 0.10 to 0.36. Again COX-2 tended to have poor binding to ligands in the library, albeit with some exceptions with acetophenone_19_benzaldehyde_89 (-12.2 kcal.mol$^{-1}$) with COX-2_A model or with acetophenone_15_benzaldehyde_45 (-10.5 kcal.mol$^{-1}$) with COX-2 B structure, for example. Correlation scores for the remaining targets appeared to be much higher. The binding energies of the acetylcholine esterase structure A (AChE_A) were very good (low in value) but the correlation of the binding energies was quite high with other targets. It would seem that chalcones that are good binders across the range of targets bind well to AChE_A, casting doubts on specificity within the library.

| | ALR2 | AChE_A | AchE_B | BchE_A | BchE_B | HSP90 | MAOA | MAOB_A | MAOB_B | PK_AB | PK_C | cox_1_A | cox_1_B | cox_2_A | cox_2_B | integrase | sLAP2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALR2 | 1 | | | | | | | | | | | | | | | | |
| AchE_A | 0,534102164 | 1 | | | | | | | | | | | | | | | |
| AchE_B | 0,537553863 | 0,985220419 | 1 | | | | | | | | | | | | | | |
| BchE_A | 0,655803785 | 0,767715122 | 0,769352917 | 1 | | | | | | | | | | | | | |
| BchE_B | 0,656056594 | 0,765874301 | 0,767624606 | 0,97376134 | 1 | | | | | | | | | | | | |
| HSP90 | 0,705548848 | 0,75481322 | 0,760131668 | 0,831884705 | 0,831427273 | 1 | | | | | | | | | | | |
| MAOA | -0,129709496 | 0,0081029 | -0,005361002 | -0,19831422 | -0,208429902 | -0,163932005 | 1 | | | | | | | | | | |
| MAOB_A | 0,425463371 | 0,77945088 | 0,773209438 | 0,684386957 | 0,68271437 | 0,644727487 | 0,176093907 | 1 | | | | | | | | | |
| MAOB_B | 0,55123502 | 0,823215138 | 0,819966925 | 0,759203302 | 0,759654619 | 0,740633468 | 0,025687964 | 0,870966055 | 1 | | | | | | | | |
| PK_AB | 0,624014984 | 0,803351317 | 0,803043124 | 0,853227472 | 0,854540381 | 0,822571613 | -0,200192906 | 0,714600964 | 0,79161038 | 1 | | | | | | | |
| PK_C | 0,483137875 | 0,718623265 | 0,71908667 | 0,706708436 | 0,706254917 | 0,679686885 | 0,037209963 | 0,698297402 | 0,707889562 | 0,766723215 | 1 | | | | | | |
| cox_1_A | 0,409958546 | 0,638501073 | 0,637088484 | 0,484664688 | 0,498038253 | 0,529547584 | 0,202443289 | 0,605124695 | 0,608534441 | 0,576417212 | 0,596889539 | 1 | | | | | |
| cox_1_B | 0,347267804 | 0,610176291 | 0,608355275 | 0,397234451 | 0,411541538 | 0,493914492 | 0,216140359 | 0,589850756 | 0,579165893 | 0,485077455 | 0,521081448 | 0,777569391 | 1 | | | | |
| cox_2_A | 0,563285515 | 0,680263495 | 0,67844804 | 0,596543051 | 0,596403159 | 0,681185119 | 0,010717895 | 0,594117512 | 0,634461318 | 0,603795903 | 0,556507958 | 0,6284282 | 0,681347853 | 1 | | | |
| cox_2_B | 0,258580204 | 0,178507889 | 0,178290494 | 0,146330548 | 0,17306819 | 0,201458655 | 0,101232593 | 0,178319339 | 0,146600935 | 0,11652568 | 0,128213857 | 0,222666064 | 0,252676466 | 0,360381638 | 1 | | |
| integrase | 0,539092287 | 0,84398888 | 0,846923634 | 0,768594858 | 0,76658521 | 0,727499389 | -0,073392756 | 0,742997585 | 0,792698441 | 0,823746356 | 0,732966564 | 0,63944249 | 0,576352774 | 0,662534777 | 0,146546456 | 1 | |
| sLAP2 | 0,499098757 | 0,839070188 | 0,835759526 | 0,748072814 | 0,749513828 | 0,742577335 | -0,012385175 | 0,747749593 | 0,792270709 | 0,826004303 | 0,743637245 | 0,716682264 | 0,641087079 | 0,654124013 | 0,118147716 | 0,80311064 | 1 |

Table 4.2: Correlation results of binding energy per protein.
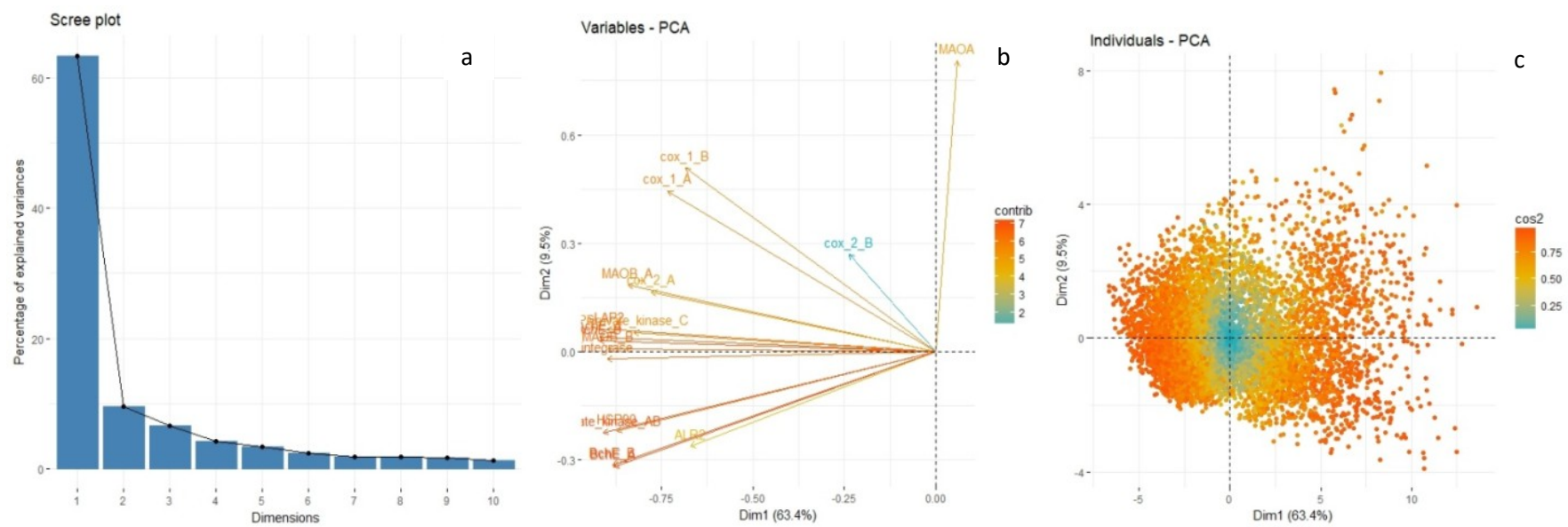
Figure 4.8: Scree plot, showing the percentage of variance explained for the PCA of binding energies per protein (a), loadings plot of the binding energy of each protein structure (b) and scores plot of the binding energy per protein (c).

Given the independence of binding to COX-2-B, and given that some ligands had good binding to COX-2, a PCA with a loadings plot was performed in order to investigate the possibility of statistical independence of binding to the COX-2 co crystal B structures. Figure 8 shows the scree plot (a graph of the percentage of variance explained by each principal component, Figure 4.8(a), the loadings plot for the first two principal components (each variable is represented by a line; the length of the line represents the importance of the variable, and the angle between the lines, a measure of the correlation between variables, Figure 4.8(b) and the scores plots for the first two principal components (each point represents a binding energy)with the individual map of the complex binding energies of each protein (Figure 4.8(c)). The first and second principal components explain 63.4% and 9.5% of the variance.  In the loadings plot, MOAO is in a quadrant of its own, and at a larger angle from all the other variables in the loadings plot and this confirms the independence of the MAOA results in relation to the cluster of the rest of the protein structures. The COX-2 co-structure B is part of the main cluster but does not contribute actively following the PCA per variable. The rest of the protein vectors are very close to each other and show a large correlation. However, the score plot of the individual points PCA, Figure 4.8(c), presents one large cluster from which it is impossible to extract any individual variables. It also presents a central point on the score plot, Figure 4.8(c), which indicates a low variable representation quality. This representation quality is indicated using the squared cosine values. These are used to provide a measure of interpretation error, which arises as a result of projection effects. A squared cosine value of close to 1 (represented as a redder colour) shows a high fidelity of the data representation. Conversely, a low variable representation quality is indicated by a value of 0. The loadings plot also shows that   COX-2-B has a lower contribution in comparison to other variables (Figure 4.8(b)).

In order to supplement these results agglomerative hierarchical cluster analysis with an average metric of the binding energy over the same full target set was performed. Figure 4.9 shows the agglomerative dendrogram of the binding energy per protein following the average metric. The distances between each protein are fairly similar with the exception of $ALR_2$ BChE_B, MAOA, pyruvate kinase AB and COX-2-B. Furthermore, with a value of 0.41, the agglomerative coefficient of the hierarchical cluster analysis is too weak to represent a good separation (close to 1 is considered good) [13]. Therefore, data were not suitable for performing a cluster analysis; however, these results are not inconsistent with the PCA and correlation analysis.

.



**Dendrogram of agnes(x = as.dist(mydata_protein), method = "average")**

as.dist(mydata_protein)
Agglomerative Coefficient = 0.41
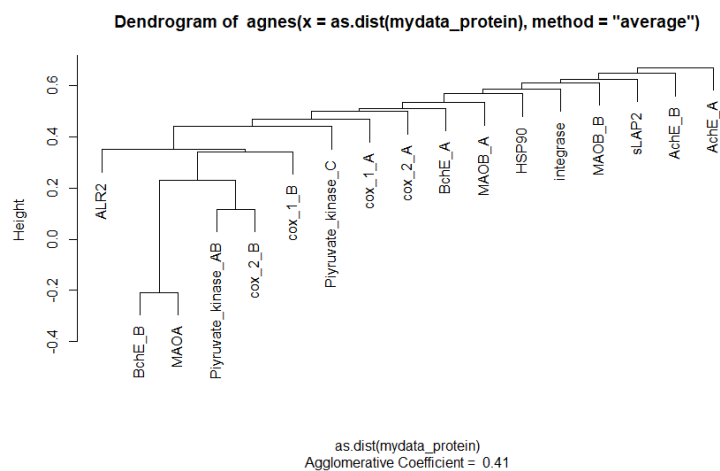
Figure 4.9: Agglomerative hierarchical cluster analysis of the binding energy per protein following the average metric.

The second attempt to analyse the full set of binding data produced by the high-throughput virtual screening (described in the previous chapters) was to investigate the binding energy per ligand instead of per protein. The aim was to identify particular ligands and their performance across the set of targets. In this case, the process of plotting an agglomerative dendrogram (Figure 4.10) was much faster than, for example, plotting an individual PCA map with all the labels. However, due to the large scale of data, we encountered the same problem as before when drawing a PCA with labels: it was unreadable. Figure 4.10 represents the agglomerative hierarchical dendrogram following the average metric of the binding energy per ligand. Other means to examine individual ligand performance were therefore explored.
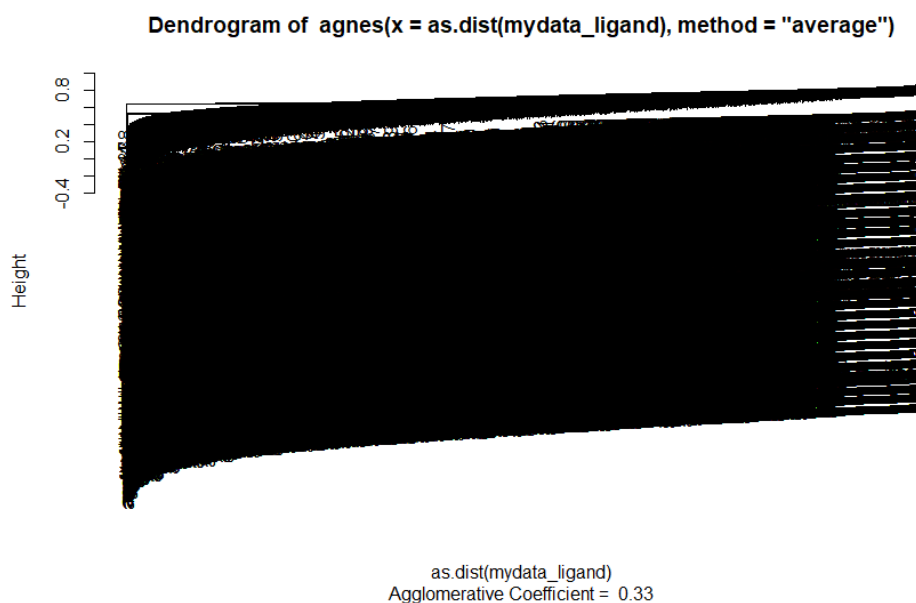
Figure 4.10: Agglomerative hierarchical dendrogram of the binding energy per ligands with the average metric.

### 1.3.2. Identification of a common denominator from chemical/physical parameters

After these attempts to analyze the data for the binding energy across the full set of targets, it was decided to separate the analysis for each target, and to include more ligand data in order to better understand the ligand properties in relation to binding energy. To this purpose physico-chemical properties for each ligand were calculated and incorporated into the dataset for each target. The selection of these physico-chemical properties was done according to the five Lipinski rules (described in the Chapter 2).

The degree of success or failure of this approach may be exemplified by the PCA calculations and clustering of ligand properties and binding to the target AChE structure A. Figure 4.11 represents the scree plot (a), the loadings plot (b) and the scores (binding energies) plot (c) of the PCA for the binding energy of the AChE structure A and the chemical/physical parameters of the chalcone library, while Figure 4.12 is the dendrogram produced in clustering this data. In terms of PCA calculations, the percentage of variance explained in the first dimension is 52.8% and the second dimension is 29.2% (Figure 11(a)). The sum of the percentage variance of the three first dimensions may explained at over 95%. The first two dimensions were used for plotting the loading graph (Figure 4.11(b)) and the score plot (Figure 4.11(c)). The loadings plot shows two clusters. The hydrogen bond donor, acceptor and the polar surface area parameters are incorporated in the first cluster (and it is expected that the presence of hydrogen bond donors and acceptors is related in some way to the polar

surface area), though the contribution of the hydrogen bond acceptor values to the PCA was weak. The second cluster included the molecular weight and the heavy atom count. Again, we expected these two sets of values to be closely related to each other. With the exception of the hydrogen bond acceptor parameter, the rest of variables presented good contributions to the variance within the data set. Upon examination of the scores plot, no true cluster was observed. It may be that the inclusion of a third dimension in the loading and score plot could have revealed any cluster presents in the data. Furthermore, the dendrogram (divisive clustering approach following Euclidean metric distance shown in Figure 4.12), confirms this analysis and shows a short distance between the first cluster (topological surface area, hydrogen bond donor and acceptor) and the binding energy. For this target, not only are these physico-chemical properties important in terms of drug-likeness (polar surface area, hydrogen bond donor/acceptor), but it is clear that these groups influence the binding in the receptor. The second isolated cluster on the dendrogram (Figure 4.12) appeared to be linked to the Mannhold logP. The analysis for the other targets follows these observations closely and their PCA maps are shown from Figure 4.13 to 4.29. One major difference is observed with respect to the targets ALR2, COX-1 (both A and B structures), COX-2 (both A and B structures) and MAOA (Figures 4.17-21 and 4.24). For these seven targets the PCA loading plots show a markedly different form, which has a tendency to be negative in the second dimension. It may be in these cases that the active site is more complex in terms of hydrophobic and polar requirements, and that the physico-chemical properties are very loosely affecting binding to fulfil these requirements.
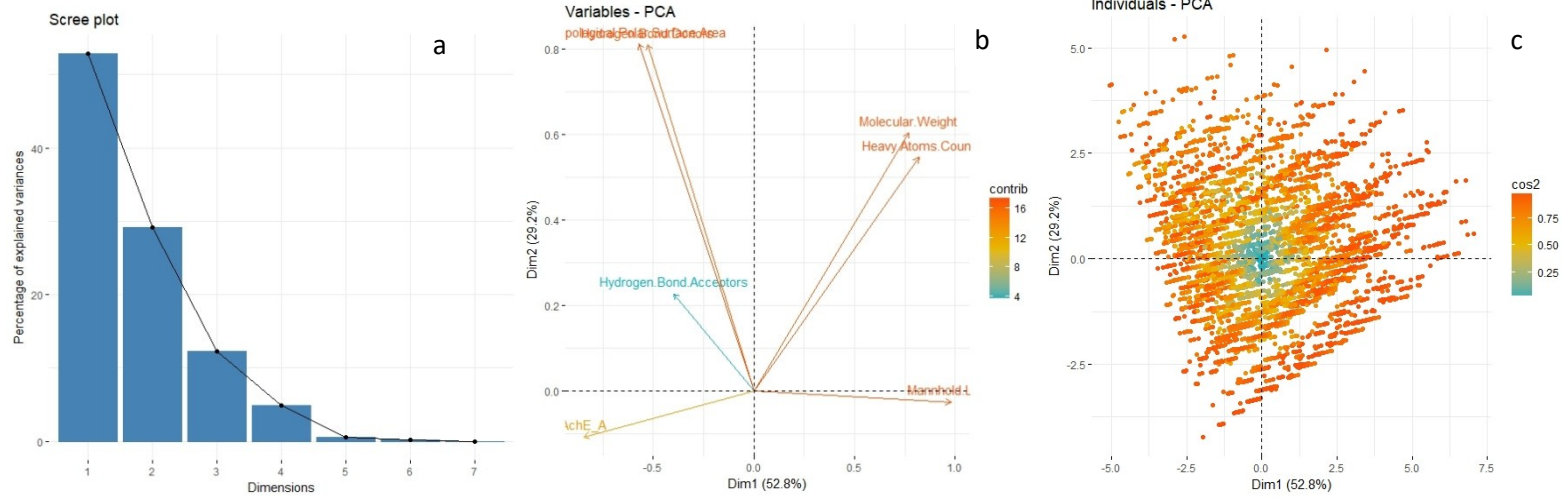
Figure 4.11: Scree plot of the explained variance for the PCA (a), loadings plot of the PCA (b) and scores plot (c) of chemical/physical properties of the chalcone and binding energies from the AChE structure A.
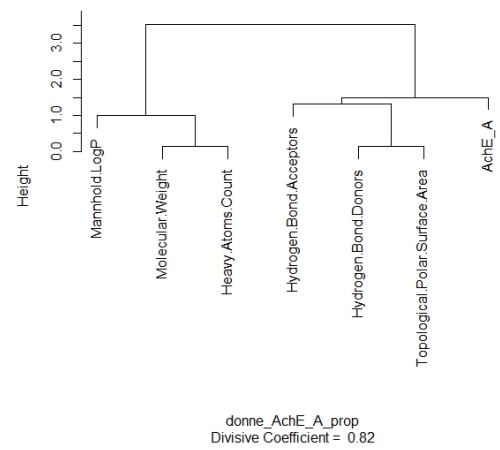
Figure 4.12: Dendrogram of divisive hierarchal clustering of the binding energies of the protein AChE_A and chemical/physical proprieties of the chalcones library following the Euclidean metric.

Figure 4.13: Scree plot, showing the percentage of variance explained for the PCA of binding energies of AChE_A (a), loadings plot of the binding energy of AChE_A structures (b) and scores plot of the binding energy (c).
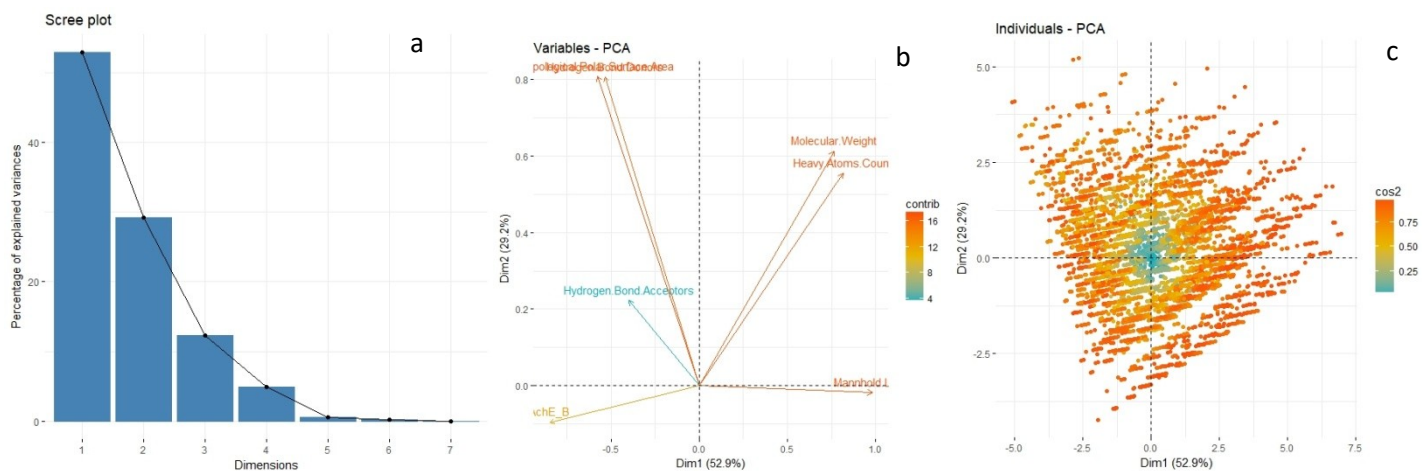


Figure 4.14: Scree plot, showing the percentage of variance explained for the PCA of binding energies AChE_B (a), loadings plot of the binding energy of AChE_B structures (b) and scores plot of the binding energy (c).
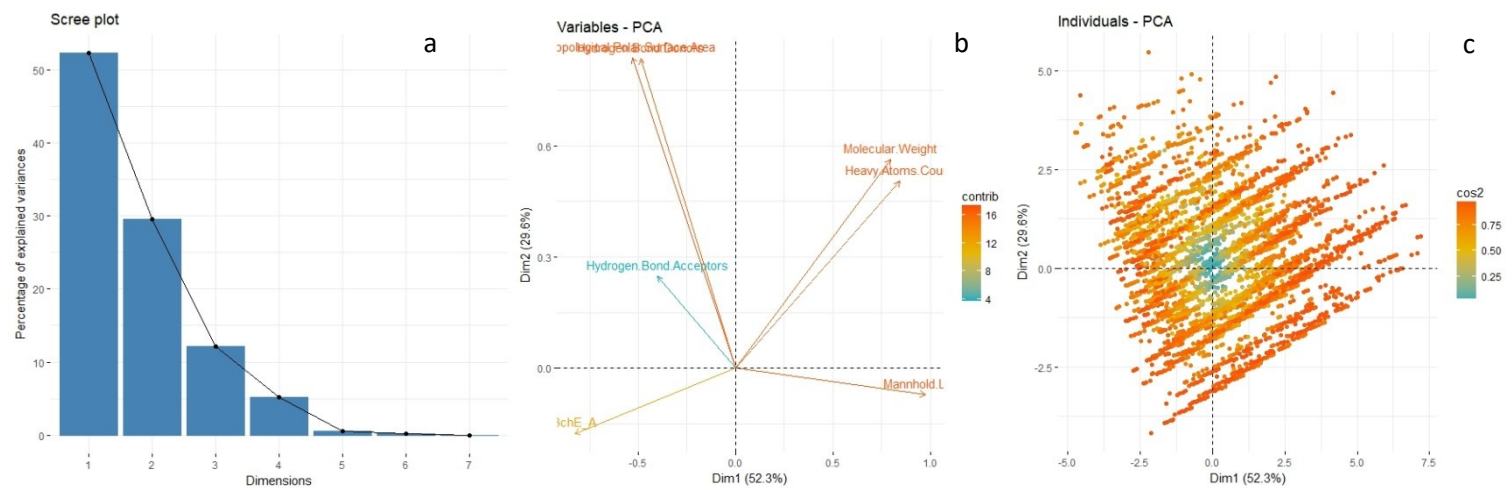
Figure 4.15: Scree plot, showing the percentage of variance explained for the PCA of binding energies of BChE_A (a), loadings plot of the binding energy of BChE_A structures (b) and scores plot of the binding energy (c).
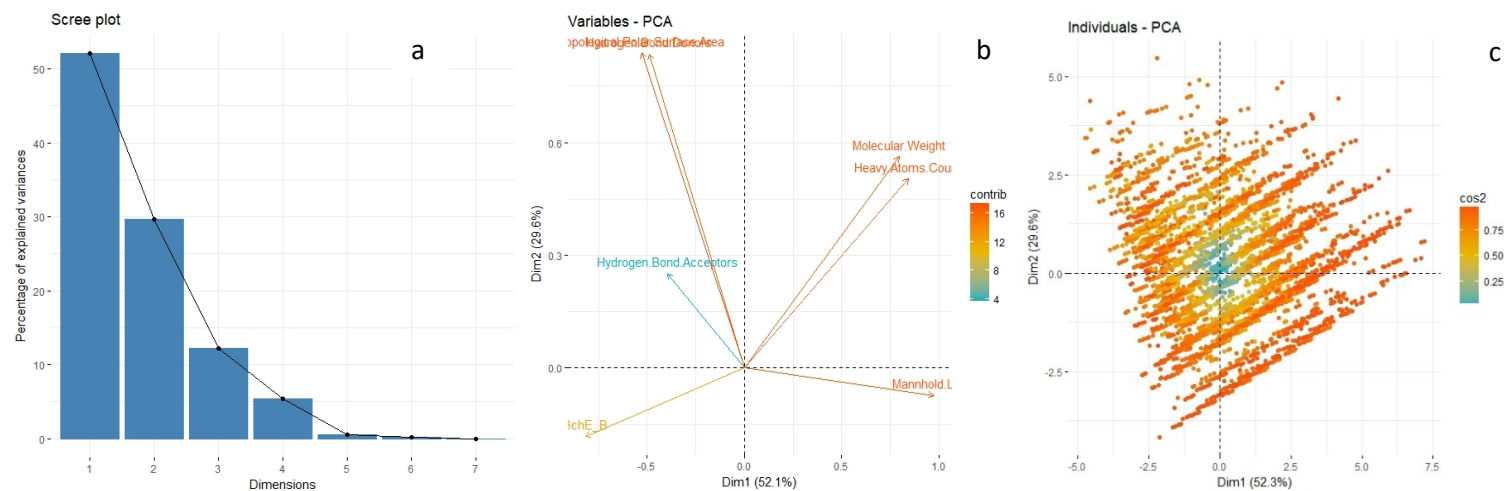
Figure 4.16: Scree plot, showing the percentage of variance explained for the PCA of binding energies of BChE_B (a), loadings plot of the binding energy of BChE_B structures (b) and scores plot of the binding energy (c).
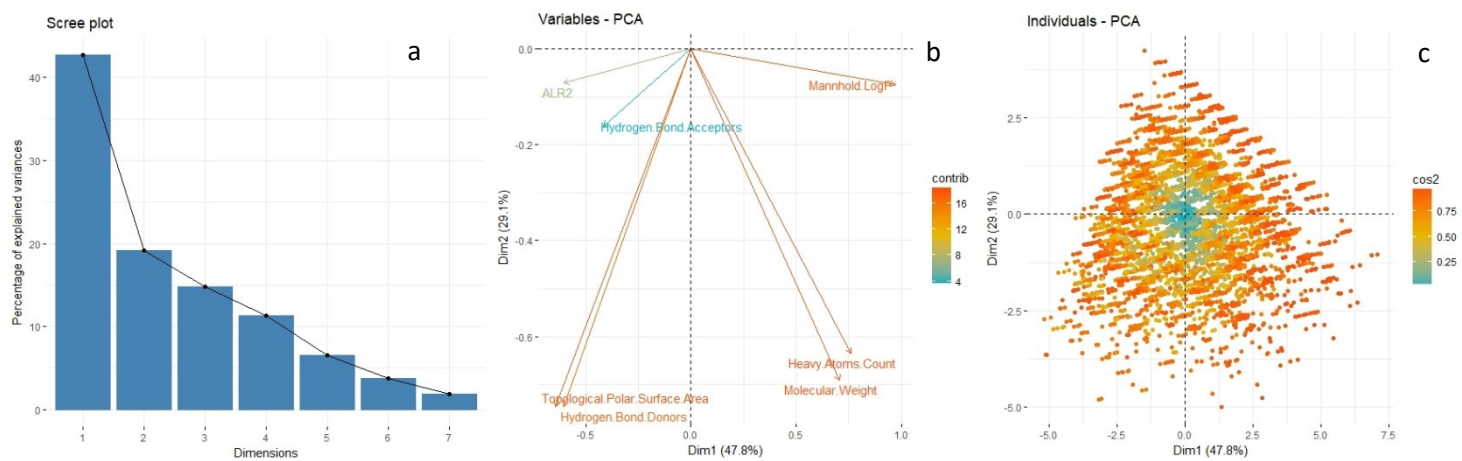


