الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ FERHAT ABBAS - SETIF1

FACULTÉ DE TECHNOLOGIE

THÈSE

Présentée au Département de Génie des Procédés

Pour l'obtention du diplôme de

DOCTORAT

Domaine : Sciences et Technologie

Filière : Génie des procédés **Option :** Génie Pharmaceutique

Par

HAMMOUDI Nour El Houda

THÈME

In-silico drug discovery of new potential NF-κB (nuclear factor-kappa B) inhibitors, using the computer aided drug design tools

Soutenue le 24/09/2022 devant le Jury :

BENANIBA Mohamed Tahar	Professeur	Univ. Ferhat Abbas Sétif 1	Président
BENGUERBA Yacine	Professeur	Univ. Ferhat Abbas Sétif 1	Directeur de thèse
BENAICHA Mohamed	Professeur	Univ. Ferhat Abbas Sétif 1	Co-Directeur
BENTOUHAMI Embarek	Professeur	Univ. Ferhat Abbas Sétif 1	Examinateur
KHEMILI TALBI Souad	Professeur	Univ. Boumerdes	Examinatrice
EL KOLLI Hayat	M.C.A.	Univ. Ferhat Abbas Sétif 1	Examinatrice
SOBHI Widad	Professeur	Univ. Ferhat Abbas Sétif 1	Invitée

Contents

CHAPTER I: The Nuclear Factor Kb (Nf-Kb) and its Role in Cancer

CHAPTER II : Quantitative Structure Activity Relationship study

CHAPTER III: Basics in Molecular Docking

CHAPTER IV: Results and Discussion

Part I: *In silico* drug discovery of IKK-β inhibitors from 2-amino-3-cyano-4-alkyl-6-(2 hydroxyphenyl) pyridine derivatives based on QSAR, docking, molecular dynamics and drug likeness evaluation studies

Part II: Comparative Study Between Many Predictive QSAR models (MLR and ANN Regressions)

CONTENTS

ACKNOWLEDGEMENT

*In the name of "**Allah**", who gave me the will, patience, and health to complete this work.*

*I would like to express my special appreciation and thanks to my supervisor **Pr**. **Yacine BENGUARBA** who gave me the golden opportunity to do this thought-provoking research project. He has been a tremendous mentor, and adviser for me. I am really thankful to him.*

*My profuse thanks go to my Co-supervisor **Pr. Mohamed BENAICHA** for his valuable advices.*

*I express my sincere gratitude and appreciation to my reading committee members, **Pr. Mohamad Tahar BENANINBA, Pr. Embarek BENTOUHAMI**. **Pr. Souad KHEMILI TALBI** and, Dr. **Hayat EL KOLL**I for their time to read this thesis and for having accepted to judge this work.*

*I would especially like to thank **Pr. Widad SOBHI** for the support she has given me, her advice that has been of assistance throughout this work.*

I should also thank all my lab members and my friends for their support.

*Finally, I must express my very profound gratitude to my **parents** and my **husband** for providing me with boundless support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them.*

Thank you!

Dedicated

To

My father

For supporting and encouraging me to believe in myself

My Mother

A strong and gentle soul who taught me to trust in Allah and believe in hard work

My Lovely sister

Sarah

My brothers Abdellatif and Rafik

My Husband

My Father-in-law

My Mother-in-law

My Sister -in-law

My best friends

and

To all who will be there during my presentation

List of Works

International Publications

1. **Hammoudi. N. E. H.**, Benguerba. Y., Sobhi W., QSAR modeling of thirty active compounds for the inhibition of the Acetylcholinesterase enzyme, proceeding of Current Research in Bioinformatics 2020
2. Lemaoui T., **Hammoudi. N. E. H**., Benguerba. Y., Molecular Docking of new active compounds towards the Acetylcholinesterase enzyme, proceeding of Current Research in Bioinformatics 2020
3. Lemaoui.T., **Hammoudi, N. E. H**., Alnashef, I. M., Balsamo, M., Erto, A., Ernst, B., & Benguerba, Y. (2020). Quantitative structure properties relationship for deep eutectic solvents using Sσ-profile as molecular descriptors. *Journal of Molecular Liquids*, 113165.
4. Lemaoui, T., Darwish, A. S., **Hammoudi, N. E. H**., Abu Hatab, F., Attoui, A., Alnashef, I. M., & Benguerba, Y. (2020). Prediction of Electrical Conductivity of Deep Eutectic Solvents Using COSMO-RS Sigma Profiles as Molecular Descriptors: A Quantitative Structure–Property Relationship Study. *Industrial & Engineering Chemistry Research*, *59*(29), 13343-13354.
5. **N. E. H, Hammoudi,** Y. Benguerba., A. Attoui, Hognon, C., Lemaoui, T., Sobhi, W., ... & Monari, A. (2020). In silico drug discovery of IKK-β inhibitors from 2-amino-3-cyano-4-alkyl-6-(2-hydroxyphenyl) pyridine derivatives based on QSAR, docking, molecular dynamics and drug-likeness evaluation studies. *Journal of Biomolecular Structure and Dynamics* 2020.
6. A. Attoui., Sobhi, W, **Hammoudi, N. E. H**, Y. Benguerba., Hognon, C., Lemaoui, 2020, Fragment-Based Drug Design of Antitumoral Molecules Polo-like kinase 1 inhibitors: in-silico approach. Journal of letters in drug design and discovery
7. **N. E. H, Hammoudi,** Sobhi, W., Attoui, A., Lemaoui, T., Erto, A., & Benguerba, Y. (2020). In silico drug discovery of Acetylcholinesterase and Butyrylcholinesterase enzymes inhibitors based on Quantitative Structure-Activity Relationship (QSAR) and drug-likeness evaluation. *Journal of Molecular Structure*, 129845.
8. Lemaoui, T., Darwish, A.S., Attoui, A., Hatab, F.A., **Hammoudi, N.E.H**., Benguerba, Y., Vega, L.F. and Alnashef, I.M., 2020. Predicting the density and viscosity of hydrophobic eutectic solvents: towards the development of sustainable solvents. *Green Chemistry*, *22*(23), pp.8511-8530.
9. Lemaoui, T., Abu Hatab, F., Darwish, A. S., Attoui, A., **Hammoudi, N. E.** H., Almustafa, G., ... & Alnashef, I. M. (2021). Molecular-Based Guide to Predict the pH of Eutectic Solvents: Promoting an Efficient Design Approach for New Green Solvents. *ACS Sustainable Chemistry & Engineering*, *9*(17), 5783-5808
10. Ferkous, H., Rouibah, K., **Hammoudi, N. E. H**., Alam, M., Djilani, C., Delimi, A., ... & Benguerba, Y. (2022). The Removal of a Textile Dye from an Aqueous Solution Using a Biocomposite Adsorbent. *Polymers*, *14*(12), 2396.

LIST OF WORKS

International Conferences

1. **N.E.H. Hammoudi**, Y. Benguerba, W. Sobhi, QSAR modeling of thirty active compounds for the inhibition of the Acetylcholinesterase enzyme, International Bioinformatics Day November 05, 2019. At the University Library UMBB, Boumerdes, Algeria
2. T. Lemaoui, **N.E.H. Hammoudi**, Y. Benguerba, Molecular Docking of new active compounds towards the Acetylcholinesterase enzyme, International Bioinformatics Day November 05, 2019. At the University Library UMBB, Boumerdes, Algeria
3. **N. E.H. Hammoudi**, M. Benaicha, Y. Benguerba, K.E ,Kanouni ,Prévision Quantitative de la toxicité des composés nitrobenzènes par modélisation QSTR (Quantitative structure Toxicity Relationship), *First International Workshop On Environmental Engineering (IWEE'19)* November 16,17 2019, setif

- National Conference

N. E.H. Hammoudi, M. Benaicha, Y. Benguerba, Etude de relation quantitative structure–Activité des composés chimiques à l'aide des descripteurs moléculaires. (Modélisation QSAR), le premier séminaire de la chimie appliqué et la modélisation moléculaire, 26 septembre 2019 Guelma.

List of Acronyms

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

CV: Cross-Validation

CD: Cluster Of differentiation

c-Flip: Cellular FLICE-like Inhibitory Protein

c-IAP: Cellular inhibitors of apoptosis

DNP: Double Numerical with Polarization

DNA: Deoxyribonucleic acid

DFT: Density functional theory

GGA: Generalized gradient approximation

HOMO: The highest occupied molecular orbitals

HTS: high throughput screening

HBD and HBA: Number of Hydrogen-Bond Donors and Acceptors

IC50: Half maximal Inhibitory Concentration

IκB: Inhibitory of kappa B

IKK: Inhibitor of kappa B kinase

IL: Interleukin

IL1R : Interleukin 1 receptor

JNK: c-Jun N-terminal kinases

LBDD: Ligand-based drug design

LOO: Leave One Out

LUMO: Lowest Unoccupied Molecular Orbital

MLR: Multiple Linear Regressions

MR: Molar Refractivity

MM: Molecular Mechanics

MRP: Multidrug Resistance Protein

MW: Molecular Weight

NSAIDs: nonsteroidal anti-inflammatory drugs

NF-kB: The nuclear factor kB

NIK: NF-κB–inducing kinase

PDB: Protein Data Bank

LIST OF ACRONYMS

PLS: Partial least squares regression

PRESS: Predictive Residual Sum of the Squares

PSA: Polar Surface Area

QSAR: Quantitative structure-activity relationship

QSPR: Quantitative property-activity relationship

QM: Quantum mechanics

RB: Number of Rotatable Bonds

RMS: Root-Mean Squared

RNA: Ribonucleic acid

SAR: Structure–Activity Relationships

SBDD: Structure-Based Drug Design

SCID: Severe Combined Immune Deficiency

SVM: Support Vector Machines

TI: topological index

TNF: Tumor necrosis factor

VS: virtual screening

List of Figures

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV (part 1)

LIST OF FIGURES

CHAPTER IV (part 2)

List of Tables

CHAPTER 1

CHAPTER II

CHAPTER III

CHAPTER IV (part 1)

CHAPTER IV (part 2)

General Introduction

GENERAL INTRODUCTION

1. General Introduction

Nuclear factor -kB transcription factor is a ubiquitous and well-characterized protein which has crucial roles in the regulation of cell growth processes, immune, apoptosis [1] as well as in inflammatory responses. It is involved in controlling cell signalling in the body under specific pathological and physiological conditions.

NF-κB controls the expression of genes encoding the pro-inflammatory cytokines (e. g., TNF-α, IL-1, IL- 2, IL-6, etc.), inducible enzymes (COX-2 and iNOS), adhesion molecules (e. g., VCAM, ICAM, E-selectin), chemokines (e. g., MIP-1α,IL-8, RANTES, MCP1, eotaxin, etc.), some of the acute phase proteins, growth factors, and immune receptors, which are included in critical roles in controlling many inflammatory processes [2]. NF-kB Activation is engaged by the IkB kinase (IKK) complex, which comprises a regulatory subunit and two catalytic subunits IKK-α and IKK-β [3].

Blocking the activation NF-κB represents an important and very attractive therapeutic target for the treatment of inflammatory diseases and cancer [4]. Therefore, the inhibition of IKK-β enzyme would be a promising target for several inflammatory diseases and cancers treatment [5]. Because of poor drug selectivity and severe adverse drug events of the marketed molecules, Mesalazine and Sulfasalazine [6-8], much focus was put on the development of new IKK-β inhibitors drugs.

The drug discovery process can take an average of 10 to 15 years [9.10]. Because of the varied physicochemical features of chemical compounds and the challenge of scaling up manufacturing, identifying a medicine and further developing it takes significant investment such as, human resources, and technological know-how [11.12], which also necessitates careful adherence to testing and production norms [13].

Efficacy or toxicity due to attrition were the main reasons that led to failing this biggest challenge and extravagant cost [14]. For a long time, in vitro pharmacological profiling played a crucial role in determining several undesirable off-target activity profiles, which would impede the discovery of new candidate drugs [14.15]. However, these methods did not reduce the time or cost needed to bring up a candidate drug to market, as the development costs now exceed $US\$2.8\ billion$ [16], Also clinical failures due to idiosyncratic toxicity, still an obstacle despite the existence of these technologies [17].

A variety of environmental issues are putting pressure on pharmaceutical companies, including significant revenue losses as a result of patent expirations, increasingly cost-constrained healthcare systems, and more stringent regulatory requirements [18].

In the past decades, as a suggested solution to possibly improve productivity, modern drug discovery involved the use of numerous technologies at various stages in order to enhance the pharmacological characteristics of lead compounds at varying degrees of complexity. These methods are based on the combination of traditional strategies in medicinal chemistry with computer-aided drug design (CADD) methods [19].

The employment of CADD techniques has become an essential tool by researchers and pharmaceutical industries for the preliminary stage of drug development in order to reduce failures in the final stage and to accelerate the process of discovery in a more cost-effective manner [20]. CAAD methods are classified into: structure-based drug design (SBDD) and ligand-based drug design (LBDD) approaches [21]. Which aim to develop therapeutically interesting small molecules. While through them, large compound libraries can be filtered and reduced into a small set of active compounds that can be synthesized and tested experimentally [22].

SBDD: consists of utilizing the 3D molecular structure of a specific target, such as a receptor or enzyme (macromolecule), to generate and screen potential candidate drugs for further synthesis and experimental tests based on their predicted interactions with the binding site of protein [20]. Examples of these methods: The Molecular Docking and De novo design [23].

In contrast, LBDD: consists of subjecting a series of compounds with a variety of structures and well-known effectiveness to computational modelling techniques used to construct theoretical predictive models. Pharmacophore Modeling and Quantitative Structure-Activity/Property relationships (QSAR/QSPR) are the common methods used in this technique [20].

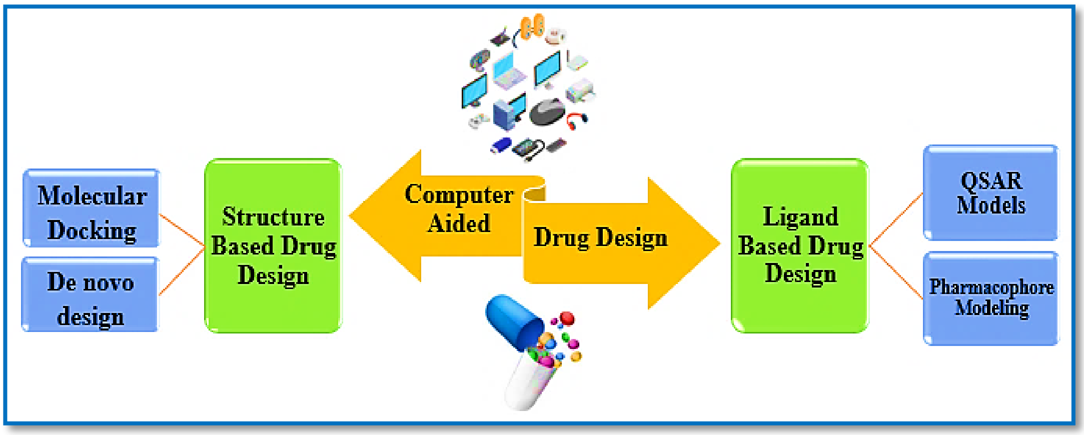


Figure 1: Computer-Aided Drug Design Methods

Contributions

NF-κB inhibition offers a promising strategy in cancer therapy, therefore, most of these inhibitors are targeting the selective inhibition IKK subunits and the individual NF-κB.

The objective of this work is twofold:

First;

- To develop new active compounds inhibitors of IKK-β with a high biological effect and minimum inhibitory concentration with better affinity to IKK-β, for the NF-κB inhibition and to study the influence of the structure on the biological activity.

- To achieve this goal, QSAR model was elaborated using Multiple linear regression model, (helpful for significant descriptors interpretation) involving thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives.

-This model allows to predict the pIC50 values of new series.

- Design new active compounds with higher biological activity

-Then a molecular docking study was carried out in order to determine the interaction between the compound and the receptor site of IKK-β protein and to predict the binding poses of the investigated compounds.

-The drug-like properties of the compounds were predicted and finally, in-silico-toxicity studies were performed to predict the toxicity of the new designed compounds.

-Note that we have decided to perform analysis taking into account only one single scaffold in order to assure a better homogeneity of the results avoiding some dispersion that may arise from considering different structures. Even if this approach limit somehow the generality and the space spanned by the hypothesized inhibitors it allows to increase the precision and the robustness of the interactions patterns that have been previewed.

Second;

To construct new robust QSAR models based on Multiple Linear Regression and Artificial Neuron Network for nuclear factor-κB (NF-κB) inhibitors prediction. Using 121 compounds. Then, the models have been assessed and evaluated. These models have been elaborated involving another series of compounds, because the ANN models require a large number of data set

Thesis Organization

To achieve the desired goal, we have organized our thesis into four chapters:

- ❖ The first chapter is devoted to general information on NF-κB (nuclear factor-kappa B) and its role in cancer diseases.

- In the second chapter we present the QSAR methods, their principle, methodology, advantages and applications.
- The third chapter entitled “Molecular Docking Study” presents the Docking method, its principle, its application.
- The fourth chapter is devoted to the methodology of the work carried out in this thesis, to the results obtained and to their discussions.

References

[1] Wang, J. L., Li, L., Hu, M. B., Wu, B., Fan, W. X., Peng, W., ... & Wu, C. J. (2019). In silico drug design of inhibitor of nuclear factor kappa B kinase subunit beta inhibitors from 2-acylamino-3-aminothienopyridines based on quantitative structure–activity relationships and molecular docking. *Computational Biology and Chemistry*, *78*, 297-305.

[2] Nam, N. H. (2006). Naturally occurring NF-κB inhibitors. *Mini reviews in medicinal chemistry*, *6*(8), 945-951.

[3] Borgatti, M., Chilin, A., Piccagli, L., Lampronti, I., Bianchi, N., Mancini, I., ... & Gambari, R. (2011). Development of a novel furocoumarin derivative inhibiting NF-κB dependent biological functions: design, synthesis and biological effects. *European Journal of Medicinal Chemistry*, *46*(10), 4870-4877.

[4] Leung, C. H., Grill, S. P., Lam, W., Han, Q. B., Sun, H. D., & Cheng, Y. C. (2005). Novel mechanism of inhibition of nuclear factor-κB DNA-binding activity by diterpenoids isolated from Isodon rubescens. *Molecular pharmacology*, *68*(2), 286-297.

[5] Lauria, A., Ippolito, M., Fazzari, M., Tutone, M., Di Blasi, F., Mingoia, F., & Almerico, A. M. (2010). IKK-β inhibitors: an analysis of drug–receptor interaction by using molecular docking and pharmacophore 3D-QSAR approaches. *Journal of Molecular Graphics and Modelling*, *29*(1), 72-81.

[6] Weber, C. K., Liptay, S., Wirth, T., Adler, G., & Schmid, R. M. (2000). Suppression of NF-κB activity by sulfasalazine is mediated by direct inhibition of IκB kinases α and β. *Gastroenterology*, *119*(5), 1209-1218.

[7] Venardos, K., Harrison, G., Headrick, J., & Perkins, A. (2004). Auranofin increases apoptosis and ischaemia–reperfusion injury in the rat isolated heart. *Clinical and experimental pharmacology and physiology*, *31*(5-6), 289-294.

[8] Mali, S. N., Pandey, A., Thorat, B. R., & Lai, C. H. (2022). Multiple 3D-and 2D-quantitative structure–activity relationship models (QSAR), theoretical study and molecular modeling to identify structural requirements of imidazopyridine analogues as anti-infective agents against tuberculosis. *Structural Chemistry*, *33*(3), 679-694.

[9] Parvathaneni, V., Kulkarni, N. S., Muth, A., & Gupta, V. (2019). Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug discovery today*, *24*(10), 2076-2085.

[10] Hammoudi, N.H., Sobhi, W., Attoui, A., Lemaoui, T., Erto, A., & Benguerba, Y. (2021). In silico drug discovery of Acetylcholinesterase and Butyrylcholinesterase enzymes inhibitors based on Quantitative Structure-Activity Relationship (QSAR) and drug-likeness evaluation. *Journal of Molecular Structure*, *1229*, 129845.

[11] Kulkarni, N. S., Parvathaneni, V., Shukla, S. K., Barasa, L., Perron, J. C., Yoganathan, S., ... & Gupta, V. (2019). Tyrosine kinase inhibitor conjugated quantum dots for non-small cell lung cancer (NSCLC) treatment. *European Journal of Pharmaceutical Sciences*, *133*, 145-159.

[12] Vaidya, B., Parvathaneni, V., Kulkarni, N. S., Shukla, S. K., Damon, J. K., Sarode, A., ... & Gupta, V. (2019). Cyclodextrin modified erlotinib loaded PLGA nanoparticles for improved therapeutic efficacy against non-small cell lung cancer. *International journal of biological macromolecules*, *122*, 338-347

[13] Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., & Tylkowski, B. (2022). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(2), e1568.

[14] Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Anthony, J.H & Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature materials*, *18*(5), 435-441.

[15] Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature reviews Drug discovery*, *11*(12), 909-922.

[16] DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics*, *47*, 20-33.

[17] Kenna, J. G. (2017). Human biology-based drug safety evaluation: scientific rationale, current status and future challenges. *Expert Opinion on Drug Metabolism & Toxicology*, *13*(5), 567-574.

REFERENCES

[18] Marrucho, I. M., Branco, L. C., & Rebelo, L. P. N. (2014). Ionic liquids in pharmaceutical applications. *Annual review of chemical and biomolecular engineering*, *5*, 527-546.

[19] VC Guido, R., Oliva, G., & D Andricopulo, A. (2011). Modern drug discovery technologies: opportunities and challenges in lead discovery. *Combinatorial chemistry & high throughput screening*, *14*(10), 830-839.

[20] Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of pharmacal research*, *38*(9), 1686-1701.

[21] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological reviews*, *66*(1), 334-395.

[22] Lambrinidis, G., & Tsantili-Kakoulidou, A. (2018). Challenges with multi-objective QSAR in drug discovery. *Expert Opinion on Drug Discovery*, *13*(9), 851-859.

CHAPTER I: The Nuclear Factor Kb (Nf-Kb) and its Role in Cancer

CHAPTER I: THE NUCLEAR FACTOR KB (NF-KB) AND ITS ROLE IN CANCER

I.1. Introduction

The nuclear factor kB (NF-kB) comprises a family of transcription factors involved in the regulation of a wide variety of biological responses. NF-kB is well-known for its involvement in the regulation of immune responses and inflammation, but new data suggests that it also plays a role in oncogenesis. NF-kB controls the expression of genes involved in a variety of processes important in the genesis and progression of cancer, including proliferation, migration, and apoptosis. Many human cancers have been shown to have abnormal or constitutive NF-kB activity. Numerous research has been conducted in recent years to elucidate the functional effects of NF-kB activation as well as it signaling pathways. NF-kB has emerged as an intriguing therapeutic target for cancer therapy [1].

NF-kB is a pleiotropic transcription factor regulating over 200 genes involved in cell function and inflammation [2]. The Rel or NF-kB family is comprised of hetero- or homo-dimeric combinations of five members, NF-kB 1 (p50 and its precursor p105), NF-kB 2 (p52 and its precursor p100), RelA (p65), RelB and c-Rel. In resting cells, NF-kB dimers are inactive because they are sequestered in the cytoplasm by inhibitory proteins. These inhibitory proteins include the inhibitory kB (α, β or γ), as well as the inactive precursors p100 and p105. NF-kB is rapidly activated in response to a wide variety of stimuli through either the traditional (canonical) pathways or the alternative (non-canonical) pathways (Figure I.1).

The classical pathway is activated by cytokines (IL-1, TNF) or bacterial products (lipopolysaccharide, LPS), which activate the inhibitory kB kinases (IKK). Phosphorylation of serine residues 32 and 36 on IB by IKKβ leads to poly-ubiquitination of the protein, which in turn triggers its destruction by the 26S proteasome. Because of this liberation of NF-kB, also known as p50: p65, from its inhibitor IB, NF-kB is now able to translocate to the nucleus. CD40 and lymphotoxin are the molecules that initiate the alternative pathway, and they do so by activating IKK α via NF-kB inducing kinase (NIK). Phosphorylation of p100 by IKKα promotes partial degradation of p100 by the proteasome to release active RelB: p52 dimers, which translocate to the nucleus to modulate gene expression.

All five NF-kB subunits contain a conserved Rel homology domain responsible for nuclear localization and DNA binding. However, only three of these subunits, p65, RelB and c-Rel contain trans-activation domains and directly promote gene transcription. Therefore p507 and p52 homodimers do not directly stimulate gene transcription. Because each subunit has distinct biological functions, various dimeric combinations of NF-kB proteins have varying impacts on cell destiny and function [3].

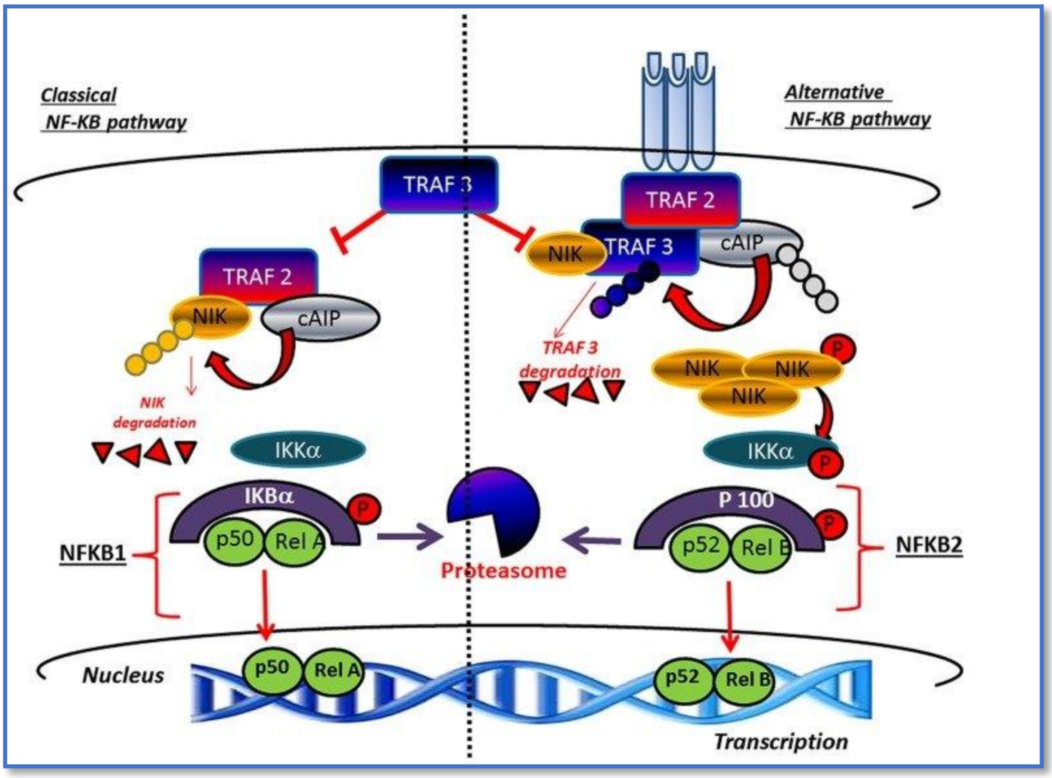


Figure I.1: Canonical and non-canonical pathways leading to the activation of NF-κB

I.2. NF-kB In Human Disease

NF-kB activation has been implicated in several human diseases– Cancer [5], AIDS (HIV-1) [6], Ageing [7], Headaches [8], Catabolic Disorders [9], Diabetes Type 1 [10], Diabetes Type 2 [11], Atherosclerosis [12], Heart Disease [13], Ischemia/Reperfusion [14], Pulmonary Disease [15], Chronic Obstructive Pulmonary Disease (COPD) [16], Renal Disease [17], Gut Diseases [18], Skin Diseases [19], Asthma [20], Arthritis [21], Crohn´s Disease [22], Ocular Allergy [23], Pancreatitis [24], Periodonitis [25], Inflamatory Bowel Disease [26],Sepsis [27], Sleep Apnoea [28], Autoimmunity [29], and Neuropathological Diseases [46.47], etc.

I.3. Role of NF-κB in cancer

I.3.1. Apoptosis

Apoptosis is necessary for the typical growth and upkeep of multicellular organisms since it enables the death and removal of individual cells without causing any damage to the organism as a whole. An intracellular protease cascade is activated during the process of apoptosis. This cleaves a large number of intracellular proteins, which ultimately results in the blebbing of the membrane, nuclear condensation, and fragmentation of the DNA [30,31]. Inhibitors of the proteases known as caspases have the ability to stop the apoptotic program from taking place. Apoptosis can be triggered by a wide variety of cellular treatments, such as DNA-damaging agents (alkylating agents, UV light, topoisomerase inhibitors, and so on), hormones, growth-factor deprivation, cellular stress (heat shock, reactive oxygen), and receptor-mediated ligands (TNF and Fasil, for example) [32].

The biological function of apoptosis that is triggered by TNF is not completely understood. A relationship between apoptosis and biological consequence has not been demonstrated despite the fact that TNF has been linked to a wide variety of physiological processes, ranging from the prevention of viral infection to the differentiation of adipocytes [33.34]. Because TNF signaling has a pleiotropic nature, it is difficult to pinpoint any one outcome as being responsible for the activation of apoptosis. It has been discovered that there are a number of signaling proteins that are located downstream of TNF-Rs. It is possible that the characterization of these proteins will shed light on the role of TNF-induced apoptosis [30].

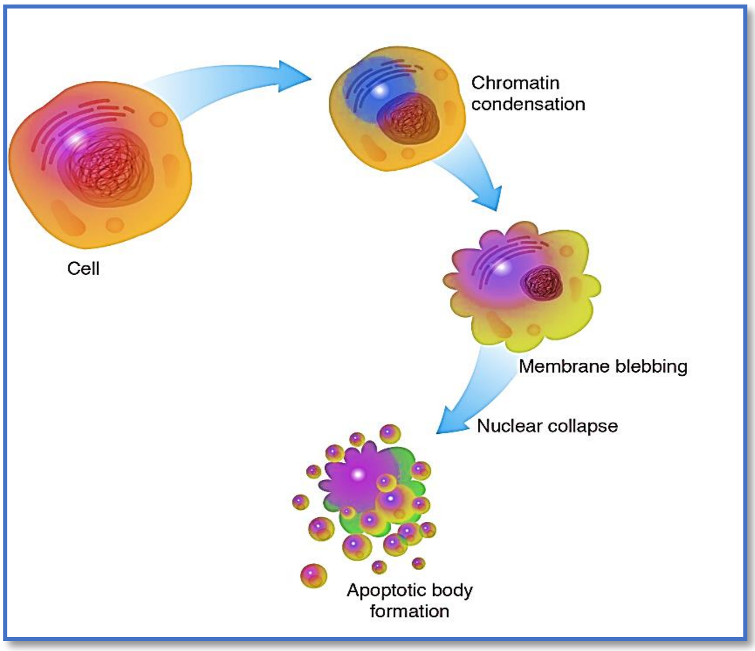


Figure I.2: Mechanism of Cell Death (Apoptosis) [35]

It is interesting to note that substances that block protein synthesis play a significant role in this process. This suggests a mechanism in which newly expressed gene products suppress the apoptosis signal. Recent research has demonstrated that one of the factors that mediate this type of gene expression is NF-kB [30].

I.3.2. Activation of NF-kB by IKK

The transcription factor NF-kB can be turned on by a wide variety of signals, the majority of which are believed to be associated with cellular stress. NF-kB is regulated by its position in the cell; when active, it is in the nucleus; when inactive, it is in the cytoplasm. An inhibitor protein family known as IKBs regulates NF-kB localization. IKB is the IKBα that regulates the immediate-early activation of NF-kB [36].

IκB kinase (IKK) is a multi-subunit protein kinase that consists of two highly homologous catalytic subunits, IKKα and IKKβ, which play a crucial role in IκB phosphorylation; it is also composed of a nonenzymatic regulatory subunit, IKKγ, which is essential for the activation of IKKα/IKKβ heterodimers in response to proinflammatory cytokines, such as TNF-α and IL-1. IBs are targeted for fast polyubiquitination and subsequent destruction by the 26S proteasome after being phosphorylated by the IKK complex in their N-terminal regulatory domain at two key serine residues (Ser32 and Ser36 in IKα, Ser19 and Ser23 in IB) [37].

The liberated NF-kB dimers go to the nucleus, where they are subject to further regulation via phosphorylation, acetylation, and interactions with coactivators and corepressors. NF-kB is responsible for regulating the transcription of a wide variety of genes, including those that code for cytokines, growth factors, cell adhesion molecules, and pro-/antiapoptotic proteins. Activated NF-kB may then be downregulated by many processes, one of which is the well-characterized feedback pathway, in which the newly generated IκBα protein attaches to nuclear NF-kB and exports it to the cytoplasm. This can be accomplished through a number of other methods as well [38.39.40].

I.3.3.NF-kB and The Regulation of Apoptosis

The transcription factor NF-kB is critical to the process of controlling apoptosis. It does this by inhibiting the expression or activity of proapoptotic proteins while simultaneously stimulating the production of various anti-apoptotic proteins, which causes apoptosis to be inhibited. An examination of RelA-/- mutant mice, who are doomed to die between 15 and 16 days into their pregnancies as a result of excessive hepatic apoptosis, provided the first evidence that NF-kB had an inhibiting influence on apoptosis. In addition to this, the embryonic mortality that is shown in RelA-/- mice may also be seen in mice that are lacking in IKK or IKKγ [40.41.42].

It was shown that TNF- α signaling in the developing liver was the insult that led to liver apoptosis. This was demonstrated by the fact that is crossing RelA-/- mice with either TNF- α /- or TNFR-/- animals was able to recover this liver phenotype.

Although TNF-α is responsible for the activation of caspases, which are responsible for the initiation and execution of apoptosis, the pro-apoptotic activity of TNF- can be inhibited by the concurrent activation of NF-kB. This is because NF-kB targets genes that code for inhibitors of caspase activation and apoptosis, which negates the effect of TNF-α.

The anti-apoptotic proteins regulated by NF-kB include inhibitor of apoptosis proteins (IAPs) 1 and 2, X-linked IAP, cellular Fas-associated death domain-like IL-1β-converting enzyme (FLICE) inhibitory protein (cFLIP), Bcl-XL, A1 (also known as Bfl-1), TNF receptor-associated factor 1 (TRAF1) and TRAF2 .IAPs (c-IAP1, c-IAP2, and XIAP) block effector caspases (caspases-3, -6, and 7) in a direct manner, which allows them to decrease apoptosis caused by both extrinsic and intrinsic routes [43.44].

Although c-FLIP and procaspase 8 have a high amount of similarity, c-FLIP does not have any catalytic activity. After being induced, c-FLIP will connect with TNFR to prevent caspase 8 activation by acting as a competitor to it [45]. Members of the Bcl-2 family that are anti-apoptotic (Bcl-XL, A1) are able to stop the release of cytochrome c and the consequent activation of caspase 9. TRAF proteins have the ability to both enhance the activation of NF-kB as well as interfere with the caspase cascade at the level of TNFR1 [46]. The capacity of NF-kB to limit prolonged c-Jun N-terminal kinases (JNK) activation and build-up of reactive oxygen species (ROS) is yet another way in which this transcription factor may prevent apoptosis from occurring [47].

