

**PEOPLE'S AND DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC
RESEARCH**

MOHAMED KHIDER UNIVERSITY OF BISKRA

FACULTY OF EXACT SCIENCES, NATURE AND LIFE SCIENCES

Department of Matter Sciences

THESIS

Presented by

ALMI Imane

Submitted for the award of the diploma of

Doctorate in Chemistry

Option :

Molecular chemistry

Entitled:

**Contribution to drug design through computational studies of
several series of bioactive heterocyclic molecules**

Publicly defended on:

24th May 2021

In front of the jury composed of:

| | | | |
|-------------------|-------|----------------------------------|------------|
| M. DIBI Ammar | Prof. | Hadj Lakhder University-Batna1 | President |
| M. BELAIDI Salah | Prof. | Mohamed Khider University-Biskra | Supervisor |
| M. MELKEMI Nadjib | Prof. | Mohamed Khider University-Biskra | Examiner |
| M. DAOUD Ismail | MCA | Mohamed Khider University-Biskra | Examiner |

Abstract:

Detoxification enzymes play an important role in cleaning the body from the toxins. These ones represent a hindrance to some drugs to fulfill their tasks, especially active compounds like anti-cancer drugs. These latter are considered to have a high degree of toxicity in the body, which makes them targeted by the previous enzymes.

We refer in this study to GSTp1-1, which is included in the detoxification enzymes class II targeted by NBD derivatives. It has dual defense feature, namely: inhibition GSTp1-1 and prevent the formation of each of the following complexes JNK1-GSTp1-1 and the TRAF2-GSTp1-1, that causes prolonged stopping of the cell cycle and facilitates apoptosis of damaged cells.

This is what made us in this study shed light on the modeling similar compounds, and to achieve this goal, we applied a set of methods adopted in the modeling of active materials of high biological quality. Among them, the QSAR Two-dimensional (2D-QSAR) coupled with a virtual examination, by using a technique similarity search. In addition, we concretized a three-dimensional stereo (3D-QSAR) which contains effective biological properties (Pharmacophore). This application resulted to determine a quantity of compounds bearing the same previously identified characteristics. Therefore, we put limits, selectivity features extracted from specialized references, to reduce and identify biologically the best. We make sure of the validity and safety of extracted models mentioned above by using several ways, namely: LOO-CV, external test set validation, fisher randomization, and cost analysis.

As a final result of this research, we identified 28 new derivatives of NBD. From both studies, at different inhibitory concentrations, micromolar unit (μM); the value of the half-maximal inhibitory concentration of a compound is $6.531 \mu\text{M}$.

ملخص:

تلعب انزيمات ازالة السمية، دورا هاما في تنظيف الجسم من السموم المتعرض لها. حيث تمثل عائقا لبعض الادوية في تادية مهامها، من بين هذه الادوية (المركبات الفعالة) نجد المضادات السرطانية. تعتبر هذه الاخيرة ذات درجة سمية عالية في الجسم، ما يجعلها مستهدفة من قبل الانزيمات المشار اليها سابقا.

نشير في هذه الدراسة، الى GSTp1-1 التي تندرج ضمن انزيمات ازالة السمية من الدرجة الثانية، والتي تستهدف من طرف مشتقات NBD- ذات الخاصية الثنائية للدفاع، المتمثلة في تثبيط GSTp1-1 ومنع تشكل كل من المعقدات التالية: JNK1-GSTp1-1 و TRAF2-GSTp1-1 التي تتسبب في توقيف دورة الخلية لفترات طويلة وتسهيل استماتة الخلايا التالفة.

هذا ماجعلنا في هاته الدراسة نسلط الضوء على نمذجة اشباه هذه المركبات. لتحقيق الهدف المرجو منه، قمنا بتطبيق مجموعة من المناهج المعتمدة، في نمذجة مواد فعالة ذات جودة بيولوجية عالية. نذكر منها، تقنية ال-QSAR ثنائية الابعاد (2D-QSAR) مقرونة بالفحص الظاهري، باستخدام تقنية similarity search. اضافة الى ذلك قمنا بتجسيد مجسم ثلاثي الابعاد (3D-QSAR). يحوي خصائص (ميزات) بيولوجية ذات قيمة فعالة (pharmacophore). اسفر هذا التطبيق عن تحديد كم من المركبات الحاملة لنفس الميزات المحددة سابقا، ماجعلنا نضع بعض الحدود-معايير انتقائية مستخلصة من مراجع متخصصة، لتقليل وتحديد الاحسن من الناحية البيولوجية. قمنا بالتأكد من مصداقية وسلامة النماذج المستخرجة والمذكورة اعلاه، باستخدام عدة طرق نذكر منها LOO-CV, external test set validation, fisher randomization, cost analysis.

كنتيجة نهائية لهذا البحث حددنا 28 مشتقا جديدا لـ NBD، من كلا الباحثين بتركيز تثبيطية متفاوتة بوحدة الميكرومولار (μM)؛ حيث بلغت قيمة التثبيط النصفى لإحدى المركبات 6.531 ميكرومولار.

Contents:

| | |
|---|-----|
| Acknowledgements | ii |
| List of works | iii |
| List of Abbreviations | iv |
| List of Figures | v |
| List of Tables | ix |
| I. Introduction | 1 |
| I.1. Contributions | 3 |
| I. 2. Organization of the dissertation | 3 |
| I.3. References | 5 |
| II. Background on cancer disease and drugs discovery | 7 |
| II.1. Life cycle of a drug | 7 |
| II.1.1. Target identification and Validation | 9 |
| II.1.2. Hit identification | 10 |
| II.1.3. Lead generator and optimization | 10 |
| II.1.4. Preclinical studies | 11 |
| II.2. Cancer | 12 |

| | |
|---|----|
| II.2.1. Cancer, a major health issue | 12 |
| II.2.2. Pathogenesis of cancer | 15 |
| II.2.2.1. Outside Body Factors (Environmental Factors) | 16 |
| II.2.2.2. Inside Body Factors | 16 |
| II.2.3. Treatment of cancer | 16 |
| II.2.3.1. Glutathione-S-Transferase (GST) | 17 |
| II.2.3.2. GST P1-1 | 19 |
| II.2.3.3. GST P1-1 physiological function (Role in cancer diseases) | 19 |
| II.2.3.4. GST π inhibitors | 20 |
| II. 3. References | 23 |
| III. Computer-Aided Drug Design and Discovery | 27 |
| III.1. Generality | 27 |
| Part I: Ligand-based drug design | 30 |
| III.2. Ligand-based drug design (LBDD) | 31 |
| III.2.1. QSAR analysis | 31 |
| III.2.1.1. Object of QSAR study | 32 |
| III.2.1.2. Steps involved in QSAR study: | 33 |
| III.2.1.2.1. Data collection and selection of training set | 33 |
| III.2.1.2.2. Molecular Descriptors used in QSAR | 34 |
| III.2.1.2.3. Variable selection methods | 35 |
| III.2.1.2.4. Development of QSAR model | 36 |
| III.2.1.2.4.1. Linear regression: | 36 |
| a. Multiple linear regression (MLR) | 36 |
| b. Partial least squares regression (PLS) | 36 |
| III.2.1.2.4.2. Non-linear regression | 37 |
| III.2.1.2.5. Validation of QSAR model | 39 |
| III.2.1.2.6. Applicability domain (AD) | 42 |
| III.2.2. Chemical similarity analysis | 43 |
| III.2.3. Ligand_ Based Pharmacophore | 44 |

| | |
|---|----|
| Part II: Structure-Based Drug Design | 47 |
| III.3. Structure-based drug design (SBDD) | 48 |
| III. 3. 1. Molecular Docking | 48 |
| III.3.1.1. Theory of docking | 49 |
| III.3.1.2. Search algorithm | 50 |
| III.3.1.3. Scoring | 52 |
| III.3.1.4. Molecular docking types | 55 |
| III.3.2. Generality on molecular dynamic | 55 |
| III. 4. References | 58 |
| IV. Contributions and result | 62 |
| Part 1. QSAR investigations and Ligand-based virtual screening on a series of nitrobenzoxadiazole derivatives targeting human glutathione-S-transferases | 63 |
| VI. 1. 1. Introduction | 64 |
| VI. 1. 2. Methodologies | 72 |
| VI. 1. 2. 1. Equilibrium structure optimizations | 72 |
| VI. 1. 2. 2. Molecular descriptors generation | 73 |
| VI. 1. 2. 3. Model development | 74 |
| VI. 1. 2. 4. Virtual screening | 75 |
| VI. 1. 3. Results and discussion | 75 |
| VI. 1. 3. 1. Equilibrium structure of the nitrobenzoxadiazole derivatives | 75 |
| VI. 1. 3. 2. Quantitative structure activity relationships (QSAR) study | 76 |
| VI. 1. 3. 3. Applicability domain of the model | 80 |
| VI. 1. 3. 4. Importance of descriptors within different QSAR models | 81 |
| VI. 1. 3. 5. Virtual Screening Application | 82 |
| Part 2. Combined 3D-QSAR based Virtual Screening and Molecular Docking study of cytotoxic agents targeting human glutathione-s transferases | 85 |
| VI. 2. 1. Introduction | 86 |

| | |
|--|-----|
| VI. 2. 2. Data collection and preparation | 86 |
| VI. 2. 3. Results and discussion | 87 |
| VI. 2. 3. 1. Generation of pharmacophore models: | 87 |
| VI. 2. 3. 2. Validation of the pharmacophore model | 89 |
| VI. 2. 3. 2. 1. Cost analysis | 90 |
| VI. 2. 3. 2. 2. Test set analysis | 90 |
| VI. 2. 3. 2. 3. Fischer Randomization Method | 91 |
| VI. 2. 3. 3. Virtual screening | 91 |
| VI. 2. 3. 4. Docking | 92 |
| VI. 3. References | 94 |
| V. Conclusion | 99 |
| Appendix | 101 |

To my dear parents

To my sister

To my brothers

Acknowledgements

In the name of "Allah", who gave me the will, patience, and health to complete this work although the circumstances surrounding us. I would like to express my gratitude to my supervisor Prof. BELAIDI Salah for his kind help, support, patience, and guidance throughout this work.

I express my sincere gratitude and appreciation to my reading committee members, Prof. DIBI Ammar of Batna University, Prof. MELKEMI Nadjib, and Dr. DAOUD Ismail of Biskra University for their time to read this thesis and for having accepted to judge this work.

I am also grateful to Prof. MELKEMI Nadjib for the support he has given me, his advice that has been of assistance throughout this work.

I would especially like to thank Prof. Magid abou-gharbia and Prof. Khaled M. Elokely for having dedicated me of the time throughout this year, as well as for their encouragement, and their distinguished reception in the Moulder Center for Drug Discovery Research laboratory and their supervision during my stay in USA. In addition, I would like to extend my warmest thanks to Dr. Mohammed Ibrahim for his help, for his kindness and patience.

Finally, I am forever indebted to my family especially my mother and my father for encouraging me during these many years, and to have always been available when I needed them. In addition, I thank my friends Rachida and Merzaka for their kind help, support, and for their valuable time.

List of works:

International publications:

[1] **Almi, I.**, Belaidi, S., Melkemi, N., & Bouzidi, D. (2018). Chemical reactivity, drug-likeness and structure activity/property relationship studies of 2, 1, 3-benzoxadiazole derivatives as anti-cancer activity. *Journal of Bionanoscience*, 12(1), 49-57.

[2] **Almi, I.**, Belaidi, S., Zerroug, E., Alloui, M., Said, R. B., Linguerri, R., & Hochlaf, M. (2020). QSAR investigations and structure-based virtual screening on a series of nitrobenzoxadiazole derivatives targeting human glutathione-S-transferases. *Journal of Molecular Structure*, 1211, 128015.

[3] Manandhar, A., Blass, B. E., Colussi, D. J., **Almi, I.**, Abou-Gharbia, M., Klein, M. L., & Elokely, K. M. (2021). Targeting SARS-CoV-2 M3CLpro by HCV NS3/4a Inhibitors: In Silico Modeling and In Vitro Screening. *Journal of chemical information and modeling*, 61(2), 1020-1032.

National publications:

[1] **Almi, I.**, Melkemi, N., Salah, T., & Daoud, I., Pharmacophore Searching, Virtual Screening and Molecular Docking for the Discovery of Novel Cytotoxic Agents Targeting Human Glutathione-S-Transferases. DOI:10.163.pcbj/2020.14.-3-225.

International conferences:

[1] Combined 3D-QSAR based virtual screening and molecular docking study of novel cytotoxic agents targeting human glutathione-S-transferases, **I. Almi**, N. Melkemi, T. Salah; International Bioinformatics- JIBioinfo 2019, 05 november 2019, Boumerdes, Algeria.

[2] QSAR investigations of series of nitrobenzoxadiazole derivatives targeting human GST and identification of novel compounds, **I. Almi**, S. Belaidi, E. Zerroug; 13th International Days of Theoretical and Computational Chemistry, 02 February 2020, Biskra, Algeria.

National conferences:

[1] Study of Electronic and Structural Property of 2, 1, 3- Benzoxadiazole, **I. Almi**, N. Melkemi, S. Belaidi; 1st day of study on molecular physical-chemistry, 6 December 2017, El Oued, Algeria.

[2] Ligand-based pharmacophore searching and molecular docking for the discovery of novel cytotoxic agents targeting human glutathione-s-transferases, **I. Almi**, N. Melkemi, T. Salah; 1st National Seminar on Applied Chemistry and Molecular Modeling, 26 September 2019, Guelma, Algeria.

List of Abbreviations

AD: Applicability domain

ADMET: Absorption, Distribution, Metabolism, Elimination, and Toxicity

ANN: Artificial Neural Network

B3LYP: Becke, three-parameter, Lee-Yang-Parr

CADD: Computer Aided Drug Design

CHARMM: Chemistry at Harvard Macromolecular Mechanics

CHEMBL: Chemical European Molecular Biology Laboratory

CV: Cross-Validation

DFT: Density Functional Theory

DNA: Deoxyribonucleic Acid

DS: Discovery studio

FDA: Food and Drug Administration

GSH: glutathione

GSTs: Glutathione S-transferases

HCA: Hierarchical cluster analysis

HOMO: Highest Occupied Molecular Orbital

HTS: High Throughput Screening

IC50: Half maximal Inhibitory Concentration

IND: Investigational New Drug

IUPAC: International Union of Pure and Applied Chemistry

JNK1: c-Jun N-terminal kinases

LBDD: Ligand-based drug design

LOO: Leave One Out

LUMO: Lowest Unoccupied Molecular Orbital

MAPEG: membrane-associated proteins in eicosanoid and glutathione metabolism

MLR: Multiple Linear Regressions

MMFF: Merck Molecular Force Field

MR: Molar Refractivity

MRP: Multidrug Resistance Protein

MW: Molecular Weight

NBD: Nitrobenzoxadiazole

NCI: National Cancer Institute

NHBD and NHBA: Number of Hydrogen-Bond Donors and Acceptors

NMR: Nuclear Magnetic Resonance

nrot: Number of Rotatable Bonds (nrotb)

OECD: Organization for Economic Cooperation and Development

PDB: Protein Data Bank

PGP: P-Glycoprotein

PLS: Partial least squares regression

PRESS: Predictive Residual Sum of the Squares

PSA: Polar Surface Area

QSAR: Quantitative structure-activity relationship

R&D: research and development

RMS: Root-Mean Squared

SAR: structure-activity relationship

SAR: Structure–Activity Relationships

SBDD: Structure-Based Drug Design

SCID: Severe Combined Immune Deficiency

SVM: Support Vector Machines

TRAF2: TNF receptor-associated factor 2

VS: Virtual screening

WHO: World Health Organization

List of figures

Figure I. 1: The basic skeleton of Nitrobenzoxadiazole.

Figure II.1. Overview of the process of drug discovery and development.

Figure II. 2. The lead optimization phase start with the detection of the lead structure ('hit') in the relevant biological assay. New analogs with surface changes are prepared and tested in the bioassay. If the outcomes of the assay increase, the modifications are retained and the cycle is repeated. If the modifications are negative, the modifications are eliminated and the cycle is repeated. This method proceeds until a potential substance with the desired properties has been identified.

Figure II. 3 improvement in three cancer survival measures, World, 1990 to 2017. This graph measures the mortality rate of cancer, the mortality rate of cancer and the age-standardized mortality rate.

Figure II. 4 Preventive targeting of cancer hallmarks.

Figure II. 5 The effect of genes and the environment on cancer growth. (a) The percentage contribution of genetic and environmental causes to cancer. (b) The figure reflect family risk ratios- the age-adjusted risk ratio for first-degree cases compared to the general population. (c) The number of cancer deaths due to the stated environmental risk factor.

Figure II. 6 Overview of xenobiotic enzymatic biotransformation. Harmful molecules can migrate through the plasma membrane and, within the cells, may be attacked by the enzymes of the so-

called Step I metabolism. The major ones belong to the Cytochrome P450 family, consisting of many enzymes that catalyze various reactions, including hydroxylation—the main reaction involved—oxidation and reduction. GSTs that catalyze the conjugation of phase I-modified xenobiotics to endogenous GSH play a key role in the resulting phase II metabolism. The conjugate obtained is then actively transported out of the cell by various transmembrane efflux pumps (Phase III). Any compounds can join the metabolism of phase II directly.

Figure II. 7 (A) GST detoxification process. (B) (left) GSH identification GST. (right) The molecular architecture of the GST covalent inhibitor.

Figure II. 8 Ligand-binding features of JNK and TRAF2.

Figure II. 9 GSTP1-1 function in the JNK signaling pathway.

Monomeric GSTP1 prevents tumor cells from apoptosis by inhibiting the JNK signaling pathway via the development of a GSTP1-JNK-cJun complex that inhibits c-Jun phosphorylation. Under conditions of stress, GSTP1 can disassociate and dimerize from the complex, allowing JNK to phosphorylate c-Jun. This event can also be caused by a GST inhibitor NBDHEX that binds GSTP1 and induces its release from the complex.

Figure III.1. Workflow of ligand -based drug design (LBDD) and structure -based drug design (SBDD).

Figure III.2. Flowchart of the methodology used in QSAR study.

Figure III.3. Graphical view of an artificial neural network with one input layer (comprising three descriptors) attached to the hidden layer with the necessary weights and the output layer.

Figure III.4. Activation functions generally used for neural networks study.

Figure III.5. a) Schematic overview of the application domain. Every ringed dot is a single data point used for model training. New Chemical structures (solid dots) that fall into the inner, darker field are close enough to the training set and the model can be used confidently. The latest substances that fall in the white region are so far from the training collection that the formula can no longer be used. b) Williams' plot for the applicability domain of QSAR model.

Figure III.6. General framework of pharmacophore methods.

Figure III.7. Enzyme Activity Model Lock-and-Key.

Figure III.8. Schematic explains the techniques used for protein ligand docking.

Figure III.9. Small-molecule conformational search methods. (A) A molecule containing two bulky groups (green and purple spheres) has its conformation defined by two internal dihedrals Φ_1 and Φ_2 ; (B) Considering Φ_2 as a frozen dihedral, the energy variation due to rotation of Φ_1 is plotted in a 1D energy landscape. The initial structure (grey spheres) is modified by changing Φ_1 , leading to a decrease in energy. The systematic search algorithm changes all structural parameters until a local (blue spheres) or global (red sphere) energy minimum is reached; (C) The stochastic search explores the conformational space by randomly generating distinct conformations, populating a broad range of the energy landscape. This procedure increases the probability of finding a global energy minimum.

Figure III.10. Molecular dynamics basic algorithm. Notes: The simulation output, the trajectory, is an ordered list of $3N$ atom coordinates for each simulation time (or snapshot). Abbreviations: E_{pot} , potential energy; t , simulation time; dt , iteration time; For each spatial coordinate of the N simulated atoms (i): x , atom coordinate; F , forces component; a , acceleration; m , atom mass; v , velocity.

Figure IV. 1. The workflow used in QSAR-based virtual screening study

Figure IV.2. Structure of 4-Nitro-2, 1, 3-Benzoxadiazole.

Figure IV.3. Hierarchical cluster analysis of descriptors (dendrogram). See Table IV. 2 for the definition of these descriptors. The vertical red dashed line corresponds to the clipping limit that takes into account the minimum number of descriptor groups without losing any information necessary for the model.

Figure IV. 4. Experimental versus calculated pIC50 values (MLR in a) and ANN in c)), and residuals (MLR in b) and ANN in d)).

Figure IV.5. Applicability domain plot for the ANN model. Horizontal lines represent $\pm 3\sigma$ and the vertical dashed line represents the warning leverage ($h^* = 0.536$).

Figure IV.6. Comparison of descriptors contribution in the ANN and MLR models.

Figure IV. 7. Schematic representation of the virtual screening process implemented in the identification of Top inhibitors.

Figure IV. 8. The best HypoGen pharmacophore model, Hypo1

Figure IV. 9. Correlation graph between experimental and estimated activities in logarithmic scale for training and test set compounds based on Hypo1.

Figure IV.10. The difference in costs between the HypoGen runs and scrambled runs. The 95% confidence level was selected.

Figure IV. 11. 2D Binding interaction representation of NBD most active compound with active site of GSTp 1-1.

List of tables

Table II.1. The clinical and preclinical trials in drug development

Table II.2 Antitumor agents targeting GST π in background.

Table III.1. Popularly known molecular descriptors dependent on various dimensions.

Table III.2. Mathematical equation of statistical validation metrics used in QSAR studies.

Table III. 3. Examples of Scoring Function Formulae.

Table IV. 1. Observed and predictive activities (and their differences) of the set of nitrobenzoxadiazole derivatives [25, 26]. * denotes the external test set for GSTP1-1.

Table IV. 2. Symbols and description of all calculated molecular descriptors

Table IV. 3. Correlation matrix for the four selected descriptors with pIC₅₀. See Table III. 2 for the definition of these descriptors.

Table IV.4. Random MLR Model Parameters.

Table IV.5. Statistical results of MLR and ANN models.

Table IV.6. Proposed structural compounds and predicted activities.

Table IV.7. Statistical results of the top 10 pharmacophore hypotheses generated by HypoGen algorithm.

Table IV.8. Experimental and estimated activity of individual training set compounds.

Table IV. 9. Docking interaction of NDB (Most active compound) and virtually screened hit compounds.

Chapter I:

Introduction

Since the 20th century, the increase of life expectancy has been associated to the increase in exposure to carcinogenic elements, particularly those in tobacco smoke, as: azo dyes, aflatoxins, asbestos, benzene and radon, as well as ionizing radiation. These causes have been well documented as leading to a wide range of human cancers. Therefore, this disease has become the second most common cause of death around the world [1]. Nevertheless, there is a way to fight against cancer. These often-complementary therapies are used individually or in conjunction, depending on the type, position and stage of the cancer. There are mainly three kind of treatment such as: Chemotherapy, radiotherapy and surgery [2].

Chemotherapy occupies an important place in the clinical treatment of cancer. It uses such medications to destroy cancer cells or stop them from developing and spreading to other areas of the body. Many drugs are used in the treatment of cancer diseases; sometimes, however, they cannot play their role due to the detoxifying enzymes, like glutathione S-transferases (GSTs), which attacks these drugs and reduce their therapeutic effect. GSTs are a family of massive, distributed phase II detoxifying enzymes, which catalyze the recombination of reactive

electrophiles to the nucleophilic sulfur of the main intracellular thiol. GSTs' inhibitors have been shown to decrease drug resistance by improving anti-cancer drug action in tumor cells [3, 4]. Several synthesized compounds have been used to suppress detoxifying enzymes, including Nitrobenzoxadiazole (NBD) compounds (Scheme 1) and their derivatives, which gained a significant attention due to their unique mode of action: at the cellular level, they induce the dissociation of the JNK1- GSTP1-1 and TRAF2-GSTP1-1 complexes (GSTP1-1 for glutathioneS-transferase), leading to prolonged cell cycle arrest and apoptosis [5, 6].

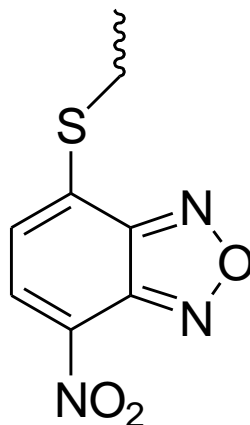


Figure I. 1: The basic skeleton of Nitrobenzoxadiazole.

Drug discovery is a long and complex process [7], both in terms of time and money invested [8]. The field of "drug design" can be explored through molecular modeling, using computer tools to design of new molecules [9]. It essentially boils down to identify new compounds (natural or synthetic molecules) which will ideally evolve into drugs acting on specific biological targets responsible for dysfunctions. The identification of therapeutic targets is linked to knowledge of molecular functioning, metabolic pathways and generally biological systems, and the cause of disease [10].

The world of pharmaceutical research is constantly optimizing all stages of its drug discovery and development process. Chemoinformatics is a tool of choice for reducing the time and cost of developing a drug. This discipline can intervene at different levels of the drug discovery process. Among the chemoinformatics techniques, we can cite the screening of chemical libraries (QSAR, docking and pharmacophores techniques). Chemoinformatics is present today in all stages of drug development [11].

By analogy with the expressions *in vivo* and *in vitro*, the term “*in silico*” has been introduced to qualify the numerical methods used to treat such systems. By its name, this term refers to silicon, the main material found in all computers’ chips. The *in silico* field brings together a very large set of numerical methods based on the laws of physics and chemistry which, using mathematical approaches, make it possible to simulate or model a biological phenomenon using the computer tool. Virtual screening is the most widely used *in-silico* strategy for the identification of compounds in the context of new drug research [12].

The advantage of virtual screening “*in-silico*” is therefore to provide a small list of molecules to be experimentally tested, thus reducing costs and saving time. We can also quickly explore many molecules and then focus, at the experimental level, on the most interesting molecules [13].

I.1. Contributions:

The objective of our work is to use virtual screening methods in the search for new bioactive molecules and to study their interactions with the enzyme glutathione-S-transferase.

Our main contributions are summed in these essential points, namely:

- Multiple Linear Regressions (MLR) and Artificial Neural Network (ANN) were applied for modeling of the studied molecules.
- Developed ANN models have reasonably predicted the GSTP1-1 inhibitory activities of 23 hits.
- Generation of a pharmacophore model based on ligands.
- Virtual screening procedure has been applied to large chemical compound database.
- Hits obtained have good predictive activities.
- Molecular docking of molecules resulting from virtual screening.

I.2. Organization of the dissertation:

To achieve our goal, we have organized our thesis into three chapters:

1. **Part I- Background on cancer disease and drug discovery:** In the chapter II we Will introduce the fundamental principle, the possess of development and discovery of drugs. In addition, it contains a description of the cancer diseases, their pathogenesis and their treatment.

2. **Part II-Computer-Aided Drug Design and Discovery:** this chapter consists of two parts: Ligand-based drug design (LBDD) and Structure-Based Drug Design (SBDD), which describe principle formulation, theories of the methods used during this research and the principle equations for each method.
3. **Part III- Contributions and result:** we will be mainly dedicated in this part to interpretations of the results obtained, which were divided into two parts. The first one, a Quantitative structure-activity relationship (QSAR) models were generated using Multiple Linear Regression (MLR) and Artificial Neural Network (ANN). These technics aim to determine the best molecular descriptors to be used in conjunction to identify the best candidates for GSTP1-1 inhibition. At last, the obtained QSAR models were employed to define biological activities of potentially novel active compounds by means of in silico screening processes. As to the second part, Ligand-based pharmacophore modeling was used to identify the chemical features responsible also for inhibiting GST p1-1. The identified features used to screen the database contain more than 200000 compounds. The last point consists of a molecular docking analysis, which recognizes that leads are possible toward the GST p1-1 enzyme and reveals the lowest energy and good associations with reduced active site of the GST. These ligands can then form stable complexes.

I. 3. References:

1. G. Lenglet, "Mécanisme d'action de nouveaux agents alkylants ciblant l'ADN ou les protéines," Université de Lille Nord de France Ecole Doctorale Biologie-Santé Mécanisme d'action de nouveaux agents alkylants ciblant l'ADN ou les protéines," 2012.
2. C. W. Thornber and A. Shaw, *Antihypertensive Agents*, vol. 12, no. C. 1977.
3. C. Adamo *et al.*, "Reproduction et environnement To cite this version : HAL Id : inserm-02102503," 2019.
4. F.Louacheni, "Développement d'un portail web pour le criblage virtuel sur la grille de calcul". 2014.
5. A. De Luca, L. Federici, M. De Canio, L. Stella, and A. M. Caccuri, "New insights into the mechanism of JNK1 inhibition by glutathione transferase P1-1," *Biochemistry*, vol. 51, no. 37, pp. 7304–7312, 2012, doi: 10.1021/bi300559m.
6. A. De Luca *et al.*, "The fine-tuning of TRAF2–GSTP1-1 interaction: effect of ligand binding and in situ detection of the complex," *Cell Death Dis.*, vol. 5, no. 1, pp. e1015–e1015, 2014, doi: 10.1038/cddis.2013.529.
7. U. D. Orleans, P. Obtenir, and L. E. Grade, "DOCTEUR DE L'UNIVERSITE D'ORLEANS Discipline : Chimie Informatique et Théorique MONGE Aurélien Création et utilisation de chimiothèques optimisées pour la recherche « in silico » de nouveaux composés bioactifs .," 2006.
8. H. Guillemain, *Evaluation et application de méthodes de criblage in silico*. 2012.
9. Y. Asses, "Conception par modélisation et criblage in silico d'inhibiteurs du récepteur c-Met.," p. 142, 2011.
10. A. Saadi and A. Cowe, "Le criblage à haut débit, en anglais High-Throughput Screening," 2007.
11. W. L. Jorgensen, "The Many Roles of Computation in Drug Discovery," *Science (80-.)*, vol. 303, no. 5665, pp. 1813–1818, 2004, doi: 10.1126/science.1096361.
12. W. H. M. Peters and H. M. J. Roelofs, "Biochemical Characterization of Resistance to Mitoxantrone and Adriamycin in Caco-2 Human Colon Adenocarcinoma Cells: A Possible Role for Glutathione S-Transferases," *Cancer Res.*, vol. 52, no. 7, pp. 1886–1890, 1992.
13. A. Hall, S. J. Proctor, A. R. Cattan, C. N. Robson, I. D. Hickson, and A. L. Harris, "Possible Role of Inhibition of Glutathione S-Transferase in the Partial Reversal of Chlorambucil Resistance by Indomethacin in a Chinese Hamster Ovary Cell Line," *Cancer Res.*, vol. 49, no. 22, pp. 6265–6268, 1989.

“God has created all diseases, and he also has created an agent or a drug for every disease. They can be found everywhere in nature, because nature is the universal pharmacy. God is the highest ranking pharmacist.” the Swiss-Austrian medical doctor and natural scientist Paracelsus

Chapter II:

Background on cancer disease and drugs discovery

II.1. Life cycle of a drug:

Throughout history, there was an almost continuous need for clinical action in the treatment of illness. There was no possibility to understand the biological origin of these diseases. The notion that treating diseases or alleviating effects might be done by smoking, consuming or adding morphine, ephedra, hemp, tobacco, salicylic acid, digitalis, coca, quinine, and a number of other medications, still in use, for a long time [1, 2]. The identification of new drugs was primarily done by changing the molecular structure of an existing drug or by serendipity. Whereas this process was a slow trial, it gives results with many errors. The techniques used to discover medicinal agents have evolved significantly over the course of human history. Now, a computer can display the molecular structure of any drug from a list of thousands in a database [3]. Computer Aided Drug Design (CADD) is expressed by in silico term; as an analogy to the Latin phrases in situ, in vitro and in vivo. This means the logical design from which medications are formulated or found using computational methods. The main aim of in silico Aided Drug Design is to identify the best chemical compound to experimental testing by reducing late stage attrition and costs [4].

The word "lifecycle" refers to the sequence of modifications that a substance, process, activity, ... etc has experienced over its lifetime. A drug's lifecycle includes two major steps, namely: *research and development (R&D)* that brings a new drug from discovery to launch, and *the marketing and sales* of medications [5]. In this part, we aim to describe the four major disciplines which have completely revolutionized the search for new drugs and resulted in the processes currently used in the early stages of research and development (R&D).

The drug discovery R&D processes are highly costly, time consuming and technology intensive. They bring together all the steps leading to the marketing of the new drug. According to published studies, it is expected that only one in ten of the compounds entering clinical development is successful, with an overall cost of USD 500-800 million and a standard time scale of 10-15 years from pre-clinical development to regulatory approval. Usually, the whole process is divided into "**Discovery**," "**Development**" and "**Registration**" stages (Figure II.1) [6, 7].

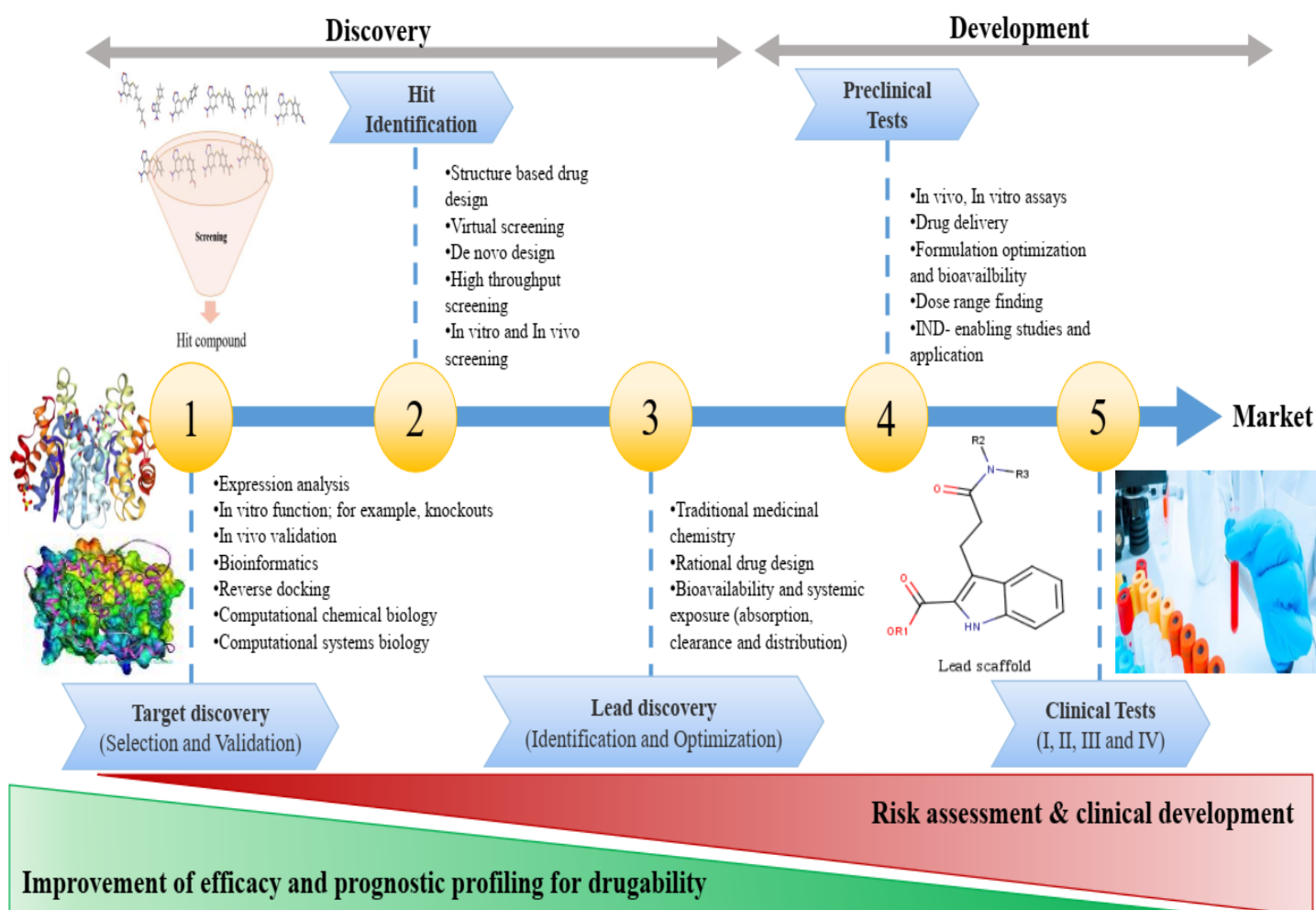


Figure II.1. Overview of the process of drug discovery and development.