Figure 4.17: Scree plot, showing the percentage of variance explained for the PCA of binding energies of $ALR_2$ (a), loadings plot of the binding energy of $ALR_2$ structures (b) and scores plot of the binding energy (c).
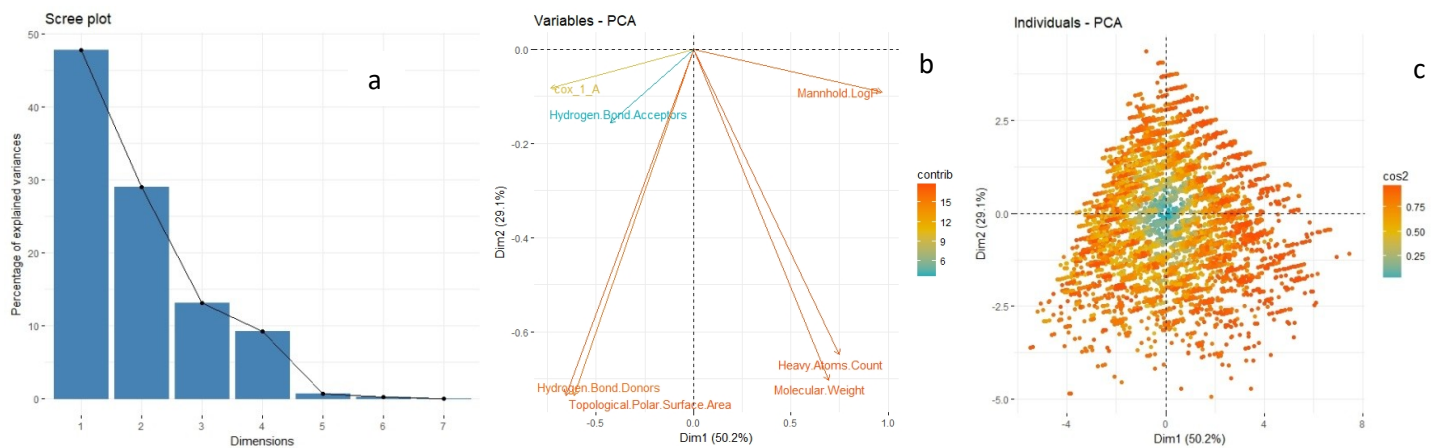
Figure 4.18: Scree plot, showing the percentage of variance explained for the PCA of binding energies of COX-1_A (a), loadings plot of the binding energy of COX-1_A structures (b) and scores plot of the binding energy (c).
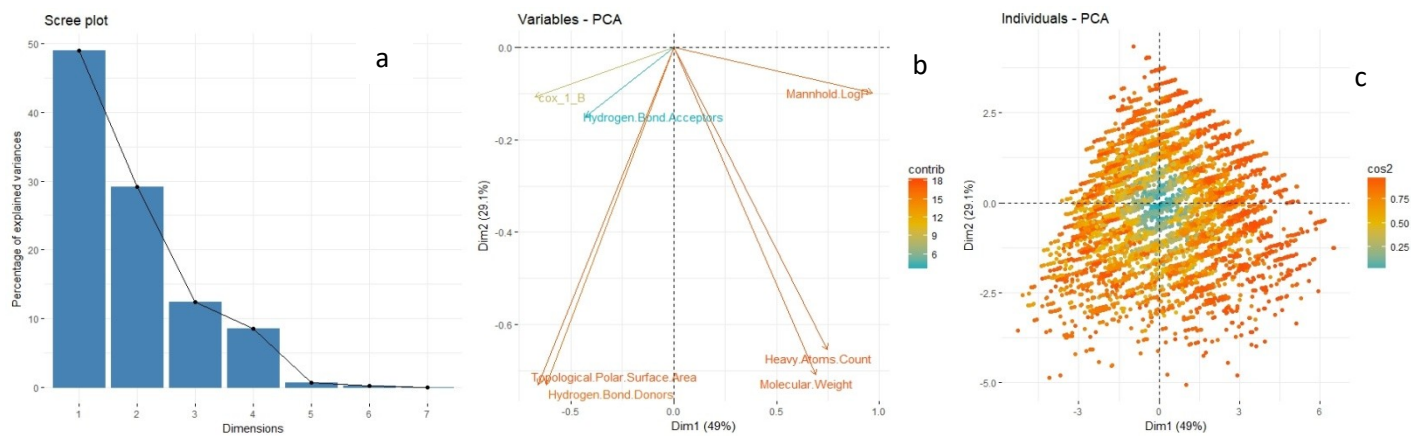


Figure 4.19: Scree plot, showing the percentage of variance explained for the PCA of binding energies of COX-1_B (a), loadings plot of the binding energy of COX-1_B structures (b) and scores plot of the binding energy (c).
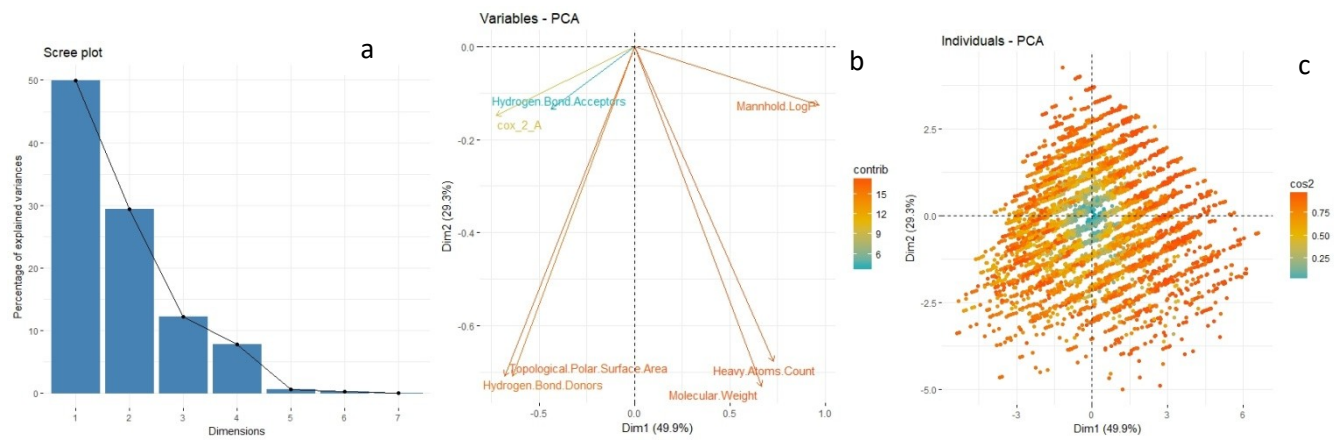
Figure 4.20: Scree plot, showing the percentage of variance explained for the PCA of binding energies of COX-2_A (a), loadings plot of the binding energy of COX-2_A structures (b) and scores plot of the binding energy (c).
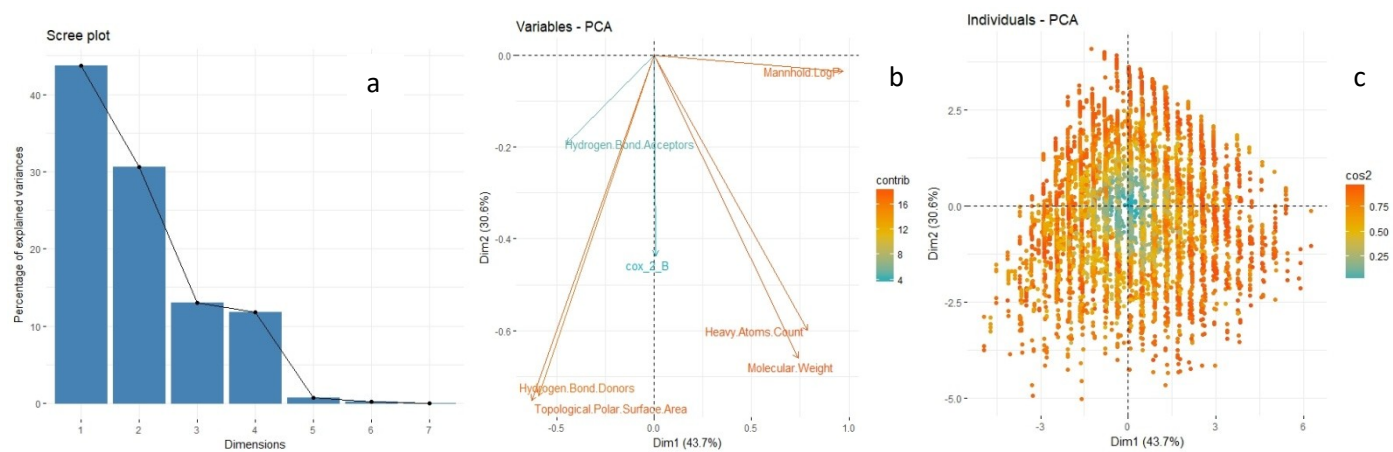


Figure 4.21 Scree plot, showing the percentage of variance explained for the PCA of binding energies of COX-2_B (a), loadings plot of the binding energy of COX-2_B structures (b) and scores plot of the binding energy (c).
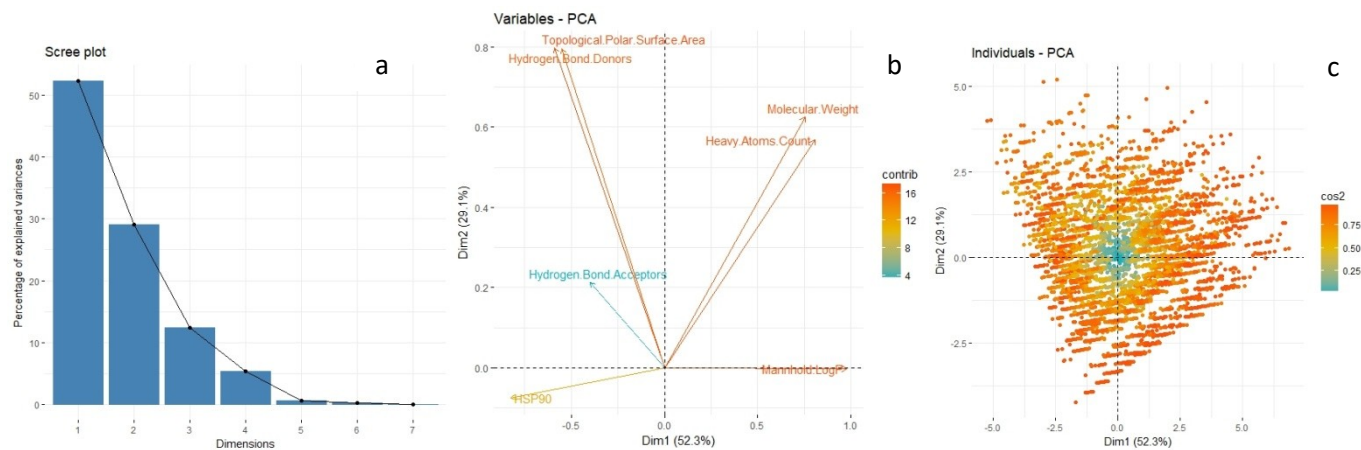
Figure 4.22 Scree plot, showing the percentage of variance explained for the PCA of binding energies of HSP90 (a), loadings plot of the binding energy of HSP90 structures (b) and scores plot of the binding energy (c).
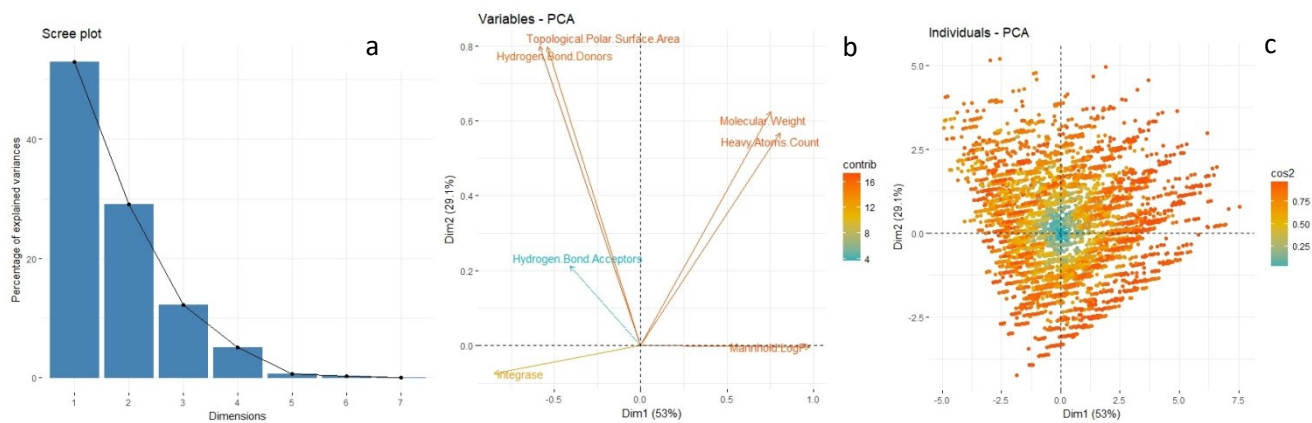


Figure 4.23 Scree plot, showing the percentage of variance explained for the PCA of binding energies of integrase (a), loadings score of the binding energy of integrase structures (b) and scores plot of the binding energy (c).

146
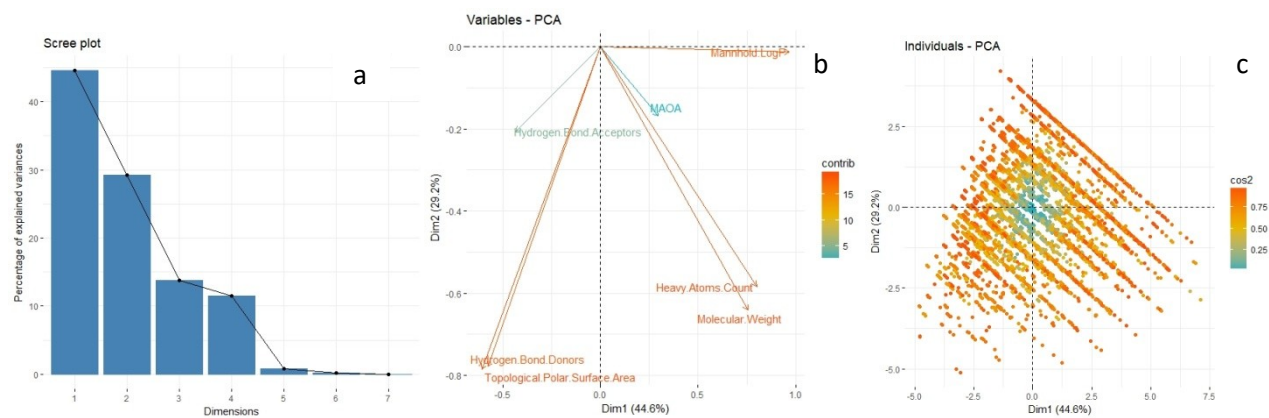
Figure 4.24 Scree plot, showing the percentage of variance explained for the PCA of binding energies of MAOA (a), loadings plot of the binding energy of MAOA structures (b) and scores plot of the binding energy (c).
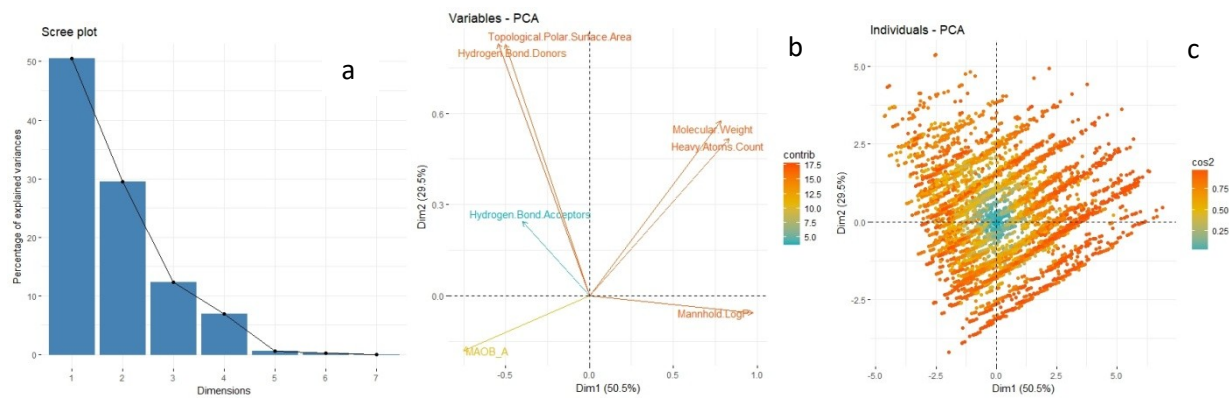


Figure 4.25: Scree plot, showing the percentage of variance explained for the PCA of binding energies of MAOB_A (a), loadings plot of the binding energy of MAOB_A structures (b) and scores plot of the binding energy (c).
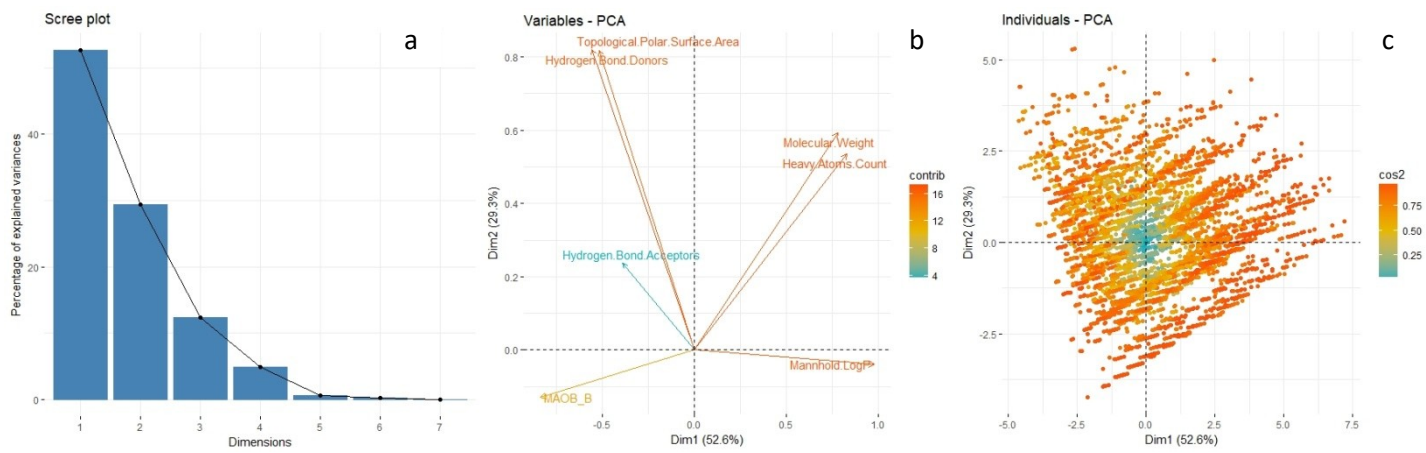
Figure 4.26: Scree plot, showing the percentage of variance explained for the PCA of binding energies of MAOB_B (a), loadings plot of the binding energy of MAOB_B structures (b) and scores plot of the binding energy (c).
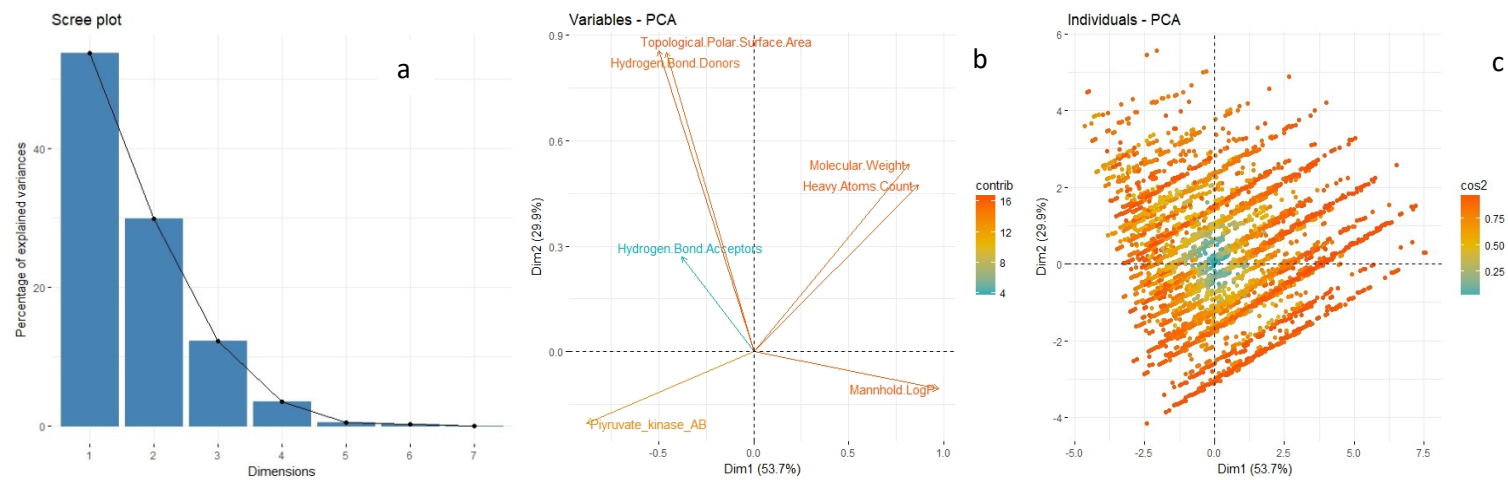
Figure 4.27: Scree plot, showing the percentage of variance explained for the PCA of binding energies of pyruvate kinase_ AB (a), loadings plot of the binding energy of pyruvate kinase_ AB structures (b) and scores plot of the binding energy (c).
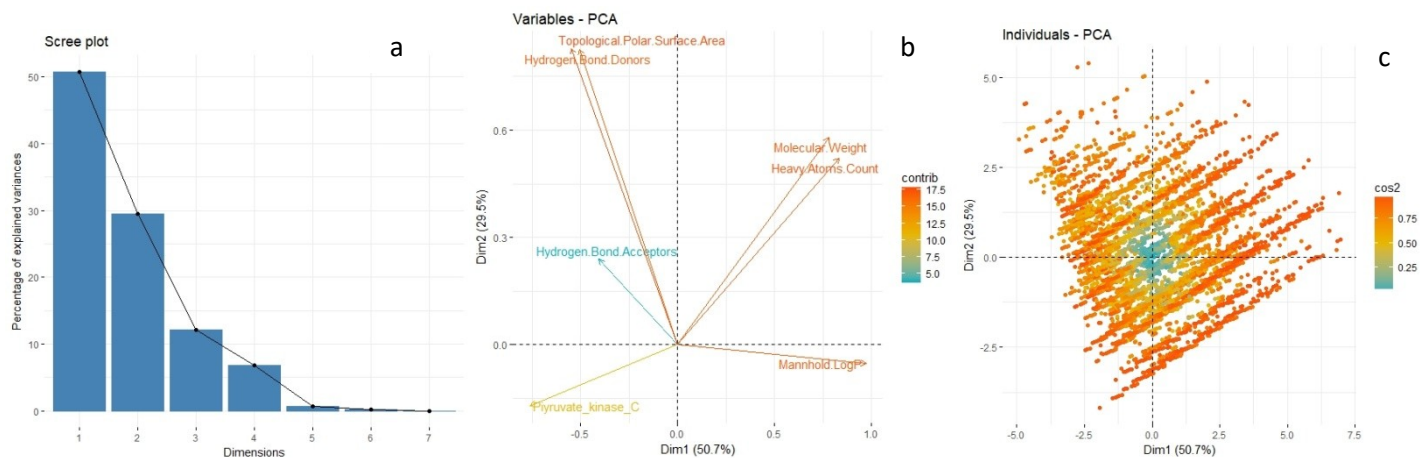
Figure 4.28: Scree plot, showing the percentage of variance explained for the PCA of binding energies of pyruvate kinase_ C (a), loadings plot of the binding energy of pyruvate kinase_ C structures (b) and scores plot of the binding energy (c).
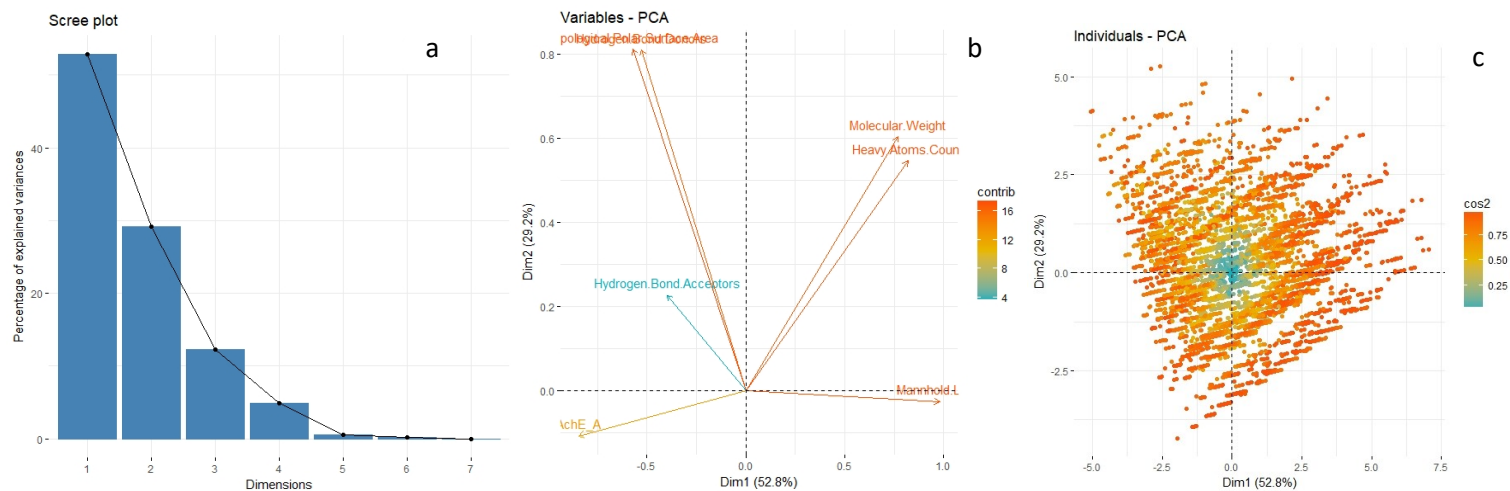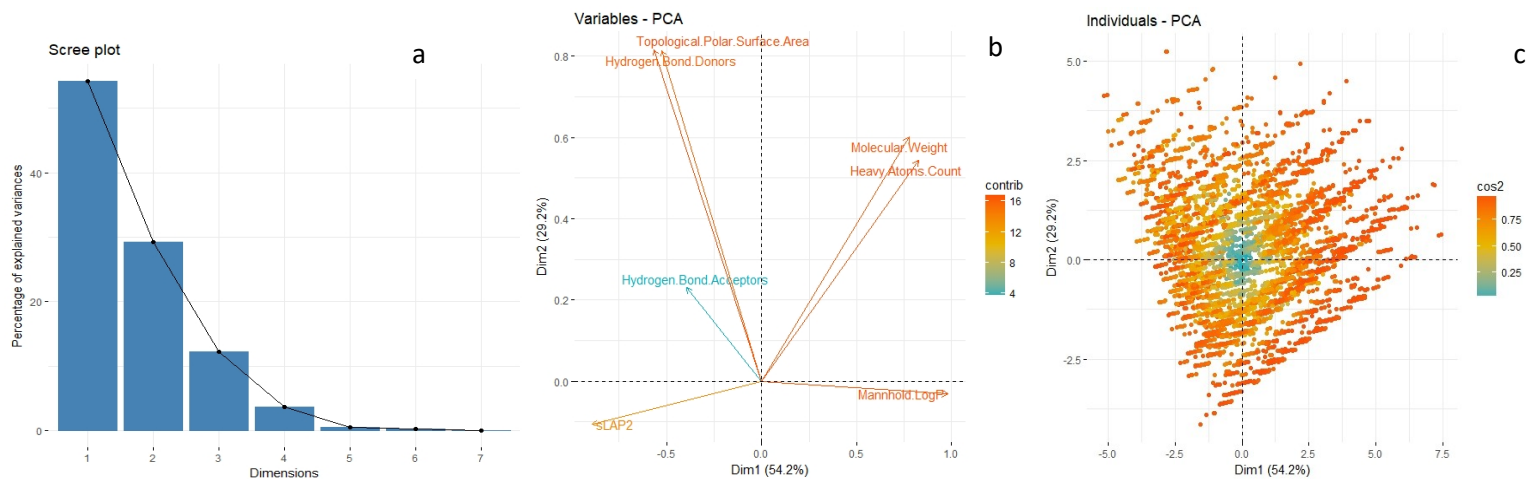