I.4. NF-kB and Cancer Therapy

Several research have been conducted to identify potential NF-kB inhibitors as cancer treatment agents. Since NF-kB activation is the result of a multi-step signaling pathway, these compounds may target different points of the signaling process. Some anti-inflammatory medicines, for example, may suppress NF-kB by interfering with IKK function. [48.49.50]. Curcumin, trans-resveratrol, and parthenolide are a few examples of other naturally occurring chemicals that have been shown to suppress IKK activity. Other natural compounds that inhibit IKK activity include: [51.52.53].

Inhibition of NF-kB can also be accomplished by directing one's attention toward the proteasome breakdown process. Proteasome inhibitors prevent NF-kB activation by blocking the degradation of IkBs, NF-kB1/p105 or NF-kB2/p100 [1].

I.4.1. Current strategies for inhibiting NF-κB

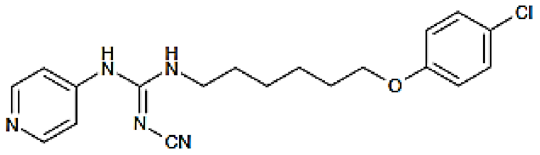
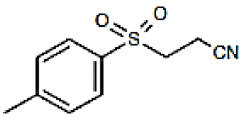
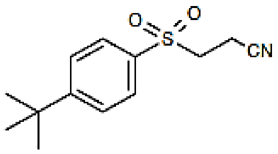
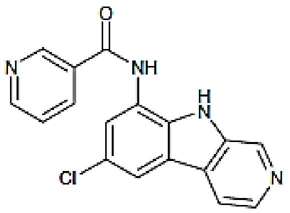
I.4.1.1. Selective IKK Inhibitors

Targeting the activity of IKK might be the method with the greatest potential for achieving a selective reduction in NF-kB activation; Recent efforts in the field of pharmaceutical research have been made to generate some very effective and selective inhibitors of catalytic activity IKKβ and/or IKKα (Table I.1). Although several agents initially identified as IKKβ inhibitors can inhibit IKKα catalytic activity in the low µmolar range, no potent IKKα-specific inhibitors have been described so far.

Although some compounds previously identified as IKKβ inhibitors can decrease IKKα catalytic activity at low molar concentrations, no effective IKKα specific inhibitors have been reported.

The particular part that IKKα plays in the activation of the alternative pathway, which is essential for the B-cell-mediated responses, the organization of lymphoid organs, and the development of the mammary gland [54], evidence suggests that IKK might be a promising therapeutic strategy for treating cancer and autoimmune diseases [55.56].

Table I.1: Selective IκB kinase inhibitors.

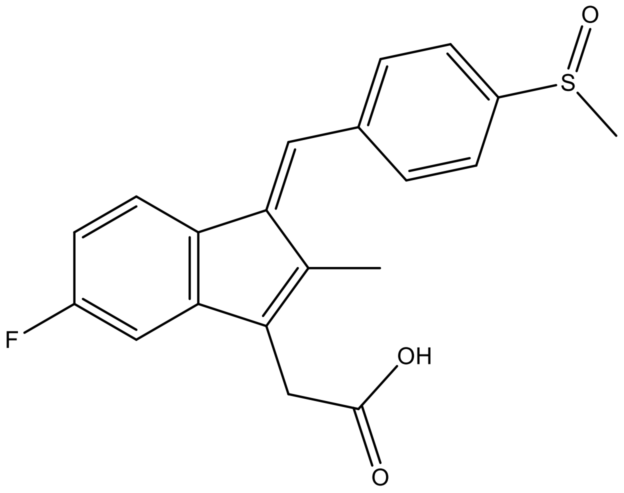
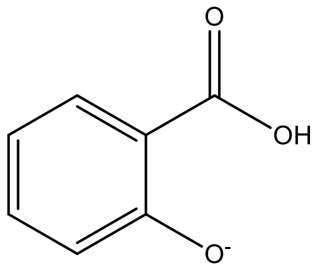
Name	Structure	Phase
CHS-828		I/II
BAY 11-7082		Preclinical
BAY 11-7085		Preclinical
PS-1145		Preclinical

I.4.1.2. NSAIDs

The suppression of cyclooxygenase activity, which results in the prevention of prostaglandin formation, is the explanation for the anti-inflammatory actions of nonsteroidal anti-inflammatory drugs (NSAIDs) that is the most widely accepted. The mechanism of traditional anti-inflammatory medicines has been re-evaluated, and it has been proven to be, at least partially, suppression of NF-kB activation.

This is due to the fact that NF-kB plays a vital role in the regulation of the inflammatory process. It has been shown that the actions of these drugs on the NF-kB pathway are not dependent on the inhibition of cyclooxygenase.
as shown by the fact that indomethacin, which is a powerful inhibitor of prostaglandin production, does not block the NF-kB pathway [57.58]. Several NSAIDs are capable of inhibiting NF-kB activation, including salicylates [59.60], sulindac and its analogs [61.62].

Table I.2: Chemical Structures of sulindac and salicylate

	
sulindac	salicylate

I.4.1.3. Proteasome inhibitors

An emerging theme for anticancer drug development is protein degradation. The multicatalytic ubiquitin–proteasome pathway is the principal mechanism responsible for the degradation of eukaryotic cellular proteins [63]. Recently, proteasome inhibitors that have shown promising anticancer responses both *in vitro* and *in vivo* have been introduced into human cancer treatment [64].

I.4.1.4. Other Approaches

It has been demonstrated that several different substances, some of which have anticancer and/or antiangiogenetic potential, also exhibit NF-kB -inhibiting action. Thalidomide, which is known as an immunomodulatory drug, has anticancer, anti-inflammatory, antiangiogenic and immunosuppressive effects. In order to explain the therapeutic effect of this compound, several distinct theories have been offered, one of which is the inhibition of NF-kB activation through the reduction of IKK activity [65].

Dehydroxymethylepoxyquinomicin (DHMEQ), an epoxyquinomicin C derivative originally isolated as a weak antibiotic and anti-inflammatory agent, was found to inhibit activation of NF-kB at the level of nuclear translocation [66]. It is believed that DHMEQ is a potential anticancer drug in the treatment of hormone-refractory prostate cancer [58].

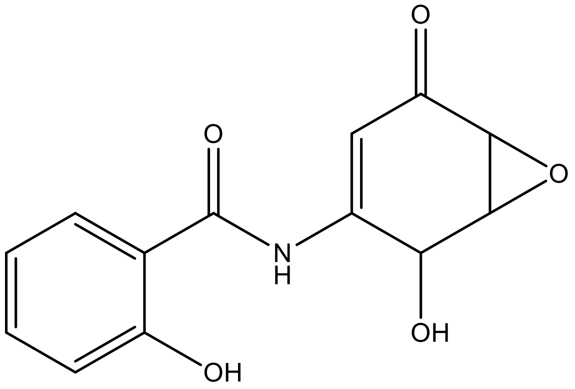


Figure I.3: Chemical structure of Dehydroxymethylepoxyquinomicin

REFERENCES

References

[1] Dolcet, X., Llobet, D., Pallares, J., & Matias-Guiu, X. (2005). NF-kB in development and progression of human cancer. *Virchows archiv*, *446*(5), 475-482.

[2] Pereira, S. G., & Oakley, F. (2008). Nuclear factor-κB1: regulation and function. *The international journal of biochemistry & cell biology*, *40*(8), 1425-1430.

[3] Hayden, M. S., & Ghosh, S. (2004). Signaling to NF-kB. Genes Dev. *Ghosh S. Signaling to NF-kB. Genes Dev*, *18*, 2195-2224.

[4] Cornelius, C., Koverech, G., Crupi, R., Di Paola, R., Koverech, A., Lodato, F., Scuto, M.,Salinaro, A.,cuzzocrea, S.,& Calabrese, V. (2014). Osteoporosis and Alzheimer pathology: role of cellular stress response and hormetic redox signaling in aging and bone remodeling. *Frontiers in Pharmacology*, *5*, 120.

[5] Tafani, M., Pucci, B., Russo, A., Schito, L., Pellegrini, L., Perrone, G. A., ... & Russo, M. A. (2013). Modulators of HIF1α and NFkB in cancer treatment: is it a rational approach for controlling malignant progression?. *Frontiers in Pharmacology*, *4*, 13.

[6] Mihm, S., Ennen, J., Pessara, U., Kurth, R., & Droge, W. (1991). Inhibition of HIV-1 replication and NF-kB activity by cysteine and cysteine derivatives. *Aids*, *5*(5), 497-503.

[7] Kanigur Sultuybek, G., Soydas, T., & Yenmis, G. (2019). NF-κB as the mediator of metformin's effect on ageing and ageing-related diseases. *Clinical and Experimental Pharmacology and Physiology*, *46*(5), 413-422.

[8] Soveyd, N., Abdolahi, M., Bitarafan, S., Tafakhori, A., Sarraf, P., Togha, M., ... & Honarvar, N. M. (2017). Molecular mechanisms of omega-3 fatty acids in the migraine headache. *Iranian journal of neurology*, *16*(4), 210.

[9] Wuertz, K., Vo, N., Kletsas, D., & Boos, N. (2012). Inflammatory and catabolic signalling in intervertebral discs: the roles of NF-κB and MAP kinases. *European Cells and Materials*, *23*, 103-119.

[10] Hofmann, M. A., Schiekofer, S., Kanitz, M., Klevesath, M. S., Joswig, M., Lee, V., ... & Nawroth, P. P. (1998). Insufficient glycemic control increases nuclear factor-κB binding activity in peripheral blood mononuclear cells isolated from patients with type 1 diabetes. *Diabetes Care*, *21*(8), 1310-1316.

[11] Coto, E., Díaz-Corte, C., Tranche, S., Gómez, J., Alonso, B., Iglesias, S., ... & Coto-Segura, P. (2018). Gene variants in the NF-KB pathway (NFKB1, NFKBIA, NFKBIZ) and their association with type 2 diabetes and impaired renal function. *Human immunology*, *79*(6), 494-498.

[12] Pan, J. X. (2017). LncRNA H19 promotes atherosclerosis by regulating MAPK and NF-kB signaling pathway. *Eur Rev Med Pharmacol Sci*, *21*(2), 322-328.

[**13]** Xiao, F., Zheng, R., Yang, D., Cao, K., Zhang, S., Wu, B., ... & Zhou, B. (2017). Sex-dependent aortic valve pathology in patients with rheumatic heart disease. *PLoS One*, *12*(6), e0180230.

[14] Kim, J. W., Jin, Y. C., Kim, Y. M., Rhie, S., Kim, H. J., Seo, H. G., ... & Chang, K. C. (2009). Daidzein administration in vivo reduces myocardial injury in a rat ischemia/reperfusion model by inhibiting NF-kB activation. *Life sciences*, *84*(7-8), 227-234.

[15] Cui, Y., Xin, H., Tao, Y., Mei, L., & Wang, Z. (2021). Arenaria kansuensis attenuates pulmonary fibrosis in mice via the activation of Nrf2 pathway and the inhibition of NF-kB/TGF-beta1/Smad2/3 pathway. *Phytotherapy Research*, *35*(2), 974-986.

[16] Cappello, F., Caramori, G., Campanella, C., Vicari, C., Gnemmi, I., Zanini, A., ... & Di Stefano, A. (2011). Convergent sets of data from in vivo and in vitro methods point to an active role of Hsp60 in chronic obstructive pulmonary disease pathogenesis. *PLoS One*, *6*(11), e28200.

[17] Kumar, D., Singla, S. K., Puri, V., & Puri, S. (2015). The restrained expression of NF-kB in renal tissue ameliorates folic acid induced acute kidney injury in mice. *PloS one*, *10*(1), e115947.

[18] Fantini, M. C., & Pallone, F. (2008). Cytokines: from gut inflammation to colorectal cancer. *Current drug targets*, *9*(5), 375-380.

REFERENCES

[19] Herron, B. J., Rao, C., Liu, S., Laprade, L., Richardson, J. A., Olivieri, E., ... & Beier, D. R. (2005). A mutation in NFkB interacting protein 1 results in cardiomyopathy and abnormal skin development in wa3 mice. *Human molecular genetics*, *14*(5), 667-677.

[20] Levine, S. J. (2003). NF-kB: A KEY SIGNALING PATHWAY IN ASTHMA. *Signal transduction and human disease*, 23.

[21] Barrow, M. (2021). An Overview of the NF-kB mechanism of pathophysiology in rheumatoid arthritis, investigation of the NF-kB ligand RANKL and related nutritional interventions. *Autoimmunity Reviews*, *20*(2), 102741.

[22] Maeda, S., Hsu, L. C., Liu, H., Bankston, L. A., Iimura, M., Kagnoff, M. F., ... & Karin, M. (2005). Nod2 mutation in Crohn's disease potentiates NF-κB activity and IL-1ß processing. *Science*, *307*(5710), 734-738.

[23] Leonardi, A., Modugno, R. L., & Salami, E. (2021). Allergy and dry eye disease. *Ocular immunology and inflammation*, *29*(6), 1168-1176.

[24] Kim, H., Seo, J. Y., & Kim, K. H. (2000). NF-kB and cytokines in pancreatic acinar cells. *Journal of Korean medical science*, *15*(Suppl), S53-S54.

[25] Arabaci, T., Cicek, Y., Canakci, V., Canakci, C. F., Ozgoz, M., Albayrak, M., & Keles, O. N. (2010). Immunohistochemical and stereologic analysis of NF-κB activation in chronic periodontitis. *European journal of dentistry*, *4*(4), 454.

[26] Bank, S., Julsgaard, M., Abed, O. K., Burisch, J., Broder Brodersen, J., Pedersen, N. K., ... & Sode, J. (2019). Polymorphisms in the NF kB, TNF-alpha, IL-1beta, and IL-18 pathways are associated with response to anti-TNF therapy in Danish patients with inflammatory bowel disease. *Alimentary pharmacology & therapeutics*, *49*(7), 890-903.

[27] Filgueiras, Jr, L. R., Martins, J. O., Serezani, C. H., Capelozzi, V. L., Montes, M. B., & Jancar, S. (2012). Sepsis-induced acute lung injury (ALI) is milder in diabetic rats and correlates with impaired NFkB activation.

[28] Israel, L. P., Benharoch, D., Gopas, J., & Goldbart, A. D. (2013). A pro-inflammatory role for nuclear factor kappa B in childhood obstructive sleep apnea syndrome. *Sleep*, *36*(12), 1947-1955.

[29] Giuliani, C., Napolitano, G., Bucci, I., Montani, V., & Monaco, F. (2001). Nf-kB transcription factor: role in the pathogenesis of inflammatory, autoimmune, and neoplastic diseases and therapy implications. *La Clinica Terapeutica*, *152*(4), 249-253.

[30] Serasanambati, M., & Chilakapati, S. R. (2016). Function of nuclear factor kappa B (NF-kB) in human diseases-a review. *South Indian Journal of Biological Sciences*, *2*(4), 368-87.

[31] Van Antwerp, D. J., Martin, S. J., Verma, I. M., & Green, D. R. (1998). Inhibition of TNF-induced apoptosis by NF-κB. *Trends in cell biology*, *8*(3), 107-111.

[32] Martin, S. J., & Green, D. R. (1995). Protease activation during apoptosis: death by a thousand cuts?. *Cell*, *82*(3), 349-352

[33] Alnemri, E. S., Livingston, D. J., Nicholson, D. W., Salvesen, G., Thornberry, N. A., Wong, W. W., & Yuan, J. (1996). Human ICE/CED-3 protease nomenclature. *Cell*, *87*(2), 171.

[34] Green, D. R., & Martin, S. J. (1995). The killer and the executioner: how apoptosis controls malignancy. *Current opinion in immunology*, *7*(5), 694-703.

[35] https://www.genome.gov/genetics-glossary/apoptosis

[36] Hatada, E. N., Nieters, A., Wulczyn, F. G., Naumann, M., Meyer, R., Nucifora, G., ... & Scheidereit, C. (1992). The ankyrin repeat domains of the NF-kappa B precursor p105 and the protooncogene bcl-3 act as specific inhibitors of NF-kappa B DNA binding. *Proceedings of the National Academy of Sciences*, *89*(6), 2489-2493.

[37] Senftleben, U., & Karin, M. (2002). The Ikk/nf-κb pathway. *Critical care medicine*, *30*(1), S18-S26.

[38] Ghosh, S., & Karin, M. (2002). Missing pieces in the NF-κB puzzle. *cell*, *109*(2), S81-S96.

REFERENCES

[39] Karin, M., & Lin, A. (2002). NF-κB at the crossroads of life and death. *Nature immunology*, *3*(3), 221-227.

[40] Lee, C. H., Jeon, Y. T., Kim, S. H., & Song, Y. S. (2007). NF-κB as a potential molecular target for cancer therapy. *Biofactors*, *29*(1), 19-35.

[41] Li, Z. W., Chu, W., Hu, Y., Delhase, M., Deerinck, T., Ellisman, M., ... & Karin, M. (1999). The IKKβ subunit of IκB kinase (IKK) is essential for nuclear factor κB activation and prevention of apoptosis. *The Journal of experimental medicine*, *189*(11), 1839-1845.

[42] Rudolph, D., Yeh, W. C., Wakeham, A., Rudolph, B., Nallainathan, D., Potter, J., ... & Mak, T. W. (2000). Severe liver degeneration and lack of NF-κB activation in NEMO/IKKγ-deficient mice. *Genes & development*, *14*(7), 854-862.

[43] Deveraux, Q. L., Roy, N., Stennicke, H. R., Van Arsdale, T., Zhou, Q., Srinivasula, S. M., ... & Reed, J. C. (1998). IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases. *The EMBO journal*, *17*(8), 2215-2223..

[44] Deveraux, Q. L., Takahashi, R., Salvesen, G. S., & Reed, J. C. (1997). X-linked IAP is a direct inhibitor of cell-death proteases. *Nature*, *388*(6639), 300-304.

[45] Kreuz, S., Siegmund, D., Scheurich, P., & Wajant, H. (2001). NF-κB inducers upregulate cFLIP, a cycloheximide-sensitive inhibitor of death receptor signaling. *Molecular and cellular biology*, *21*(12), 3964-3973.

[46] Kucharczak, J., Simmons, M. J., Fan, Y., & Gelinas, C. (2003). To be, or not to be: NF-κB is the answer–role of Rel/NF-κB in the regulation of apoptosis. *Oncogene*, *22*(56), 8961-8982.

[47] Arch, R. H., Gedrich, R. W., & Thompson, C. B. (1998). Tumor necrosis factor receptor-associated factors (TRAFs)—a family of adapter proteins that regulates life and death. *Genes & development*, *12*(18), 2821-2830.

[48] Luo, J. L., Kamata, H., & Karin, M. (2005). IKK/NF-κB signaling: balancing life and death–a new approach to cancer therapy. *The Journal of clinical investigation*, *115*(10), 2625-2632..

[49] Yamamoto, Y., & Gaynor, R. B. (2001). Role of the NF-kB pathway in the pathogenesis of human disease states. *Current molecular medicine*, *1*(3), 287-296.

[50] Yin, M. J., Yamamoto, Y., & Gaynor, R. B. (1998). The anti-inflammatory agents aspirin and salicylate inhibit the activity of IκB kinase-β. *Nature*, *396*(6706), 77-80.

[51] Holmes-McNary, M., & Baldwin Jr, A. S. (2000). Chemopreventive properties of trans-resveratrol are associated with inhibition of activation of the IκB kinase. *Cancer Research*, *60*(13), 3477-3483.

[52] Keifer, J. A., Guttridge, D. C., Ashburner, B. P., & Albert Jr, S. (2001). Inhibition of NF-κB activity by thalidomide through suppression of IκB kinase activity. *Journal of Biological Chemistry*, *276*(25), 22382-22387.

[53] Plummer, S. M., Holloway, K. A., Manson, M. M., Munks, R. J., Kaptein, A., Farrow, S., & Howells, L. (1999). Inhibition of cyclo-oxygenase 2 expression in colon cells by the chemopreventive agent curcumin involves inhibition of NF-κB activation via the NIK/IKK signalling complex. *Oncogene*, *18*(44), 6013-6020.

[54] Ghosh, S., & Karin, M. (2002). Missing pieces in the NF-κB puzzle. *cell*, *109*(2), S81-S96.

[55] Karin, M., Yamamoto, Y., & Wang, Q. (2004). The IKK NF-κB system: a treasure trove for drug development. *Nature reviews Drug discovery*, *3*(1), 17-26.

[56] Melisi, D., & Chiao, P. J. (2007). NF-κB as a target for cancer therapy. *Expert opinion on therapeutic targets*, *11*(2), 133-144.

[57] Yin, M. J., Yamamoto, Y., & Gaynor, R. B. (1998). The anti-inflammatory agents aspirin and salicylate inhibit the activity of IκB kinase-β. *Nature*, *396*(6706), 77-80.

[58] Yamamoto, Y., Yin, M. J., Lin, K. M., & Gaynor, R. B. (1999). Sulindac inhibits activation of the NF-κB pathway. *Journal of Biological Chemistry*, *274*(38), 27307-27314.

[59] Kopp, E., & Ghosh, S. (1994). Inhibition of NF-κB by sodium salicylate and aspirin. *Science*, *265*(5174), 956-959.

REFERENCES

[60] Pierce, J. W., Read, M. A., Ding, H., Luscinskas, F. W., & Collins, T. (1996). Salicylates inhibit I kappa B-alpha phosphorylation, endothelial-leukocyte adhesion molecule expression, and neutrophil transmigration. *The Journal of Immunology*, *156*(10), 3961-3969.

[61] Berman, K. S., Verma, U. N., Harburg, G., Minna, J. D., Cobb, M. H., & Gaynor, R. B. (2002). Sulindac enhances tumor necrosis factor-α-mediated apoptosis of lung cancer cell lines by inhibition of nuclear factor-κB. *Clinical Cancer Research*, *8*(2), 354-360.

[62] Yasui, H., Adachi, M., & Imai, K. (2003). Combination of tumor necrosis factor-α with sulindac augments its apoptotic potential and suppresses tumor growth of human carcinoma cells in nude mice. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *97*(6), 1412-1420.

[63] Adams, J. (2004). The development of proteasome inhibitors as anticancer drugs. *Cancer cell*, *5*(5), 417-421.

[64] Rajkumar, S. V., Richardson, P. G., Hideshima, T., & Anderson, K. C. (2005). Proteasome inhibition as a novel therapeutic target in human cancer. *Journal of Clinical Oncology*, *23*(3), 630-639.

[65] Keifer, J. A., Guttridge, D. C., Ashburner, B. P., & Albert Jr, S. (2001). Inhibition of NF-κB activity by thalidomide through suppression of IκB kinase activity. *Journal of Biological Chemistry*, *276*(25), 22382-22387.

[66] Ariga, A., Namekawa, J. I., Matsumoto, N., Inoue, J. I., & Umezawa, K. (2002). Inhibition of tumor necrosis factor-α-induced nuclear translocation and activation of NF-κB by dehydroxymethylepoxyquinomicin. *Journal of Biological Chemistry*, *277*(27), 24625-24630.

CHAPTER II : Quantitative Structure Activity Relationship study

II.1. Introduction

QSAR model is a mathematic equation that correlates the biological activity of chemical compounds (dependent variable) with their chemical structure information (molecular descriptors) [1]. In recent decades, Artificial Intelligence, Machine learning, in particular, have become widely used to explore the chemical structure-activity relationship and to determine new chemical compounds that have a medical relevance [2]. The main objectives of this model are: to predict the biological activity of new compounds. Determination of significant parameters that control the biological effect [1] and improving biological activities of existing leads [3.4]. In addition to its effectiveness in drug discovery, it was also applied successfully in several fields, such as toxicity prediction of the aquatic toxicity of chemical compounds to environmental species [5], Toxicity of pesticides and dyes [6.7], and prediction of corrosion inhibitors effect [8].

These models originated way back in the 19^{th} century. In 1868, Crum-Brown and Fraser defined the fundamental principle that underpins formalism is that, physicochemical properties and biological activities of molecules strongly depend on their chemical structures [9]; however, they did not demonstrate the representation of chemical structure quantitatively. Then Richardson has suggested a function that correlates the chemical structures of compounds with solubility [10]. Since then, several successful predictive QSAR/QSPR models have been built and evaluated for the prediction of biological activities (anti-bacterial [11], anti-diabetes [12], anti-cancer [13] …. etc.) of therapeutical compounds against several targets, also for the prediction of Physico-chemical properties (density, Viscosity, Melting Point [14]) and toxicity.

Table II.1 represents the summary of the application of several developed QSAR models and the used statistical methods. The present paper covers the essential steps involved in QSAR model development and validation based on machine learning models. A case study on 121compounds as potent inhibitors of nuclear factor-κB (NF-κB). Has been carefully explained as a clarifying example. Since this study includes a comparative study between many predictive QSAR models, the first was based on multiple linear regression, and the others were based on non-linear regression (ANN). Nuclear factor-κB (NF-κB) is a potential therapeutic target for a variety of immunoinflammatory and cancer diseases.

II.2. Principle of QSAR model

The principle of QSAR methods is to set up a mathematical relationship quantitatively relating molecular structure encoded by molecular properties called descriptors of small compounds with their biological activities using data analysis methods. As a result of these relationships, new predictive models can be generated with the following general form [15].

Activity = *f (D1, D2, D3……)*

D1, D2, D3: **Molecular Descriptors**

Table II.1: the summary of the application of several developed QSAR models and the used statistical methods

Author /Date	Application of QSAR/QSPR models	Statistical Method	Ref
Zahouily et al,2005	QSAR models have been elaborated for **anti-malarial** activity prediction, using a data set of 63 active compounds.	MLR/ANN	[16]
Song-Qing et al, 2009	QSAR model has been developed for **anti-corrosion** behavior towards hydrogen sulfide and CO_2 prediction.	MLR	[17]
Shukla et al, 2014	2D quantitative structural activity relationship (QSAR) model has been elaborated for **anti-inflammatory** activity prediction, using a total of 146 tumor necrosis factor- TNF-a inhibitors.	MLR	[18]
Ben Ghanem et al, 2016	Two different QSAR models have been constructed for the **ecotoxicity** prediction of a series of ionic liquids towards the bioluminescent bacterium Vibrio fischeri.	MLR and MLP	[5]
Darnag et al, 2012	Several comparative QSAR models have been elaborated for **HIV Protease inhibitors** prediction using 38 active compounds.	SVM/ANN And MLR	[19]
Shola Elijah et al,2020	Authors have elaborated a robust QSAR model for **anti-tubercular** activity prediction, using a series of 27 active compounds.	MLR	[20]
Abdulrahman et al,2020	QSAR model has been elaborated using the **anti-proliferative activities** of novel series of Parviflorons against MCF-7 breast cancer cells. The model has been used to develop new derivatives with greater efficacy against breast cancer (MCF-7).	MLR	[21]
David et al,2020	Authors have elaborated a robust QSAR model for **anti-diabetic** activity prediction, using a data set of 97 compounds as protein tyrosine phosphatase 1B (PTP 1B) inhibitors.	MLR	[22]
Mozhgan et al., 2020	A series of DAPY-like derivatives as new non-nucleoside reverse transcriptase inhibitors was used for the development of an accurate and robust QSAR model for **anti-HIV** activity prediction.	MLR and ANN	[23]
Almi et al,2020	Two QSAR models have been developed for **anti-cancer** activity prediction, using a series of 38 compounds as human glutathione-S-transferases inhibitors (GSTP1-1). Through the adopted model, 23 hits have been identified as a new potent inhibitor of GSTP1-1.	MLR and ANN	[24]
Hammoudi et al 2020	QSAR model has been elaborated for **anti-cancer activity** prediction, using a series of 30 compounds as Ikk-b Inhibitors. Authors have	MLR	[25]

	suggested a new series of compounds that have promising potential for IKK-b, based on the developed QSAR Model.		
Lemaoui et al 2021	Authors have developed two novel QSPR models for **viscosity** and **density** of hydrophobic deep eutectic solvents prediction using 530 and 606 experimental data points, respectively.	MLR	[26]
Hammoudi et al,2021	Tow QSAR models have been elaborated for **anti-Alzheimer** activity prediction, using 50 compounds as Acetylcholinesterase and Butyrylcholinesterase inhibitors.	MLR	[2]
Zaki et al,2021	QSAR-Based Virtual Screening has been used in the identification of novel compounds with a high inhibitory potential against the main protease (Mpro) of **SARS-CoV-2**.	MLR	[27]

MLP : Multilayer perceptron technique**, MLR** : Multiple Linear Regression, **ANN** : Artificial Neural Network, **SVP** : Support Vector Machine

II.3. Classification of QSAR Methodologies

Based on the structural representation of chemical compounds, the QSAR methods are classified into the following categories [10].

1D-QSAR: Several parameters, such as steric, electronic, and hydrophobic constraints, are used for the selection of the potential molecular descriptor that describes the conformer molecular properties. highest occupied molecular orbital/lowest unoccupied molecular orbital (HOMO-LUMO) energies, Log P, pK_a, are some of the used descriptors that correlate the biological activity [9].

2D-QSAR: Molecular descriptors that correlate the biological activity include molecular fingerprints, total polar surface area, topology, constitutional, quantum chemical, electrostatic, geometrical properties of compounds.

3D-QSAR: this model focused broadly on all such non-covalent interaction fields that surround the molecules [9.10].

4D-QSAR: In this Approach**,** During the molecular dynamics simulation (MDS) time, the occupancy frequencies of the various atom types generated in the cubic grid cells are used as descriptors [28].

5D-QSAR: different induced-fit models in 4D-QSAR are explicitly represented [29]

6D-QSAR: Incorporating multiple solvation scenarios in 5D-QSAR [29].

II.4. QSAR Model Development Steps

Several steps and restrictions are required to construct a reliable, robust, and significant statistical QSAR model. Figure II.1 represents the principal steps for QSAR model development. - Starting with a large experimental data collection with their experimental biological activity [15], the number of compounds should be sufficient, more than 20, with biological activity obtained through a common experimental protocol so that the potency values are comparable [12]. Many free programs and academic databases are interesting and useful resources for collecting databases, among them: Scopus, Wok, Sciencedirect and Springer [31].- Compounds should be subdivided into a training set (2/3 N compounds) for QSAR model elaboration and a test set (1/3 N compounds) for external validation [30] - A large number of molecular descriptors generation for ligands [12]- Selection of significant molecular descriptors, which act as the independent variables in QSAR model using feature selection methods [32].

Molecular descriptors should not be autocorrelated in order to avoid overfitting. The correlation coefficient r is the statistical parameter that evaluates the degree of cross-correlation. If r > 0.97 descriptors are considered to be strongly correlated [33]-Once the model is built using linear regression techniques, e.g., multiple linear regression (MLR) or non-linear machine learning techniques, e.g., artificial neural network ANN, it must be evaluated by internal and external methods to estimate its predictive power and robustness. -Finally, the applicability domain of obtained model determination [30].

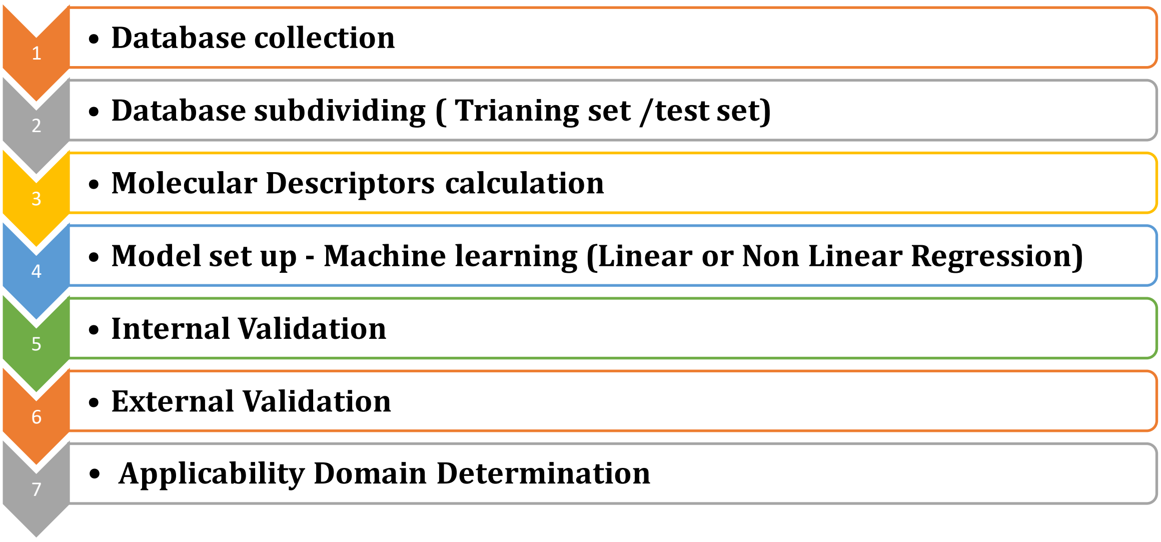

1
• Database collection
2
• Database subdividing (Trianing set /test set)
3
• Molecular Descriptors calculation
4
• Model set up - Machine learning (Linear or Non Linear Regression)
5
• Internal Validation
6
• External Validation
7
• Applicability Domain Determination

Figure II.1: the principal steps for QSAR model development

II.4.1. Data Preparation

A robust QSAR model relies heavily on experimental reference data. Therefore, to be of quality, the choice of the database should be precise, accurate, consistent, and obtained by following a single protocol. Indeed, the experimental conditions generally have a strong impact on the obtained values [15].

It is necessary to check the chemical structures of compounds since errors in structures lead to bad descriptors generation and, as a result, poor models [30]. The efficiency of a QSAR model also depends on the type of molecules included in it. Since models elaborated with compounds characterized by similar structures are highly performing [33], the distribution of data should be homogeneous and regular as possible because most statistical methods are based on this type of distribution. When compounds are highly efficient, the biological data is typically represented in reverse logarithms Log (1/C). Figure II.2 shows some examples of biological data utilized in QSAR analysis [15].

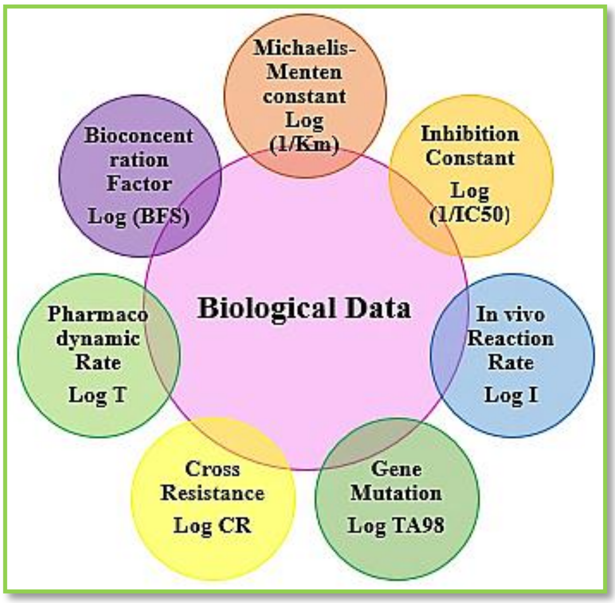

Biological Data
Michaelis-Menten constant Log (1/Km)
Inhibition Constant Log (1/IC50)
In vivo Reaction Rate Log I
Gene Mutation Log TA98
Cross Resistance Log CR
Pharmaco dynamic Rate Log T
Bioconcent ration Factor Log (BFS)

Figure II.2: examples of biological data utilized in QSAR analysis

II.4.2 Molecular Descriptors

The second step of the QSAR study consists of calculating the molecular descriptors of compounds incorporated in the training and test sets. Molecular descriptors describe quantitatively the information contained in the molecular structure to be directly related to biological activity [34]. Many descriptors can be calculated and generated using several software such as MATERIALS STUDIO, CADESSA, OASIS, ADAPT, DRAGON, Mol ConnZ, MOPAC, …etc. [15].