Drug analysis can be divided into a variety of smaller tasks and roles. The mechanism can be systematically separated into two different parts at the highest level, namely: *discovery* and *development* [8]. Drug discovery is a lengthy arduous method. It is broadly packaged with biological target recognition and associated disease, target validity, high throughput identification of hits and leads, lead optimization, and preclinical and clinical analysis [9]. Each step of drug development should aim to create a scientific connection between a biological target (e.g.: enzyme, ion channel, G-protein-coupled receptor, etc.) and a disease-state model designed to simulate human disease. Drug development aims to evaluate the safety/toxicity and efficacy of new drug substances. The key aim of drug production is to create a research database that confirms the potency and safety profile of the drug and its dose regiment(s) for marketing purposes [8].

Early stages in drug discovery are the initial process of target identification and progress to the later phase of lead optimization. Many resources, including the private market, clinical work and academic research, assist in identifying the best disease target. The selected target is then used by the pharmaceutical industry and more recently by several research centers to select molecules for the production of suitable drugs. The process requires a numerous early stages [10].

We will describe the four principal stages drug research: (i) target determination, (ii) model establishment, (iii) discovery of lead compounds, and (iv) optimization of the lead compounds.

II. 1. 1. Target identification and Validation

Target identification and validation are the starting point of new drug R&D process where the method of selecting a potential drug candidate starts, with the determination of a disease state that can be solved or changed by the use of effective chemotherapeutic action [11, 12]. Once the disease has been defined, the next step is to identify a potential biological target. A good target needs to be active, safe, meet clinical and business needs and, most importantly, be 'druggable'. A 'druggable' target is accessible to a putative drug substance, i.e. a small molecule or larger biologics and, upon binding, generates a biological response that can be measured both in vitro and in vivo [13]. Following identification of the drug target, a systematic validation should show that a molecular target is directly involved in a disease process, and that modulation of the target is likely to have a therapeutic effect [10]. Good target selection and verification encourages improved confidence in the interaction between target and disease, and

enables them to explore where target modulation can contribute to mechanism-based side effects [13].

II. 1. 2. Hit identification:

Following the process of target identification, hit identification aims to identify where the small molecule Hits suitable for use in a medical environment. Of course, this comparatively simple statement is actually a representation of an extremely complex and multi-faceted problem. To identify the small molecule (hits), there are some variety of screening paradigms exist that have been developed in order to provide some guidance as to look for biologically useful molecules. Among these paradigms, we mention: firstly, High Throughput Screening. The primary role of HTS is to detect lead compounds and supply directions by testing, in an automated fashion, for activity as inhibitors (antagonists) or activators (agonists) of a particular biological target. Secondly, Virtual Screening [14-16]. Virtual screening or VS is an alternative method to the computational screening of large chemical libraries. It is a modern technique attracts an increasing degree of interest in the pharmaceutical industry as a productive and cost-effective technology in the quest for novel lead chemicals substances [17, 18].

In view of their activity, but also of additional criteria such as their originality or their stability, the compounds to become drugs are most likely selected as hits and then optimized [17, 18].

II. 1. 3. Lead generator and optimization:

The goal of this stage of the work is to change each hit list in order to try to produce more potent and selective compounds by iterative synthesis and to analyze their efficacy in any available in vivo models (Figure II.2). In each step of the "lead optimization" process, new data are created as adjustments in the molecular structure of the "lead". These details are used to develop the next generation of compounds. This step of generating Structure-Activity Relationship data persists until appropriate chemicals substances have been developed for clinical assessment [1]. All these chemical modifications around a common scaffold aimed to elucidate SAR, to establish consistent correlations of structural features, or groups, with the biological activity of compounds in a given biological assay. This SAR aimed to maximize efficacy and potency while keeping adequate ADMET properties and selectivity profile [19].

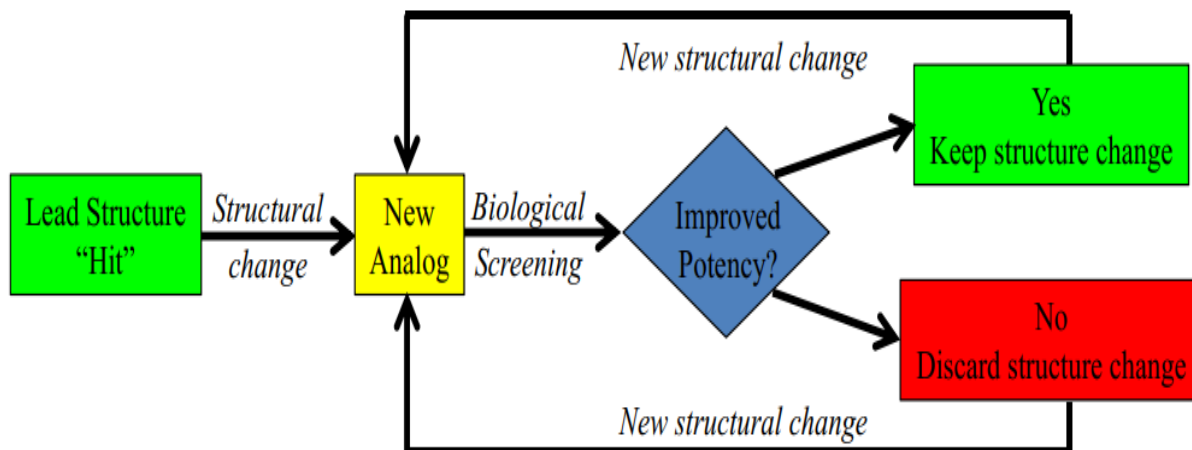


Figure II. 2. The lead optimization phase start with the detection of the lead structure ('hit') in the relevant biological assay. New analogs with surface changes are prepared and tested in the bioassay. If the outcomes of the assay increase, the modifications are retained and the cycle is repeated. If the modifications are negative, the modifications are eliminated and the cycle is repeated. This method proceeds until a potential substance with the desired properties has been identified [1].

II. 1. 4. Preclinical studies:

Following a series of *in vitro* and *in vivo* experiments to find out the best drug candidate and before clinical trials, preclinical studies try to provide information on the preparation process, protection, dosage, acute and chronic toxicity, allergic reactions, formulation and components, pharmacokinetics, stability, effectiveness, mutagenicity and local irritation tests, hemolytic, reproductive toxicity, and so on and so forth [2]. Preclinical studies must comply with the guidelines, laid down by Good Laboratory Practice to ensure consistent results and required by authorities, such as the FDA, before submitting an IND approval [21].

The clinical phases I, II, III, and IV studies consist to evaluate drug safety for human beings, with a small and a large group of participants, and identify the dose range and side effects (Table II.1).

Table II.1. The clinical and preclinical trials in drug development [21]

| Contents | Preclinical trials | Clinical trials | | | |
|--|---|--|------------------------------------|------------------------------------|---|
| | | Phase I | Phase II | Phase III | Phase IV |
| Objects | Lab study In vitro assays In vivo animals | Human 20-80 people | Human 100-300 people | Human >1000 people | Human >1000 people |
| Goal | Effects Safe dose administration routes drug distribution etc | Safety Side effects Administration routes Dose escalation | Efficacy Safety Side effects | Efficacy Safety Side effects | Efficacy Long-term Safety Side effects Mortality rate |
| Time to last Chance to pass | 1-3 years NA | Months to years 70% | 1-2 years 30% | Several years 25-30% | Long-term Rare withdrawal |

II. 2. Cancer:

The human body is made up of over than 60000 billion cells. These are the units which make up tissues and then our organs, including heart, liver and lungs. If the body wants, our cells are doubled to destroy any cells that are damaged or their lifetime is ended. This makes it possible for our tissues to retain their shape and function with the flow of time. Therefore every cell is conditioned to multiply and die. This organized but complex program is regulated by the center of the cell, the nucleus, which includes chromosomes containing several genes made up of DNA. At times, few any of these genes are modified. The nucleus gives out irregular orders, and the cell goes incorrect. It multiplies abnormally, taking on a life of its own; each new cell generated contains the same error. Cells proliferate chaotically to form a tumor. This period may be short, but it is always 10 to 30 years long and may distinguish the birth of the first irregular cell from the creation of a tumor of around one cubic centimeter in which several blood vessels are formed to survive, which will supply the tumor with oxygen and nutrients, allowing it to survive and developed. That's what we call the concept of angiogenesis. However the tumor is only really risky when cancer cells start to invade the surrounding regions through the vessels and spread to the surrounding organs. These cells can then invade other parts of the body to multiply and generate new cancer cells. Metastasis is the word used for this process of spread. But why does a cell become cancerous? In addition, is there a treatment for it?

II. 2. 1. Cancer, a major health issue:

Cancer is a vast family of diseases caused by irregular cell formation, growth rate and capacity to invade other organs [22]. Every sixth death in the world is due to cancer, making it the second leading cause of mortality worldwide. World health organization estimated that 42

million people around the world suffered from different types of cancer cells. This percentage has more increased since 1990, when an estimated 19 million patients had cancer, with an approximate 9.6 million patients dying from cancer during 2017 (Figure II. 3). Thus, cancer is a significant issue impacting the welfare of all human cultures [23, 24]. Unfortunately, it is a form of disease at the stage of the tissues and this variation is a significant problem for its particular diagnosis, followed by the effectiveness of the medication [25, 26]. Cancers as a whole accounted for 30.0 per cent of man deaths and 24.8 per cent of woman deaths in 2015. When viewed independently, 4 of the 10 leading causes of death of both sexes were cancers. Lung, colorectal cancer, leukemia and lymphomas are among top leading causes of mortality for both genders. Breast cancer is the most common type of cancer among female accounting for about 30% of all woman cancers [27]. Over the last 10 years, the incidence of breast cancer has increased by 1.5 per cent per year. Mortality, however, does not rise. Prostate cancer accounts for approximately 35% of all male cancers. The occurrence has risen by an average of 4% annually over the last 10 years, primarily due to early detection. However, the increase in prevalence is not followed by an increase in mortality [28, 29].

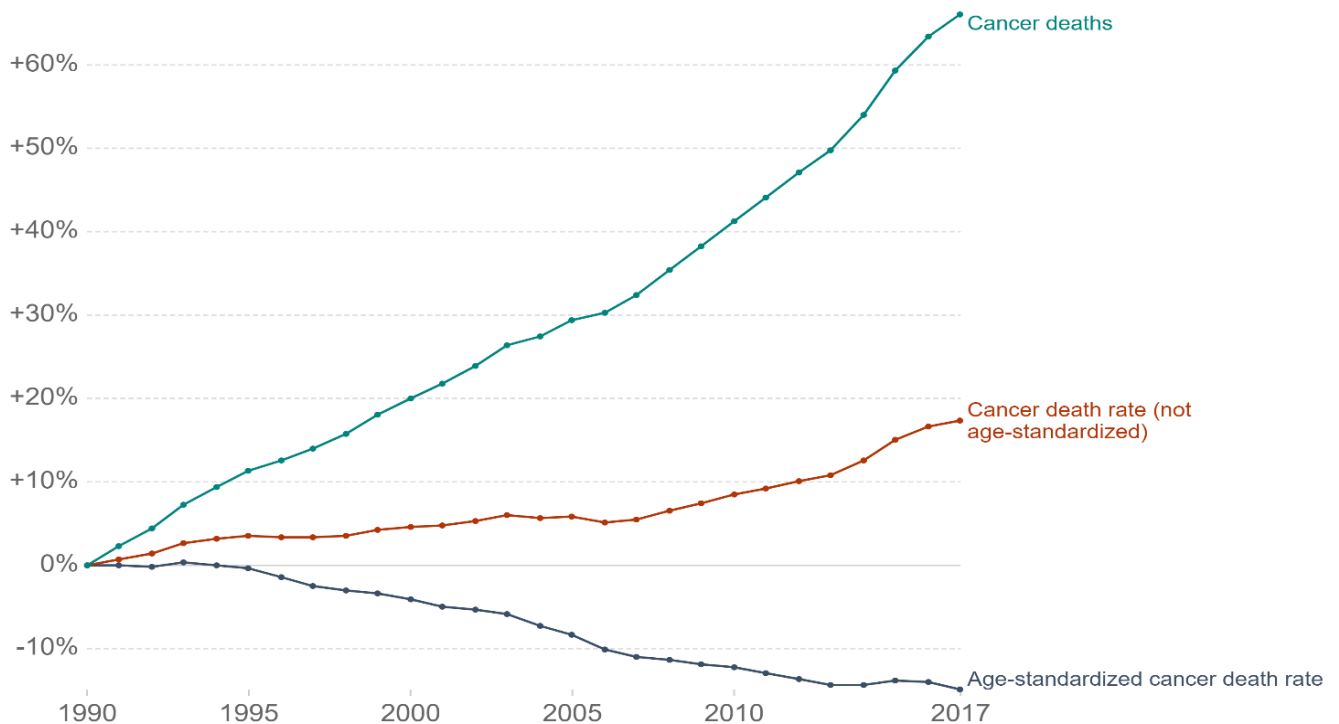


Figure II. 3 improvement in three cancer survival measures, World, 1990 to 2017. This graph measures the mortality rate of cancer, the mortality rate of cancer and the age-standardized mortality rate.

Recently, scientists summarized the ten biological hallmarks of cancer (Figure II. 4) as sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, avoiding immune destruction, tumor promoting inflammation, deregulating cellular energetics, genome instability and mutation [30].

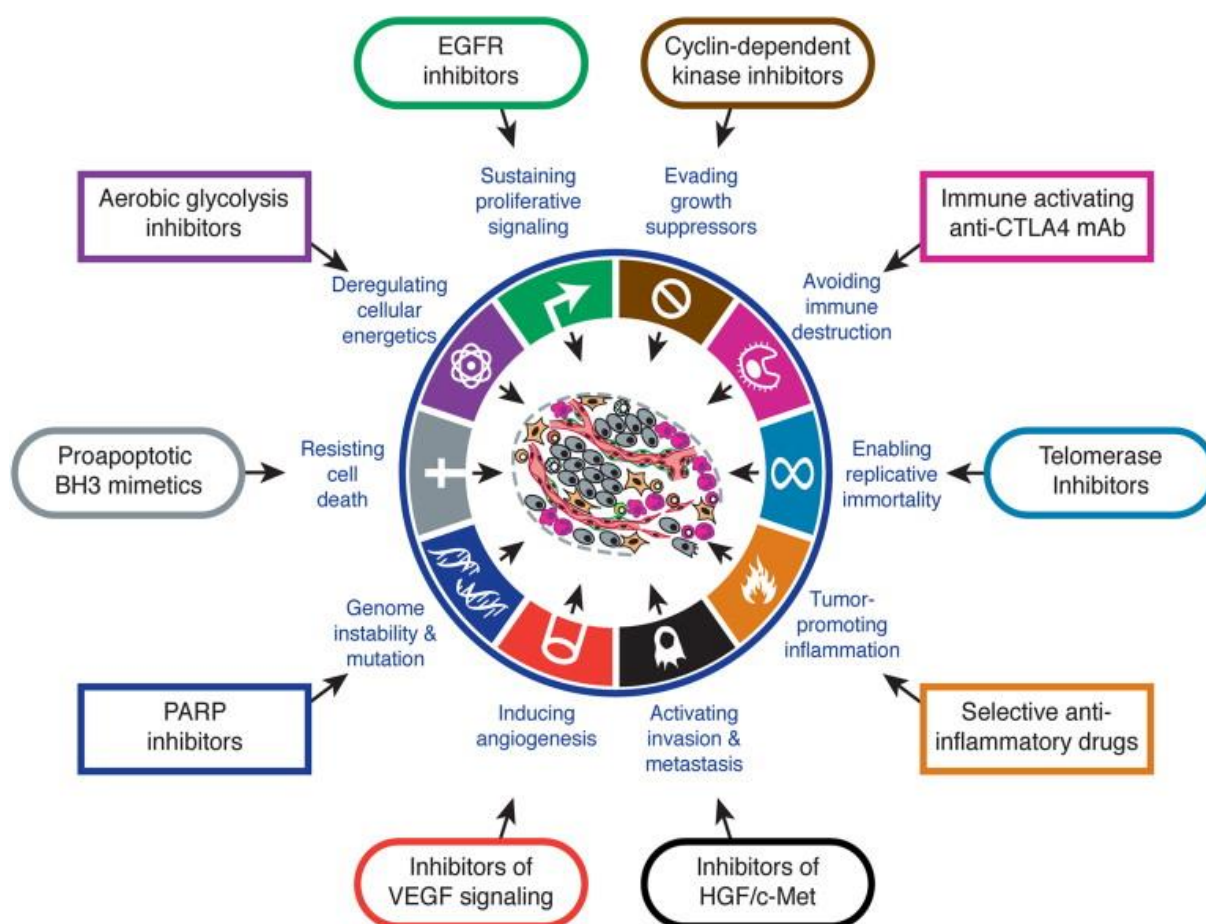


Figure II. 4 Preventive targeting of cancer hallmarks [30].

Medications, which interact with acquired ability required for cancer cells growth and development, have been developed, incorporated in clinical trials or, in certain cases, licensed for clinical use in the treatment of some types of human cancer. In contrast, the investigational medicines are designed to target one of the enabling characteristics and evolving features shown in Figure II. 4, which still retain promise as cancer treatment. Drugs mentioned are just illustrative examples; for each of these hallmarks, there is a deep pipeline of candidate drugs with various molecular targets and modes of production action [30].

II. 2. 2. Pathogenesis of cancer:

Cancer has been developed for many years and has several causes. These ones can exist in varying degrees, both inside and outside of the body, to contribute to the development of cancer [31]. Knowing the causes of cancer offers a basis for recognizing the potential for cancer prevention. If a reason is known, it is much easier to know whether it can or cannot be easily avoided [32]. Scientists typically divide these factors into two categories: those inside the body and those outside the body; environmental factors. It is estimated, however, that only 5-10 per cent of cancer is caused by inherited traits and the remaining 90-95 per cent is either caused or sustained by environmental factors Figure II. 5 [33].

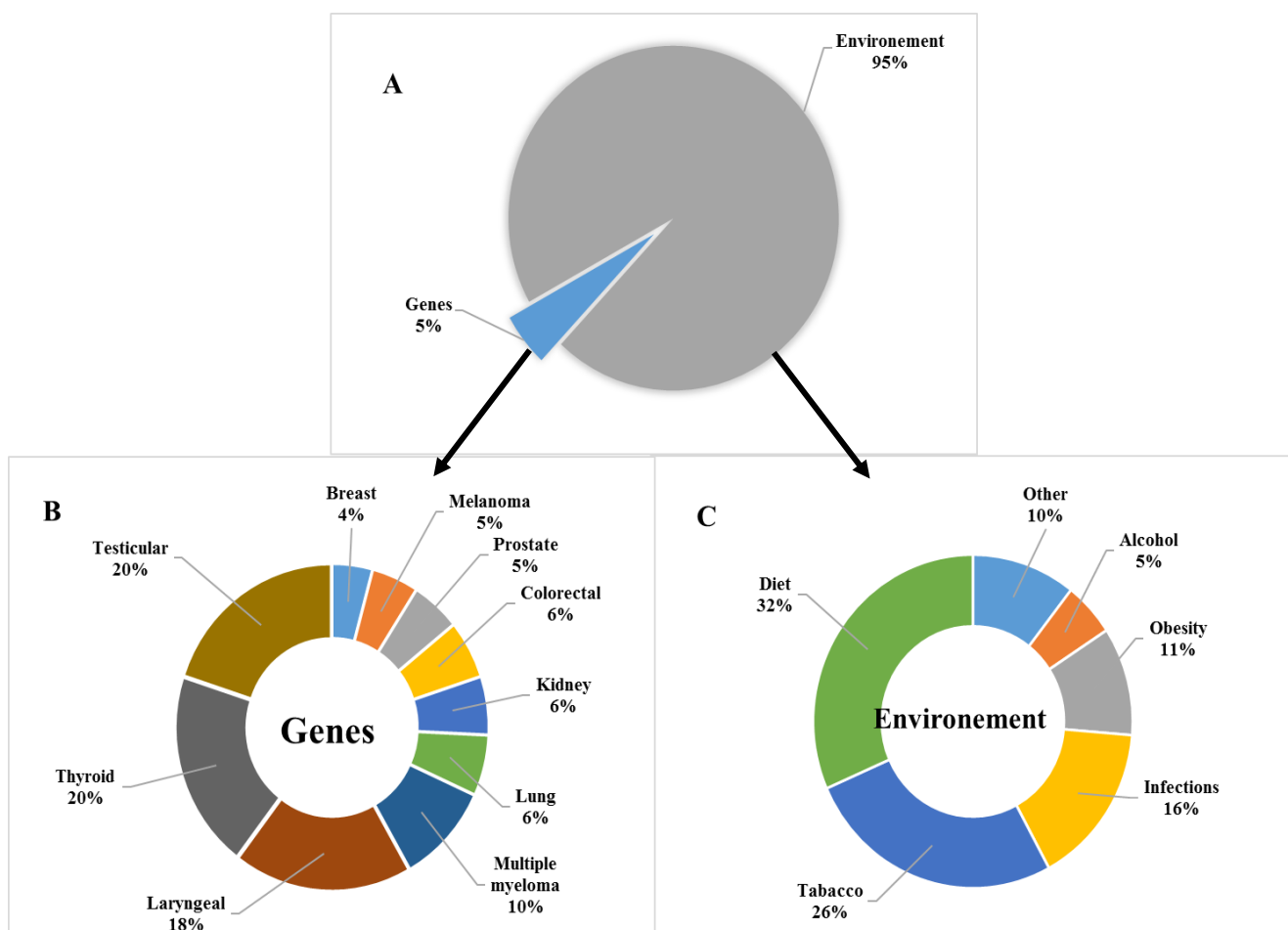


Figure II. 5 The effect of genes and the environment on cancer growth. (a) The percentage contribution of genetic and environmental causes to cancer. (b) The figure reflect family risk ratios-the age-adjusted risk ratio for first-degree cases compared to the general population. (c) The number of cancer deaths due to the stated environmental risk factor.

II. 2. 2. 1. Outside Body Factors (Environmental Factors):

Exposure to a wide range of natural and man-made chemical compounds in the environment accounts for at least two thirds of all cancer cases worldwide. These external factors involve lifestyle habits, such as: unhealthy diet, excessive alcohol intake, cigarette smoking, excessive exposure to sunlight, lack of exercise, and increased exposure to some viruses. Other considerations include exposure to certain pharmaceutical products, viruses, radiation, hormones, bacteria and environmental contaminants that may be present in the air, water, food and the workplace. Analysis of occupational groups with greater exposures to these chemicals compared to the general population has identified the cancer dangers related with certain environmental chemical agents [31].

II.2.2.2. Inside Body Factors:

Many conditions inside the body make certain individuals more likely to develop cancer than others. For example, certain individuals either inherit or develop the following conditions: changed genes in the cells of the body, increased hormone levels in the bloodstream, e.g. estrogens, which are supposed to contribute to human breast cancer, and testosterone and its metabolites are the cause of human prostate cancer or compromised immune systems, e.g. in the case of Severe Combined Immune Deficiency (SCID). Some of these factors can make some people more susceptible to cancer disease [31].

II.2.3. Treatment of cancer:

There are however means to combat cancer disease. These complementary therapies are sometimes used on their own or in conjunction, depending on the type of cancer and its status. The purpose of these therapies is to make possible to remove the tumor and heal a patient with early stage cancer or like a chronic disease in order to monitor its growth. Common and newer forms of medication (surgery, radiation therapy, chemotherapy, targeted therapy, and immunotherapy) are predominantly associated with adverse outcomes which have a detrimental impact on quality of life. Thus, the battle for more successful, more tolerable anti-cancer therapy continues [34].

Chemotherapy is known to be the most effective and commonly used modality in most forms of cancer. Tumor cells have an improved capacity to divide and the standard of immortality because they are not controlled by apoptosis. Cell proliferation to cell death ratio is therefore high. Chemotherapy prevents tumor growth by killing off their ability to divide and

enforce apoptosis. The two branches of the chemotherapeutic drugs are also cytostatic (biological drugs) and cytotoxic [35].

II. 2. 3. 1. Glutathione-S-Transferase (GST):

Living organisms are constantly exposed to exogenous and endogenous toxic chemical species, which can cause harmful and often lethal effects. The ability of living organisms to survive the danger posed by such compounds is a fundamental biological adaptation for survival. Cells have implemented various methods to combat the effects of toxic substances and their metabolites. In this context, special enzymatic and non-enzymatic mechanisms are in place to protect cells from the destructive effects of toxic chemical species. Detoxification enzymes have a crucial function to play; making them less biologically active, more water-soluble and more easily removed from the body. Among others, the GST target was chosen for our analysis. Figure II.6 reflects the GST detoxification mechanism [36, 37].

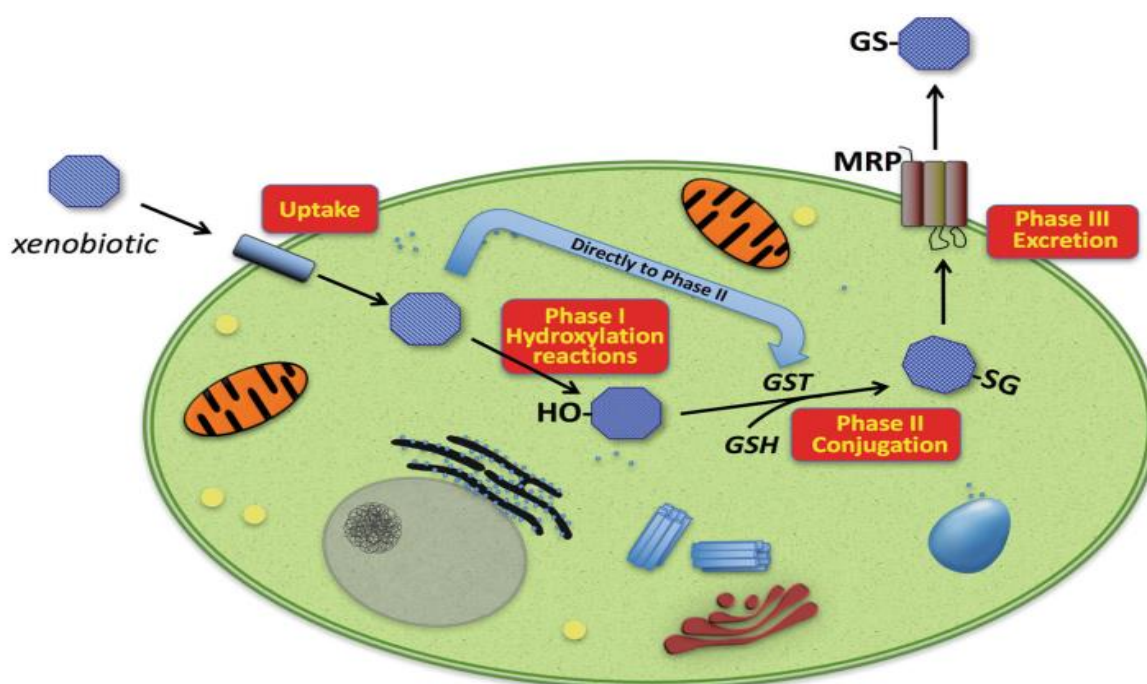


Figure II. 6 Overview of xenobiotic enzymatic biotransformation. Harmful molecules can migrate through the plasma membrane and, within the cells, may be attacked by the enzymes of the so-called Step I metabolism. The major ones belong to the Cytochrome P450 family, consisting of many enzymes that catalyze various reactions, including hydroxylation—the main reaction involved—oxidation and reduction. GSTs that catalyze the conjugation of phase I-modified xenobiotics to endogenous GSH play a key role in the resulting phase II metabolism. The conjugate obtained is then actively transported out of the cell by various transmembrane efflux pumps (Phase III). Any compounds can join the metabolism of phase II directly [37].

GSTs have many biological functions to play. They were defined as the most important enzymes involved in the metabolism of electrophilic compounds. They are classified as a family of phase II detoxification enzymes that metabolize a broad range of xenobiotic and end-of-

obiotic toxic compounds, which were classically defined as catalyzing the conjugation of glutathione (GSH) to electrophilic compounds through thio-ether linkages [38, 39].

GSTs are 50 000 Da, they are proteins formed by homodimers or heterodimers, each monomer has an active center consisting of 210 amino acids and two binding sites: A G-site in which glutathione (GSH) is bound and an H-site for an electrophilic substrate [40]. The GST detoxification reaction occurs by the following mechanism (Figure II. 7). In eukaryotes, there are three distinct families of GSTs separated by their cellular location: cytosolic, mitochondrial and microsomal (also known as membrane-associated proteins in eicosanoid and glutathione metabolism or MAPEG) [41]. Cytosolic GSTs are the most complex and closely related to the development of human diseases and are distributed and categorized into seven subtypes on the basis of their chemical, physical and structural properties. These subtypes are α , π , μ , θ , ω , σ , and δ . The cytosolic α , π , and μ classes are abundant and the most widely studied GST classes [37].

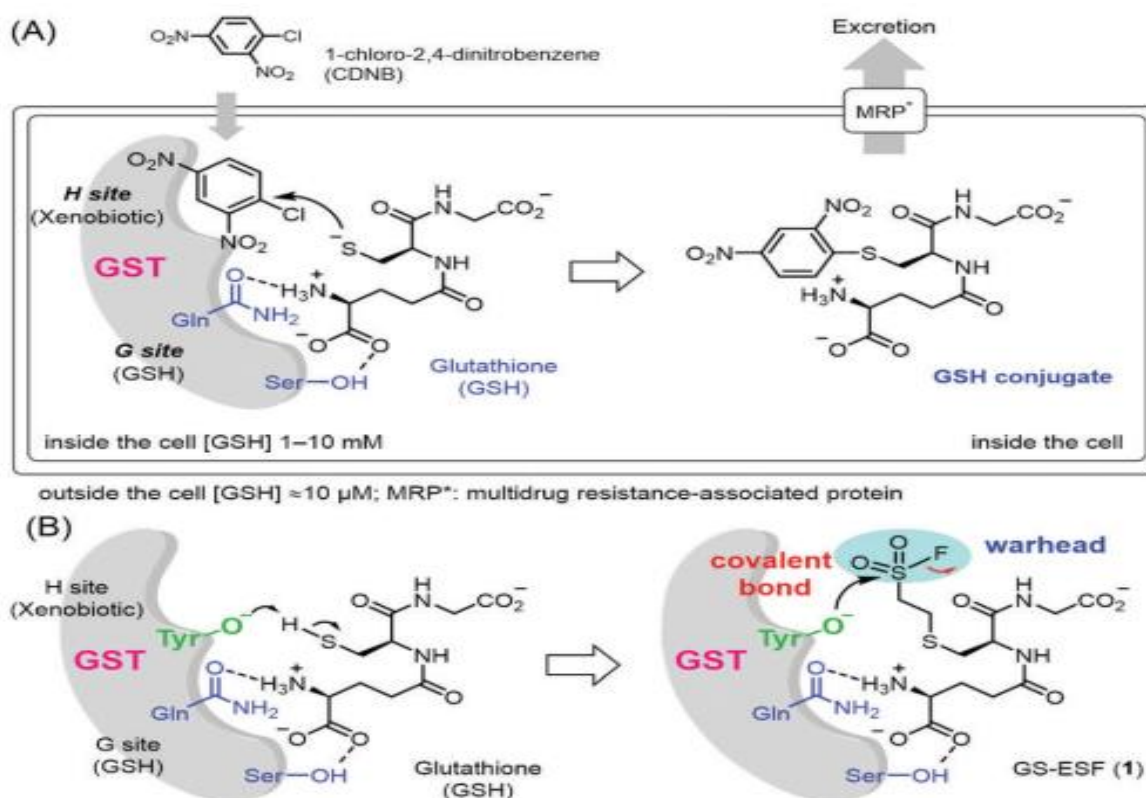


Figure II. 7 (A) GST detoxification process. (B) (left) GSH identification GST. (right) The molecular architecture of the GST covalent inhibitor [42].

II. 2. 3. 2. GST P1-1:

II. 2. 3. 2. 1. GST P1-1 physiological function (Role in cancer diseases):

The resistance of various human tumors to anti-cancer agents has been specifically associated with the conjugation of GST enzymes to GSH and the over expression of these enzymes. GSTs are implicated in resistance to many anticancer drugs in a wide spectrum of cancers. GST Pi (GSTP1-1) is the most prevalent, widely studied and highly expressed in several types of cancer cells (especially pancreatic, non-small cell lung, colon, liver, ovarian, breast, and lymphoma). Over expression of GST π can contribute to the defense of cancer cells against an attack by anticancer drugs. In which, tumor cells use GST π to form a GSH – X complex between antitumor drugs and GSH; the complex is excreted by Pgp and MRP out the cell. Synergistic interactions between GSTs and Pgp or MRPs are guiding the production of multidrug resistance in tumor cells [37, 43].

Recent literature has established GSH and other associated metabolic enzymes as essential to the cells safety from ROS via oxidation and redox mechanisms [46, 47]. Its enzymatic function is based on two aspects: the catalytic activity of Cys47 and Cys10, and the auto-S-glutathionylation of Cys47 and Cys10, both of which disrupt the subsequent interaction with C-Jun NH2-terminal kinase (JNK), causing the formation of a GST π multimer. Other members of the GSH-redox system, such as glutamate cysteine ligase, glutathione peroxidase and glutathione reductase, also play a significant role in this phase [46, 47].

In addition to metabolite detoxification, the first GST π described was initially defined as ligand binding properties due to its ability to interact covalently and non-covalently with different compounds, resulting in inhibition of conjugation activity [48]. GST π can induce cellular apoptosis by activating MAPK, MKK4, downstream JNK-signal components and p38 kinase, in the setting of cellular stress. Normal cells have low basal JNK activity to maintain optimum cell growth conditions. However, in the presence of oxidative or nitrosative stress, GST π can form homodimers to alter the reduced state of cysteine residues in its structure, resulting in JNK dissociation from the hetero-GST π – JNK complex and ensuing the subsequent activation of the c-Jun protein. These sequences of reactions will ultimately cause apoptotic pathways (Figure II. 8) [49, 50]. Further research suggests that GST π can affect the MAPK direction through both JNK and TRAF2 modulations [51].

Table II.2 Antitumor agents targeting GST π in background [42]

| Drugs | Selected examples | Functional significance |
|---------------------------------------|---|---|
| GSTπ inhibitors | EA and its analogs | Inhibiting the detoxification activity. Usually by binding to GST π substrate-binding sites |
| | TLK117/TLK199 NBDHEX and its analogs | Promoting tumor-cell apoptosis by activating the MAPK pathway and blocking the combination of JNK and GST π |
| GSTπ prodrugs | GSH or GSH derivatives (TLK286) | Catalyzed by GST π to release nitrogen mustard segment to induce tumor-cell apoptosis |
| | NO prodrugs (JS-K) | Catalyzed by GST π to release high-concentration NO to kill tumor cells directly |

Abbreviations: GST, glutathione S-transferase; EA, ethacrynic acid; NO, nitric oxide.

As already reported, GSTP1-1 is over expressed in several cancers where it protects cells from cell death by blocking the effects of JNK or its upstream activation. Indeed, the formation of both GSTP1-JNK and GSTP1-1-TRAF2 complexes has been identified in vivo [58]. NBDHEX (6-[7-nitro-2, 1, 3-benzoxadiazol-4-ylthio] hexanol) is designed as a "mechanism-based inhibitor" which has a potent effect on GST π . A number of compounds identified by ROTILI et al., containing NBD scaffolds which are not GSH peptidomimetics, are capable of inhibiting GSTP1-1 with a specific mechanism of action compared to other GST inhibitors [56].

NBDHEX binds the GSTP1-1 H-site and forms a GSH complex to inactivate the enzyme (Figure II. 9). Importantly, NBDHEX is also able to isolate GSTP1-1 from its JNK and TRAF2 complexes, thus allowing their activation. Drug combination studies have shown that NBDHEX is significantly active in cisplatin-resistant human osteosarcoma cells [57].