Figure 4.29: Scree plot, showing the percentage of variance explained for the PCA of binding energies of PLA$_2$ (a), loadings plot of the binding energy of PLA$_2$ structures (b) and scores plot of the binding energy (c).

Ways to identify clearer trends in the data were explored. It was hypothesized that the data was too diverse for meaningful PCA and dendrogram analysis. To this end, not all physico-chemical properties were included in a subset of data for a specific protein target. The chosen subset was subjected to hierarchical cluster analysis following the Euclidean metric. Figure 4.30 represents the dendrogram of a subset of data for the target HSP90. The data selected included only the binding energy of the HSP90-ligand complexes, and the molecular weight and Mannhold LogP physico-chemical properties of the library chalcones. These variables were selected randomly. The binding energy of the complex and the molecular weight of the chalcone combined into a single cluster, with the MLogP values separated by a large distance. However, the divisive coefficient is too low (0.3) to consider this to be an efficient separation (good values are close to 1) of the different values, and therefore to make any conclusions, for example about the effect of molecular weight on binding [13].
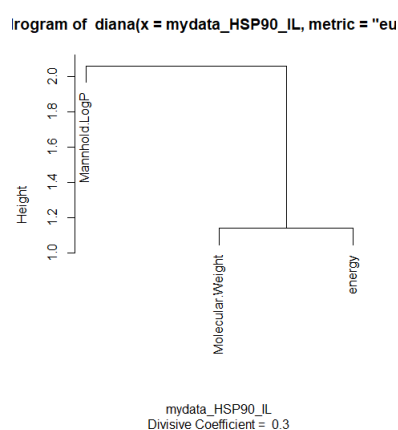


Figure 4.30: Dendrogram of divisive hierarchical clustering of the binding energies from the protein HSP90 and chemical/physical proprieties (molecular weight and MLogP) of the chalcones library following the Euclidean metric.

A second approach was a comparative approach, exploring targets in pairs. Within this approach a small number of chemical/physical parameters were kept, but the binding energy for two targets were included. Both agglomerative and divisive hierarchical cluster analyses were performed. Figure 4.31 shows the agglomerative dendrogram following the average metric of binding energies from AChE_A and BChE_A including only the physico-chemical properties molecular weight, MlogP and topological polar surface area. The graph presents a distinct separation between two clusters, with the surface area being on its own. MLogP and the binding energy from AChE_A (and remotely the topological polar surface area) present the first cluster, while the second is composed of the binding energy from BChE_A and the molecular weight. This approach provided for greater separation of clusters compared to previously. However, the agglomerative coefficient this time was too high (9.1) (this was

151

supposed to be close to 1) to make conclusions on how to tailor ligand selectivity by change of physico-chemical properties. This number could be attributed to the fact that a large dataset was used.[13] In this case, divisive hierarchical clustering proved to be a better method than agglomerative clustering.
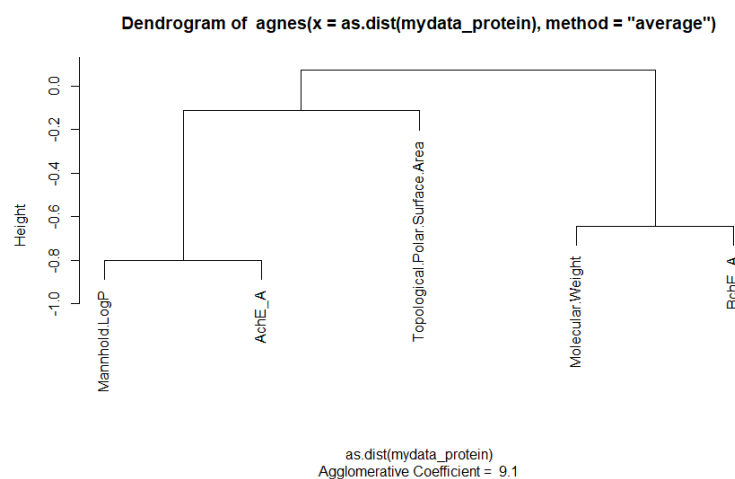


Figure 4.31: Dendrogram of agglomerative hierarchical clustering of the binding energies from proteins AChE_A and BChE_A and chemical/physical properties (molecular weight, MLogP and topological surface area) of the chalcones library following the average metric.

Given the interesting clustering, the question was raised as to what would be observed if a third set of binding energies was added to the data, but from a very dissimilar target. Acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) are from the same family of enzymes and hydrolysing the acetylcholine while the aldose reductase (ALR$_2$) is from the aldo-ketone reductase family. Thus the binding energies from ALR$_2$ were added to the subset of data, and this, larger subset, was analysed in terms of the evolution of the hierarchical cluster. Figure 4.32 represents the agglomerative dendrogram of the binding energies from AChE_A, BChE_A and ALR$_2$ and the physico-chemical properties including molecular weight, MlogP and topological polar surface area. The dendrogram remained unchanged, of course with the exception of the addition of the binding energy of ALR$_2$ which is excluded from the original clusters. The agglomerative coefficient did decrease to 4.28, but this is still not a good value.

**Dendrogram of agnes(x = as.dist(mydata_protein), method = "average")**

as.dist(mydata_protein)
Agglomerative Coefficient = 4.28
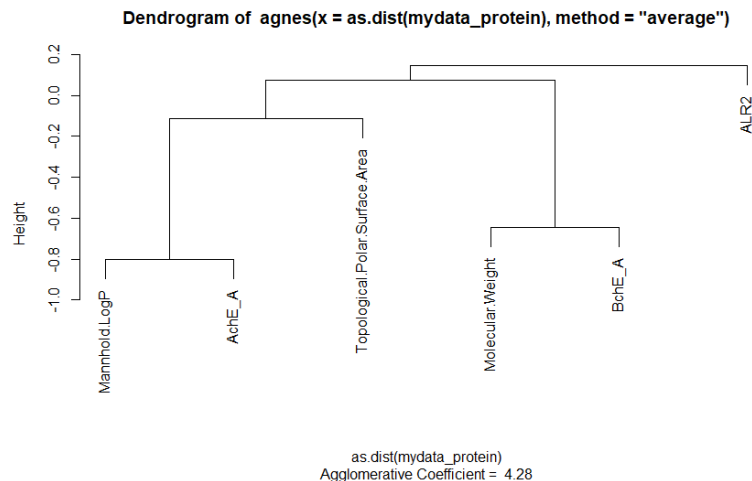
Figure 4.32: Dendrogram of agglomerative hierarchical clustering of the binding energies from proteins AChE_A, BChE_A and ALR$_2$ and chemical/physical proprieties (molecular weight, MLogP and topological surface area) of the chalcones library following the average metric.

Re-clustering of this data, using a divisive cluster separation together with a Euclidean metric, provides for the dendrogram illustrated in Figure 4.33. Again, the binding energy from the ALR$_2$ protein is largely separate, as is the topological surface area, although they are loosely part of a cluster composed of MLogP and AChE_A. The divisive coefficient is still high but in this case is close to one. These analyses were not conclusive, illustrating only slight connections and not definite physico-chemical properties that are exploitable in the tuning of the chalcone library to a particular target. However the methodology used in the preliminary clustering analysis (divisive hierarchical cluster with Euclidean metric) proved useful for later clustering using very different sets of data (geometrical rather than physico-chemical). It was concluded, that although we may observe very loose trends with regard to physico-chemical properties, these were not significant to follow in explaining the binding energies and the performance and specificity of the chalcone library against the series of protein targets.
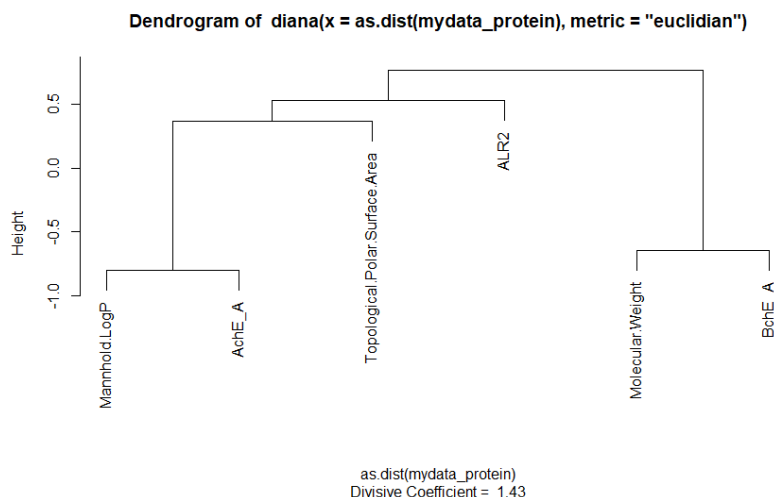
Figure 4.33: Dendrogram of agglomerative hierarchical clustering of the binding energies from proteins AChE_A, BChE_A and ALR$_2$ and chemical/physical proprieties (molecular weight, MLogP and topological surface area) of the chalcones library following the Euclidian metric.

### 1.3.3. Identification of common denominator from spherical polar coordinates of ligands molecule

A different, structural, approach was therefore followed in the analysis of the performance of ligands within the chalcone library. The hypothesis for this was that structural trends within the chalcone library affected the binding energy in a predictable fashion, or more specifically, that the spherical polar coordinates of key sites on the ligand within the active site influence the binding energy of the ligand-protein complex. Should this hypothesis prove correct, then it would imply that we would know structurally how to modify ligands in order to improve binding. In this approach the dendrograms were explored and analysed prior to exploration and analysis of the PCA maps. At this stage we were looking for the shortest distance within clusters on the dendrogram between the binding energy values and structural parameters that may cluster with the binding energy, for particular targets individually.

As such, the polar coordinates for the key benzaldehyde derived and acetophenone derived centres were calculated from the binding pose and incorporated into datasets, per target, including the binding energy. The polar coordinates were calculated according to Figure 4.3 and the discussion associated with this figure.

Figures 4.34 and 4.35 represent the separate dendrograms for each protein target, using the divisive (DIANA) clustering method together with a Euclidean metric. Interestingly, it was observed that each dendrogram gave different results even for the co-crystal structure of the

same protein. It seems that each co-crystal structure of the same protein is independent in the determination of a common denominator explaining the binding energy value, most likely due to conformational differences. For all co-crystal structures (with the exception of the pyruvate kinase), the divisive coefficient varied from 0.35 to 0.67. The pyruvate kinase data gave the best coefficient with 0.83 for the AB co-crystal structure and 0.7 for the C models. The lowest divisive coefficient was from the COX-2 (structure A), with a value of 0.3.
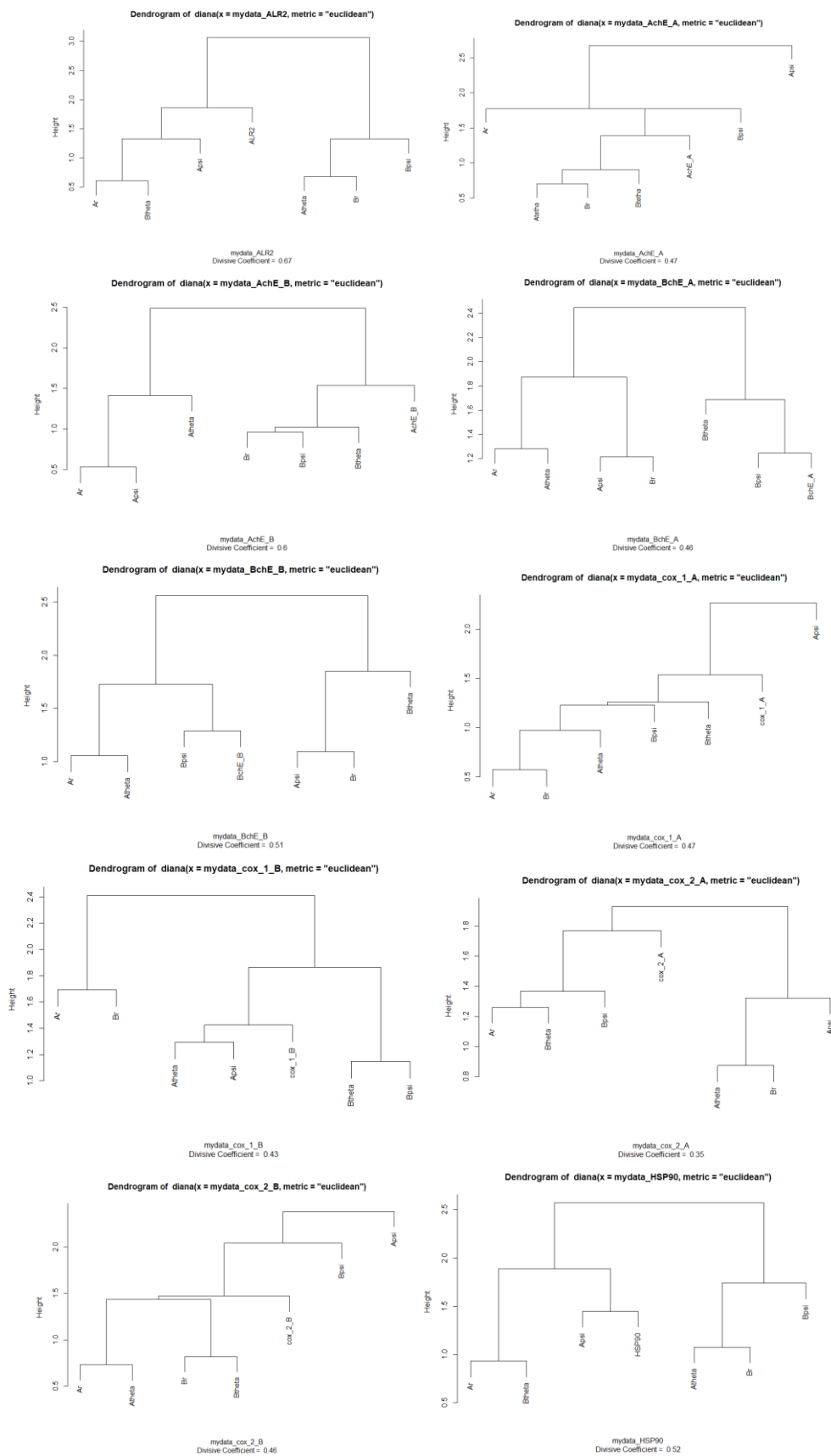
Figure 4.34: Divisive hierarchical clustering dendrograms of the spherical polar coordinates of the chalcones and the binding energies per proteins (1st part) following the Euclidean metric. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.
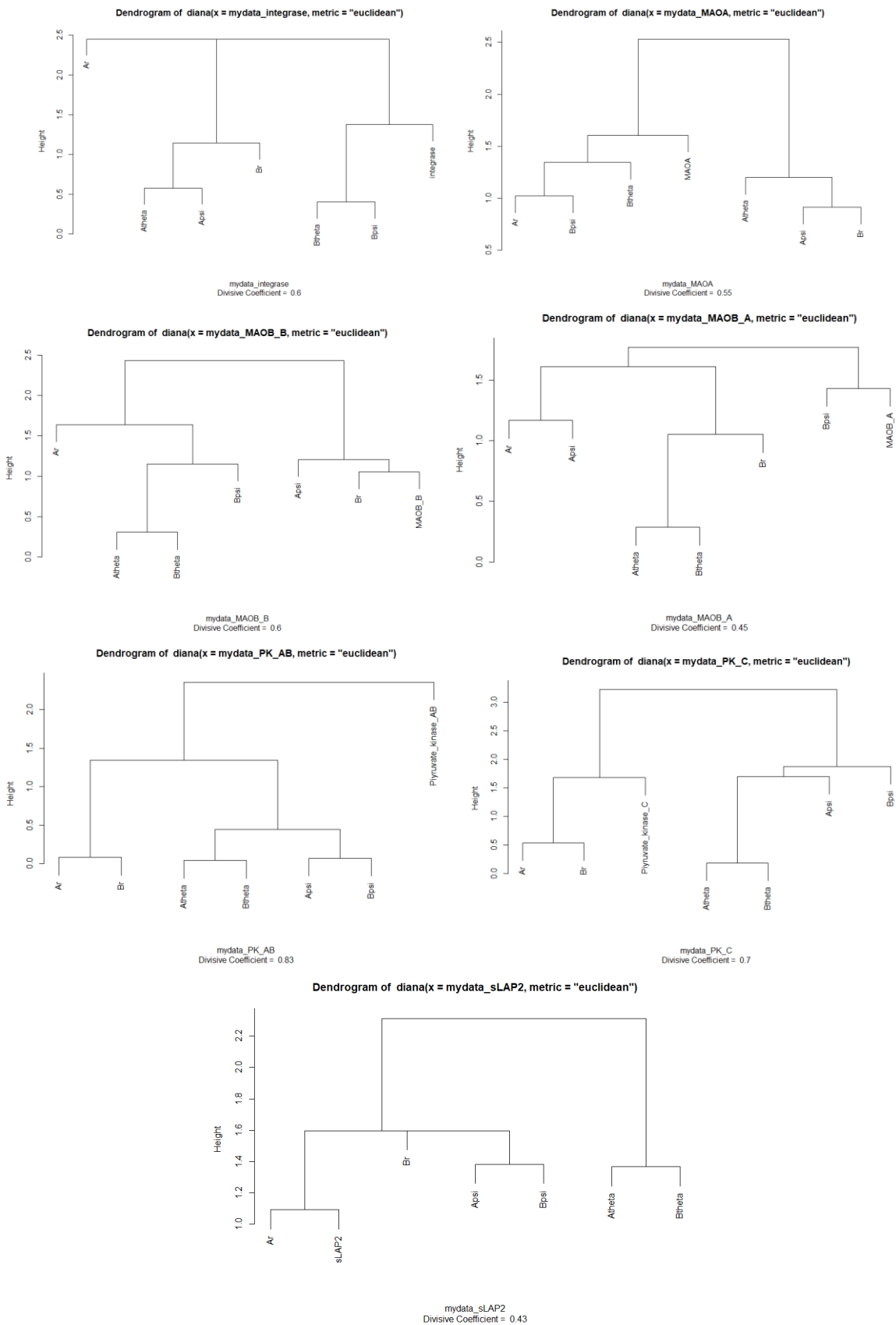
Figure 4.35: Divisive hierarchical clustering dendrogram of the spherical polar coordinates of the chalcones and the binding energies per proteins following the Euclidean metric (2nd parts). r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

The analysis of the dendrogram of AChE_A (Figure 4.36) showed the proximity of the binding energy of a cluster which is composed of the radial distance and the polar angle of the benzaldehyde carbon ($B_r$, $B_{theta}$) and the polar angle of the acetophenone carbon ($A_{theta}$). However, for the AChE_B dendrogram, the binding energy value is linked to the radial distance and the azimuthal and polar angle of the benzaldehyde carbon ($B_r$, $B_{psi}$ and $B_{theta}$).
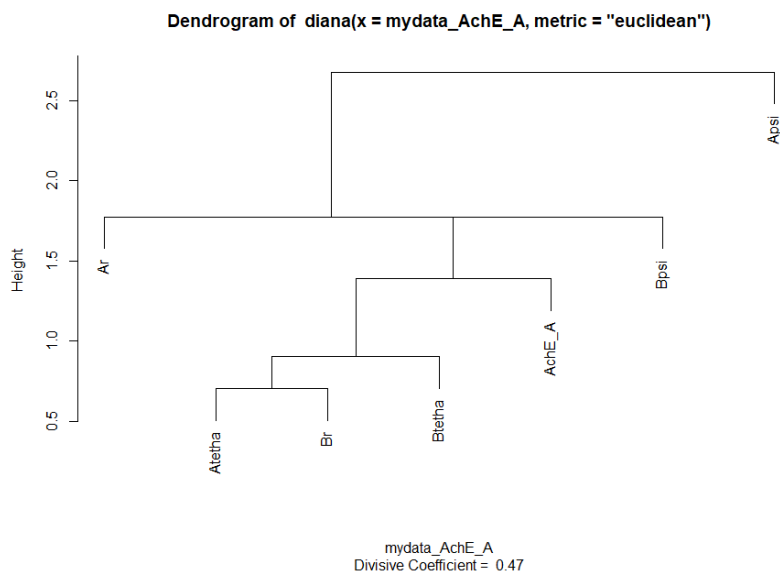


Figure 4.36: Divisive hierarchical clustering dendrograms of the spherical polar coordinates of the chalcones and the binding energies of AChE_A following the Euclidean metric. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

For BChE_A (shown in Figure 4.37), only the orientation of the benzaldehyde ($B_{psi}$) may explain the binding energy of the complexes. The binding energy of its co crystal structure is also closely associated with the azimuthal angle of the benzaldehyde carbon ($B_{psi}$).