The descriptors may fall into many classes based on the defined algorithm for its calculation and the kind of molecular representation. Some of these include topological, constitutional, geometrical, thermodynamic, and electronic descriptors [35].

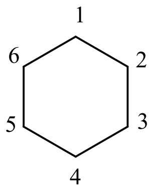
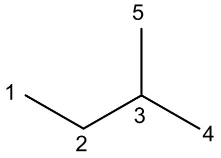
II.4.2.1 Topological Descriptors

The introduction of a mathematical method, called "Graphic Theory," to chemistry was a significant development in automated computer treatment for chemical structures and QSAR. In Chemical Graph Theory, Molecular structures are shown as graphs without hydrogen atoms, usually called molecular graphs. It depicts the atoms as vertices and bonds as edges [36]. The connectivity of atoms in molecules can be described by various types of matrices such as adjacency matrix (Table II.2), distance matrix [36], reciprocal distance matrix [37], Szeged matrix [38], Laplacian matrix [39], Burden matrix [40], reverse Wiener matrix, distance path matrix, distance complement matrix resistance distance matrix and detour matrix [41], that may be mathematically operated to obtain one generally termed as an invariant graph, topological index (TI) or graph-theoretical index. [36].

Topological index has proven to be in good correlation with various biological and physicochemical characteristics, indicating that they are rich in information that can be useful in QSAR and QSPR models analysis [42].

In the previous fifty years, there has been a big effort to produce several hundred of TIs that offer the ability to measure molecular branching, size, and shape [36]. Most of them can be calculated using several softwares such as ChemDes, CODESSA, and JOElib……. Etc. A variety of descriptors are available, among which the more widely applied are: Wiener Index, Information Content Indices, The Hosoya Index, The Zagreb Indices, Topochemical Atom Indices, The Centric Index, Triplet Indices, The Randić Index, Molecular Connectivity Indices, Flexibility Indices, The Variable Connectivity Index [43].

Table II.2 : Adjacency Matrix

Chemical structures	Adjacency Matrices
 Cyclohexane	$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$
 Isopentane	$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

The major applications of these indices are to distinguish molecules based on their degree of branching, size, overall shape, and flexibility [35]. Randic has proposed several features that TIs should present; (1) They must be well correlated with at least one property;(2) they should be independent and do not present any complexity; (4) they are easy to use for a local structure;(5) They should be independent of experimental properties and of others descriptors. Fortunately, the topological descriptors meet most of these characteristics. Hence, they have been widely used in QSAR/ QSPR studies for structural similarity or dissimilarity of chemical compounds characterization [35].

II.4.2.2. Constitutional Descriptors

Constitutional descriptors are the simplest employed descriptors that characterize the molecular composition of compounds, with no necessity for any topological or geometrical information. These descriptors do not differentiate among isomers. Therefore, they are inert to any conformation changes [35]. Examples of these descriptors are mentioned in Table II.3.

Table II.3: Example of some constitutional descriptors

Descriptor	Description
MW	*Molecular weight*
Si	*sum of first ionization potentials (scaled on Carbon atom)*
Me	*mean atomic Sanderson electronegativity (scaled on Carbon atom)*
Mi	*mean first ionization potential (scaled on Carbon atom)*
nSk	*number of non − H atoms*
nBO	*number of non − H bonds*
SCBO	*sum of conventional bond orders (H − depleted)*
RBN	*Number of rotatable bonds*
nDB	*Number of double bonds*
nAB	*Number of aromatic bonds*
nC	*Number of Carbon atoms*
np	*Number of Phosphorous atoms*
H %	*Percentage of H atoms*
nCsp	*Number of sp hybridized Carbon atoms*

II.4.2.3. Quantum Chemical Derived Descriptors

Quantum-chemical methods and molecular modelling techniques are thus an appealing source of novel molecular descriptors that may, in theory, represent all of the geometric and electronic features that characterize the shape, reactivity, and binding properties, corresponding to the minimum energy of a complete molecule as well as of molecular fragments [44,45].

Recently, quantum chemically derived descriptors have gained a lot of traction in QSAR/QSPR modelling due to the versatility and reliability of their prediction. The use of this kind of descriptors in QSAR/QSPR studies helps to clarify the mechanism of action directly in terms of the chemical reactivity of the studied compounds. Accordingly, information about the nature of intermolecular forces that influence the biological activity will be included in the obtained QSAR models [46,47,48]. A summary of several used quantum chemical descriptors is present in Table II. 4 [44].

Table II. 4: several used quantum chemical descriptors

HOMO and LUMO Energies	**Description**
$\boldsymbol{E_{HOMO}}$	$HOMO$: *energy of the highest occupied molecular orbitals*
$\boldsymbol{E_{LUMO}}$	$LUMO$: *energy of the lowest unoccupied molecular orbitals*
$\boldsymbol{I = -E_{HOMO}}$	*Ionisation potential* (I)
$\boldsymbol{A = -E_{LUMO}}$	*Electron affinity* (A)
$\boldsymbol{E_{LUMO} - E_{HOMO}}$	*Energy band gap* (ΔE)
$\eta = \frac{1}{2}\,(\boldsymbol{E_{LUMO} - E_{HOMO}})$	*Global hardness* (η)
$\boldsymbol{\mu} = \frac{1}{2}\,(\boldsymbol{E_{HOMO} + E_{LUMO}})$	*Chemical potential* (μ)
$\boldsymbol{\omega} = \frac{\boldsymbol{\mu^2}}{2\eta}$	*Electrophilicity index* $(\boldsymbol{\omega})$
Orbital Electron Densities	**Description**
$\boldsymbol{q_{A,ó}}$, $\boldsymbol{q_{A,\pi}}$	ó $-$ *and* $\pi -$ *electron densities of the atom A*
$\boldsymbol{Q_{A,H}}$, $\boldsymbol{Q_{A,H}}$	$HOMO/LUMO$ *electron densities on the atom A*
$\boldsymbol{f_r^E = \sum(C_{HOMO,n})^2}$	*Electrophilic atomic frontier electron densities,* $\boldsymbol{C_{HOMO,n}}$*are the coefficients of the atomic orbital* X_n *in the HOMO.*
$\boldsymbol{f_r^N = \sum(C_{LUMO,n})^2}$	*nucleophilic atomic frontier electron densities,* $\boldsymbol{C\_(LUMO,n)}$ *a coefficients of the atomic orbital* X_n *in the LUMO*
$\boldsymbol{F_r^E = f_r^E / E_{HOMO}}$ $\boldsymbol{F_r^N = f_r^N / E_{LUMO}}$	*Indices of frontier electron density*

Charges	Description
$\boldsymbol{Q_A}$	*Net atomic charge on atom A*
$\boldsymbol{Q_{min}}_{,}\boldsymbol{Q_{max}}$	*Net charges of the most negative and most positive atoms*
$\boldsymbol{Q_{A,B}}$	*Net group charge on atoms A, B*
$\sum \boldsymbol{q_A^2}$	*sum of squared charge densities on atoms of type A*
Superdelocalizabilities	**Description**
$\boldsymbol{S_{E,A}}$ $,\boldsymbol{S_{N,A}}=2\ \sum_j \sum_{m=1}^{N_A} (C_{Jm}^A)^2/E_j$ $\sum \boldsymbol{S_{E,A}}\quad ,\sum \boldsymbol{S_{N,A}}$	*Electrophilic and nucleophilic superdelocalizabilities* *(sum over the* *occupied (E) or unoccupied (N) MO (j) and over the number of* *AO in the atom A* $(m\ = 1,\ldots,\boldsymbol{N_A})$ *Sums of electrophilic and nucleophilic superdelocalizabilities*
Atom-Atom Polarizabilities	**Description**
$\pi_{AA}\ ,\pi_{AB}\ =$ $4\sum_i \sum_a \sum_p \sum_r \frac{c_{pI}^A\ c_{pa}^A c_{rI}^B c_{ra}^B}{£_i - £_a}$	self-atom polarizabilities and atom-atom polarizabilities (sum over MOs (*i*, *a*) and over valence AOs (*p*, *r*))
Energies	**Description**
$\boldsymbol{E_T}$	Total Energy
$\boldsymbol{E_b}=\sum_I^N \boldsymbol{E_{Ai}} - \boldsymbol{E_T}$	Binding Energy
ΔH_f^0	Heat of formation
$\Delta(\Delta H_f^0)$	Relative heat of formation ionization potential
IP	Ionization potential
EA	Electron affinity is the difference in total energy between the neutral and anion radical species
ΔE	The energy of protonation, the difference between the total energy of the protonated and neutral forms

Dipole Moments and Polarity Indices	Description
$\boldsymbol{\mu}$	*Molecular dipole moment*
$\boldsymbol{\mu_{char}}$,$\boldsymbol{\mu_{hyber}}$	Charge and hybridization components of the dipole moment
$\boldsymbol{\mu^2_{D_x\ D_Y\ D_Z}}$	square of the molecular dipole moment components of dipole moment along inertia axes
$\boldsymbol{\Delta}$	Submolecular polarity parameter (largest difference in electron charges between two atoms)

II.4.2.4. Geometrical Descriptors

The 3D coordinates of the atoms in a particular molecule are used to calculate geometrical descriptors. Compared to topological descriptors, these descriptors provide more information and discriminating power for similar chemical structures and molecular conformations [35]. Despite the fact that they contain a lot of information, these descriptors frequently have disadvantages since they are only obtained after a geometry optimization of structures, therefore the cost of calculating them [15]. In addition to their need for alignment rules for performing molecule comparability. Examples of these descriptors: MoRSE (3D-Molecular representation of Structures based on Electron), WHIM (Weighted Holistic Invariant Molecular) [35].

II.4.2.5. Physicochemical descriptors

There are several descriptors that can be measured experimentally and are widely used to study structure-activity correlations, among them: the partition coefficient (octanol-water) (log P), permeability, and hydro solubility. These characteristics have a significant influence on determining the drug's concentration in the body [35].

Lipophilicity: Lipophilicity is the sum of intermolecular forces, including a solute and two immiscible solvents between which it partitions [49]. It plays a crucial role in the pharmacological process of drugs, that is, absorption, distribution, metabolism, and excretion) [25] as well as the biological membrane penetration and hydrophobic interactions with receptors. A drug's lipophilicity refers to its affinity for a lipophilic or an aqueous environment [50]. Careful attention to physicochemical characteristics can enhance the chance of effective delivery and therapeutic efficacy for a promising candidate drug molecule. Since recent findings show that, in addition to defining pre-clinical ADMET (absorption, distribution, metabolism, elimination, and toxicity) characteristics, drugs with optimum lipophilicity within a defined optimal range [50] may have a higher probability of development success [51]. In many particular situations, the introduction of more complex quantitative structure-activity relationships (SARs) pioneered by Hansch in the 1960s continued to highlight the relevance of the lipophilicity term [52].

Lipophilicity is measured using the partition coefficient Log P, which is defined as the logarithm of solute concentration in octanol over unionized solute concentration in water, or log D, the distribution coefficient in octanol-water at a specific pH. In general, the computed log P (clog P) is frequently used to determine lipophilicity rather than the observed log P, with measured partition coefficients obtained on key compounds through a project's progression [53]. Lipinski suggested that the desirable lipophilicity values for drug candidates are universal to some extent, emphasizing the importance of logP values in particular. (logP) < 5 [52]. For the design and more sophisticated lead optimization efforts, it has been proposed that regions of optimal lipophilicity be addressed [52].

II.4.2.5.1 Experimental Measuring logP values

-The shake-flask method

This traditional method allows the determination of the partition coefficient of neural or ionic solutes by controlling the PH of the aqueous phase. The shake-flask method has several practical constraints, including microemulsion generation, phase-volume ratio accuracy, solute impurities, solute stability or volatility, and time consumption [54].

-The potentiometric method

The ionization constants and logP values of ionizable solutes can be determined using the potentiometric method. The technique is based on the fact that when an ionizable molecule is titrated in an aqueous solution with an organic phase present, the titration curve moves to the right for acids and to the left for bases. The potentiometric technique has the benefit of being able to use a number of solvents and measure a wide range of logP values. However, it is not always easy to understand the results [55].

-Cyclic voltammetry at the ITIES

In medicinal chemistry, cyclic voltammetry was recently used to assess the lipophilicity of ions and investigate their processes of transfer at the interface. The major benefit of cyclic voltammetry at the ITIES is that, unlike other approaches that do not regulate the Galvani potential difference, the potentials are controlled here, resulting in standard logP values that are independent of experimental variables except temperature and solvents [56].

-Calculating and estimating of log P Methods

Ghose and Viswanadhan method, Rekker method, Klopman and Iroff method and Hansch method are the oldest used fragmentally techniques, in which a molecule is divided into predefined fragments, and the corresponding contributions are summed to lead to an estimated value of log P [30].

II.4.2.5.2 hydro-solubility

The water-solubility of drugs plays a crucial role in drug design and development. This parameter has a significant impact on many processes such as pharmaceutical formulation, bio pharmacy as well as bioavailability. The importance of drug solubility was generally overlooked; it would be assessed once a possible development candidate has been discovered, usually as part of drug metabolism research.

The situation has changed in recent years since the relevance of solubility is recognized, and it is tested or predicted at an early stage in drug development. In recent years, advances in combinatorial chemistry and high-throughput screening (HTS) techniques have increased the ability to synthesize and test thousands upon thousands of chemicals in a short amount of time. It has been noted that compounds generated in HTS drug discovery are more drug-like molecules with a desirable aqueous solubility and lipophilicity. Many significant computational models have been developed by Johnson and Zheng for aqueous solubility of compounds prediction [57.58].

II.4.3. Descriptor Selection Methods in QSAR Study

selection of meaningful descriptors and correct analysis allows the construction of reliable QSAR models. Although a QSAR model can be generated including all calculated descriptors, but there are many reasons to choose only a small series of descriptors which typically contain the information required to construct a robust mathematical model. When the number of calculated descriptors is marked with n, The variable selection technique is frequently characterized as picking the m< n descriptors that allow the construction of the optimal QSAR model. Feature selection approaches reduce the number of degrees of freedom in the model by removing unneeded features.

As a result, the ratio of data points to degrees of freedom rises, resulting in more predictive models [13].

- Eliminating unnecessary and redundant descriptors will enhance the prediction accuracy of resulted models.

- A developed model containing a small set of descriptors generally is simple and can be interpreted easily, also will reduce the complexity and time requirement of machine learning methods.
- The accuracy of the QSAR model decries when the number of molecular descriptors is more than the number of molecules of interest.

In QSAR modelling: classical feature selection methods, feature selection employing artificial intelligence algorithms (wrapper approach), and miscellaneous methods are the main used classes for variable selection (Figure II.3) [35].

II.4.3.1. Classical Methods

II.4.3.1.1. Forward Selection (FS) Method

The forward selection approach consists of adding all molecular descriptors at one time. The descriptor with the highest fitness function is the first to be included in the regression. The initial descriptor chosen is used in all subsequent QSAR models. New descriptors are gradually added to the regression, with each descriptor chosen because it provides the best fitness function when combined with those already picked. As a stopping criterion, many rules can be used. FS has the benefit of being relevant even when the initial data set has more explanatory factors than sites [59]

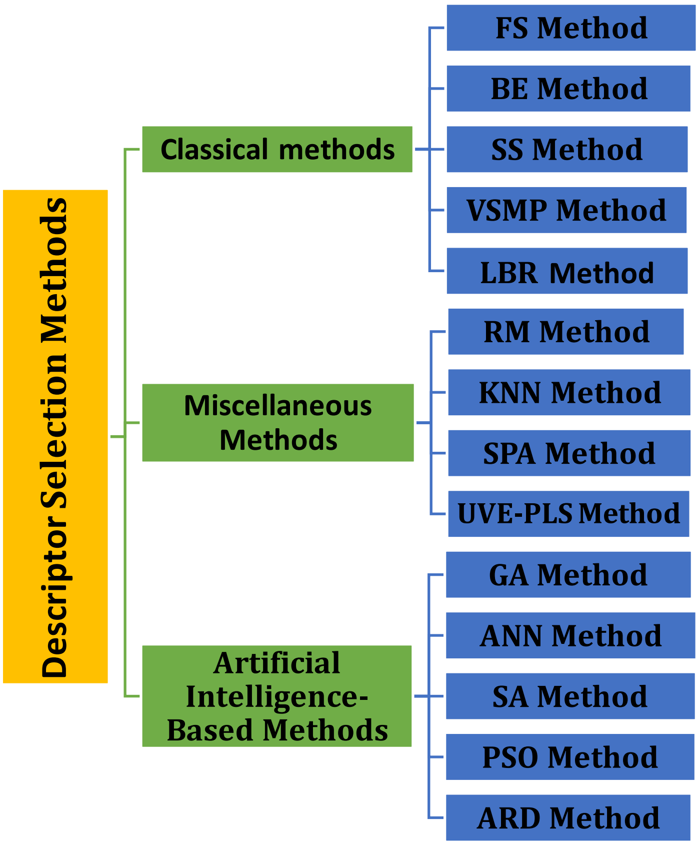


Figure II.3: Descriptor Selection Methods

II.4.3.1.2. Backward Elimination (BE) Method

When compared to the forward selection technique, the first stage of this strategy is to include all of the variables in the QSAR model. In the next phases, the dimensionality of descriptors will be reduced by deleting variables one by one based on the criterion (descriptors with the least contribution to the reduction of predicted residual error sum of squares) (PRESS). When all of the relevant descriptors are significant, or all but one predictor variable has been eliminated, the exclusion procedure is terminated [60].

II.4.3.1.3. Stepwise Regression Method

The stepwise descriptor selection technique is a well-known variable selection methodology that has long been used in QSAR investigations. It is based on the combination of the Backward elimination and forward selection methods [32]. In stepwise method, a variable that entered the model in the early phases of selection may be eliminated in the later stages. The computations used for variable inclusion and exclusion are the same as those used for forward and backward selection. That is, the stepwise technique is fundamentally a forward selection procedure, but the potential of eliminating a variable is evaluated at each stage, as in backward elimination. The number of significant parameters that remained in the obtained model is determined by the levels of significance assumed for the inclusion and exclusion of variables from the model [61].

II.4.3.1.4. Leaps-and-Bounds Regression

This technique can provide quickly varied sizes of the optimal subsets of variables Without considering all feasible subsets. It is based on the fundamental inequality, $RSS\ (A) \leq RSS\ (A_i)\ where, RSS: is\ the\ residual\ sum\ of\ square$

$A: is\ any\ set\ of\ independent\ variables\ and\ A_i: is\ a\ subset\ of\ A.$

The usage of the inequality can limit the number of subsets assessed in a search for the optimal subset regression [61]. For example, set A_1 contains three variables with $RSS, 596$; set A_2 contains four variables with $RSS, 605$. Thus, all of the subsets of A_2 will be ignored because these subsets have RSS greater than that for A2 and also for A1.

II.4.3.1.5. Variable Selection and Modeling based on the prediction

Tow statistics parameters: The interrelated coefficient (r_{int}) between the variables and the correlation coefficient in the leave–one–out (LOO) cross-validation (q) have been introduced into the all-subset regression (ASR) procedure for a novel computer program construction, which allows the variable selection and modelling based on the prediction. To the aim of accelerating the classical ASR and getting a better variable subset based on the predictive quality [62].

How to choose and analyze the optimal subset from a big independent variable matrix involving K molecules, each with L descriptors, x (k, l)? In the VSMP software, the optimal selection job is completed in two major parts are well explained in the following reference [62].

II.4.3.2. Miscellaneous Methods

II.4.3.2.1. Replacement Method

This novel approach consists of substituting a selected variable from the set with another that minimizes the total standard deviation (Stot), thus its name (Replacement Method; herein after RM)

To do this, d descriptors X_1, X_2,….. $X_3$are picked at random, then linear regression is run [35].

II.4.3.2.2. Nearest Neighbors (kNN) method

Tropsha and his colleagues created kNN as a method for variable selection in QSAR experiments. This group utilized this method to create QSAR models for a variety of data types. In brief, a series of molecular descriptors is chosen at random as a hypothetical descriptor pharmacophore. The number of selected variables is set to various values in order to achieve the best LOO q2 feasible [35].

II.4.3.2.3. Successive projections algorithm (SPA)

Successive Projections Algorithm is a new variable selection method which has been suggested for multivariate calibration. This technique uses basic operations in a vector space to decrease variable collinearity [63].

II.4.3.2.4. Uninformative variable elimination-partial least square (UVE-PLS)

Centner et al. suggested the uninformative variable elimination-partial least squares (UVE-PLS) technique, which was adopted for quality model improvement. Based on the exclusion of uninformative variables with a high variance but low correlation with the end point [64].

II.4.3.2.5. Factor analysis (FA)

Factor analysis is a technique for reducing data. Thus, the primary uses of FA: are to minimize the number of variables and to find the structure of the relationships between variables. FA can also be used as a data-pre-processing step for the identification of significant predictors variables that influence the response variable and to prevent collinearities among them [65].

II.4.3.3. Artificial Intelligence-Based Methods

II.4.3.3.1. Genetic algorithm (GA) method

GA method has been employed in QSAR investigation for the first time by Rogers and Hopfinger. They noted a variety of benefits over other descriptor selection techniques. The GA used four fundamental phases to determine the best set of descriptors: (1) initialization, (2) selection, (3) genetic operator, and (4) fitness of evaluation [76].

II.4.3.3.2. Artificial Neural Network (ANN) Method

ANNs are frequently employed as a regression approach in combination with feature selection optimization strategies. It was also utilized to determine descriptors that were most related to biological activity. In descriptor selection by ANN, an approach can be suggested to variable selection that uses a neural network model, the tool to identify which descriptors are to be eliminated [32].

II.4.3.3.3. Simulated Annealing (SA) Method

Simulated annealing (SA) is a global search technique developed by Metropolis et al. SA derives its fundamental understanding from metallurgy. The proposed algorithm allows SA to avoid the local optimality and the improvement of results of several search iterations [67].

II.4.3.3.4. Particle swarm optimization (PSO)

Particle swarm optimization (PSO) is a non-linear method. This approach belongs to the category of evolutionary computation techniques. It involves a population of particles which are flown through a multi-dimensional search space defined by an original matrix of descriptors [68.69].

II.4.3.3.5. Automatic Relevance Determination (ARD)

The ARD can be coupled with a Bayesian neural network for optimal QSAR model development. These BRANN-ARD networks have the ability to handle a variety of difficulties that emerge in QSAR modelling, including the following: model selection; robustness of model; size of validation effort; optimization of network architecture and choice of validation set [70].

II.4.4. Modelling Algorithms

The final stage in developing the QSAR model, given the specified descriptors, is to generate the mapping between the activity and the values of the descriptors. Simple but effective techniques represent activity as a linear regression of the selected features. Linear models have served as the foundation of QSAR analysis. They are easy to comprehend and reasonably accurate for small datasets of identical chemicals, particularly when the descriptors are properly chosen for a certain activity.

Other methods that are non-linear expand this technique to more complicated relationships**.** Such models may improve in accuracy, particularly for vast and varied datasets. However, they are typically more difficult to interpret. The most common linear and non-linear regression-based modelling algorithms are presented in Figure II.4 [71].

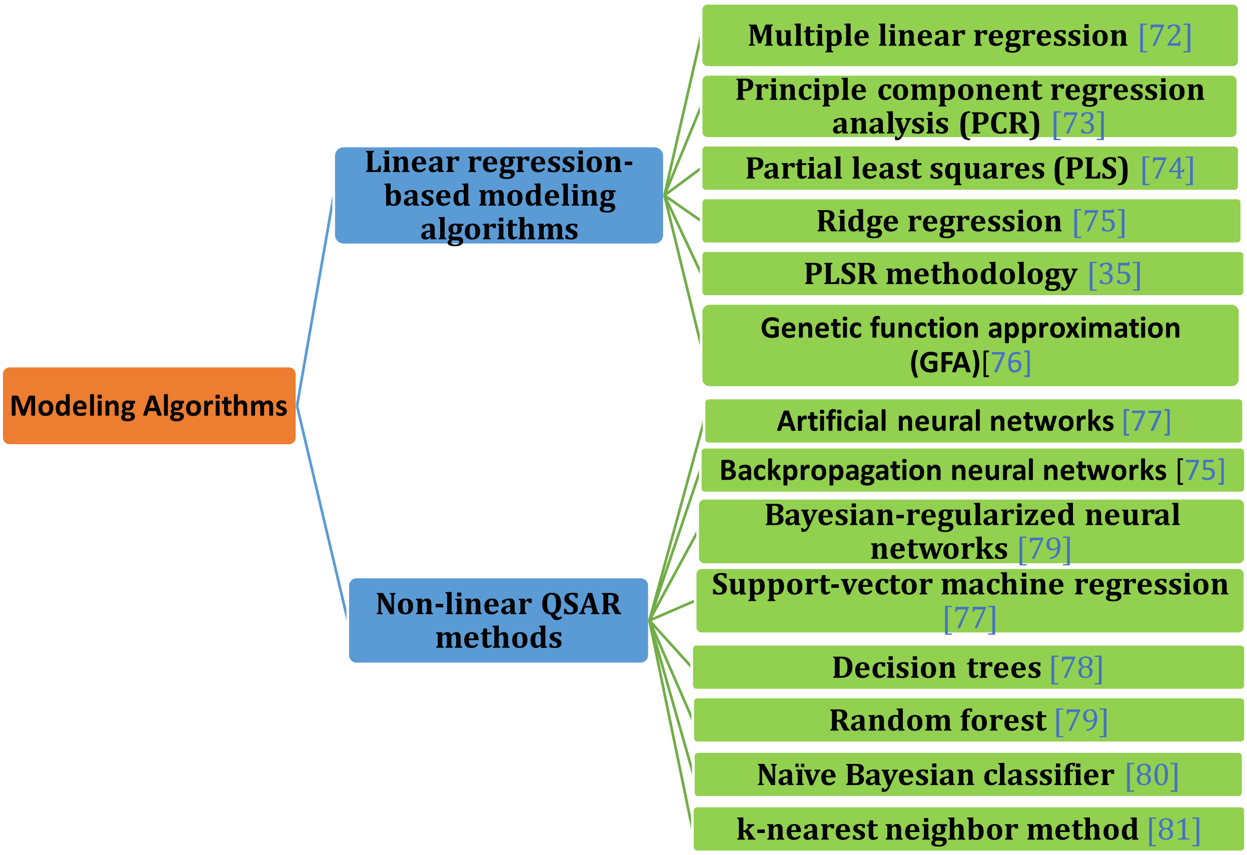

Modeling Algorithms
Linear regression-based modeling algorithms
Multiple linear regression [72]
Principle component regression analysis (PCR) [73]
Partial least squares (PLS) [74]
Ridge regression [75]
PLSR methodology [35]
Genetic function approximation (GFA)[76]
Non-linear QSAR methods
Artificial neural networks [77]
Backpropagation neural networks [75]
Bayesian-regularized neural networks [79]
Support-vector machine regression [77]
Decision trees [78]
Random forest [79]
Naïve Bayesian classifier [80]
k-nearest neighbor method [81]

Figure II.4: QSAR Modeling Algorithms

II.4.4.1 Multiple linear Regression (MLR)

Multiple linear regression has been one of the most extensively utilized mapping approaches in QSAR during the last few decades [71]**.** MLR is a statistical approach that involves a series of variables to predict the outcome of an answer variable. The object of (MLR) is to determine how the dependent variable (Y) is correlated to the independent variables (X_i) by fitting a linear regression [85.86]. According to [87], the MLR approach is used when: (1) The number of examples is far more than the number of parameters to be evaluated; (2) The data shows a consistent pattern of behaviour; (3) There are a few missing data points; (4) A limited number of independent variables are adequate (after transformations if necessary) to predict output variables linearly allowing for an interpretable representation [88].

The application of the MLR approach necessitates the verification of the related assumptions. The primary assumptions to be examined are as follows: -Linearity: the relationship between each variable and the response is linear, thus, the model accurately represents the data's behaviour. –The error component is a variable that is independent and normally distributed, having a constant variance and a mean value of zero [89].

Multiple linear regression is expressed by the following model, which is characterized by k predictor variables $x_1 ,x_2 \ldots\ldots x_k$ [85].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_k + \varepsilon \qquad \text{Eq1}$$

ε : $is\ the\ error\ term$

$\boldsymbol{\beta_0}$: $Is\ the\ intercept$

$\boldsymbol{\beta_1 - \beta_k}$: $partial\ regression\ coefficients,$ e.g., $\boldsymbol{\beta_i}$ When $1 \leq i \leq k$ reflects the change in the mean response corresponding to a unit change in x_iWhen all other variables are maintained constant.

MLR's goal is to find the set of coefficients $(-) = \{\boldsymbol{\beta_0}, \boldsymbol{\beta_1} \ldots . \boldsymbol{\beta_k})$ given observations X and targets Y. The MLR problem is frequently solved with lease squares. Assume that each predictor variable $x_1; x_2; \ldots ; x_k$contains n observations. Then x_{ij} represents the ith observation of the jth predictor variable x_i. For example, x_{41} represents the first value of the fourth observation. To be more specific, the preceding equation can be written as:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} + \ldots\ldots + \beta_{jk} + \varepsilon \qquad \text{Eq2}$$

$1 \leq j \leq n$

$\boldsymbol{y_j}$: $is\ the\ jth\ target\ value$

The n-equation system may therefore be expressed in matrix notation as follows:

$$y = X\beta + \varepsilon \qquad \text{Eq3}$$

The matrix is known as the design matrix. It includes information about the predictor variables' levels at which the observations are acquired. All of the regression coefficients are contained in the vector. To produce the regression model, **β** is calculated using the least-squares method, as shown below.

$$\hat{\boldsymbol{\beta}}= (XX^T)^{-1}X_y^T \qquad \text{Eq4}$$

$$\text{Wher: } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ . \\ . \\ . \\ . \\ \beta_i \end{bmatrix}, \quad \text{X=} \begin{bmatrix} 1 & x_{11} & x_{12} & . . . . & x_{1i} \\ 1 & x_{21} & x_{22} & . . . . & x_{2i} \\ : & : & : & ::::: & \\ . & . & . & ..... & \\ . & . & . & ..... & \\ & x_{j1} & x_{j2} & . . . . & x_{ji} \end{bmatrix}, \text{Y =} \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

After obtaining $\hat{\beta}$ the estimated value of y is calculated by the following equations:

$\hat{y}$ =X$\hat{\beta}$+ ε	Eq5
$\varepsilon = y - \hat{y}$	Eq6

II.4.4.2 Artificial Neural Networks (ANN)

In QSAR, artificial neural networks (ANN) are a more flexible modelling tool that allows for the depiction of more intricate interactions between a high-dimensional descriptor space and the provided retention data [90]. Neural networks are characterized by their ability for non-linear data modelling. This is a significant benefit because so many scientific interactions are non-linear. Previously, linear QSARs were used to quantitatively correlate molecular descriptors, such as shape or dipole moment, to biological and other types of activity, but there is little theoretical support for the view that linear models are the most effective. As a result, this has proven to be a fruitful field for networks' non-linear abilities [91].

An artificial neural network (ANN) (Figure II.5), more commonly referred to as a "neural network" (NN), is a mathematical or computational model that is inspired way biological neural networks (Figure II.6), such as the brain, or in other words, a modelisation of a biological brain system. It is composed of a network of interconnected neurones that analyses data using a connectionist approach to computation. Generally, an ANN is an adaptive system that modifies its structure in response to external or internal data that flows through the network during the learning phase [92].

Neural networks essentially can be defined based on the following three characteristics: The Architecture specifies the number of layers and the total number of processing elements or nodes in each layer; The learning algorithm used to update the connection weights; and the activation functions that are employed in different layers [93].

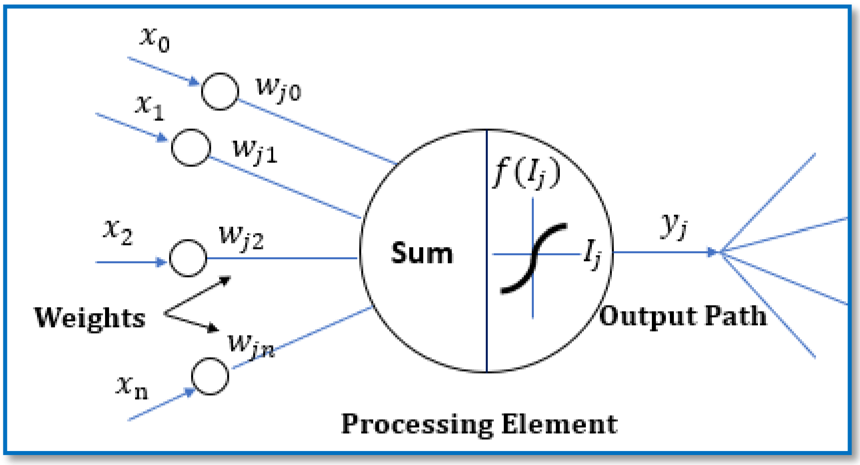


Figure II.5: Artificial neuron [92]

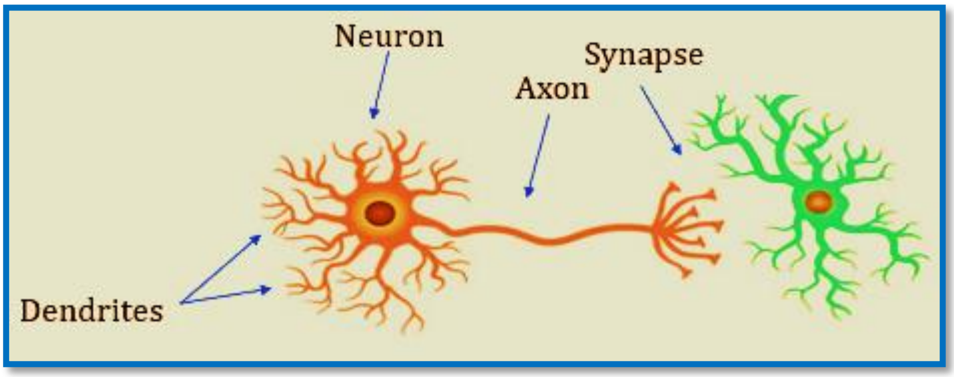


Figure II.6: Biological Neural [92]

A neural network is made up of a systematic arrangement of "neurons," which are again structured by the number of layers. A layer's neurons are all connected to the following layer's neurons in a weighted way. A neural network consists of three types of layers “input” layer, one or many “hidden” layer(s), and the “output” layer, as illustrated in Figure II.7. The input layer contains input neurons (nodes) that provide data to the hidden layer via synapses, and then the hidden layer, in turn, sends data to the output layer via additional synapses. Weights are stored in the synapses, which allow them to manipulate the input and output to multiple layers [93].