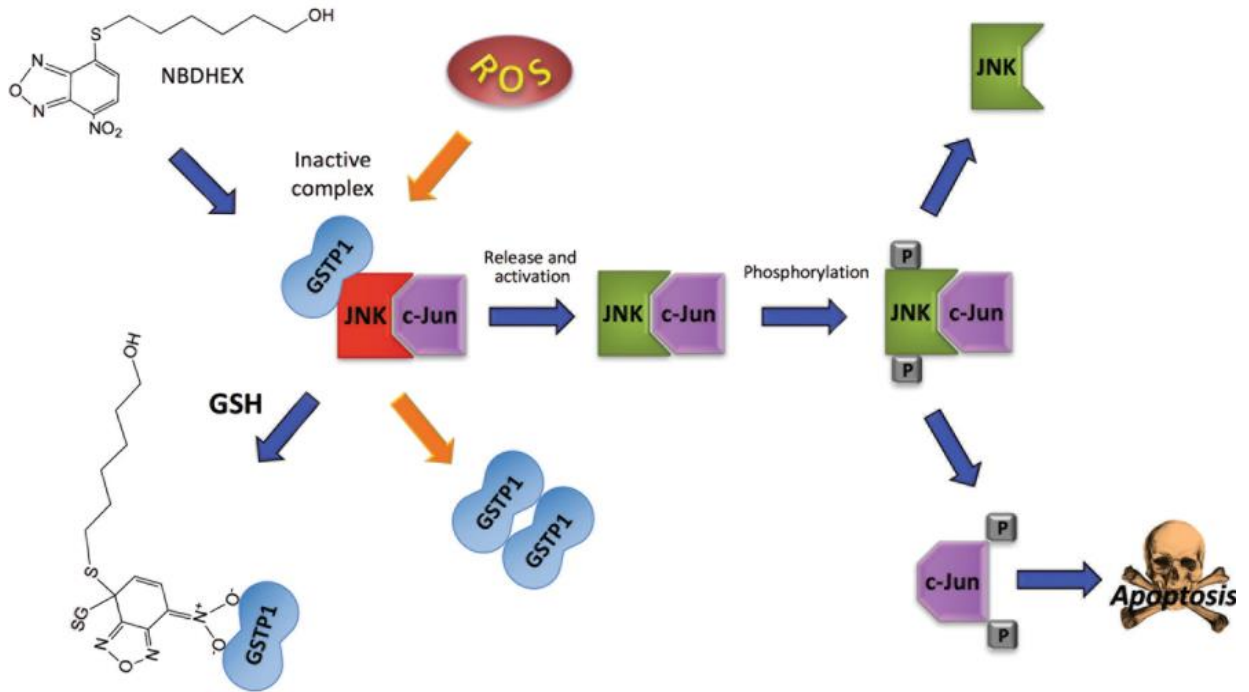


Figure II. 9 GSTP1-1 function in the JNK signaling pathway.

Monomeric GSTP1 prevents tumor cells from apoptosis by inhibiting the JNK signaling pathway via the development of a GSTP1-JNK-cJun complex that inhibits c-Jun phosphorylation. Under conditions of stress, GSTP1 can disassociate and dimerize from the complex, allowing JNK to phosphorylate c-Jun. This event can also be caused by a GST inhibitor NBDHEX that binds GSTP1 and induces its release from the complex [37].

II.3. References:

1. B. E. Blass, "Basic Principles of Drug Discovery and Development," *Basic Princ. Drug Discov. Dev.*, pp. 1–574, 2015, doi: 10.1016/C2012-0-06670-7.
2. P. Krogsgaard-Larsen, T. Liljefors, and U. Madsen, *Textbook of Drug Design and Discovery*, vol. 53, no. 9. 2002.
3. G. Kalyani, D. Sharma, and Y. Vaishnav, "A review on drug designing, methods, its applications and prospects.," *Int J Pharm Res Dev*, vol. 5, no. 5, pp. 15–30, 2013.
4. I. M. Kapetanovic, "Computer aided drug discovery and development: in silico-chemico-biological approach," *Chem. Biol. Interact.*, vol. 171, no. 2, pp. 165–176, 2008, doi: 10.1016/j.cbi.2006.12.006.COMPUTER-AIDED.
5. W. Loscher, H. Klitgaard, R. E. Twyman, and D. Schmidt, New avenues for anti-epileptic drug discovery and development. *Nat. Rev. Drug. Discov.*, vol. 12, no. 757. 2013.
6. N. Vernaz *et al.*, "Patented Drug Extension Strategies on Healthcare Spending: A Cost-Evaluation Analysis," *PLoS Med.*, vol. 10, no. 6, pp. 2012–2014, 2013, doi: 10.1371/journal.pmed.1001460.
7. B. M. Seddon and P. Workman, "The role of functional and molecular imaging in cancer drug discovery and development," *Br. J. Radiol.*, vol. 76, no. SPEC. ISS. 2, 2003, doi: 10.1259/bjr/27373639.
8. R. Panchagnula and N. S. Thomas, "Biopharmaceutics and pharmacokinetics in drug research," *Int. J. Pharm.*, vol. 201, no. 2, pp. 131–150, 2000, doi: 10.1016/S0378-5173(00)00344-6.
9. V. K. Gombar, I. S. Silver, and Z. Zhao, "Role of ADME Characteristics in Drug Discovery and Their In Silico Evaluation: In Silico Screening of Chemicals for their Metabolic Stability," *Curr. Top. Med. Chem.*, vol. 3, no. 11, pp. 1205–25, 2003, doi: 10.2174/1568026033452014.
10. <https://www.sigmaaldrich.com/technical-documents/articles/biology/target-identification-and-validation-for-early-drug-discovery.html>. (Accessed July 2020)
11. Quan Y, Xu H, Han Y, Mesplede T, Wainberg MA. JAK-STAT signaling pathways and inhibitors affect reversion of envelope-mutated HIV-1. *J Virol* 2017; 91.
12. G. Hessler and K. H. Baringhaus, "Artificial intelligence in drug design," *Molecules*, vol. 23, no. 10, 2018, doi: 10.3390/molecules23102520.
13. Y. Yang, S. J. Adelstein, and A. I. Kassis, "Target discovery from data mining approaches," *Drug Discov. Today*, vol. 17, no. SUPPL., pp. S16–S23, 2012, doi: 10.1016/j.drudis.2011.12.006.
14. S. Fox *et al.*, "High-throughput screening: Update on practices and success," *J. Biomol. Screen.*, vol. 11, no. 7, pp. 864–869, 2006, doi: 10.1177/1087057106292473.
15. T. L. Lemke and D. A. Williams, *Foye's Principles of Medicinal Chemistry*, no. 6. 2007.
16. Z. Huang, A. Inazu, A. Nohara, T. Higashikata, and H. Mabuchi, "Cholesteryl ester transfer protein inhibitor (JTT-705) and the development of atherosclerosis in rabbits with severe hypercholesterolaemia," *Clin. Sci.*, vol. 103, no. 6, pp. 587–594, 2002, doi: 10.1042/cs1030587.
17. H. R. Noori and R. Spanagel, "In silico pharmacology: drug design and discovery's gate to the future," *Silico Pharmacol.*, vol. 1, no. 1, pp. 1–2, 2013, doi: 10.1186/2193-9616-1-1.
18. B. K. Shoichet, S. L. McGovern, B. Wei, and J. J. Irwin, "Lead discovery using molecular docking," *Curr. Opin. Chem. Biol.*, vol. 6, no. 4, pp. 439–446, 2002, doi: 10.1016/S1367-5931(02)00339-3.
19. T. D. Y. Chung, D. B. Terry, and L. H. Smith, "In Vitro and In Vivo Assessment of ADME and PK Properties During Lead Selection and Lead Optimization – Guidelines, Benchmarks and Rules of Thumb," *Assay Guid. Man.*, no. Md, pp. 1–14, 2004, [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/26561695>.
20. W. Löscher, H. Klitgaard, R. E. Twyman, and D. Schmidt, "New avenues for anti-epileptic

- drug discovery and development,” *Nat. Rev. Drug Discov.*, vol. 12, no. 10, pp. 757–776, 2013, doi: 10.1038/nrd4126.
21. J. Chen, X. Luo, H. Qiu, V. Mackey, L. Sun, and X. Ouyang, “Drug discovery and drug marketing with the critical roles of modern administration,” *Am. J. Transl. Res.*, vol. 10, no. 12, pp. 4302–4312, 2018.
 22. M. H. Bailey *et al.*, “Comprehensive Characterization of Cancer Driver Genes and Mutations,” *Cell*, vol. 173, no. 2, pp. 371–385.e18, 2018, doi: 10.1016/j.cell.2018.02.060.
 23. M. Naghavi *et al.*, “Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: A systematic analysis for the Global Burden of Disease Study 2016,” *Lancet*, vol. 390, no. 10100, pp. 1151–1210, 2017, doi: 10.1016/S0140-6736(17)32152-9.
 24. <https://www.who.int/news-room/fact-sheets/detail/cancer>. (Accessed July 2020)
 25. C. E. Meacham and S. J. Morrison, “Tumour heterogeneity and cancer cell plasticity,” *Nature*, vol. 501, no. 7467, pp. 328–337, 2013, doi: 10.1038/nature12624.
 26. R. Fisher, L. Pusztai, and C. Swanton, “Cancer heterogeneity: Implications for targeted therapeutics,” *Br. J. Cancer*, vol. 108, no. 3, pp. 479–485, 2013, doi: 10.1038/bjc.2012.581.
 27. <https://www.gov.uk/government/publications/health-profile-for-england/chapter-2-major-causes-of-death-and-how-they-have-changed>. (Accessed July 2020)
 28. <http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-figures-2012>. (Accessed July 2020)
 29. J. Ma and A. Jemal, “Breast cancer statistics,” *Breast Cancer Metastasis Drug Resist. Prog. Prospect.*, vol. 9781461456476, no. 6, pp. 1–18, 2013, doi: 10.1007/978-1-4614-5647-6_1.
 30. D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: The next generation,” *Cell*, vol. 144, no. 5, pp. 646–674, 2011, doi: 10.1016/j.cell.2011.02.013.
 31. Susan Sieber Fabro, *CANCER AND THE ENVIRONMENT*. 2004.
 32. A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, C. J. L. Murray, and S. Asia, *Global Burden of Disease and Risk Factors Editors AND PACIFIC THE CARIBBEAN*. 2006.
 33. B. Kaczkowski, “Computational Cancer Biology: From Carcinogenesis to Metastasis,” 2012.
 34. E. J. Mun, H. M. Babiker, U. Weinberg, E. D. Kirson, and D. D. Von Hoff, “Tumor-treating fields: A fourth modality in cancer treatment,” *Clin. Cancer Res.*, vol. 24, no. 2, pp. 266–275, 2018, doi: 10.1158/1078-0432.CCR-17-1117.
 35. V. T. DeVita and E. Chu, “A history of cancer chemotherapy,” *Cancer Res.*, vol. 68, no. 21, pp. 8643–8653, 2008, doi: 10.1158/0008-5472.CAN-07-6611.
 36. P. Descartes, “Camille Savary Étude de la toxicité chronique et du potentiel cancérigène de contaminants de l ’ environnement séparément et en mélange sur les cellules HepaRG,” 2014.
 37. N. Allocati, M. Masulli, C. Di Ilio, and L. Federici, “Glutathione transferases: Substrates, inhibitors and pro-drugs in cancer and neurodegenerative diseases,” *Oncogenesis*, vol. 7, no. 1, 2018, doi: 10.1038/s41389-017-0025-3.
 38. E. Boyland and L. F. Chasseaud, “Enzyme-catalysed conjugations of glutathione with unsaturated compounds,” *Biochem. J.*, vol. 104, no. 1, pp. 95–102, 1967, doi: 10.1042/bj1040095.
 39. R. N. Armstrong, *Compr. Toxicol.* 2010, 4, 295–321.
 40. D. F. A. R. Dourado, P. A. Fernandes, B. Mannervik, and M. J. Ramos, “Glutathione transferase: New model for glutathione activation,” *Chem. - A Eur. J.*, vol. 14, no. 31, pp. 9591–9598, 2008, doi: 10.1002/chem.200800946.
 41. B. Mannervik and H. Jensson, “Binary combinations of four protein subunits with different catalytic specificities explain the relationship between six basic glutathione S-transferases in rat liver cytosol,” *J. Biol. Chem.*, vol. 257, no. 17, pp. 9909–9912, 1982, doi: 10.1016/s0021-9258(18)33960-7.
 42. Y. Shishido *et al.*, “A covalent G-site inhibitor for glutathione S-transferase Pi (GSTP1-1),” *Chem. Commun.*, vol. 53, no. 81, pp. 11138–11141, 2017, doi: 10.1039/c7cc05829b.

43. S. C. Dong et al., "Glutathione S-transferase π : A potential role in antitumor therapy," *Drug Des. Devel. Ther.*, vol. 12, pp. 3535–3547, 2018, doi: 10.2147/DDDT.S169833.
44. V. I. Lushchak, "Glutathione Homeostasis and Functions: Potential Targets for Medical Interventions," *J. Amino Acids*, vol. 2012, pp. 1–26, 2012, doi: 10.1155/2012/736837.
45. Z. Ye, J. Zhang, D. M. Townsend, and K. D. Tew, "Biochimica et Biophysica Acta Oxidative stress , redox regulation and diseases of cellular differentiation ☆," *BBA - Gen. Subj.*, vol. 1850, no. 8, pp. 1607–1621, 2015, doi: 10.1016/j.bbagen.2014.11.010.
46. K. D. Tew, Y. Manevich, C. Grek, Y. Xiong, J. Uys, and D. M. Townsend, "The role of glutathione S-transferase P in signaling pathways and S-glutathionylation in cancer," *Free Radic. Biol. Med.*, vol. 51, no. 2, pp. 299–313, 2011, doi: 10.1016/j.freeradbiomed.2011.04.013.
47. T. Seefeldt et al., "Characterization of a novel dithiocarbamate glutathione reductase inhibitor and its use as a tool to modulate intracellular glutathione," *J. Biol. Chem.*, vol. 284, no. 5, pp. 2729–2737, 2009, doi: 10.1074/jbc.M802683200.
48. K. D. Tew, D. M. Townsend, E. Therapeutics, S. Carolina, B. Sciences, and S. Carolina, "To Detoxification," vol. 43, no. 2, pp. 179–193, 2015, doi: 10.3109/03602532.2011.552912.Regulatory.
49. L. Gate, R. S. Majumdar, A. Lunk, and K. D. Tew, "Increased Myeloproliferation in Glutathione S-Transferase π -deficient Mice Is Associated with a Deregulation of JNK and Janus Kinase/STAT Pathways," *J. Biol. Chem.*, vol. 279, no. 10, pp. 8608–8616, 2004, doi: 10.1074/jbc.M308613200.
50. A. Sau et al., "Targeting GSTP1-1 induces JNK activation and leads to apoptosis in cisplatin-sensitive and -resistant human osteosarcoma cell lines," *Mol. Biosyst.*, vol. 8, no. 4, pp. 994–1006, 2012, doi: 10.1039/c1mb05295k.
51. L. Zhang, K. Blackwell, A. Altaeva, Z. Shi, and H. Habelhah, "TRAF2 phosphorylation promotes NF- κ B-dependent gene expression and inhibits oxidative stress-induced cell death," *Mol. Biol. Cell*, vol. 22, no. 1, pp. 128–140, 2011, doi: 10.1091/mbc.E10-06-0556.
52. J. D. Hayes, J. U. Flanagan, and I. R. Jowsey, "Glutathione transferases," *Annu. Rev. Pharmacol. Toxicol.*, vol. 45, no. February 2005, pp. 51–88, 2005, doi: 10.1146/annurev.pharmtox.45.120403.095857.
53. D. Burg, D. V. Filippov, R. Hermanns, G. A. Van der Marel, J. H. Van Boom, and G. J. Mulder, "Peptidomimetic Glutathione Analogues as Novel γ GT Stable GST Inhibitors," *Bioorganic Med. Chem.*, vol. 10, no. 1, pp. 195–205, 2002, doi: 10.1016/S0968-0896(01)00269-3.
54. W. Harshbarger et al., "Structural and biochemical analyses reveal the mechanism of glutathione S-transferase Pi 1 inhibition by the anti-cancer compound piperlongumine," *J. Biol. Chem.*, vol. 292, no. 1, pp. 112–120, 2017, doi: 10.1074/jbc.M116.750299.
55. L. Federici et al., "Structural basis for the binding of the anticancer compound 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol to human glutathione S-transferases," *Cancer Res.*, vol. 69, no. 20, pp. 8025–8034, 2009, doi: 10.1158/0008-5472.CAN-09-1314.
56. A. Ascione et al., "The glutathione S-transferase inhibitor 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol overcomes the MDR1-P-glycoprotein and MRP1-mediated multidrug resistance in acute myeloid leukemia cells," *Cancer Chemother. Pharmacol.*, vol. 64, no. 2, pp. 419–424, 2009, doi: 10.1007/s00280-009-0960-6.
57. A. Sau et al., "Targeting GSTP1-1 induces JNK activation and leads to apoptosis in cisplatin-sensitive and -resistant human osteosarcoma cell lines," *Mol. Biosyst.*, vol. 8, no. 4, pp. 994–1006, 2012, doi: 10.1039/c1mb05295k.

**"Knowledge always win in the end, but not unless and until it is known." – Professor John
McMurtry**

Chapter 3:

Computer-Aided Drug Design and Discovery

III.1. Generality:

As mentioned in chapter II (Section.1), medical chemists have often struggled with the difficult problem of determining which compounds to synthesize. There are several ways to classify hits, which can then be used as a starting point for hit to lead optimization. Computer Aided Drug Design (CADD) techniques provide a time-consuming and economical tool for the discovery of novel active compounds. It is a theoretical methodology that uses computer-based techniques. This strategy has become the most commonly used technique to substantially reduce the number of compounds to be synthesized and tested in vitro by predicting which would be inactive and active.

CADD methods can be divided into two main strategies, in which the drug target or known active compounds are used to find novel compounds having likely the desired effect (Figure III.1),

namely: (a) ligand-based methods which depend on the similarity of compounds of interest to active compounds, and (b) receptor-based methods which focus on the complementarity of the compounds of interest with the binding site of the target protein [1].

The computational chemist has the laudable aim of developing these two different forms of CADD approaches by using some kind of computer program capable of automatically evaluating very large compound libraries. However, the combination of different structural and ligand-based design techniques in drug discovery efforts has been established to be more successful than any single strategy, as both approaches are capable of complementing their strengths and weaknesses [2].

CADD approaches are currently very popular, trying to identify new hits in the R&D process for new drugs. They streamline the discovery process of new compounds, when data are available on one or more reference ligands, or a 3D-structure of protein–ligand complex. It is possible to cite numerous successful examples of these approaches, which are contributed in particular to the marketing of an anti-cancer drug, gefitinib (Iressa®), and LY-517717 compound, a factor Xa inhibitor from a virtual structure-based screening, which reaches phase II of clinical trials [3].

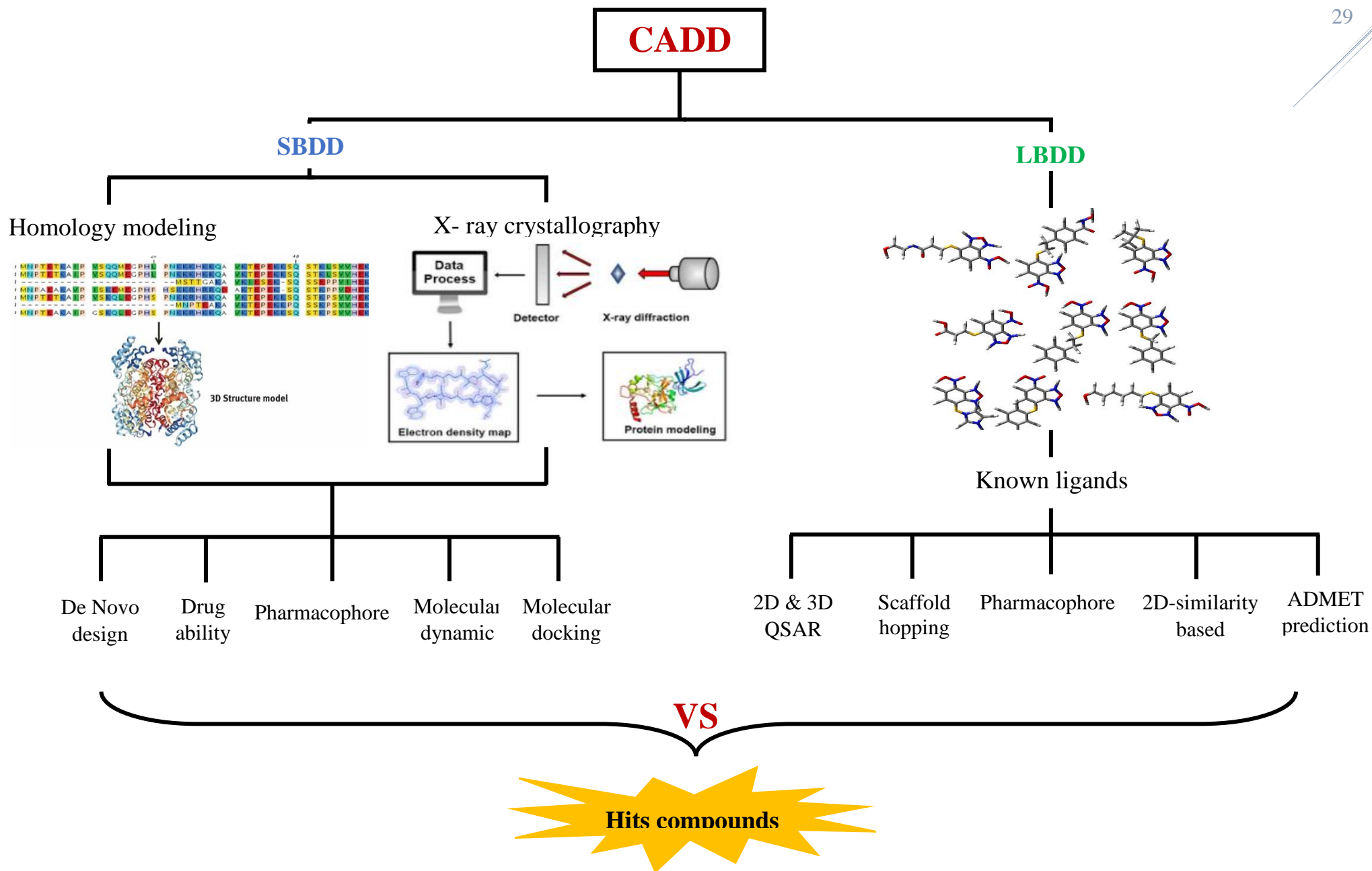


Figure III.1. Workflow of ligand -based drug design (LBDD) and structure -based drug design (SBDD).

Part I: Ligand-based drug design

III.2. Ligand-based drug design (LBDD):

III.2.1. QSAR analysis:

Quantitative Structure–Activity Relationship (QSAR) Analysis is one of the commonly used methods in ligand-based drug design processes to explain the quantitative relationship between structural molecular properties (descriptors) and their functions, e.g. biological activities, physicochemical properties, toxicity, or other kinds of activities on their molecular characteristics [4].

The structure of any chemical compound is determined by its properties. The main concept of QSAR is that identical or more precisely similar molecules have similar properties. In other words, a "small" change in the chemical structure of any compound leads to a change in their biological activities [5]. QSAR proposes that if a group of chemical compounds shows the same mechanism of action against the target, the modification of biological activity often changes chemical, structural and physical properties [6].

The basic formalism of the QSAR method will result from statistical analyses. The simple mathematical relationship is defined as follows (Eq III.1) [7]:

$$\text{Function} = f(\text{structural molecular or fragment properties}) \quad \text{Eq III.1}$$

During the QSAR analysis, the creation of models follows a general workflow, starting with dataset collection and the generation of chemical descriptors to be used as independent variables. After the removal of descriptors, which value varies little or not, across all molecules, the Multivariate study finally conducts a statistical validation of the model(s) to ensure its reliability (Figure III.2) [8].

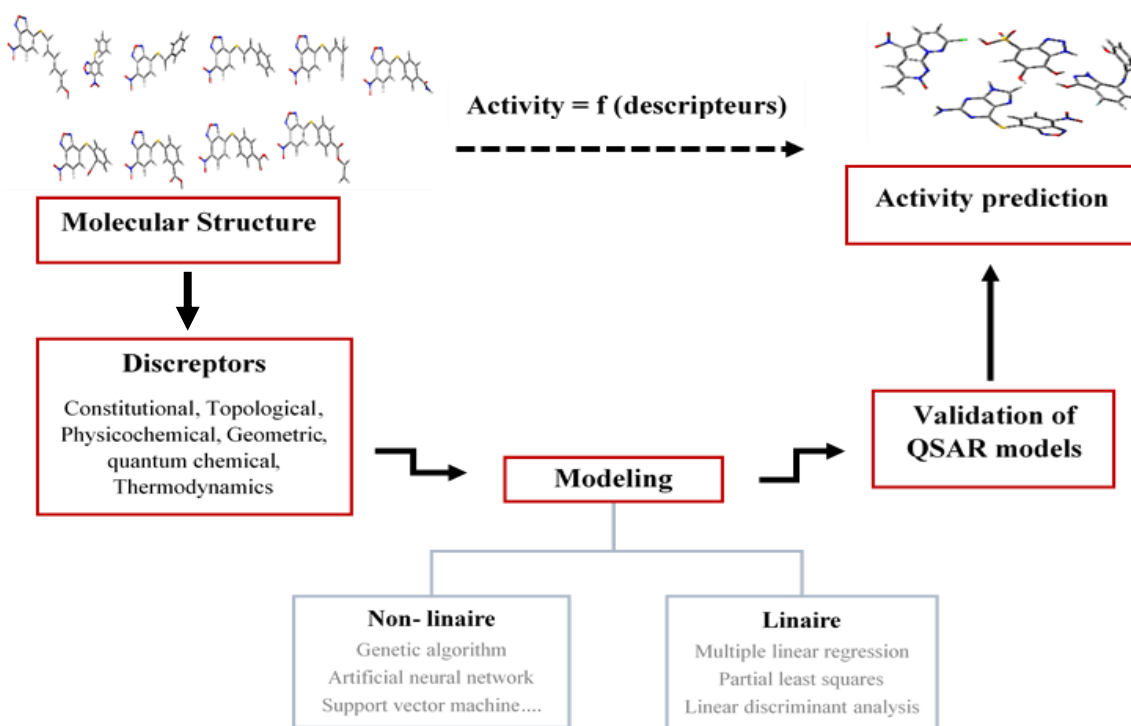


Figure III.2. Flowchart of the methodology used in QSAR study.

Currently there are more than 5000 descriptors that can be used in QSAR studies. They may be classified into various groups that can be derived from the chemical structure or from the use of suitable software, based on the various dimensions of the molecular descriptor; this could be divided into 2D-QSAR, 3DQSAR, 4D-QSAR, and so on. It can also generally be categorized as QSAR receptor-independent (RI) and QSAR receptor-dependent (RD), based on the availability of target receptors in the model construction process. So, this division decides the form or the medical modalities [9].

III.2.2. Object of QSAR study:

The main objective of the QSAR study is the rational creation of a mathematical model, followed by an examination of the involved chemical information, in order to gain insight into the mechanism and behavior of the system to be studied [10].

It is also useful in identifying alternative modes of action, in selecting useful structural features, in preparing new design methodologies, in developing new drugs and in helping to formulate new hypotheses for future research studies. As a result, QSAR reduces costs, time and

human capital to make the pharmaceutical product available to patients. QSAR models are also used in anticipation of pharmacokinetic and pharmacodynamics properties. QSAR also predicts properties, such as: permeability, and solubility. In this thesis, the objective of the QSAR model is to enable the estimation of the biological activities of unknown or novel chemical compounds to provide insight into the specific and consistent chemical properties or descriptors (2D/3D) describing the biological activity [11].

III.2.3. Steps involved in QSAR study:

The production of a good quality QSAR model depends on several factors, such as: the quality of biological data, the selection of descriptors and the used statistical analysis. Given the technical advancements and the wider availability of different statistical methods and types of descriptors, it is now relatively easy and straightforward to create a statistically accurate model [12].

III.2.3.1. Data collection and selection of training set:

The process of identifying accurate, initial, meaningful and potentially useful data arrangement is called data mining. This can include data collection, data cleaning, data engineering, algorithm engineering, algorithm running, result assessment, and information utilization . At the time of data collection, the same test procedure must be followed to bypass inter-laboratory shifts. [13- 15]. In order to retrieve a good collection of QSAR data, the following steps should be considered, namely:

- The number of chemical substances required should be appropriate.
- The biological activity of the chemical substances should be evenly distributed.
- The activity spectrum of action should be spread between the least active and extreme active chemical substances.
- The list of data set should have a diversity dose response relationship [16].

It is critically important for any QSAR model that the training set chosen to calibrate the model shows a well-balanced distribution and contains representative compounds. This calibration can be accomplished by a systematic collection of the training set, where the key structural features are systematically and simultaneously varied. In addition, there should be a proper ratio between the number of chemical substances in the training and the test set lists. The statistical molecular

design, self-organizing map, clustering, selection of Kennard-Stone, exclusion of spheres, and so on are few of the different techniques available to divide the data into training and test sets [17].

III.2.3.2. Molecular Descriptors used in QSAR:

Molecular descriptors are concepts defining the specific information of the chemical substances being studied. They are the result of a logical and mathematical process, which translates chemical information encoded within the symbolic representation of a substance into a useful number. The useful number should therefore be correlated with different physical properties, chemical reactivity or biological activity. This mathematical representation must be invariant with the size of the molecule and the number of atoms in order to allow modeling using statistical methods [18].

There are generally different types of descriptors being used during QSAR. Descriptors may be classified in a number of ways, including: constitutional, topological, geometric, quantum chemical and physicochemical one. Therefore, the majority of QSAR scientists prefer to classify the types of descriptors in terms of their dimensions. In view of this element, Table III.1 offers a valuable example of largely used molecular descriptors depending on dimensions [5]. The key advantage of calculating theoretical descriptors using sophisticated software is that they can be produced even for those compounds which are not yet synthesized [19]. Table III.1 shows molecular descriptors widely used depending on various dimensions.

Table III.1. Popularly known molecular descriptors dependent on various dimensions [20].

| Dimension of descriptors | Parameters |
|--------------------------|--|
| 0D | Constitutional indices, molecular property, atom, and bond count. |
| 1D | Fragment counts, fingerprints. |
| 2D | Topological, structural, physicochemical parameters including thermodynamic descriptors. |
| 3D | Electronic, spatial parameters, MSA parameters, MFA parameters, RSA parameters. |
| 4D | Volsurf, GRID, Raptor, etc. derived descriptors. |
| 5D | These descriptors consider induced-fit parameters and aim to establish a ligand-based virtual or pseudoreceptor model. These can be explained as 4D-QSAR 1 explicit representation of different induced-fit models. Example: flexible-protein docking. |
| 6D | These are derived using the representation of various solvation circumstances along with the information obtained from 5D descriptors. They can be explained as 5D-QSAR 1 simultaneous consideration of different solvation models. Example: Quasar. |
| 7D | They comprise real receptor or target-based receptor model data. |

III.2.3.3. Variable selection methods:

The performance of QSAR thus depends not only on the consistency of the initial collection of active/inactive compounds, but also on the choice of descriptors and the capacity to produce an acceptable mathematical relationship [21].

However, a single molecule can be represented in several ways by computing thousands computational descriptors using many algorithms and sophisticated software. Several of these descriptors catch the same details at times and are directly linked to each other. Therefore, the selection of descriptors requires a great deal of expertise for the QSAR modeler to choose the right ones for model creation [22]. Models can be constructed, using all the measured descriptors, but there may be several explanations to choose only a subset of them, such as:

- Prediction of model accuracy can be enhanced by eliminating obsolete and unnecessary descriptors.
- The QSAR model to be developed is also easier and theoretically quicker when fewer input descriptors are used and the interpretability of the correlation between the descriptors and the observable activity may be improved.
- If the number of selected descriptors is high relative to the number of chemical substances of interest, the effective number of degrees of freedom could be too large to determine the accurate calculation of the parameters of the QSAR model.
- Several machine learning approaches are more time-consuming than linear in the number of chemical substances and/or the number of descriptors, which prevents the study of data sets of several hundred descriptors [23].

For this purpose, specializing in the QSAR modeler; descriptors with a constant value for all observations and descriptors with a very low variance may be omitted. Just one descriptor for those exhibiting a strong degree of reciprocal correlation should be maintained. Descriptors which display a very poor connection with the biological activity can also be omitted in order to thin the descriptor pool. In certain situations, an effective scaling of the descriptors might also be necessary [24, 19].

III.2.3.4. Development of QSAR model:

After the elimination of correlated and obsolete descriptors, the next step is to pick the descriptors to be used in the created model. In general, methods for designing the QSAR model could be divided into two groups: (i) Classical variable selection and (ii) Variable selection by artificial intelligence algorithms [25]. The first group focuses on linear methods by considering a linear interaction between independent variables (descriptors) and dependent variables (biological activity). However, in the case of artificial intelligence methods, non-linear techniques are used to pick independent variables (descriptors) and help to solve some of the shortcomings of classical methods. The selection methods are grouped into two categories as showing in the follow section [26].

III. 2. 3. 4. 1. Linear regression:

a. Multiple linear regression (MLR):

MLR is one of the most common and basic methods used to create QSAR models, making it easy to understand the features used in model creation. In the MLR method, a linear relationship is formed between the compounds (activity/property/toxicity), Y, and the number of independent variables, X, typically molecular descriptors [27]. The simplified expression (Eq III. 2) of the MLR equation will be as follows:

$$Y = a_0 + a_1 \times X_1 + a_2 \times X_2 + a_3 \times X_3 + \dots + a_n \times X_n \quad \text{Eq III. 2}$$

Where X_1, X_2, \dots, X_n are independent variables or molecular descriptors present in the model with the associated regression coefficients a_1, a_2, \dots, a_n (for molecular descriptors 1 to n) and a_0 is the constant term of the model [28].

The primary drawback of MLR is that it can require collinear descriptors, which may refer to a regression model with incorrect regression coefficients. In contrast, the number of features chosen does not exceed the number of observations used for model development [29].

b. Partial least squares regression (PLS):

PLS is the simplest method of quantitative multivariate modeling. It implies a linear relationship between two data matrices, X (dependent variables) and Y (independent variables) (target variable). In comparison to MLR, PLS provides advantages such as it can be useful in the

study of data with highly collinear, noisy and several X variables, as well as in the simultaneous simulation of several target variables Y [30]. PLS is based on the premise that the analyzed system relies on the latent variables.

PLS eliminates the difficulty of collinear features by removing these latent factors from a wide collection of descriptors, which provide the critical information needed to model the target (response variable). The latent variables T (known as X-score) and U (Y-score) are derived from the large collection of descriptors and the responses (biological activity). The obtained latent variable T (X-score) is used to predict the U (Y-score) and, then, the U (Y-score) is used to predict the response (biological activity) [31].

The number of latent factors used in PLS is a key factor for QSAR modeling and is typically accomplished by the use of cross-validation approaches, such as: n-fold cross-validation and leave-one-out methods, where a portion of the samples is used as a training set, while the other portion is set aside as a test set to verify the model which was developed from the training set [32].

III. 2. 3. 4. 2. Non-linear regression:

Artificial neural networks (ANN):

ANN are the most common and widely studied soft computing techniques. They are a family of mathematical models being focused on the workings of the human brain. However, in addition to some neurological understanding, it has been shown to be an important method for solving nonlinear problems in much scientific research, ranging from technology to biological applications [33]. To each of these entries (inputs) is associated a weight (W_i), representative of the forces of the connection. Each elementary processor has a unique output, which then branches out to supply a variable number of downstream neurons [34]. The network consists of several basic units called neurons, grouped in a certain topology, and connected to each other. Neurons have been arranged into layers. Based on their location, there are three layers so-called input layers, hidden layers and output layers. Each of them has separate relations and functions, namely the transfer function, the learning rule and the connection formula [35].