**Dendrogram of diana(x = mydata_BchE_A, metric = "euclidean")**

mydata_BchE_A
Divisive Coefficient = 0.46
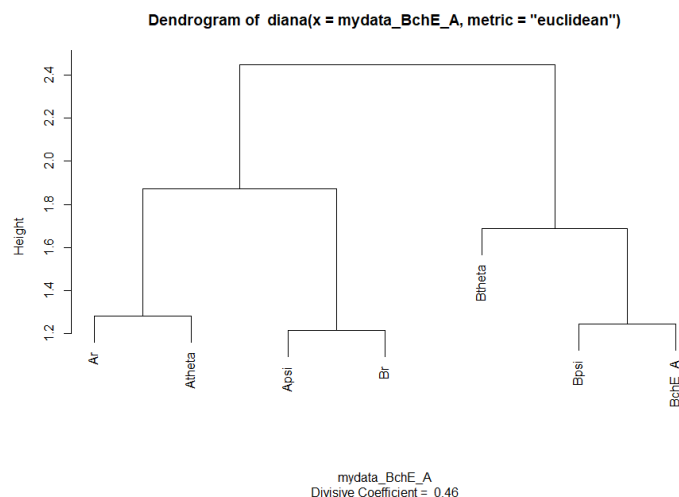
Figure 4.37: Divisive hierarchical clustering dendrograms of the spherical polar coordinates of the chalcones and the binding energies of BChE_A following the Euclidean metric. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

In another interesting result, the values for the binding energy of COX-1_A (Figure 4.38) were found to be linked to a cluster of the other parameters except for the azimuthal angle of the acetophenone centre ($A_{psi}$). This may demonstrate its relative importance when analysing structural effects on the docking score.



**Dendrogram of diana(x = mydata_cox_1_A, metric = "euclidean")**

mydata_cox_1_A
Divisive Coefficient = 0.47
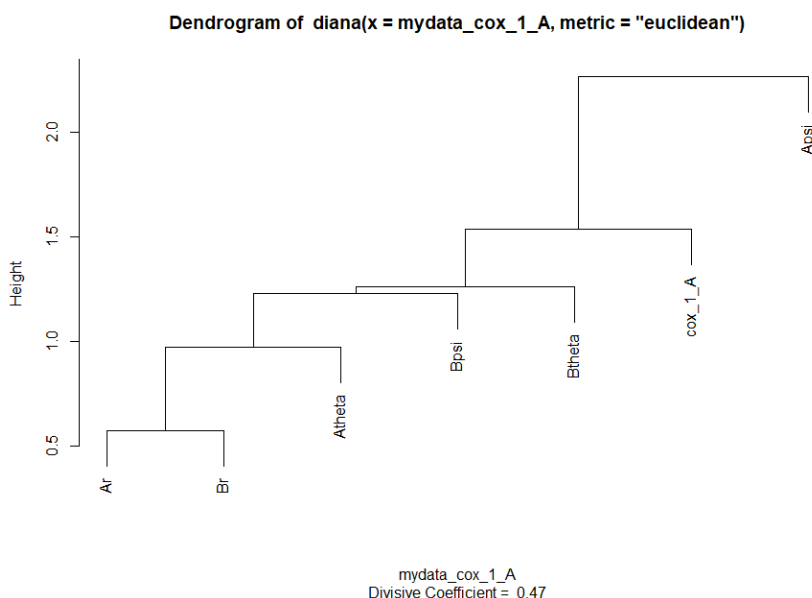
Figure 4.38: Divisive hierarchical clustering dendrograms of the spherical polar coordinates of the chalcones and the binding energies of COX_1_A following the Euclidean metric. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

The most promising result was obtained for the pyruvate kinase model C. The radial distances of both the acetophenone and benzaldehyde ($A_r$ and $B_r$) are closely associated with the binding score (shown in Figure 4.39). Furthermore it seems the polar angles have a weak impact on the result of the binding energy. Despite these results, other pyruvate kinase domains show no preference for a specific parameter in clustering with the binding energy.



**Dendrogram of  diana(x = mydata_PK_C_wo_A56, metric = "euclidean")**

mydata_PK_C_wo_A56
Divisive Coefficient =  0.7
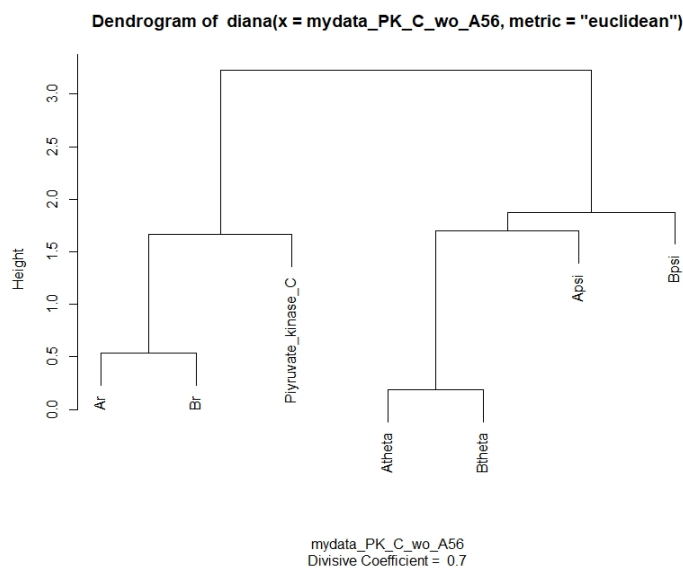
Figure 4.39:  Divisive hierarchical clustering dendrogram of the spherical polar coordinates of the chalcones and the binding energies of the protein pyruvate kinase domain C following the Euclidean metric. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

It was decided to perform PCA calculations with these structural parameters for the pyruvate kinase C domain. The plots for this PCA are presented in Figure 40. With 46.5% and 19.5% (total 66%) of the variance explained by the first two components respectively, the PCA of the variables confirm the presence of at least two different clusters, which include, in the first cluster the radial distance parameters ($A_r$ and $B_r$), and in the second cluster the polar angle of the acetophenone and benzaldehyde ($A_{theta}$ and $B_{theta}$). Opposite on the second dimension (y-axis), azimuthal angles of the acetophenone and benzaldehyde derivated ($A_{psi}$ and $B_{psi}$) may be independent of the different variables. These parameters contribute strongly to the PCA in opposition to the binding energy. The scores plot of the PCA (Figure 4.40 (c)) presents four distinct clusters. It is interesting to note that the centre of the first cluster is surrounded by data which is not well-represented. Careful examination of several docking poses from each of the four clusters did not reveal a clear common denominator defining each cluster.

From the PCA and clustering analysis of the pyruvate kinase C, we decided to focus on this particular set of pyruvate kinase screening results, exploring the importance of the radial distance and the polar angles of acetophenone and benzaldehyde-derived moieties in influencing binding and the observed the binding energy. The strategy followed in order to do this, is discussed in Chapter 5.
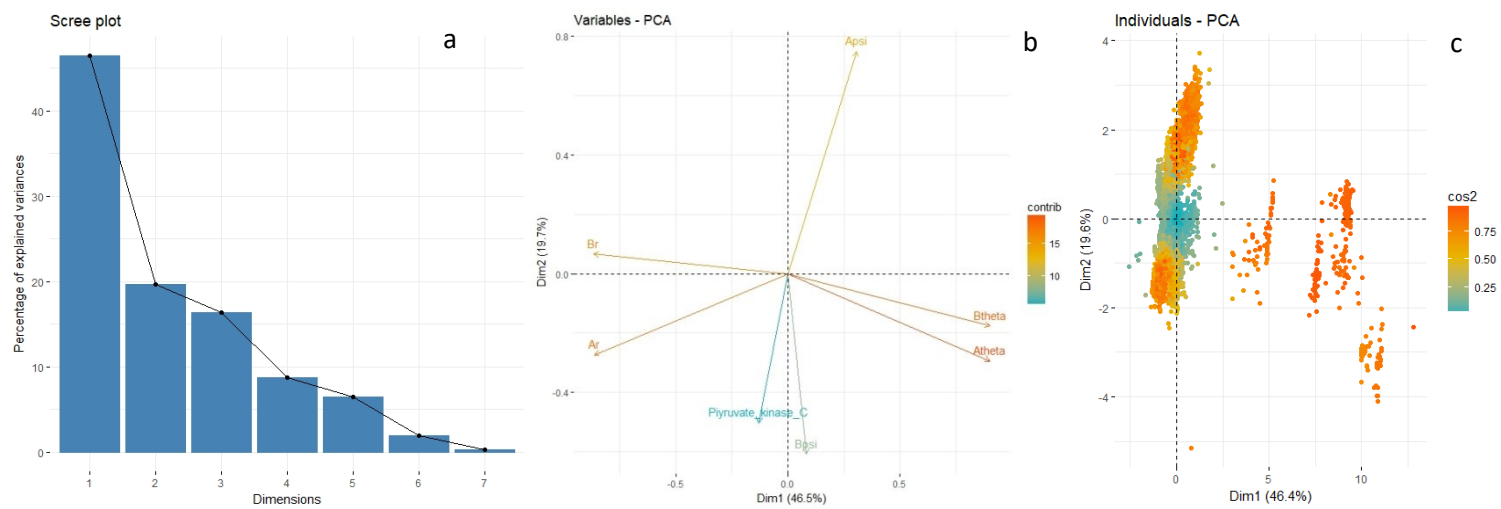
Figure 4.40:  Scree plot, showing the percentage of variance explained for the PCA (a), loadings plot of the PCA (b) and scores plot (c) of chemical/physical proprieties of the chalcone and

binding energies from the AChE structure A. r, radial distance; theta, polar angle; psi, azimuthal angle; A, acetophenone; B, benzaldehyde.

## 2.  Machine learning with Knime

Given that it is possible to find associations between structural parameters and binding energies, it was of interest to create a machine learning procedure based solely on ligand characteristics (which, if successful would outperform the high throughput screening using docking) and test this model, using simply the energy data from the pyruvate kinase C crystal structure.

### 2.1.  Data preparation

An identical file (of csv format) to the one described in the spherical polar coordinate PCA calculation was utilised, although not all values were used. An additional column was introduced which flagged the activity of the chalcone derivatives. This flag was set to be active if the binding energy of the chalcone-pyruvate kinase C complex was calculated to be less than or equal to -8 kcal.mol$^{-1}$ flagged as inactive otherwise. The structures of the chalcone library were saved into one unique SDF file.

Figure 4.41 represents the Knime workflow used for the machine learning. The SDF of the chalcone structure was read into the workflow and converted to RDKit molecular format. From this, a canonical structure of SMILES was generated and added as a column to the table. The results of this were merged with the data from the csv file containing the spherical polar coordinates and the binding energy from the pyruvate kinase C crystal model.  The column of the canonical SMILES was typecast to SMILES format using the Molecule Type Cast node. The table rows were next grouped by the canonical SMILES column (as a precaution with respect to possible presence of duplicates, there were no duplicates in the data). Using the SMILES, RDKit fingerprints were calculated and added as a new column to the table. The type of fingerprint used was a Morgan Fingerprint utilizing 2048 bits with a radius of 2. These fingerprints, together with the corresponding activity data were submitted to a random Forest learner from which a decision tree (100 decision trees for this study) was generated. In order to visualize the results, three nodes were installed. The scorer helped to compare the original activity of the library against the predicted values from the random Forest learner and gave the proportion of true and false positive and negatives. The Receiver Operating Characteristic (ROC) curve indicated the quality of the learning method.  The last node prepared a summarized table with all results.
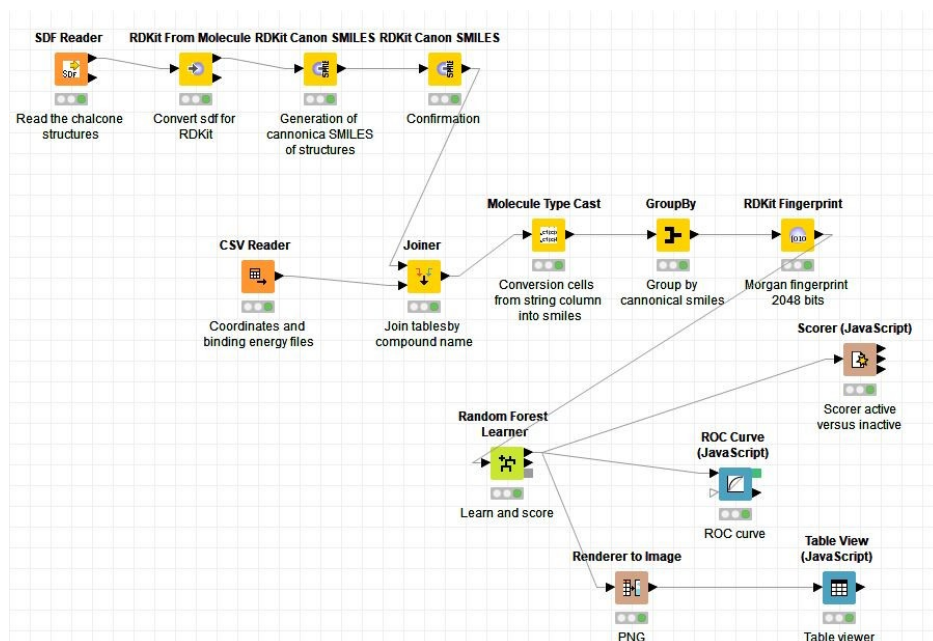
Figure 4.41: Machine learning workflow with Knime.

## 2.2. Results

The results focus on the Scorer and the ROC curve. The matrix confusion is presented in Table 4.3. The number of true positive and true negative results represented are high (2270 active and 4386 inactive respectively) whilst false positives and negatives are low (275 active and 494 inactive.

| Row ID | active | inactive |
|---|---|---|
| active | 2270 | 494 |
| inactive | 275 | 4386 |

Table 4.3: Confusion matrix.

However, in order to fully assess the decision tree performance, further quality indicators have to be considered. To this end the sensitivity, specificity, recall, precision and the F-measure were calculated and these are presented in Table 4.4. It is interesting to note that for both active and inactive compounds, the values for recall, precision, sensitivity, specificity and the F-measure are all close to 1. The recall and the sensitivity for active compounds are both equal to 82.1% and represent the probability of correct detection of an active compound inside of chalcone library. The specificity measures the proportion of inactive molecules which are correctly filtered out as not active. The measured specificity in this row is over 94%. The precision represents the probability that a random molecule of the library is truly

active and is equal to 89.2 %. The F-measure, a harmonic mean of the recall and the precision of the active compounds, is high at 85.5%. The inactive section presents similar results.

| Row ID | I True Po... | I False P... | I True Negatives | I False N... | D Recall | D Precision | D Sensitivity | D Specificity | D F-measure |
|--------|--------------|--------------|------------------|--------------|----------|-------------|---------------|---------------|-------------|
| active | 2270 | 275 | 4386 | 494 | 0.821 | 0.892 | 0.821 | 0.941 | 0.855 |
| inactive | 4386 | 494 | 2270 | 275 | 0.941 | 0.899 | 0.941 | 0.821 | 0.919 |

Table 4.4: Class statistic table.

Figure 4.42 represents the graph of the ROC curve of the machine learning. With a representation of the activity equal to 0.962. Active chalcones represents the best performance.
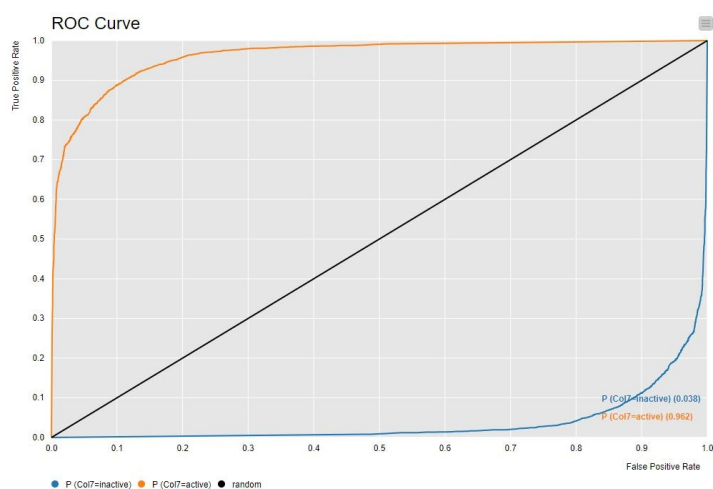


Figure 4.42: ROC curve; Col7, activity/inactivity.

The random forest learner exhibits excellent efficiency in differentiating between active and inactive chalcones in the library, in terms of activity against the pyruvate kinase C crystal structure.

# Conclusion

In this chapter, trends and associations within the plethora of data have been explored. Statistical methodologies were implemented with success and allowed us to improve the analysis of large amounts of data. PCA and cluster analysis, which are complementary statistical methods have confirmed the following. Although only loose associations have been uncovered in general, there appear to be, in the case of the pyruvate kinase, links between structural features in the chalcone ligands and the observed binding energy. This link is explored in Chapter 5. Finally, Random Forest Learning is successful in predicting active or inactive ligands in terms of binding against the pyruvate kinase crystal structure.

# REFERENCES

1.      Jolliffe, I.T., *Principal Component Analysis, Second Edition*, Springer, Editor. 2002.
2.      J. E. V. Ferreira, A. F. de Figueiredo, J. P. Barbosa, and J.C. Pinheiro, *Chemometric Study on Molecules with Anticancer Properties*, in *CHEMOMETRICS IN PRACTICAL APPLICATIONS*, K. Varmuza, Editor. 2012.
3.      L. V. Pérez-Arribas, M. E. Leõn-González, and N. Rosales-Conrado, *Learning Principal Component Analysis by Using Data From Air Quality Networks.* J Chem Edu, 2017. **94**(4): p. 458-464.
4.      Boley, D., *Principal Direction Divisive Partitioning.* Data Min Knowl Disc, 1998. **2**(4): p. 325-344.
5.      Wilks, D.S., *Statistical Methods in the Atmospheric Sciences*, D.S. Wilks, Editor. 2011.
6.      S. R. Dorling, T. D. Davies, and C.E. Pierce, *Cluster analysis: a technique for estimating the synoptic meteorological control on air and precipitation chemistry-method and applications.* Atmospheric Environment, 1992. **26A**(14): p. 2575-2581.
7.      B. Wang, L. Yang, L. Zhang, and F. Li, *Robust Multi-view Features Fusion Method Based on CNMF*, in *Lecture Notes in Computer Science*, L. Cheng, A. Leung, and S. Ozawa, Editors. 2018, Springer. p. 27-39.
8.      A.Kassambara, *Practical Guide to Cluster Analysis in r: Unsupervised Machine Learning*, C.I.P. Platform, Editor. 2017. p. 188.
9.      M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T.Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, *KNIME: The Konstanz Information Miner*, in *Data Analysis, Machine Learning and Applications*, C. Preisach, et al., Editors. 2008, Springer-Verlag Berlin Heidelberg. p. 319-326.
10.     S. Beisken, T. Meinl, B. Wiswedel, L. F. de Figueiredo, M. Berthold, and C. Steinbeck, *KNIME-CDK: Workflow-driven cheminformatics.* BMC Bioinformatics, 2013. **14**(257): p. 1-4.
11.     S. Lê, J. Josse, and F. Husson, *FactoMineR: An R Package for Multivariate Analysis.* J Stat Softw, 2008. **25**(1).
12.     Team, R., *RStudio: Integrated Development for R.* , RStudio, Editor. 2016: Boston, MA.
13.     L. Kaufman and P.J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, Editor. 1990, John Wiley & Sons Inc.

# Chapter 5: Modification of leads in two cases, based on either the Topliss decision tree, or based on the roles of the radial distance and the polar angles in the observed binding energies.

This chapter discusses two hypotheses which were developed in order to attempt, firstly, to improve the binding energy and, secondly, to prove the importance of some physical parameters in the binding energy. The first experiment involved modifying the substituent of one of the best binding chalcones for the HSP90, inspired by the Topliss decision tree and re-docked against HSP90. The second experiment was based on the statistical results from the previous chapter allowing us to volunteer the following hypothesis: "The radial distances of the acetophenone and benzaldehyde moieties and the polar angles may have an impact on the binding energy of the chalcone-pyruvate kinase complex". In order to confirm or to refute this hypothesis, we decided to design a specific library for each of the physical parameters and re-dock all structures of the pyruvate kinase used previously.

## 1. Attempt to improve the binding energy of a hit in the chalcone library against HSP90

### 1.1. Design of the library

The first experiment was also undertaken in order to investigate for a lead by improving the binding energy. It was decided to test the new library against the HSP90 crystal structure [PDB ID: 5FNC]. [1] We decided not to use the best lowest binding energy molecule (acetophenone_22_benzaldehyde_78 with a binding energy of -13.5 kcal.mol$^{-1}$) as the base for the modified library. This choice is based on an attempt to significantly improve the binding energy of the chalcones which were not in the top five of the best (lowest energy binding) molecules tested against HSP90. Our decision led us to select the acetophenone_22_benzaldehyde_17 which is ranked in 6$^{th}$ position with a binding energy of -13 kcal.mol$^{-1}$. This molecule was subject to modifications. The protons of the benzaldehyde-derived moiety (represented in Figure 5.1) were replaced by a selection of atoms inspired by the Topliss tree. This decision tree, when applied to a molecule, helps to develop it into a modified system that is better in binding by consideration of the nature and properties of each

atom. A template for this decision tree from the literature is shown in Figure 5.2.[2] The substituents selected for this new library were chloride, iodide, methyl, methyl ether and hydroxyl groups. All four protons were replaced each time. Instead of using our template scripts to replace the substituents, a decision was made to use Schrödinger software and two modules in particular: the reagent preparation (which helped to localize the protons which we wanted to replace) and the combinatorial library enumeration (which facilitated the design of the library). Instead of modifying the molecules on an individual basis, the decision was made to generate all of the possible modified chalcones in a single procedure, comparing the binding energy at each level of the Topliss tree. A library of 625 modified acetophenone_22_benzaldehyde_78 molecules was created.



Acetophenone_22_benzaldehyde_17
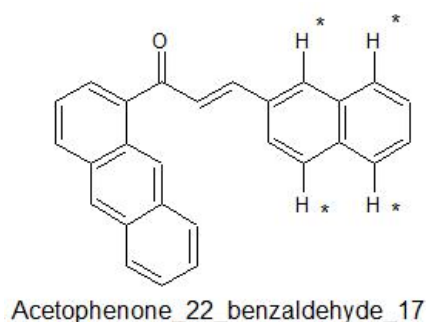
Figure 5.1:  Acetophenone_22_benzaldehyde_17. The protons with the star were submitted to modifications.
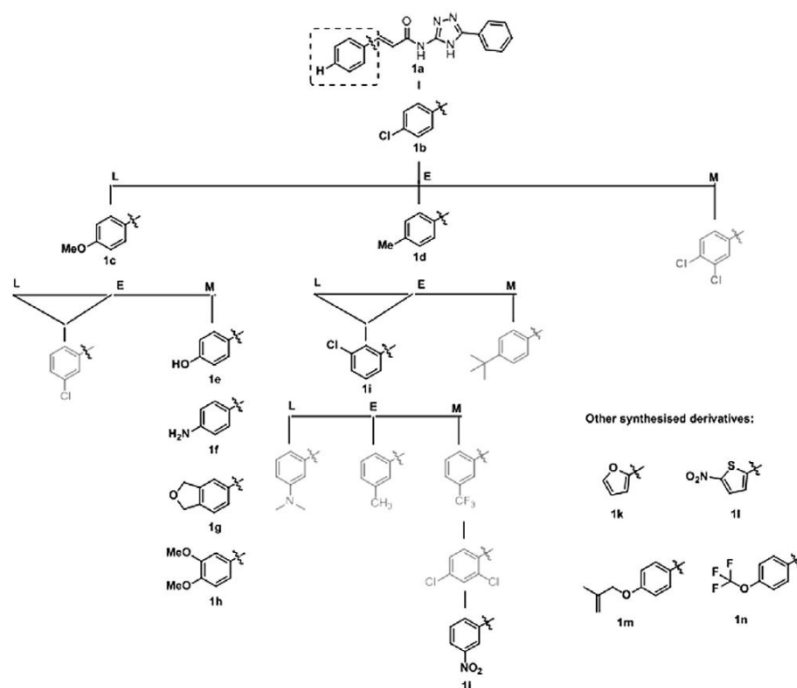


Figure 5.2: Template of Topliss tree used N-(3-aryl-1,2,4-triazol-5-yl) cinnamamide derivatives as antimicrobial agents (with the permission of the publisher). L, lowest activity; E, equivalent activity; M, Major activity. [2]

A geometric optimization of all 625 structures was necessary to obtain the final library. The unique file, containing the entire library, was imported into KNIME. To do this, the library was first converted to SDF, a Babel compatible extension. Babel (within KNIME) was used to convert and split each molecule into com format (as text in a table). Three lines were added to each row in order to specify memory used (7GB, using the %mem=7GB directive), number of processors shared (8, using the %nprocshared=8 directive) and the task asked (geometric optimization with the method AM1). As such each row in the KNIME table contained the full text of a workable Gaussian 09 input file. In order to write these to individual ".com" files, a loop in KNIME was created in order to read the name of the molecule, create an appropriate file name and use the csv file writer to write to the files. Writing as csv files did work, since the generated Gaussian text was as a single column, and no commas were written as would be if a second column was present to provide for comma-separated values. Figure 5.3 represents the workflow created for the preparation of the library for the geometric optimization. The ".com" were submitted for geometric optimization with Gaussian 09 using the departmental cluster. After the calculation, files were converted to pdb and mol2 formats, using scripts in the traditional manner.
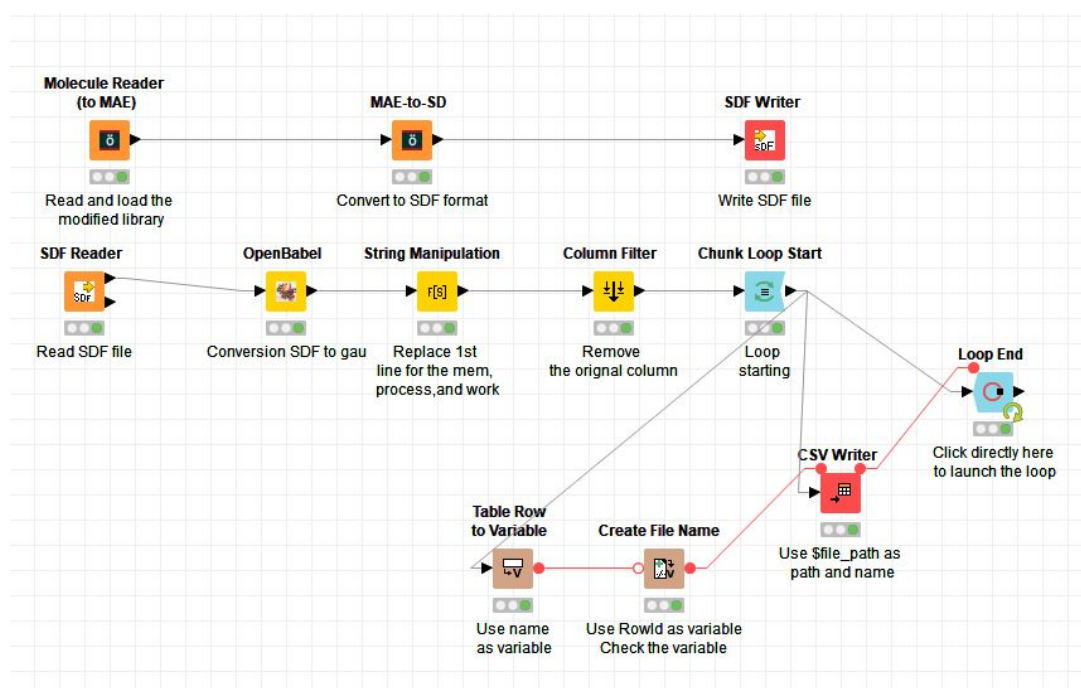


Figure 5.3: Workflow of the preparation of the molecule for a geometric optimization.