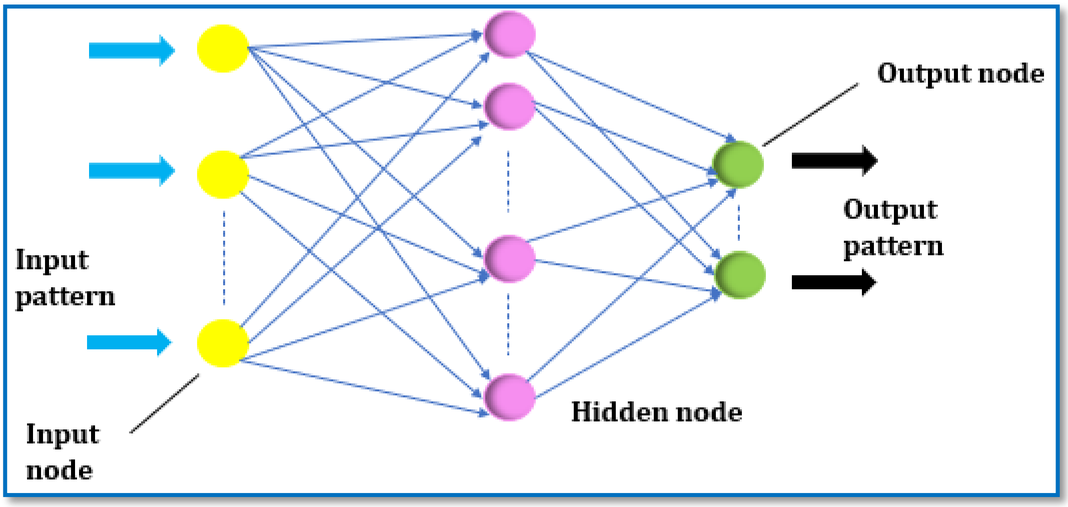


Figure II.7: A typical structure of ANN [94].

The input from each node in the preceding layer (x_i) is multiplied by an adjustable connection called weight (w_{ij}). At each node, the weighted input signals are summed, and a threshold value (Θ_j) is added to each PE. This combined input (I_j) is then processed by a non-linear transfer function (f (.)) to yield the node output of the PE (y_j). The output of one node serves as the input to the nodes in the following layer. Equations 1,2, and Figures 7 and 8 summarize this procedure [94].

$$I_j = \sum w_{ij}\, x_i + \Theta_j \qquad \text{Eq7} \qquad \text{summation}$$

$$y_i = f\,(I_j) \qquad \text{Eq8} \qquad \text{transfer}$$

At the input layer, which contains the input data, the propagation of information in ANNs begins. The network changes its weights based on the presentation of a training data set. It then uses a learning rule to find a set of weights that will give the input/output mapping with the smallest possible error, which is what the network does [94].

II.5. Validation Techniques

After the regression equation is found, it is also important to look at how well the model fits and how stable it is. Also, it is important to look at how accurate it is. To validate a method is to make sure that the method is both reliable and useful for a specific use. Validation of a QSAR model is the process of determining a QSAR's predictive ability and mechanistic basis for practical purposes.

Validation determines whether the model accurately depicts reality from the perspective of the model's intended application [95]. In other words, validation procedures are acknowledged as critical processes for evaluating the statistical acceptability and applicability of built models on a new set of data in order to determine the predictive confidence of the models [96].

II.5.1. Internal Validation

II.5.1.1. Measure of Goodness of Fit

The coefficient of multiple regression (R^2) is a statistical parameter that is used for a given data fit assessment. It describes the amount of variation experienced by the response (dependent variable) in the regression. $R^2 = 0.00$ indicates the non-presence of any relationship between the independent variables, whilst a value of 1.00 indicates that the variables are well fitted. ${R^2}_{adj}$(adjusted) is a comparable statistical method that may also be used to assess the quality of fit.

Furthermore, the standard error of estimate (se) may be used to assess the quality of fit. The standard error of estimate, which is determined from both the predicted and observed values, describes the distribution of observed values from the regression fit. The model is considered more reliable when the (se) value is less; furthermore, it should be more than the experimental error [97].

chi-squared (x^2) and root-mean-squared error are used to determine if the model has the predictive quality indicated by the R^2. The usage of RMSE represents the error between the mean of the observed and predicted values. The chi-squared value x^2demonstrates the gap between observed and predicted biological activities [98].

II.5.1.2. Cross-Validation

Cross-validation (CV) is a widely used technique for internal QSAR model validation. CV is a technique for evaluating a model's prediction performance using a smaller amount of structural data. Typically, at each time, one compound of the set is removed, and a new model based on the reduced dataset is developed, which is then used the activity of removed compounds prediction. Once the technique has been completed, it is repeated many times until all compounds have been eliminated and predicted once. This is referred to as the "leave-one-out (LOO) approach."

Similarly, the leave-n-out or leave-many-out CV approach consists of eliminating more than 1 compound from the dataset at each time. As a result of the LOO technique, the cross-validated correlation coefficient r_{cv}^2(or q^2) is calculated, which is a measure of both the robustness and the predictive capacity of the model.

$$r_{cv}^2 = (PRESS_0 - PRESS)/PRESS_0 \qquad \text{Eq9}$$

$PRESS_0: the\ mean\ of\ the\ observed\ biological\ activity;$
$PRESS: the\ sum\ of\ the\ squares\ of\ the\ differences\ between\ the\ predicted\ and\ the\ observed\ activity$

Several scholars believe that a high q^2 is the final confirmation of a great predictive capacity of the QSAR model, although this is not the case. When test sets with known biological activities values were provided for prediction, it was discovered that there was no link between the q^2 and the r^2 values. The coefficient q^2 should be seen more like an internal consistency metric than like a real prediction metric for the model that was developed. It should be emphasized that, while fitting the experimental data is simpler than predicting them from the QSAR model, the model's r^2 is always greater than q^2. Cross-validation is not without flaws. As a result, a high value of q^2 is not sufficient to determine if a QSAR model is highly predictive [99].

II.5.1.3. Y- Randomization (Scrambling model)

The y-randomization test is used to find and quantify chance correlations between the dependent and descriptor variables. In this case, the phrase "chance correlation" refers to the fact that although the actual model may include descriptors that are statistically highly correlated with y, in reality, there is no causal link encoded in the corresponding correlations with y since they are unrelated to the mechanism of action. The purpose of this test is to ensure that the obtained model is not coincidental.

Scrambled Model development is a unique approach to validating the descriptors used in the model since the bioactivities are randomized, assuring the new model is produced from a false data set. This approach is used to validate the initial QSAR model and to confirm that the descriptors used are adequate. These new models (Scram-models) are constructed with the same descriptors as the original model but with scrambled Y. each Scram-model should be of inferior quality and devoid of significance, remaining with the same descriptors with no changing since y-randomization is not dependent on parameter optimization [100]. After repeating this process several time, if the random model is similar to the original one, that indicates the non-sufficient of data for model support.

How should the results of each randomization run be analyzed, and how many runs should be conducted? There are several methods for determining if a true model is characterized by random correlation. Eriksson and Wold approach is a simple method, which is defined as a collection of decision inequalities determined by the following values: Q^2_{Yrand}, R^2_{yrand} and their relationship

$R^2_{yrand} > Q^2_{Yrand}$

$R^2_{yrand} < 0.2,\ Q^2_{Yrand} < 0.2\ \ indicates\ the\ non-correlation\ chance$

$any\ Q^2_{Yrand}\ \ and\ \ 0.2 < R^2_{yrand} < 0.3\ \ negligible\ chance\ correlation$

$any\ Q^2_{Yrand}\ \ and\ \ 0.3 < R^2_{yrand} < 0.4\ \ tolerable\ chance\ correlation$

$any\ Q^2_{Yrand}\ \ R^2_{yrand} > 0.4\ \ recognized\ chance\ correlation.$

The second method identifies the shortest distance in units of Q^2 or R^2 between the true model and all randomized models. This minimal distance is then represented in terms of the standard deviation for each randomization run. The real model is distinguished from randomized models by having a sufficient confidence level for the normal distribution. Another way that can also be used to measure chance correlation is to calculate the absolute value of the Pearson correlation coefficient, r, between randomized vectors y and the original vector y. For both randomized and real models, 2 y randomization plots are created, $r - R^2_{yrand}$ and $r - Q^2_{Yrand}$. and the linear regression lines are obtained as follows:

$$Q^2_{Yrand} = a_Q + b_Q \mathrm{r} \qquad \text{Eq10}$$

$$R^2_{yrand} = a_R + b_R \mathrm{r} \qquad \text{Eq11}$$

the $intercepts\ a_{R\ and}\ a_Q$ measures the random correlation If $a_Q < 0.05\ and\ a_R < 0.3$ indicates the non-correlation chance of the model. The intrinsic random correlation contained in X is observable when the statistical effects of randomizing the y vector are removed, i.e., when the correlation between the real and randomized y vectors r = 0 [100.-104].

II.5.1.4. Bootstrapping

Bootstrapping is more often employed internal validation method. It creates confidence bounds for individual model parameters. Model validation using the bootstrap technique does not involve extra data gathering. This approach is generally used to determine the model's generalizability. In this technique, compounds (samples) are chosen randomly from the database. The simplest version of bootstrapping. Rather than repeatedly studying subsets of data, subsamples of data are studied repeatedly.

Each sub sample is a random sample with a replacement from the whole sample. Typically, bootstrap validation generates K groups of size n by repeated randomly selecting n items from the original data set. Some of these items can be chosen again and again in the same random sample, but other items may never be chosen. The model constructed using n randomly chosen items can be used for the properties prediction of excluded compounds. In bootstrap validation, a model is considered robust when Q^2 value is hight [105].

II.5.2. External Validation

The primary goal of external validation of a QSAR model is to make a straightforward assessment of the mistakes that were encountered during the prediction of a test set [106].

II.5.2.1. Choosing Training and Test Sets:

It is important to note that the selection of suitable training and validation sets is a step of QSAR that is sometimes disregarded. This involves determining how to pick the sets such that they satisfy the essential statistical requirement of representativity. Compounds in the training and validation sets must be representative of the class of compounds from which they are derived, which means they must be selected in such a way that they appropriately span the chemical and structural characteristics of the compounds under consideration [107].

Optimal splitting is ensured when each member of the test set is close to at least one point of the training set. The development of reasonable methodologies for the selection of training and test sets is a current topic of research interest. The straightforward random selection, activity sampling, Kennard stone are some of the used approaches for training and test set creation. These strategies aid in the achievement of desired statistical features in both the training and test sets to varying degrees.

II.5.2.2. Evaluating the prediction potential of QSPR models

Generally, a valid QSAR model is characterized by generalization ability that makes it capable of predicting the activity of the test set similar to those of the training set model [108]. In this stage, it is required to make a comparison between the observed and the predicted values of suitably large external series of molecules that were not implemented in the model construction. The performance accuracy prediction of the QSAR on this involved external series determines the real predictive power of a QSAR model [109]. To measure the prediction capacity of a QSPR model, an external q^2may be calculated as follows:

$$q^2{}_{ext} = 1 - (\sum(y_i - \hat{Y}_i)^2 / \sum(y_i - y_{mean})^2) \qquad \text{Eq12}$$

y_i : is the measured value of the test set compound,$\hat{Y}_i$ is the predicted value of the test set compounds and y_{mean}: is the average activity of all molecules of the training set. The following statistical parameters are also suggested: (1) correlation coefficient R between the predicted and observed values; (2) coefficients of determination (predicted versus observed activities R^2And observed versus predicted activitiesR'^2); (3) slopes k and k' of the regression lines through the origin.

II.6. Applicability domain

II.6.1. Definition

When a QSAR model is used, the applicability domain is defined as the response and chemical structure space in which the model generates predictions with a particular level of confidence. In this definition, Chemical structure may be represented by fragmental and/or physicochemical information, and the dependent variable can be any predicted environmental, biological, or physicochemical consequence. The applicability domain concept demands users to identify the model's limits in terms of its structural domain and response space [110].

In general, valid QSAR predictions are confined to compounds that are structurally similar to those used to develop that model. The Compounds that fall within the scope of the model are regarded to be inside the AD and are classed as interpolated, whilst the remainder is extrapolations and are thus labelled as outside the AD. The reliability of a particular model is stronger for predictions that fall within the AD, and it is more likely to be inaccurate for extrapolations that go outside the AD. The degree to which a predictive model is generalizable is determined by how extensive the realm of application is for the model. It is possible that the domain is too narrow, which means that the model is only capable of making solid predictions for a small number of chemical structures [111].

II.6.2. The importance of the AD in the (Q)SAR life cycle

The AD plays a crucial role in all stages of the QSAR life-cycle (development, validation and application), as it is depicted in Figure II.8**.** The concept should be used during model building to guarantee that a domain is described as widely as feasible in order to achieve the required degree of predictability. In general, when it comes to developing a model, there are two options: either design a model that can be used broadly and sacrifice some predictability or construct a model that can be applied specifically (for example, a certain class of chemicals) and have a higher degree of predictability.

The AD is significant during QSAR validation because it allows a predetermined AD to be checked and maybe adjusted. If external validation is conducted, Predictions must be made for samples that are not included in the training set. It is necessary to verify that the external "validation" series is adequate for model validation by ensuring that the test chemical structures fall inside the AD of the model. One of the primary reasons for having a well-defined AD is to make it easier to apply QSARs to specific compounds in the regulatory process. The choice to employ a QSAR for regulatory reasons will often need to determine if the studied compounds fall within the AD of the model [112].

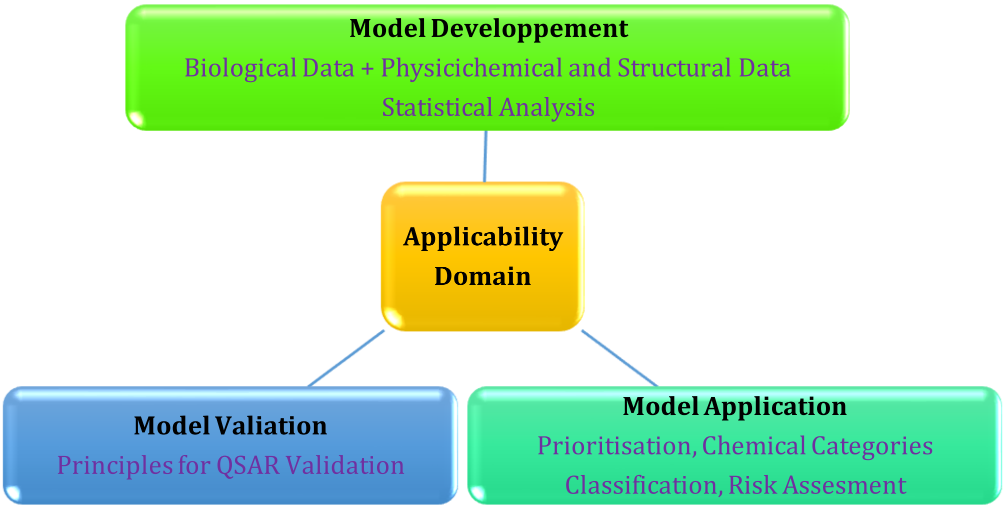

Model Developpement
Biological Data + Physicichemical and Structural Data
Statistical Analysis
Applicability Domain
Model Valiation
Principles for QSAR Validation
Model Application
Prioritisation, Chemical Categories
Classification, Risk Assesment

Figure II.8: The applicability domain's essential role in different phases of QSAR

II.6.3. Methods for Applicability Domain Determination

There are many different approaches for QSAR Applicability Domain determination. The following are the methodologies that are most often used for estimating the interpolation zones in a multivariate space: (1) Ranges in the descriptor space and (2) Distance-based methods. (3) Geometrical methods, (4). Probability density distribution, and (5) Range of the response variable. The first four techniques in this section are based on the methodology used for the characterization of interpolation space in model descriptor space. The latter, on the other hand, is exclusively dependent on the response space of the training set molecules. Figure II.9 shows the available approaches for calculating AD.

There is no way that can be deemed generally optimal among the present methods. Each strategy has advantages and disadvantages. The leverage technique (William’s plot), DModX, and other similarity evaluation approaches have been widely used in QSAR articles over the past decade to detect outliers and compounds lying beyond the AD zone [113].

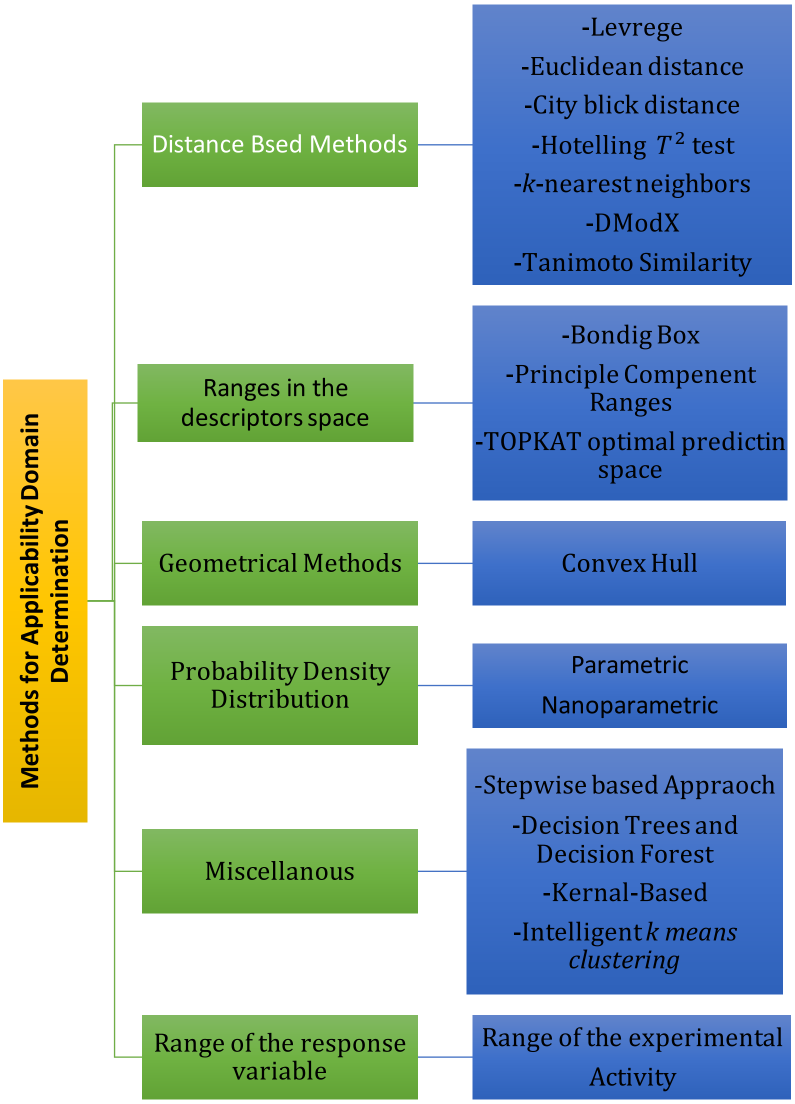

Methods for Applicability Domain Determination
Distance Bsed Methods
-Levrege
-Euclidean distance
-City blick distance
-Hotelling T^2 test
-*k*-nearest neighbors
-DModX
-Tanimoto Similarity
Ranges in the descriptors space
-Bondig Box
-Principle Compenent Ranges
-TOPKAT optimal predictin space
Geometrical Methods
Convex Hull
Probability Density Distribution
Parametric
Nanoparametric
Miscellanous
-Stepwise based Appraoch
-Decision Trees and Decision Forest
-Kernal-Based
-Intelligent *k means clustering*
Range of the response variable
Range of the experimental Activity

Figure II.9: Methods for Applicability Domain Determination

References

[1] Veselovsky, A. V., & Ivanov, A. S. (2003). Strategy of computer-aided drug design. *Current Drug Targets-Infectious Disorders*, *3*(1), 33-40.

[2] Sebastián-Pérez, V., Martínez, M. J., Gil, C., Campillo, N. E., Martínez, A., & Ponzoni, I. (2018, May). QSAR modelling for drug discovery: predicting the activity of LRRK2 inhibitors for parkinson's disease using cheminformatics approaches. In *International Conference on Practical Applications of Computational Biology & Bioinformatics* (pp. 63-70). Springer, Cham.

[3] Hansch, C. (1969). Quantitative approach to biochemical structure-activity relationships. *Accounts of chemical research*, *2*(8), 232-239.

[4] Kubinyi, H. (1977). Quantitative structure-activity relations. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *Journal of medicinal chemistry*, *20*(5), 625-629.

[5] Ghanem, O. B., Mutalib, M. A., Lévêque, J. M., & El-Harbawi, M. (2017). Development of QSAR model to predict the ecotoxicity of Vibrio fischeri using COSMO-RS descriptors. *Chemosphere*, *170*, 242-250.

[6] Toropov, A. A., Toropova, A. P., Marzo, M., Dorne, J. L., Georgiadis, N., & Benfenati, E. (2017). QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environmental toxicology and pharmacology*, *53*, 158-163.

[7] Mon, J., Flury, M., & Harsh, J. B. (2006). A quantitative structure–activity relationships (QSAR) analysis of triarylmethane dye tracers. *Journal of hydrology*, *316*(1-4), 84-97.

[8] Liu, Y., Guo, Y., Wu, W., Xiong, Y., Sun, C., Yuan, L., & Li, M. (2019). A machine learning-based QSAR model for benzimidazole derivatives as corrosion inhibitors by incorporating comprehensive

feature selection. *Interdisciplinary Sciences: Computational Life Sciences*, *11*(4), 738-747.

[9] G Damale, M., N Harke, S., A Kalam Khan, F., B Shinde, D., & N Sangshetti, J. (2014). Recent advances in multidimensional QSAR (4D-6D): a critical review. *Mini reviews in medicinal chemistry*, *14*(1), 35-55.

[10] Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design-a review. *Current topics in medicinal chemistry*, *10*(1), 95-115.

[11] Karthikeyan, K., Sivakumar, P. M., Doble, M., & Perumal, P. T. (2010). Synthesis, antibacterial activity evaluation and QSAR studies of novel dispiropyrrolidines. *European journal of medicinal chemistry*, *45*(8), 3446-3452.

[12] Abuhammad, A., & Taha, M. O. (2016). QSAR studies in the discovery of novel type-II diabetic therapies. *Expert opinion on drug discovery*, *11*(2), 197-214.

[13] Scotti, L., Jaime Bezerra Mendonca Junior, F., Rodrigo Magalhaes Moreira, D., Sobral da Silva, M., R Pitta, I., & Tullius Scotti, M. (2012). SAR, QSAR and docking of anticancer flavonoids and variants: a review. *Current topics in medicinal chemistry*, *12*(24), 2785-2809.

[14] Lemaoui, T., Hammoudi, N. E. H., Alnashef, I. M., Balsamo, M., Erto, A., Ernst, B., & Benguerba, Y. (2020). Quantitative structure properties relationship for deep eutectic solvents using Sσ-profile as molecular descriptors. *Journal of Molecular Liquids*, *309*, 113165.

[15] M. SAIHI Youcef Etude de la relation quantitative structure-activité inhibitrice des enzymes hydrolytiques : cas des alpha-glucosidasesThesis.

[16] Zahouily, Mohamed, et al. "QSAR for anti-malarial activity of 2-aziridinyl and 2, 3-bis (aziridinyl)-1, 4-naphthoquinonyl sulfonate and acylate derivatives." *Journal of Molecular Modeling* 12.4 (2006): 398-405.

[17] HU, S. Q., HU, J. C., SHI, X., ZHANG, J., & GUO, W. Y. (2009). QSAR and molecular design of imidazoline derivatives as corrosion inhibitors. *Acta Physico-Chimica Sinica*, *25*(12), 2524-2530.

[18] Shukla, A., Sharma, P., Prakash, O., Singh, M., Kalani, K., Khan, F., ... & Srivastava, S. K. (2014). QSAR and docking studies on capsazepine derivatives for immunomodulatory and anti-inflammatory activity. *PLoS One*, *9*(7), e100797.

[19] Darnag, R., Minaoui, B., & Fakir, M. (2017). QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry*, *10*, S600-S608.

REFERENCES

[20] Adeniji, S. E., Shallangwa, G. A., Arthur, D. E., Abdullahi, M., Mahmoud, A. Y., & Haruna, A. (2020). Quantum modelling and molecular docking evaluation of some selected quinoline derivatives as anti-tubercular agents. *Heliyon*, *6*(3), e03639.

[21] Abdulrahman, H. L., Uzairu, A., & Uba, S. (2021). QSAR, ligand-based design and pharmacokinetic studies of parviflorons derivatives as anti-breast cancer drug compounds against MCF-7 cell line. *Chemistry Africa*, *4*(1), 175-187.

[22] Arthur, D. E., Ejeh, S., & Uzairu, A. (2020). Quantitative structure-activity relationship (QSAR) and design of novel ligands that demonstrate high potency and target selectivity as protein tyrosine phosphatase 1B (PTP 1B) inhibitors as an effective strategy used to model anti-diabetic agents. *Journal of Receptors and Signal Transduction*, *40*(6), 501-520.

[23] Beglari, M., Goudarzi, N., Shahsavani, D., Arab Chamjangali, M., & Dousti, R. (2020). QSAR modeling of anti-HIV activity for DAPY-like derivatives using the mixture of ligand-receptor binding information and functional group features as a new class of descriptors. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1), 1-15.

[24] 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.

[25] Hammoudi, N. E. H., Benguerba, Y., Attoui, A., Hognon, C., Lemaoui, T., Sobhi, W., Benaicha., Badawi, K., & Monari, A. (2020).In silico drug discovery of IKK-β inhibitors from 2-amino-3-cyano-4-alkyl-6-(2-hydroxyphenyl) pyridine derivatives based on QSAR, docking, molecular dynamics and drug-likeness evaluation studies. *Journal of Biomolecular Structure and Dynamics*, 1-17.

[26] Lemaoui, T., Darwish, A. S., Attoui, A., Hatab, F. A., Hammoudi, N. E. H., Benguerba, Y., ... & Alnashef, I. M. (2020). Predicting the density and viscosity of hydrophobic eutectic solvents: Towards the development of sustainable solvents. *Green Chemistry*, *22*(23), 8511-8530

[27] Zaki, M. E., Al-Hussain, S. A., Masand, V. H., Akasapu, S., Bajaj, S. O., El-Sayed, N. N., ... & Lewaa, I. (2021). Identification of Anti-SARS-CoV-2 compounds from food using QSAR-based virtual screening, molecular docking, and molecular dynamics simulation analysis. *Pharmaceuticals*, *14*(4), 357.

[28] Andrade, C. H., Pasqualoto, K. F., Ferreira, E. I., & Hopfinger, A. J. (2010). 4D-QSAR: perspectives in drug design. *Molecules*, *15*(5), 3281-3294.

[29] Polanski, J. (2009). Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. *Current medicinal chemistry*, *16*(25), 3243-3257.

[30] ERRAHOUI née BELLIFA KHADIDJA Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires.Modélisation QSAR thesis 08/10/2015

[31] Chen, B., Zhang, T., Bond, T., & Gan, Y. (2015). Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of methods and resources. *Journal of hazardous materials*, *299*, 260-279.

[32] Shahlaei, M. (2013). Descriptor selection methods in quantitative structure–activity relationship studies: a review study. *Chemical reviews*, *113*(10), 8093-8103.

[33] Thi Ngoc Phuong Huynh Synthèse et études des relations structure/activitéquantitatives (QSAR/2D) d'analyse benzo[c]phénanthridiniques thesis 29 /06/2007

[34] Khan, P. M., & Roy, K. (2018). Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert opinion on drug discovery*, *13*(12), 1075-1089.

[35] Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug discovery today*, *21*(8), 1291-1302.

[36] Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). Application of topological descriptors in QSAR and drug design: history and new trends. *Current Drug Targets-Infectious Disorders*, *2*(1), 93-102.

[37] Burden, F. R. (1989). Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences*, *29*(3), 225-227.

[38] Diudea, M. V. (1997). Indices of reciprocal properties or Harary indices. *Journal of chemical information and computer sciences*, *37*(2), 292-299.

[39] Pisanski, T., & Shawe-Taylor, J. (2000). Characterizing graph drawing with eigenvectors. *Journal of Chemical Information and Computer Sciences*, *40*(3), 567-571.

[40] Burden, F. R. (1989). Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences*, *29*(3), 225-227.

[41] Roy, K. (2004). Topological descriptors in drug design and modeling studies. *Molecular Diversity, 8*(4), 321-323.

[42] Adnan, M., Bokhary, S. A. U. H., Abbas, G., & Iqbal, T. (2022). Degree-Based Topological Indices and QSPR Analysis of Antituberculosis Drugs. *Journal of Chemistry*, *2022*.

[43] Dearden, J. C. (2017). The use of topological indices in QSAR and QSPR modeling. In *Advances in QSAR modeling* (pp. 57-88). Springer, Cham.

[44] Karelson, M., Lobanov, V. S., & Katritzky, A. R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical reviews*, *96*(3), 1027-1044.

[45] Thanikaivelan, P., Subramanian, V., Rao, J. R., & Nair, B. U. (2000). Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chemical Physics Letters*, *323*(1-2), 59-70.

[46] Karabunarliev, S., Mekenyan, O. G., Karcher, W., Russom, C. L., & Bradbury, S. P. (1996). Quantum-chemical Descriptors for Estimating the Acute Toxicity of Electrophiles to the Fathed minnow (Pimephales promelas): An Analysis Based on Molecular Mechanisms. *Quantitative Structure-Activity Relationships*, *15*(4), 302-310.

[47] Hemmateenejad, B., & Sanchooli, M. (2007). Substituent electronic descriptors for fast QSAR/QSPR. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *21*(3-4), 96-107.

[48] Parthasarathi, R., Subramanian, V., Roy, D. R., & Chattaraj, P. K. (2004). Electrophilicity index as a possible descriptor of biological activity. *Bioorganic & medicinal chemistry*, *12*(21), 5533-5543.

[49] Rutkowska, E., Pajak, K., & Jóźwiak, K. (2013). Lipophilicity--methods of determination and its role in medicinal chemistry. *Acta poloniae pharmaceutica*, *70*(1), 3-18.

[50] Giaginis, C., & Tsantili-Kakoulidou, A. (2007). Current state of the art in HPLC methodology for lipophilicity assessment of basic drugs. A review. *Journal of liquid chromatography & related technologies*, *31*(1), 79-96.

[51] Lobo, S. (2020). Is there enough focus on lipophilicity in drug discovery?. *Expert opinion on drug discovery*, *15*(3), 261-263.

[52] Waring, M. J. (2010). Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery*, *5*(3), 235-248.

[53] Arnott, J. A., & Planey, S. L. (2012). The influence of lipophilicity in drug discovery and design. *Expert opinion on drug discovery*, *7*(10), 863-875.

[54] Dearden, J. C., & Bresnen, G. M. (1988). The measurement of partition coefficients. *Quantitative Structure-Activity Relationships*, *7*(3), 133-144.

[55] Barzanti, C., Evans, R., Fouquet, J., Gouzin, L., Howarth, N. M., Kean, G., ... & Kraft, A. (2007). Potentiometric determination of octanol–water and liposome–water partition coefficients (log P) of ionizable organic compounds. *Tetrahedron letters*, *48*(19), 3337-3341.

[56] Steyaert, G., Lisa, G., Gaillard, P., Boss, G., Reymond, F., Carrupt, P. A., & Testa, B. (1997). Intermolecular forces expressed in 1, 2-dichloroethane–water partition coefficients. *Journal of the Chemical Society, Faraday Transactions*, *93*(3), 401-406.

[57] Huuskonen, J., Livingstone, D. J., & Manallack, D. T. (2008). Prediction of drug solubility from molecular structure using a drug-like training set. *SAR and QSAR in Environmental Research*, *19*(3-4), 191-212.

[58] Hou, T. J., Xia, K., Zhang, W., & Xu, X. J. (2004). ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *Journal of chemical information and computer sciences*, *44*(1), 266-275.

[59] Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Forward selection of explanatory variables. *Ecology*, *89*(9), 2623-2632.

[60] Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265-282.

[**61]** Xu, L., & Zhang, W. J. (2001). Comparison of different methods for variable selection. *Analytica Chimica Acta*, *446*(1-2), 475-481.

[62] Liu, S. S., Cui, S. H., Shi, Y. Y., & Wang, L. S. (2002). A novel variable selection and modeling method based on the prediction for QSAR of cyclooxygenase-2 inhibition by thiazolone and oxazolone series. *Internet Electron J Mol Des*, *1*, 610-619.

[63] Araújo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and intelligent laboratory systems*, *57*(2), 65-73.

[64] Daszykowski, M., Stanimirova, I., Walczak, B., Daeyaert, F., De Jonge, M. R., Heeres, J., ... & Massart, D. L. (2005). Improving QSAR models for the biological activity of HIV Reverse Transcriptase inhibitors: Aspects of outlier detection and uninformative variable elimination. *Talanta*, *68*(1), 54-60.

[65] Ojha, P. K., & Roy, K. (2011). Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection. *Chemometrics and Intelligent Laboratory Systems*, *109*(2), 146-161.

[66] Lucasius, C. B., & Kateman, G. (1993). Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and intelligent laboratory systems*, *19*(1), 1-33.

[67] Lin, S. W., Lee, Z. J., Chen, S. C., & Tseng, T. Y. (2008). Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, *8*(4), 1505-1512.

[68] Lü, J. X., Shen, Q., Jiang, J. H., Shen, G. L., & Yu, R. Q. (2004). QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. *Journal of pharmaceutical and biomedical analysis*, *35*(4), 679-687.

[69] Wang, Z., Durst, G. L., Eberhart, R. C., Boyd, D. B., & Miled, Z. B. (2004, April). Particle swarm optimization and neural network application for QSAR. In *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.* (p. 194). IEEE.

[70] Burden, F. R., Ford, M. G., Whitley, D. C., & Winkler, D. A. (2000). Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *Journal of Chemical Information and Computer Sciences*, *40*(6), 1423-1430.

[71] Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening*, *9*(3), 213-228.

[72] Leonard, J. T., & Roy, K. (2006). QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. *Bioorganic & medicinal chemistry*, *14*(4), 1039-1046.

[73] Valle, S., Li, W., & Qin, S. J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, *38*(11), 4389-4401.

[74] Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., & Clementi, S. (1993). Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quantitative Structure-Activity Relationships*, *12*(1), 9-20.

[75] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

[76] Rogers, D. (1996). Some theory and examples of genetic function approximation with comparison to evolutionary techniques. In *Genetic algorithms in molecular modeling* (pp. 87-107). Academic Press.

[77] Niculescu, S. P. (2003). Artificial neural networks and genetic algorithms in QSAR. *Journal of molecular structure: THEOCHEM*, *622*(1-2), 71-83.

[78] Baskin, I. I., Ait, A. O., Halberstam, N. M., Palyulin, V. A., & Zefirov, N. S. (2002). An approach to the interpretation of backpropagation neural network models in QSAR studies. *SAR and QSAR in Environmental Research*, *13*(1), 35-41.

[79] MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural computation*, *4*(3), 448-472.

[80] Cui, W., & Yan, X. (2009). Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. *Chemometrics and Intelligent Laboratory Systems*, *98*(2), 130-135.

[81] Andres, C., & Hutter, M. C. (2006). CNS permeability of drugs predicted by a decision tree. *QSAR & Combinatorial Science*, *25*(4), 305-309.

[82] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[83] Fox, T., & Kriegl, J. M. (2006). Machine learning techniques for in silico modeling of drug metabolism. *Current Topics in Medicinal Chemistry*, *6*(15), 1579-1591.

[84] Kauffman, G. W., & Jurs, P. C. (2001). QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *Journal of chemical information and computer sciences*, *41*(6), 1553-1560

[85] Zhaobin,Z., Yue, L., Zhu, L., Shan, L., Multiple Linear Regression for High Efficiency Video Intra Coding (2019). ICASSP 2019–2019 IEEE.

[86] Fedotova, O., Teixeira, L., & Alvelos, H. (2013). Software Effort Estimation with Multiple Linear Regression: Review and Practical Application. *J. Inf. Sci. Eng.*, *29*(5), 925-945.