Four procedures (Figure III.3) are done in a computer neuron. The first one is the input and output process, which compares the input signals from the former layer neurons, decides the strength of each input and transmits the output signal toward the next layer neurons. The next one

is the sum function, which measures the number of the cumulative input signals according to equation III.3 [36]:

$$i_j = \sum w_{ij} \times o_i \quad \text{Eq III.3}$$

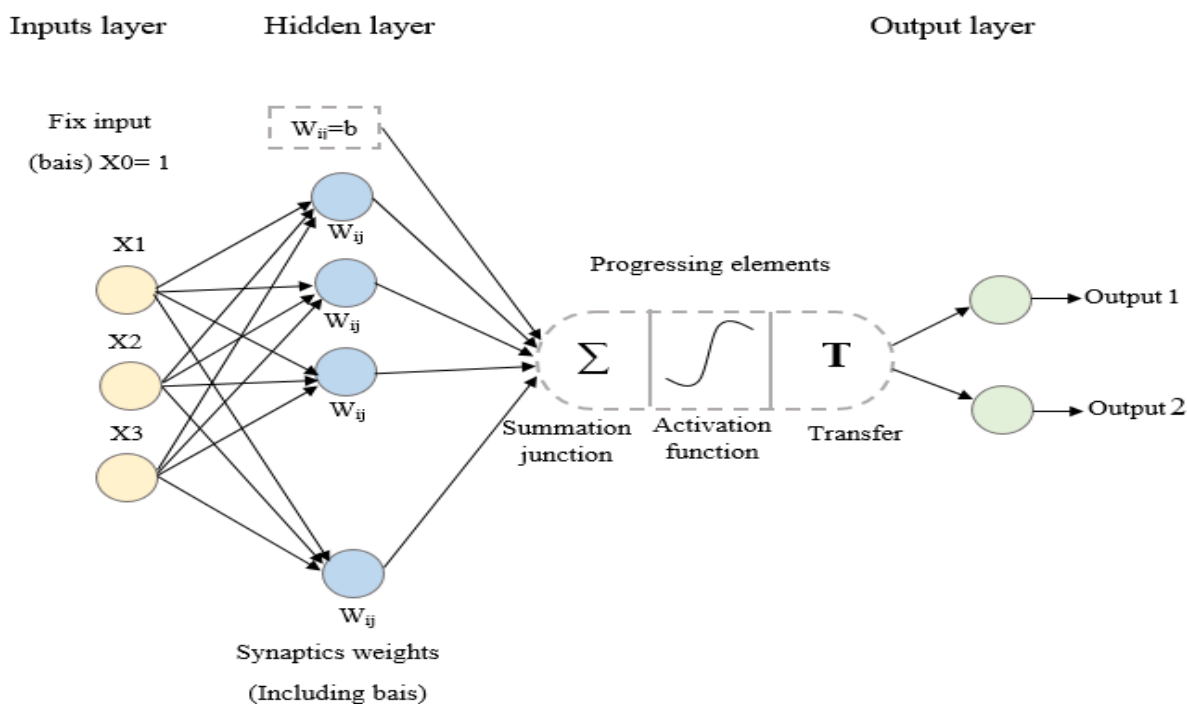


Figure III.3. Graphical view of an artificial neural network with one input layer (comprising three descriptors) attached to the hidden layer with the necessary weights and the output layer.

Where i_j is the net entry in node j (of, say, layer λ), while o_i is the output of node I in the previous layer ($\lambda-1$); and w_{ji} is the weight associated with nodes i and j . The third one is the activation function, which causes outputs to change from time to time. The overview result is transferred to this function before the conversion function is entered. The final factor is the transfer function that maps the summed input to the output value. There are many possible forms for the transfer function, which are threshold functions, sigmoid functions and linear functions (figure III. 4). The sigmoid logistic function is the most used, because it represents a good compromise between the threshold and linear functions [37], and incorporates the nonlinearity feature in the mapping process:

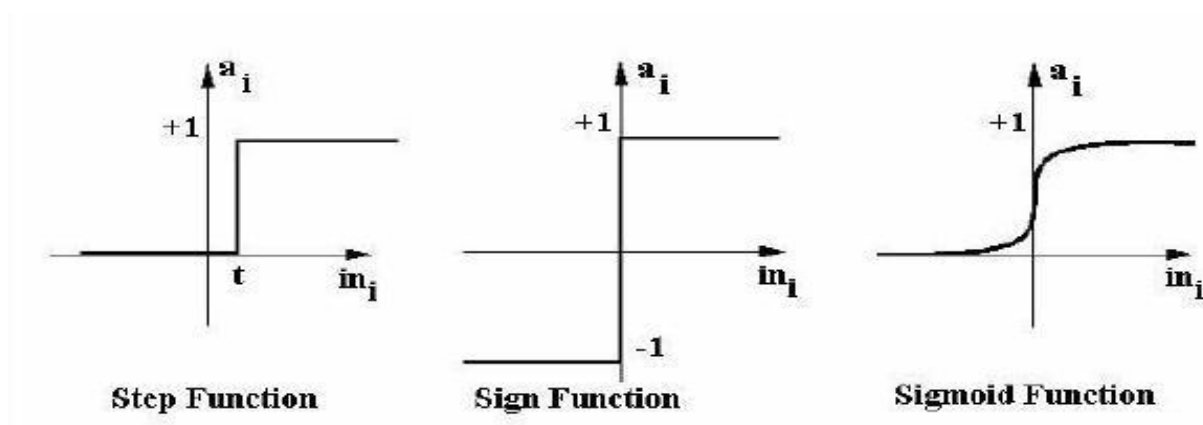


Figure III.4. Activation functions generally used for neural networks study.

An ANN will contain a variety of hidden layers. These units are composed of a regression equation that transforms input information into non-linear output data.

In order to promote the development of alternative approaches, the OECD (Organization for Economic Cooperation and Development) has recently developed rules for the validation of QSAR/QSPR models [18, 19]. The assessment of each of the five concepts is an essential step for presenting models relevant to the experimental plan, which was the purpose of this study.

- Principle 1—A defined endpoint
- Principle 2—An unambiguous algorithm
- Principle 3—A defined domain of applicability
- Principle 4—Appropriate measures of goodness-of-fit, robustness and predictivity
- Principle 5—A mechanistic interpretation, if possible [38].

III. 2. 3. 5. Validation of QSAR model:

The next progress, following the development of the QSAR model, is to verify the acceptability and reliability of the QSAR model predictions. The evaluation of QSAR regression model success in fitting, robustness and external prediction is crucially important [39]. The requisite condition for the validity of the regression model is that the multiple correlation coefficients R^2 is as close as possible to one and the standard error of the estimation is small; although the former is not essentially a very good predictor of fitness. Apart from the use of fitness parameters, the validation of the QSAR models consists of four major parameters [40] (1) Internal validation. (2) Validation by data division in training and testing samples. (3) External validation

applying the model to outside data. (4) Data randomization. Table III.2. Demonstrate Validation parameters and their threshold values.

Table III.2. Mathematical equation of statistical validation metrics used in QSAR studies [41].

| parameter | Formula | Threshold |
|--------------|---|-----------------------------------|
| R^2 | $1 - \frac{\sum(Y_{obs} - Y_{cal})^2}{\sum(Y_{obs} - \bar{Y}_{tran})^2}$ PRESS = $\sum(Y_{obs} - Y_{pred})^2$, $SSY = \sum(Y_{obs} - \bar{Y}_{tran})^2$ \bar{Y}_{tran} is the mean observed activity of the training set compounds | $R^2 > 0.6$ |
| R^2_{adj} | $\frac{\{(n-1)R\} - P}{n-p-1}$ N est le nombre des observations (les molécules) ; est le nombre de variables indépendantes (les descripteurs) ; est le coefficient de détermination du modèle. | $R^2_{adj} > 0.6$ |
| Q^2_{LOO} | $1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{tra})^2}$ Y_{obs} is the observed response, Y_{pred} is the calculated response, n defines the total number of compounds and predictor variables is denoted as p. | $Q^2_{LOO} > 0.6$ |
| F | $\frac{\frac{\sum(Y_{cal} - \bar{Y})^2}{p}}{\frac{\sum(Y_{obs} - Y_{cal})^2}{n-p-1}}$ Y_{obs} is the observed response, Y_{calc} is the calculated response, n defines the total number of compounds and predictor variables is denoted as p. | $F > F$ of fisher table |
| SE | $\sqrt{\frac{\sum(Y_{obs} - Y_{cal})^2}{n-p-1}}$ Y_{obs} and Y_{calc} are the observed (experimental) and estimated scores respectively, while n is the number of compounds and p is the number of descriptors | SE should be low for a good model |
| R^2_{pred} | $1 - \frac{\sum(Y_{obs} - Y_{cal})^2}{\sum(Y_{obs} - \bar{Y}_{tran})^2}$ | $R^2_{pred} > 0.6$ |
| Q^2_{F1} | $1 - \frac{\sum(Y_{obs}(test) - Y_{cal}(test))^2}{\sum(Y_{obs}(test) - \bar{Y}_{tran})^2}$ | $Q^2_{F1} > 0.5$ |
| Q^2_{F2} | $1 - \frac{\sum(Y_{obs}(test) - Y_{cal}(test))^2}{\sum(Y_{obs}(test) - \bar{Y}_{test})^2}$ | $Q^2_{F2} > 0.5$ |
| Q^2_{F3} | $1 - \frac{[\sum(Y_{obs}(test) - Y_{pred}(test))^2] / n_{test}}{[\sum(Y_{obs}(tran) - \bar{Y}_{tran})^2] / n_{train}}$ n_{train} and n_{test} denote the number of training set and test set compounds, respectively | $Q^2_{F3} > 0.5$ |

| | | |
|--------------------|--|---|
| CCC | $\frac{2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})}$ | CCC < 1 |
| | $Q_{tran}^2 > 0.5$ $R_{test}^2 > 0.6$ | |
| | $\frac{r^2 - r_0^2}{r^2} < 0.1 \text{ and } 0.85 \leq k \leq 1.15 \text{ or } \frac{r^2 - r_0'^2}{r^2} < 0.1 \text{ and } 0.85 \leq k' \leq 1.15$ $ r_0^2 - r_0'^2 < 0.3$ | |
| | $\overline{r_m^2} = (r_m^2 + r_m'^2)/2 \text{ and } \Delta r_m^2 = r_m^2 - r_m'^2 $ <p>Where $r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2})$</p> $r_m'^2 = r^2 \times (1 - \sqrt{r^2 - r_0'^2})$ <p>The parametre r^2 and r_0^2 are defined as follows:</p> $r_0^2 = 1 - \frac{\sum(Y_{obs} - k \times Y_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{obs})^2} \text{ \& } r_0'^2 = 1 - \frac{\sum(Y_{pred} - k' \times Y_{obs})^2}{\sum(Y_{pred} - \bar{Y}_{pred})^2}$ <p>The term k and k' are defined as:</p> $k = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum(Y_{pred})^2} \text{ \& } k' = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum(Y_{obs})^2}$ | $\Delta r_m^2 < 0.2$ $r_m^2 > 0.5$ |
| rm2 metric | <p>The Y_{obs} and Y_{pred} values have been scaled at the beginning using the following formula:</p> $Y_{i(scaled)} = \frac{Y_i - Y_{\min(obs)}}{Y_{\max(obs)} - Y_{\min(obs)}}$ <p>rand r_0^2 are the squared correlation coefficients between the observed and (leave-one-out) predicted values of the compounds with and without intercept respectively.</p> | |
| $r_{m^2}^{(rank)}$ | $r_{(rank)}^2 \times \left(1 - \sqrt{r_{(rank)}^2 - r_0^2(r_{rank})}\right)$ | |
| \overline{Rr} | An average of the correlation coefficient for randomized data | $\overline{Rr} < 0.5$ |
| $\overline{Rr^2}$ | An average of determination coefficient for randomized data | $\overline{Rr^2} < 0.5$ |
| $\overline{Qr^2}$ | An average of leave one out cross-validated determination coefficient for randomized data | $\overline{Qr^2} < 0.5$ |
| ${}^c R_p^2$ | $R^2 \times (1 - \sqrt{ R^2 - \overline{Rr^2} })$ | ${}^c R_p^2 > 0.6$ |
| MAE | $\frac{1}{n} \times \sum Y_{obs} - Y_{pred} $ | <ul style="list-style-type: none"> • Good predictions: $MAE \leq 0.1 \times \text{training set range}$, AND $MAE + 3 \times \sigma \leq 0.2 \times \text{training set range}$ • Bad prediction: $MAE > 0.15 \times \text{training set range}$; OR $MAE + 3 \times \sigma > 0.25 \times \text{training set range}$ |

III. 2. 3. 6. Applicability domain (AD):

The application of every QSAR model to new compounds is indirectly limited by the fact that the model is derived from a certain finite set of molecules: the training set. Therefore the prediction of a modeled response using QSAR is only valid if the compound being predicted falls beyond the model AD, since it is hard to anticipate the entire universe of compounds using a specific QSAR model [42]. The general description of the applicability domain (AD) was coined by Netzeva and colleagues [43]: "The applicability domain of the QSAR model is the response and chemical structure space in which the model makes predictions with certain reliability". The AD is a valuable tool for the accurate implementation of QSAR models, while the characterization of the interpolation space is relevant in the definition of the AD. AD (Figure III. 5. a) is an area in the chemical space containing physicochemical, electronic or biological knowledge on which the model training set is developed [44]. Various methods are in place to assess the AD of QSAR models. From the QSAR publications of the last decade, the most widely used method for estimating interpolation regions is the leverage approach (Williams plot Figure III. 5. b). A compound would be found outside the applicability domain if the leverage value is greater than the critical value of $3p/n$, where p is the number of model variables plus 1 and n is the number of objects used to construct the model [45].

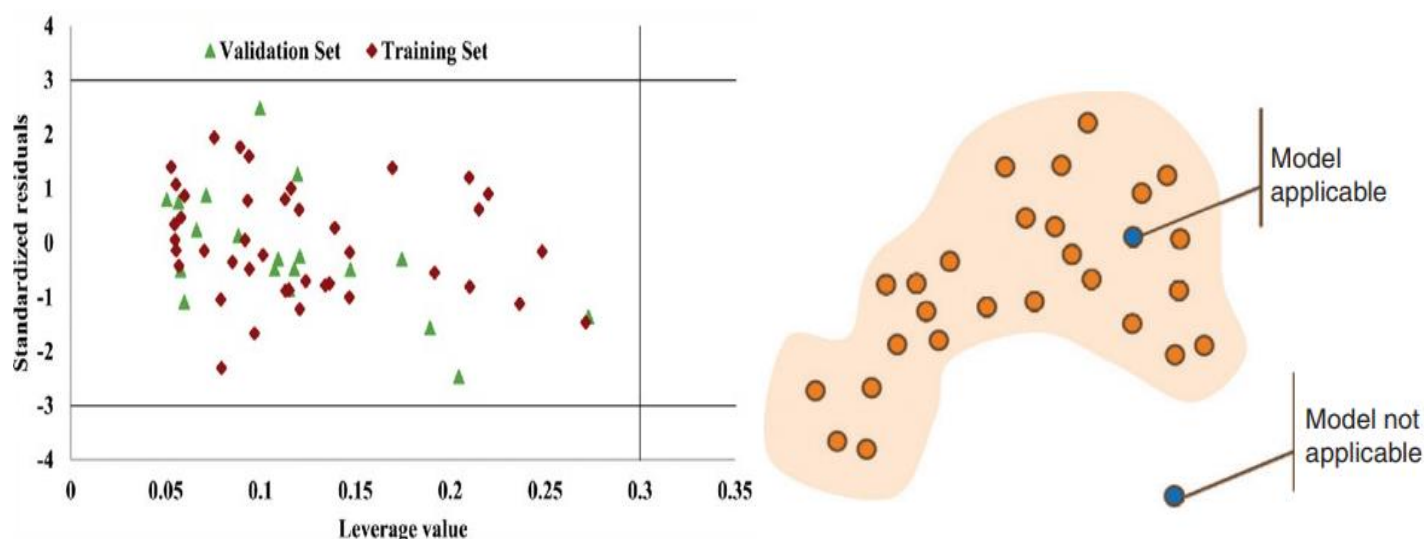


Figure III. 5. a) Schematic overview of the application domain. Every ringed dot is a single data point used for model training. New Chemical structures (solid dots) that fall into the inner, darker field are close enough to the training set and the model can be used confidently. The latest substances that fall in the white region are so far from the training collection that the formula can no longer be used. b) Williams' plot for the applicability domain of QSAR model [46].

III.2.2. Chemical similarity analysis

The quest for molecular similarities is a key concept for the drug discovery based in Ligand. Its aim is to identify and discovery new chemical substances with identical structures and bioactivity to query compounds. Chemical similarity is based on the idea that two identical molecules are likely to share similar bioactivity and physical properties. This search strategy was used to narrow large datasets of chemical substances to a smaller number by measuring and comparing the similarity coefficients between the known active compound and the compounds being screened. Molecular similarity has also been used to optimize the efficacy and pharmacokinetic properties of lead compounds, based on their structure–activity relationship [33]. A number of computational chemical similarity search algorithms have been developed. Chemical substructure fingerprints, Non-hashed structural fingerprints, such as: MACCS keys or Obabel FP3 are the most widely used methods [48].

Some forms of molecular representations have been used in similarity searches, including physiochemical properties, topological indices, molecular diagrams, pharmacophore characteristics, molecular shapes, molecular fields, and so on. When the molecular descriptors for the molecules of interest have been determined and the compounds translated to suitable data representations, the next step is to measure the chemical similarity using a distance metric. These metrics can be measured using one of the different methods. However, in the case of binary chemical fingerprints, the most common is the Dice coefficient, the Cosine coefficient, and the easiest and most straightforward distance calculation is the Tanimoto coefficient [49]. The coefficients are so-called "association" when they take, each one, their value in the interval [0; 1], thus making it possible, by simple subtraction, to convert a distance metric into a similarity metric and vice versa [50].

Focused on structural representation, molecular similarity techniques can be mainly divided as 2D or 3D similarity techniques. 2D similarity methods rely only on 2D structural knowledge and are among the easiest, most effective and most common similarity search methods. In order to overcome shortcomings associated with 2D similarity techniques, several techniques have been developed that account for 3D molecular conformation when performing similarity searches [51]. These techniques include form similarity, 3D fingerprints, field-based molecular methods and pharmacophore modeling, which were considered to be the most commonly used tool. A description of the pharmacophore modeling is included in the next section.

III.2.3. Ligand_ Based Pharmacophore:

Pharmacophore simulation is another important and effective approach used today in CADD. This method allows researchers to conduct simulated screening on vast ligand databases and to accomplish the key goals of the CADD, which are meant to identify and/or design new drug candidates for use as new medicines or to design new drugs that are supposed to be superior to current treatments [52].

The definition of a pharmacophore was previously proposed by Pual Ehrlich in 1909. It is known as "a molecular framework which carries (phoros) the essential features responsible for a drug's (pharmacon's) biological activity" [17] This description has been revised by Peter Gund after a century of creation: "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" [18]. According to the very recent description of the IUPAC [2], there is a pharmacophore model "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" [15, 39]. One which explains it more easily: a pharmacophore is a spatial arrangement of functional groups important for biological activity; a pattern arising from a collection of molecules having biological activity. As a result:

- The pharmacophore identifies the basic, electronic and steric function identifying points required for an optimum relationship with the related pharmacological objective.
- A pharmacophore is not a real combination of functional groups nor a particular compound, but a purely abstract term which represents for the common compounds a potential interaction of a group of compounds against their binding site [53].

Focused on the superposition of a series of inactive and active compounds, either a ligand-based pharmacophore model may be created. The goal of these methods can be summarized as the recognition of important features to be found in active compounds and, therefore, not found in inactive compounds. In a structure-based approach, by exploring potential contact sites between the biological active site and the molecules, Pharmacophore methods have been widely utilized in

scaffold hopping, 3D database search, VS, ligand profiling, fragment modeling, pose filtering and pharmacological predictive activities Figure III. 6 [54].

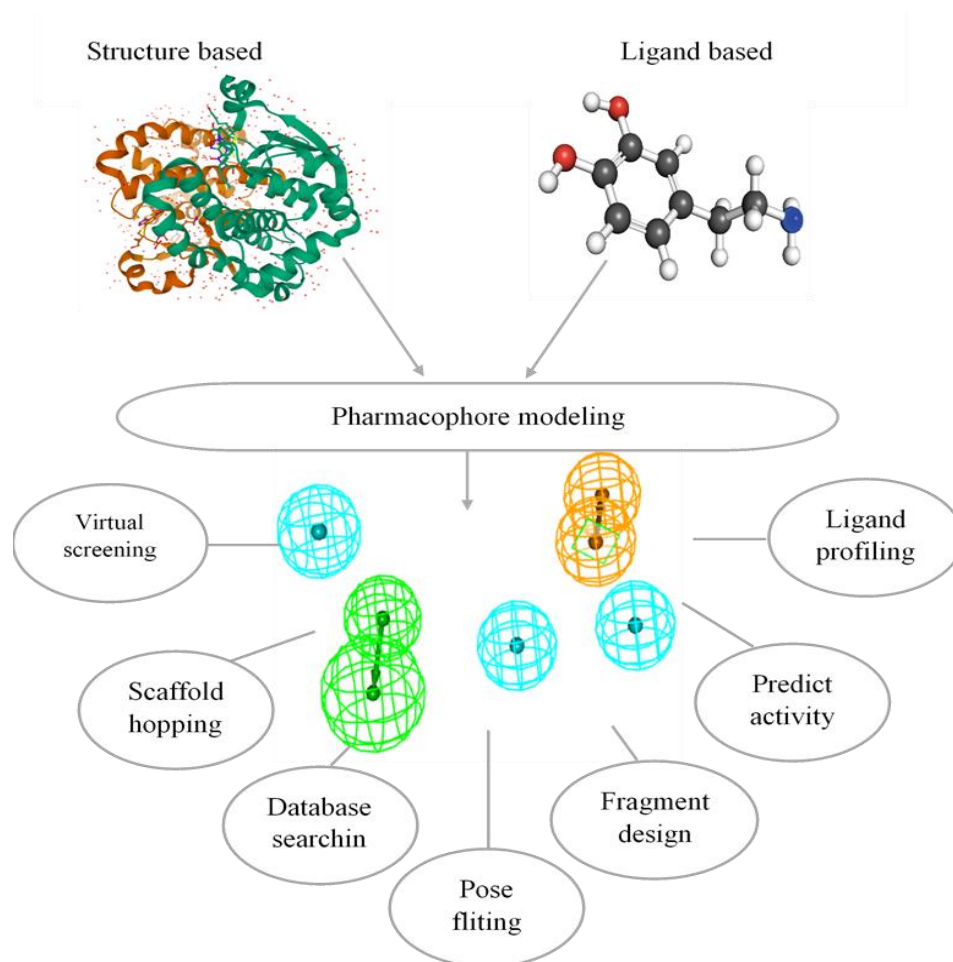


Figure III.6. General framework of pharmacophore methods.

Ligand-based pharmacophore simulation is an efficient strategy to promote drug development and discovery where the target structure is unknown, indicating potential pharmacophore queries based on a collection of aligned active chemical structures, in which the choice of these ligands has a high effect on the subsequent pharmacophore model. Therefore, these recognized ligands are compared in the next step, and the typical chemical features of their 3D structures are removed, reflecting the critical interaction between the compounds and the macromolecular structure. This technique can be used to produce a pharmacophore model from a variety of ligands (training set chemical structures) and typically includes the following key steps: (i) A collection of suitable conformers have been created for each ligand in the training set list to reflect conformational versatility within the DS Diverse conformation generation module, using

the conformational analysis, conformational algorithm, and (ii) alignment of multiple ligands in the training set and generation of the ligand-based pharmacophore model [21, 55] .

There are two different kinds of techniques which are documented in the published works to define the important common chemical characteristics for the design of pharmacophore models. Similar Features Pharmacophore Modeling [56] uses similar chemical features is found on the most active chemical structure; 3D QSAR Pharmacophore Modeling [57] is determined by the chemical properties of the most active and inactive chemical structures and their subsequent biological action. The 3D-QSAR technique differs from the Common Feature Pharmacophore method, since there is no restriction on the number of training chemical structures set and the technique does not need empirical biological activity measures in comparable bioassay conditions [58].

(iii) The final stages are the evaluation of the pharmacophore model by: Fischer's randomization test, cost analysis and the estimation of biological activity of the test set list. The performance of the model is defined in terms of fixed costs, total costs and null costs. The fixed cost is the simplest model which suits the data perfectly [59].

Part II: Structure-Based Drug

Design

III.3. Structure-based drug design (SBDD):

III. 3. 1. Molecular Docking:

Many enzymes or proteins are targets for essential bioactive chemical agents in the treating of plant, animal and human related diseases. The interaction between biologically important small chemical compounds (so-called ligands) and protein or enzyme targets provides an important role in the maintenance of protein functions. Molecular docking is an analysis of how two or even more molecular structures (e.g. drugs and enzyme or proteins) interacting together [60]. Molecular docking is among the most widely used techniques in SBDD once 3D-target protein and binding site are available. This methodology is able of predicting and defining, with a high level of precision, the conformation and low-energy binding mode of small-molecule ligands within the required protein or enzyme target binding site.

Fischer suggested first ever docking strategy for binding ligand receptor research (Figure III. 7) demonstrates the "lock-and-key model"[3], which corresponds to a rigid docking. Ligand moves into the bending site of the target as a key, and the target acts as a lock to detriment the perfect position for the "key" to unlock the "lock". This model underlines the significance of geometric complementarity [61]. Therefore, the actual docking mechanism is so fluid that targets and ligands need to modify their orientation to match each other well. So, computer-simulated ligand binding attempts to determine the current best ligand binding mode for a proteins partner. It consists of producing a variety of potential conformations/orientations, i.e. ligand poses, and within the macromolecule binding site (here just proteins are considered). The existence of the three-dimensional structure of the protein is therefore a sufficient condition. For reliable docking analysis, an experimental structure (e.g. X-ray crystallography or NMR) or a structure extracted through computational methods (e.g. homology modeling) with high resolution is needed [62].

A chemical substance or ligand which is strongly joined by hydrogen bonds, van der Waal bonds or any potential electrostatic attractions connected with the receptor or protein target of the disease can block the role and then behave as a drug. Hydrogen bonds are localized electrostatic connections between atoms that contribute an important role in the detection of ligand binding to the protein. Calculation of correct protein-ligand interactions is the main concept behind structural drug development [63].

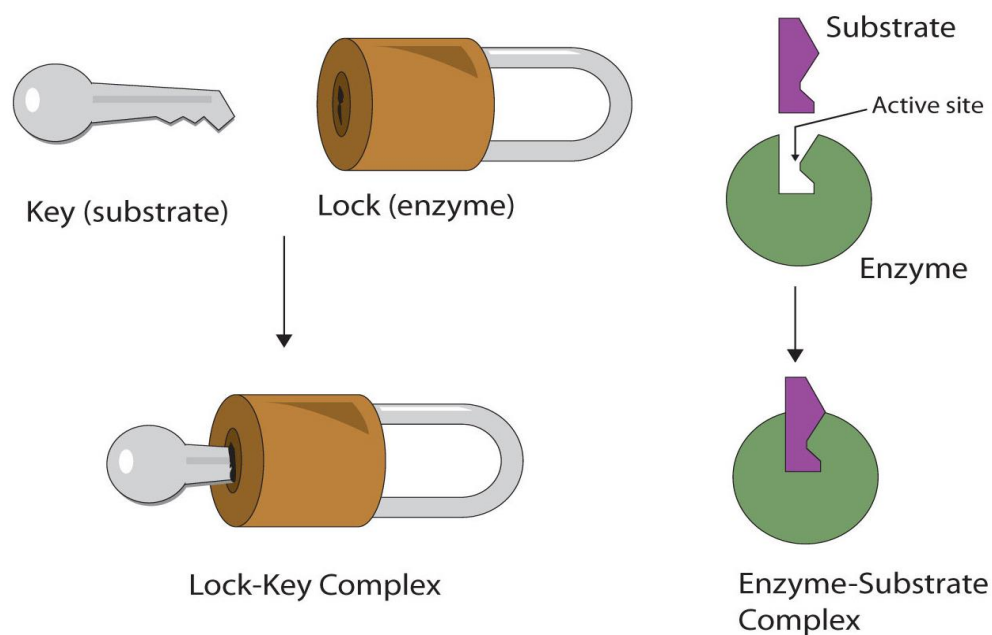


Figure III.7. Enzyme Activity Model Lock-and-Key [64].

III.3.1.1. Theory of docking:

Basically, the determination of the most possible binding orientations consists primarily of two parts: firstly, the use of a search algorithm to determine ligand orientations at the active site of the macromolecule, and, secondly, the scoring function, which correlates the score for each orientation (figure III.8). Preferably, search algorithms must be able to achieve the experimental interaction mode, and the score function will also be the best of all the generated orientations [65]. These tasks are performed by the software of this approach in a cyclical process.

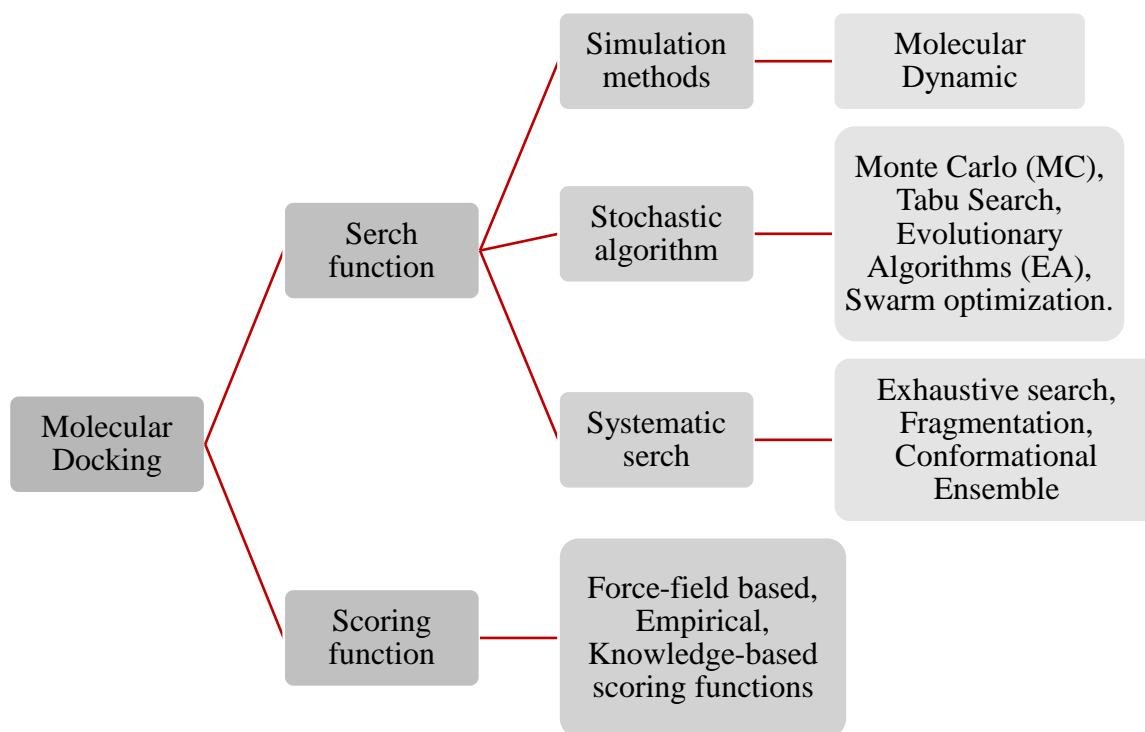


Figure III.8. Schematic explains the techniques used for protein ligand docking.

III.3.1.2. Search algorithm:

The algorithm can establish an optimal number of orientations where the structural properties of the compounds, such as torsional (dihedral), translational and rotational degrees of freedom, are adjusted progressively. A range of algorithms used for docking study could make the treatment of ligand flexible, and can be categorized into three essential categories: systematic methods (incremental construction, conformational search, databases); random or stochastic methods (Monte Carlo, genetic algorithms, tabu search); and simulation methods (molecular dynamics, energy minimization). The degrees of flexibility of the molecules involved in the calculation control the classification of molecular docking methods [66].

- The goal of these algorithms is to discover all the degrees of freedom in a ligand by rotating it from 0 to 360 ° for all single bonds using a picked incremental step. All the degrees of freedom of each coordinate are discussed in a combinatorial manner. As a consequence, the number of potential molecular conformations is measured by Eq III.4:

$$N_{\text{conformations}} = \prod_{i=1}^N \prod_{j=1}^{n_{\text{inc}}} \frac{360}{\theta_{i,j}} \quad \text{Eq III.4}$$

After multiple cycles of search and assessment, the minimal energy solution referring to the most probable binding mode will converge (Figure III.9.b). Despite the good quality of this system, it can converge to a local minimum instead of a global minimum. This downside can be solved by running simultaneous searches beginning from various points of the energy landscape (i.e., distinct conformations). These strategies are divided up into: Exhaustive Scan, conformational Ensemble, fragmentation [67].

- *Stochastic algorithms*: Adjust the values of the degrees of freedom randomly instead of systematically. The benefit of these strategies is speed, so they might theoretically find an optimal solution quickly. As a downside, they do not guarantee that the conformational space is fully checked, meaning that the real solution can be lost. The absence of convergence is partially overcome by increasing the number of iterations of the algorithm [62].

Stochastic algorithms work by creating random improvements to a specific ligand or a ligand group. For this purpose, the algorithm produces a series of molecular orientations and occupies a wide variety of energy landscapes (Figure III.9.c). This approach removes trapping the final solution at a minimum of local energy and raises the chance of a global minimum. As the algorithm supports wide coverage of the energy environment, the expense of processing associated with this technique is a significant constraint [33, 35]. The benefit of these strategies is efficiency, so they might theoretically find an optimal solution quickly. The most popular stochastic algorithms in the world are: Tabu search, Swarm optimization, Evolutionary Algorithms (EA), and Monte Carlo (MC) [67].

- *Simulation methods*: Molecular Dynamics are the most popular simulation technique, a process that explains the evolution of the system over time. This method is accomplished by changing each atom independently in the region of the rest of atoms, whereas the MD simulation reflects the versatility of both ligand and macromolecule more accurately than other algorithms. A wider explanation will be given in section III.3.2.

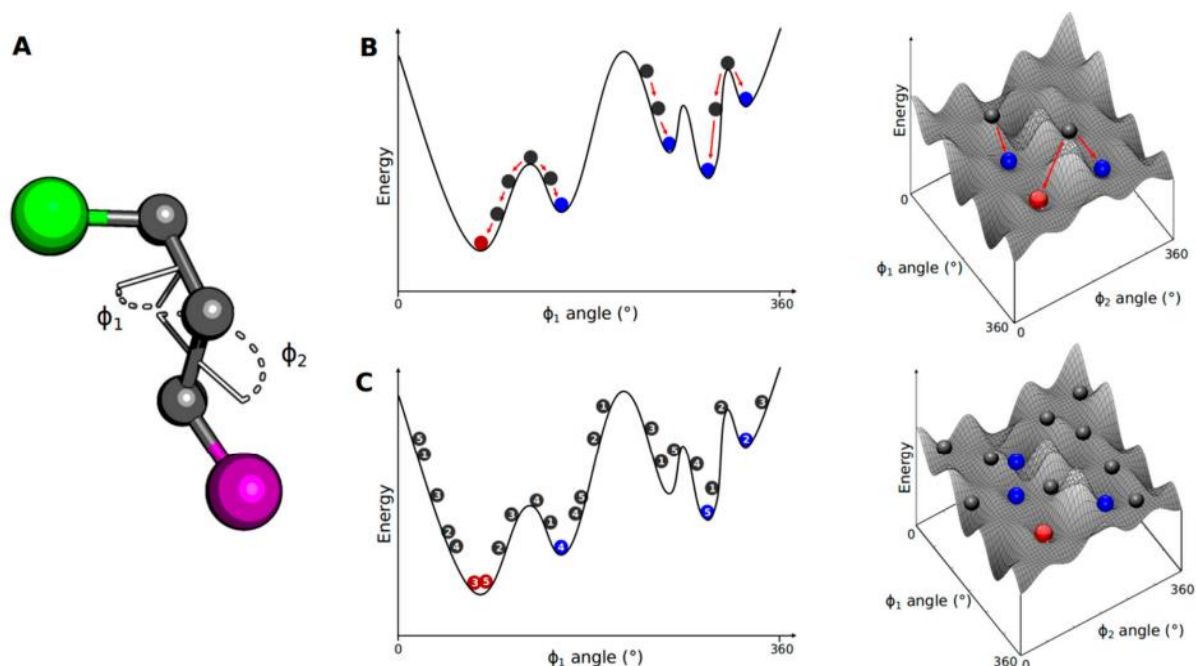


Figure III.9. Small-molecule conformational search methods. (A) A molecule containing two bulky groups (green and purple spheres) has its conformation defined by two internal dihedrals Φ_1 and Φ_2 ; (B) Considering Φ_2 as a frozen dihedral, the energy variation due to rotation of Φ_1 is plotted in a 1D energy landscape. The initial structure (grey spheres) is modified by changing Φ_1 , leading to a decrease in energy. The systematic search algorithm changes all structural parameters until a local (blue spheres) or global (red sphere) energy minimum is reached; (C) The stochastic search explores the conformational space by randomly generating distinct conformations, populating a broad range of the energy landscape. This procedure increases the probability of finding a global energy minimum [67].