Following the geometric optimization, the molecules were prepared in pdbqt format in order to be used in AutoDock Vina (using the AutoDock Tools script, prepare_ligand4.py individually for each structure).

The HSP90 had already been prepared, using the procedures described in Chapter 2. The next step was the specification of the search area and preparation of the AutoDock Vina input files. With the exception of the number of processors (the departmental cluster has nodes with fewer cores than the CHPC), all other parameters specified for docking to HSP90 remained the same as described in Chapter 3 and are presented in Table 5.1 below.

| Crystal structure | Cartesian Coordinates | | | Size of the grid box | Energy range | Exhaustiveness | CPU |
|---|---|---|---|---|---|---|---|
| HSP90 | 2.269 | 10.687 | 24.493 | 48/44/46 | Normal | 120 | 4 |

Table 5.1: Parameters using for the high-throughput virtual screening of the modified chalcone library against HSP90.

Binding energy and poses were collected in excel and pdb format respectively.

## 1.2. Results

### 1.2.1. Binding energy analysis

The results of the 50 best binding ligands are represented in Table 5.2. By comparing all of the binding energies of the modified library to the binding energy of the original acetophenone_22_benzaldehyde_17 (-13 kcal.mol$^{-1}$) it was noted that no improvement in binding was obtained for this system. It was interesting to note the presence of molecules containing chlorine, methyl and iodine moieties are within the top 50 performing ligands. In this particular case, the Topliss decision tree failed to improve the binding energy, a result that was not expected. Although this methodology did not work with modifying this particular starting ligand, it is possible that this failure is not general, and that there are ligands within the original chalcone library for which improvement in binding may be effected. AutoDock Vina was used for docking for consistency (within KNIME, using GLIDE to dock would have been an easier option). The problem, though, with using AutoDock Vina with these modified ligands relates to the heavy presence of aromatic halides (particularly of the heavier halides including iodine) in the ligand set. AutoDock Vina cannot account for halogen bonding in these systems, and it is likely that halogen bonding will affect binding particularly for this modified set of ligands. An option within other docking software such as Schrödinger GLIDE allows for the consideration of the halogen as a donor or acceptor of electrons. High throughput virtual screening of the original library of chalcones and this subset of modified ligands using docking software such as Schrödinger GLIDE therefore may well resolve this issue; however this is beyond the scope of this work.

| name | energy (kcal.mol-1) |
|---|---|
| acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl | -12,9 |
| acetophenone_22_benzaldehyde_17_Cl_Cl_Me_Cl | -12,9 |
| acetophenone_22_benzaldehyde_17_Cl_Me_Me_Cl | -12,9 |
| acetophenone_22_benzaldehyde_17_I_Cl_Cl_Cl | -12,9 |
| acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Me | -12,8 |
| acetophenone_22_benzaldehyde_17_Cl_Me_I_OH | -12,8 |
| acetophenone_22_benzaldehyde_17_I_Cl_Me_Me | -12,8 |
| acetophenone_22_benzaldehyde_17_I_I_I_Cl | -12,8 |
| acetophenone_22_benzaldehyde_17_I_OH_I_Cl | -12,8 |
| acetophenone_22_benzaldehyde_17_Me_Cl_I_Cl | -12,8 |
| acetophenone_22_benzaldehyde_17_Cl_Me_I_Me | -12,7 |
| acetophenone_22_benzaldehyde_17_Cl_OH_I_Me | -12,7 |
| acetophenone_22_benzaldehyde_17_Cl_OH_Me_Cl | -12,7 |
| acetophenone_22_benzaldehyde_17_I_Cl_I_OH | -12,7 |
| acetophenone_22_benzaldehyde_17_I_Cl_Me_OH | -12,7 |
| acetophenone_22_benzaldehyde_17_I_Me_Cl_Cl | -12,7 |
| acetophenone_22_benzaldehyde_17_I_OH_Me_Cl | -12,7 |
| acetophenone_22_benzaldehyde_17_Me_Cl_OMe_Me | -12,7 |
| acetophenone_22_benzaldehyde_17_Me_Me_I_Cl | -12,7 |
| acetophenone_22_benzaldehyde_17_OH_Cl_I_Cl | -12,7 |
| acetophenone_22_benzaldehyde_17_Cl_Cl_OH_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_Cl_Me_I_OMe | -12,6 |
| acetophenone_22_benzaldehyde_17_Cl_Me_OH_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_Cl_Me_OMe_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_I_Cl_Cl_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_I_Cl_I_OMe | -12,6 |
| acetophenone_22_benzaldehyde_17_I_Cl_OMe_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_I_Me_Me_Cl | -12,6 |
| acetophenone_22_benzaldehyde_17_I_OH_Cl_Cl | -12,6 |
| acetophenone_22_benzaldehyde_17_I_OH_Me_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_Me_Cl_OH_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_Me_Me_Me_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_Me_Me_OMe_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_OH_Cl_Me_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_OH_Me_Cl_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_OH_Me_Me_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_OH_Me_OMe_OH | -12,6 |
| acetophenone_22_benzaldehyde_17_OMe_Cl_I_Cl | -12,6 |
| acetophenone_22_benzaldehyde_17_OMe_Cl_Me_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_OMe_OH_Me_Me | -12,6 |
| acetophenone_22_benzaldehyde_17_Cl_Cl_Me_OMe | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_Cl_OH_Cl | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_I_Cl_Cl | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_I_OH_Me | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_Me_Me_OMe | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_OH_I_OH | -12,5 |
| acetophenone_22_benzaldehyde_17_Cl_OH_Me_Me | -12,5 |
| acetophenone_22_benzaldehyde_17_I_Cl_Cl_OMe | -12,5 |
| acetophenone_22_benzaldehyde_17_I_Cl_Me_OMe | -12,5 |

Table 5.2: Top 50 best lowest energy complex ligands-protein. The binding energy of the original ligand acetophenone_22_benzaldehyde_17 was -13 kcal.mol$^{-1}$.

## 1.2.2. Interaction ligand-protein analysis

It was interesting to explore the reasons behind the poor performance of this modified library, and to do this the protein ligand interactions were explored with the comparison of the interactions between non-modified acetophenone_22_benzaldehyde_17 and the best performing ligand in the modified library (which possesses four chlorine substituents) against HSP90. Figure 5.4 represents the poses of the original acetophenone_22_benzaldehyde_17 and the modified tetrachloro ligand. The acetophenone-derived moieties of the chalcones are very closely aligned while the modified benzaldehyde-derived portion of the modified chalcone shows a clear rotation between the two ligands.
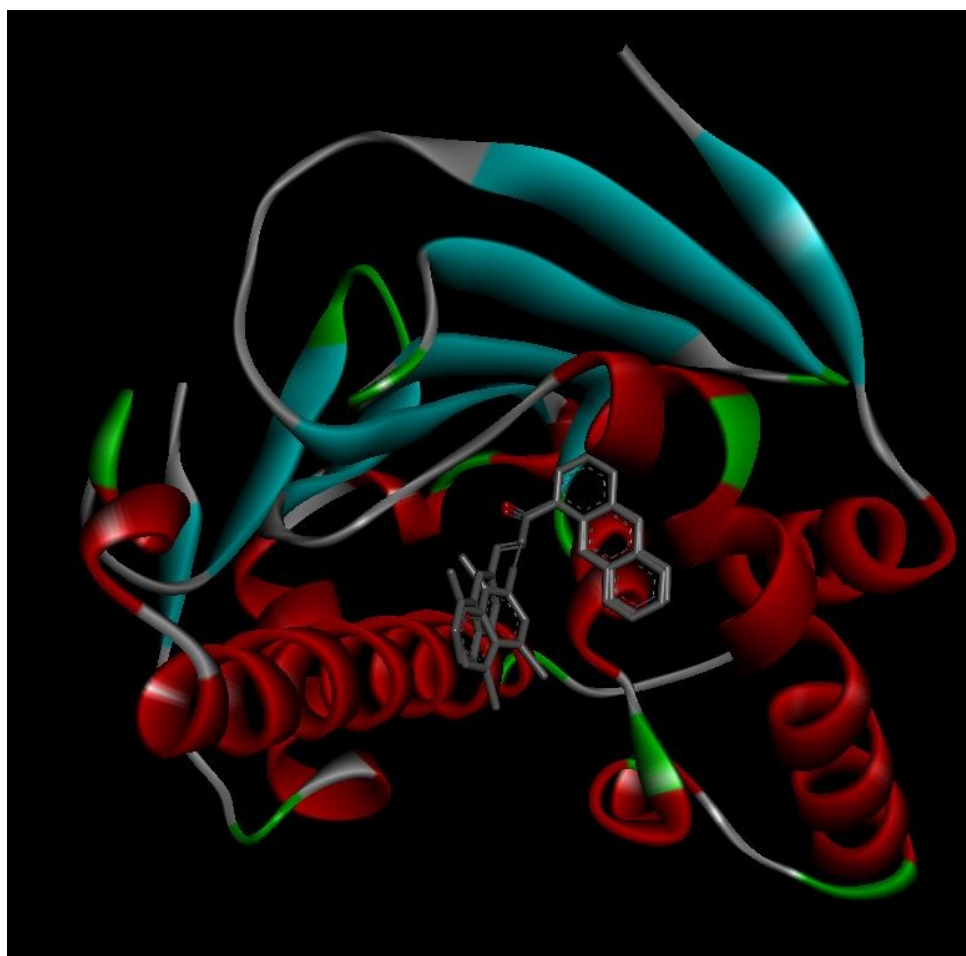


Figure 5.4: Poses of the original acetophenone_22_benzaldehyde_17 and the modified one with 4 chlorines.
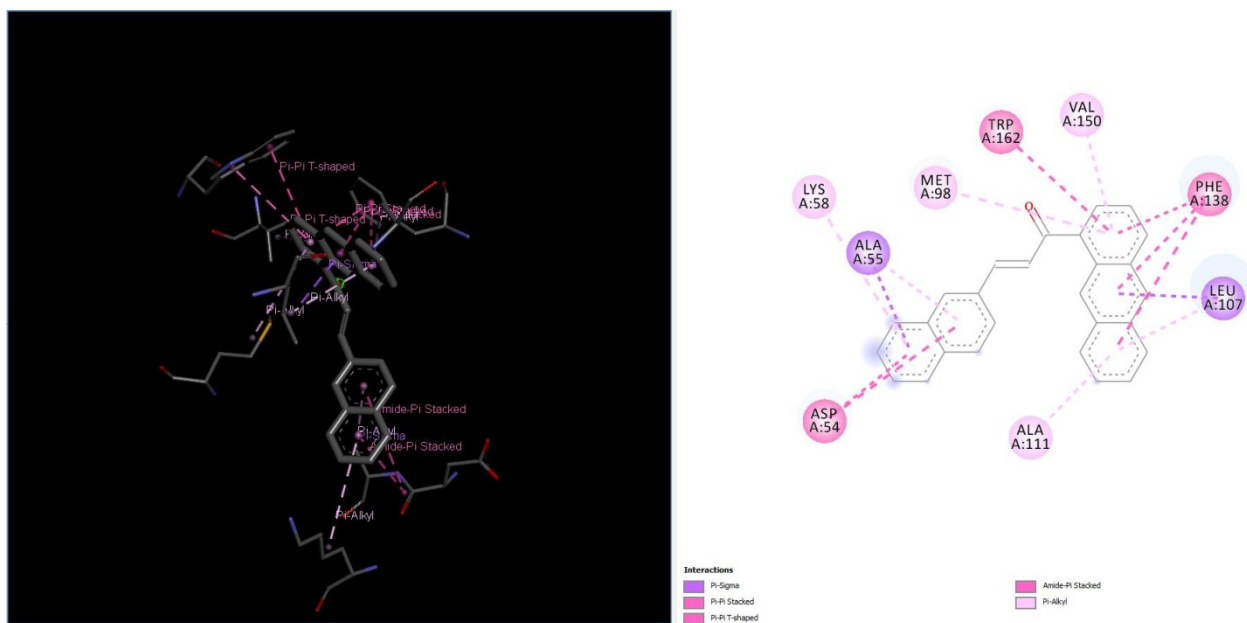
Figure 5.5: 3D (*left*) and 2D (*right*) interaction maps between acetophenone_22_benzaldehyde_17 and HSP90 crystal structure.

Figure 5.5 represents the 2D and 3D interaction map between the original ligand acetophenone_22_benzaldehyde_17 and the HSP90 protein. There are four π-π interactions including involvement by TRP162 and PHE138 residues concentrated on the anthracene ring of this chalcone. Interesting is the π-σ interaction between ALA55 and the naphthalene ring.
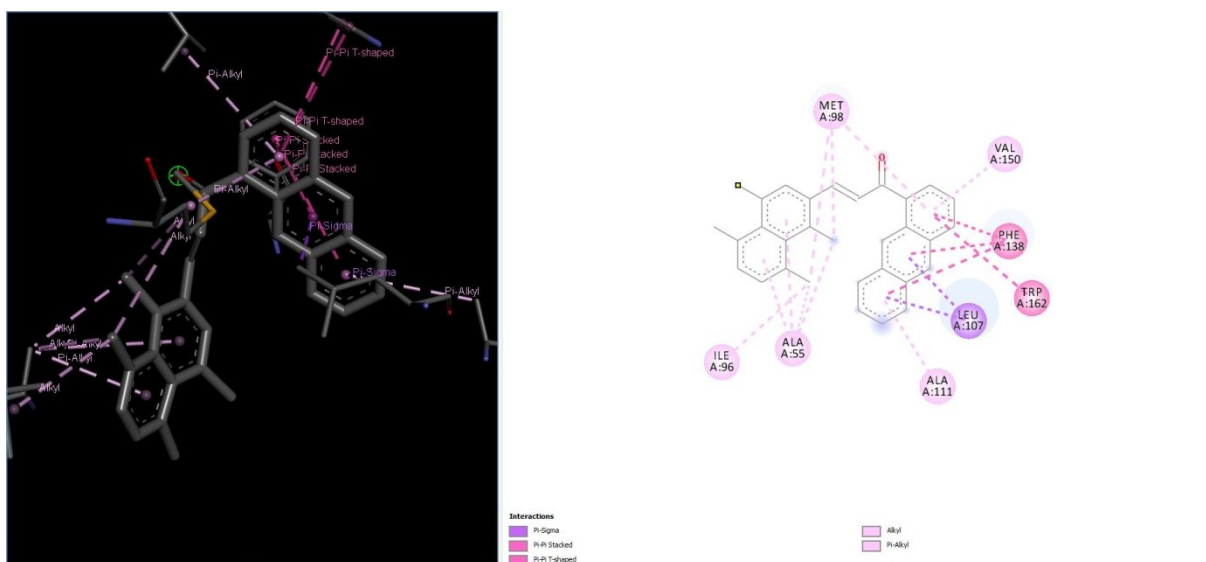


Figure 5.6: 3D (*left*) and 2D (*right*) interaction maps between the acetophenone_22_benzaldehyde_17_Cl_Cl_Cl_Cl and HSP90 crystal structure. The four chlorine substituents are attached to the naphthalene system.

Figure 5.6 shows the interactions in binding of the modified system. Although there appear to be fewer interactions to the naphthalene system, some key interactions are maintained, including the anthracene π-π interactions and the π -σ interaction between ALA55 and the

174

naphthalene ring. This reduction in interactions on the naphthalene subsystem may be a result of the high electronegativity of the chlorine compared to the unsubstituted parent ligand. This reduction in interaction mirrors the rotation of this system observed during binding, compared to the parent ligand.

Although the modifications of the parent ligand may have a positive effect on its physico-chemical properties and its drug likeness, the aim of this section was to improve binding. This was not successful, but this does not preclude that this may be possible with other ligands in the chalcone library. Further it paves the way for extension of the library of 2-3 orders of magnitude. Further work may involve the use of software that correctly describes and models halogen bonding, given the extensive use of halogens in our modification of the original ligand.

## 2. Confirmation of the important role of the radial distance inside the chalcone molecules for the binding energy with the pyruvate kinase.

### 2.1. Design of the library

In the statistical analysis of the performance of the chalcone library against a range of targets, an interesting observation, unique to the pyruvate kinase AB target was that radial extents of key atoms in the chalcone were linked to the binding energy.

In order to further investigate and provide more evidence for this, it was necessary to modulate the radial distance of the acetophenone and the benzaldehyde-derived moieties of the chalcones. First, three acetophenones and three benzaldehydes were randomly selected, this subset may give rise to 9 chalcones present in our library. Their binding energies for these 9 chalcones to pyruvate kinase AB and C are presented in Table 5.3. The pyruvate kinase AB and C structures show, respectively, a variation from -7.7 to -9.2 kcal.mol$^{-1}$ and from -7.4 to -8.9 kcal.mol$^{-1}$. The specific acetophenone (2, 37 and 7) and benzaldehyde (61, 7 and 88) precursors were the primary precursors for a new, focused library and are shown in Figure 5.7. These precursors are monocyclic (except for the benzaldehyde 88, which includes a naphthalene ring) and include a variety of substituents (methyl, hydroxyl and amino groups). All of the acetophenones and all of the benzaldehydes were already spatially aligned (in the same plane) making generation of derivative 3-dimensional structures straightforward.

| Chalcone names | Energy (kcal.mol$^{-1}$) | |
| --- | --- | --- |
| | Pyruvate kinase AB | Pyruvate kinase C |
| acetophenone_2_benzaldehyde_61 | -8 | -8 |
| acetophenone_2_benzaldehyde_7 | -8,1 | -7,9 |
| acetophenone_2_benzaldehyde_88 | -9 | -8,9 |
| acetophenone_37_benzaldehyde_61 | -8,1 | -7,7 |
| acetophenone_37_benzaldehyde_7 | -7,7 | -7,4 |
| acetophenone_37_benzaldehyde_88 | -8,7 | -7,7 |
| acetophenone_7_benzaldehyde_61 | -8 | -7,6 |
| acetophenone_7_benzaldehyde_7 | -8 | -7,8 |
| acetophenone_7_benzaldehyde_88 | -9,2 | -7,8 |

Table 5.3: Molecules retained for the extended library with their binding energies for the pyruvate kinase AB and C domains.
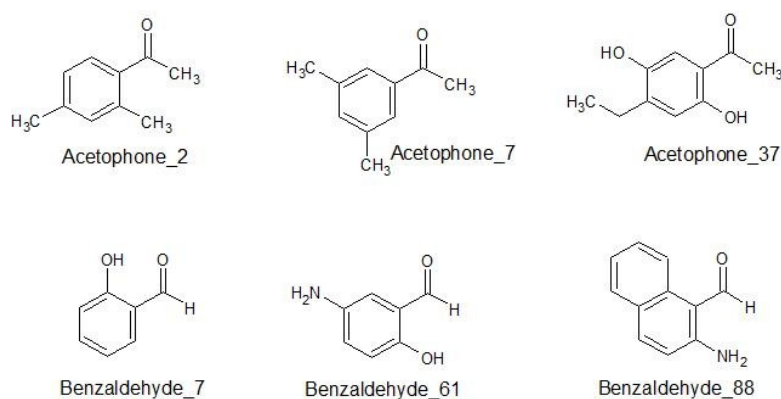


Figure 5.7: Acetophenone and benzaldehyde derivates selected for the new library.

However, for the focused library, in order to change to systems where radial distance varies, a third series was introduced in the form of an unsaturated poly-ene chains of various lengths which linked the acetophenone-based to the benzaldehyde-based moieties. We modulated this chain by the incrementing of one carbon per step and we alternated by using an unsaturated bond (alkene chain) and ketone function in order to maintain conjugation though the system with unbroken chains of sp$^2$ carbon atoms. From 0 atoms initially (no atoms were placed between acetophenone and benzaldehyde-based moieties) the length of the linker was varied up to 6 atoms. All linkers were superimposed in the same manner as for the alkyl chain described in Chapter 2. Figure 5.8 represents the linker selected for the extended library.
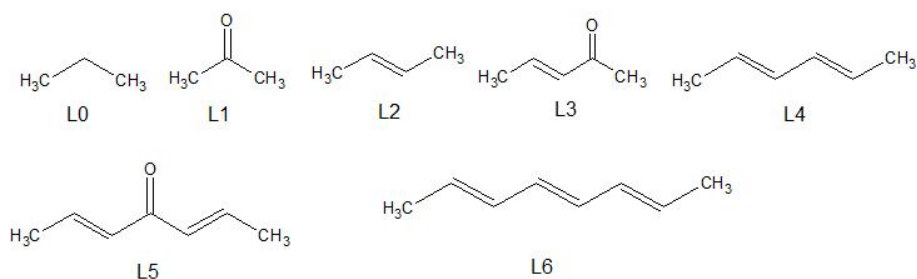
Figure 5.8: Linkers selected for the extended library. The number corresponds to the number of atoms between the two terminal methyl groups. L =linker.

A script was written, using Python, in order to merge one acetophenone system, one benzaldehyde and one linker to form one unique molecule. This required some modification to the process described in Chapter 2 to produce this focused library. After reading in the respective pdb files, the Cartesian coordinates of the first carbon of the linker and of the benzaldehyde moiety were identified. A vector was then calculated from the variation between these two coordinates; this vector provided a reference for appropriate translation in 3d space of the benzaldehyde moiety, from which superimposition of all three sections was possible, followed by deletion of appropriate atoms in each case to form the focussed library. The pseudo-code for library generation is displayed in Figure 5.9.

```
assign variable linker_coordinate
assign variable benzaldehyde_coordinate
assign variable vector

for linker_file in linker_directory:
    if linker_file.endswith(".pdb")
        for benzaldehyde_file in benzaldehyde_directory:
            if benzaldehyde_file.endswith(".pdb")
                for acetophenone_file in acetophenone directory:
                    if acetophenone_file.endswith(".pdb")
                        #preparation of the name files
                        molecule_file=linker_file"_"acetophenone_file"_"benzaldehyde_file
                        read(linker_file,benzaldehyde_file,acetophenone_file)
                        #calculation of translation vector
                        extract_coordinates(linker_file, "C0") as linker_coordinate
                        extract_coordinates(benzaldehyde_file, "C2") as benzaldehyde_coordinate
                        calculate_vector(linker_coordinate,benzaldehyde_coordinate) as vector
                        #writing library compound file
                        open new molecule_file for writing
                        for aline in acetophenone_file:
                            if(valid(aline)): #omit deleted atoms
                                write(aline,molecule_file) #write all remaining acetophenone atoms
                        for lline in linker_file:
                            if(valid(lline)): #omit deleted atoms
                                write(lline,molecule_file) #write all remaining linker atoms
                        for bline in benzaldehyde_file:
                            if(valid(bline)): #omit deleted atoms
                                translate(bline,vector) #translate atoms by vector

                                write(bline,molecule_file) #write all remaining benzaldehyde atoms

                        close(molecule_file)
```

Figure 5.9: Pseudo code used for the extended library generations.

The 63 molecules of this new library were submitted to geometric optimization with Gaussian 09 at the semi-empirical AM1 level. Again scripts were used firstly to convert the original pdb files to Gaussian ".com" format, and appropriate directives edited in prior to optimization. Further scripting extracted the Gaussian ".log" files to pdb format, and AutoDock Tools (prepare_ligand4.py) was used to prepare all ligands for docking.

## 2.2. Parameters

The new library was docked again against the pyruvate kinase AB and C co crystal structures (code PDB: 3T05) as the protein responsible for MRSA[3] which are described in previous chapters. AutoDock Vina was used to dock this library against this protein. The same parameters as described in Chapter 3 were used we continued to use a blind docking approach. The coordinates of the active site were retained. A script was written to submit jobs to the departmental cluster in the context of the job scheduler. Table 5.4 shows the parameters used for the docking experiment.

| Cristal structure | Cartesian Coordinates for Center | | | Size of the grid box | Energy range | Exhaustiveness | CPU |
|---|---|---|---|---|---|---|---|
| AB | 0.579 | 33.8 | 23.22 | 126/126/126 | 4 | 240 | 24 |
| C | 31.494 | 11.371 | 31.974 | 126/126/126 | 4 | 240 | 24 |

Table 5.4: Parameters used for the docking experiment.

Calculations were not submitted to the CHPC due to the small size of the library, and the departmental cluster provided results in adequate time. Again, binding energy results were extracted through the use of custom python scripts and the results were adequately plotted on a graph using Excel.

## 2.3. Principal Component Analysis codes updated with R

The PCA code is useful to visualize variables for all entries; however, locating specific chalcones inside of the different clusters was not possible. In order to solve the problem, the code described in the previous chapter was updated. The individual PCA map is composed of 4 clusters (A, B, C and D). A, B, C and D are respectively situated at ]-∞; 2.5], [2.5; 5], [5; 10] and [10; +∞[ following the dimension 1 axis. A code line including the function which produced all individual compounds which are members of the cluster following their domains

was used. These lines are implemented in the original code. The new program is presented in Figure 5.10. The PCA remains the same and is presented in Chapter 4.

```
set the default directory
read the csv file of the data
normalize the data file
correlation analysis of the data
perform a principle component analysis of the data
extract and visualize the eigenvalue and the variance of dimension and plot the percentage explained of
variance for the PCA
plot the graph of the pca variables with quality indicator of the data
plot the graph of the pca individual with contribution indicator of the data
create a group A with data which values are lowest or equal to 2.5 in the x-axis
create a group B with data which values are between to 2.5 and 5 in the x-axis
create a group C with data which values are between to 5 and 9.5 in the x-axis
create a group D with data which values are highest or equal to 10 in the x-axis
visualize the group A, B, C, D
```

Figure 5.10: Pseudo-code for the separation of the clusters from the individual PCA map and the visualization of the entries which were composed for RStudio.