[87] Abdellatif, T. M. (2018, May). A comparison study between soft computing and statistical regression techniques for software effort estimation. In *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)* (pp. 1-5). IEEE.

[88] Jørgensen, M. (2004). Regression models of software development effort estimation accuracy and bias. *Empirical Software Engineering*, *9*(4), 297-314.

[89] Ott, R. L., & Longnecker, M. T. (2015). *An introduction to statistical methods and data analysis*. Cengage Learning.

[90] Wu, J., Mei, J., Wen, S., Liao, S., Chen, J., & Shen, Y. (2010). A self‐adaptive genetic algorithm‐artificial neural network algorithm with leave‐one‐out cross validation for descriptor selection in QSAR study. *Journal of computational chemistry*, *31*(10), 1956-1968.

[91] Livingstone, D. J. (Ed.). (2008). *Artificial neural networks: methods and applications* (pp. 185-202). Totowa, NJ, USA: Humana Press.

[92] Singh, Y., & Chauhan, A. S. (2009). NEURAL NETWORKS IN DATA MINING. *Journal of Theoretical & Applied Information Technology*, *5*(1).

[93] Zakaria, M., Al-Shebany, M., & Sarhan, S. (2014). Artificial neural network: a brief overview. *International Journal of Engineering Research and Applications*, *4*(2), 7-12.

[94] Shahin, M. A., Jaksa, M. B., & Maier, H. R. (2001). Artificial neural network applications in geotechnical engineering. *Australian geomechanics*, *36*(1), 49-62.

[95] Roy, K., & Mitra, I. (2011). On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Combinatorial chemistry & high throughput screening*, *14*(6), 450-474.

[96] Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., & Tropsha, A. (2006). A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *Journal of chemical information and modeling*, *46*(5), 1984-1995.

[97] Damme, S. V., & Bultinck, P. (2007). A new computer program for QSAR‐analysis: ARTE‐QSAR. *Journal of Computational Chemistry*, *28*(11), 1924-1928.

[98] Pecka, J., & Ponec, R. (2000). Simple analytical method for evaluation of statistical importance of correlations in QSAR studies. *Journal of Mathematical Chemistry*, *27*(1), 13-22.

[99] Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design-a review. *Current topics in medicinal chemistry*, *10*(1), 95-115.

REFERENCES

[100] Kiralj, R., & Ferreira, M. (2009). Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *Journal of the Brazilian Chemical Society*, *20*(4), 770-787.

[101] Baumann, K., & Stiefl, N. (2004). Validation tools for variable subset regression. *Journal of computer-aided molecular design*, *18*(7), 549-562.

[102] Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives*, *111*(10), 1361-1375.

[103] Clark, R. D., & Fox, P. C. (2004). Statistical variation in progressive scrambling. *Journal of computer-aided molecular design*, *18*(7), 563-576.

[104] Clark, R. D., & Fox, P. C. (2004). Statistical variation in progressive scrambling. *Journal of computer-aided molecular design*, *18*(7), 563-576.

[105] Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov*, *3*, 511-519.

[106] Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18-33.

[107] Tropsha, A. (2005). Application of predictive QSAR models to database mining. *Chemoinformatics in Drug Discovery*, 437-455.

[108] Roy, K., & Mitra, I. (2011). On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Combinatorial chemistry & high throughput screening*, *14*(6), 450-474.

[109] Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Current pharmaceutical design*, *13*(34), 3494-3504.

[110] Weaver, S., & Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, *26*(8), 1315-1326.

[111] Nikolova-Jeliazkova, N., & Jaworska, J. (2005). An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *Alternatives to Laboratory Animals*, *33*(5), 461-470.

[112] Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., ... & Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals*, *33*(2), 155-173.

[113] Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791-4810.

CHAPTER III: Basics in Molecular Docking

CHAPTER III: BASICS IN MOLECULAR DOCKING STUDY

III.1. Introduction

A growing number of new therapeutic targets for use in drug discovery have been identified since the completion of the human genome project. Simultaneously, thanks to the following advances: high-throughput protein purification, nuclear magnetic resonance, crystallography, nuclear magnetic resonance, and spectroscopy.

Numerous structural details of proteins and protein-ligand complexes have been revealed. These technological advancements make it possible for computational strategies to permeate every aspect of drug development today, including the virtual screening (VS) techniques for hit identification and the methods for lead optimization [1-6]. VS is a more direct and rational approach to drug discovery when compared to the conventional experimental high throughput screening (HTS) method. Additionally, VS has the advantage of being low-cost effective and screening more compounds [6.7].

VS is categorized into structure-based and ligand-based methods. The application of ligand-based methods, such as pharmacophore modeling and quantitative structure-activity relationship (QSAR) methods, is possible in situations in which a collection of active ligand molecules is already known, but there is either very little or no structural information available for the targets. Regarding structure-based drug design, the molecular docking technique is by far the most common and has been in widespread use ever since the early 1980s [8].

The first strategy of ligand-receptor binding research was suggested for the first time by Fisher. Figure III displays the lock-and-key model, which corresponds to rigid docking. For the planning and design of new drugs, the use of molecular docking methodologies is of the utmost importance. The goal of these methods is to make a prediction about the experimental binding mode and affinity of small compounds incorporated in the binding site of an important receptor target [9].

Molecular docking is a computer simulation procedure that consists of three principal objectives: virtual screening, pose prediction, and binding affinity estimation of a receptor-ligand complex where the ligand is either a small molecule or another protein. The receptor can be a protein or a nucleic acid molecule (DNA or RNA). Another way to define it is as a simulation process in which the position of a ligand is estimated within a binding site that has either been predicted or pre-defined [8.9].

Molecular docking simulations can also be utilized to reproduce experimental data employing docking validation algorithms. The obtained in silico confirmations of the studied complex (protein-ligand or protein-protein) will be compared to the chemical structures found from X-ray crystallography or nuclear magnetic resonance [8].

There are two basic steps in docking: predicting the ligand conformation and its position and orientation within these sites (often referred to as pose) and evaluating the binding affinity. These two steps are related to sampling methods and scoring schemes [9-11].

A reliable docking methodology is capable of accurately estimating the pose of the incorporated ligand in the receptor-binding site. (Achieve an acceptable level of precision in terms of the experimental ligand geometry achieved) as well as the associated physical-chemical molecular interactions. In addition, when researching large compound libraries, the method needs to be able to successfully differentiate binding compounds from nonbinding compounds and correctly rank these ligands among the most effective compounds in the database [12]. These tasks are carried out by molecular docking programs in the form of a cyclical process, during which the conformation of the ligand is evaluated based on a set of specific scoring functions. This process is repeated in a recursive manner until [13].

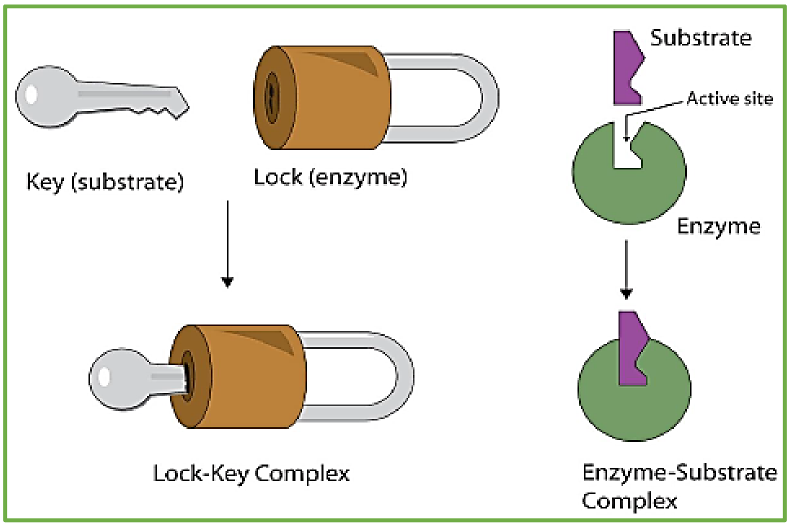
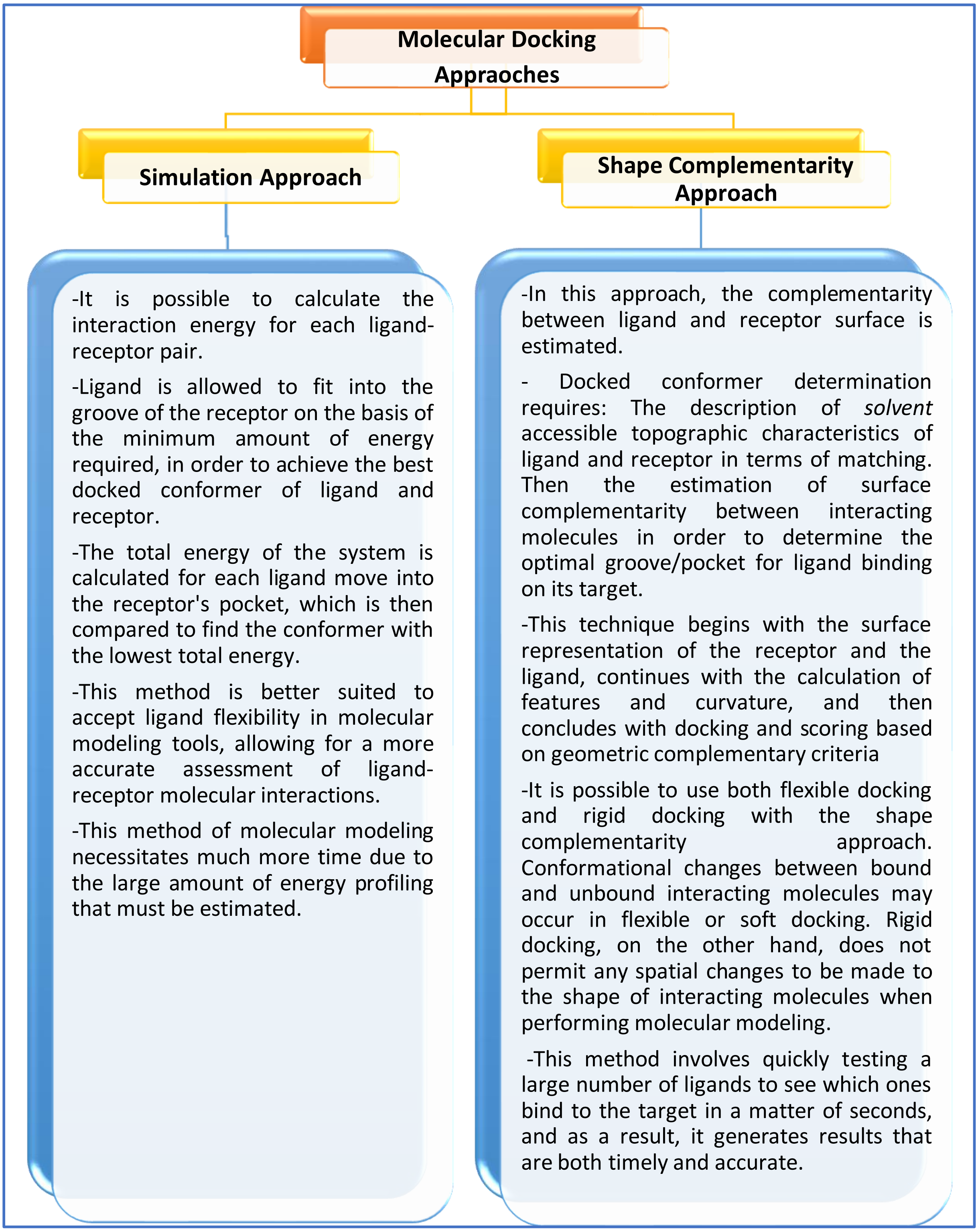


Figure III.1: Enzyme Activity Model Lock-and-Key [14]

III .2. Approaches of Molecular Docking

There are primarily two different kinds of approaches that are utilized when carrying out molecular docking. Computer simulations are used in one of the approaches; during these simulations, energy profiling for the ligand-target conformer that is docked is estimated. In contrast, the second approach makes use of a method that determines surface complementarity between ligand and target. The main properties of the two approaches are presented in Figure III.2 [15-18].



Molecular Docking Appraoches

Simulation Approach

-It is possible to calculate the interaction energy for each ligand-receptor pair.

-Ligand is allowed to fit into the groove of the receptor on the basis of the minimum amount of energy required, in order to achieve the best docked conformer of ligand and receptor.

-The total energy of the system is calculated for each ligand move into the receptor's pocket, which is then compared to find the conformer with the lowest total energy.

-This method is better suited to accept ligand flexibility in molecular modeling tools, allowing for a more accurate assessment of ligand-receptor molecular interactions.

-This method of molecular modeling necessitates much more time due to the large amount of energy profiling that must be estimated.

Shape Complementarity Approach

-In this approach, the complementarity between ligand and receptor surface is estimated.

- Docked conformer determination requires: The description of *solvent* accessible topographic characteristics of ligand and receptor in terms of matching. Then the estimation of surface complementarity between interacting molecules in order to determine the optimal groove/pocket for ligand binding on its target.

-This technique begins with the surface representation of the receptor and the ligand, continues with the calculation of features and curvature, and then concludes with docking and scoring based on geometric complementary criteria

-It is possible to use both flexible docking and rigid docking with the shape complementarity approach. Conformational changes between bound and unbound interacting molecules may occur in flexible or soft docking. Rigid docking, on the other hand, does not permit any spatial changes to be made to the shape of interacting molecules when performing molecular modeling.

-This method involves quickly testing a large number of ligands to see which ones bind to the target in a matter of seconds, and as a result, it generates results that are both timely and accurate.

Figure III.2: Molecular Docking Approaches and Their Main Properties

III.3. Applications of Molecular Docking

Molecule docking is a pressing issue in today's scientific endeavors. If it is carried out prior to the experimental part of any investigation, it can demonstrate the feasibility of any task. Molecular docking has revolutionized research in some areas. In particular, ligand-protein interactions can be used to predict whether an enzyme can be activated or inhibited. This kind of information has the potential to serve as a source of raw material for rational drug design. Some of the major applications of molecular docking are presented in Figure III.3 [19].

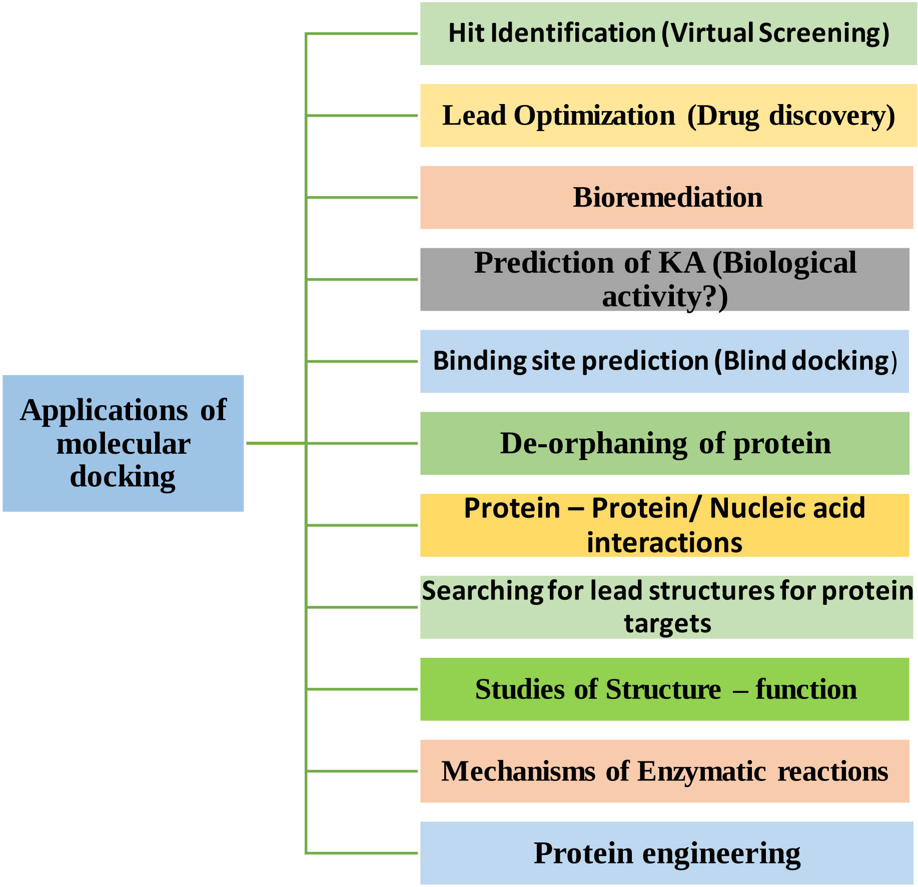

Applications of molecular docking
Hit Identification (Virtual Screening)
Lead Optimization (Drug discovery)
Bioremediation
Prediction of KA (Biological activity?)
Binding site prediction (Blind docking)
De-orphaning of protein
Protein – Protein/ Nucleic acid interactions
Searching for lead structures for protein targets
Studies of Structure – function
Mechanisms of Enzymatic reactions
Protein engineering

Figure III.3: Molecular Docking Applications

III.4. Theory of docking

Docking can be accomplished through the completion of two steps that are intertwined with one another. The first step involves collecting conformations of the ligand while it is bound to the active site of the protein. The second step involves using a scoring function to rank the collected conformations [20].

III.4.1. Search Algorithms

Search algorithms are utilized in the process of molecular docking in order to investigate the free energy landscape and locate the most favorable ligand poses. When the entropic and enthalpic effects of the system are accurately modelled by the energy function, it is expected that the experimental receptor-ligand conformation will correspond to the global minimum of the energy landscape. Regrettably, taking into account the entropic effects is not a straightforward process, and the docking methods that are currently in use employ rough approximations. As a result, there is no assurance that the native binding mode will correspond to the global minimum that is associated with the energy landscape that was investigated by a docking methodology [21].

Docking methods make use of the following three primary strategies with regard to the flexibility of proteins and ligands:

1. the protein and ligand are considered to be rigid (i.e., It is assumed that the ligand is a rigid body that does not possess any internal degrees of freedom).
2. the protein is rigid, and all degrees of freedom of the ligand are explored (i.e., Translational, rotational, and conformational)
3. The protein is considered totally or partially flexible, and all degrees of freedom of the ligand are also investigated.

III.4.1.1. Ligand Sampling

The vast majority of algorithms treat the protein as if it were a rigid object, whereas the ligand is treated as though it were flexible. This means that sp3 bonds are allowed to spin, but bond lengths and bond angles are assumed to remain the same [21.22].

In this chapter, we review the main strategies associated with the second and the third approaches. Systematic, stochastic, and deterministic searches are the three primary categories that serve as the basis for classifying search algorithms. This categorization is based on the approach that is used to investigate ligand flexibility. In addition, some algorithms use a hybrid technique, which combines two or all three of these strategies. During the search process, systematic algorithms investigate all of the ligand's degrees of freedom [22].

III.4.1.1.1 Systematic Methods

Methods that take this approach can be further categorized as exhaustive, incremental construction, or conformational ensemble.

a) Exhaustive searches

Exhaustive searches systematically explore the values of each degree of freedom in a combinatorial manner, rotating all dihedral angles of the ligand according to a predetermined range of values and a set of initial restraints, e.g., geometrical and chemical constraints. It stands to reason that, ligands that are more flexible will have a greater number of rotatable bonds; as a result, the complexity of the optimization problem will be greatly increased [23].

b) Incremental Construction

The method of fragmentation known as incremental construction begins with the separation of the ligand into smaller fragments. Next, a base fragment is chosen and docked into the binding site of the receptor. After that, the ligand is pieced back together one fragment at a time by forming covalent bonds between the various other pieces and the base group. This method is utilized quite frequently in de novo ligand creation, which seeks to find new compounds by joining the segments that dock in the receptor-binding site in the best possible positions [24.25.26].

III.4.1.1.2 Stochastic methods

In Stochastic methods, every time a ligand's translational, rotational, and conformational (degrees of freedom) are randomly changed, at each step wide range of possible outcomes is generated. There are several different classes of stochastic algorithms, including Monte Carlo (MC), Tabu Search (TS), Swarm Optimisation (SO), and Evolutionary Algorithms (EAs) [27]. The fact that there is no assurance that these heuristic methods will converge on the optimal solution is the main associated disadvantage. In addition, multiple and independent iterations of the algorithm are essential for the probability of locating the global energy minimum maximization.

MC Methods performs modifications in all of the ligand degrees of freedom that are, and energy minimization is often performed for each produced conformation. After then, the poses are judged to be accepted or not based on the Boltzmann factor, which takes into account the energies of the solutions before and after the random modifications as well as the absolute temperature. Because raising the temperature may make it possible for some energy barriers to be overcome, one significant version of this process, which is known as Simulated Annealing, employs temperature fluctuations in order to boost the likelihood of locating the global minimum [28].

EAs are a type of optimization approach that is based on the notion of the evolution of biological populations through the process of natural selection. The goal of these methods is to find the most effective solutions to a specific issue. These methods are categorized into three main classes: evolutionary programming, evolution strategies, and genetic algorithms. The advantage of these methods allows to quickly get out of local minima and identify many different low-energy solutions at the same time [29.30].

The TS strategy investigates the search space by making arbitrary adjustments to the ligand's degrees of freedom. In addition to preserving the solution with the least amount of energy, this technique employs “Tabu list” which contains the lowest energy solutions previously found, for the similarity of a recently generated, non-lowest energy pose analysis [31].

The SO method takes its cues from the cooperative actions of organisms like birds and ants. According to this approach, any adjustments that are made to an existing solution will be made so that it more closely resembles the optimal pose of the population. The SO method can be implemented in a few different ways, two of which are the Particle Swarm Optimisation and the Ant Colony Optimisation [21].

III.4.1.1.3 Deterministic Methods

In deterministic methods, the actual state of the system determines the modifications to be made, leading to its next state. The final result is highly dependent on the initial input structure because, given exactly the same initial system configuration and a particular set of parameters, the final state will always be the same [32].

III.4.1.2. Protein flexibility

In the realm of molecular docking, one of the ongoing challenges is to account for the flexibility of proteins throughout the searching process. When ligands bind to proteins in solution, there can be minor to substantial conformational changes in the protein as a result of induced fit effects. In addition, it is now generally acknowledged that proteins do not reside in a single native configuration but rather in a collection of many conformations. Based on this theory, the binding of the ligand to one of these conformations shifts the population equilibrium towards a specific ligand-bound conformation [33.34.35].

It is, therefore, a very difficult task to build search algorithms and sampling procedures that adequately deal with the several degrees of freedom associated with protein flexibility. Taking into consideration protein flexibility has necessitated the development of a variety of approaches, the most important of which may be categorized into the following five categories: collective degrees of freedom, soft docking, molecular relaxation, ensemble docking, and side-chain flexibility. Small protein motions can have a big impact on molecular docking outcomes [36].

It is possible for the side chain of a single amino acid residue to rearrange itself in a way that will change the profile of the binding site. This will result in a reduction in the accuracy of the docking process during posture prediction as well as binding affinity estimate.

III.4.1.2.1. The Soft Docking Method

In order to take into account these relatively minor conformational (i.e., atomic variations of up to 1 angstrom), the soft docking method has been applied. In a more general sense, this strategy reduces the repulsive term of the Lennard-Jones potential, which makes it possible for there to be modest overlaps between the atoms of the ligand and the protein. The main advantage of this method is its speed, but it should only be employed to consider local receptor motions. The quickness of this procedure is its primary benefit; nonetheless, it shouldn't be considered for anything other than the motions of local receptors [37.38.39].

III.4.1.2.2. Side-Chain Flexibility Method

It is possible to use the side-chain flexibility approach in situations in which the generated fit effect is significant and goes beyond the soft-docking limits (but continues to be a local effect). Using this strategy, the backbone of the protein is held steady while various conformations of the side chains are investigated by adjusting the important torsional degrees of freedom of those chains [40].

III.4.1.2.3. Molecular Relaxation

Molecular relaxation is the third strategy for optimizing the conformations of protein-ligand complexes. This is accomplished through the use of rigid-protein docking methods, which take into account the motions of the protein's backbone and side chains in addition to all of the ligand's degrees of freedom. The methods of energy minimization, Monte Carlo methods, and molecular dynamics are the ones that are typically used, and they are typically tailored to treat receptor flexibility during docking. Ensemble docking implicitly considers the receptor flexibility by docking the ligand on a set of protein conformations instead of a single conformation [41].

III.4.1.2.4. Ensemble Docking

By docking the ligand on a group of protein conformations rather than a single conformation, ensemble docking takes into account the flexibility of the receptor. Experimental data (such as nuclear magnetic resonance (NMR) and X-ray crystallographic structures, for example) or computational techniques (such as comparison modeling, normal mode analysis, and molecular dynamics-derived frames, among others) can be used to derive distinct receptor states. The ensemble docking methodologies also vary in the manner in which the docking findings received from numerous structures are analyzed, such as the rescoring process and the prediction of the binding affinity [42.43.44.45].

III.4.1.2.5. Collective Degrees of Freedom

The procedures that are based on collective degrees of freedom take into account the complete protein's flexibility. In general, the goal of these methodologies is to decrease the original high-dimensional representation of protein motion that was used to a lower-dimensional version that simply accounts for the predominant motion modes. This method is used in several applications to take into consideration the flexibility of proteins, such as principal component analysis and normal mode analysis. The primary issue with these approaches is that the degrees of freedom that are taken into account are not the native ones; rather, it is the collective motions that are derived from them. This might result in mistakes in the prediction of the binding mode [21.46.47].

III.4.2. Scoring Functions

It is feasible to comprehend crucial aspects connected to the binding process if one predicts the binding mode and affinity of a small molecule within the binding site of an interesting target. This will allow one to better grasp how the binding process works. For the purposes of pose and affinity prediction, empirical scoring functions are utilized extensively. The main goals of the scoring function are [48]

- During the process of the docking procedure, the search algorithm analyses a large number of different conformations for each molecule in the compound library. At this stage, the scoring functions will evaluate the quality of the generated docking poses in order to direct the search strategies toward relevant ligand conformations.

-The classification of active and inactive compounds (VS).

-The prediction of the absolute binding affinity and the accurate ranking of compounds in accordance with their effectiveness.

III.4.2.1. Physics-Based Scoring Functions

Physics-based SFs include the scoring functions based on solvation models, force field, and quantum mechanics methods. The binding energy is calculated using the classical force field-based SF by adding up the van der Waals and electrostatic interactions between the protein-ligand atom pairs according to the following equation:

$$E_{bin} = E_{vdw} + E_{elec} \qquad \text{Eq.1}$$

It takes into account the role that enthalpy plays in the production of total energy. The performance of the force field-based SF is not enough because it does not take into account entropy and the solvent effect. Therefore, the force field-based SF is made better by the incorporation of the torsion entropy of ligands, as well as the solvation/desolvation impact and implicit solvent models Eq. 2.

$$E_{bin} = E_{vdw} + E_{elec} + E_{\Delta G} \quad \text{Eq.2}$$

However, because this kind of scoring function is based on the force field, the predictive accuracy for the binding energy is significantly affected by the functional form of the potential energy and related parameters, which are difficult to find. This is because of the fact that the force field is the basis for the scoring function. Recent research has developed the SF, which is based on quantum mechanics (QM), in order to address the difficulties of covalent contacts, polarization, and charge transfer in docking. On the other hand, the QM-based SF not only has a greater computing cost than the force field-based SF, but it also has a greater degree of precision than the force field-based SF. As a suggested solution for the computational cost and predictive accuracy comprising, a hybrid quantum mechanical and molecular mechanics (QM/MM) technique (Eq. 3) has been developed [49.50.51.52.53.54].

$$E_{bin} = E_{QM/MM} + E_{\Delta G} \quad \text{Eq.3}$$

III.4.2.2. Empirical Scoring Function

The second category of scoring functions is known as the empirical scoring function, and it has been increasingly popular in recent years. The binding energy can be stated as a weighted sum of explicit hydrogen bonding and hydrophobic contact components. There is a great deal of extra terminology, as well as other functional forms that have been selected. Some functions solely use distance terms to describe hydrogen bonds, but the vast majority of others take into account and penalize any angular deviation from the geometries of idealized hydrogen bonds. The weighting terms of these functions are typically derived by fitting to experimentally determined Ki values for protein-ligand complexes whose crystal structures have been determined for compilations of such data [58.59.60].

III.4.2.3. knowledge-based scoring function

A knowledge-based scoring function is one that is dependent on particular characteristics of known protein-protein interactions that are maintained in a database. In most cases, the score is determined by either chemical, physical, or biological characteristics. As ranking scoring functions, they have been utilized but with varying degrees of effectiveness. In the beginning, statistical potentials were developed in order to differentiate a right protein fold (also known as near-native) of a model from a multitude of generated solutions. It has been described and investigated how many different statistical possibilities there are. In order to evaluate the potential of contacts between protein molecules, specific potentials for the interactions between macromolecules were derived [60].

III.4.2.4. Machine Learning-Based

Machine learning-based represents the new type of scoring function for protein-ligand interaction evaluation. Machine learning can be employed when the characteristics of both the ligand and the protein, in addition to the patterns of their interactions, are capable of being encoded using specific descriptors. The results of this analysis can then be used to derive statistical models that compute protein-ligand binding scores. Figure III.4 depicts the typical process that is followed when training SF based on machine learning [49.61].

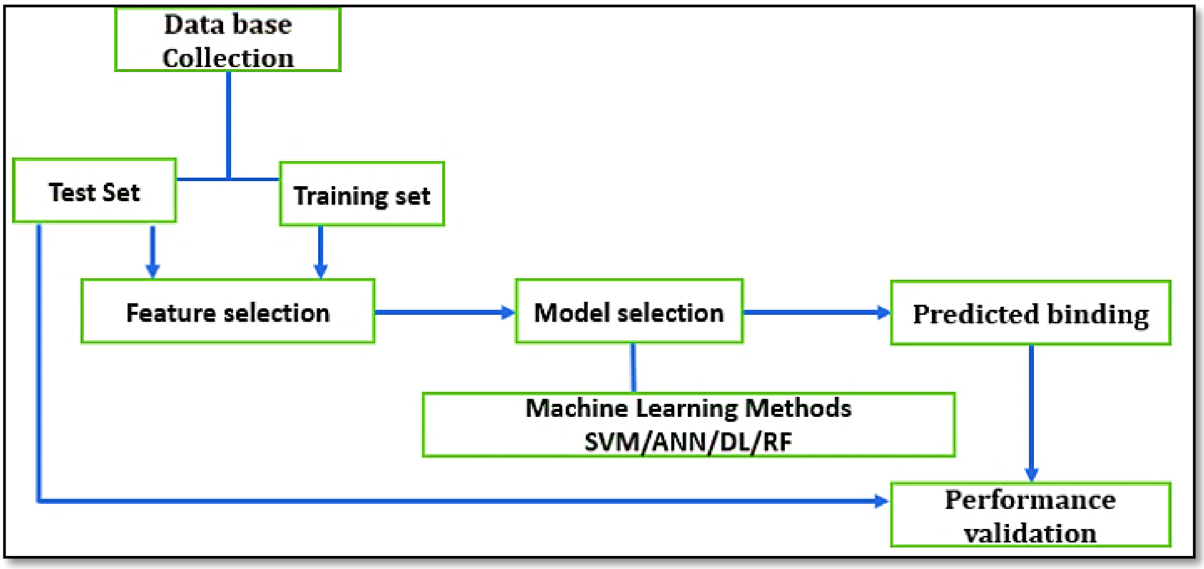


Figure III.4: Typical process that is followed when training SF based on machine learning.

III.5. Molecular Dynamics Simulation

When performing molecular docking, it is vital to take into account the flexibility of the target binding site; yet, this feature is commonly neglected. During the molecular recognition process, enzymes and receptors are capable of undergoing conformational modifications. In certain instances, these structural rearrangements are rather minor, and the ligand is able to fit into a binding site with little mobility. Otherwise, certain proteins will go through considerable conformational changes, which may involve aspects of their secondary and tertiary structures.

Problems of this nature with regard to flexibility are amenable to solutions with methods such as MD [62.63.64]. In most cases, the ligand is responsible for stabilizing a subset of the receptor's potential conformations, which moves the equilibrium closer to the structures with the lowest amount of energy [65]. In situations like these, MD simulations have the ability to develop alternate conformational states that correlate to the structures that are ligand-induced. In addition, in the case of the no existing of crystallographic structures for a given molecular target that are suitable, MD can be used to construct a set of docking structures that are convenient. In addition, MD can be utilized for the purpose of determining the stability of a ligand-receptor complex that was predicted via molecular docking [66.67]

In molecular dynamics, Newton's equations of motion, as outlined in classical mechanics, are utilized to determine the location and velocity of each atom in the system that is being investigated. As a direct consequence of this, it is possible to investigate the path that a ligand-receptor complex takes throughout time [68].

III.6. Molecular Docking Methodology

III.6.1. Ligands preparation

The concept of ligand preparation involves taking 2D structures and producing corresponding low-energy 3D structures.

III.6.2. Protein Preparation and grid generation

Removing water molecules, fill in messing loops and side chains, deleting the alternate conformations and adding hydrogen, and inserting the missing atoms in incomplete residues are elementary and indispensable steps that were used for the protein refinement and preparation. Then, receptor grid generation was performed in order to determine the position and size of the receptor active site area where binding interaction between the ligand and the residues can occur [69]

III.6.3. Docking

Ligand is docked against the protein (Figure III.5), and the interactions are analyzed. The scoring function gives a score on the basis of the best-docked ligand complex picked out [19].

III.6.4. Docking Validation

The root-mean-square deviation (RMSD) value is a crucial and necessary statistical parameter for docking validation which indicates the accuracy and the reliability of the docking. This value is obtained from the determination of the deviation between two positions: The first is the initial pose of the co-crystallized ligand in its original site, the second is the same co-crystallized ligand re-docked to the binding site of the protein pose. When the RMSD is less than 2 Å), the docking is considered reliable.

III.7. Molecular docking databases

The Protein Data Bank, which is open to the public, is often considered the most reliable resource for protein structure information (PDB). Additionally, it is not necessary to pay to access public databases such as the PubChem Compound Database and ZINC Besides; there are a great number of significant commercial databases, such as the Cambridge Structural Database (CSD) and the Compound Database (Aced) [71].

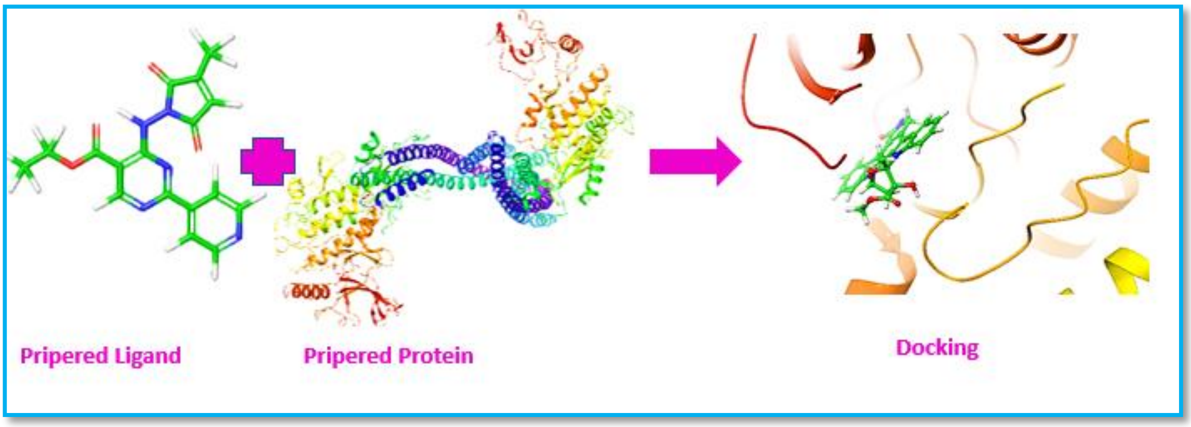


Figure III.5: Molecular docking flow chart.