III.3.1.3. Scoring:

After the creation of thousands of ligand configurations, the scoring functions are another significant feature that needs to be discussed since they have an important role in the choice of poses. It is used to distinguish putative correct poses from incorrect poses produced by the sampling engine or binders from inactive compounds in a fair computational period. The goal of any scoring method is to measure the free energy change of the creation of the ligand-receptor complex pose [60]. This could be described by the fundamental thermodynamics (Eq III.5) given by the binding constant (K_d) and the free energy of Gibbs (ΔG_L).

$$\Delta G = \Delta H - T\Delta S \quad \text{Eq III.5}$$

Where ΔH is the enthalpy change, T is the temperature of the system in Kelvin and ΔS is the entropy change [68].

Free-energy simulation methodologies have been improved for the computational modeling of protein-ligand interactions and the estimation of binding affinity. However, these costly measurements remain inefficient for the measurement of large quantities of protein-ligand complexes and are not always reliable. Scoring functions introduced in molecular docking programs make numerous assumptions and simplifications in the evaluation of modeled complexes and do not adequately account for a variety of physical processes that determine molecular recognition, e.g. entropic effects [68]. The conventional classification proposed by Wang et al. (2002) is classified into three major groups: force field-based, knowledge-based and empiric. In 2015, Liu and Wang suggest a new classification as following: physics-based, empirical, knowledge-based and machine learning-based. For that study of Jin li et al. indicates that the conventional classification is more general and is capable of categorizing the functions of score according to the major development plan embraced.

- *Force field based scoring functions:* It measures binding energy by integrating energy concepts from the classical force field of the bond (angle bending, dihedral variation and bond stretching) and non-bonded terms (electrostatic and van der Waals force) using equations of molecular mechanics [69]. It calculates both ligand internal energy and protein-ligand interaction energy by the electrostatic interactions described in the Coulomb function and van der Waals energy described in the Lennard Jones potential, where a distance-dependent dielectric can be added to mimic the solvent effect. Drawbacks include overestimation of binding affinity and arbitrary collection of non-bonded cutoff names, and the benefits of force field-based score functions include solvent accounting [70].
- *Empirical scoring functions:* are designed to replicate experimental affinity data; they are a separate type of assessment techniques. These operations are the sum of different empiric energy concepts, such as apolar interactions, ionic and hydrogen bonding, as well as entropic effects and desolvation, etc. For a first stage in the creation of an empiric function, a sequence of protein-ligand aggregates with defined binding affinities are used as a training set list to run a multiple linear regression study. The weight constants produced by the statistical model are then used as coefficients to modify the terms of the equation [71].
- *Knowledge-based scoring functions:* are other techniques used to test ligand-receptor binding energy. These approaches use mathematical analysis of interactive atom pairs from protein-ligand complex structure with accessible three-dimensional structures. These

potentials are built by considering the frequency where two separate atoms are observed over a specified distance in the structural data source. The various types of interactions found in the dataset are categorized and weighted by their frequency of occurrence. The total score value is the sum of these individual interactions. As knowledge-based functions do not rely on reproducing linking affinities (empirical methods) or ab initio measurements (force-field methods), they have an adequate balance between precision and speed [72].

Table III. 3. Examples of Scoring Function Formulae [66].

Scoring function formulate

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{hbond} \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{solv} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}$$

Extended force-field-based scoring function from AutoDock.

For two atoms i, j , the pair-wise atomic energy is evaluated by the sum of van der Waals, hydrogen bond, coulomb energy and desolvation. W are weight factor to calibrate the empirical free energy.

$$\Delta G = \Delta G_0 + \Delta G_{rot} \times N_{rot} + \Delta G_{hb} \sum_{neutral\ H-bond} f(\Delta R, \Delta \alpha) + \Delta G_{io} \sum_{ion\ int} f(\Delta R, \Delta \alpha) + \Delta G_{aro} \sum_{aro\ int} f(\Delta R, \Delta \alpha) + \Delta G_{lipo} \sum_{lipo\ cont} f^*(\Delta R)$$

Empirical scoring function from FlexX.

ΔG is the estimated free energy of binding; ΔG_0 is the regression constant; ΔG_{rot} , ΔG_{hb} , ΔG_{io} , ΔG_{aro} and ΔG_{lipo} are regression coefficients for each corresponding free energy term; $f(\Delta R, \Delta \alpha)$ is scaling function penalizing deviations from the ideal geometry; N_{rot} is the number of free rotate bonds that are immobilized in the complex.

$$PMF_{score} = \sum_{KI} A_{ij}(r) \quad A_{ij}(r) = -k_B T \ln \left[f_{vol-corr}^j(r) \frac{\rho_{seg}^{ij}(r)}{\rho_{bulk}^{ij}} \right]$$

Knowledge-based scoring functions PMF.

k_B is the Boltzmann constant; T is the absolute temperature; r is the atom pair distance. $f_{vol-corr}^j(r)$ is the ligand volume correction factor;

$\frac{\rho_{seg}^{ij}(r)}{\rho_{bulk}^{ij}}$ designates the radial distribution function of a protein atom of type i and a ligand atom of type j .

III.3.1.4. Molecular docking types:

- *Rigid docking*: This approach considers both ligand and protein to be like rigid structures, for that the search space is very small and limited. It includes only three translational and three rotational degrees of freedom during the searching process. This approximation is close to the "lock-key" binding model and is mostly used for docking protein where there are many orientation degrees of freedom too large to be looked for [46]. Throughout this case, ligand mobility may be handled with the use of a pre-computed range of ligand orientations or by allowing a degree of atom-atom overlap between the ligand and protein. In the general case of this technique, the binding site and the ligand are approximated by "hot" dots in these approaches and the overlap of the matching point is evaluated. Computations for rigid docking are easier to complete and do not reflect correct poses or global minimum orientation [73].
- *Semi-flexible docking*: In comparison to the previous approach, the flexibility of the ligand is reacquired where the orientation is changed; unlike the protein is solid entity. This type of docking is ideal for docking small molecules and macromolecules such as proteins or nucleic acids and small ligand compounds [74]. These approaches presume that the unaltered conformation of the protein may correspond to the one capable of recognizing the ligands to be docked. This hypothesis, as already stated, is not always confirmed. Since 1980s, various docking algorithms have been developed (table III.3) [75].
- *Flexible Docking* Its principle idea based on, during binding, a protein is not a passive solid agent and it considers ligand and protein to be flexible equivalents. During this method, the docking method allowed the orientation of the docking process (receptor and ligand) to be easily changed. In view of the fact that receptor and ligand variables rise in accordance with the number of atoms, multiple additional factors considerations need to be addressed. The fact is that the measurement is significant and that the docking procedure is too complex. Flexible docking is commonly used to reliably analyze the relationship (interaction) between compounds (generally ligand-receptor) accurately [76].

III.3.2. General on molecular dynamic:

Molecular dynamics (MD) methods play a key role in understanding and forecasting the function, structure and properties of molecular systems. They are a crucial method for predictive molecular design. They were suggested in the being by Alder and Wainwright in the last of 1950s

to model the interactions of hard spheres. Nowadays, however, MD strategies are used to research almost any form of nucleic acids, macromolecule—proteins, carbohydrates—of biological or medicinal importance. In short, the MD approach is based on Newton's second law or the motion equation (Eq III.5) for a group of atoms [77, 75].

$$m_i \frac{\delta^2 r_i}{\delta t^2} = F_i \quad \text{Eq III.5}$$

Where, F_i is the component of the net force acting on the i^{th} atom with a mass, m_i , and r_i denotes the position of the atom at time t . The force can then be computed as (Eq III.6):

$$F_i = - \frac{\delta U(r_1, r_2, \dots, r_n)}{\delta r_i} \quad \text{Eq III.6}$$

Where, $U(r_1, r_2, \dots, r_n)$ is the potential energy function of the specific conformation and can be described by using the concept of a force field with predefined parameters [78].

The concept "force field" is a statistical method used to incorporate the mathematical formula and the related parameters. There are complicated equations, but they are easy to determine. It consists of an empirical form of the interatomic potential energy U , and a series of parameters used to characterize the energy of the protein as a function of its atomic coordinates [77]. The product of the MD simulation is a series of snapshots or orientations named the trajectory of the system after a certain period of time; typically tens to a few hundred nanoseconds. These snapshots can be used to explain device dynamics and to measure macroscopic properties using statistical mechanics principles, some of which can be directly related to experimental results [77]. We carry out computer simulations in the hope of understanding the properties of compounds assemblies in terms of their structure and the microscopic interactions between them. The key benefit of the MD method is its ability to simulate the laboratory conditions in which a conventional biological issue is answered [79].

In this formalism, the atoms of the system are modeled as points with a given mass and charge. Charges are used to measure the electrostatic force field by which the force of each atom in the system can be measured. The force is then used to update the position of each atom using classical mechanics. This method is then iterated to adjust the configuration of the system. This approach enables us to acquire knowledge not only on the conformations explored by protein systems, but also on their dynamics Figure III.10 [80].

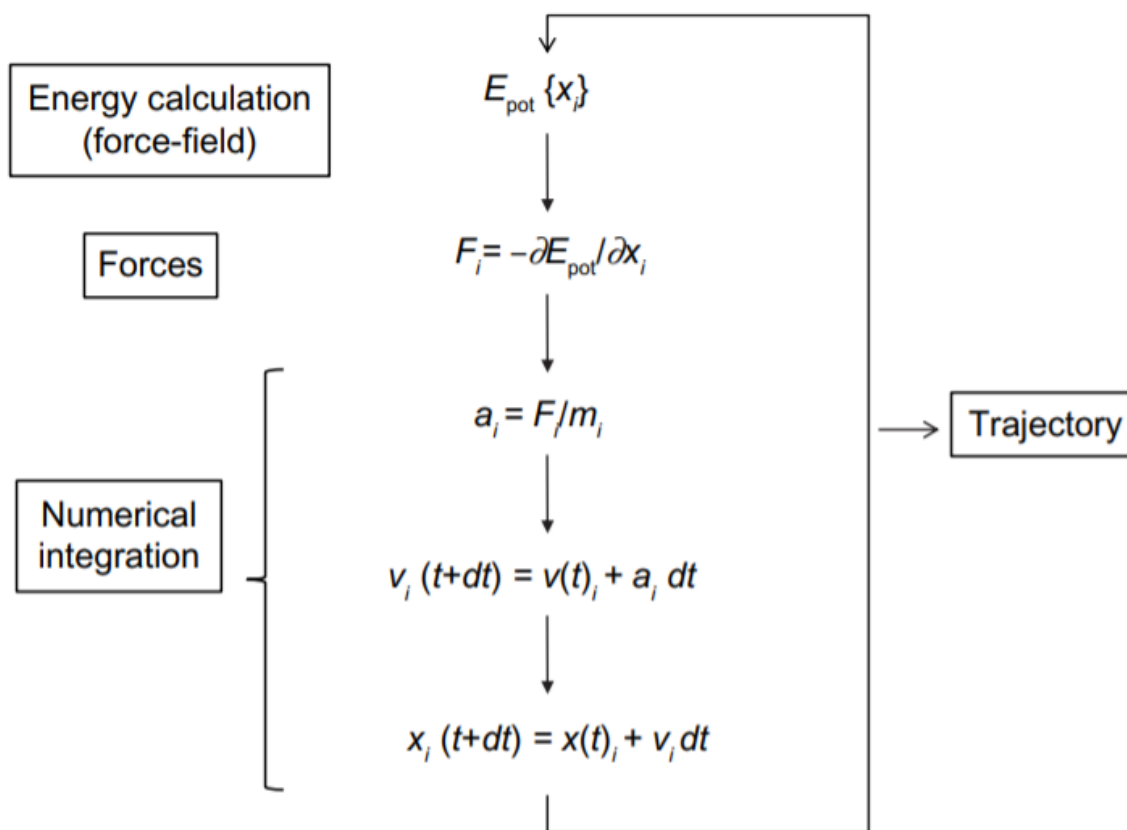


Figure III.10. Molecular dynamics basic algorithm. Notes: The simulation output, the trajectory, is an ordered list of $3N$ atom coordinates for each simulation time (or snapshot). Abbreviations: E_{pot} , potential energy; t , simulation time; dt , iteration time; For each spatial coordinate of the N simulated atoms (i): x , atom coordinate; F , forces component; a , acceleration; m , atom mass; v , velocity [80].

III.4. References:

1. A. V Veselovsky and A. Ivanov, "Strategy of Computer-Aided Drug Design Strategy of Computer-Aided Drug Design," no. September 2014, 2003, doi: 10.2174/1568005033342145.
2. C. Liao, M. Sitzmann, A. Pugliese, and M. C. Nicklaus, "Software and resources for computational medicinal chemistry," *Future Med. Chem.*, vol. 3, no. 8, pp. 1057–1085, 2011, doi: 10.4155/fmc.11.63.
3. N. Lagarde, "Importance De L ' Évaluation Et Application À," 2014.
4. I. M. Kapetanovic, "Computer aided drug discovery and development: in silico-chemico-biological approach," *Chem. Biol. Interact.*, vol. 171, no. 2, pp. 165–176, 2008, doi: 10.1016/j.cbi.2006.12.006.COMPUTER-AIDED.
5. C. Nantasenamat, C. Isarankura-na-ayudhya, and T. Naenna, "Review article : A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP," pp. 74–88, 2009.
6. M. V Putz, C. Duda-seiman, D. Duda-seiman, and A. Putz, "Chemical Structure-Biological Activity Models for Pharmacophores ' 3D-Interactions," 2016, doi: 10.3390/ijms17071087.
7. H. M. Patel, M. N. Noolvi, and P. Sharma, "CHEMISTRY Quantitative structure – activity relationship (QSAR) studies as strategic approach in drug discovery," pp. 4991–5007, 2014, doi: 10.1007/s00044-014-1072-3.
8. M. Ahmadi and M. Shahlaei, "Quantitative structure–activity relationship study of P2X7 receptor inhibitors using combination of principal component analysis and artificial intelligence methods," *Res. Pharm. Sci.*, vol. 10, no. 4, pp. 307–325, 2015.
9. L. C. Yee and Y. C. Wei, "Current Modeling Methods Used in QSAR / QSPR," 2012.
10. J. J. Kraker, D. M. Hawkins, S. C. Basak, R. Natarajan, and D. Mills, "Quantitative Structure – Activity Relationship (QSAR) modeling of juvenile hormone activity: Comparison of validation procedures," vol. 87, pp. 33–42, 2007, doi: 10.1016/j.chemolab.2006.03.001.
11. M. Hiv-, "crossm JAK-STAT Signaling Pathways and," vol. 91, no. 9, pp. 1–15, 2017.
12. J. Huang and X. Fan, "Why QSAR Fails : An Empirical Evaluation Using Conventional Computational Approach," pp. 600–608, 2011.
13. Y. Yang, S. J. Adelstein, and A. I. Kassiss, "Target discovery from data mining approaches," *Drug Discov. Today*, vol. 17, no. SUPPL., pp. S16–S23, 2012, doi: 10.1016/j.drudis.2011.12.006.
14. H. J. GEORGE, "Enhancements to the data mining process," Stanford University, 1997.
15. D. J. Abraham, *MEDICINAL CHEMISTRY AND DRUG DISCOVERY*, 6th ed. 2000.
16. R. O. P. Erkins, H. O. N. G. F. Ang, W. E. T. Ong, and W. I. J. W. Elsh, "Annual Review QUANTITATIVE STRUCTURE – ACTIVITY RELATIONSHIP METHODS : PERSPECTIVES ON DRUG DISCOVERY AND TOXICOLOGY," vol. 22, no. 8, pp. 1666–1679, 2003.
17. S. García, Intelligent Systems Reference Library 72 Data Preprocessing in Data Mining.
18. K. Sattelmeyer, "Computer Software Re V iews Book Re V iews," vol. 123, no. 29, pp. 7196–7198, 2001.
19. K. Roy, S. Kar, and R. N. Das, *A primer on QSAR/QSPR modeling: fundamental concepts (SpringerBriefs in Molecular Science)*. 2015.
20. A. U. Khan, "Descriptors and their selection methods in QSAR analysis : paradigm for drug design," *Drug Discov. Today*, vol. 21, no. 8, pp. 1291–1302, 2016, doi: 10.1016/j.drudis.2016.06.013.
21. M. Ibrahim, N. A. Saleh, W. M. Elshemey, and A. A. Elsayed, "Fullerene Derivative as Anti-HIV Protease Inhibitor : Molecular Modeling and QSAR Approaches," vol. 2, pp. 447–451, 2012.

22. R. Guha and E. Willighagen, "A Survey of Quantitative Descriptions of Molecular Structure," *Curr Top Med Chem*, vol. 12, no. 18, pp. 1946–1956, 2013.
23. M. Shahlaei, "Descriptor Selection Methods in Quantitative Structure – Activity Relationship Studies : A Review Study," 2012.
24. A. Ra, "Intercorrelation Limits in Molecular Descriptor Preselection for QSAR / QSPR," vol. 1800154, pp. 2–7, 2019, doi: 10.1002/minf.201800154.
25. S. Guest and E. Section, "Feature Selection Methods in QSAR Studies," *J. AOAC Int.*, vol. 95, no. 3, pp. 636–651, 2012, doi: 10.5740/jaoacint.SGE.
26. I. Furxhi, F. Murphy, M. Mullins, and A. Arvanitis, "Practices and Trends of Machine Learning Application in Nanotoxicology," pp. 1–32, 2020.
27. P. Gramatica, N. Chirico, E. Papa, S. Cassani, and S. Kovarich, "QSARINS : A New Software for the Development , Analysis , and Validation of QSAR MLR Models," pp. 1–12, 2013, doi: 10.1002/jcc.23361.
28. K. A. Marill, "Advanced Statistics : Linear Regression , Part II : Multiple Linear Regression," pp. 94–102, doi: 10.1197/S1069-6563(03)00601-8.
29. P. M. Khan and K. Roy, "Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR)," *Expert Opin. Drug Discov.*, vol. 13, no. 12, pp. 1075–1089, 2018.
30. S. Wold and M. Sjostrom, "PLS-regression : a basic tool of chemometrics," pp. 109–130, 2001.
31. P. Notions, "Partial least squares regression and projection on latent structure regression (PLS Regression)," 2010, doi: 10.1002/wics.51.
32. D. M. Hawkins, S. C. Basak, and D. Mills, "Assessing Model Fit by Cross-Validation," pp. 579–586, 2003.
33. F. Anctil and D. G. Tape, "An exploration of artificial neural network rainfall-runoff forecasting combined with wavelet decomposition 1," vol. 128, 2004, doi: 10.1139/S03-071.
34. C. Touzet, C. Touzet, L. E. S. Reseaux, D. E. N. Artificiels, and I. A. U. Connex-, "HAL Id : hal-01338010 INTRODUCTION AU Claude TOUZET Juillet 1992," 2016.
35. S. Bhattacharyya, "Neural Networks :," pp. 450–452, doi: 10.4018/978-1-61350-429-1.ch024.
36. D. Zhang et al., "All Spin Artificial Neural Networks Based on Compound Spintronic Synapse and Neuron," vol. 10, no. 4, pp. 828–836, 2016.
37. O. Kharroubi, O. Blanpain, E. Masson, and S. Lallahem, "Application du réseau des neurones artificiels à la prévision des débits horaires : Cas du bassin versant de l ' Eure , France," *Hydrol. Sci. J.*, vol. 61, no. 3, pp. 541–550, 2016, doi: 10.1080/02626667.2014.933225.
38. "OECD principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models," *Transport*, vol. 2, no. February, pp. 1–154, 2007.
39. K. Roy and P. Ambure, Author ' s Accepted Manuscript. Elsevier, 2016.
40. S. C. Basak and D. Mills, "Quantitative structure-activity relationships for cycloguanil analogs as PfdHFR inhibitors using mathematical molecular descriptors," *SAR QSAR Environ. Res.*, vol. 21, no. 3, pp. 215–229, 2010, doi: 10.1080/10629361003770951.
41. F. J. Luque, *Frontiers in computational chemistry for drug discovery*, vol. 23, no. 11. 2018.
42. K. Roy, "Importance of Applicability Domain of QSAR Models," pp. 180–182, doi: 10.4018/978-1-4666-8136-1.ch005.
43. T. I. Netzeva, E. C. Agency, A. Worth, E. Commission, and T. Aldenberg, "Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure – Activity Relationships," no. May 2005, 2017, doi: 10.1177/026119290503300209.
44. K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. 2015.

45. D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti, and O. Nicolotti, "Applicability Domain for QSAR Models," *Int. J. Quant. Struct. Relationships*, vol. 1, no. 1, pp. 45–63, 2016, doi: 10.4018/ijqspr.2016010102.
46. P. Gedeck, C. Kramer, and P. Ertl, *Computational analysis of structure-activity relationships*, vol. 49, no. 10. Elsevier B.V., 2010.
47. C. Liao, Z. Liu, A. Hagler, and Q. Gu, "Chemical Structure Similarity Search for Ligand-Based Virtual Screening: Methods and Computational Resources," no. November, 2015, doi: 10.2174/1389450116666151102095555.
48. R. T. Sataloff, M. M. Johns, and K. M. Kost, *Handbook of chemoinformatics algorithms*. 2010.
49. B. Y.-C. Lo and J. Z. Torres, "Chemical Similarity Networks for Drug Discovery," *Spec. Top. Drug Discov.*, no. 53, 2016.
50. M. S. Concepts and I. Applications, "Molecular Similarity Concepts for Informatics Applications," vol. 1526, 2017, doi: 10.1007/978-1-4939-6613-4.
51. T. Steinbrecher, A. Labahn, P. Chemie, and U. Freiburg, "Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies," pp. 767–785, 2010.
52. P. Willett, "Universities of Leeds, Sheffield and York Similarity-Based Virtual Screening Using 2D Fingerprints," vol. 11, pp. 1046–1053, 2006.
53. M. A. Lill and M. L. Danielson, "Computer-aided drug design platform using PyMOL," no. August 2010, pp. 13–19, 2011, doi: 10.1007/s10822-010-9395-8.
54. T. Kaserer, K. R. Beck, M. Akram, A. Odermatt, D. Schuster, and P. Willett, "Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases," *Molecules*, vol. 20, no. 12, pp. 22799–22832, 2015, doi: 10.3390/molecules201219880.
55. S. Yang, "Pharmacophore modeling and applications in drug discovery : challenges and recent advances," *Drug Discov. Today*, vol. 15, no. 11–12, pp. 444–450, 2010, doi: 10.1016/j.drudis.2010.03.013.
56. D. Schaller, Š. Dora, T. Noonan, D. Machalz, M. Bermudez, and G. Wolber, "Next generation 3D pharmacophore modeling," no. November 2019, pp. 1–20, 2020, doi: 10.1002/wcms.1468.
57. M. N. Drwal, K. Agama, L. P. G. Wakelin, Y. Pommier, and R. Griffith, "Exploring DNA Topoisomerase I Ligand Space in Search of Novel Anticancer Agents," vol. 6, no. 9, pp. 1–12, 2011, doi: 10.1371/journal.pone.0025150.
58. Y. Pommier, E. Leo, H. Zhang, and C. Marchand, "Review DNA Topoisomerases and Their Poisoning by Anticancer and Antibacterial Drugs," *Chem. Biol.*, vol. 17, no. 5, pp. 421–433, 2010, doi: 10.1016/j.chembiol.2010.04.012.
59. S. Pal, V. Kumar, B. Kundu, D. Bhattacharya, and N. Preethy, "Ligand-based Pharmacophore Modeling, Virtual Screening and Molecular Docking Studies for Discovery of Potential Topoisomerase I Inhibitors," *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 291–310, 2019, doi: 10.1016/j.csbj.2019.02.006.
60. Y. Li, L. Han, Z. Liu, and R. Wang, "Comparative Assessment of Scoring Functions on an Updated Benchmark : 2 . Evaluation Methods and General Results Part I . Summary of other published comparative studies of docking / scoring methods."
61. N. Yanamala, K. C. Tirupula, and J. Klein-seetharaman, "Preferential binding of allosteric modulators to active and inactive conformational states of metabotropic glutamate receptors," vol. 14, pp. 1–12, 2008, doi: 10.1186/1471-2105-9-S1-S16.
62. V. Salmaso and S. Moro, "Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process : An Overview," vol. 9, no. August, pp. 1–16, 2018, doi: 10.3389/fphar.2018.00923.

63. D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S. M. Dann, and T. C. Savidge, "Regulation of protein-ligand binding affinity by hydrogen bond pairing," no. March, 2016.
64. https://saylordotorg.github.io/text_the-basics-of-general-organic-and-biological-chemistry/s21-06-enzyme-action.html
65. X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, "Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery," *Curr. Comput. Aided-Drug Des.*, vol. 7, no. 2, pp. 146–157, 2012, doi: 10.2174/157340911795677602.
66. N. Brooijmans, A. Review, "Molecular Recognition and Docking Algorithms," no. January, 2014, doi: 10.1146/annurev.biophys.32.110601.142532.
67. L. G. Ferreira, R. N. Santos, G. Oliva, and A. D. Andricopulo, *Molecular Docking and Structure-Based Drug Design Strategies*. 2015.
68. X. Du *et al.*, "Insights into Protein – Ligand Interactions : Mechanisms , Models , and Methods," no. i, pp. 1–34, doi: 10.3390/ijms17020144.
69. N. L. Allinger, Y. H. Yuh, and J.-H. Lii, "Molecular mechanics. The MM3 force field for hydrocarbons," *J. AMERIGW Chem. Soc.*, vol. 11, no. 23, pp. 8551–8566, 1989.
70. X. Shen, S. Chen, K. Dai, and Z. Chen, *Translational Bioinformatics and Its Application Series editors*. 2017.
71. I. A. Guedes, F. S. S. Pereira, and L. E. Dardenne, "Empirical Scoring Functions for Structure-Based Virtual Screening : Applications , Critical Aspects , and Challenges," vol. 9, no. September, pp. 1–18, 2018, doi: 10.3389/fphar.2018.01089.
72. E. Feliu, P. Aloy, and B. Oliva, "On the analysis of protein – protein interactions via knowledge-based potentials for the prediction of protein – protein docking," vol. 20, no. 88, pp. 529–541, 2011, doi: 10.1002/pro.585.
73. R. D. Taylor, P. J. Jewsbury, and Jonathan W. Essex, "A review of protein-small molecule docking methods.," *J. Comput. Aided. Mol. Des.*, vol. 16, no. 3, pp. 151–166, 2002, doi: <https://doi.org/10.1023/A:1020155510718>.
74. G. Tiwari and D. Mohanty, "An In Silico Analysis of the Binding Modes and Binding Affinities of Small Molecule Modulators of PDZ-Peptide Interactions," vol. 8, no. 8, 2013, doi: 10.1371/journal.pone.0071340.
75. V. Salmaso and S. Moro, "Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process : An Overview," vol. 9, no. August, pp. 1–16, 2018, doi: 10.3389/fphar.2018.00923.
76. A. Roy, P. Seal, J. Sikdar, and S. Banerjee, "Underlying molecular interaction of bovine serum albumin and linezolid : A biophysical outlook," *J. Biomol. Struct. Dyn.*, vol. 1102, no. January, pp. 0–1, 2017, doi: 10.1080/07391102.2017.1278721.
77. A. Ganesan, M. L. Coote, and K. Barakat, "Molecular dynamics-driven drug discovery: leaping forward with confidence," *Drug Discov. Today*, vol. 22, no. 2, pp. 249–269, 2017, doi: 10.1016/j.drudis.2016.11.001.
78. A. D. Mackerell, "Empirical Force Fields for Biological Macromolecules: Overview and Issues," 2004, doi: 10.1002/jcc.20082.
79. M. O. Steinhauser and S. Hiermaier, "A review of computational methods in materials science: Examples from shock-wave and polymer physics," *Int. J. Mol. Sci.*, vol. 10, no. 12, pp. 5135–5216, 2009, doi: 10.3390/ijms10125135.
80. A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpi, "Molecular dynamics simulations: Advances and applications," *Adv. Appl. Bioinforma. Chem.*, vol. 8, no. 1, pp. 37–47, 2015, doi: 10.2147/AABC.S70333.

Chapter IV:

Contributions and results

IV. 1. QSAR investigations and Ligand-based virtual screening on a series of nitrobenzoxadiazole derivatives targeting human glutathione-S-transferases

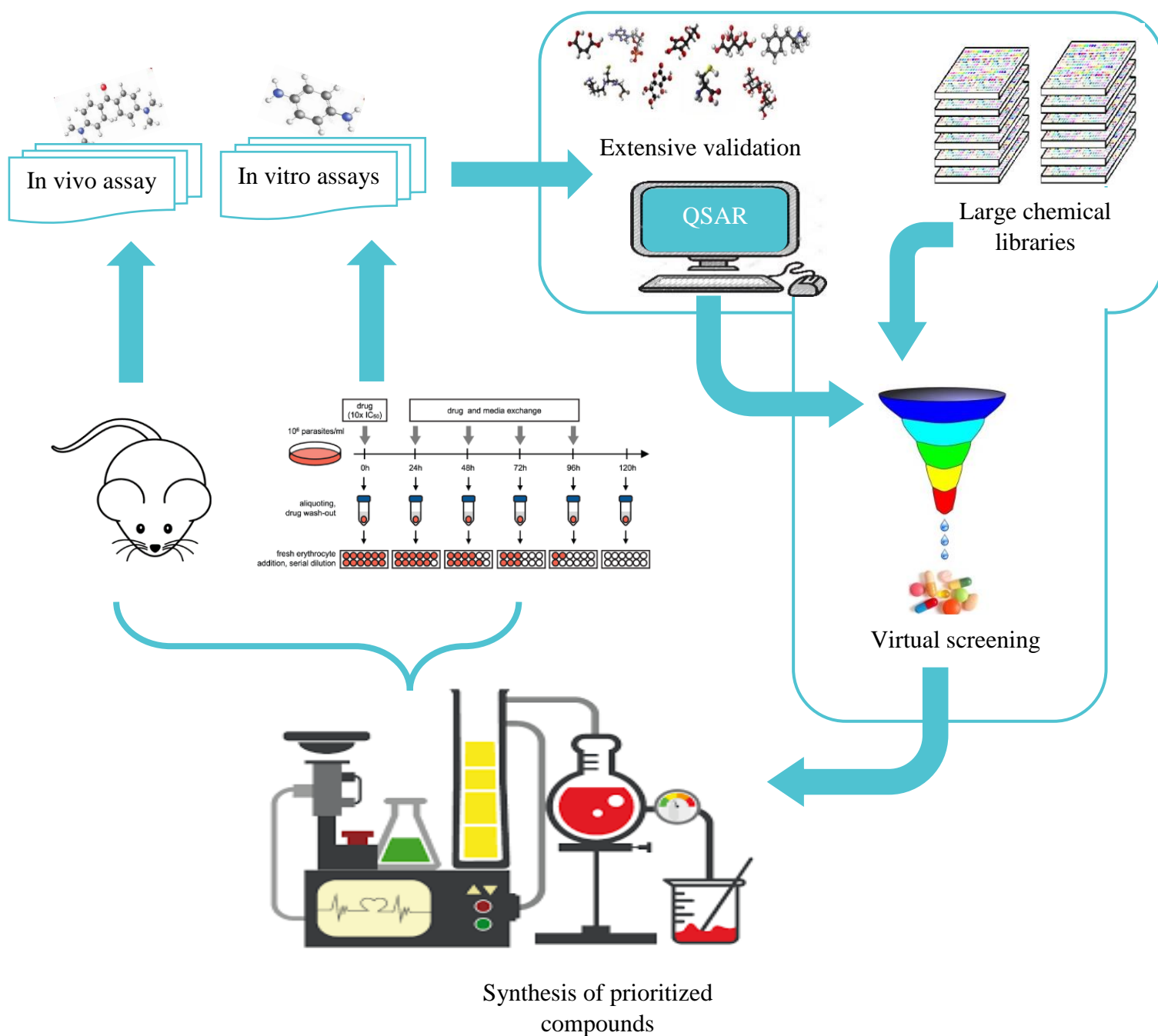


Figure IV. 1. The workflow used in QSAR-based virtual screening study

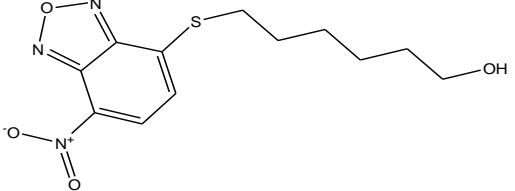
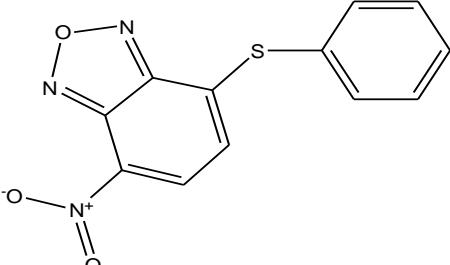
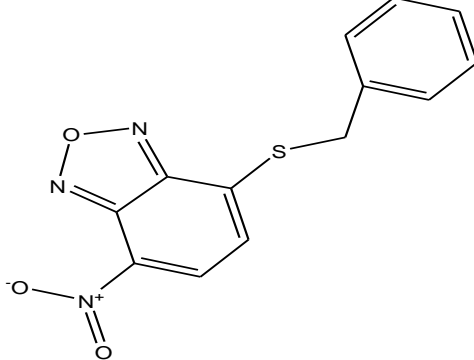
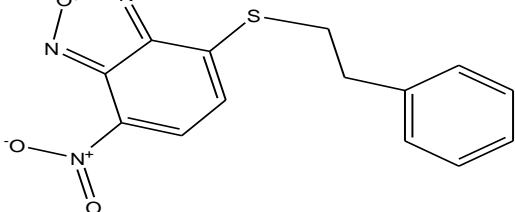
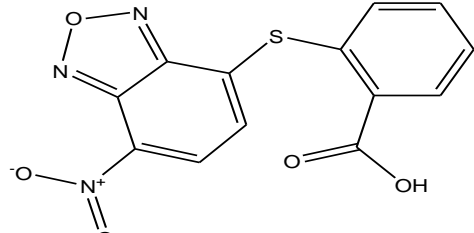
IV. 1. 1. Introduction

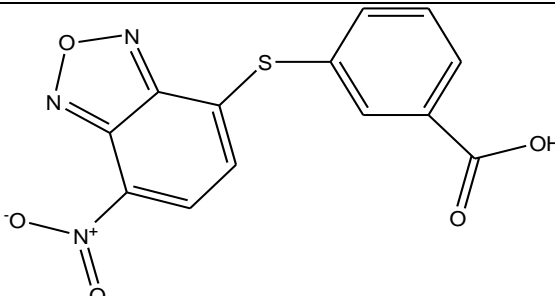
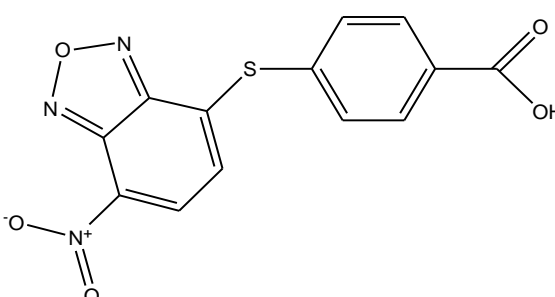
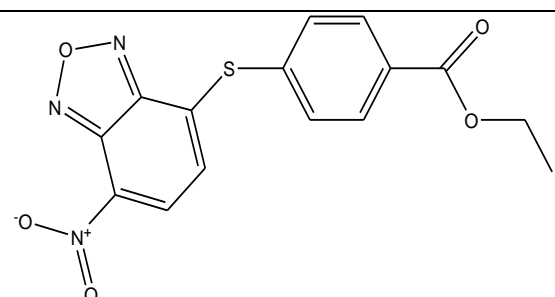
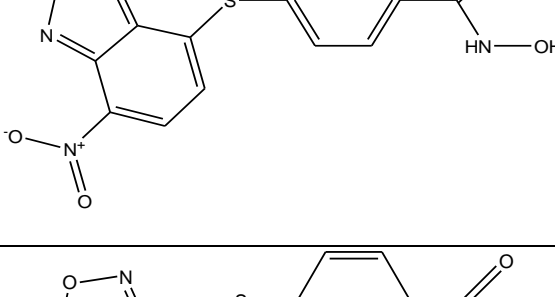
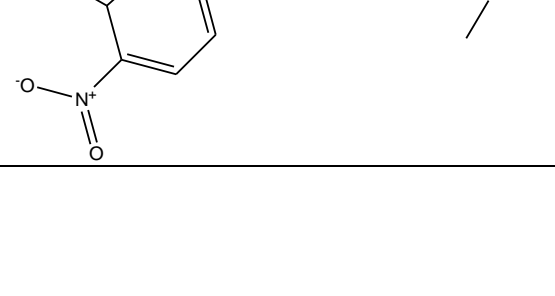
For the rational design of novel GSTP1-1 inhibitors, the quantitative structure-activity relationship (QSAR) [1-5] has considered to be an important method for estimating the biological activities of molecular structure compounds and experimental data [6]. Thanks to QSAR, the biological properties and activities can be easily estimated *in silico* without any experimental effort for the synthesis and evaluation of potentially novel compounds [7, 8]. The secret to success in this type of research is the proper choice of molecular descriptors related to the biological behavior, the chosen statistical models and the consistency and availability of biological data [9]. Common QSAR approaches include partial least squares (PLS) [11], multiple linear regression (MLR) [10], artificial neural network (ANN) [13], genetic algorithms (GA) [12], and support vector machine (SVM) learning method [14].