## 2.4. Results

### 2.4.1. Binding energy analysis

The variation of the binding energy for this set of structures varying linker length was found to be in the range -7.1 to -8.6 kcal.mol$^{-1}$ for the pyruvate kinase C and between -7.3 to -9.5 kcal.mol$^{-1}$ for the AB model. Table 5.5 presents the results of the binding, ordered by parent benzaldehyde/acetophenone and linker length in terms of the number of carbon atoms. Interestingly, the binding energy of each complex for the pyruvate kinase closely matches the binding energy of their chalcone reference (linker length 3) as shown in Table. 5.1. However with only four exceptions (acetophenone_2_benzaldehyde_88, acetophenone_7_benzaldehyde_7, acetophenone_37_benzaldehyde_7 and acetophenone_37_benzaldehyde_61), the binding energy decreases (improves – greater negative) when the atom count in the linker reaches 3. This is marked in the binding to pyruvate kinase C. Increasing the linker to 4 atoms again raises the energy, and the next decrease (improvement) is when the linker length reaches 5 carbon atoms. The improvement in binding for 5 atom linkers, especially when the molecule is derived from acetophenone_2, is marked. The binding energies of the compounds derived from acetophenone_7 and acetophenone_37 weakly decrease and increase respectively. The observed effects are less dramatic in the binding to pyruvate kinase structure AB. The acetophenone_2_benzaldehyde_88 derived system with a linker of 5 carbons binds best in this set against both pyruvate kinase structures. Also, by consideration of the specificity for the AB domain, the same derivation of acetophenone_2_benzaldehyde_88 with a linker of 3

carbons shows some specificity for this over the C model. By a small margin the best single docking complex is with this system with linker of 5 atoms, to the pyruvate kinase C structure. To a small extent one can therefore, not only alter binding energies, but also tune the selectivity of target by changing the linker length.

| Pyruvate kinase C | Binding energy (kcal.mol$^{-1}$) | | | | | | |
|---|---|---|---|---|---|---|---|
| Name of the molecule | Number of linker carbon | | | | | | |
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| acetophenone_2_benzaldehyde_7 | -7,6 | -7,8 | -7,8 | -8,6 | -7,9 | -8,6 | -7,5 |
| acetophenone_2_benzaldehyde_61 | -7,9 | -8 | -8,2 | -8,5 | -8,1 | -8,6 | -7,4 |
| acetophenone_2_benzaldehyde_88 | -8,1 | -8,1 | -8,6 | -7,8 | -8,3 | -9,6 | -8,3 |
| acetophenone_7_benzaldehyde_7 | -7,1 | -7,3 | -7,6 | -7,7 | -7,2 | -7,6 | -7,6 |
| acetophenone_7_benzaldehyde_61 | -7,2 | -7,1 | -7,6 | -7,7 | -8 | -8 | -7,8 |
| acetophenone_7_benzaldehyde_88 | -7,9 | -7,8 | -7,8 | -7,9 | -7,5 | -7,8 | -7,5 |
| acetophenone_37_benzaldehyde_7 | -7,4 | -7,3 | -7,6 | -7,1 | -7,1 | -7,4 | -7,4 |
| acetophenone_37_benzaldehyde_61 | -7,3 | -7,5 | -7,3 | -7,3 | -7,5 | -7,4 | -7,4 |
| acetophenone_37_benzaldehyde_88 | -7,9 | -7,8 | -7,8 | -7,9 | -7,9 | -7,6 | -7,5 |

| Pyruvate kinase AB | Binding energy (kcal.mol$^{-1}$) | | | | | | |
|---|---|---|---|---|---|---|---|
| Name of the molecule | Number of linker carbon | | | | | | |
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| acetophenone_2_benzaldehyde_7 | -7,4 | -7,8 | -8 | -8,5 | -7,9 | -8,1 | -8,3 |
| acetophenone_2_benzaldehyde_61 | -7,9 | -7,9 | -8,3 | -8,4 | -8,1 | -8,3 | -7,9 |
| acetophenone_2_benzaldehyde_88 | -9,2 | -9,3 | -9,1 | -9,3 | -9,1 | -9,5 | -9,1 |
| acetophenone_7_benzaldehyde_7 | -7,3 | -7,6 | -7,8 | -7,7 | -7,8 | -8 | -8 |
| acetophenone_7_benzaldehyde_61 | -7,8 | -7,8 | -8,3 | -8,2 | -8,1 | -8,3 | -8 |
| acetophenone_7_benzaldehyde_88 | -8,7 | -9,1 | -9,1 | -9,1 | -9 | -9,1 | -9,4 |
| acetophenone_37_benzaldehyde_7 | -7,6 | -7,6 | -7,6 | -7,6 | -7,5 | -7,8 | -7,8 |
| acetophenone_37_benzaldehyde_61 | -7,6 | -7,5 | -7,8 | -7,7 | -7,5 | -8 | -7,9 |
| acetophenone_37_benzaldehyde_88 | -8,8 | -8,9 | -8,8 | -9 | -8,8 | -8,6 | -9 |

Table 5.5: Binding energy (expressed in kcal.mol$^{-1}$) of molecule per number of carbons present inside of the linker for the pyruvate kinase C domain (*left*) and AB co-crystal structures (*right*).
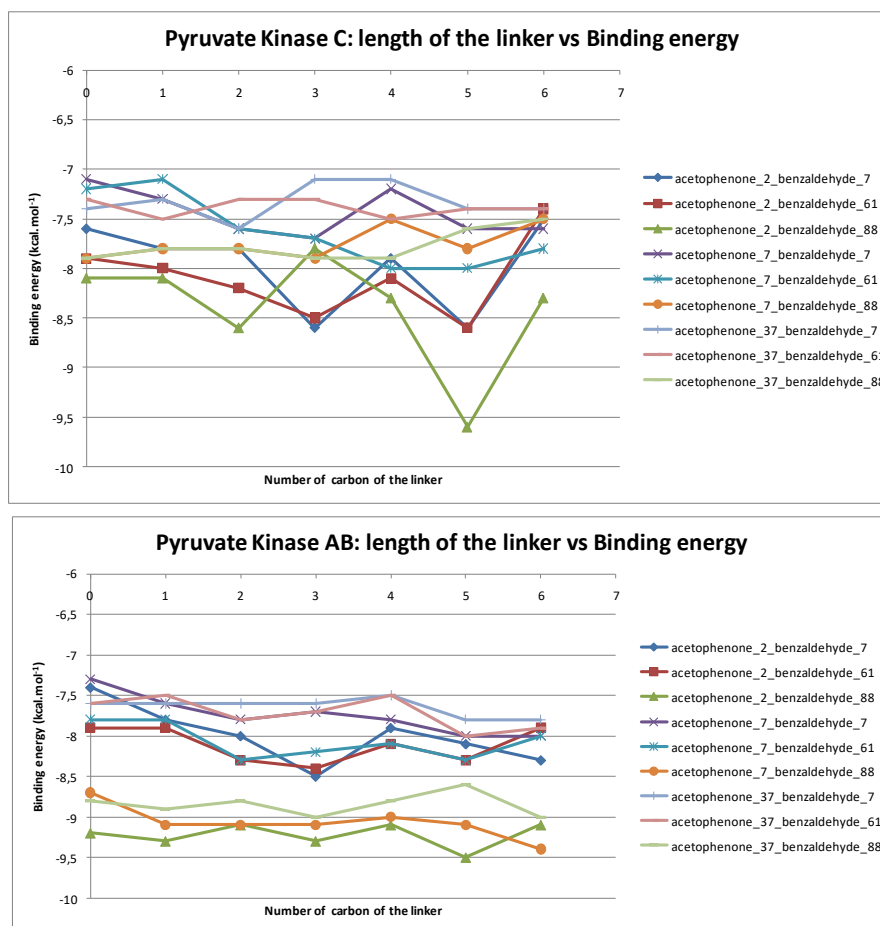
Figure 5.11: Evolution of the binding energy (expressed in kcal.mol$^{-1}$) with the number of carbons of the linker per molecule for the pyruvate kinase C (*top*) and AB (*down*) co-crystal structure.

For better visualisation, graphs were plotted to represent the evolution of the binding energy according to the number of carbon atoms in the linker, per system binding to the pyruvate kinase C and AB co-crystal structures. The graphs, shown in Figure 5.11, confirm (for most systems) that when the count of linker atoms equals 3 or 5 the observed binding energy decreases.

Figure 5.12 illustrates the docking of all of the acetophenone_2_benzaldehyde_88-derived systems within the pyruvate kinase C structure. All molecules were localized inside of the same area of the protein which corresponded to the active site of pyruvate kinase.[3] All of the acetophenone_2 moieties (even with the different linkers) appeared to be localized on a single axis while the benzaldehyde moieties appeared to have the freedom to rotate to many orientations within the binding site.

After analysing the poses of the different acetophenone_2_benzaldehyde_88-derived systems binding inside the pyruvate kinase AB model, and, informed by the literature, the following observations were made. Acetophenone_2_benzaldehyde_88-derived systems possessing linker lengths of 1, 3 and 6 carbons were found to dock in the effector site of pyruvate kinase B co-crystal structure while the system with a linker of 5 carbons dock within the active site of the A co crystal structures.[3] The model with a linker of 0 and 2 carbons was located in the effecter site of the A structure. The model with a length of 4 carbons in the library appeared to dock to random positions. Figure 5.13 presents the different orientations of the acetophenone_2_benzaldehyde_88 systems of different linker-length within the pyruvate kinase AB model.

The analysis shows a trend between the molecules wherein the best energy was seen when the new linker chain was comprised of 5 carbons. With the original chalcone library major variation in these radial position measurements would have to be due to the relative positioning of ligands. Yet this has informed this focused library in which radial distances are "forced", and this has confirmed the hypothesis that these radial distance positions do to have an impact on the binding energy. The effect of radial position was seen for docking of the focussed library to both AB and C co crystal structures. However, since the docking to structure C is consistently within the active site, while the docking to AB model varies between the effector and active site, and random positions, it is not possible to confirm the important role of the radial distance on binding energy against the pyruvate kinase AB co crystal structure.

It was later observed from the PCA (Figure 5.14), the original ligands used for this experiment are members of the same cluster A inside of the scores plot of this PCA. They were identified by removing one cluster at a time if they are not included at least one of these molecules. Therefore the radial distance variable must be in this cluster.

Figure 5.12: Orientations of the different acetophenones_2_benzaldehyde_88 systems docked to the pyruvate kinase C structure. All systems dock to the active site during blind docking.
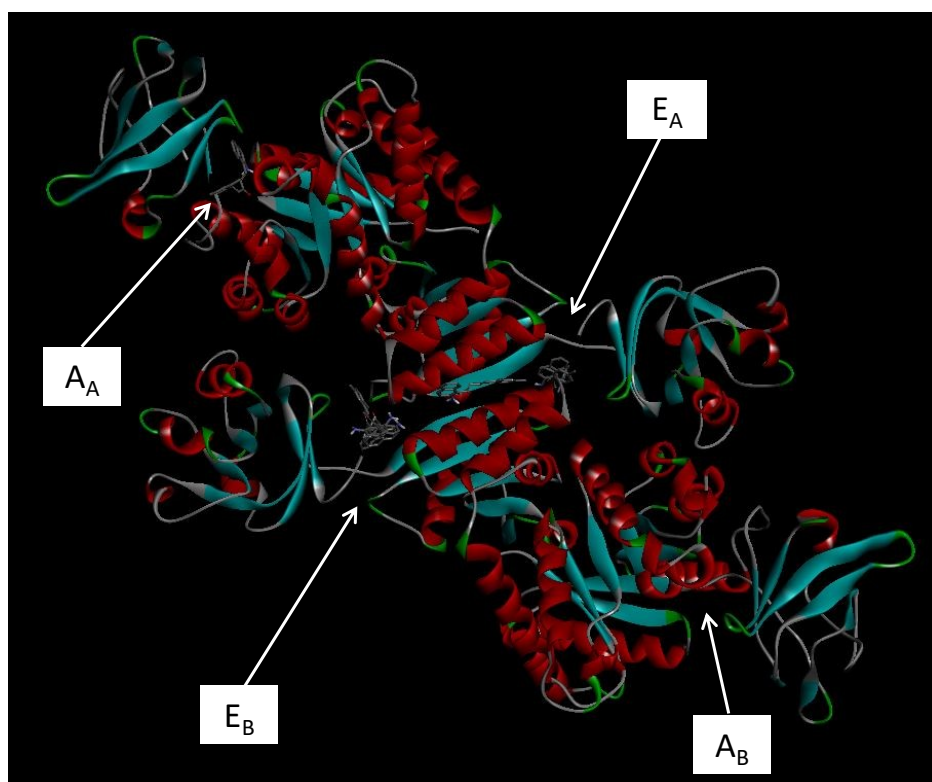


Figure 5.13: Orientations of the different acetophenones_2_benzaldehyde_88 systems inside of the pyruvate kinase AB co-crystal structure. AA, active site of model A; AB, active site of model B; EA, effector site of model A; $E_B$, effector site of model B.
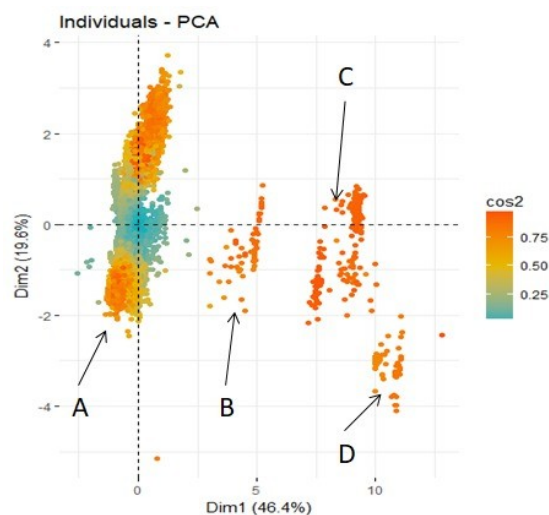
### 2.4.2. Interaction ligand-protein analysis

Acetophenone_2_benzaldehyde_88 is the most promising system (when the linker atom count is equal to 3 or 5, L3 and L5) for binding the pyruvate kinase C structure. However, on the other hand, in order to attempt to understand why these molecules docked in different locations for the pyruvate kinase AB model, it was decided to compare their interactions with Discovery Studio Visualizer. The 3D and 2D interaction map between L3 and L5 acetophenone_2_benzaldehyde_88 for the pyruvate kinase C and AB are represented in Figures 5.15 to 5.18 and Figures 5.19 to 5.22 respectively.

The analysis of the interaction between the L3 and L5 acetophenone_2_benzaldehyde_88 and the pyruvate kinase C co crystal structure showed that the two molecules had in common a $\pi$-anion interaction with GLU71 residue and the naphthalene ring of the benzaldehyde-derived moiety. Both also formed a hydrogen bond from the amine group present to LYS68. Further, the methyl group in the *para* position of the acetophenone-derived aromatic system is linked by alkyl and $\pi$-alkyl interaction to the ILE72, LEU160, PHE193 and VAL165 amino acid residues. In short, for both L3 and L5 systems there are a plethora of hydrophobic interactions between the ligand and protein. There are minor variations, such as the second hydrogen bond forming between the second proton of the amine system and SER36 for L5, or the additional hydrophobic interaction with an *ortho*-methyl substituent and PHE193 for L3. Perhaps the two most interesting of these variations are the hydrogen bond that forms between SER192 and the carbonyl of L5, and the $\pi$-$\pi$ stacked interactions between PHE193 and the benzene ring in L5. Neither of these, hydrogen bond nor the $\pi$-$\pi$ stacking interaction are observed with

the L3 system. It is likely that these two effects account for the improved binding in L5 over L3.

π-π stacking is a short-distance interaction with weak interaction energy, while carbonyl hydrogen bond interactions are medium distance interactions with average interaction energies respectively. The two are synonymous with recognition and orientation of molecules generally within the active sites of proteins.[4] The increase in the length of the linker from L3 to L5 atoms has not only positioned the carbonyl oxygen such that a hydrogen bond is possible, but has extended the reach of the benzene ring such that the π-π stacking interaction is also possible. Given the greater energetic value, it is of interest to explore the potential role of the SER192 in ligand binding.



Figure 5.15: Pose of the acetophenones_2_benzaldehyde_88_L3 inside of the pyruvate kinase C structure.

Figure 5.16: Interaction between L3_acetophenone_2_benzaldehyde_88 and the protein pyruvate kinase C in 3D (*left*) and 2D map (*right*).



Figure 5.17: Pose of the acetophenones_2_benzaldehyde_88_L5 inside of the pyruvate kinase C structure.

For binding to the AB co-crystal structure, the situation was quite different. A comparison of the interaction analysis for acetophenone_2_benzaldehyde_88 L3 and L5 molecules was deemed not possible, since these bound to very different positions in the pyruvate kinase AB structure. As such, it was more interesting to analyse them individually, detailing their position within the domain and their protein interactions separately (Figures 5.19 and 5.20 for L3, and Figure 5.21 and 5.22 for L5).

The L3 acetophenone_2_benzaldehyde_88 complex with pyruvate kinase AB possessed 3 types of interactions: the π-anion interactions between the naphthalene ring and GLU460 and GLU475, the hydrogen bond between the amine and the carbonyl of VAL456, and the π-alkyl interaction between the two methyl groups on the benzene ring, the *ortho* methyl group interacts with LYS576 and PHE578, while the *para* methyl group interacts with LEU471, LEU355 and LEU449 (again a plethora of hydrophobic interactions).

Also of interest is the high similarity of the interaction between the L5-acetophenone_2_benzaldehyde_88 and both pyruvate kinase models (AB and C). With the exception of an additional hydrogen bond interaction with the amine, the types of interactions and amino acid residues involved are the same.

After further examination of the literature, L5_acetophenone_2_benzaldehyde_88 was confirmed to be docked within the active site for the pyruvate kinase A substructure AB

model (Figures 5.21 and 5.22). This contrasts with all molecules of the focussed library docking to the active site for pyruvate kinase C.[3]

We confirm from this that there is a link between the radial positions of key points on the chalcones or the focussed derivative, and the calculated binding energy. Further, chalcones active against pyruvate kinase may well have dual action, with primary action at the active site, while for some it is possible that action is also at the allosteric site of this enzyme.
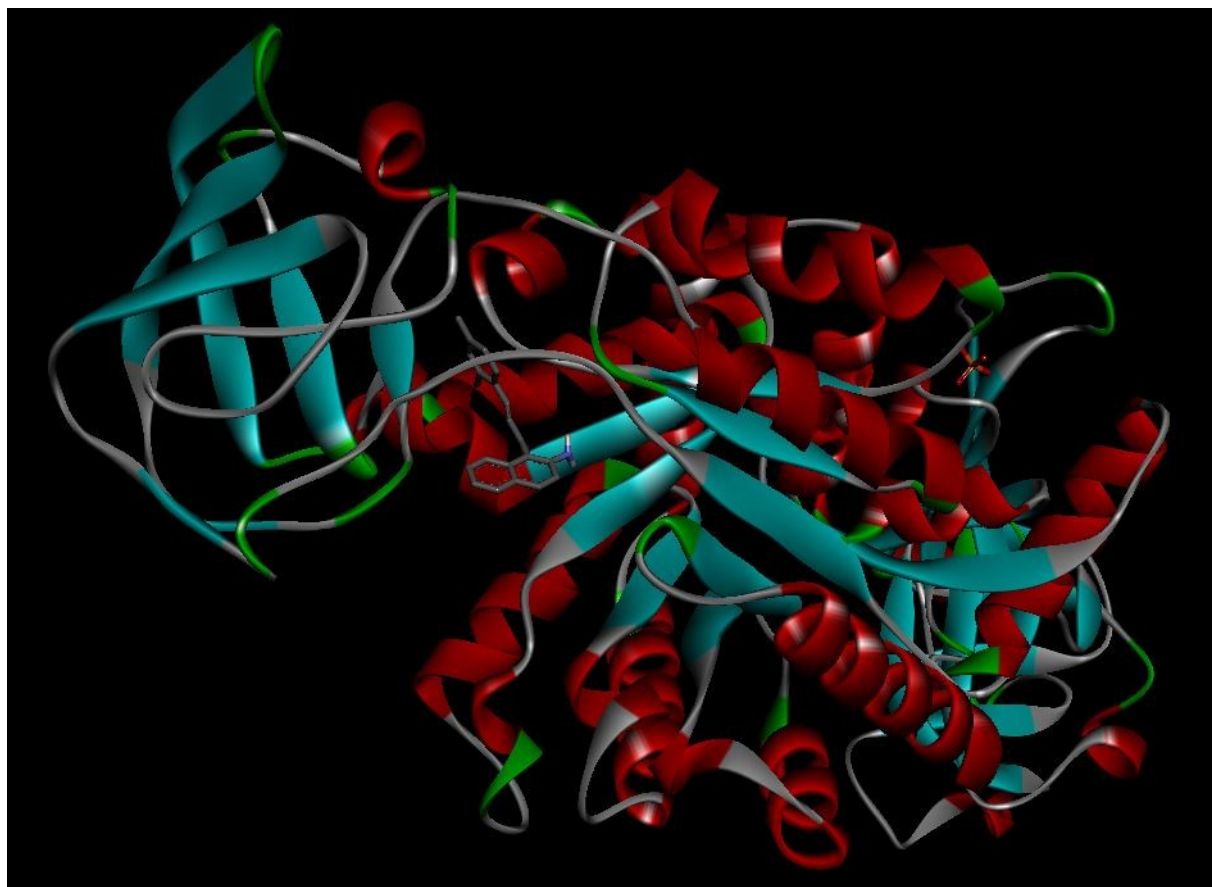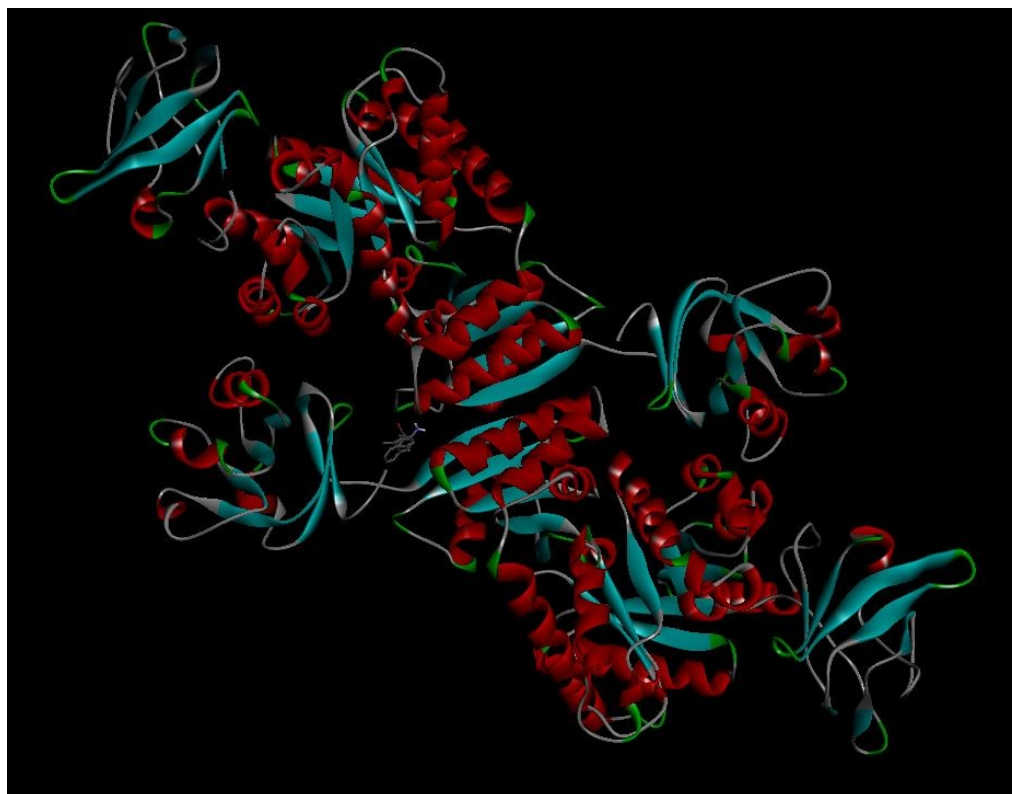