III.8. Molecular docking software

The most popular types of molecular docking software, together with their associated algorithms, assessment methodologies, features, and application domains, are outlined in Table 1.

Table II.1: Representative software for molecular docking

Name	Search Algorithm	Evaluation method	Speed	Feature and Application areas
Flex X [72]	Fragmentation algorithm	Semi-empirical calculation on free energy	Fast	Flexible-rigid docking. Utilizing an incremental construction technique enables it to be utilized for the virtual screening of small molecule databases.
Glide [73]	Exhaustive systematic Search	Semi-empirical calculation on free energy	Medium	Flexible docking. This software narrows the search range using domain knowledge via XP (extra precision), SP (standard precision), and high throughout virtual.
AutoDock [74]	GA (genetic algorithm)	Semi-empirical calculation on free energy	Medium	Flexible-rigid docking. This software is always used in conjunction with Autodock-tools and is free for academic usage.
RDOCK [75]	GA (genetic algorithm) MC (monte carlo)	Molecular force Field	Medium	In addition to predicting the manner of binding, it is particularly well suited for high throughput virtual screening (HTVS) campaigns.

REFERENCES

References

[1] Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science*, *303*(5665), 1813-1818.

[2] Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, 2002, *1*, 882-894.

[3] Walters, W.P.; Stahl, M.T.; Murcko, M.A. Virtual screening – an overview. *Drug Discov. Today*, 1998, *3*, 160-178

[4] Langer, T.; Hoffmann, R.D. Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.*, 2001, *7*, 509-527.

[5] Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 2004, *3*, 935-949.

[6] Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, *7*(2), 146-157.

[7] Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C.R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.*, 2008, *153(Suppl 1)*, S7-26.

[8] Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **1982**, *161*, 269-288.

[9] Dias, R., de Azevedo, J., & Walter, F. (2008). Molecular docking algorithms. *Current drug targets*, *9*(12), 1040-1047.

[10] Laskowski, R.A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **1995**, *13*, 323-330, 307-328.

[11] Glaser, F.; Morris, R.J.; Najmanovich, R.J.; Laskowski, R.A.; Thornton, J.M. A method for localizing ligand binding pockets in protein structures. *Proteins*, **2006**, *62*, 479-488.

[12] Brady, G.P. Jr.; Stouten, P.F. Fast prediction and visualization of

[13] Kolb P, Irwin JJ D2009] Docking screens: right for the right reasons? Curr Top Med Chem 9:755–770 protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **2000**, *14*, 383-401.

[14]https://saylordotorg.github.io/text_the-basics-of-general-organic-and-biological-chemistry/s21-06-enzyme-action.html

[15] Mukesh, B., & Rakesh, K. (2011). Molecular docking: a review. *Int J Res Ayurveda Pharm*, *2*(6), 1746-51.

[16] Guedes, I. A., de Magalhães, C. S., & Dardenne, L. E. (2014). Receptor–ligand molecular docking. *Biophysical reviews*, *6*(1), 75-87.

[17] Agarwal, S., Chadha, D., & Mehrotra, R. (2015). Molecular modeling and spectroscopic studies of semustine binding with DNA and its comparison with lomustine–DNA adduct formation. *Journal of Biomolecular Structure and Dynamics*, *33*(8), 1653-1668.

[18] Agarwal, S., & Mehrotra, R. J. J. C. (2016). An overview of molecular docking. *JSM chem*, *4*(2), 1024-1028.

[19] Chaudhary, K. K., & Mishra, N. (2016). A review on molecular docking: novel tool for drug discovery. *Databases*, *3*(4), 1029.

[20] 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,

[21] Guedes, I. A., de Magalhães, C. S., & Dardenne, L. E. (2014). Receptor–ligand molecular docking. *Biophysical reviews*, *6*(1), 75-87.

[22] Brooijmans, N., & Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annual review of biophysics and biomolecular structure*, *32*(1), 335-373.

[23] Huang, Z., Wong, C. F., & Wheeler, R. A. (2008). Flexible protein–flexible ligand docking with disrupted velocity simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, *71*(1), 440-454.

[24] Kuhl, F. S., Crippen, G. M., & Friesen, D. K. (1984). A combinatorial algorithm for calculating ligand binding. *Journal of Computational Chemistry*, *5*(1), 24-34.

[25] Nishibata, Y., & Itai, A. (1991). Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron*, *47*(43), 8985-8990.

[26] Nishibata, Y., & Itai, A. (1993). Confirmation of usefulness of a structure construction program based on three-dimensional receptor structure for rational lead generation. *Journal of medicinal chemistry*, *36*(20), 2921-2928.

[27] Huang, S. Y., Grinter, S. Z., & Zou, X. (2010). Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics*, *12*(40), 12899-12908.

[28] Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, *8*(3), 195-202.

[29] Clark, D. E., & Westhead, D. R. (1996). Evolutionary algorithms in computer-aided molecular design. *Journal of Computer-Aided Molecular Design*, *10*(4), 337-358.

[30] Clark, D. E. (1999). Evolutionary algorithms in computer-aided molecular design: a review of current applications and a look to the future.

[31] Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., & Eldridge, M. D. (1998). Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Structure, Function, and Bioinformatics*, *33*(3), 367-382.

[32] Wu, G., Robertson, D. H., Brooks III, C. L., & Vieth, M. (2003). Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMm-based MD docking algorithm. *Journal of computational chemistry*, *24*(13), 1549-1562.

[33] Changeux, J. P., & Edelstein, S. (2011). Conformational selection or induced fit? 50 years of debate resolved. *F1000 biology reports*, *3*.

[34] Tsai, C. J., Ma, B., & Nussinov, R. (1999). Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences*, *96*(18), 9970-9972.

[35] Kar, G., Keskin, O., Gursoy, A., & Nussinov, R. (2010). Allostery and population shift in drug discovery. *Current opinion in pharmacology*, *10*(6), 715-722.

[36] Teodoro, M. L., & Kavraki, L. E. (2003). Conformational flexibility models for the receptor in structure-based drug design. *Current pharmaceutical design*, *9*(20), 1635-1648.

[37] Jiang, F., & Kim, S. H. (1991). “Soft docking”: matching of molecular surface cubes. *Journal of molecular biology*, *219*(1), 79-102.

[38] Apostolakis, J., Plückthun, A., & Caflisch, A. (1998). Docking small ligands in flexible binding sites. *Journal of Computational Chemistry*, *19*(1), 21-37.

[39] Mizutani, M. Y., Takamatsu, Y., Ichinose, T., Nakamura, K., & Itai, A. (2006). Effective handling of induced-fit motion in flexible docking. *Proteins: Structure, Function, and Bioinformatics*, *63*(4), 878-891.

[40] Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *Journal of molecular biology*, *235*(1), 345-356.

[41] Huang, S. Y., & Zou, X. (2010). Advances and challenges in protein-ligand docking. *International journal of molecular sciences*, *11*(8), 3016-3034.

[42] Novoa, E. M., Pouplana, L. R. D., Barril, X., & Orozco, M. (2010). Ensemble docking from homology models. *Journal of Chemical Theory and Computation*, *6*(8), 2547-2557.

[43] Sperandio, O., Mouawad, L., Pinto, E., Villoutreix, B. O., Perahia, D., & Miteva, M. A. (2010). How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *European Biophysics Journal*, *39*(9), 1365-1372.

[44] Nichols, S. E., Baron, R., Ivetac, A., & McCammon, J. A. (2011). Predictive power of molecular dynamics receptor structures in virtual screening. *Journal of chemical information and modeling*, *51*(6), 1439-1446.

[45] Carlson, H. A. (2002). Protein flexibility is an important component of structure-based drug discovery. *Current Pharmaceutical Design*, *8*(17), 1571-1578.

[46] Teodoro, M. L., Phillips Jr, G. N., & Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *Journal of Computational Biology*, *10*(3-4), 617-634.

[47] Zacharias, M., & Sklenar, H. (1999). Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: Application to DNA minor groove ligand complex. *Journal of Computational Chemistry*, *20*(3), 287-300.

[48] Guedes, I. A., Pereira, F. S., & Dardenne, L. E. (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Frontiers in pharmacology*, 1089.

[49] Li, J., Fu, A., & Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, *11*(2), 320-328.

[50] Meng, E. C., Shoichet, B. K., & Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. *Journal of computational chemistry*, *13*(4), 505-524.

[51] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, *79*(2), 926-935.

[52] Raha, K., Peters, M. B., Wang, B., Yu, N., Wollacott, A. M., Westerhoff, L. M., & Merz Jr, K. M. (2007). The role of quantum mechanics in structure-based drug design. *Drug discovery today*, *12*(17-18), 725-731.

[53] Senn, H. M., & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angewandte Chemie International Edition*, *48*(7), 1198-1229.

[54] Kramer, B., Rarey, M., & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. *Proteins: Structure, Function, and Bioinformatics*, *37*(2), 228-241.

[55] Yang, Y., Lightstone, F. C., & Wong, S. E. (2013). Approaches to efficiently estimate solvation and explicit water energetics in ligand binding: the use of WaterMap. *Expert Opinion on Drug Discovery*, *8*(3), 277-287.

[56] Kumar, A., & Zhang, K. Y. (2013). Investigation on the effect of key water molecules on docking performance in CSARdock exercise. *Journal of chemical information and modeling*, *53*(8), 1880-1892.

[57] Schulz-Gasch, T., & Stahl, M. (2004). Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, *1*(3), 231-239.

[58] Böhm, H. J., & Stahl, M. (2003). The use of scoring functions in drug discovery applications. *Reviews in computational chemistry*, *18*, 41-87.

[59] Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W., & Shoichet, B. K. (2002). A model binding site for testing scoring functions in molecular docking. *Journal of molecular biology*, *322*(2), 339-355.

[60] Feliu, E., Aloy, P., & Oliva, B. (2011). On the analysis of protein–protein interactions via knowledge-based potentials for the prediction of protein–protein docking. *Protein Science*, *20*(3), 529-541.

[61] Liu, J., & Wang, R. (2015). Classification of current scoring functions. *Journal of chemical information and modeling*, *55*(3), 475-482.

[62] Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules*, *20*(7), 13384-13421.

[63] Lin, J. H. (2011). Accommodating protein flexibility for structure-based drug design. *Current topics in medicinal chemistry*, *11*(2), 171-178.

[64] Salsbury Jr, F. R. (2010). Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current opinion in pharmacology*, *10*(6), 738-744.

[65] Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology*, *9*(1), 1-9.

[66] Harvey, M. J., & De Fabritiis, G. (2012). High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug discovery today*, *17*(19-20), 1059-1062.

[67] Alonso, H., Bliznyuk, A. A., & Gready, J. E. (2006). Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews*, *26*(5), 531-568.

[68] Nichols, S. E., Baron, R., Ivetac, A., & McCammon, J. A. (2011). Predictive power of molecular dynamics receptor structures in virtual screening. *Journal of chemical information and modeling*, *51*(6), 1439-1446.

[69] Vora, J., Patel, S., Sinha, S., Sharma, S., Srivastava, A., Chhabria, M., & Shrivastava, N. (2019). Molecular docking, QSAR and ADMET based mining of natural compounds against prime targets of HIV. *Journal of Biomolecular Structure and Dynamics*, *37*(1), 131-146.

[70] Bell, E. W., & Zhang, Y. (2019). DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *Journal of Cheminformatics*, *11*(1), 1-9.

[71] Fan, J., Fu, A., & Zhang, L. (2019). Progress in molecular docking. *Quantitative Biology*, *7*(2), 83-89.

[72] Kramer, B., Rarey, M., & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. *Proteins: Structure, Function, and Bioinformatics*, *37*(2), 228-241.

[73] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., ... & Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*, *47*(7), 1739-1749.

[74] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), 2785-2791.

[75] Li, L., Chen, R., & Weng, Z. (2003). RDOCK: refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics*, *53*(3), 693-707.

CHAPTER IV: Methodology, Results and Discussion

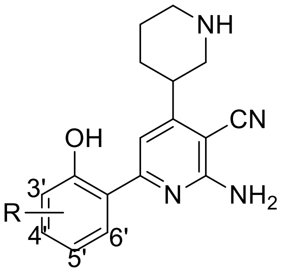
Part 1: *In silico* drug discovery of IKK-β inhibitors from 2-amino-3- piperidin -4-alkyl-6-(2 hydroxyphenyl) pyridine derivatives based on QSAR, docking, molecular dynamics and drug likeness evaluation studies

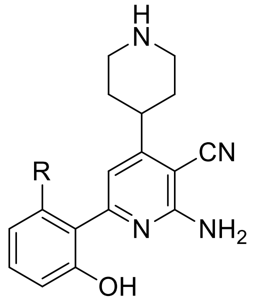
IV.1. QSAR Methodology

IV.1.1 Data set

A series of thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives known for their ability to inhibit the IKK-β enzyme, with their biological activities IC_{50} (i.e., concentration that is required for 50% inhibition) were taken from the literature [1.2]. The 30 compounds were divided into training and test sets. 75% were selected as a training set for the QSAR model elaboration the remaining 7 compounds were used as test sets [3]. The corresponding half-maximal inhibitory concentration (IC_{50}values were transformed to pIC_{50} (-logIC_{50}) values. The chemical structures and experimental activity data are shown in Table IV.1.1.

Table IV.1.1. The selected 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives with their experimental p IC50 values.

<table>
<tr><td colspan="3">
</td></tr>
<tr><th>Compound</th><th>R</th><th>IC_{50} (nM)</th></tr>
<tr><td>Comp1</td><td>-H</td><td>25</td></tr>
<tr><td>Comp2</td><td>-CH3</td><td>1300</td></tr>
<tr><td>Comp3</td><td>-OCH3</td><td>20.000</td></tr>
<tr><td>Comp4</td><td>-OCH3</td><td>560</td></tr>
<tr><td>Comp5</td><td>-CH3</td><td>2400</td></tr>
<tr><td>Comp6</td><td>-NH2</td><td>7800</td></tr>
<tr><td>Comp7</td><td>-OH</td><td>15</td></tr>
<tr><td>Comp8</td><td>-OHCH3</td><td>34</td></tr>
<tr><td>Comp9</td><td>-OCH2CH3</td><td>14</td></tr>
</table>

Comp10	-O(CH2)2CH3	5
Comp11	-O(CH2)4CH3	6
Comp12	-O(CH2)6CH3	14
Comp13	-OCH(CH3)2	81
Comp14	-OCH2CH(CH3)2	5
Comp15	OCH2-cyclopropyl	3
Comp16	-OCH2-cyclobutyl	4
Comp17	-OCH2-cyclohexyl	26
Comp18	-OCH2-phenyl	9
	 	
Compound	**R**	IC_{50} **(nM)**
Comp19	-H	300
Comp20	-OH	270
Comp21	$-OCH_2CH_3$	120
Comp22	$-O(CH_2)_2CH_3$	24
Comp23	$-O(CH_2)_3CH_3$	15
Comp24	$-O(CH_2)_4CH_3$	20
Comp25	$-O(CH_2)_5CH_3$	25
Comp26	$-O(CH_2)_6CH_3$	50
Comp27	$-OCH_2CH(CH_3)2$	15
Comp28	OCH_2-cyclopropyl	8.5
Comp29	$-OCH_2$-cyclobutyl	12
Comp30	$-OCH_2$-phenyl	110

IV.1.2. Compound preparation and calculation methods

Two-dimensional (2D) structures of the investigated compounds were drawn using ChemDraw ultra 12.0 software. The structures were saved in the sdf file format and then transformed to 3D structures directly when input into Materials studio 8.0 software. Geometric optimization of these compounds was carried out as a first step to get the quantum chemical parameters (Molecular descriptors). The quantum chemical calculations were performed using the GGA (generalized gradient approximation) within the framework of the density functional theory (DFT), using software package DMol3 in Materials Studio. The geometric optimization of all investigated compounds was achieved by the Triple Numerical with Polarization (TNP) basis set and the functional exchange-correlation (BP) [4.5].

IV.1.3. QSAR models analysis

The best QSAR model resulted for the estimation of the pIC_{50} value is represented by the following MLR equation:

$$pIC50 = 60.838 + 1.293 * HBA + 0.854 * ChC + 5.453 * MF - 27.82 * BI\_JX - 0.013 * VD_E - 0.965 * ESKS\_ssCH2$$

(1)

Where HBA is the hydrogen bound acceptor, ChC is the chiral center, MF is the molecular flexibility, BI_JX is the Balaban Index, VDE is the vertex distance /equality and ESK S_ssCH2 is the E-state keys (sums): S_ssCH2.

Suitable statistical methods were adopted to determine the best fitting of the mathematical model, which expresses the variation of the inhibitory concentration as a function of the descriptors mentioned above. From the obtained values, Table IV.1.2 shows the coefficient of determination (R^2=0.938), the root-mean-square error (RMSE=0.257) and the Adjusted $R_A^2 = 0.917$, the MLR model provides an accurate fit of the experimental data set and it is characterized by a high predictivity. The cross validation (Q^2=0.931) insures this information.

Table IV.1.2. Statistical results of MLR model.

R^2	0.938
R_A^2	0.917
RMSE	0.257
Q^2	0.931
$\boldsymbol{R^2_{pred}}$	0.430

The most important significant parameters were chosen to be used in QSAR model construction. Out of the 159 descriptors considered for each compound in the dataset, only 6 proved to be significant: ChC, HBA, MF, BI_JX, IC, VDE, ESK S_ssCH2, showing high correlations with pIC_{50} values of compounds. These descriptors influence the inhibitory activity of the compounds toward the Ikk-B enzyme. The selection of these parameters was performed according to the coefficient of determination and the P-value of the regression

Table IV.1.3, while the obtained values of the parameters are represented in together with their statistical properties. The parameters are considered statistically significant if P-value <0.05.

Table IV.1.3. Predictor coefficients of the MLR algorithm

Terme	Estimate	Standard Error	T- ratio	Prob. > |t|
Constant	60.838	4.962	12.26	<.0001*
HBA	1.293	0.183	7.05	<.0001*
ChC	0.854	0.121	7.04	<.0001*
MF	5.453	0.474	11.50	<.0001*
BI	-27.820	2.293	-12.13	<.0001*
VDE	-0.013	0.001	-11.72	<.0001*
ESK S_ssCH2	-0.965	0.098	-9.80	<.0001*

Table IV.1.4 contains the predicted and observed values of the compounds (training set and test set). The contribution of each parameter to the investigated biological activity depends on the corresponding coefficient value and on the sign that precedes it.BI_JX, IC, VDE, ESK S_ssCH2, are preceded by a negative sign. Consequently, these parameters have a diminishing effect on the dependent variable (pIC_{50}), while, HBA, ChC, MF, which are preceded by positive sign, have an increasing effect on the dependent variable value. Therfore it can be concluded that the most active predicted compounds are characterized by HBA, BI_JX, IC, VDE, ESK S_ssCH2, which should not be elevated. The more the chiral center, molecular flexibility and Hydrogen Bond Acceptors values increase the more the ($pIC50_{AChE}$) value increases.

Table IV.1.4. Experimental and predicted values of the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives

Compounds	Observed activity (pIC_{50})	Predicted activity	Residual
$\boldsymbol{Cmp1^{-}}$	7.60	7.41	-0.19
Cmp2	5.88	5.97	0.09
$\boldsymbol{Cmp3^{-}}$	4.69	4.13	0.34
$\boldsymbol{Cmp4^{-}}$	6.25	8.25	2.00
Cmp5	5.61	5.84	0.22
Cmp6	5.10	5.07	-0.03
$\boldsymbol{Cmp7^{-}}$	7.82	7.18	-0.64
Cmp8	7.46	7.61	0.14
Cmp9	7.85	7.93	0.08
Cmp10	8.30	8.11	-0.19
$\boldsymbol{Cmp11^{-}}$	8.22	8.35	0.13

Cmp12	7.85	8.19	0.34
Cmp13	7.09	7.26	0.17
Cmp14	8.30	7.97	-0.32
Cmp15	8.52	8.93	0.41
Cmp16	8.39	8.16	-0.23
Cmp17	7.58	7.25	-0.33
Cmp18	8.04	7.87	-0.169
Cmp19	6.52	6.65	0.12
Cmp20	6.56	6.43	-0.13
Cmp21	6.92	7.18	0.26
Cmp22	7.61	7.35	-0.26
Cmp23	7.82	7.50	-0.31
Cmp24	7.69	7.60	-0.09
Cmp25	7.60	7.59	-0.00
Cmp26	7.30	7.44	0.14
$\boldsymbol{Cmp27^{-}}$	7.82	7.22	-0.59
Cmp28	8.07	8.18	0.11
Cmp29	7.92	7.41	-0.50
$\boldsymbol{Cmp30^{-}}$	6.958	7.13	0.172

- Test set

The parity plot shows a linear regression which confirms that the MLR model matches well the experimental data, except for few scattered points and it can be used for the prediction of pIC_{50} values with reliability Figure IV1.1.

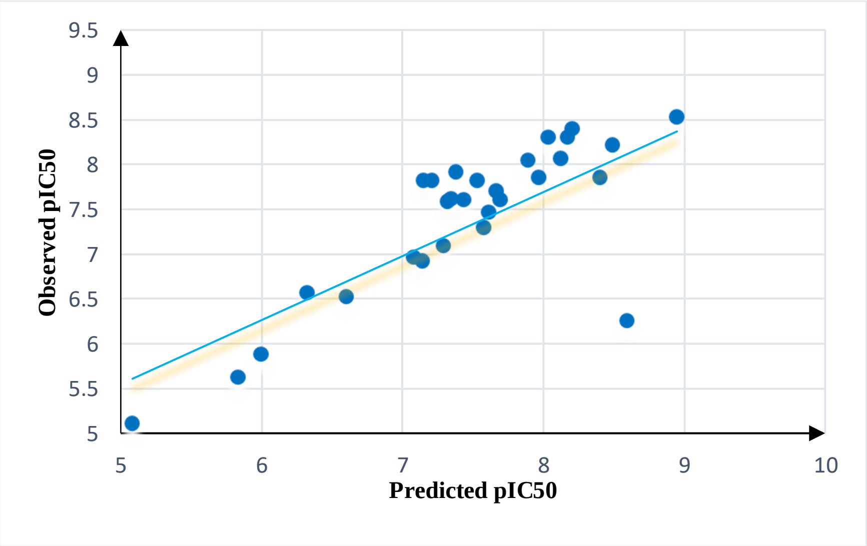


Figure IV.1.1. Parity diagram (observed values *vs* predicted values) of the inhibitory concentration using MLR model.

The analysis of the residual confirms this information Figure IV.2 where most of data points are less than an absolute value of 0.4 and are distributed around the zero line.

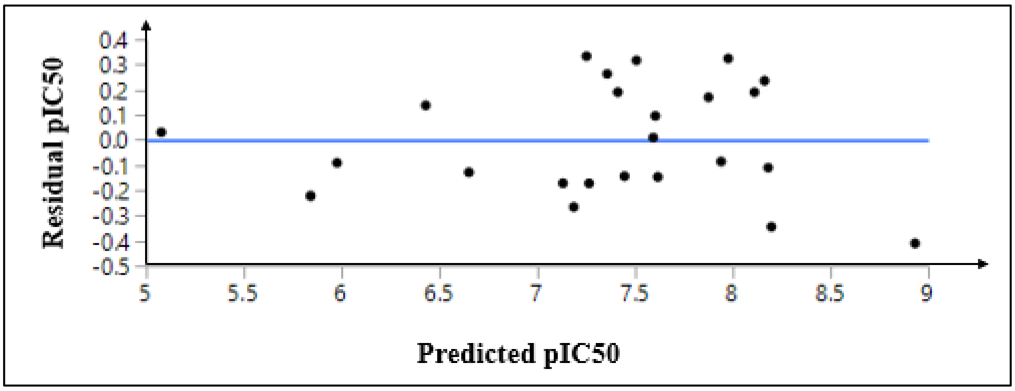


Figure IV.1.2: Graph of residuals versus predicted pIC_{50} values using MLR model.

IV.1.4. Applicability domain and compound conception

A small library of new molecules possessing (2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine was generated based on analogue design strategy. Their inhibitory concentration was predicted by the obtained model. Then the applicability domain of the elaborated model allowed to exclude the A11p compounds and 2 other molecules which were outside the domain,

The horizontal limits of this domain are the standard deviation of this model $\pm$3and the vertical line represents the warning leverage (h* = 0.83). All compounds (training set, test set, and new designed compounds) present leverage values less than the warning h* value (0.83) and standardized residuals values less than the thresholds. except the two outliers compounds as shown in Figure IV.1.3. Among the new designed compounds inside the applicability domain, 21 compounds showed higher biological activities compared to the (2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives. The results are shown in Table IV.1.5.

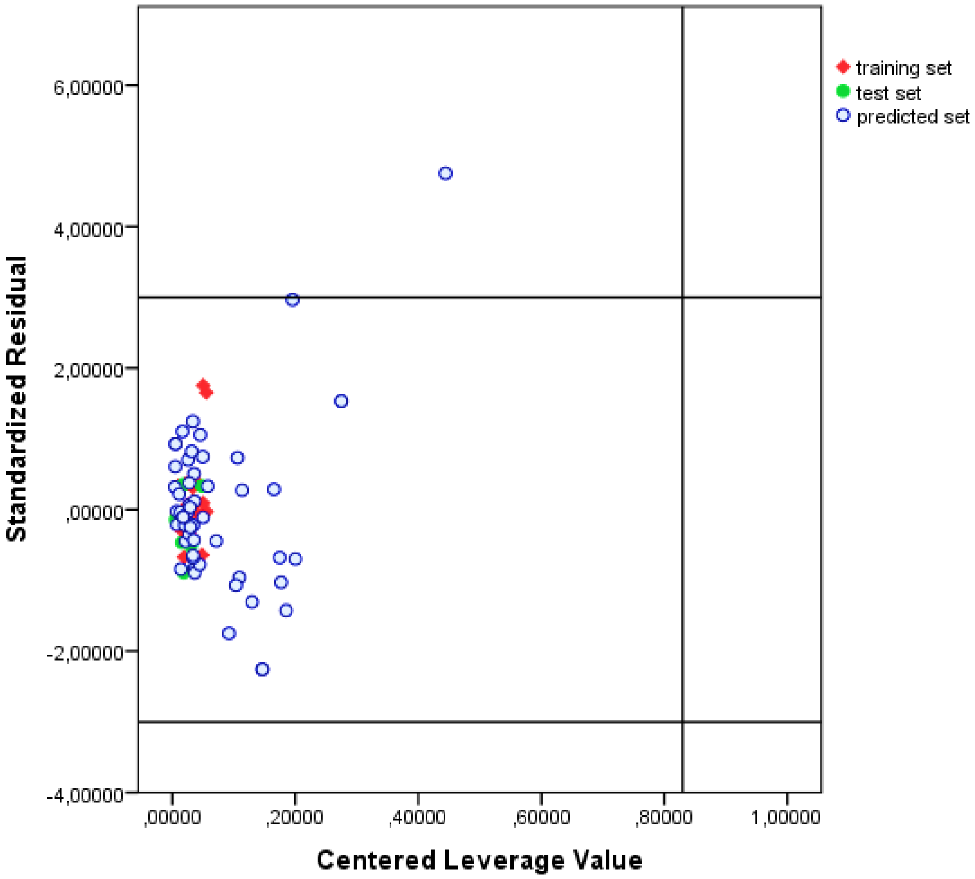
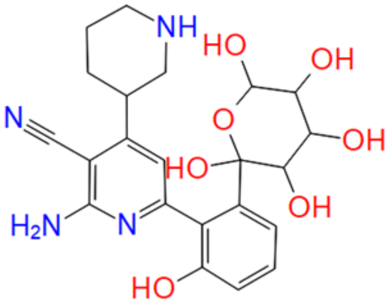
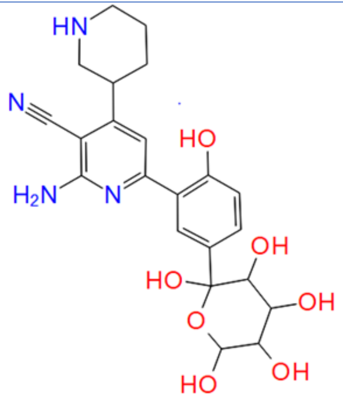
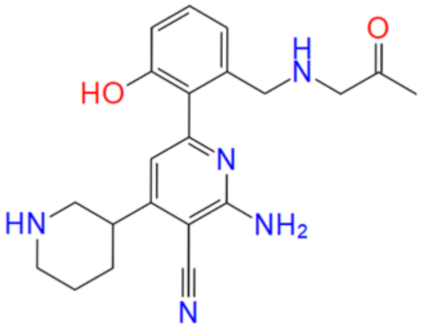
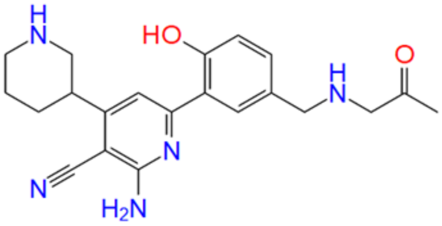
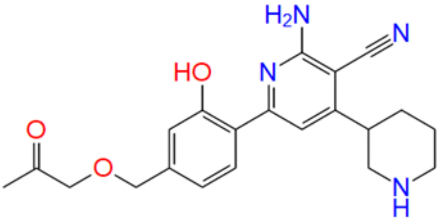
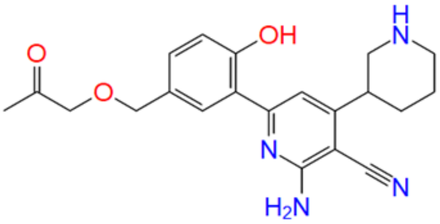
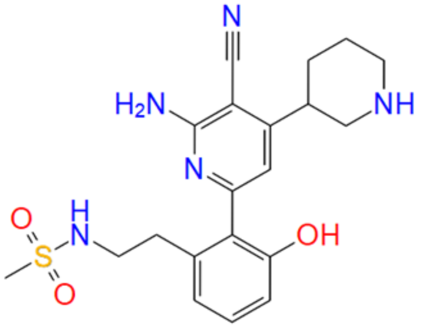
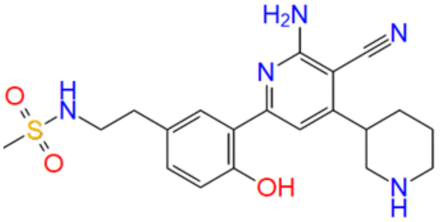


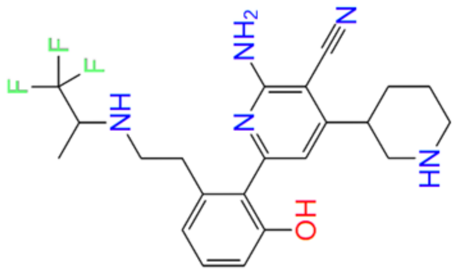
Figure IV.1.3: Applicability domain plot of the MLR model

The Table IV.1.5 shows the predicted pIC 50 values of the new designed compounds. Each two molecules have the same chemical structure but they are different at the substitution position.

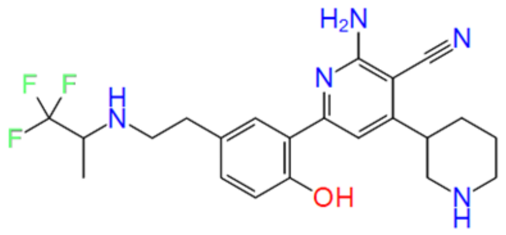
For example: The compound A1p and A1m are two molecules with a same structure where their substitutions are at para, meta positions of the phenyl ring respectively. From the obtained values it was noticed that, the substitutions at different positions (meta and para) of phenyl ring influenced the biological activity. Therefore, all compounds which are characterized by a substitution at para position showed better and higher pIC_{50} except (A10 and A3) compared to the molecules characterizing by a substitution at meta position. It can be concluded that the para positions may increase the IC_{50} values and consequently, the drug efficacy increases at low concentration.

Table IV.1.5. New designed compounds with their predicted pIC_{50} values.

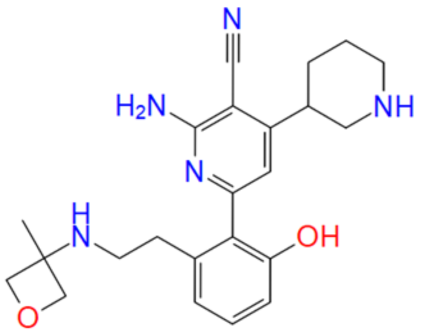
A1m	pIC_{50} = 10.41	**A1p**	pIC_{50} = 10.69
			
A2m	pIC_{50} = 8.621	**A2p**	pIC_{50} = 9.39
			
A3m	pIC_{50}= 10.76	**A3p**	pIC_{50} = 10.079
			
A4m	pIC_{50} = 8.59	**A4p**	pIC_{50} = 9.31
			
A5m	pIC_{50} = 9.89	**A5p**	pIC_{50} = 10.465



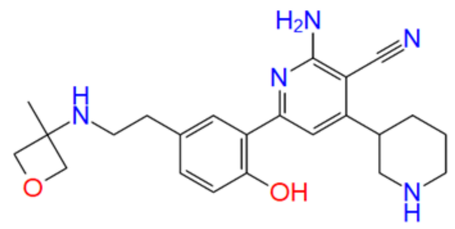
A6m pIC_{50} = 8.23



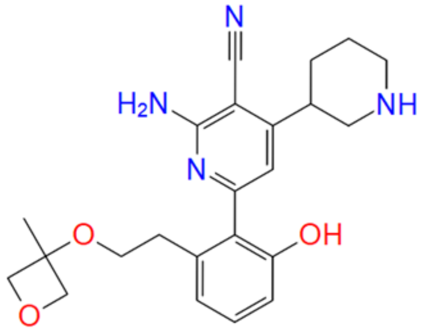
A6p pIC_{50} = 8.55



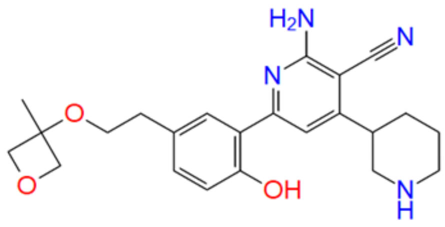
A7m pIC_{50} = 8.96



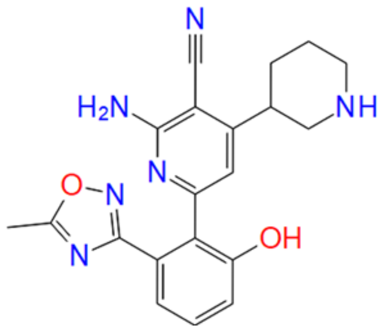
A7p pIC_{50} = 8.27



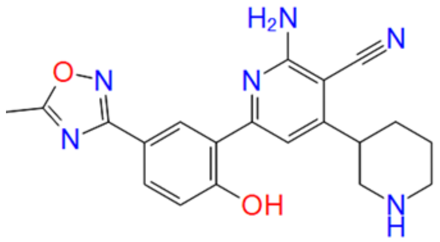
A8m pIC_{50} = 7.30



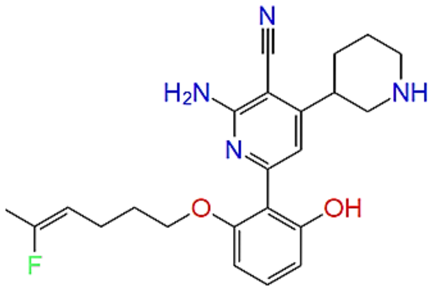
A8p pIC_{50} = 8.231



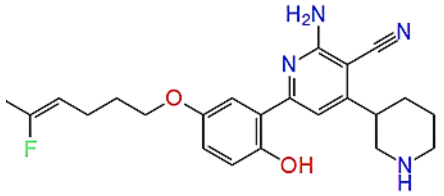
A9m pIC_{50} = 7.84



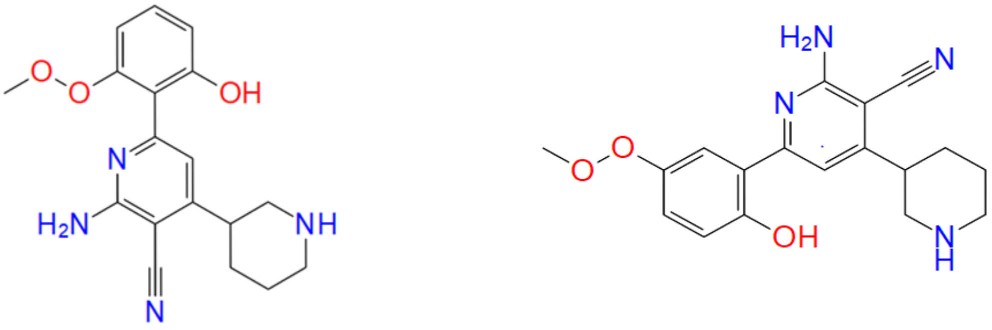
A9p pIC_{50} = 10.46



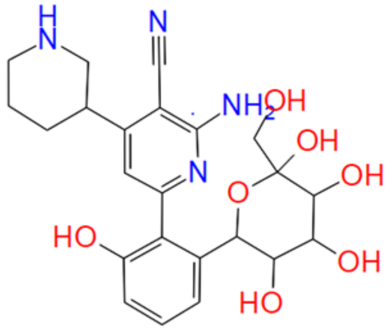
A10m pIC_{50} = 9.97



A10p pIC_{50}= 9.89



A11m pIC_{50} = 10.98



P : para position ; m : meta position

IV.2. Molecular Docking Methodology

IV. 2.1. Ligands preparation

Ligprep module was used to prepare the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives and the new designed compounds which is considered as a necessary step for an appropriate docking. The concept of ligand preparation involves taking 2D structures producing corresponding low energy 3D structures. This minimization of energy was achieved by applying the OPLS3 force field.