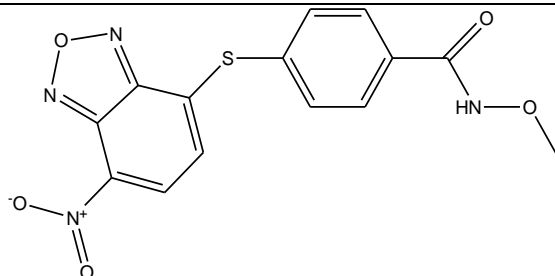
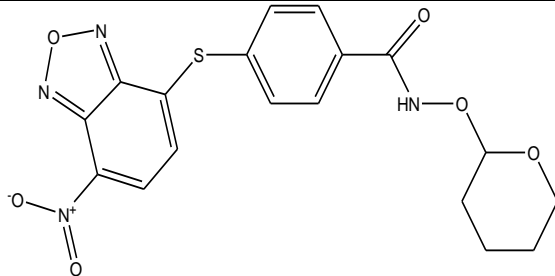
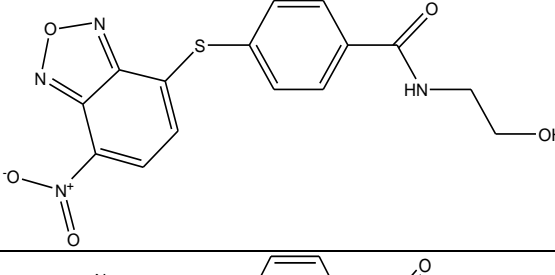
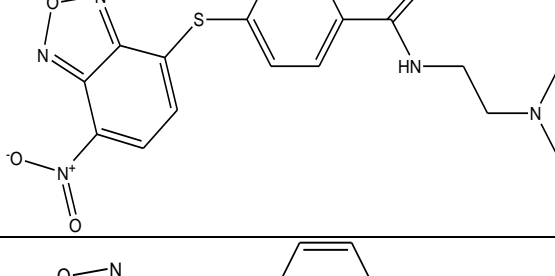
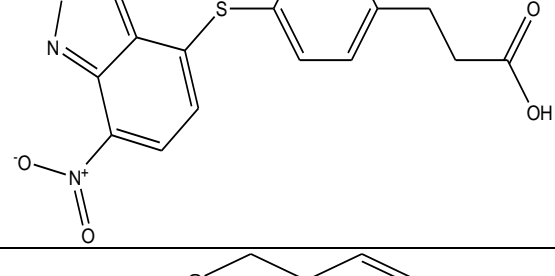
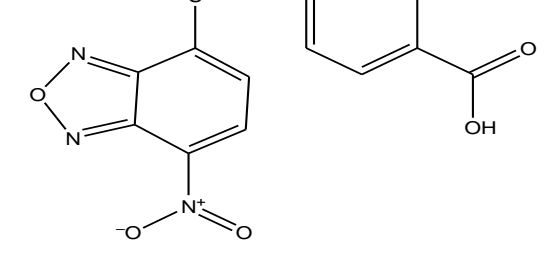
The objective of this study (Figure IV. 1) is to validate an efficient strategy for the accurate prediction of molecular geometries and electronic properties of potentially active compounds and to identify the best molecular descriptors to be used in combination with the linear (MLR) and nonlinear (ANN) QSAR models to identify the best GSTP1 inhibition candidates. Towards this reason, the biological data used in this study concerned cytotoxic agents targeting human glutathione-S-transferases as defined by Caccuri and co-workers. [15, 16]. These authors researched, examined, synthesized and tested a series of 38 nitrobenzoxadiazole derivatives for their *in vitro* GSTP1-1 inhibitory activity. Their chemical compositions are shown in Table IV. 1. They correspond to the substituted thiol group fixed in para-position with respect to the group nitro of nitrobenzoxadiazole. This family of thiol-activated anticancer drugs is promising for cancer treatment. However, most of them are either in pre-clinical creation or clinical trial phases [17]. In-depth studies of these compounds are therefore required.

We began our analysis by optimizing the composition geometry of interest derivatives of nitrobenzoxadiazole. These optimizations are carried out at the theoretical stage of B3LYP/6-311++ G (d, p). Subsequently, the validation of the accuracy and reliability of the QSAR models adopted was carried out using the LOOCV, Y-randomization and external test set validation techniques. The QSAR models obtained were finally used to classify the biological activities of potentially novel active compounds by means of *in silico* screening procedures.

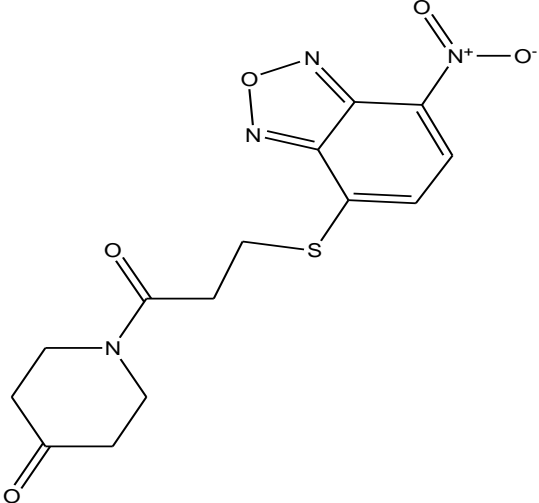
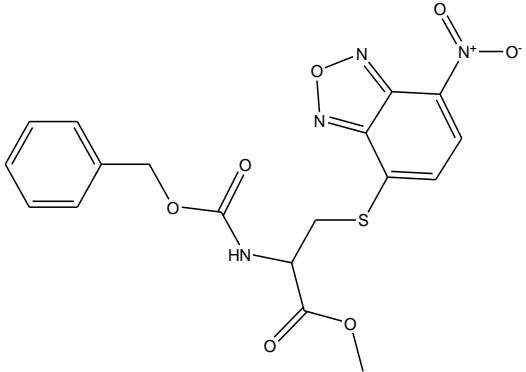
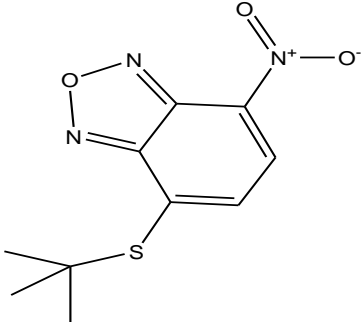
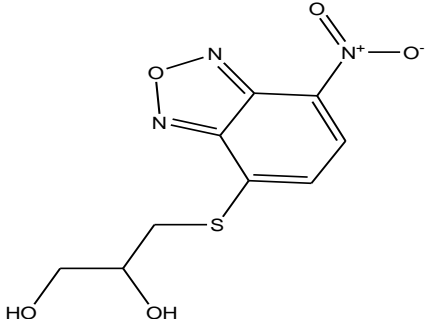
Table IV. 1. Observed and predictive activities (and their differences) of the set of nitrobenzoxadiazole derivatives [15, 16]. * denotes the external test set for GSTP1-1.

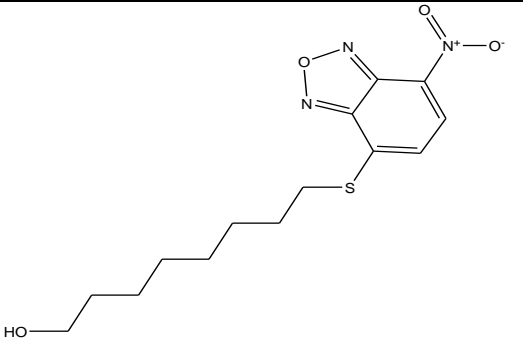
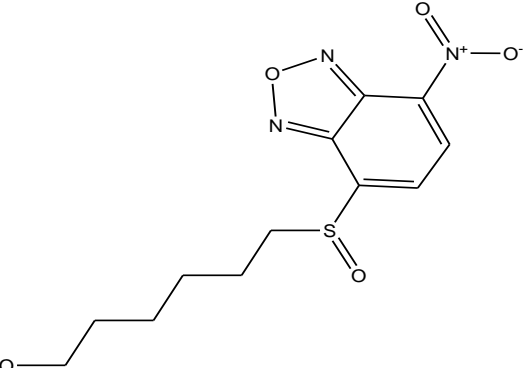
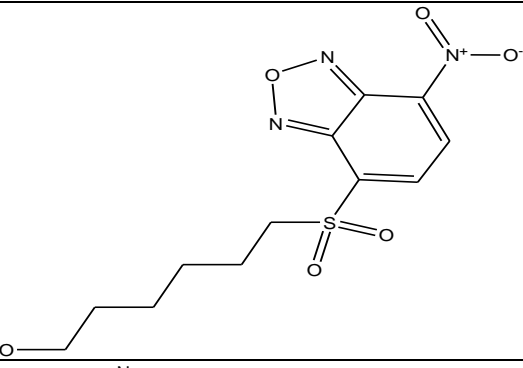
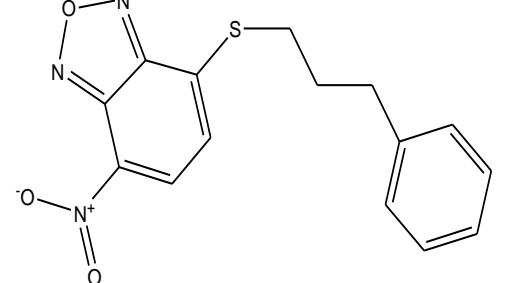
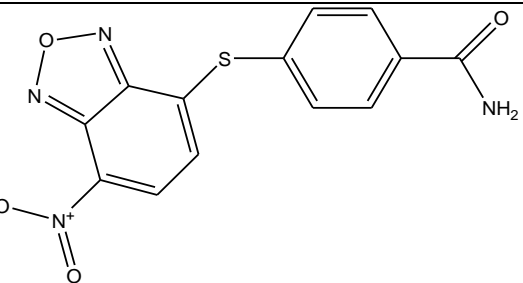
| No. | Compound structure | pIC50 (μM) | Pred. (MLR) | Δ_{MLR} | Pred. (ANN) | Δ_{ANN} |
|-----|---|----------------------------|----------------|-----------------------|----------------|-----------------------|
| 1 |  | 6.097 | 5.880 | -0.217 | 6.178 | 0.081 |
| 2 |  | 6.222 | 6.330 | 0.108 | 6.499 | 0.277 |
| 3 |  | 6.000 | 6.470 | 0.47 | 6.565 | 0.565 |
| 4 |  | 6.699 | 6.630 | -0.069 | 6.616 | -0.083 |
| 5 |  | 5.509 | 5.640 | 0.131 | 5.799 | 0.29 |

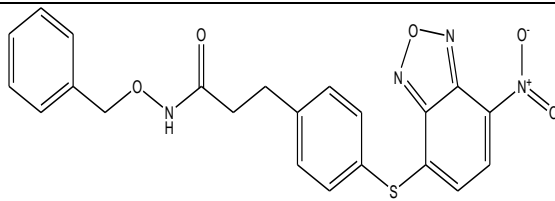
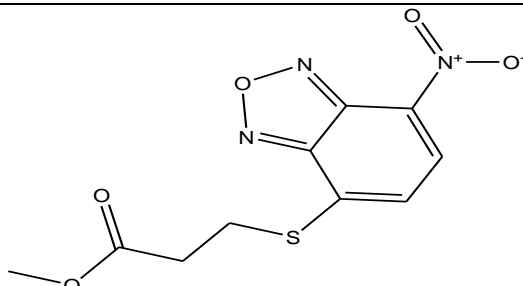
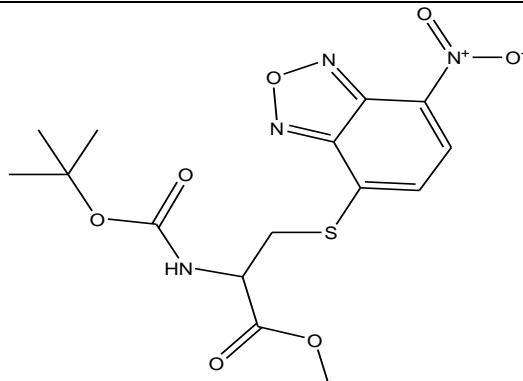
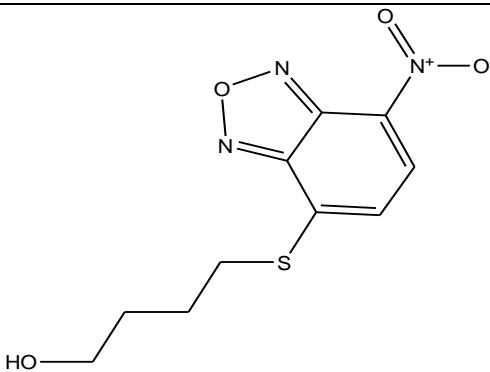
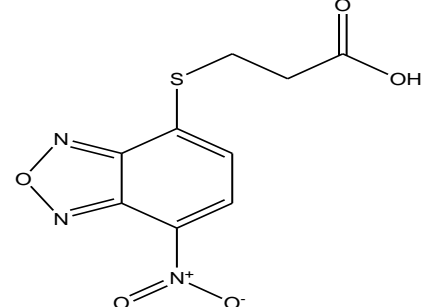
| | | | | | | |
|----|---|-------|-------|--------|-------|--------|
| 6 |  | 5.854 | 5.650 | -0.204 | 5.816 | -0.038 |
| 7 |  | 5.770 | 5.630 | -0.14 | 5.773 | 0.003 |
| 8 |  | 5.745 | 6.130 | 0.385 | 5.935 | 0.19 |
| 9 |  | 5.921 | 6.030 | 0.109 | 5.880 | -0.041 |
| 10 |  | 6.222 | 6.130 | -0.092 | 5.935 | -0.287 |

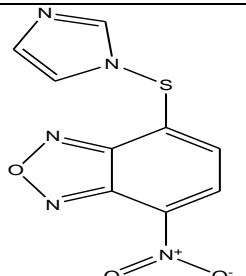
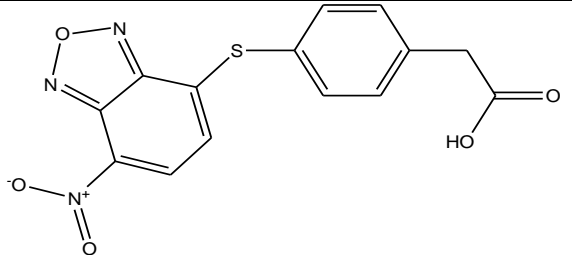
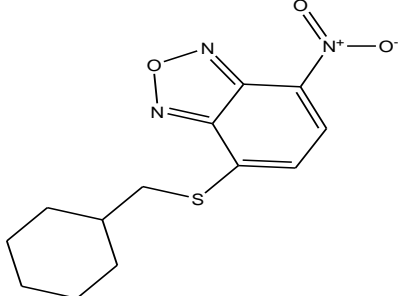
| | | | | | | |
|----|---|-------|-------|--------|-------|--------|
| 11 |  | 6.301 | 5.940 | -0.361 | 6.305 | 0.004 |
| 12 |  | 5.921 | 5.940 | 0.019 | 5.789 | -0.132 |
| 13 |  | 5.678 | 5.870 | 0.192 | 5.717 | 0.039 |
| 14 |  | 6.000 | 5.960 | -0.04 | 6.049 | 0.049 |
| 15 |  | 6.155 | 5.930 | -0.225 | 6.207 | 0.052 |
| 16 |  | 5.796 | 5.780 | -0.016 | 5.985 | 0.189 |

| | | | | | | |
|----|--|-------|-------|--------|-------|--------|
| 17 | | 6.155 | 6.300 | 0.145 | 6.182 | 0.027 |
| 18 | | 6.699 | 6.780 | 0.081 | 6.727 | 0.028 |
| 19 | | 5.745 | 5.500 | -0.245 | 5.733 | -0.012 |
| 20 | | 5.161 | 5.300 | 0.139 | 5.202 | 0.041 |
| 21 | | 5.319 | 5.230 | -0.089 | 5.309 | -0.01 |

| | | | | | | |
|----|---|-------|-------|--------|-------|--------|
| 22 |  | 5.187 | 5.450 | 0.263 | 5.328 | 0.141 |
| 23 |  | 6.398 | 6.020 | -0.378 | 6.184 | -0.214 |
| 24 |  | 6.523 | 6.060 | -0.463 | 6.290 | -0.233 |
| 25 |  | 5.131 | 5.360 | 0.229 | 5.399 | 0.268 |

| | | | | | | |
|-----|---|-------|-------|--------|-------|--------|
| 26 |  | 6.301 | 6.150 | -0.151 | 6.354 | 0.053 |
| 27 |  | 4.852 | 5.340 | 0.488 | 5.388 | 0.536 |
| 28 |  | 5.046 | 4.960 | -0.086 | 5.074 | 0.028 |
| 29* |  | 6.301 | 6.770 | 0.469 | 6.647 | 0.346 |
| 30* |  | 6.097 | 5.710 | -0.387 | 5.933 | -0.164 |

| | | | | | | |
|-----|---|-------|-------|--------|-------|--------|
| 31* |  | 7.000 | 7.050 | 0.05 | 6.756 | -0.244 |
| 32* |  | 5.770 | 5.420 | -0.35 | 5.367 | -0.403 |
| 33* |  | 6.222 | 5.700 | -0.522 | 5.834 | -0.388 |
| 34* |  | 5.854 | 5.620 | -0.234 | 5.892 | 0.038 |
| 35* |  | 5.244 | 5.070 | -0.174 | 5.277 | 0.033 |

| | | | | | | |
|-----|--|-------|-------|--------|-------|-------|
| 36* |  | 5.201 | 5.170 | -0.031 | 5.314 | 0.113 |
| 37* |  | 6.398 | 5.860 | -0.538 | 6.148 | -0.25 |
| 38* |  | 6.523 | 6.470 | -0.053 | 6.566 | 0.043 |

IV. 1. 2. Methodologies:

IV. 1. 2. 1. Equilibrium structure optimizations:

Accurate molecular geometry predictions are subject to the choice of the electronic structure method and the atomic basis set used for the classification of the atoms. We began our investigations by selecting the necessary methods to be used to evaluate the equilibrium structures of the derivatives of nitrobenzoxadiazole under analysis. Our approach is to conduct benchmark computations on the subunit of the series. We have shown that this technique ensures a reasonable balance between accuracy and computational resources to describe the properties of the nitrobenzoxadiazole derivatives considered. In Refs, [18, 19] benchmarks for the 2,1,3-benzoxadiazole subunit, using semi-empirical AM1 and PM3 methods, Hartree-Fock and MollerPlesset (MP2) ab initio techniques, and BLYP and B3LYP DFTs, in combination with different base sets, are applied. When compared to experimental structural parameters, it turns out that B3LYP/6-311G(d, p) is accurate enough to predict a 2,1,3-benzoxadiazole balance structure. From this, we conclude that B3LYP, in conjunction with the 6-311++G (d, p) base set, is suitable

for the analysis of 2,1,3-benzoxadiazole derivatives. Thus, this stage of theory will be used for the estimation of the balance geometry of interest derivatives of nitrobenzoxadiazole. Geometries of all molecules were initially pre-optimized by molecular mechanics (MM) approach. The minimized structures were further optimized using the semi-empirical Austin Model 1 (AM1) method as implemented in HyperChem (version 7.0) [20], which was used for these calculations. In the optimization of geometry, an RMS gradient of 0.01 kcal Å⁻¹ mol⁻¹ was adopted as a convergence threshold. These structures were further refined, without restrictions, by using the DFT B3LYP/6-311++G(d, p) method, which was also used to measure some theoretical descriptors, as implemented in Gaussian 09 [21]. The MarvinSketch [22] software was used to measure other molecular descriptors, such as topological ones.

IV. 1. 2. 2. Molecular descriptors generation:

In order to obtain reliable QSAR models, various molecular descriptors, encoding different molecular size, hydrophilicity [23], electronic and topological properties, were computed from three separate programs: MarvinSketch[22], HyperChem[20], and Gaussian[21]. The various molecular descriptors selected for the 2D QSAR are shown in Table IV. 2 along with their symbols and descriptions. The 2D and 3D MESP maps of 2,1,3-benzoxadiazole are given in Ref. [19]. Areas which characterized with low electrostatic potential are found around nitrogen and oxygen atoms. They relate to the excess of electronic charges favoring electrophilic attacks on these sites. Whereas the areas characterized with high potential are situated around the four hydrogen atoms of the six-member aromatic ring (Figure IV.2). They are characteristic of an electron deficiency where nucleophilic attacks may occur.

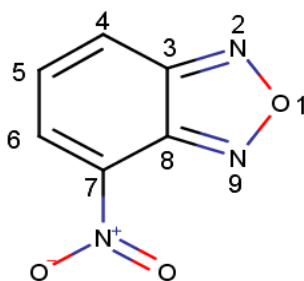


Figure IV.2. Structure of 4-Nitro-2, 1, 3-Benzoxadiazole.

Table IV. 2. Symbols and description of all calculated molecular descriptors

| Notation | Description | Notation | Description |
|------------------------|-----------------------------|------------------------|---|
| V | Volume | n_{rot} | Number of Rotatable Bonds |
| MW | Molecular weight | Wiener | Wiener indexes |
| S | Surface | nC=O | Number of carbonyl fragments |
| Log P | Partition Coefficient | Q_{max} | The highest positive partial charge on the molecule |
| HE | Hydration energy | Q_{min} | The lowest negative charge on the molecule |
| SAS | Solvent accessible surface | HOMO | The energy of the highest occupied Molecular Orbital |
| Pol | Polarizability | LUMO | The energy of the Lowest Unoccupied Molecular Orbital |
| PSA | Polar Surface Area | qS | Charge of the Sulfur atom |
| HBD | Hydrogen Bond Donor | μ | Dipole moment |
| HBA | Hydrogen Bond Acceptor | E-Sol | Solvation Energy |
| a_{hyd} | Number of hydrophobic atoms | HF | Heat of formation |

IV. 1. 2. 3. Model development:

The first pre-selection of the most appropriate set of descriptors for the anti-cancer model operation was performed using the hierarchical cluster analysis (HCA) [24]. Within this process, a dendrogram is created by an iterative coupling procedure in which clusters are developed on the basis of similarity and grouping criteria [25]. The Pearson coefficient R was used to test the association between molecular descriptors and biological responses. These coefficients were used to evaluate the representative descriptors of each cluster, where only one with the highest correlation with the negative Log of the half-maximum inhibitory concentration (pIC₅₀) was chosen.

Table IV. 3. Correlation matrix for the four selected descriptors with pIC₅₀. See Table IV. 2 for the definition of these descriptors.

| | HBA | HBD | Q _{max} | a _{hyd} | VIF |
|-------------------|--------|--------|------------------|------------------|-------|
| HBA | 1 | | | | 1.677 |
| HBD | 0.586 | 1 | | | 1.973 |
| Q _{max} | -0.441 | -0.580 | 1 | | 1.558 |
| a _{hyd} | 0.115 | -0.148 | 0.133 | 1 | 1.103 |
| pIC ₅₀ | -0.509 | -0.313 | 0.484 | 0.579 | |

We used a forward selection method to produce a subset of specific low-intercorrelation descriptors [26]. Then, as implemented in IBM SPSS Statistics 21[28], we used the MLR [27] method to construct linear QSAR models. The objective of this part was to determine the optimum collection of descriptors which generate the most significant QSAR models linking and interpreting the chemical structure of small molecules with their functional activity [25]. The correlation matrix

for defined descriptors could be seen in Table IV. 3. The quality of the MLR approach was compared to that of the ANN approach, another accurate and predictive QSAR model that was well suited for the treatment of non-linear relations between descriptors and activity [29,30], especially for highly non-linear cases [31]. In particular, back propagation (BP) is the ANN algorithm used in this analysis [32]. All ANN analyses were carried out with MATLAB [33]. For the validation of our model, we used cross-validation and validation leave-one-out (LOO) via an external test set and Y-randomization procedures Table IV. 4 [34–36]. Information of the validation parameters computed for the model are shown in Table IV. 5.

IV. 1. 2. 4. Virtual screening:

Thanks to developments in computational technology over the last decade, virtual screening has now become a useful tool for drug development, enabling rapid and accurate detection of large numbers of possible hit structures. Thus, by focusing only on a small number of target substances, the high cost of experimental research is substantially reduced [37, 38]. Molecular similarity is a key concept in drug modeling and medicinal chemistry [39]. This was done by implementing the extended connectivity fingerprints (ECFPs) scheme in this research. The molecules are represented in this method as binary vectors, where a molecular fragment is represented by bit. "1" and "0" indicate the presence or absence of a given fragment, respectively. Therefore it is possible to infer the similarity between two molecules by comparing the number of common bits between their structures via the Dice coefficient [40, 41].

IV. 1. 3. Results and discussion:

IV. 1. 3. 1. Equilibrium structure of the nitrobenzoxadiazole derivatives:

At the B3LYP/6-311++G(d,p) stage of the theory, the balance structures of the set of nitrobenzoxadiazole derivatives (Table IV. 1) were obtained. The comparison between these structures shows that the part of benzoxadiazole closely resembles the isolated part of 2,1,3-benzoxadiazole [18,19].

For instance, this specific skeleton's distances and angles are O1-N2~1.355 Å; N9-C8~1.315 Å; C8-C7~1.439 Å; C7-C6~1.38 Å; C5-C6~1.420 Å; C5-C4~1.37 Å; C4-C3~1.429 Å; C3-N2~1.318 Å; O1-N9~1.37 Å and O1-N9-C8~104.6°; N9-C3-C8~108.4°; N9-C8-C7~133.1°. In the sequence, the nitro group attached to this subunit also has the same structural characteristics.

The heterocyclic nitrobenzoxadiazole is thus just marginally affected by the replacement of the sulfur atom.

IV. 1. 3. 2. Quantitative structure activity relationships (QSAR) study:

The QSAR study was done by evaluating the values of $pIC_{50} = -\text{Log}(IC_{50})$ for the 38 selected compounds Table IV. 2, established for their capacity to inhibit human S-transferase glutathione[15, 16]. As shown in Figure IV. 3, the application of hierarchical cluster analysis to the set of chemical descriptors has led to seven main clusters.

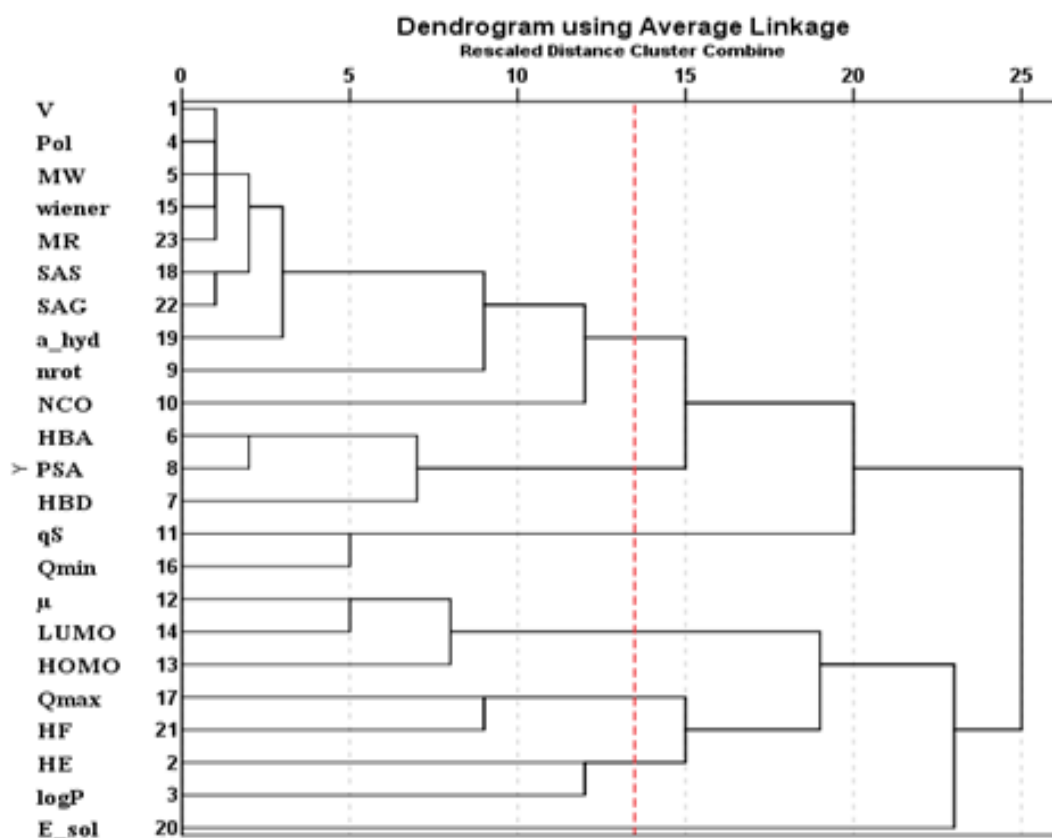


Figure IV.3. Hierarchical cluster analysis of descriptors (dendrogram). See Table IV. 2 for the definition of these descriptors. The vertical red dashed line corresponds to the clipping limit that takes into account the minimum number of descriptor groups without losing any information necessary for the model.

When we used HCA, descriptors were grouped according to their pair descriptors, which form a specific cluster and have a high correlation between each other. We have therefore chosen one descriptor as a representation of each cluster to prevent duplication. The representative descriptors were chosen so that to reduce their coefficient of correlation with descriptors representing other groups. Finally, these representative descriptors have been selected: a hyd, HBD,

HBA, Q_{min} , HOMO, Q_{max} , Log P and E-sol. This is compatible with the docking study [15], which proposed choosing at least the descriptors of HBD, HBA and a hyd. The data set was randomly split into two subsets after choosing the independent descriptors and the dependent variable (pIC_{50}). The training and test data sets consist, respectively, of 28 and 10 molecular compounds. Multiple linear regression enables the structural descriptors to be related to the activity of each of the 28 compounds in order to measure quantitatively the impact of their substituents. The following MLR model was developed by applying the FS-SWR (Forward Selection Step Wise Regression):

$$pIC_{50} = 6.627 - (0.389 \times HBA) + (0.258 \times HBD) + (1.393 \times Q_{max}) + (0.140 \times a_{hyd})$$

Eq IV.1

Where Q_{max} (the highest positive partial charge on the molecule), a_{hyd} (number of hydrophobic atoms) have a positive impact, while HBD (the number of hydrogen bond donor), and HBA (the number of hydrogen bond acceptor) has a negative impact on the activity.

The model was developed to estimate activity values for both training and test data sets. Table IV. 1 reports the observed and theoretical activities of the pIC_{50} , as well as their variations. The plot of the calculated versus observed activity (Figure IV.4) shows a linear relationship, indicating a satisfactory internal predictability of the produced model, regardless of the method used (MLR or ANN). In addition, the plot of the measured residuals against the observed activity values in Figure IV.4 shows that the residuals are uniformly distributed along the zero axis, thereby confirming the absence of systematic errors in the model.

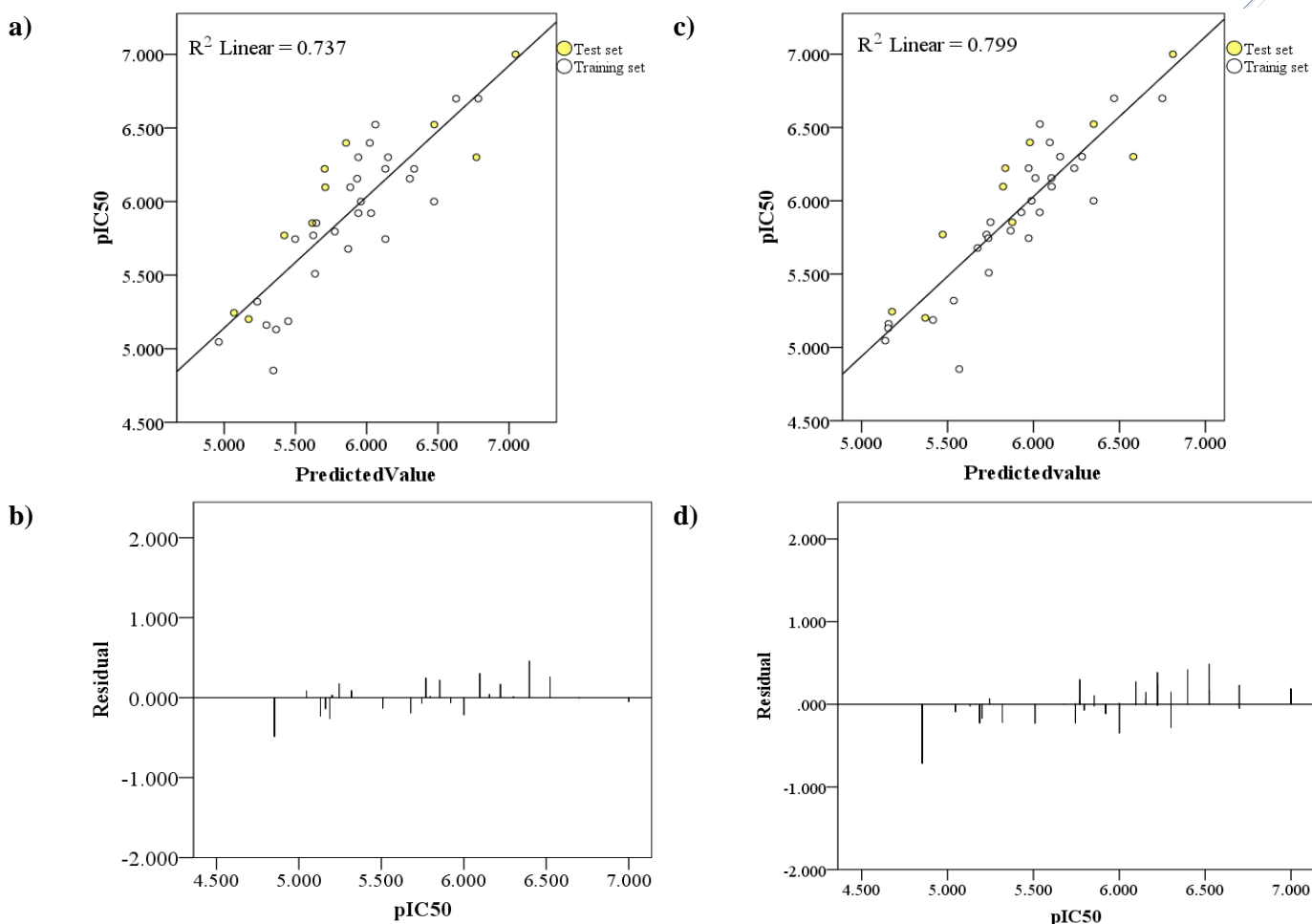


Figure IV. 4. Experimental versus calculated pIC50 values (MLR in a) and ANN in c)), and residuals (MLR in b) and ANN in d)).

Multi-collinearity was observed by the measurement of the inflation variance factors (VIF) for the selected descriptors [42]. The corresponding VIF values and correlation matrix for each descriptor are shown in Table IV. 3. From this table, it can be shown that the chosen descriptors are practically independent, because their R coefficients are less than 0.6. Also all variables have a VIF value of less than 5 and demonstrate that the model does not contain any multi-collinearity and has a simple statistical significance.

Table IV.4. Random MLR Model Parameters.

| | |
|---------------------------------|--------|
| Average R | 0.374 |
| Average R^2 | 0.156 |
| Average Q^2 | -0.268 |
| $^cR_p^2$ | 0.685 |

Further validation of the model was done by the implementation of the Y-randomization test. Several random shuffles of the Y vector were performed and small mean values of 0.156 for R^2 and -0.268 for Q^2 were obtained after 1,000 randomized trials, thus showing that the successful results (Table IV.4) of our original model were not due to a chance association or structural dependence of the training set. As a second step, the presence of a non-linear relationship seen between pIC₅₀ and the four descriptors chosen was investigated. For this reason, a BP artificial neural network was built using the identified MLR descriptors as inputs. Parameter 2n+1 was used to calculate the number of hidden layers, where n represents the number of input layers that play a key role in deciding the best artificial neural network architecture [43]. After optimization, the architecture of the selected ANN model was 4-3-1, i.e. 4 descriptors in the first layer three neurons in the hidden layer, and one neurons in the output layer for the pIC₅₀ results. The three-layer ANN was trained using the Levenberg–Marquardt training algorithm.