Figure 5.19: Pose of the acetophenones_2_benzaldehyde_88_L3 inside of the pyruvate kinase AB co-crystal structure.
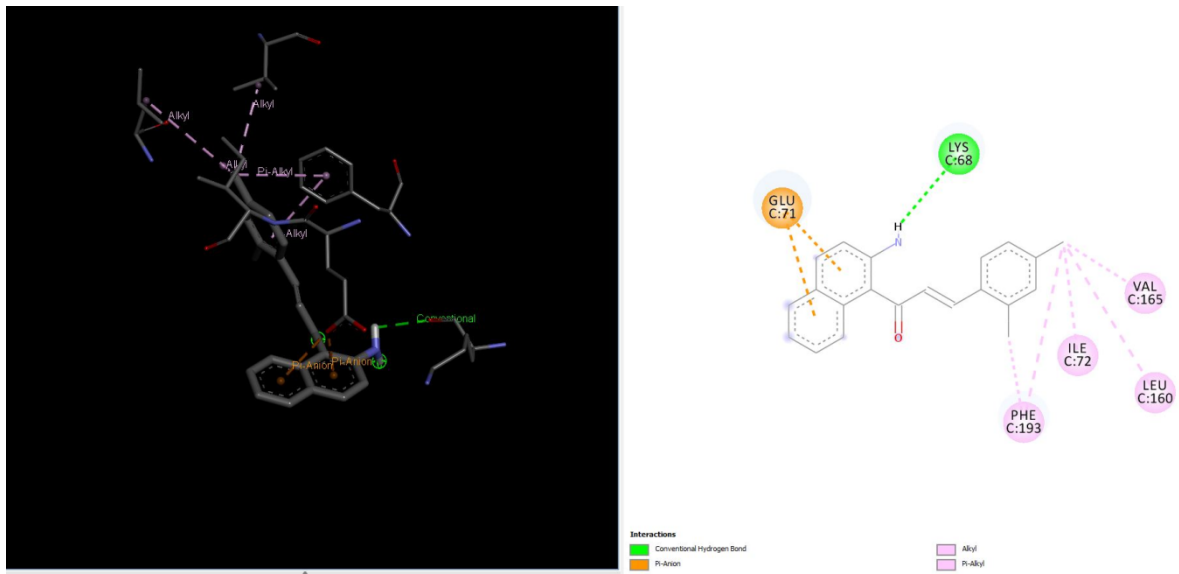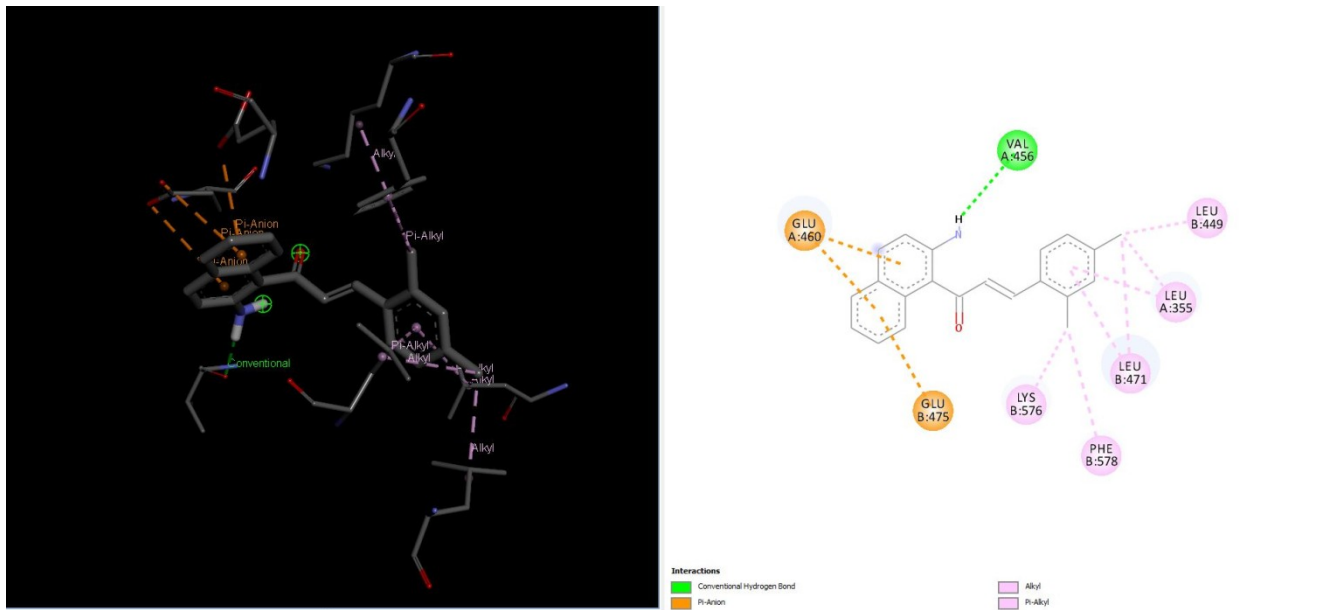
Figure 5.20: Interaction between L3_acetophenone_2_benzaldehyde_88 and the protein pyruvate kinase AB in 3D (*left*) and 2D map (*right*).
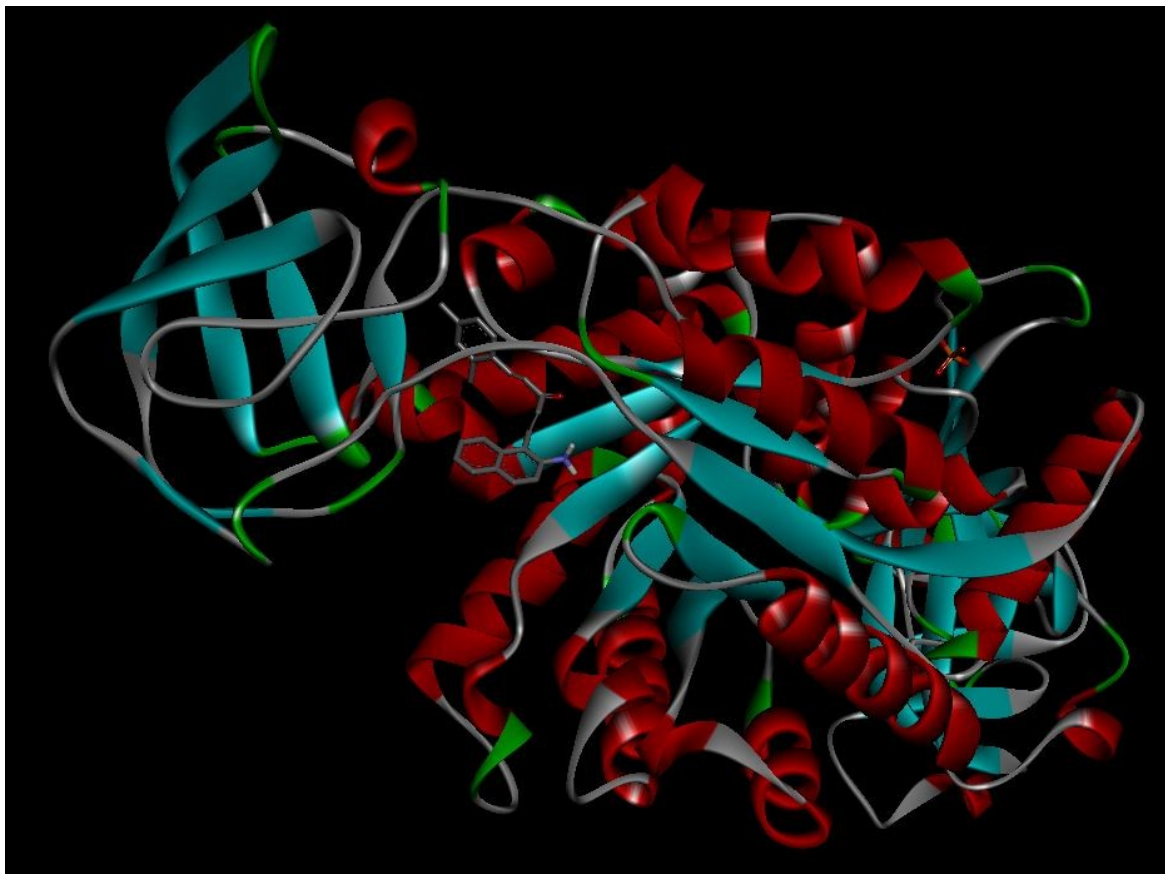
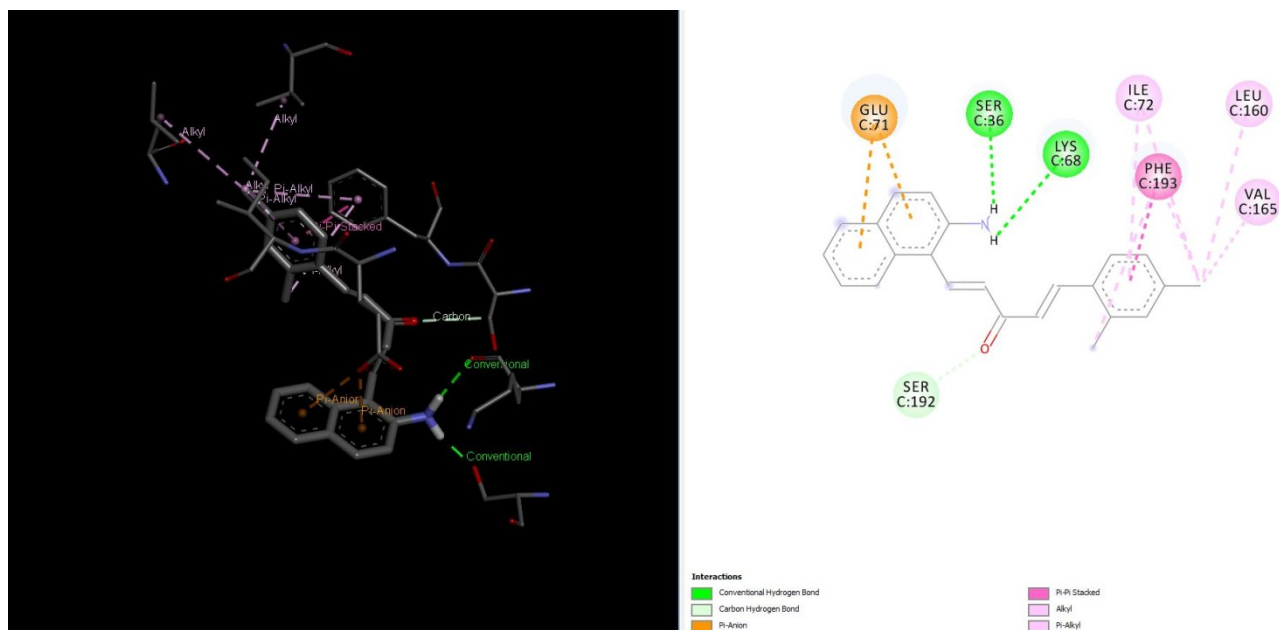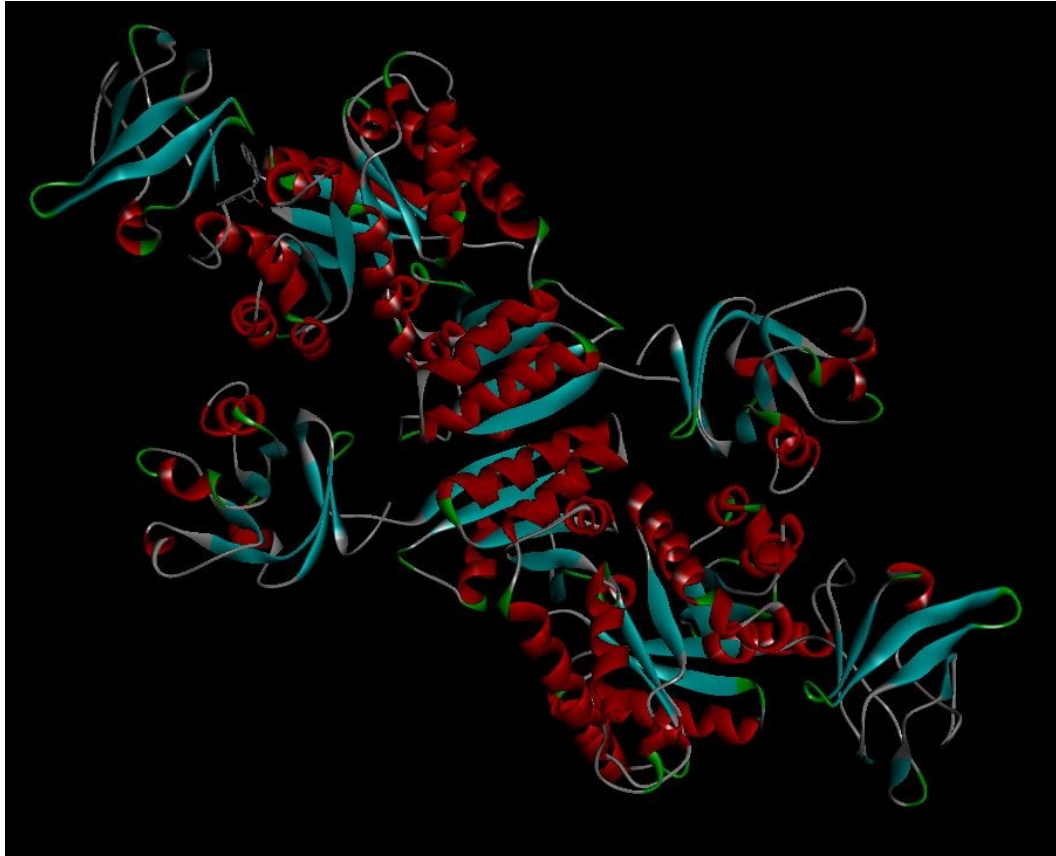Figure 5.21: pose of the acetophenones_2_benzaldehyde_88_L5 inside of the pyruvate kinase AB co-crystal structure.
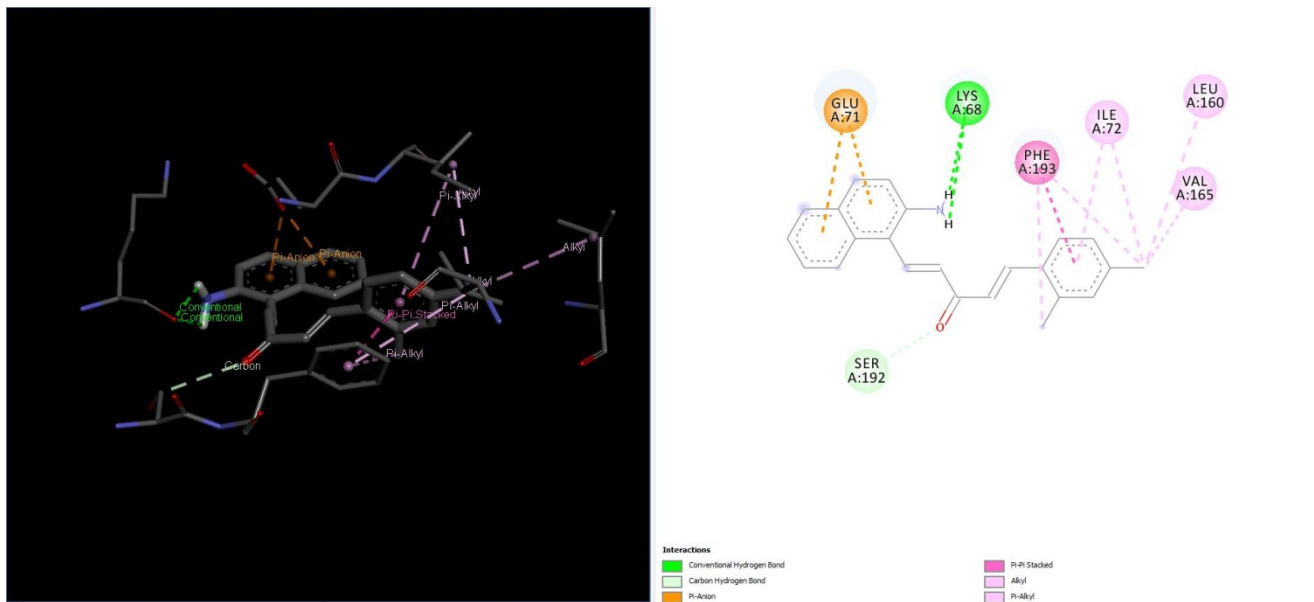


Figure 5.22: interaction between L5_acetophenone_2_benzaldehyde_88 and the protein pyruvate kinase AB in 3D (*left*) and 2D map (*right*).

## 3. Confirmation or refutation of the importance of the role of the polar angles of the chalcone molecules for the binding energy with the pyruvate kinase

### 3.1. Conformational analysis of chalcones

In order to better understand the conformational space available to chalcones when binding to the various enzymes, conformational analysis on a representative system was performed. What was of interest was not changes between s-*cis* or s-*trans* arrangements of the α,β-unsaturated system, but rather the rotation of the aromatic systems about the peripheral torsions of the linker region. The key interactions are with the α-hydrogen relative to the ketone and, the β-hydrogen and the ketone oxygen, which interact sterically with *meta*-hydrogen substituents as the aromatic groups rotate. It was decided to use only one of the best binding molecules (in terms of binding to pyruvate kinase C) as a base for this conformational analysis: the acetophenone_15_benzaldehyde_89 (-11.2 kcal.mol$^{-1}$). Two torsions were explored for 72 steps per full rotation (a resolution of 5° per step). Figure 5.23 represents the acetophenone_15_benzaldehyde_89 with the location of the two chosen dihedral angles $T_1$ and $T_2$ chosen for conformational exploration.
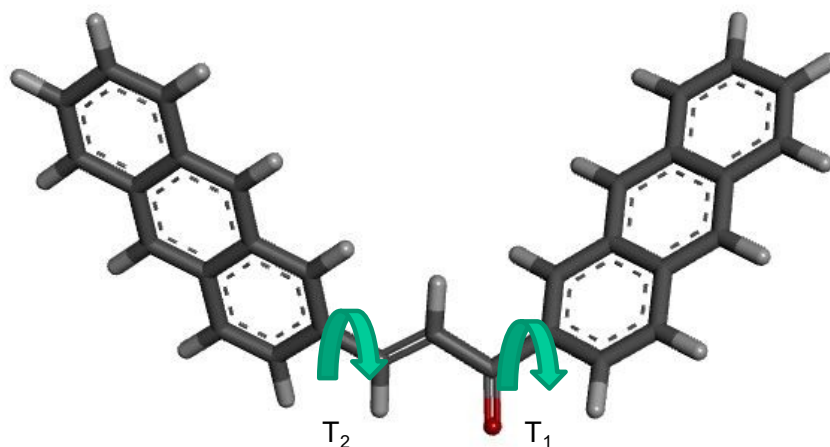


Figure 5.23: Acetophenone_15_benzaldehyde_89. The green arrows represent the dihedral angles. $T_1$; first torsion, $T_2$; second torsion.

For the s-*trans* configuration, the acetophenone_15_benzaldehyde_89 was already geometrically optimized as discussed in Chapter 2. The s-*cis* configuration of the same molecule was obtained through modification of the dihedral angle of the α,β unsaturated bond of the chalcone by 180°, prior to geometry optimization. The ammp module in VegaZZ was used to generate all 72x72 conformations, for each of the s-*cis* and s-*trans* configurations.

## 3.2. Results

The energies of the various s-*cis* and s-*trans* acetophenone_15_benzaldehyde_89 conformations were collated and matched, together with the corresponding pairs of dihedral angles, in order to generate two potential energy surfaces. From these two potential energy surfaces or heatmaps (s-*cis* and s-*trans*-configuration) the low energy conformations for the acetophenone_15_benzaldehyde_89 system were identified. The horizontal axis represents the angles of the $T_1$ axis (close to the acetophenone-derived group) and vertical axis represents the angles of the $T_2$ axis (close to the benzaldehyde derived side of the chalcone). On the potential energy surfaces, an arbitrary gradient color code was implemented, where red represents the high energy, unstable conformations while the more stable conformations are represented by green. Figure 5.24 represents the heatmap of the potential surface energy of the s-*cis*-configuration of the acetophenone_15_benzaldehyde_89. The potential surface energy varied from 87.5 kcal.mol$^{-1}$ to 185.1 kcal.mol$^{-1}$. The heatmap presents height zones of stability (represented in green). The central vertical band of high energy divides in the map into two groups (left and right) of which each have available eight minima. There is a characteristic band of high energy associated with rotation about $T_1$ which causes a meta-hydrogen atom of the benzaldehyde-derived subsystem to come unfavourably close to the α-hydrogen relative to the ketone. There is an energy barrier (vertically down the center of the potential energy surface) with a barrier of about 20 kcal.mol$^{-1}$, separating systems that would not be separable as rotamers at biological temperatures.



Figure 5.24: Heatmap of the potential surface energy for the s-cis configuration of the acetophenone_15_benzaldehyde_89. Row, dihedral angle close to the α-β unsaturated bond; Column; dihedral angle close to the carbonyl moieties; Red, high energy, green, low energy. (Angle and energy expressed respectively in degrees and kcal.mol$^{-1}$).

Eight conformations (corresponding to the minima on the left) were extracted, as were the eight minima on the right of the potential energy surface. The mirror image conformations

from the left and right of the potential energy surface are illustrated in Figure 5.26. In Table 5.6 and 5.7, provide the arbitrary name of each conformation, together with the angles of both torsions.

| Name | Dihedral angle close the α-β unsaturated bond (in degree) $T_2$ | Dihedral angle close to the carbonyl function (in degree) $T_1$ |
|---|---|---|
| 12-6 | -54.99697 | 149.45125 |
| 12-30 | -54.95369 | 29.43976 |
| 12-43 | -54.99178 | -35.59796 |
| 12-70 | -55.00555 | -170.52553 |
| 29-6 | -139.99260 | 149.43948 |
| 29-30 | -139.94824 | 29.48524 |
| 29-43 | -139.97908 | -35.53554 |
| 29-70 | -139.99328 | -170.56922 |

Table 5.6: Molecule name with their dihedral angles expressed in degrees from the left of the heatmap of the *trans*-configuration of acetophone_15_benzaldehyde_88.

| Name | Dihedral angle close the α-β unsaturated bond (in degree) $T_2$ | Dihedral angle close to the carbonyl function (in degree) $T_1$ |
|---|---|---|
| 48-6 | 125.03913 | 149.38891 |
| 48-30 | 125.00203 | 29.37468 |
| 48-43 | 125.03100 | -35.52408 |
| 48-70 | 124.98243 | -170.56913 |
| 63-6 | 49.98648 | 149.45751 |
| 63-30 | 50.03293 | 29.44074 |
| 63-43 | 50.04363 | -35.56504 |
| 63-70 | 49.95949 | -170.55945 |

Table 5.7: Molecule name with their dihedral angles expressed in degrees from the right of the heatmap of the *trans*-configuration of acetophone_15_benzaldehyde_88.
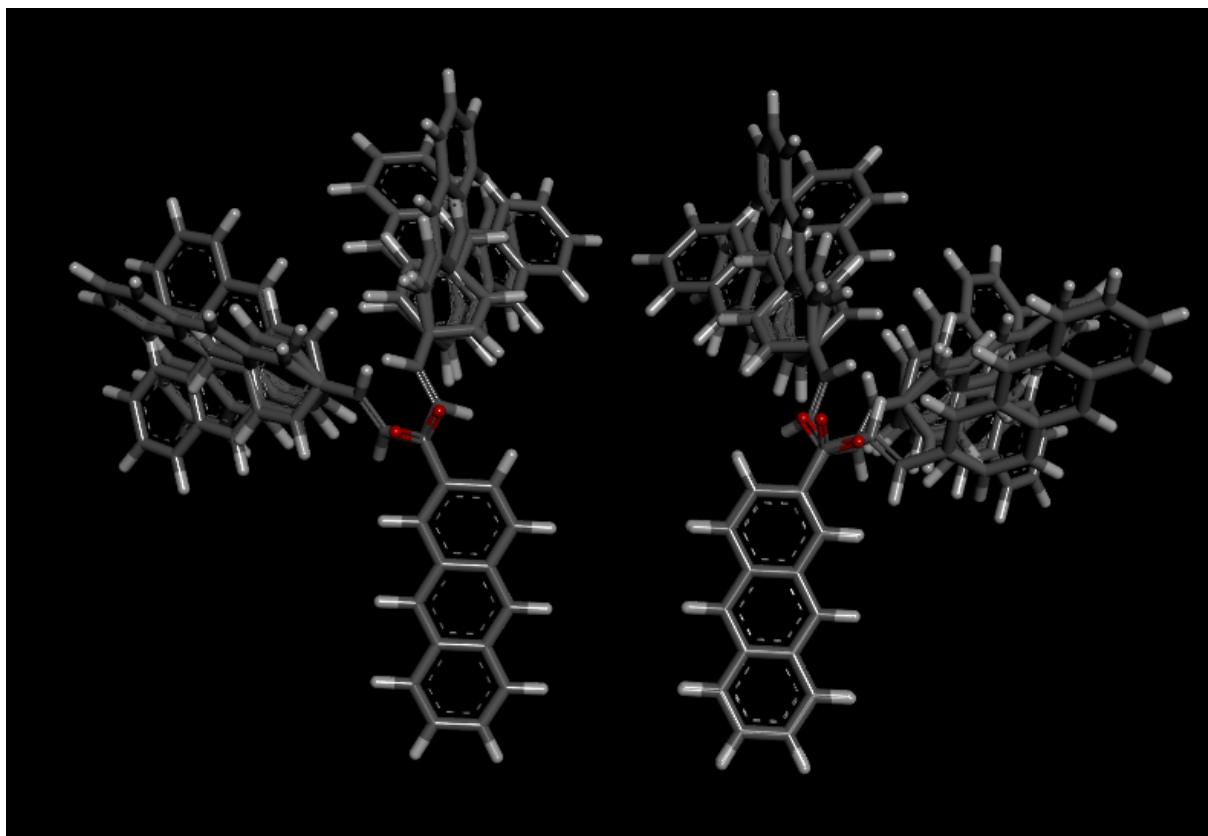
Figure 5.26: Enantiomeric conformations of s-*cis* acetophenone_15_benzaldehyde_89.

Figure 5.27 shows the origin of the band of instability at the left and the right of the potential energy surface, as being due to steric interactions between *meta*-hydrogen atoms and the α-hydrogen in the α,β-unsaturated system.
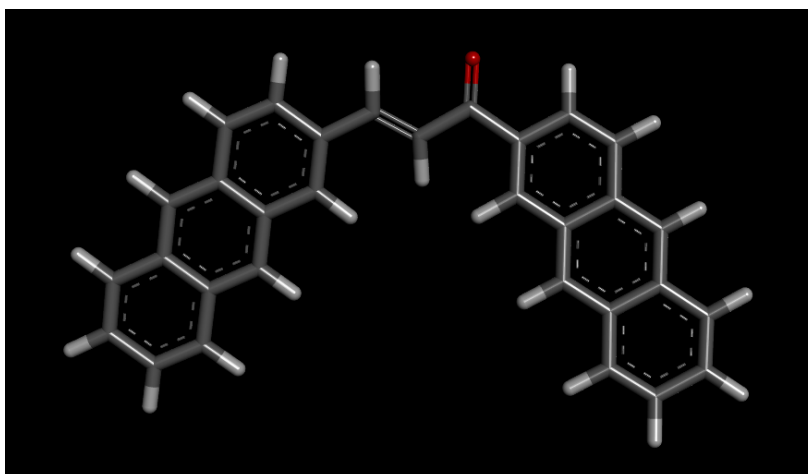


Figure 5.27: Origin of the high energy ridge in the potential energy surface for s-cis acetophenone_15_benzaldehyde_89, as due to a *meta*-hydrogen (right hand side) interacting with the α-hydrogen (center).