IV. 2.2. Protein Preparation and grid generation

Removing water molecules, fill in messing loops and side chains, deleting the alternate conformations and adding hydrogen, inserting the missing atoms in incomplete residues are elementary and indispensable steps which were used for the protein refinement and preparation. Then, receptor grid generation was performed in order to determine the position and size of the receptor active site area where binding interaction between the ligand and the residues can occur [6].

IV. 2.3. Docking Study

The molecular docking study was carried out in order to determine the binding mode between ligands and macromolecular targets and to investigate the potential of the new candidate drugs. The x-ray crystal structure of IKK-β complexed with K252-A compound was taken from the protein data bank (PDB code: 4KIK) https://www.rcsb.org/structure/4KIK. In this study the XP molecular docking method was performed to identify the different poses of molecules into the active site of the protein. The evaluation of the obtained docked poses and ranking the investigated compounds were based on the XP-*GScore*. This scoring function is based on an energy model which can be approximated by a sum of different types of energies [7], which are: van deer waals energy. coulomb energy, lipophilic contacts energy, hydrogen bonds energy, metallic bonds energy, penalty for buried polar groups, penalty for immobilizing ligand's rotatable bonds, and polar contacts energy in the active site. The best predicted affinity corresponds to the negative maximum (XP-*GScore)* value and the more negative the (XP-*GScore) is,* the more the compound is expected to be well binded to IKK-β enzyme. The results are shown in Tables IV.1.6 and IV.1.7.

IV.2.4 Docking Validation

The root-mean-square deviation (RMSD) value is a crucial and necessary statistical parameter for docking validation which indicates the accuracy and the reliability of the docking. This value is obtained from the determination of the deviation between two positions: The first is the initial pose of the co-crystallized ligand in its original site, the second is the same co-crystallized ligand re-docked to the binding site of the protein pose [2.8]. When the RMSD is less than 2 Å) the docking is considered reliable [8].

IV.2.5. All Atom Molecular dynamics (MD)

To characterize and analyze the stability of the docking poses we performed classical MD on five different systems: namely IKK-β interacting with A1M, A11M, A1P, A5P or A9P. The force field for the drug was parameterized following the Amber antechamber procedure. Restricted electrostatic potential (RESP) [2] procedure was used to assign point charges, and the ground state geometry of the five drugs was optimized at the density functional theory level using the standard 6-31G basis set and the B3LYP functional [9] IKK-β contains one phosphoserine residue that was modeled using the force field developed by Homeyer *et al.* [10] The rest of the protein was described with amberf99 amber force field

The different systems have been solvated in a cubic water box described by a TIP3P force field [11,12], and K$^+$ cations were added to ensure electroneutrality of the simulation box. To speed up the simulation, the H Mass Repartition (HMR) algorithm (Hopkins et al. 2015) has been used so we increased the time step to integrate Newton's equations of motion to 4 fs by artificially scaling the mass of all non-water hydrogen atoms from 1.008 to 3.024 Da.

After an equilibration and thermalization of 18 ns, 100 ns have been run in the constant pressure and temperature (NPT) ensemble at 1 atm and 300K. All MD simulations have been performed using the NAMD 2.13 code [13] and results have been analyzed and visualized with VMD

IV.2.6. Molecular docking study

The root-mean-square deviation (RMSD) value between the initial pose and docking of the co-crystallized ligand [14] was 0.0131 which implied that the docking method showed a highest precision and could reproduce the crystal binding model [15]. The representation of the superposed poses is shown in Figure IV.1.4.

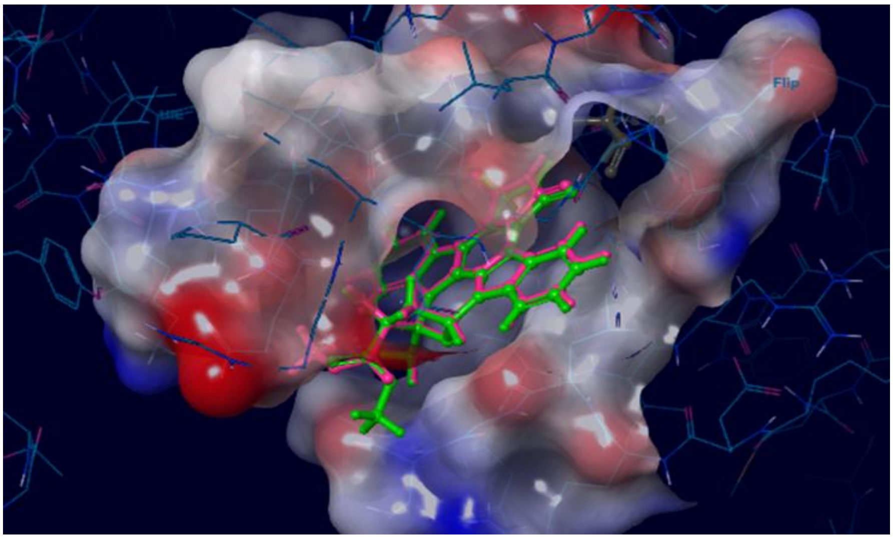


Figure IV.1.4: Spatial overlap of the docking position (pink) with the experimental position (green) of K252-A into the active site.

Once the docking protocol is validated, it can be used for the prediction of the binding modes between new IKK-β inhibitors compounds and IKK-β enzyme. The screened 21 designed compounds showing a higher biological effect, and the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives were docked into the active pocket of the IKK-βusing the same approach. The obtained results showed that all the investigated compounds have poses in the cavity and they interacted with residues via H bond interactions. The best pose of each molecule was chosen based on the *Glide* score (XP-*GScore)*.

Docking of the co-crystallized compound K252-A reproduced the same interaction types (Hydrogen bond interaction) observed experimentally. The visual analysis of the best pose showed that the best complex was obtained when the K252-A interacts with the following residues GLU 97, CYS 99, GLU 149 via three H-bond interactions. These interactions ensured the best orientation of the molecule towards the cavity with high affinity which allows it to trigger the inhibition of the IKK-β enzyme. The XP GScore of this pose is -14.287 (Figure IV.1. 5).

Also, the docking results indicated that most compounds shared the same binding mode with CYS 99 residue as the co-crystallized compound K252-A, which is characterized by high affinity to the active site and they contain several hydroxyl groups. The docking score values of the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives are ranging from -5.710 to -8.441 Table 6. The comp12 showed the highest binding affinity in this series, where the (XP-*GScore)* of this pose was estimated by -8.439. The interactions of K252-A and comp12 compounds with the protein are shown in Figures IV.1.5 and IV.1.6.

Table IV.1.6. Docking score of 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives and Interacting residues

Compound	XPG Score	Docking score	Interacting residues
K252-A	-14.287	-14.287	GLU 97, CYS 99, GLU 149
Comp12	-8.441	-8.439	TYR98, CYS99, GLN100
Comp18	-7.784	-7.783	ASN150, ASP166
Como4	-7.701	-7.700	CYS99, ASP103, LYS44
Comp1	-7.585	-7.584	LEU21, CYS 99, ASP103
Comp16	-7.531	-7.529	LEU21
Comp23	-7.516	-7.516	CYS99, ASP 103
Comp26	-7.463	-7.463	GLU19, CYS99, ASP 103
Comp17	-7.463	-7.462	LEU21, CYS99, ASP103, ASP166
Comp3	-7.326	-7.325	CYS99, ASP103, GLU149
Comp2	-7.209	-7.208	CYS99, ASP103, GLU149
Comp19	-7.128	-7.128	CYS99, ASP103
Comp15	-7.105	-7.104	CYS99, ASP103
Comp14	-7.098	-7.097	CYS99, ASP103
Comp20	-7.018	-7.018	CYS99, ASP103, ASP166

Comp27	-6.98	-6.988	ASP103, ASP166
Comp13	-6.903	-6.901	CYS99, ASP103
Comp24	-6.864	-6.864	CYS99, ASP103
Comp7	-6.850	-6.847	CYS99, ASP103, ASP166
Comp10	-6.762	-6.760	THR23, LYS44
Comp6	-6.744	-6.744	CYS99, ASP103
Com25	-6.681	-6.681	ASP103, ASP166
Comp5	-6.681	-6.680	LEU21
Comp28	-6.677	-6.677	ASP103
Comp21	-6.551	-6.551	ASP166
Comp9	-6.488	-6.486	LYS44, ASN150, ASP166
Comp8	-6.202	-6.201	CYS99
Comp11	-6.344	-6.342	ASN150, ASP166
Comp30	-6.308	-6.308	ASN150, ASP166
Comp22	-5.711	-5.711	ASN150, ASP166
Comp29	-5.705	-5.705	ASP66

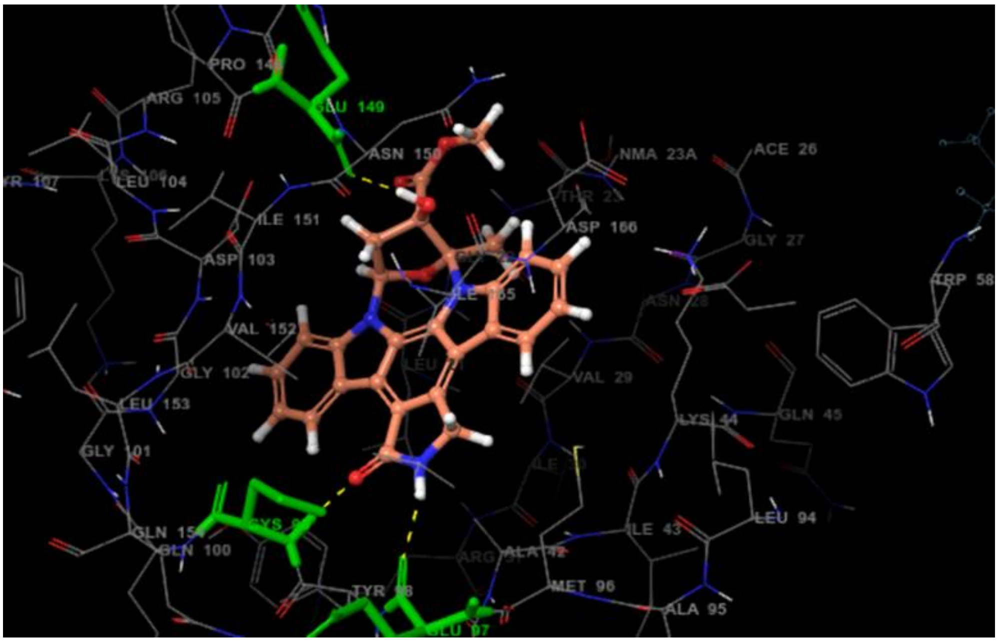


Figure IV.1. 5: Ligand–protein interaction scheme of K252-A compound. Interacting residues are colored in green

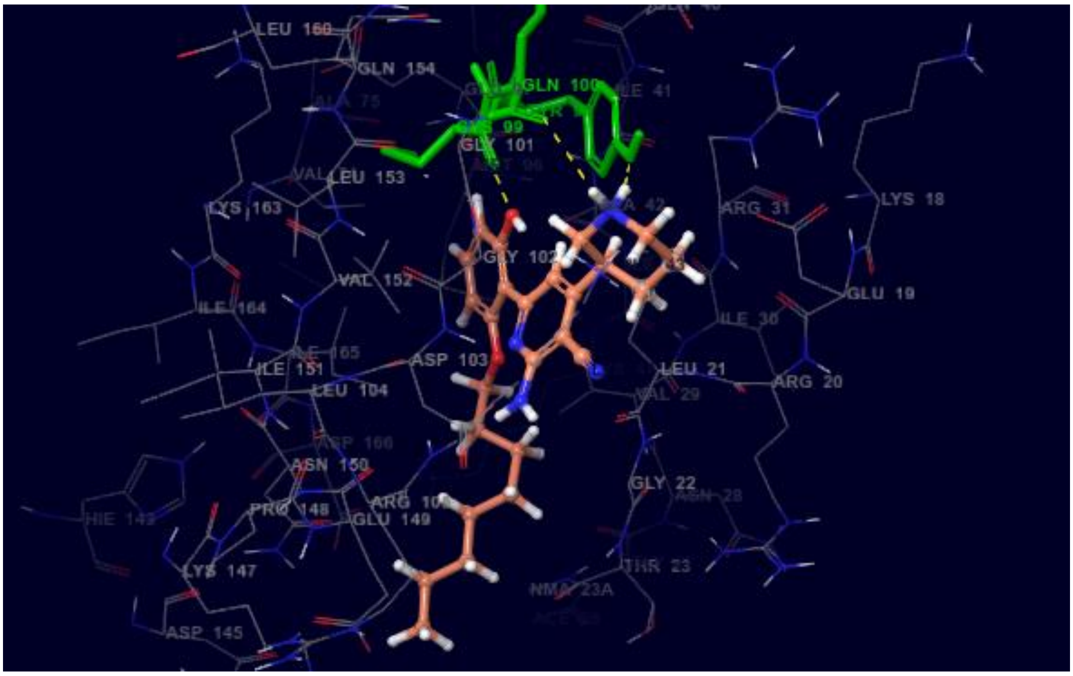


Figure IV.1.6: Ligand–protein interaction scheme of comp12. Interacting residues are colored in green

The suitable position of the comp 12 which allows to form the most stable complex, has been ensured by three H- bond interactions. The first occurred between the hydroxyl group of the compound phenyl ring and CYS99 residue while the other two were between the amino hydrogen group of the pyridine and GLN150, TYR98 residues. The XP-*GScore* this pose was estimated by -8.444.

As for the new designed compounds their docking score is ranging from -6.365 to -12.375. **(**Table IV.1.7**)**. Out of 21 compounds four compounds showed a better binding affinity to the active site compared with the comp12. They share the same binding mode with CYS99 and GLU149 amino acid residues as the co-crystallized compound K252-A. These amino acid residues made the molecules more stable in the pocket. So, they might be important for IKK-β inhibition.

The XP-*GScore* value of the best new designed compound which is characterized by the highest affinity to the active site (A1m compound) was estimated by -12.375 (Figure IV.1.7). The proposed substitution added at para position of the phenyl ring allows the structure to adopt an appropriate orientation in the active site so that this molecule can interact with the following residues LEU21, GLU97, CYS99, ASP103, ASN155, ASP166 via six H-bond interactions. The docking results showed that this molecule shares the same binding mode with GLU97 and CYS 99 residues as the co-crystallized compound K252-A.

As for the A1p compound, changing the substitution at para position led to the modification of the molecule orientation into the active site and the interacted residues. When it docked into the active site, it was able to form hydrogen bonds with the following residues GLU149, THR23, ASP103. It also led to the minimization of the affinity of the molecule to the pocket, where its XP-*GScore* value became -10. 763 (Figure IV.1.8).

The compound A11m was found to dock into the active site of IKK-β with good affinity, XPGscore (-10.633), interacting with five residues (TYR98, CYS99, ASP103, GLU19, GLU149) forming six hydrogen bond interactions via two hydroxylic groups of the added substitution and the hydroxylic group of the phenyl ring in addition to its amino hydrogen groups. The interactions are well visualized in ligand interaction diagram (Figure IV.9). Hence, the proposed substitution improved the orientation of this compound in the active site, and consequently improved its affinity. The fourth best compound is A9p. Its conformation was stabilized via three H-bonds. When the amino hydrogen of the pyridine ring interacts with amino acid residue CYS99, this latter also interacts with the hydroxylic group of the phenyl ring. The piperidine ring of this molecule also contributed for this suitable conformation via its hydrogen amino group, which interacted with the ASP 103. The docking poses of these compounds are shown in (Figures IV.1.9, IV.1.10).

Table IV.1.7. Docking score of design compounds and interacting residues.

Compound	XPG Score	Docking Sore	Interacting residues
A1p	-12.375	-12.375	LEU21 GLU97, CYS99, ASP103, ASN155, ASP166
A1m	-10.713	-10.711	THR23, ASP103, GLU149
A11m	-10.633	-10.631	TYR98, CYS99, ASP103, GLU19, GLU149
A9p	-9.381	-9.381	CYS99, ASP103, GLU149
A5p	-8.698	-8.614	CYS99, GLN100
A6p	-8.526	-8.512	CYS99, ASP103
A7m	-8.479	-8.478	CYS99, TYR98
A6m	-8.430	-8.430	GLU97, CYS99, ASP103
A4m	-8.279	-8.279	CYS99
A2p	-8.199	-8.159	CYS99, ASP103, LEU21
A5m	-7.936	-7.552	CYS99, ASP103, GLU149
A4p	-7.623	-7.622	LEU21, CYS99, GLU 97, ASP103

A10m	-7.546	-7.545	CYS99, GLN150
A8m	-7.447	-7.443	CYS99, ASP103
A7p	-7.410	-7.409	CYS99
A10p	-7.393	-7.393	CYS99, ASP103, GLU149
A8P	-6.989	-6.970	LYS44
A3m	-6.923	-6.922	LYS106, ASN150, ASP166
A3p	-6.864	-6.863	LYS106, ASN150, ASP166
A9m	-6.512	-6.510	ASP103, TYR98, CYS99, GLU19
A2m	-6.365	-6.412	ASP103, CYS99, GLU97

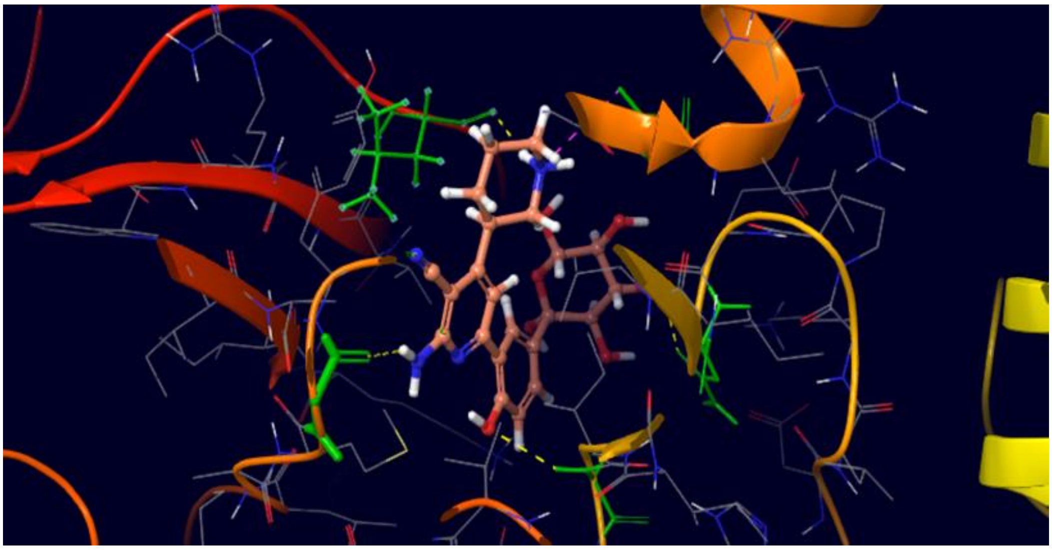


Figure IV.1.7: Ligand–protein interaction scheme of A1p compound. Interacting residues are colored in green

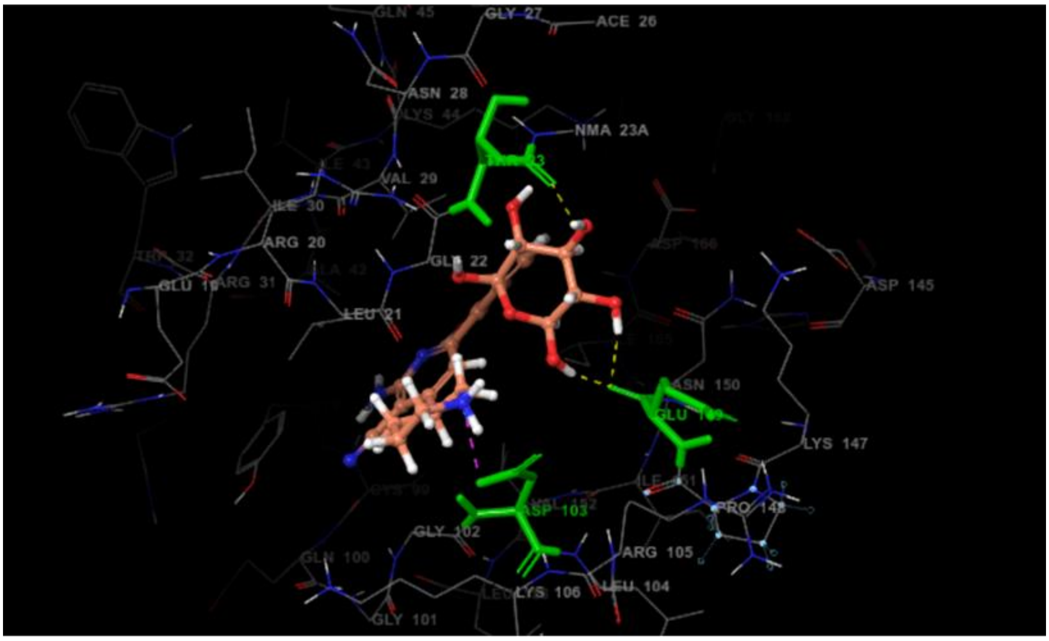


Figure IV.1.8: Ligand–protein interaction scheme of A1m compound. Interacting residues are colored in green

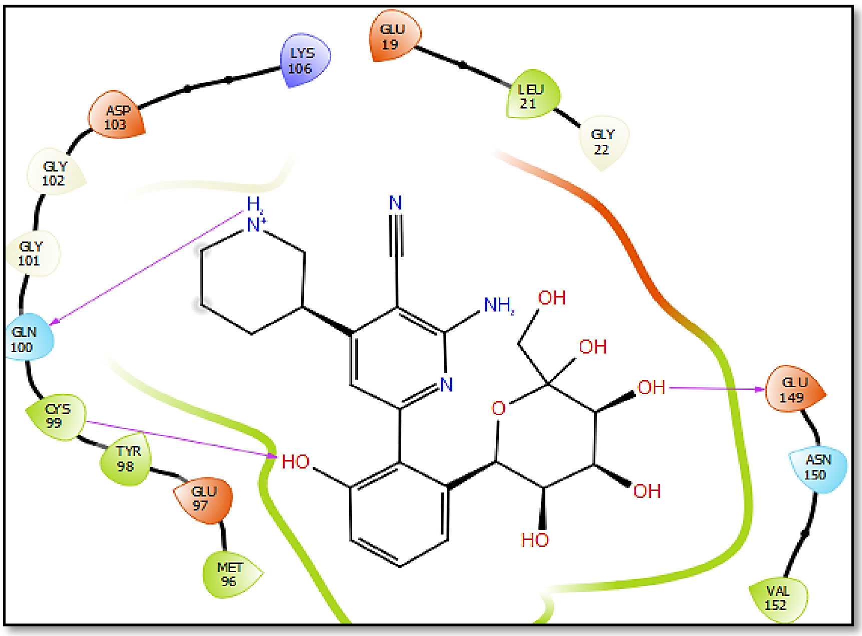


Figure IV.1.9: Ligand interaction diagram of the molecule A11m.

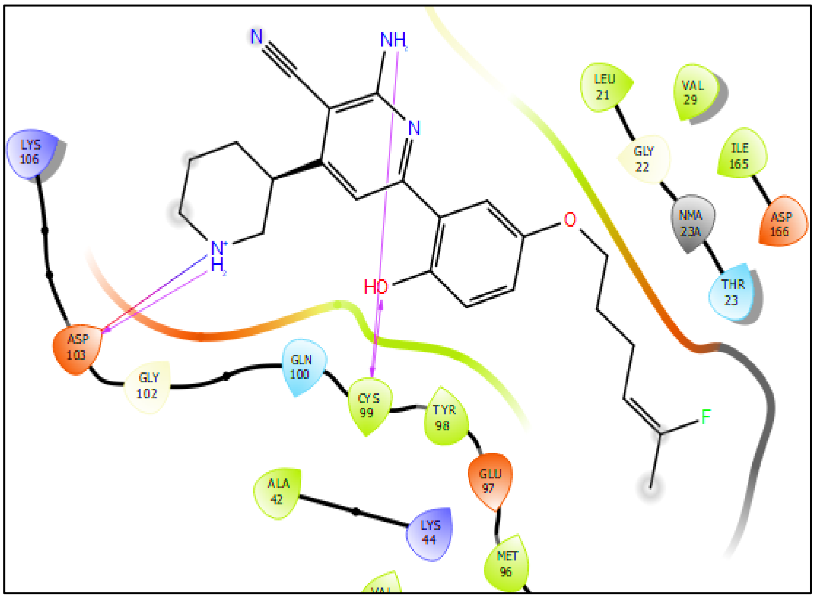


Figure IV.1.10: Ligand interaction diagram of the molecule A9p.

IV.2.7. Molecular Dynamics Simulations

In order to assess for the stability and validate the pose issued from the molecular docking approach we have performed all atom MD simulations for five leading compounds (Due to the impossibility of achieving the molecular dynamics for all compounds only five of them have been selected randomly for this study) . The results are pictorially reported in Figure IV.1.11. It appears that all the drugs remain persistently into the binding pocket, as shown by the snapshot extracted at 100 ns, i.e. at the end of the simulation. As evidenced by the low value root mean square deviation (RMSD) not exceeding 4 Å reported in Supplementary Information the aggregate is globally stable and the structure of the protein is conserved, even if a slight enlargement of the binding pocket may be observed. However, a moderate sliding of the drug inside the pocket is also evident that in the case of A11M leads to the rearrangement of the interaction patterns with the protein amino acid, namely the weakening of the interaction with Asp103 and Leu21.

A similar, albeit less pronounced effect is also evidenced for A5P that experiences the partial breaking of the interaction with Cys99 while the concomitant interaction with Glu19 is strongly reinforced. Globally, the results definitively point to the persistence of the drug protein aggregate and hence confirm and validate the docking poses.

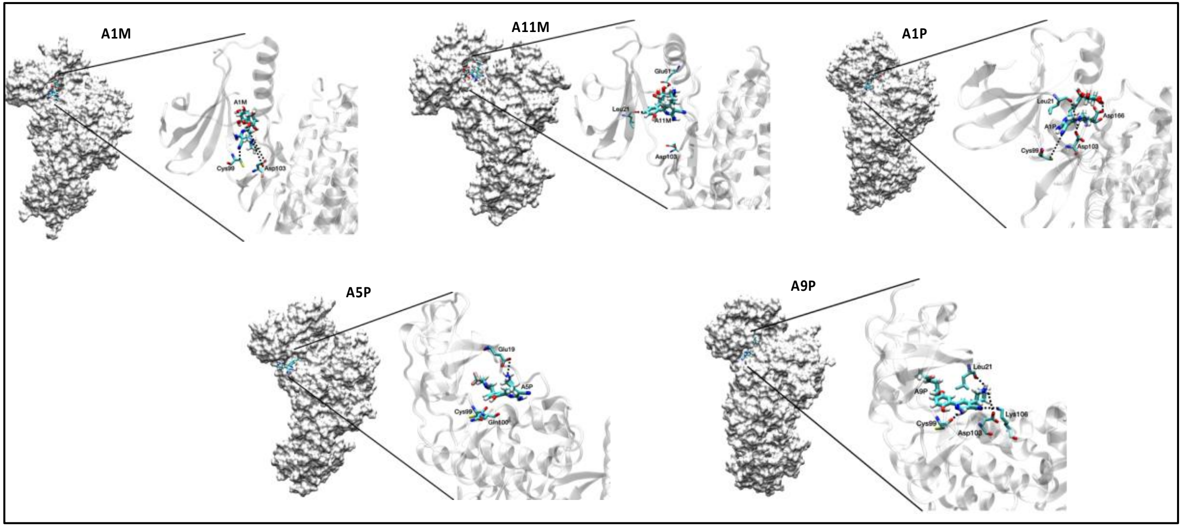


Figure IV.1.11: Snapshots extracted at 100ns of the MD trajectory showing the stability of the drug/protein complex. The specific interaction taking place are also highlighted. Larger figures are also available in Supplementary Information together with the initial snapshot (0 ns) corresponding to the docking pose.

IV.3. Druglikness properties

Pharmacokinetics parameters such as (Absorption, distribution, Metabolism-Excretion) play a crucial role in drug discovery [16]. High and low variable bioavailability is one of the pharmacokinetics properties which is indeed the main reason that impedes the development of drugs [17]. Understanding the molecular properties that limit oral bioavailability facilitates the design of new effective candidates' drug [18].

Lipinski /Veber rules and the analysis of the golden triangle allow the prediction of drug-likeness character by analyzing the Absorption, Distribution, Metabolism, Excretion (ADME) properties which are affected by the physicochemical properties of the drug compounds [16] Lipinski 'Rule-of-Five' describes the importance of lipophilicity (octanol-water partition), molecular weight (MW), number of hydrogen bond donors and number of hydrogen bond acceptors in the analysis of the structures of orally administered drugs, and candidate drugs [17.18].

The empirical conditions which suit Lipinski's rule and demonstrate good oral bioavailability imply an equilibrium between the capacity of a compound to diffuse passively across the various biological barriers and its aqueous solubility [19]. According to Lipinski's rule of five, (Figure IV.1.12) compounds which do not validate at least two of the following criteria will probably have absorption or permeability problems.

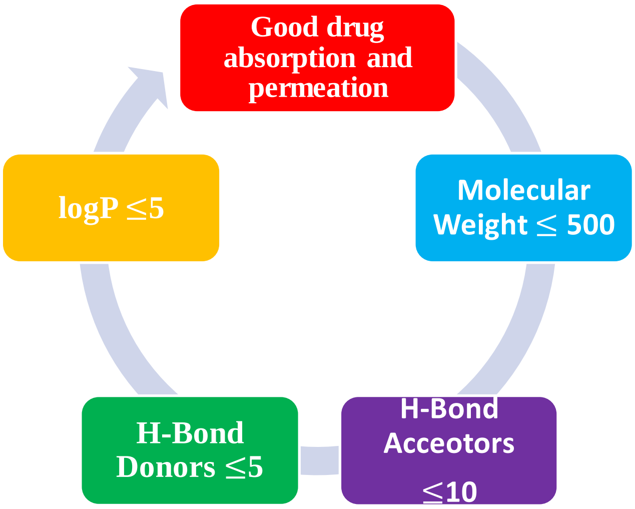


Figure IV.1.12: Lipenski's rule (rule of five).

-Lipophilicity is considered as an important element in the process of solubility, absorption, distribution, metabolism and excretion, as well as pharmacological activity. According to Hansch and Leo, very lipophilic molecules are distributed and stored inside lipid layers of cell membranes. Drugs have a low solubility aqueous when there is a high log P. When the log P is very low, the drugs find it difficult to penetrate into the lipid bilayers of cell membranes. Unlike the hydration energy, lipophilicity increases proportionally with the hydrophobic nature of the substituent groups [19].

- Increasing the number of hydrogen bonds have also a crucial role (Hydrogen bond donors, Hydrogen bond acceptors) in the increasing of the aqueous solubility of compounds (Almi, meanwhile it leads to poor permeability through the membrane bi-layer [20].
- Molecular weight (MW) is one of the factors that determine drug permeability. It is related to the size of molecule. The solubilization of the drugs necessitates the formation of a larger cavity in water. When the size of the molecule increases, it can prevent passive diffusion through the tightly packed aliphatic side chains of the bilayer membrane [16] Johnson et al, 2009).
- Topological polar surface area TPSA and the number of rotatable bonds NRB are two other parameters introduced by Veber which are often used in addition to the rule of "5. For ideal drug absorption the TPSA must be less than 140 Å^2 and the NRB ≤ 10 [19] Warring rule and golden triangle suggest that, the molecular weight and LogD (the distribution coefficient at pH 7.4) are two essential parameters which may affect the permeability, bioavailability and the clearance of candidate drugs [20]. According to Warring rule, compounds with MW < 414 and logD7.4 >1.3 are more likely to be characterized with a high permeability.

Golden Triangle is a visualization tool with a baseline of log D7.4 = _2.0 to log D7.4 = 5.0 at MW= 200 and a peak at log D7.4 = 1, 0–2,0 and MW= 450.These limits were established from measurements of in vitro clearance, in vitro permeability, and computational designed data. Golden Triangle helps chemists to discover metabolically stable, permeable and potent drug candidates, by restricting their molecular weight and coefficient of distribution to locate them within the golden triangle (reside inside the golden triangle) [20].

In this study Lipinski, Veber rules, and golden triangle were used to identify the druglikness properties of the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives series and the new designed the results are shown in Tables IV.8 and IV.9.

Through Table IV.8, the lipophilicity of all compounds is less than 5. All the values are ranging from 1.560 to 4.48. It means that these compounds can be easily solubilized in aqueous and organic solutions [19]. The number of hydrogen bond donors of all compounds is ranging from 4 to 6. All values are less than 5 except comp 6 which has 6 HBD. The number of hydrogen bond acceptors of all compounds is less than 10, and the values range from 4 to 5. The Molecular weight of all compounds is less than 500 Da, however, the minimum value is 294.35 and the maximum value is 408.54. All compounds, except the comp 6, have a null violation for Lipinski's rule. This indicates that these compounds are orally bioavailable and would not have problems with the permeability, because they possess different groups which act as substrates for transporters [7].