Table IV.5. Statistical results of MLR and ANN models.

| | Parameter | MLR | ANN | | Parameter | MLR | ANN |
|---------------------|-------------|--------|--------|-----------------|--------------------|-------|-------|
| Training set | R^2 | 0.758 | 0.812 | Test set | R^2_{pred} | 0.795 | 0.828 |
| | R^2_{adj} | 0.716 | 0.779 | | $RMSE_{ext}$ | 0.338 | 0.258 |
| | SE | 0.265 | 0.244 | | r_m^2 | 0.530 | 0.788 |
| | F | 18.019 | 24.835 | | $r^2 - r'^2/r^2$ | 0.007 | 0.020 |
| | RMSE | 0.240 | 0.221 | | K' | 0.971 | 0.975 |
| | Q^2_{Loo} | 0.654 | 0.689 | | $ r_0^2 - r_0'^2 $ | 0.105 | 0.015 |

The ANN model, built with the same descriptors as the MLR model, given the evaluation metrics shown in Table IV. 5. This table below lists the evaluation metrics of both models, including their correlation coefficient (R^2), variance ratio (F), standard error (SE), root-mean-square error (RMSE), adjusted R^2 (R^2_{adj}), leave-one cross-validated Q^2_{LOO} , and r^2_m for external validation. A detailed overview of r^2_m is available in the literature [44]. A comparative analysis between the values in Table IV.5 and those extracted from the MLR method confirms the enhanced performance of ANN over MLR, suggesting the presence of a non-linear relationship between the four selected descriptors and the pIC₅₀ of the studied compounds. The higher values of R^2 and R^2_{adj} and the smaller root-mean-square error (RMSE) suggest that the proposed model is predictive and accurate. F-test values with p just under 0.005 (see Table IV.5) indicate that the model is statistically important. The larger Q^2_{LOO} and R^2_{pred} and the smaller $RMSE_{ext}$ values also show the strong predictive capabilities of the ANN model and demonstrate its robustness. Values of r^2 greater than

0.5 and values of $r^2 - r'^2/r^2$ and $|r_0^2 - r'_0{}^2|$ smaller than 0.1 and 0.3, respectively indicate the strong predictive efficiency of the model.

IV. 1. 3. 3. Applicability domain of the model:

The Applicability Domain (AD) is an area within the chemical space involving physicochemical, electronic or biological information on which the model training set is built. The molecular AD have a fundamental role in estimating the uncertainty in the similarity test between the substance and those used to create the model [45]. The common definition of the AD is based on the Eq IV.2 following to leverage values.

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, 2, \dots, n) \quad \text{Eq IV.2}$$

For each structure, where I is the descriptor row vector of the query compound, and X is the matrix of k model descriptor values for n training chemical structure [46]. Substances with $h > h^*$ (h^* being a threshold value equal to $3p/n$, where p is the number of model descriptors plus one and n is the number of compounds included in the training set) may be considered to be chemically different from the training set compounds and thus outside the AD [47].

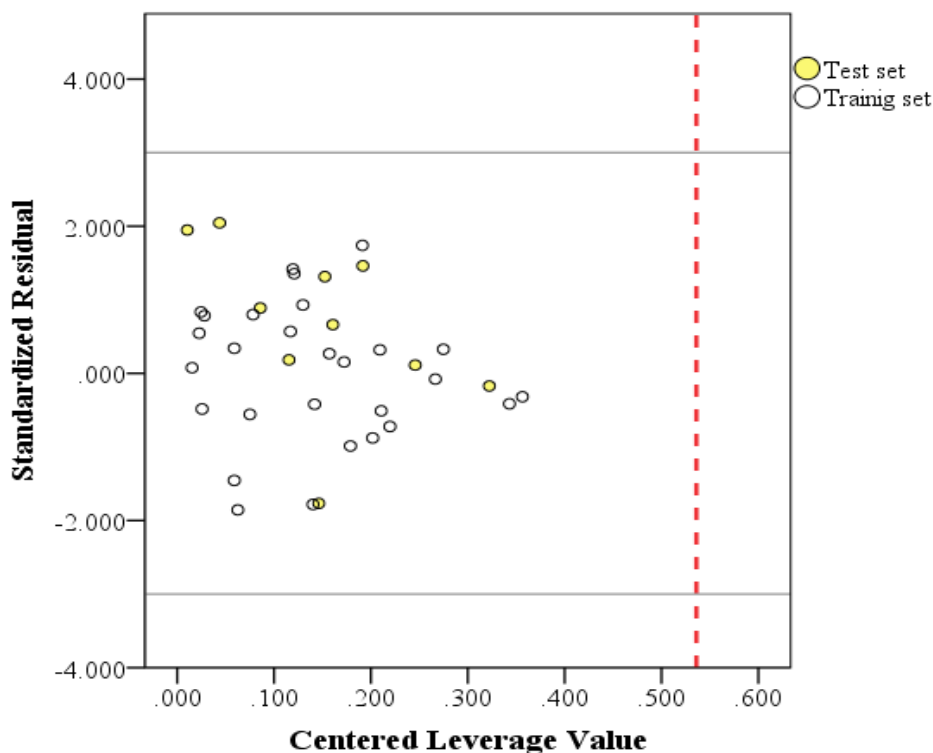


Figure IV.5. Applicability domain plot for the ANN model. Horizontal lines represent $\pm 3\sigma$ and the vertical dashed line represents the warning leverage ($h^* = 0.536$).

To view the AD of the QSAR model, William's plot is mapped (Figure IV.5). In this map, the AD is defined within a squared area within the standard deviation $\pm x$ (in this study $x = 3$; "three

sigma rule" [44]). Molecules with uniform residues three times larger than the standard deviation of the model are marked outliers. Careful analysis of Figure IV. 5 shows that all chemicals compounds in the data set fall within the AD of the proposed ANN (warning leverage) model. Neither of the compounds have leverage values greater than the h^* alert value and none of them have standardized residues greater than the threshold. As a result, the model shows the best statistical parameters and strong predictive properties and can be used with a high degree of confidence in this AD.

IV. 1. 3. 4. Importance of descriptors within different QSAR models:

A randomization technique was used to evaluate the relative value of each descriptor used to create the MLR and ANN models. This also helps the best molecular descriptors to be identified. After the models were constructed, the first column corresponding to the first descriptor used in the model was deleted, leaving the remaining descriptor matrix and the Y-column identical. The mean absolute deviations (Δm_i) between the experimental and the calculated activities for all compounds were defined for the four descriptors. The contribution of each descriptor ($C_i\%$) is given by the following Eq IV. 3 [49].

$$C_i\% = \frac{\Delta m_i}{\sum_{j=1}^4 \Delta m_j} \cdot 100 \quad \text{Eq IV. 3}$$

Where the sum runs over the four descriptors (HBA, HBD, Q_{\max} , a_{hyd}).

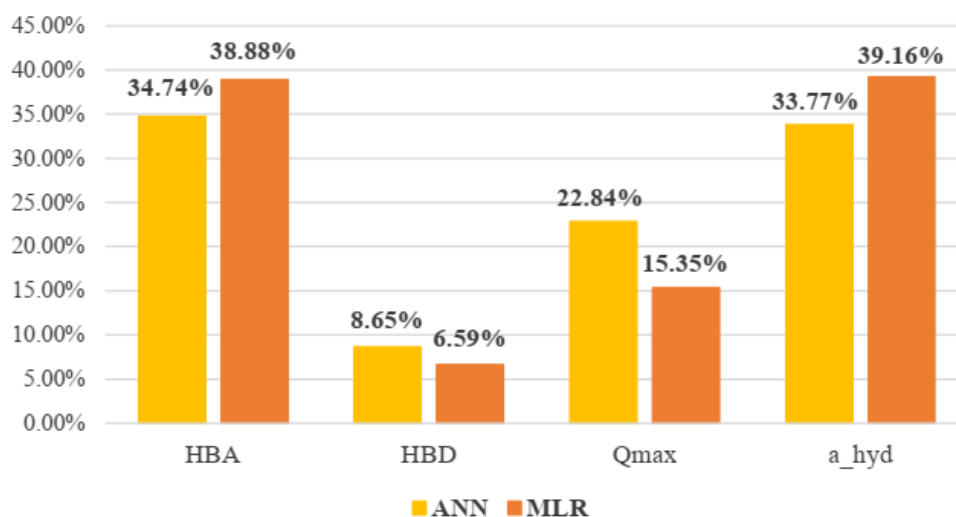


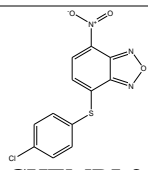
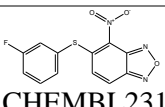
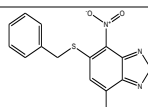
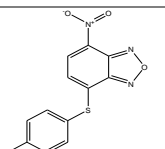
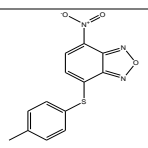
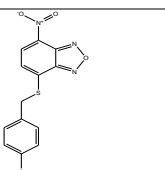
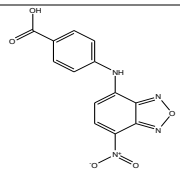
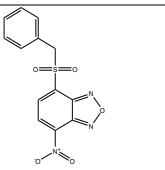
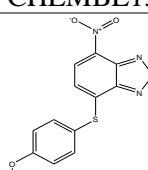
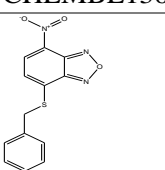
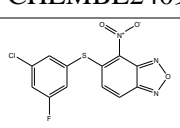
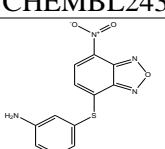
Figure IV.6. Comparison of descriptors contribution in the ANN and MLR models.

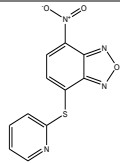
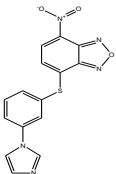
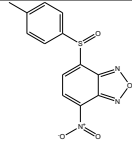
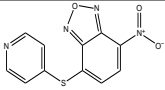
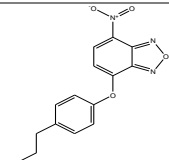
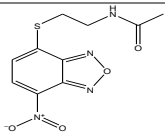
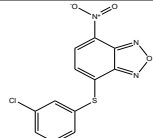
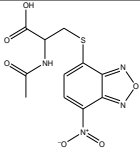
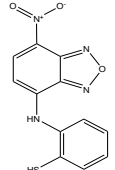
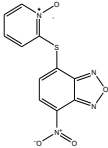
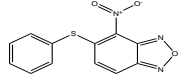
The increase in the use of scrambled descriptor values is a measure of the significance of the descriptor in the model, where larger increments refer to higher importance. This method has

been extended to all four descriptors and the findings are shown in Figure IV. 6. It follows from this diagram that the hydrogen bond acceptors (HBA) and the number of hydrophobic atoms (a hyd) descriptors are of particular importance in the MLR and ANN models. These results are corroborated by the association between these descriptors and the behavior (see Table IV. 3), suggesting that steric interactions are the prevailing forces in ligand-protein complex creation.

IV. 1. 3. 5. Virtual Screening Application:

Table IV.6. Proposed structural compounds and predicted activities.

| No. | Compound structure & ID | pIC ₅₀ | Leverage (<0.536) | No. | Compound structure & ID | pIC ₅₀ | Leverage (<0.536) |
|-----|--|-------------------|-------------------|-----|---|-------------------|-------------------|
| 39 |  CHEMBL2430538 | 6.351 | 0.116 | 51 |  CHEMBL2311954 | 6.383 | 0.127 |
| 40 |  CHEMBL3416322 | 6.532 | 0.158 | 52 |  CHEMBL2430539 | 6.356 | 0.117 |
| 41 |  CHEMBL2409283 | 6.348 | 0.115 | 53 |  CHEMBL2409281 | 6.044 | 0.064 |
| 42 |  CHEMBL1336889 | 5.185 | 0.193 | 54 |  CHEMBL1360942 | 5.409 | 0.142 |
| 43 |  CHEMBL2409289 | 5.939 | 0.063 | 55 |  CHEMBL2430534 | 6.463 | 0.123 |
| 44 |  CHEMBL2311955 | 6.488 | 0.138 | 56 |  CHEMBL2430542 | 5.893 | 0.154 |

| | | | | | | | |
|-----------|---|-------|-------|-----------|---|-------|-------|
| 45 |  | 5.725 | 0.092 | 57 |  | 5.976 | 0.052 |
| | CHEMBL2426721 | | | | CHEMBL3947563 | | |
| 46 |  | 5.921 | 0.092 | 58 |  | 5.659 | 0.102 |
| | CHEMBL1702248 | | | | | | |
| 47 |  | 6.263 | 0.059 | 59 |  | 5.732 | 0.116 |
| | CHEMBL2430448 | | | | | | |
| 48 |  | 6.145 | 0.131 | 60 |  | 5.582 | 0.098 |
| | CHEMBL3416323 | | | | | | |
| 49 |  | 6.391 | 0.317 | 61 |  | 5.670 | 0.084 |
| | CHEMBL2430452 | | | | | | |
| 50 |  | 6.268 | 0.124 | | | | |
| | CHEMBL2311953 | | | | | | |

In order to determine GSTP1-1 inhibitors, the ChEMBL database [50] was searched for substances with more than 70% similarity to the most active compound in the study data set (Compound 31, Table IV.6). The biological activity of some compounds identified in the literature [59] was also estimated (Compound 58-61). These molecules have a significant degree of similarity, as they have the identical basic skeleton as Compound 31. In a previous docking study [15], the successful inhibitory potency of Compound 31 against GSTP1-1 was attributed to additional hydrogen bond interactions between the carbonyl moiety of Compound 31 and Gln39 residues of GSTP1-1, and to hydrophobic interactions with its amino acids Phe8, Val35, and Gly205, which help to stabilize the complex. We apply the Lipinski, Veber, Ghose, and Golden Triangle rules [19] to get drug-like substances. Then for all compounds, molecular descriptors were produced. The best-performing model, ANN, was used to estimate the inhibition activity of GSTP1-1 of the

compounds considered. The results of the predictions were acknowledged only when the compound was contained inside the AD. Table IV.6 shows that all the compounds contained are $h < h^*$, ($h^* = 0.536$, Figure IV.6), and the structures and activities of these compounds are stated in Table IV. 6. These compounds mainly lead to the para-substituted thio-nitrobenzoxadiazole discovered earlier by Caccuri and co-workers [15, 16]. We have also identified such compounds where thiol is in ortho position with respect to the nitrobenzoxadiazole nitro group and compounds where the nitrobenzoxadiazole six-member aromatic ring is replaced by an amine or etheroxy organic role in para with respect to the nitro group. Most of these compounds have a pIC_{50} closest to that of compound 31.

IV.2. Combined 3D-QSAR based Virtual Screening and Molecular Docking study of cytotoxic agents targeting human glutathione-s transferases

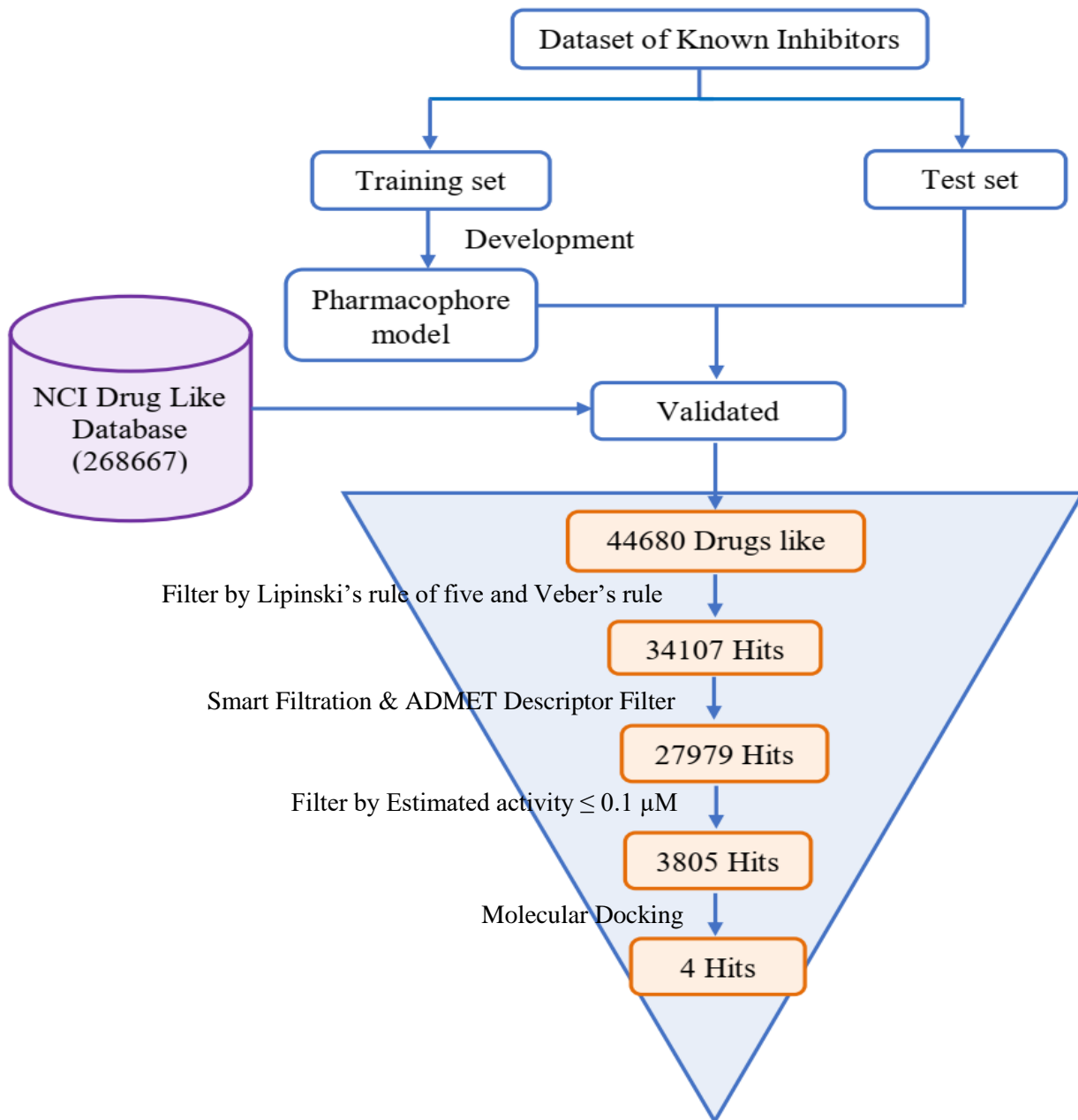


Figure IV. 7. Schematic representation of the virtual screening process implemented in the identification of Top inhibitors.

IV. 2. 1. Introduction:

Virtual screening and modeling based on pharmacophore has reached maturity and has been extensively reviewed in past literature, and is very well recognized in the Medicinal Chemistry Laboratory [52, 53]. Pharmacophores are classified as part of a molecular structure that represents a collection of steric and electronic features responsible for a specific biological or pharmacological interaction with a specific target structure and for inhibiting its biological response [54].

Recently, Docking tools show promising applications for hit discovery, lead optimization and target-based library design [55]. The state of the art molecular docking is a computational technique that aims to predict the non-covalent binding of a macromolecule (Receptor) obtained from data banks or MD simulations, etc. with a small molecule (Ligand) as a lead for further drug development [56]. These lead candidates can be found using a docking algorithm that attempts to classify the optimal binding mode of a small molecule to an active biological target site [57]. Molecular docking can be used to predict affinity, bound conformation and binding energy [55]. The aim of drug discovery is therefore to extract drugs that bind more strongly than the natural substrate to a given protein target [15].

In the present investigation (Figure IV. 7), we have produced 3D QSAR-Pharmacophore models, beginning with a series of cytotoxic agents targeting human Glutathione-S-Transferase. The best pharmacophore model was validated using three different methods and then used in the virtual screening of the NCI chemical library data-base containing more than 200 thousand compounds. In addition, in order to maximize the balance of drug-like properties of selected molecules, several filters have been used, such as Lipinski's rule of five and Veber's rule, Smart Filtration & ADMET Descriptor, and Filter by Estimated activity $\leq 0.1 \mu\text{M}$. Afterwards, we expanded our research by applying Docking-Based Virtual Screening by studying the Ligand-receptor binding affinity to inhibit the functioning of the GST enzyme. Results such as Pharmacophore hypothesis, scoring, docking study, binding mode, and so on, have been determined and discussed in the present section.

IV. 2. 2. Data collection and preparation:

Computational drug design includes ligand-based and structure-based drug design. Pharmacophore model has become an excellent computational tool for searching of novel Hit/Lead compounds in various disease areas. The 3D-QSAR method is considered as one of important

ligand-based pharmacophore-modeling approaches [58]. In this study, 3D-QSAR pharmacophore model was generated while basing on data set of 45 published compounds, which were extracted from European Bioinformatics Institute database ChEMBL and the literatures [15, 16, 59]. These compounds were tested with similar bioassay protocol to allow proper QSAR correlation. The in vitro bioactivity of the inhibitors collected was expressed as the concentration of the test compound required for 50% inhibition of GSTP1-1 enzyme, i.e., IC₅₀. Among the 45 compounds, 16 different inhibitors were selected for training from 0.1 to 14 μM (table IV. 1). The remaining inhibitors have been taken as a test set. The selection of the two data sets of training and testing carried out in accordance with the following rules: 1- same binding mode and structural diversity of molecules; 2- both data sets most cover a wide range of activities; 3- the highest active compounds are included in the training set, because they provide crucial information for generating pharmacophores.

The molecular structures of all compounds have been sketched and constructed using Accelrys Discovery Studio 4.1 [60] from their smile format. The optimization of these structures were done using the steepest descent algorithm with a convergence gradient value of 0.001 kcal / mol and a group of representative orientations were generated by fast conformational analysis methods using polling minimize algorithm [61] and CHARMM force field parameters [62]. A large number of orientations for each compound were generated within an energy threshold of 20.0 kcal / mol above the global energy minimum.

IV. 2. 3. Results and discussion:

IV. 2 .3 .1. Generation of pharmacophore models:

Two types of ligand-based pharmacophore modeling are reported in literature, one of which is a common feature of pharmacophore modeling and the other is 3D-QSAR based pharmacophore modeling, which differs from the first approaches as there is a limitation of the number of training compounds and a requirement for experimental biological activity values predicted in similar bioassay conditions [63].

The HypoGen algorithm in Discovery Studio 4.1 (DS) from Accelrys [69] was used to produce a 3D-QSAR pharmacophore model, quantitatively predicting the biological activity (IC₅₀) of the compounds studied against GSTP1-1. The behaviors of the study compounds ranging from 0.1 to 14 μM in the training set were used to produce the pharmacophore models (Table IV.7). The resulting conformations (255 for each compound) were used to generate pharmacophoric

hypotheses. In order to classify the important pharmacophore features of the training set, the DS feature mapping was carried out, resulting in hydrogen bond donor (HBD), hydrogen bond acceptor feature (HBA), hydrophobic (HYD) and ring aromatic (RA) features. The HypoGen algorithm gives the product of the training set, their pharmacophore features, table IV. 7 display the statistical parameters of the 10 top-score, hypothetical pharmacophore models developed.

Table IV.7. Statistical results of the top 10 pharmacophore hypotheses generated by HypoGen algorithm.




| Hypo. No. | Total cost | Cost difference | RMSD | R training | Max fit | Features | R test |
|-----------|------------|-----------------|------|------------|---------|---------------|--------|
| 1 | 67.40 | 95.31 | 1.05 | 0.96 | 9.83 | HBA, 3HYD, RA | 0.51 |
| 2 | 73.00 | 89.71 | 1.35 | 0.94 | 10.19 | HBA, 3HYD, RA | 0.41 |
| 3 | 79.29 | 83.43 | 1.60 | 0.91 | 11.33 | 2HBA, 3HYD | 0.76 |
| 4 | 79.87 | 82.84 | 1.63 | 0.90 | 9.00 | HBA, 3HYD, RA | 0.64 |
| 5 | 80.58 | 82.13 | 1.66 | 0.90 | 9.27 | HBA, 3HYD, RA | 0.65 |
| 6 | 81.26 | 81.45 | 1.69 | 0.90 | 9.70 | HBA, 3HYD, RA | 0.68 |
| 7 | 81.43 | 81.28 | 1.69 | 0.90 | 9.92 | HBA, 3HYD, RA | 0.67 |
| 8 | 81.68 | 81.04 | 0.89 | 1.70 | 9.74 | HBA, 3HYD, RA | 0.67 |
| 9 | 81.88 | 80.83 | 1.67 | 0.90 | 7.81 | HBA, 3HYD, RA | 0.44 |
| 10 | 81.92 | 80.79 | 1.67 | 0.90 | 7.70 | 2HBA, 3HYD | 0.45 |

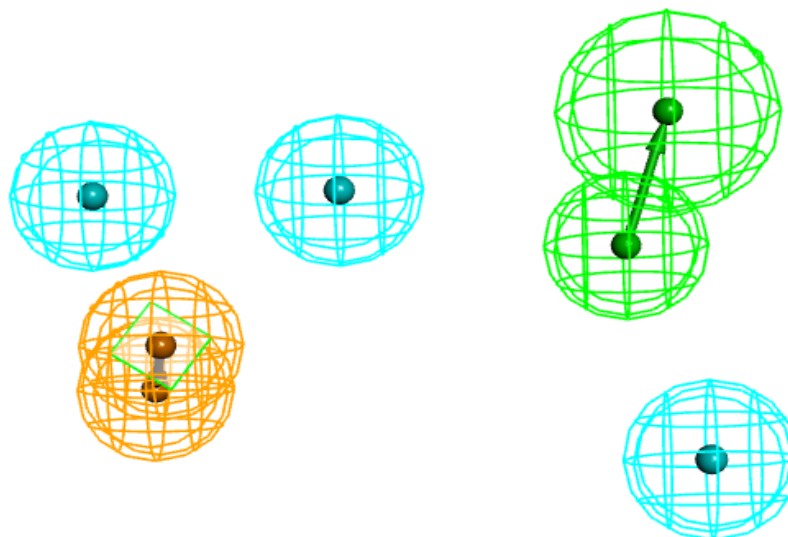
Null cost = 162.712, Fixed cost = 58.5.

Over these 10 models, we must pick one of them as a relevant model, with the lowest total score, the highest cost difference, the low RMSD value and the high correlation coefficient [7]. The product of the table IV. 8 Among the 10 pharmacophore models produced, the first model (Hypo1) had the highest cost difference of 95.31 bits and the total cost value was much closer to the fixed cost compared to other models. The highest cost difference value of Hypo1 means that it can estimate the experimental IC₅₀ value of training compounds with a statistical significance of > 90 per cent (figure IV.8). This model has also shown that it has the highest correlation coefficient value of 0.9609 and the lowest RMSD variance of 1.054, which means that it has a better ability to predict the experimental behavior of training compounds.

Table IV.8. Experimental and estimated activity of individual training set compounds.

| Compound NO. | IC50 value (μM) | | Errors | Fit value | Activity scale | |
|--------------|------------------------------|-----------|--------|-----------|----------------|-----------|
| | Experimental | Estimated | | | Experimental | Estimated |
| 1 | 0.1 | 0.062 | 1.6 | 7.70681 | ++++ | ++++ |
| 2 | 0.1 | 0.22 | 2.2 | 7.16137 | ++++ | +++ |
| 3 | 0.2 | 0.35 | -1 | 6.92014 | +++ | +++ |
| 4 | 0.4 | 1.4 | 1 | 6.51369 | +++ | +++ |
| 5 | 0.5 | 4.3 | 2.1 | 6.38302 | +++ | +++ |
| 6 | 0.5 | 1 | -2 | 6.15953 | +++ | +++ |
| 7 | 0.6 | 13 | 4.7 | 6.00051 | +++ | +++ |
| 8 | 0.6 | 2.4 | -1.2 | 5.88577 | +++ | +++ |
| 9 | 0.7 | 3.1 | -1.2 | 5.5328 | +++ | +++ |
| 10 | 1 | 9.3 | 1.3 | 5.48017 | ++ | ++ |
| 11 | 1.6 | 11 | -1.6 | 5.39279 | ++ | ++ |
| 12 | 1.7 | 32 | 1.6 | 5.371 | ++ | ++ |
| 13 | 3.1 | 14 | -4.3 | 5.01444 | ++ | + |
| 14 | 5.7 | 400 | 2.1 | 3.92986 | ++ | + |
| 15 | 9 | 390 | -1.1 | 3.92914 | ++ | + |
| 16 | 14 | 390 | -2.6 | 3.91655 | + | + |

 Ring aromatic feature,
  Hydrophobic feature,
  Hydrogen bond acceptor feature.

**Figure IV. 8.** The best HypoGen pharmacophore model, Hypo1.

IV. 2. 3. 2. Validation of the pharmacophore model:

The Pharmacophore model could be evaluated using different validation methods; in this part we perform three different validation methods: a) cost analysis, b) test set analysis, and c) Fischer randomization test.

IV. 2. 3. 2. 1. Cost analysis:

The cost parameter produced by HypoGen algorithm in DS such as total cost, fixed cost, and null cost Table IV. 7. Hypo1 (figure IV. 8) model has the cost difference 95.31; correlation coefficient value 0.96 and the RMSD value 1.05 bits (Table IV. 7).

IV. 2. 3. 2. 2. Test set analysis:

Following the selection of the pharmacophore model (Hypo 1), which gives positive results in the cost analysis, this model was validated using 29 test-set compounds, which are different from the training-set compounds (table IV. 8). Using the same training set preparation protocol, the test set was prepared and used to assess if the hypothesis was capable of predicting active compounds other than the training set molecules. The obtained correlation coefficient value for the test compounds set is 0.511 and for the training compounds set is 0.961 (Figure IV. 9).

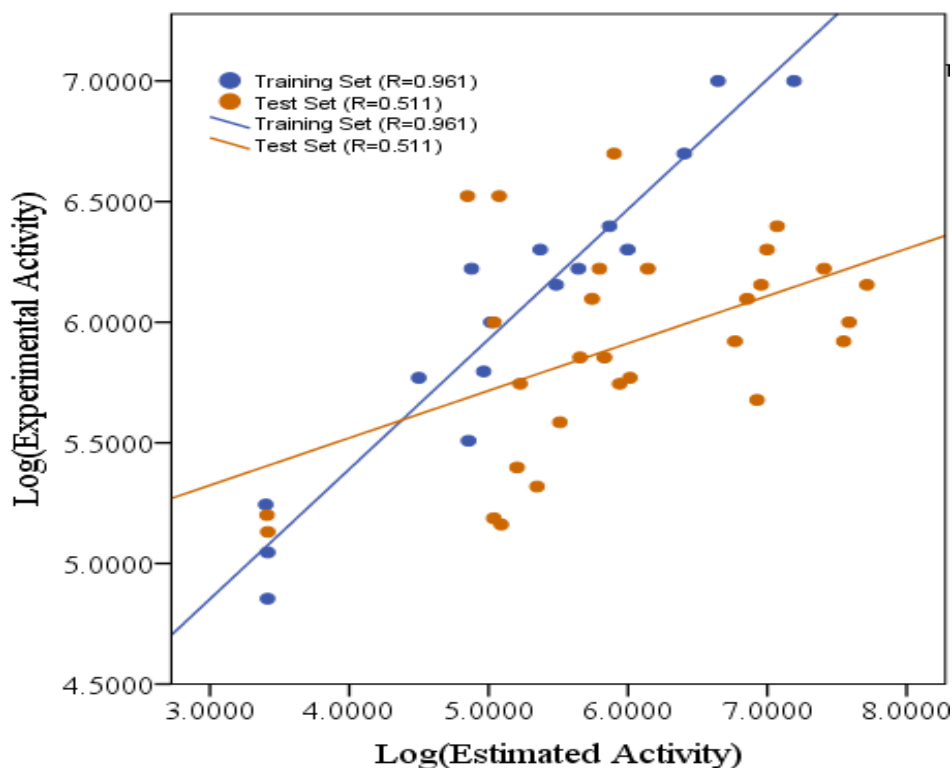


Figure IV. 9. Correlation graph between experimental and estimated activities in logarithmic scale for training and test set compounds based on Hypo1.

IV. 2. 3. 2. 3. Fischer Randomization Method:

The Fisher randomization test used to testify and evaluate Hypo 1. This approach suggests that the pharmacophore model did not produce a random association between the behaviors studied and the training structures of the data set. In order to demonstrate that the Hypo1 model was not created by chance at a 95 % confidence level, out of a total of 19 hypotheses generated by the 14 scramble run, there were no valid hypotheses. This was also immediately omitted from the data tables. None of the five remaining randomly generated hypotheses had a lower total cost than that of Hypo1 (figure IV. 10). This Fischer randomization result indicates clearly that Hypo1 is statistically robust and not randomly generated because Hypo1 represented a true correlation in the training set.

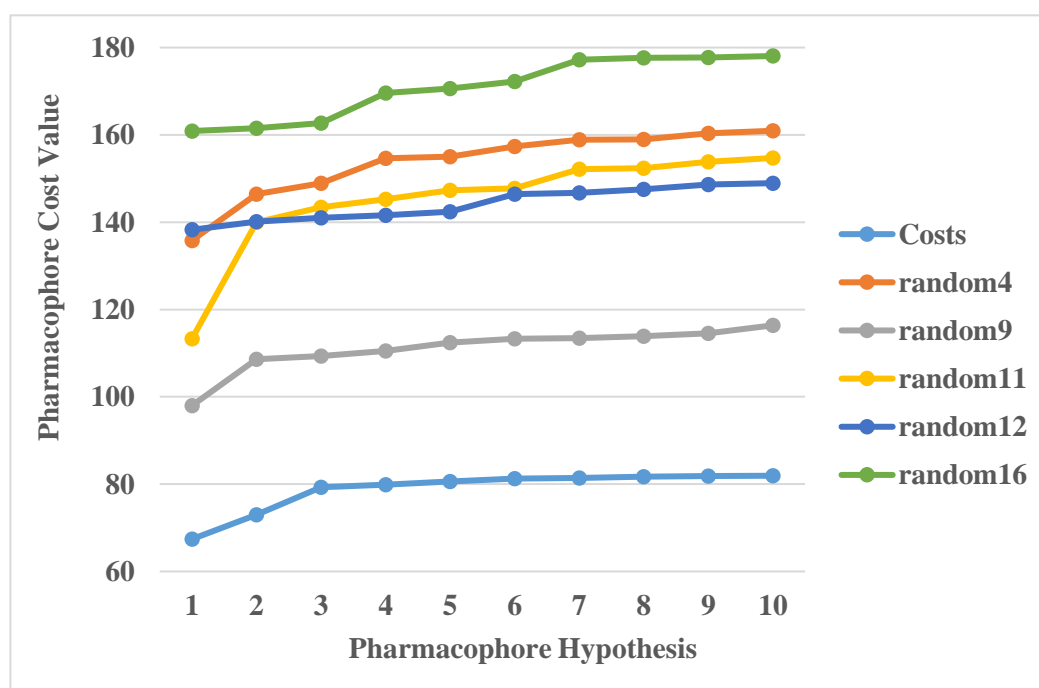


Figure IV.10. The difference in costs between the HypoGen runs and scrambled runs. The 95% confidence level was selected.

IV. 2. 3. 3. Virtual screening:

In this part, we have successfully used virtual screening to classify novel compounds that enhance the function of GSTp1-1 enzymes with their ability to inhibit the formulation of chemically induced cancer. For this reason, we download NCI Database SDF files containing 268667 compounds, including both chemical and natural products [67]. These compounds were primarily

filtered on the basis of the Lipinski rule of five [67], the SMARTS properties of filtration and the ADMET properties predicted by the ADMET Descriptors of the DS [69, 70]. The best conformer generation method was used to produce conformers for each molecule in the NCI database, allowing a maximum energy of 10 kcal/mol over that of the most stable conformation. The 3D query model validated in the section was used to screen the dataset. The Ligand Pharmacophore Mapping option, which is coupled with the best/flexible search method, was used for the screening of the database in order to find novel hit compounds that match all the pharmacophore features. Finally, the collected compounds were further filtered by the criterion that the compounds had an estimated activity value of less than 0.1 μM .