Figure 5.28 represents the heatmap of the potential surface energy of the s-*trans* configuration of the acetophenone_15_benzaldehyde_89. The potential surface energy varied from 91.2 kcal.mol$^{-1}$ to 400.0 kcal.mol$^{-1}$ (the maximum energy was truncated to 400.0 kcal.mol$^{-1}$). The heatmap presents zones of stability (represented in green). An imaginary horizontal line through the center may be drawn that divides the potential energy surface into two groups (top and bottom) which each are composed of four stable groups. Again $T_2$ (the torsion adjacent to the double bond) marks the *x* axis, while $T_1$ (adjacent to the carbonyl) marks the *y* axis. In this potential energy surface, it may be seen that conformations in the top half are separated from those in the bottom half by a barrier of about 50 kcal.mol$^{-1}$. This may indicate that rotamers are possible.
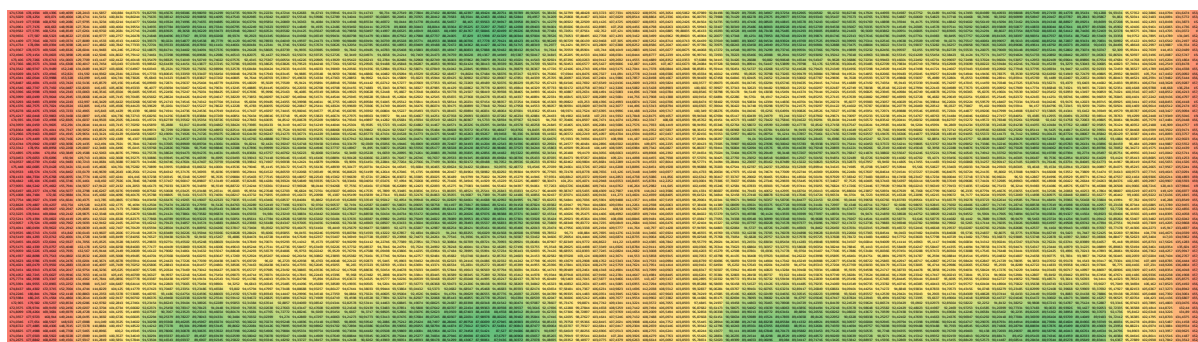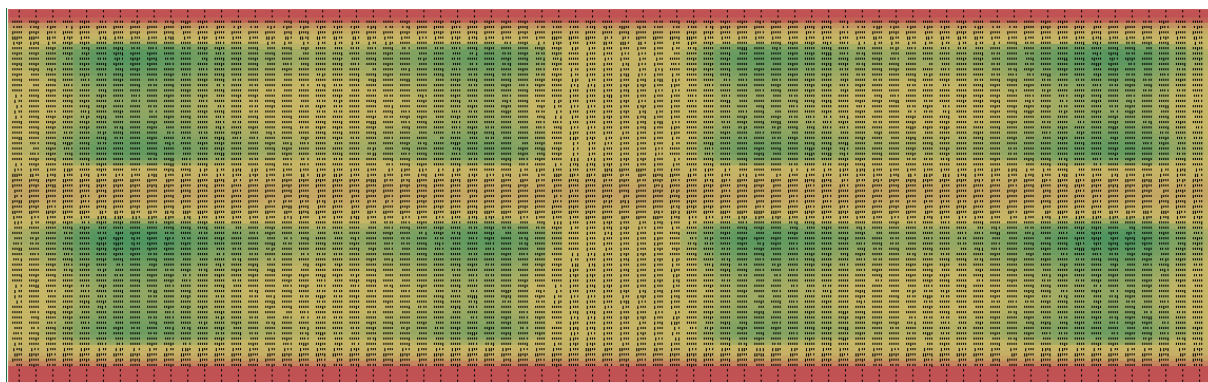


Figure 5.28: heatmap of the potential surface energy for the trans-configuration of the acetophenone_15_benzaldehyde_89. Row, dihedral angle close to the α-β unsaturated bond; Column; dihedral angle close to the carbonyl moieties; Red, high energy, green, low energy. (Angle and energy expressed respectively in degrees and kcal.mol$^{-1}$).

Eight minima were selected from the top half of the PES and from the bottom half of the PES as before. Figure 5.29 shows the enantiomeric rotameric conformations. Again, in Table 5.8 and 5.9, the arbitrary name of each conformation is provided following the step position with regards to rows and columns on the potential energy surface.
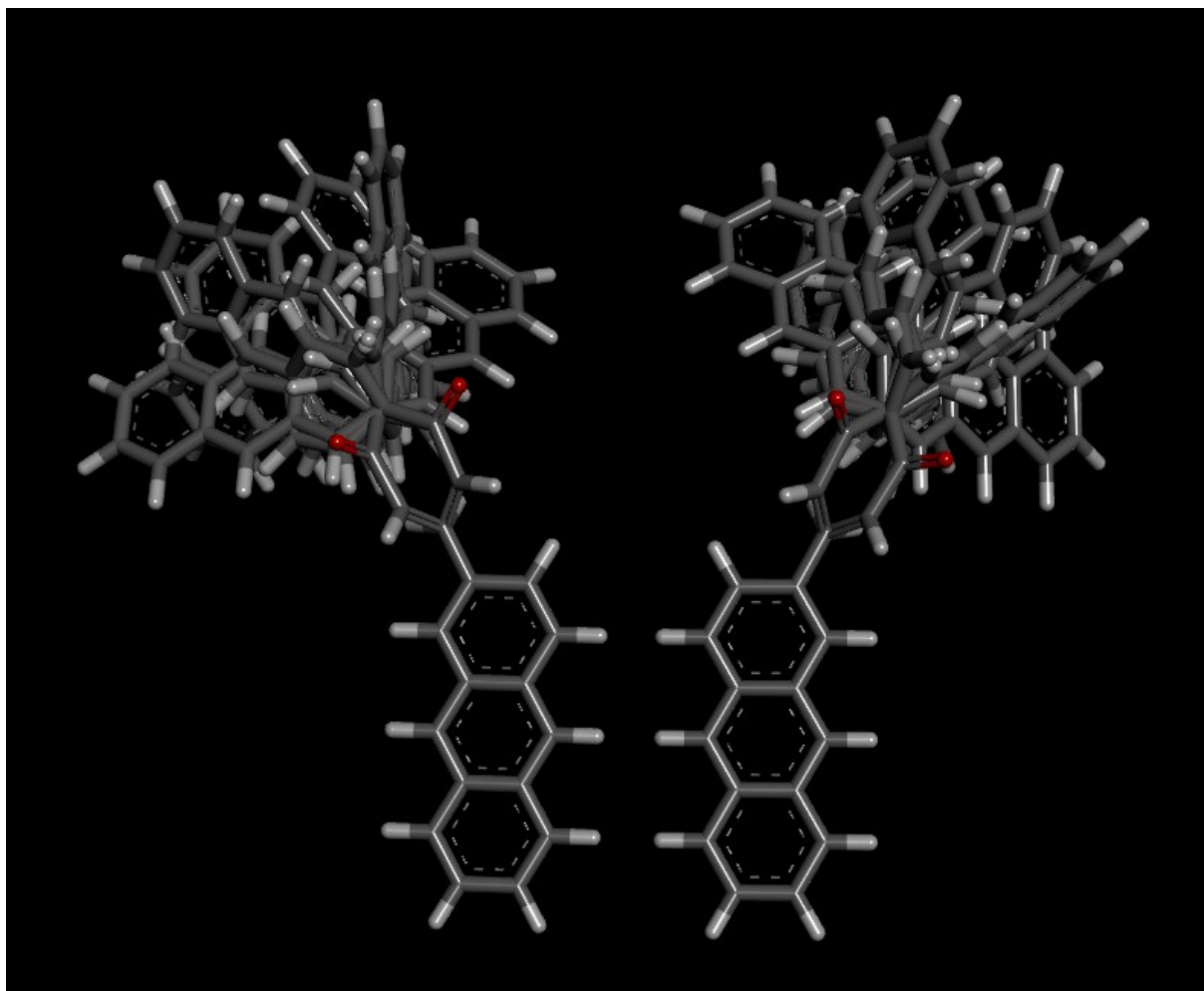
Figure 5.29: Overlay of s-trans acetophenone_15_benzaldehyde_89 conformations, separated as enantiomeric conformations with a 50 kcal.mol$^{-1}$ energy barrier separating them.

| Name | Dihedral angle close the α-β unsaturated bond (in degree) T$_2$ | Dihedral angle close to the carbonyl function (in degree) T$_1$ |
|---|---|---|
| 12-6 | -29.97399 | 125.13037 |
| 12-30 | -150.00576 | 125.04026 |
| 12-43 | 145.02735 | 125.1153 |
| 12-70 | 10.00286 | 125.10225 |
| 29-6 | -30.04241 | 40.08239 |
| 29-30 | -149.92571 | 40.03121 |
| 29-43 | 145.00601 | 40.13032 |
| 29-70 | 10.04145 | 40.07139 |

Table 5.8: Molecule name with their dihedral angles expressed in degrees from the left of the heatmap of the *trans*-configuration of acetophone_15_benzaldehyde_88.

| Name | Dihedral angle close the α-β unsaturated bond (in degree) T$_2$ | Dihedral angle close to the carbonyl function (in degree) T$_1$ |
|---|---|---|
| 48-6 | -29.95088 | -54.90733 |
| 48-30 | -149.9836 | -54.92896 |
| 48-43 | 144.97960 | -54.87250 |
| 48-70 | 10.06674 | -54.89566 |
| 63-6 | -30.00952 | -129.92286 |
| 63-30 | -149.99485 | -129.91703 |
| 63-43 | 145.00914 | -129.85965 |
| 63-70 | 9.97365 | -129.87760 |

Table 5.9: Molecule name with their dihedral angles expressed in degrees from the right of the heatmap of the *trans*-configuration of acetophone_15_benzaldehyde_88.

Figure 5.30 shows the origin of instability on the potential energy surface as due to unfavourable steric interactions between *meta*-hydrogen atoms of aromatic substituents and the β-hydrogen atom of the α,β-unsaturated system.
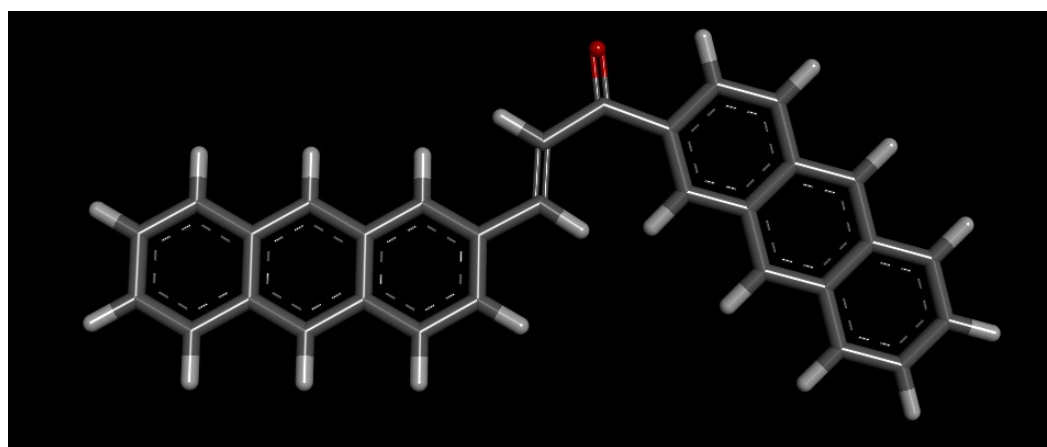


Figure 5.30: The origin of high energy conformations for s-*trans* acetophenone_15_benzaldehyde_89 with the steric interaction between a *meta*-hydrogen atom and the β-hydrogen of the α,β-unsaturated system.

# Conclusion

Modification of ligands and the formation of a library following the Topliss decision tree, in the case of binding to HSP90, did not improve binding at all. This may well be a peculiarity of this particular chalcone-HSP90 complex, with modification of the chalcone chosen disrupting a good fit to the active site.

From the statistical analysis of the previous chapter, a particular hypothesis was made regarding a radial distance of key points in a chalcone system (the key points being key atoms in the acetophenone and the benzaldehyde-derived moieties), and the impact of these distances on the binding energy of chalcone to pyruvate kinase C. To this end a new virtual library of compounds was designed and generated specifically perturbing this distance, by modulation of the length of the linker region. This library was docked against the pyruvate kinase AB and C co-crystal structures. The results showed an evolution of the binding energy relative to the length of the linker region. Low energy (good binding) was observed when the length of linker was 3 or 5 carbon atoms long. Analysis of the ligand-protein interactions of the compounds with a linker of 3 and 5 carbons, revealed in one case a relationship between the length of the linger and the type and diversity of interactions. This reinforces one conclusion from the previous chapter, that one cluster within a PCA plot (from the previous chapter) is a carrier of radial distance information.

Another aspect of interest was the conformational space available to these chalcones, as this has an impact on binding. Conformational studies were performed on both s-*cis* and s-*trans* configurations of the chalcone acetophenone_15_benzaldehyde_89. This indicated the presence of many rotameric conformations; in biological systems there will be an energy cost for transformation to the appropriate s-*cis* or s-*trans* configuration, and on top of this to the appropriate rotameric form for binding.

# REFERENCES

1.      S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N . Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, and X. Barril, *Dynamic undocking and the quasi-bound state as tools for drug discovery.* Nat Chem, 2017. **9**(3): p. 201-206.

2.      Kakwani, M.D., N.H.P. Desai, A.C. Lele, M. Ray, M.G.R. Rajan, and M.S. Degani, *Synthesis and preliminary biological evaluation of novel N-(3-aryl-1,2,4-triazol-5-yl) cinnamamide derivatives as potential antimycobacterial agents: An operational Topliss Tree approach.* Bioorg Med Chem Lett, 2011. **21**(21): p. 6523-6526.

3.      R. Zoraghi, L. Worral, R. H. See, W. L. Popplewell, H. Gong, T. Samaai, R. D. Swayze, S. Kaur, M. Vuckovic, B. B. Finlay, R. C. Bruham, W. R. McMaster, M. T. Davies-Coleman, N. C. Strynadka, R. J. Andersen, and N.E. Reiner, *Methicillin-resistant Staphylococcus aureus (MRSA) Pyruvate Kinase as a Target for Bis-indole Alkaloids with Antibacterial Activities.* J Biol Chem, 2011. **286**(52): p. 44716-44725.

4.      Farce, A. *Interaction ligands-cible: le drug design direct*. in *Master Conception du Médicament*. 2011.

# Chapter 6: Conclusion of the thesis

An in-house virtual library was designed following the Claisen-Schmidt reaction. Reactants were chosen according to several criteria such as commercial availability and short alkyl chain length in order to facilitate the future synthesis of these chalcone derivatives. Composed of 8 063 molecules, this library was successfully constructed with all models geometrically optimized at the QM level. The physicochemical characteristics linked to the Lipinski rules were analyzed. Very few presented any violations.

This library was submitted to high-throughput virtual screening against several proteins responsible for a range of conditions. These proteins (HIV1 integrase, pyruvate kinase, HSP90, COX-1 and 2, $ALR_2$, MAOA, MAOB, Acetylcholine esterase, butyrylcholinesterase, phospholipase $A_2$) were identified using the literature existence *in vitro* assays of chalcone derivatives against the various targets. AutoDock Vina was used to calculate the binding energy of the ligand-protein complexes using the Center for High Performance Computing (CHPC). The results of the calculation were visualized using a heatmap. The complexes generally presented good binding values except for MAOA, for which there were complexes with positive binding energies. COX-1 and pyruvate kinase showed the strongest binding across the virtual library. With the exception of pyruvate kinase AB crystal structure, the experiment was validated by comparison of the docking of known chalcone inhibitors for each particular target, and comparison with crystal structure ligands within the active site. Molecular dynamics focussing on HSP90 with good binding ligands, and subsequently modified ligands (including the crystal structure ligand, one of the best binding chalcones and its modified version following a Topliss decision tree) was performed. A comparison was made with respect to protein motion during dynamics for the protein alone and the protein in the presence of the various ligands.

A statistical analysis of the binding energy was undertaken to investigate parameters which might explain the results of the binding energy. First a PCA and then a cluster analysis relating properties of the ligands (from CDK and structural/pose information) to the observed binding energy were performed. The PCA for the virtual screens showed interesting aspects of binding to the targets, including some observations about selectivity. For instance, it was concluded that the binding energy of each individual ligand from the library to MAOA was not correlated in any way to binding to any other target. The polar spherical coordinates of the

ligand poses for each protein were calculated using a custom script and incorporated into the study. It was suggested through this analysis, for the pyruvate kinase C crystal structure, that the radial distances and the polar angles of two key structural points in each chalcone was correlated to the binding energy of the ligand-protein complex. A machine learning process was constructed producing similar results to the PCA.

Two separate focussed libraries were constructed. The first one was based on increasing the distance by increments between rings A and B relative to chalcones; this library was created by modification of five random chalcones molecules with specified linker segments. This library was docked against pyruvate kinase AB and C crystal structures. The results showed a decrease in the binding energy when the linker region between rings A and B of the ligand contained exactly five atoms. This confirmed PCA analysis that radial distances play important roles in explaining the binding energy to the pyruvate kinase C crystal structure.

Inspired by the Topliss decision tree, a modified library was made in attempt to improve the binding energy of ligands to the HPS90 protein. Due to choice of ligand, no improvement was observed with the modified library.

A conformational search of a representative chalcone revealed regions of stability for both the s-*cis* and s-*trans*-configuration. This provides further insight as to conformational cost associated with binding of a chalcone to a target.

Future work will involve the synthesis of select chalcones for the purposes of *in vitro* and Nuclear Magnetic Resonance (water-logsy and Saturation-Transfer-Difference) tests to confirm the binding to select targets.

# ANNEXES CHAPTER III

```python
protein_list = []
protein = []
acetophenone = []
acetophenone_list = []
ligand_directory= []
ligand_list= []

protein_list= os.listdir("path of the protein folder")
ligand_directory = os.listdir("path of the ligand folder")

for protein in protein_list:
    for ligand in ligand_directory:
        prot= remove the extension ".pdbqt" of the name of the protein
        lig=remove the extension ".pdbqt" of the name of the ligand
        os.system("mkdir create a directory with the name of prot)
        vina_file_name="path of the name of the protein plus prot_ligand.vina"
        print "name is "+vina_file_name

        vinafile=open(vina_file_name, 'w')
        vinafile.write("receptor = file of the protein with its full path\n")
        vinafile.write("ligand = file of the protein with its full path\n")
        vinafile.write("out = name for the output file with the extensation ".all.pdbqt"\n")
        vinafile.write("log =  name for the log with the extension ".log"\n")

        if prot name == "AchE_4m0e_A_apo":
            vinafile.write("center_x = -13.683\n")
            vinafile.write("center_y = -42.076\n")
            vinafile.write("center_z = 25.453\n")
            vinafile.write("size_x = 26.25\n")
            vinafile.write("size_y = 18.75\n")
            vinafile.write("size_z = 20.25\n")
            vinafile.write("energy_range = 4\n")
            vinafile.write("exhaustiveness = 120\n")
            vinafile.write("cpu = 24\n")

        elif prot name == "AchE_4m0e_B_apo":
            vinafile.write("center_x = 9.262\n")
            vinafile.write("center_y = -54.442\n")
            vinafile.write("center_z = -22.451\n")
            vinafile.write("size_x = 26.25\n")
            vinafile.write("size_y = 24.75\n")
            vinafile.write("size_z = 20.25\n")
            vinafile.write("energy_range = 4\n")
            vinafile.write("exhaustiveness = 120\n")
            vinafile.write("cpu = 24\n")

        elif prot == "piyruvte_kinase_3t05_AB_apo":
            vinafile.write("center_x = 0.579\n")
            vinafile.write("center_y = 33.8\n")
            vinafile.write("center_z = 23.22\n")
            vinafile.write("size_x = 126\n")
            vinafile.write("size_y = 126\n")
            vinafile.write("size_z = 126\n")
            vinafile.write("energy_range = 4\n")
            vinafile.write("exhaustiveness = 240\n")
            vinafile.write("cpu = 24\n")

        elif prot == "piyruvte_kinase_3t05_C_apo":
            vinafile.write("center_x = 31.494\n")
            vinafile.write("center_y = 11.371\n")
            vinafile.write("center_z = 31.974\n")
            vinafile.write("size_x = 126\n")
            vinafile.write("size_y = 126\n")
            vinafile.write("size_z = 126\n")
            vinafile.write("energy_range = 4\n")
            vinafile.write("exhaustiveness = 240\n")
            vinafile.write("cpu = 24\n")

        else:
            print "$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$ ERROR Cannot find the name of the protein: "+prot

        vinafile.close()
```

Figure 1: part of *pseudo code of the generation of vina files including parameters for AChE_A and B and for pyruvate kinase AB.*

# ANNEXES CHAPTER V

```python
rootdir="path of the library"
linkerdir=rootdir+"linker/aligned/"
benzdir=rootdir+"core/Benzaldehyde/"
acetodir=rootdir+"core/Acetophenone/"
outdir=rootdir+"output/"

linkerfiles=os.listdir(linkerdir)
benzfiles=os.listdir(benzdir)
acetofiles=os.listdir(acetodir)

#set the default coordinates of the linker, benzaldehydes and vectors
default_coordinates of linker_x=0.0
default_coordinates of linker_y=0.0
default_coordinates of linker_z=0.0

benz_x=0.0
benz_y=0.0
benz_z=0.0

vector_x=0.0
vector_y=0.0
vector_z=0.0

for linkerfile in linkerfiles:
    if linkerfile.endswith(".pdb"):
        for benzfile in benzfiles:
            if benzfile.endswith(".pdb"):
                for acetofile in acetofiles:
                    if acetofile.endswith(".pdb")
                        #preparation of the name files
                        libname=linkerfile+"_"+acetofile+"_"+benzfile
                        libname=libname.replace(".pdb","",2)
                        linkerdata=open(linkerdir+linkerfile,"r")
                        acetodata=open(acetodir+acetofile,"r")
                        benzdata=open(benzdir+benzfile,"r")
                        print "output "+ outdir+libname
                        outfile=open(outdir+libname,"w")
                        for acetoline in acetodata:

                            if(not(("REMARK" in acetoline)or("CONECT" in acetoline)or("TER" in acetoline) or
                            ("END" in acetoline)or("O9" inacetoline)or("C7" in acetoline)or("C8" in acetoline)
                            or("H6" in acetoline)or("H7"in acetoline)or("H8" in acetoline))): # do not write
                            these specific lines
                                outfile.write(acetoline)
                        for linkline in linkerdata:
                            if(("C0" in linkline)):
                                link_x=linkline[30:38] #localisation of the X coordinates
                                link_y=linkline[38:46] #localisation of the Y coordinates
                                link_z=linkline[46:54] #localisation of the Z coordinates


                            if(not(("REMARK" in linkline)or("CONECT" in linkline)or("TER" in linkline) or ("END
                            " in linkline) or ("C1" inlinkline) or("H1" in linkline) or("H2" in linkline)or("H3
                            " in linkline) )): # do not write these specific lines
                                outfile.write(linkline)
                        for benzline in benzdata:
                            if("C2 " in benzline):
                                benz_x=benzline[30:38] #localisation of the X coordinates
                                benz_y=benzline[38:46] #localisation of the X coordinates
                                benz_z=benzline[46:54] #localisation of the X coordinates
                        benzdata.seek(0)
                                    ### L2 L$ L6 need to rotate
                        vector_x=float(link_x)-float(benz_x) #calculation of the vector for X
                        vector_y=float(link_y)-float(benz_y) #calculation of the vector for Y
                        vector_z=float(link_z)-float(benz_z) #calculation of the vector for Z
                        for benzline in benzdata:

                            if(not(("REMARK" in benzline)or("CONECT" in benzline)or("H8" in benzline)or("C7"
                            in benzline)or("O9" inbenzline)or("C2 " in benzline))): # do not write these
                            specific lines
                                if("HETATM" in benzline):
                                    before=benzline[0:30] #information of the atom that we do not change
                                    after=benzline[54:] #information of the atom that we have to change
(coordinates)
                                    benz_x=benzline[30:38] # old X coordinates
                                    new_x=float(benz_x)+vector_x # new X coordinates
                                    benz_y=benzline[38:46] old Y coordinates
                                    new_y=float(benz_y)+vector_y # new Y coordinates
                                    benz_z=benzline[46:54]  old Z coordinates
```

```
                              new_z=float(benz_z)+vector_z # new Z coordinates
                              outfile.write("%s%8.3f%8.3f%8.3f%s" % (before,new_x,new_y,new_z,after)) #
write the new coordinates
                         else:
                              outfile.write(benzline)
                    outfile.close()
```

Figure 25: code of the extended library generation

```
   ################## INSTALLING PACKAGE FOR PCA  ##################################################
install.packages("FactoMineR")
install.packages("factoextra")
install.packages("devtools")
install.packages("clusterSim")

################ Activate the library ###########################################################
library(FactoMineR)
library(factoextra)
library(cluster)
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
library(clusterSim)

################ Read the table from previously ##################################################
setwd("D:/PhD/Chalcones/vina/Results_stats/")
donne_PK_C_wo_A56 <- read.csv("PK_C_Results_coord_exclude_A56.csv",sep = ",", dec = ".", rownames=1,
header = T)


################# Correlation analysis for protein #############################################
mydata_PK_C_wo_A56 <-cor(donne_PK_C_NORM_wo_A56)

################# Normalization of the data ####################################################
donne_PK_C_NORM_wo_A56 <- data.Normalization(donne_PK_C_wo_A56, type = "n1", normalization = "column")  #####
n1 corresponding to normalization cf help

################## PCA analysis with individual and factor maps ###############################
PK_C_NORM_wo_A56.pca <- prcomp(donne_PK_C_NORM_wo_A56, scale. = FALSE)
fviz_eig(PK_C_NORM_wo_A56.pca)
fviz_pca_ind(PK_C_NORM_wo_A56.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, label ="none") ###### test in progress
fviz_pca_var(PK_C_NORM_wo_A56.pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))

################ Separation of the different cluster ##############################
groupA = which(PK_C_NORM_wo_A56.pca$x[,1]<=2.5)
groupB = which(PK_C_NORM_wo_A56.pca$x[,1] >2.5 & PK_C_NORM_wo_A56.pca$x[,1] <=5)
groupC = which(PK_C_NORM_wo_A56.pca$x[,1] >5 & PK_C_NORM_wo_A56.pca$x[,1] <=9.5)
groupD = which(PK_C_NORM_wo_A56.pca$x[,1]>=10)

############ visualization of the entries of the differents cluster #####################################
groupA
groupB
groupC
groupD
```

Figure 26: Code for the separation of the clusters from the individual PCA map and the visualization of the entries which are
composed for RStudio