IV. 2. 3. 4. Molecular Docking:

In this section, molecular docking was carried out to identify compounds that were able to fit well into the binding site of GST p1-1 enzyme. As well as molecular docking generates a score for each compound based on the binding affinities of protein-ligand complexes. The Libdock algorithm [70] in DS was used to perform this study. Co-crystallized structure complex of GSTp1-1 with NBD obtained at 1.53Å was downloaded from the protein data bank (PDB ID: 3GUS) [72]. The active site was defined, based on the co-crystallized inhibitor, N11211. The most active compound structure was docked on the same active site of the GSTp1-1 protein. The PMF 04 score and docking interaction of NBD were enlisted in the Table IV. 9. The score (PMF 04) of this compound was 31.4. The NBD inhibitor (most active compound) was able to form 5 hydrogen bonds with Gln 51, Gln 64, Ser 65 and Tyr108, and tree hydrophobic interaction ILE 104, ARG 13 and PRO 53, which is shown in Figure IV. 11. The molecules which showed better PMF 04 score than that of NBD were considered as the potential 'hit' GST p 1-1 poison. Overall, 41 compounds were found to satisfy all of the parameters selected. Meticulous visual inspection and examination of the bindings showed 4 compounds with a better overall docking profile compared to the reset list of NCI. The predicted activity of the 4 lead selected molecules NCI 767370, NCI 750299, NCI 749387 and NCI 750300 were 0.060364 μM , 0.062314 μM , 0.063227 μM and 0.06983 μM respectively.

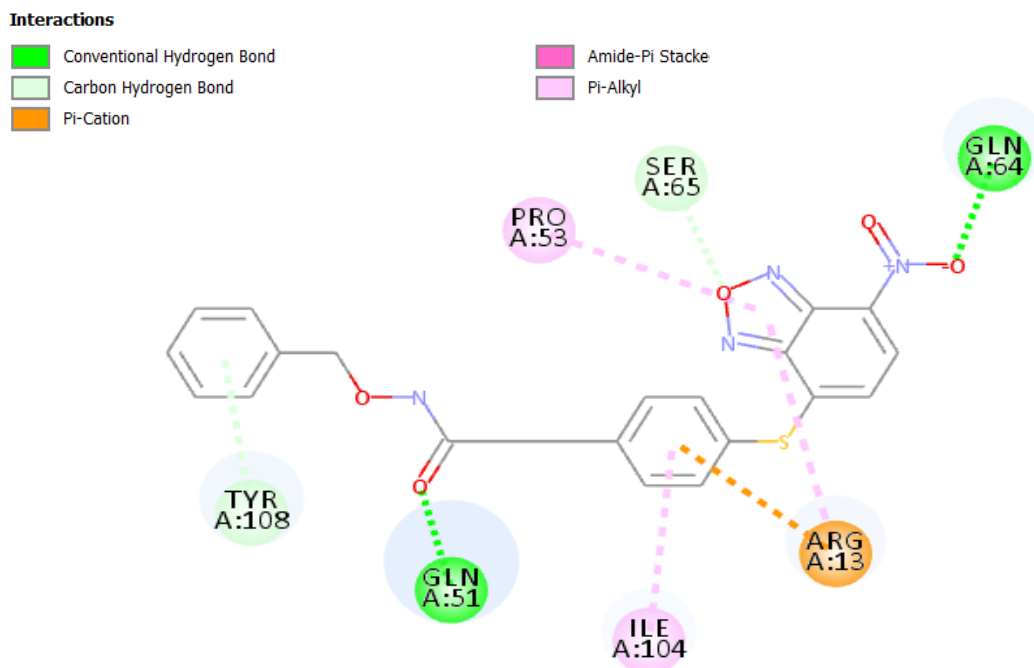


Figure IV. 11. 2D Binding interaction representation of NBD most active compound with active site of GSTp 1-1.

Table IV. 9. Docking interaction of NDB (Most active compound) and virtually screened hit compounds.

| Compound name | Hydrophobic interacting groups | H-bond monitoring | H-bond distance (Å) | PMF_04 score |
|-----------------------------|--------------------------------|-------------------|---------------------|--------------|
| Most active compound | Ile 104, Arg 13, Pro 53 | Gln 51:HE22-O1 | 1.5336 | 31.4 |
| | | Gln 64: HN-O16 | 1.736 | |
| NCI 767370 | Tyr 108, Arg 13 | Ser 65:HN-O4 | 1.67312 | 30.68 |
| | | Ser 65:HG-O4 | 2.05418 | |
| | | Leu 52:O-H28 | 1.57378 | |
| NCI 750299 | Tyr 108, Ile 104 | Ser 65:HN -O28 | 1.80941 | 15.45 |
| | | Ser 65:HG-O28 | 2.1301 | |
| | | Gln 64:OE1-H49 | 2.331 | |
| NCI 749387 | Tyr 108 | Ser 65:HN-O1 | 1.79903 | 26.33 |
| | | Ser 65:HG-O1 | 1.5007 | |
| NCI 750300 | Tyr 49 | Ser 65:HN-O26 | 2.12504 | 20.61 |

III.3. References:

1. C. Hansch and T. Fujita, "p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure," *J. Am. Chem. Soc.*, vol. 86, no. 8, pp. 1616–1626, 1964, doi: 10.1021/ja01062a035.
2. T. J. Hou, J. M. Wang, and X. J. Xu, "Applications of genetic algorithms on the structure-activity correlation study of a group of non-nucleoside HIV-1 inhibitors," *Chemom. Intell. Lab. Syst.*, vol. 45, no. 1–2, pp. 303–310, 1999, doi: 10.1016/S0169-7439(98)00135-X.
3. H. Kubinyi, "3D QSAR in Drug Design: Theory Methods and Applications." Springer US, 1993.
4. S. Peter, J.K. Dhanjal, V. Malik, N. Radhakrishnan, M. Jayakanthan, D. Sundar, *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, Oxford, 2018, pp. 661e676.
5. Y. X. Zhou, L. Xu, Y. P. Wu, and B. L. Liu, "A QSAR study of the antiallergic activities of substituted benzamides and their structures," *Chemom. Intell. Lab. Syst.*, vol. 45, no. 1–2, pp. 95–100, 1999, doi: 10.1016/S0169-7439(98)00092-6.
6. S. Riahi, E. Pourbasheer, M. R. Ganjali, and P. Norouzi, "Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine," *J. Hazard. Mater.*, vol. 166, no. 2–3, pp. 853–859, 2009, doi: 10.1016/j.jhazmat.2008.11.097.
7. E. Pourbasheer, R. Aalizadeh, M. R. Ganjali, and P. Norouzi, "QSAR study of α1β4 integrin inhibitors by GA-MLR and GA-SVM methods.," *Struct. Chem.*, vol. 25, no. 1, pp. 355–370, 2014.
8. W. Li, Y. Tang, Y. L. Zheng, and Z. B. Qiu, "Molecular modeling and 3D-QSAR studies of indolomorphinan derivatives as kappa opioid antagonists.," *Bioorg. Med. Chem.*, vol. 14, no. 3, pp. 601–610, 2006.
9. N. Goudarzi, M. Goodarzi, T. Chen, *Med. Chem. Res.* 21 (2012) 437e443.
10. A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, O. Igglessi-Markopoulou, and G. Kollias, "A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs," *Mol. Divers.*, vol. 14, no. 2, pp. 225–235, 2010, doi: 10.1007/s11030-009-9163-7.
11. H. Khajehsharifi, H. Tavallali, M. Shekoochi, and M. Sadeghi, "Spectrophotometric simultaneous determination of orotic acid, creatinine and uric acid by orthogonal signal correction-partial least squares in spiked real samples," *Drug Test. Anal.*, vol. 5, pp. 353–360, 2013.
12. J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.
13. S. J. Kwon, "Artificial neural networks," *Artif. Neural Networks*, pp. 1–426, 2011, doi: 10.15864/jmscm.1104.
14. H. Khajehsharifi, M. Sadeghi, and E. Pourbasheer, "Spectrophotometric simultaneous determination of ceratine, creatinine, and uric acid in real samples by orthogonal signal correction-partial least squares regression," *Monatshefte fur Chemie*, vol. 140, no. 6, pp. 685–691, 2009, doi: 10.1007/s00706-009-0155-1.
15. D. Rotili *et al.*, "Synthesis and structure-activity relationship of new cytotoxic agents targeting human glutathione-S-transferases," *Eur. J. Med. Chem.*, vol. 89, pp. 156–171, 2014, doi: 10.1016/j.ejmech.2014.10.033.
16. G. Ricci *et al.*, "7-Nitro-2,1,3-benzoxadiazole derivatives, a new class of suicide inhibitors for glutathione S-transferases: Mechanism of action of potential anticancer drugs," *J. Biol. Chem.*, vol. 280, no. 28, pp. 26397–26405, 2005, doi: 10.1074/jbc.M503295200.
17. D. Dalzoppo *et al.*, "Thiol-activated anticancer agents: the state of the art. Anti-Cancer Agents in Medicinal Chemistry," *Former. Curr. Med. Chem. Agents*, vol. 7, no. 1, pp. 4–20., 2017.

18. I. Almi, S. Belaidi, N. Melkemi, and D. Bouzidi, "Chemical reactivity, drug-likeness and structure activity/property relationship studies of 2,1,3-benzoxadiazole derivatives as anti-cancer activity," *J. Bionanoscience*, vol. 12, no. 1, 2018, doi: 10.1166/jbns.2018.1503.
19. R. P. Magisetty, A. Shukla, and B. Kandasubramanian, "Dielectric, Hydrophobic Investigation of ABS/NiFe₂O₄ Nanocomposites Fabricated by Atomized Spray Assisted and Solution Casted Techniques for Miniaturized Electronic Applications," *J. Electron. Mater.*, vol. 47, no. 9, pp. 5640–5656, 2018, doi: 10.1007/s11664-018-6452-x.
20. HyperChem (Molecular Modelling System), Hypercube, Inc., 2008, 1115NW, Gainesville, FL 32601, USA.
21. Gaussian09, R. A. (2009). 1, mj frisch, gw trucks, hb schlegel, ge scuseria, ma robb, jr cheeseman, g. Scalmani, v. Barone, b. Mennucci, ga petersson et al., gaussian. *Inc.*, Wallingford CT, 121, 150-166.
22. Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions, Marvin 17.1.2, 2017, ChemAxon (<http://www.chemaxon.com>)
23. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review.," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.
24. H. Mesa and G. Restrepo, "On dendrograms and topologies," *Match*, vol. 60, no. 2, pp. 371–384, 2008.
25. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
26. P. Liu and W. Long, "Current mathematical methods used in QSAR/QSPR studies," *Int. J. Mol. Sci.*, vol. 10, no. 5, pp. 1978–1998, 2009, doi: 10.3390/ijms10051978.
27. SPSS 21 for Windows, SPSS software packages, SPSS Inc., 444 North Michigan Avenue, Suite 3000, Chicago, Illinois, 60611, USA.
28. B. J. Wythoff, "Backpropagation neural networks. A tutorial," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 2, pp. 115–155, 1993, doi: 10.1016/0169-7439(93)80052-J.
29. J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design: an Introduction*, Wiley-VCH, 1999.
30. I.A. Basheer, M. Hajmeer, *J. Microbiol. Methods* 43 (2000) 3e31.
31. S. Mukherjee, K. Ashish, N. B. Hui, and S. Chattopadhyay, "Modeling Depression Data: Feed Forward Neural Network vs. Radial Basis Function Neural Network," *Am. J. Biomed. Sci.*, pp. 166–174, 2014, doi: 10.5099/aj140300166.
32. K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. 2015.
33. Dorf, R.C. and R.H. Bishop, *Modern Control Systems*, Addison-Wesley, Menlo Park, CA, 1998.
34. N. Chirico and P. Gramatica, "Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient," *J. Chem. Inf. Model.*, vol. 51, no. 9, pp. 2320–2335, 2011, doi: 10.1021/ci200211n.
35. K. Roy, "On some aspects of validation of predictive quantitative structure-activity relationship models," *Expert Opin. Drug Discov.*, vol. 2, no. 12, pp. 1567–1577, 2007, doi: 10.1517/17460441.2.12.1567.
36. A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods*, vol. 71, no. C, pp. 58–63, 2015, doi: 10.1016/j.ymeth.2014.08.005.
37. K. H. Kim, N. D. Kim, and B. L. Seong, "Pharmacophore-based virtual screening: A review of recent applications," *Expert Opin. Drug Discov.*, vol. 5, no. 3, pp. 205–222, 2010, doi: 10.1517/17460441003592072.

38. J.L. Medina-Franco, G.M. Maggiora, *Chemoinformatics for Drug Discovery*, John Wiley and Sons, 2013.
39. H. Koeppen, J. Kriegl, U. Lessel, *Virtual Screening. Principles, Challenges and Practical Guidelines*, John Wiley & Sons, 2011.
40. J. Gasteiger, T.E. Eds (Eds.), *Molecular Modelling Neural Networks in Chemistry and Drug Design 150 and More Basic*, Wiley-VCH, 2003.
41. V. K. AGRAWAL and P. V KHADIKAR, "QSAR prediction of toxicity of nitrobenzenes," *Bioorg. Med. Chem.*, vol. 9, no. 11, pp. 3035–3040, 2001.
42. A. P. Piotrowski and J. J. Napiorkowski, "Author's personal copy Optimizing neural networks for river flow forecasting – Evolutionary Computation methods versus the Levenberg – Marquardt approach," doi: 10.1016/j.jhydrol.2011.06.019.
43. P. P. Roy and K. Roy, "On some aspects of variable selection for partial least squares regression models," *QSAR Comb. Sci.*, vol. 27, no. 3, pp. 302–313, 2008, doi: 10.1002/qsar.200710043.
44. S. Weaver, M.P. Gleeson, *J. Mol. Graph. Model.* 26 (2008) 1315e1326. S. Weaver and M. P. Gleeson, "The importance of the domain of applicability in QSAR modeling," *J. Mol. Graph. Model.*, vol. 26, no. 8, pp. 1315–1326, 2008, doi: 10.1016/j.jmgm.2008.01.002.
45. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, and P. Gramatica, "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs," *Environ. Health Perspect.*, vol. 111, no. 10, pp. 1361–1375, 2003, doi: 10.1289/ehp.5758.
46. A. GISSI, O. NICOLOTTI, A. CAROTTI, and E. Al, "Integration of QSAR models for bioconcentration suitable for REACH," *Sci. Total Environ.*, vol. 456, pp. 325–332, 2013.
47. P. Gramatica, P. Pilutti, and E. Papa, "QSAR prediction of ozone tropospheric degradation," *QSAR Comb. Sci.*, vol. 22, no. 3, pp. 364–373, 2003, doi: 10.1002/qsar.200390026.
48. F. Zheng *et al.*, "QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release," *Bioorganic Med. Chem.*, vol. 14, no. 9, pp. 3017–3037, 2006, doi: 10.1016/j.bmc.2005.12.036.
49. I. Cortés-ciriano and A. Bender, "Reliable Prediction Errors for Deep Neural Networks Using Test-Time Dropout."
50. M. Bem *et al.*, "7-nitrobenzo[c] [1, 2, 5]oxadiazole (nitrobenzofurazan) derivatives with a sulfide group at the 4-position. Synthesis and physical properties," *Rev. Roum. Chim.*, vol. 63, no. 2, pp. 149–155, 2018.
51. K. H, "Virtual screening - what does it give us?," *Curr. Opin. Drug Discov. Devel.*, vol. 12, no. 3, pp. 397–407, 2009.
52. C. McInnes, "Virtual screening strategies in drug discovery," *Curr. Opin. Chem. Biol.*, vol. 11, no. 5, pp. 494–502, 2007, doi: 10.1016/j.cbpa.2007.08.033.
53. J. H. Duffus, M. Nordberg, and D. M. Templeton, "Glossary of terms used in toxicology, 2nd edition (IUPAC recommendations 2007)," *Pure Appl. Chem.*, vol. 79, no. 7, pp. 1153–1344, 2007, doi: 10.1351/pac200779071153.
54. C. N. Cavasotto and A. J. Orry, "Ligand docking and structure-based virtual screening in drug discovery," *Curr. Top. Med. Chem.*, vol. 7, no. 10, pp. 1006–1014, 2007, doi: 10.2174/156802607780906753.
55. O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading," *Comput Chem*, vol. 31, no. 2, pp. 455–461, 2010, doi: 10.1002/jcc.21334.
56. T. Salah, S. Belaidi, N. Melkemi, I. Daoud, and S. Boughdiri, "In silico investigation by conceptual DFT and molecular docking of antitrypanosomal compounds for understanding

- cruzain inhibition,” *J. Theor. Comput. Chem.*, vol. 15, no. 3, 2016, doi: 10.1142/S0219633616500218.
57. P. Aparoy, K. Kumar Reddy, and P. Reddanna, “Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors,” *Curr. Med. Chem.*, vol. 19, no. 22, pp. 3763–3778, 2012, doi: 10.2174/092986712801661112.
 58. B. Chandrasekaran, N. Agrawal, and S. Kaushik, “Pharmacophore development,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 677–687, 2018, doi: 10.1016/B978-0-12-809633-8.20276-8.
 59. V. Di Paolo *et al.*, “Synthesis and characterisation of a new benzamide-containing nitrobenzoxadiazole as a GSTP1-1 inhibitor endowed with high stability to metabolic hydrolysis,” *J. Enzyme Inhib. Med. Chem.*, vol. 34, no. 1, pp. 1131–1139, 2019, doi: 10.1080/14756366.2019.1617287.
 60. BIOVIA, Dassault Systèmes, Discovery Studio, 4.1, San Diego: Dassault Systèmes, 2016.
 61. H. V. Prabhu and G. S. Nagaraja, “Quality of service guaranteed delay sensitive polling algorithm for WiMax network: PQ-Poll,” *Proc. 2017 2nd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2017*, pp. 0–5, 2017, doi: 10.1109/ICECCT.2017.8117841.
 62. A. Allouche, “Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares,” *J. Comput. Chem.*, vol. 32, pp. 174–182, 2012, doi: 10.1002/jcc.
 63. S. Pal *et al.*, “Ligand-based Pharmacophore Modeling, Virtual Screening and Molecular Docking Studies for Discovery of Potential Topoisomerase I Inhibitors,” *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 291–310, 2019, doi: 10.1016/j.csbj.2019.02.006.
 64. Singh and Singh, 2013
 65. S. John, S. Thangapandian, M. Arooj, J. C. Hong, K. D. Kim, and K. W. Lee, “Development, evaluation and application of 3D QSAR Pharmacophore model in the discovery of potential human renin inhibitors.,” *BMC Bioinformatics*, vol. 12 Suppl 14, no. Suppl 14, 2011, doi: 10.1186/1471-2105-12-S14-S4.
 66. S. G. Rohrer and K. Baumann, “Impact of benchmark data set topology on the validation of virtual screening methods: Exploration and quantification by spatial statistics,” *J. Chem. Inf. Model.*, vol. 48, no. 4, pp. 704–718, 2008, doi: 10.1021/ci700099u.
 67. T. Grkovic *et al.*, “National Cancer Institute (NCI) Program for Natural Products Discovery: Rapid Isolation and Identification of Biologically Active Natural Products from the NCI Prefractionated Library,” *ACS Chem. Biol.*, vol. 15, no. 4, pp. 1104–1114, 2020, doi: 10.1021/acscchembio.0c00139.
 68. L. Z. Benet, C. M. Hosey, O. Ursu, and T. I. Opreab, “BDDCS, the Rule of 5 and Drugability,” *Adv Drug Deliv Rev*, vol. 101, no. 2, pp. 89–98, 2016, doi: 10.1016/j.addr.2016.05.007.
 69. J. A. Pradeepkiran, K. K. Kumar, Y. N. Kumar, and M. Bhaskar, “Modeling, molecular dynamics, and docking assessment of transcription factor rho: A potential drug target in brucella melitensis 16M,” *Drug Des. Devel. Ther.*, vol. 9, pp. 1897–1912, 2015, doi: 10.2147/DDDT.S77020.
 70. P. Ponnann *et al.*, “ 2D-QSAR, Docking Studies, and In Silico ADMET Prediction of Polyphenolic Acetates as Substrates for Protein Acetyltransferase Function of Glutamine Synthetase of Mycobacterium tuberculosis ,” *ISRN Struct. Biol.*, vol. 2013, pp. 1–12, 2013, doi: 10.1155/2013/373516.
 71. D. J. Diller and K. M. Merz, “High throughput docking for library design and library prioritization,” *Proteins Struct. Funct. Genet.*, vol. 43, no. 2, pp. 113–124, 2001, doi: 10.1002/1097-0134(20010501)43:2<113::AID-PROT1023>3.0.CO;2-T.

72. L. Federici *et al.*, “Structural basis for the binding of the anticancer compound 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol to human glutathione S-transferases,” *Cancer Res.*, vol. 69, no. 20, pp. 8025–8034, 2009, doi: 10.1158/0008-5472.CAN-09-1314.

Chapter V:

Conclusion

Multidrug resistance to chemotherapy drugs represents an obstacle in human cancer treatment. It encouraged extensive research into the discovery of new and novel mechanisms that could overcome this obstacle. Many recent researches have mentioned the importance role of GST in diverse cellular processes as well as in conferring resistance to chemotherapy. For that purpose, we implemented elaborate ligand-based and structure-based computational workflows to explore the structural features necessary for potent inhibition of GSTp 1-1 using 45 different inhibitors.

In this work, the QSAR analyses were carried out using MLR and ANN methodologies. We have identified four critical descriptors which successfully predict the GSTP1-1 inhibitory activity. The results of validation indicate the accuracy and robustness of the proposed QSAR model. Based on the proposed QSAR model coupled with similarity search technique, we have identified a series of potential novel compounds. This series has been used as a primary step for

predicting the GSTP1-1 inhibitory activity. It is ought to test the reliability of these predictions in vitro.

The next part, provided development of ligand-based pharmacophore model by 3D-QSAR Pharmacophore Generation protocol. The best quantitative pharmacophore (Hypo1) was chosen among 10 other pharmacophores. The Hypo1 model was used as a 3D query for the virtual screening of 268667 drug-like molecules from NCI database. By applying selective parameters number of molecules funnel down to 3805 hits, docked at the active sites of GSTp1-1 (PDB ID: 3GUS) by LibDock protocol on DS. Finally, four hits were selected based on the molecular interaction, structure and scoring.

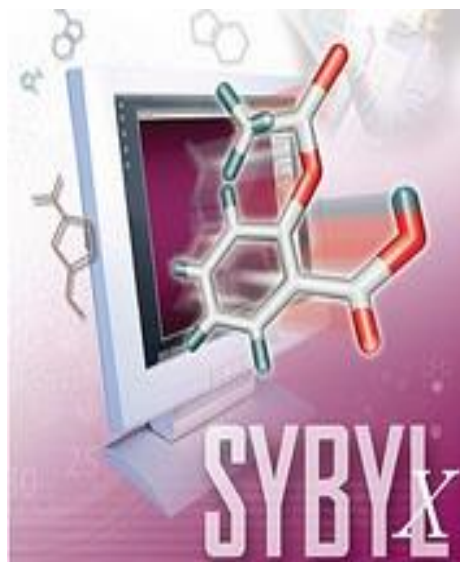
In this work, we focused more on the use of computer Ligand-Based drug design. For that and as future work we aim to use Molecular Dynamic, one of the techniques of computer Structure-Based drug design, to confirm the quality of our extracted compounds; then we will be able to test their activity in vitro and in vivo.

Appendix

Appendix A: table of calculated descriptors.

| comp | pIC50 | HE | Pol | HBA | HBD | PSA | n _{rot} | N(C=O) | q(s) | μ | HOMO | LUMO | wiener | Qmin | Qmax | AM1_HF | Pol | a _{hyd} | E _{sol} | logP(o/w) | V | MW | ASA |
|------|-------|--------|-------|-----|-----|--------|------------------|--------|--------|--------|--------|--------|--------|-------|--------|---------|--------|------------------|------------------|-----------|---------|---------|---------|
| 1 | 6.097 | -17.57 | 28.84 | 5 | 1 | 102.29 | 8 | 0 | -0.156 | 8.364 | -0.257 | -0.135 | 943 | 0.862 | -0.724 | 43.503 | 40.530 | 14 | -13.369 | 2.971 | 307.077 | 297.335 | 527.195 |
| 2 | 6.222 | -14.63 | 26.86 | 4 | 0 | 82.06 | 3 | 0 | -0.182 | 9.143 | -0.259 | -0.134 | 702 | 0.911 | -0.496 | 154.486 | 34.394 | 14 | -10.894 | 3.551 | 293.449 | 273.272 | 486.633 |
| 3 | 6 | -14.07 | 28.69 | 4 | 0 | 82.06 | 4 | 0 | -0.102 | 9.175 | -0.258 | -0.134 | 851 | 0.894 | -0.496 | 148.287 | 37.487 | 15 | -8.211 | 3.685 | 301.693 | 287.299 | 521.864 |
| 4 | 6.699 | -14.19 | 30.53 | 4 | 0 | 82.06 | 5 | 0 | -0.182 | 8.995 | -0.258 | -0.135 | 1020 | 0.861 | -0.485 | 143.245 | 40.581 | 16 | -13.211 | 3.773 | 321.707 | 301.326 | 531.716 |
| 5 | 6.301 | -13.77 | 32.36 | 4 | 0 | 82.06 | 6 | 0 | -0.179 | 9.141 | -0.254 | -0.132 | 1210 | 0.861 | -0.483 | 141.057 | 43.674 | 17 | -13.662 | 4.215 | 319.575 | 315.353 | 545.500 |
| 6 | 5.509 | -17.77 | 29.42 | 6 | 1 | 119.36 | 4 | 1 | -0.153 | 10.717 | -0.255 | -0.13 | 1035 | 0.89 | -0.622 | 213.810 | 37.758 | 14 | -20.162 | 3.226 | 306.052 | 317.281 | 546.316 |
| 7 | 5.854 | -19.77 | 29.42 | 6 | 1 | 119.36 | 4 | 1 | -0.168 | 9.417 | -0.263 | -0.137 | 1074 | 0.876 | -0.616 | 67.158 | 37.758 | 14 | -16.425 | 3.265 | 287.020 | 317.281 | 496.255 |
| 8 | 5.77 | -19.52 | 29.42 | 6 | 1 | 119.36 | 4 | 1 | -0.177 | 7.742 | -0.265 | -0.138 | 1113 | 0.884 | -0.631 | 68.651 | 37.758 | 14 | -14.860 | 3.228 | 284.716 | 317.281 | 496.688 |
| 9 | 5.745 | -14.33 | 33.09 | 5 | 0 | 108.36 | 6 | 1 | -0.177 | 8.914 | -0.262 | -0.136 | 1470 | 0.891 | -0.562 | 69.307 | 43.945 | 16 | -9.601 | 3.833 | 354.675 | 345.335 | 594.345 |
| 10 | 6.097 | -19.07 | 30.13 | 5 | 1 | 125.15 | 4 | 1 | -0.183 | 9.289 | -0.263 | -0.137 | 1113 | 0.899 | -0.85 | 118.278 | 38.722 | 14 | -18.192 | 2.494 | 309.326 | 316.297 | 520.857 |
| 11 | 5.921 | -27.47 | 30.77 | 6 | 2 | 131.39 | 4 | 1 | -0.168 | 5.599 | -0.267 | -0.141 | 1280 | 0.884 | -0.525 | 115.295 | 39.524 | 14 | -20.003 | 2.549 | 329.861 | 332.296 | 574.577 |
| 12 | 6.222 | -13.44 | 33.8 | 5 | 0 | 102.37 | 4 | 1 | -0.164 | 7.996 | -0.262 | -0.137 | 1449 | 0.892 | -0.562 | 130.267 | 44.910 | 16 | -17.223 | 3.047 | 329.933 | 344.351 | 557.249 |
| 13 | 6.301 | -20.61 | 32.6 | 6 | 1 | 120.39 | 5 | 1 | -0.168 | 7.774 | -0.265 | -0.138 | 1470 | 0.878 | -0.504 | 115.223 | 42.618 | 15 | -18.655 | 2.992 | 341.250 | 346.323 | 576.164 |
| 14 | 5.921 | -18.82 | 39.81 | 7 | 1 | 129.62 | 6 | 1 | -0.186 | 11.014 | -0.262 | -0.136 | 2654 | 0.862 | -0.525 | 71.233 | 54.461 | 18 | -24.517 | 4.082 | 414.963 | 416.414 | 678.978 |
| 15 | 6 | -20.45 | 42.26 | 6 | 1 | 120.39 | 7 | 1 | -0.175 | 8.203 | -0.263 | -0.138 | 2997 | 0.88 | -0.528 | 151.489 | 55.845 | 21 | -26.090 | 4.780 | 404.869 | 422.421 | 667.128 |
| 16 | 5.678 | -22.57 | 34.44 | 6 | 2 | 131.39 | 6 | 1 | -0.181 | 8.31 | -0.263 | -0.137 | 1684 | 0.895 | -0.841 | 73.148 | 45.712 | 16 | -28.447 | 2.156 | 344.086 | 360.350 | 578.221 |
| 17 | 6 | -15.03 | 38.82 | 6 | 1 | 114.4 | 7 | 1 | -0.172 | 8.552 | -0.233 | -0.137 | 2164 | 0.892 | -0.792 | 125.241 | 52.863 | 18 | -15.201 | 2.727 | 390.059 | 387.420 | 637.308 |
| 18 | 6.398 | -21.48 | 31.25 | 6 | 1 | 119.36 | 5 | 1 | -0.161 | 7.314 | -0.268 | -0.141 | 1298 | 0.883 | -0.566 | 72.943 | 40.851 | 15 | -27.148 | 3.316 | 356.958 | 331.308 | 595.967 |
| 19 | 6.155 | -20.98 | 33.09 | 6 | 1 | 119.36 | 6 | 1 | -0.173 | 5.09 | -0.262 | -0.138 | 1506 | 0.888 | -0.61 | 49.430 | 43.945 | 16 | -15.638 | 3.404 | 319.921 | 345.335 | 556.227 |
| 20 | 7 | -20.52 | 45.93 | 6 | 1 | 120.39 | 9 | 1 | -0.186 | 7.978 | -0.259 | -0.135 | 3794 | 0.907 | -0.513 | 133.960 | 62.032 | 23 | -13.829 | 4.956 | 450.175 | 450.475 | 734.730 |
| 21 | 5.796 | -19.21 | 31.25 | 6 | 1 | 119.36 | 5 | 1 | -0.101 | 7.652 | -0.263 | -0.138 | 1318 | 0.879 | -0.623 | 59.593 | 40.851 | 15 | -23.080 | 3.362 | 329.101 | 331.308 | 553.827 |
| 22 | 6.155 | -15.23 | 34.92 | 5 | 0 | 108.36 | 7 | 1 | -0.107 | 5.002 | -0.264 | -0.139 | 1716 | 0.883 | -0.539 | 62.282 | 47.038 | 17 | -17.045 | 3.967 | 361.441 | 359.362 | 629.708 |
| 23 | 6.699 | -18.93 | 44.1 | 6 | 1 | 120.39 | 8 | 1 | -0.154 | 6.639 | -0.257 | -0.139 | 3384 | 0.881 | -0.601 | 143.939 | 58.939 | 22 | -24.159 | 4.914 | 448.109 | 436.448 | 735.218 |
| 24 | 7 | -19.37 | 36.27 | 6 | 1 | 120.39 | 7 | 1 | -0.169 | 6.96 | -0.261 | -0.138 | 1987 | 0.868 | -0.598 | 113.083 | 48.805 | 17 | -15.319 | 3.214 | 365.299 | 374.377 | 602.257 |
| 25 | 5.77 | -14.62 | 25.26 | 5 | 0 | 108.36 | 6 | 1 | -0.17 | 8.727 | -0.261 | -0.137 | 758 | 0.868 | -0.571 | 33.328 | 33.811 | 11 | -11.255 | 1.754 | 264.362 | 283.264 | 474.574 |
| 26 | 5.745 | -14.73 | 25.97 | 5 | 1 | 111.16 | 5 | 1 | -0.141 | 8.418 | -0.261 | -0.137 | 772 | 0.86 | -0.701 | 82.146 | 34.776 | 11 | -25.028 | 1.150 | 306.268 | 282.280 | 535.557 |
| 27 | 5.161 | -21.72 | 28.45 | 6 | 2 | 131.39 | 7 | 1 | -0.161 | 9.758 | -0.256 | -0.138 | 1058 | 0.89 | -0.852 | 27.793 | 38.672 | 12 | -23.332 | 0.418 | 302.324 | 312.306 | 524.474 |
| 28 | 5.319 | -12.76 | 30.71 | 6 | 0 | 102.37 | 5 | 1 | -0.168 | 7.408 | -0.252 | -0.139 | 1180 | 0.865 | -0.728 | 93.697 | 42.723 | 14 | -11.051 | 1.153 | 311.738 | 322.345 | 537.035 |
| 29 | 5.187 | -13.36 | 32.63 | 6 | 0 | 119.44 | 5 | 2 | -0.173 | 6.461 | -0.262 | -0.138 | 1518 | 0.87 | -0.572 | 48.988 | 45.285 | 14 | -21.856 | 0.688 | 357.362 | 350.355 | 593.184 |
| 30 | 6.222 | -15.97 | 36.51 | 6 | 1 | 146.69 | 9 | 2 | -0.177 | 9.359 | -0.255 | -0.132 | 1980 | 0.988 | -0.674 | -46.980 | 51.316 | 15 | -12.332 | 2.536 | 395.410 | 398.396 | 653.170 |
| 31 | 6.398 | -18.8 | 40.67 | 6 | 1 | 146.69 | 10 | 2 | -0.161 | 12.838 | -0.254 | -0.129 | 2778 | 1.044 | -0.747 | -3.100 | 55.263 | 18 | -21.174 | 3.114 | 443.567 | 432.413 | 737.162 |
| 32 | 6.523 | -11.03 | 29.27 | 4 | 0 | 82.06 | 4 | 0 | -0.157 | 9.587 | -0.255 | -0.133 | 851 | 0.888 | -0.495 | 97.082 | 41.488 | 15 | -8.928 | 4.042 | 317.507 | 293.347 | 523.099 |
| 33 | 6.523 | -11.16 | 24.54 | 4 | 0 | 82.06 | 3 | 0 | -0.268 | 9.427 | -0.255 | -0.132 | 494 | 0.87 | -0.491 | 94.436 | 33.541 | 12 | -13.105 | 3.623 | 287.975 | 253.282 | 476.996 |
| 34 | 5.131 | -22.54 | 23.98 | 6 | 2 | 122.52 | 5 | 0 | -0.148 | 11.361 | -0.253 | -0.13 | 624 | 0.879 | -0.703 | 19.728 | 32.051 | 11 | -18.039 | 0.630 | 249.266 | 271.253 | 441.841 |
| 35 | 5.854 | -18.41 | 25.17 | 5 | 1 | 102.29 | 6 | 0 | -0.179 | 9.273 | -0.258 | -0.134 | 652 | 0.881 | -0.715 | 56.911 | 34.343 | 12 | -14.088 | 2.087 | 264.200 | 269.281 | 474.861 |
| 36 | 6.301 | -16.95 | 32.51 | 5 | 1 | 102.29 | 10 | 0 | -0.184 | 9.51 | -0.257 | -0.134 | 1314 | 0.879 | -0.734 | 32.825 | 46.717 | 16 | -10.382 | 3.855 | 359.550 | 325.389 | 617.329 |
| 37 | 4.852 | -18.39 | 26.41 | 6 | 1 | 119.36 | 8 | 0 | 0.254 | 5.46 | -0.268 | -0.152 | 1034 | 0.882 | -0.732 | 13.792 | 41.332 | 13 | -16.763 | 0.984 | 325.755 | 313.334 | 565.666 |
| 38 | 5.046 | -18.9 | 26.98 | 7 | 1 | 136.43 | 8 | 0 | 1.14 | 3.55 | -0.288 | -0.16 | 1127 | 1.14 | -0.728 | -6.841 | 42.134 | 13 | -29.038 | 1.577 | 342.215 | 329.333 | 573.884 |

Appendix B: Computer aided drug design software.



GaussView



Gaussian, Inc.

Carnegie Office Park - Building 6
Pittsburgh PA 15106 USA

ChemOffice.Com

CambridgeSoft
Life Science Enterprise Solutions
Lösungen für biochemische Unternehmen
Solutions d'entreprise Life Science
ライフサイエンス・エンタープライズソリューション




ChemDraw
Chemical Structure Drawing Standard



MOE
Molecular
Operating
Environment



HyperChem



Microsoft
Windows Vista
compatible