According to Veber rule parameters, the thirty 2-amino-3-piperidin-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives except comp 12 and comp 26, are flexible molecules because all their number of rotatable bonds is less than 10. As for the TPSA values of all compounds are less than 140 $Å^2$which indicates that these compounds are characterized by a good cell membrane permeability [2.14].

All the new designed compounds within the range set by Veber’s rule and Lipinski’s rule except the following compounds A1m, A1p, A11m which their HBD are more than five and their TPSA is more than 140 $Å^2$.

Table IV.1.8. Veber and Lipinski properties of the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives

Compound	Lipinski's rule				Veber' rule	
	log *P*	MW	HBD	HBA	TPSA	RB
Cmp 1	2.3436	294.35	4	4	94.96	4
Cmp 2	2.8108	308.38	4	4	94.96	4
Cmp 3	2.0909	324.38	4	5	104.19	5
Cmp 4	2.0909	324.38	4	5	104.19	5
Cmp 5	2.5581	308.38	4	5	94.96	5
Cmp 6	1.5604	309.37	6	4	120.98	4
Cmp 7	2.0592	310.35	5	5	115.19	5
Cmp 8	2.0909	324.38	4	5	104.19	5
Cmp 9	2.4334	338.4	4	5	104.19	6
Cmp 10	2.9020	352.43	4	5	104.19	7
Cmp 11	3.6946	380.48	4	5	104.19	9
Cmp 12	4.4872	408.54	4	5	104.19	11
Cmp 13	2.8465	352.43	4	5	104.19	6
Cmp 14	3.3048	366.46	4	5	104.19	7
Cmp 15	2.8008	364	4	5	104.19	7
Cmp 16	3.1971	378.47	4	5	104.19	7
Cmp 17	3.5934	406.52	4	5	104.19	7
Cmp 18	3.8675	400.47	4	5	104.19	7
Cmp 19	2.2713	294.35	4	4	94.96	4
Cmp 20	1.9869	310.35	5	5	115.19	5
Cmp 21	2.3611	338.4	4	5	104.19	6
Cmp 22	2.8297	352.43	4	5	104.19	7
Cmp 23	3.226	366.46	4	5	104.19	8
Cmp 24	3.6223	380.48	4	5	104.19	9
Cmp 25	4.0186	394.51	4	5	104.19	10
Cmp 26	4.4149	408.54	4	5	104.19	11
Cmp 27	3.2325	366.46	4	5	104.19	7
Cmp 28	2.7285	364.44	4	5	104.19	7
Cmp 29	3.1248	378.47	4	5	104.19	7
Cmp 30	3.7952	400.47	4	5	104.19	7

Table IV.1. 9. Veber and Lipinski properties of the new designed compounds.

Compound	Lipinski's rule				Veber' rule	
	log *P*	MW	HBD	HBA	TPSA	RB
A1m	1.158	458.46	9	4	205.34	10
A1p	1.158	458.46	9	4	205.34	10
A2m	1.95	379.46	4	10	124.06	8
A2p	1.95	379.46	4	10	124.06	8
A3m	2.17	380.44	5	6	121.26	8
A3p	2.17	380.44	5	6	121.26	8
A4m	1.08	415.15	5	7	149.51	8
A4p	1.08	415.15	5	7	149.51	8
A5m	3.5	433.47	5	7	106.99	8
A5p	3.5	433.47	5	7	106.99	8
A6m	1.92	407.51	5	6	116.22	8
A6p	1.92	407.51	5	6	116.22	8
A7m	2.14	408.49	5	6	113.42	8
A7p	2.14	408.49	5	6	113.42	8
A8m	3.27	376.41	4	7	133.88	5
A8p	3.27	376.41	4	7	133.88	5
A9m	3.01	410.48	4	6	104.19	9
A9p	3.01	410.48	4	6	104.19	9
A10m	2.36	340.38	4	6	113.42	6
A10p	2.36	340.38	4	6	113.42	6
A11m	0.89	472.49	9	10	205.34	11

The analysis of golden triangle showed that five compounds among the thirty 2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives reside outside the golden triangle: comp6, comp12, comp20, comp21 and comp26. Therfore, these derivatives are characterized by poor clearance and mild membrane permeability. The rest of derivatives which are located inside the golden triangle can cross the cell membrane easily and with no clearance and toxicity problems. As depicted in Figure IV.1.13, the new designed compounds are located outside the golden triangle except compounds A9m and A9p, which means that, these compounds risk to be subject to clearance and cell membrane problems.

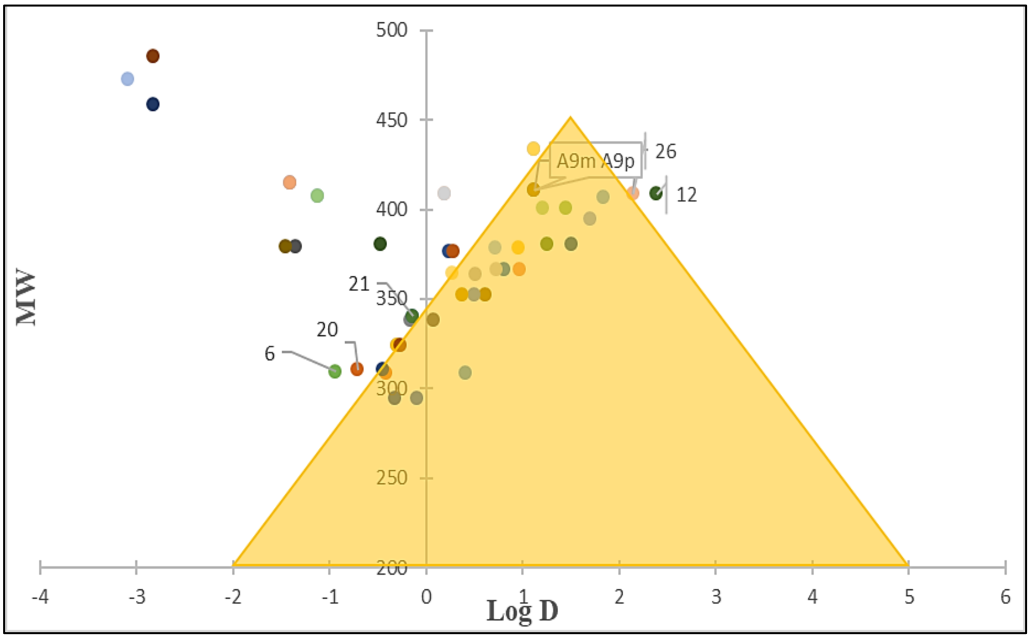


Figure IV.1.13. In vitro permeability and clearance trends across MW and LogD.

IV.4 Conclusion

In this study QSAR model of thirty2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives against IKK-β enzyme were performed based on theoretical molecular descriptors which were selected by stepwise regression. This model was used to predict the biological activity of new designed compounds. The applicability domain of the elaborated QSAR model succeeded to screen 21 compounds with higher biological activity, and most of compounds which were characterized by a substitution at para position showed a higher pIC_{50}values than those were characterized by a substitution at meta position of the phenyl ring. Docking simulation was carried out to position the investigated compounds into the IKK-β enzyme active site. The results revealed that all compounds had poses in the pocket. Among the new designed compounds, four compounds exhibited better binding interaction toIKK-β which are A1m, A1p, A9m, A11, with better binding affinity to the active site. Their XP GScore are respectively -12.375, -10.713, -10.633, -9. 381. Docking study also showed that CYS 99 the GLU149 amino acid residues might be crucial for IKK-β inhibition and the hydroxylic groups of compounds may enhance the affinity of compounds. The compounds to be considered as drug like molecule were evaluated on the basis of Veber rule and Lipinski rule. most of the designed compounds showed no violations as per Veber and Lipinski rule.

CHAPTER IV: Methodology, Results and Discussion

Part 2: Comparative Study Between Many Predictive QSAR models (MLR and ANN Regression)

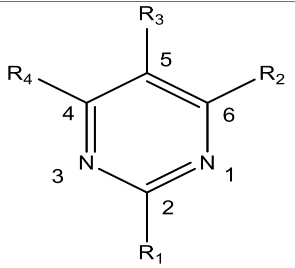
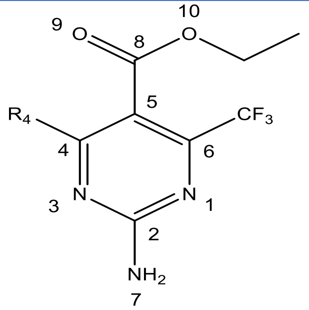
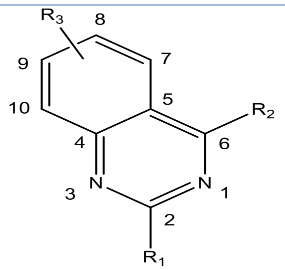
IV.1. QSAR Methodology

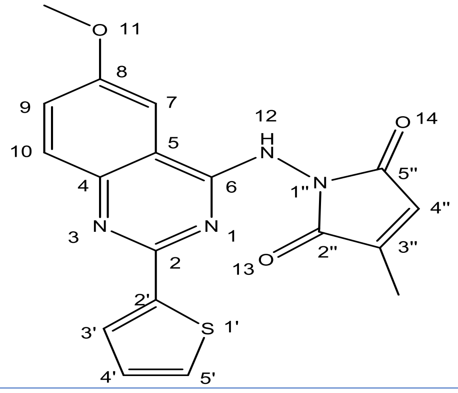
IV.1.1. Data set

One hundred twenty-one compounds were identified in the literature as having the ability to inhibit nuclear factor-κB (NF). The IC_{50} values for these compounds (the concentration required to achieve 50% inhibition) were obtained from the literature and used in this study for QSAR model determination. The chemical structures of the studied compounds are presented in Table IV.2.1.

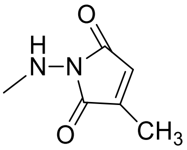
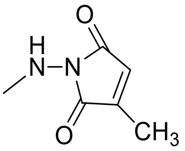
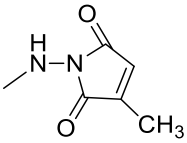
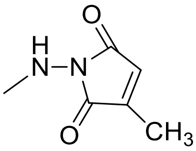
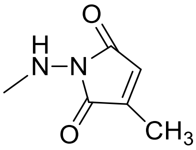
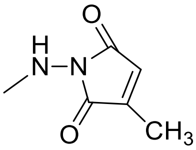
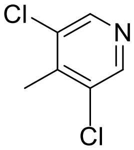
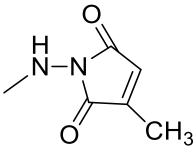
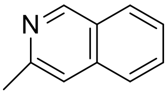
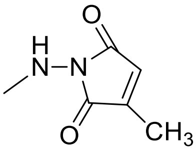
The 121 compounds were randomly divided into two groups: training compounds and test compounds. Sixty-six percent of the compounds were chosen as training sets for the QSAR model's development, with the remaining 39 compounds serving as test sets. It was necessary to transform the corresponding half-maximal inhibitory concentration (IC_{50}) values into pIC_{50} (-logIC_{50}) values in order to perform this analysis.

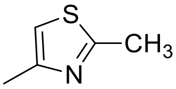
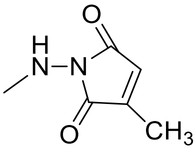
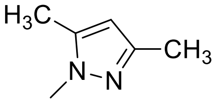
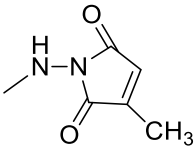
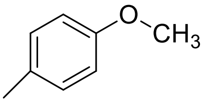
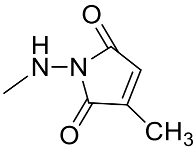
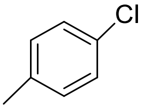
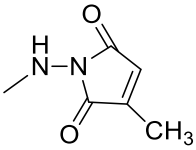
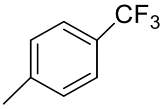
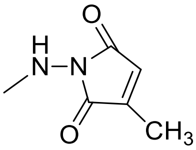
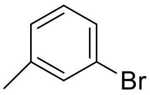
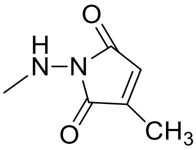
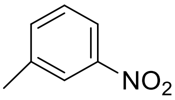
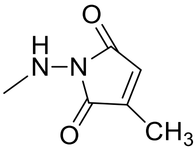
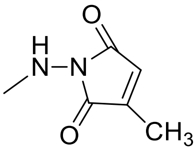
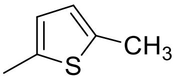
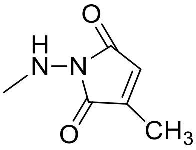
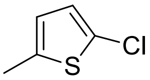
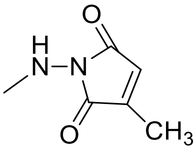
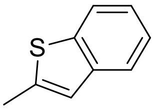
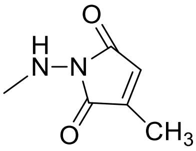
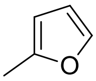
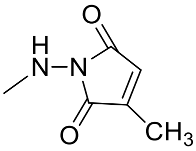
Table IV.2.1: Chemical structures of the investigated compounds.

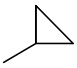
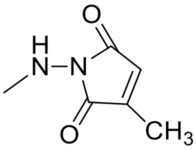
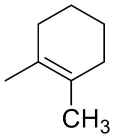
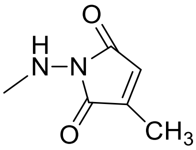
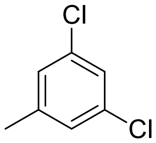
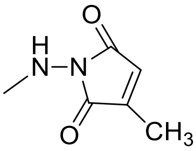
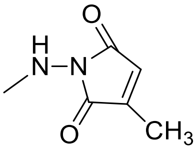
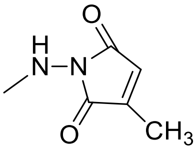
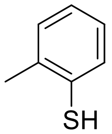
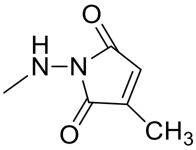
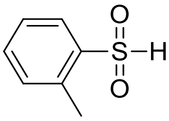
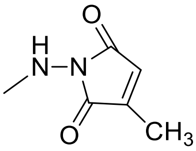
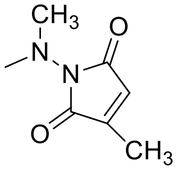
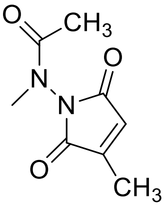
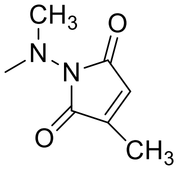
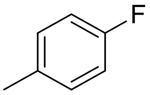
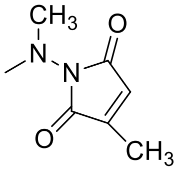
a) General structural formula and numbering of pyrimidine (Compounds from 1-46 and 48-100)
 R3 5 R4 R2 4 6 N N 3 1 2 R1
b) Compound 47
 10 9 O O 8 5 R4 CF3 4 6 N N 3 1 2 NH2 7
c) General structural formula and numbering of quinazoline derivatives (compounds from 101 -103 and 105 to 121)
 R3 8 9 7 10 5 R2 4 6 N N 3 1 2 R1
d) Compound 104

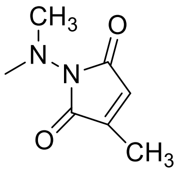
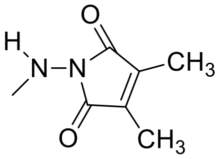
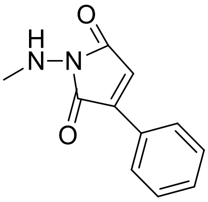
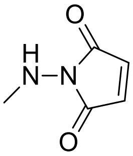
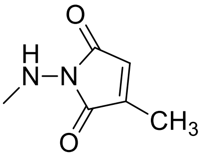
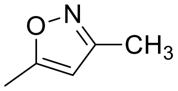
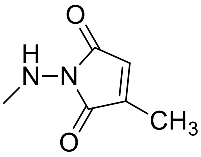
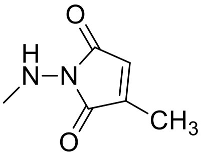
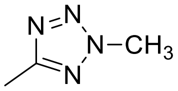
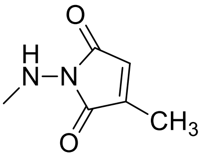
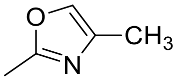
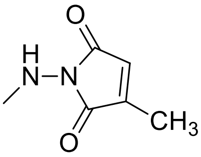
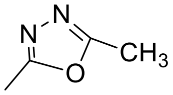
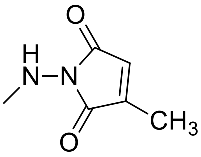
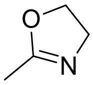
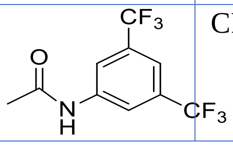


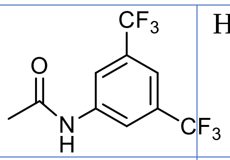
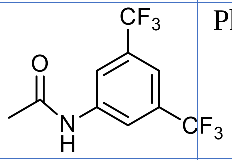
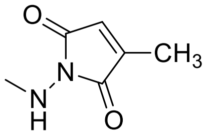
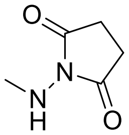
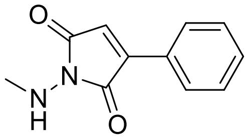
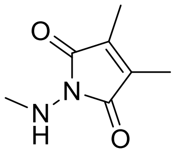
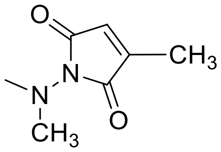
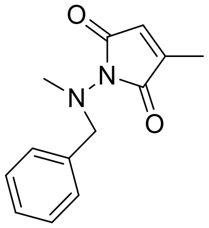
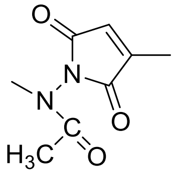
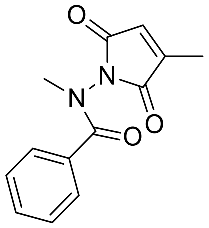
Continued

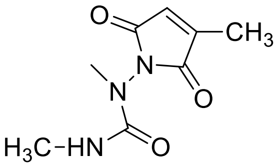
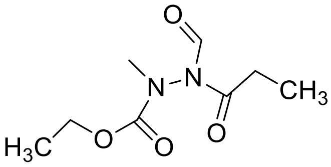
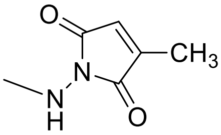
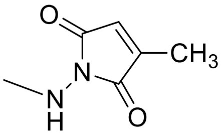
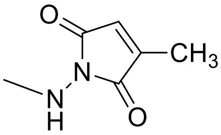
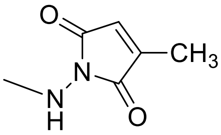
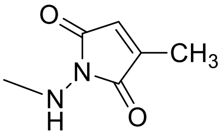
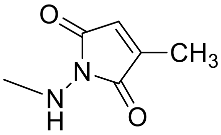
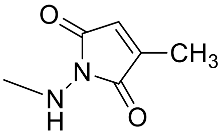
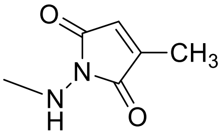
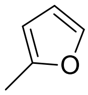
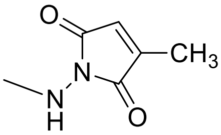
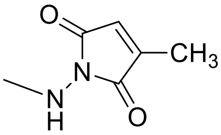
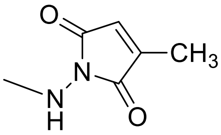
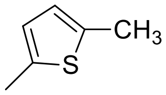
Compounds	R1	R2	R3	R4	R5
Comp 1	CF3		CO2Et	H	-
Comp 2	Me		CO2Et	H	
Comp 3	Et		CO2Et	H	
Comp 4	t-Bu		CO2Et	H	
Comp 5	SMe		CO2Et	H	
Comp 6	4-Pyridyl		CO2Et	H	
Comp 7			CO2Et	H	
Comp 8			CO2Et	H	

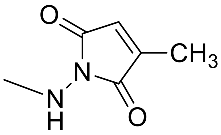
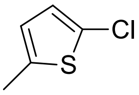
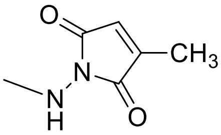
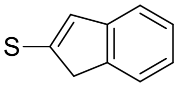
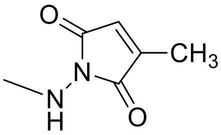
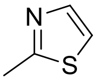
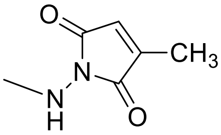
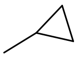
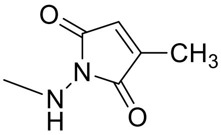
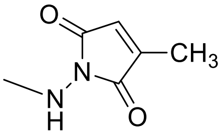
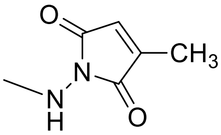
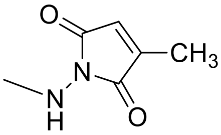
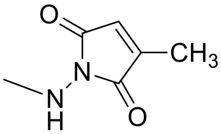
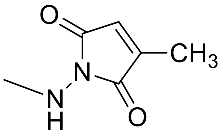
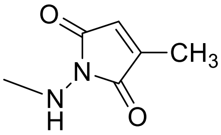
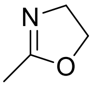
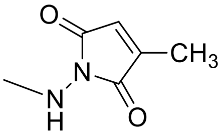
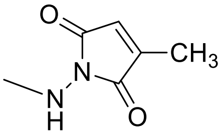
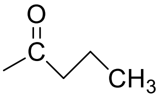
Comp 9			CO2Et	H	
Comp 10			CO2Et	H	
Comp 11			CO2Et	H	
Comp 12			CO2Et	H	
Comp 13			CO2Et	H	
Comp 14			CO2Et	H	
Comp 15			CO2Et	H	
Comp 16	3-Thienyl		CO2Et	H	
Comp 17			CO2Et	H	
Comp 18			CO2Et	H	
Comp 19			CO2Et	H	
Comp 20			CO2Et	H	

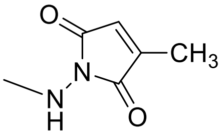
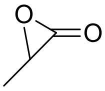
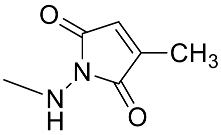
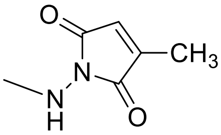
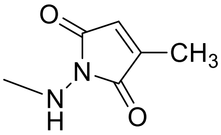
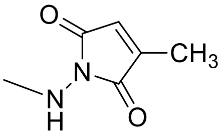
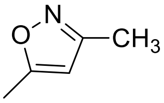
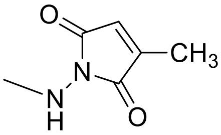
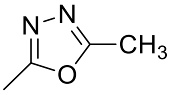
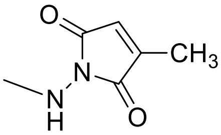
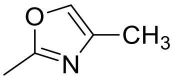
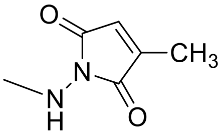
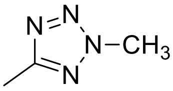
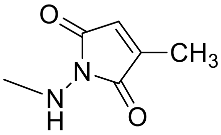
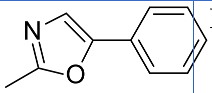
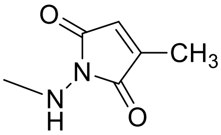
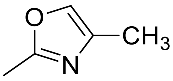
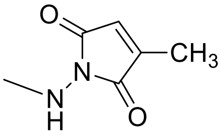
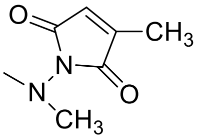
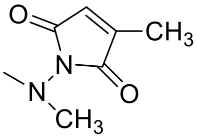
Comp 21			CO_2Et	H	
Comp 22			CO_2Et	H	
Comp 23			CO_2Et	H	
Comp 24	Benzyl		CO_2Et	H	
Comp 25	Phenoxy		CO_2Et	H	
Comp 26			CO_2Et	H	
Comp 27			CO_2Et	H	
Comp 28	Phenyl		CO_2Et	H	
Comp 29	Phenyl		CO_2Et	H	
Comp 30	2-Thienyl		CO_2Et	H	
Comp 31			CO_2Et	H	

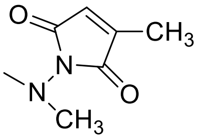
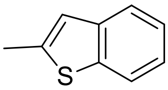
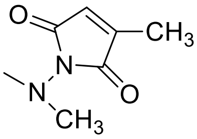
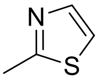
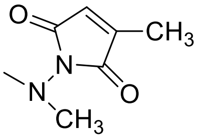
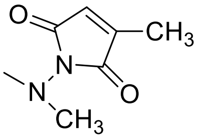
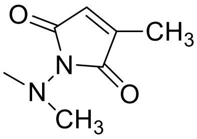
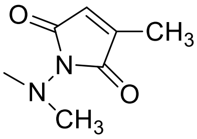
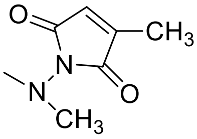
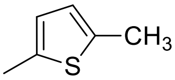
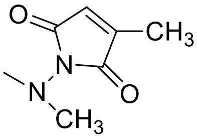
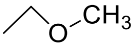
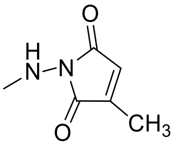
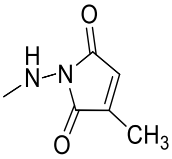
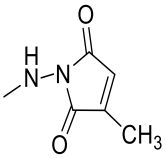
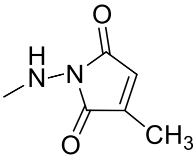
Comp 32	Et		CO_2Et	H	
Comp 33	CF_3		CO_2Et	H	
Comp 34	CF_3		CO_2Et	H	
Comp 35	CF_3		CO_2Et	H	
Comp 36	2-Thienyl			H	
Comp 37	2-Thienyl		CN	H	
Comp 38	2-Thienyl			H	
Comp 39	2-Thienyl			H	
Comp 40	2-Thienyl			H	
Comp 41	2-Thienyl			H	
Comp 42	Cl	H		CF_3	

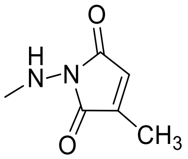
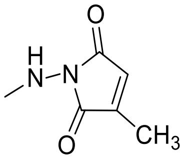
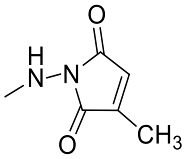
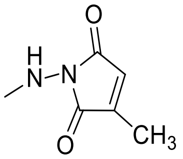
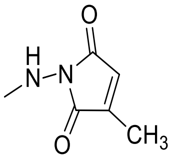
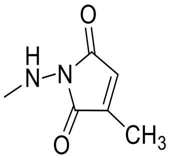
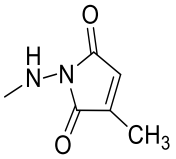
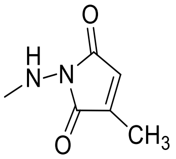
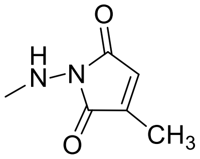
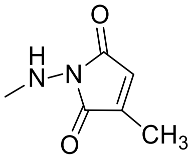
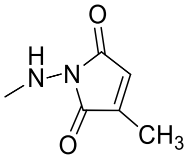
Comp 43	Cl	H		H	
Comp 44	Cl	H		Phenyl	
Comp 45		CF_3	CO_2Et	H	
Comp 46		CF_3	CO_2Et	H	
Comp 47	NH_2	CF_3	CO_2Et	H	
Comp 48		CF_3	CO_2Et	H	
Comp 49		CF_3	CO_2Et	H	
Comp 50		CF_3	CO_2Et	H	
Comp 51		CF_3	CO_2Et	H	
Comp 52		CF_3	CO_2Et	H	
Comp 53		CF_3	CO_2Et	H	

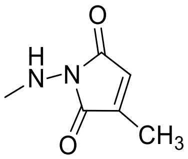
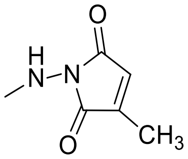
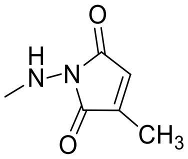
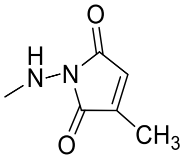
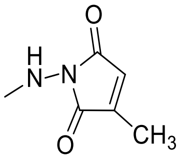
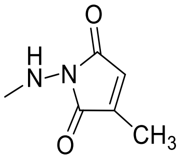
Comp 54		CF_3	CO_2Et	H	
Comp 55		CF_3	CO_2Et	H	
Comp 56		H	CO_2Et	H	
Comp 57		Me	CO_2Et	H	
Comp 58		Et	CO_2Et	H	
Comp 59		CF_2CF_3	CO_2Et	H	
Comp 60		Phenyl	CO_2Et	H	
Comp 61		Benzyl	CO_2Et	H	
Comp 62		CH_2OMe	CO_2Et	H	
Comp 63			CO_2Et	H	
Comp 64		2-Thienyl	CO_2Et	H	
Comp 65		3-Thienyl	CO_2Et	H	
Comp 66			CO_2Et	H	

Comp 67			CO_2Et	H	
Comp 68			CO_2Et	H	
Comp 69			CO_2Et	H	
Comp 70			CO_2Et	H	
Comp 71		CF_3	$COCMe_3$	H	
Comp 72		CF_3	COOH	H	
Comp 73		CF_3	$CONH_2$	H	
Comp 74		CF_3	$CONMe_2$	H	
Comp 75		CF_3	Acetyl	H	
Comp 76		CF_3	Phenoxy	H	
Comp 77		CF_3		H	
Comp 78		Et	Phenoxy	H	
Comp 79		Et		H	

Comp 80		Et		H	
Comp 81		Et	CH₂OH	H	
Comp 82		Et	CN	H	
Comp 83		Me	Acetyl	H	
Comp 84		Et		H	
Comp 85		Et		H	
Comp 86		Et		H	
Comp 87		Et		H	
Comp 88		Et		H	
Comp 89		2-Thienyl		H	
Comp 90		2-Thienyl	CO-t-Bu	H	
Comp 91		CF₂CF₃	CO₂Et	H	
Comp 92		Et	CO₂Et	H	

Comp 93			CO_2Et	H	
Comp 94			CO_2Et	H	
Comp 95		Et	CO_2Et	H	
Comp 96		CF_2CF_3	CO_2Et	H	
Comp 97		2-Thienyl	CO_2Et	H	
Comp 98		3-Thienyl	CO_2Et	H	
Comp 99			CO_2Et	H	
Comp 100		Et		H	
Comp 101	2-Thienyl				
Comp 102	Phenyl				
Comp 103	CF_3				
Comp 104	2-Thienyl				7-OMe

Comp 105	2-Thienyl				8-OMe
Comp 106	2-Thienyl				9-OMe
Comp 107	2-Thienyl				10-OMe
Comp 108	2-Thienyl				8,9-Di-OMe
Comp 109	2-Thienyl				8,9,10-Tri-OMe
Comp 110	2-Thienyl				7-F
Comp 111	2-Thienyl				8-Cl
Comp 112	2-Thienyl				7-Me
Comp 113	2-Thienyl				9-(NMe₂)
Comp 114	CF₃				7-OMe
Comp 115	CF₃				8-OMe

Comp 116	CF_3				7-OMe
Comp 117	CF_3				7-Me
Comp 118	CF_3				8-SMe
Comp 119	CF_3				8-OH
Comp 120	CF_3				9-5(1-Piperridyl)
Comp 121	CF_3				9-(NMe_2)

IV.1.2. Compound preparation and calculation methods

The two-dimensional (2D) structures of the compounds under investigation were drowned using the ChemDraw ultra 12.0 software. The structures were saved in the SDF file format and then directly converted to 3D structures when they were imported into the Materials studio 8.0 software program. To obtain the quantum chemical parameters, the geometric optimization of these compounds was carried out as a first step (Molecular descriptors). The quantum chemical calculations were carried out using the GGA (generalized gradient approximation) within the framework of the density functional theory (DFT), with the help of the software package DMol3 in the Materials Studio software. In order to achieve geometric optimization of all investigated compounds, the Double Numerical with Polarization (DNP) basis set and the functional exchange-correlation were used in conjunction (BP) [21].

IV.1.3. Molecular Descriptors Calculation and selection

Materials studio software and the Swiss ADME online tool were used to generate a total of more than 100 molecular descriptors example of them is mentioned in Table IV.2.2. The QSAR model was developed on the basis of 82 compounds. Multiple Linear Regression

(MLR) was used to quantify this model, which correlates the dependent variables pIC_{50} and the significant parameters (independent variables). Stepwise regression was used to derive the QSAR model coefficients.

the parameter cost function (stopping rule) selected as minimum AICc, can be described as follow:

$$AIC = 2K - 2Ln\,L \qquad \text{Eq 1}$$

$$AIC_c = AIC + \frac{2k^2+2k}{p-k-1} \qquad \text{Eq 2}$$

K: *the number of estimated parameters in the model*
L: *represents the maximum value of the likelihood function in the model*
P: *represents the number of training experimental data points.*

$When: p \rightarrow \infty$, AICc becomes AIC because the extra corrected penalty term in AICc converges to zero. Using the stepwise AICc algorithm, only important descriptors that improved the model's information criteria were added to the model, while descriptors that had a negligible influence were removed from the model using the stepwise approach [22].

Following that, an analysis of variance was used to identify molecular descriptors with a high statistical significance for predicting the NF-κB inhibitory concentration. The objectives of this part are to develop a simplified MLR model with fewer terms that accurately predicts the NF-κB inhibitory concentration and to estimate the coefficients of the model's significant terms. Table IV.2.3 summarizes the findings of this analysis.

Table IV.2.2: Calculated Molecular Descriptors and their types

Molecular Descriptors	Type
Wiener Index, Information Content Indices, Topological Polar Surface Area (TPSA)	Topological Descriptors
Number of Rotatable Bonds (RTB), Molecular Weights	Geometrical Descriptors
Energy of the highest occupied molecular orbitals (HOMO) Energy of the lowest unoccupied molecular orbitals (LUMO) Ionization potential (I), Electron affinity (A), Energy band gap (ΔE) Global hardness (η), Chemical potential (μ), Electrophilicity index (ω)	Electronic descriptors
Molar Reactivity (MR), Number of hydrogen bond Doner (HBD), Number of Hydrogen Bond Acceptors, Lipophelicity LogP, Hydro solubility (Log S).	Physicochemical descriptors

Table IV.2.3: Predictor coefficients of the MLR algorithm

Terme	*Estimation*	*Standard Erreur*	$\lvert t_{ratio} \rvert$	p_{value}
$\boldsymbol{Intercept}$	0.2872344	1.375232	0.21	0.8352
$\boldsymbol{LUMO}$	0.4568486	0.178515	2.56	0.0127*
$\boldsymbol{\omega}$	0.1975915	0.045106	4.38	<.0001*
$\boldsymbol{AlogP98}$	0.5338582	0.057273	9.32	<.0001*
$\boldsymbol{Weiner\ index}$	-0.002733	0.000542	5.05	<.0001*
$\boldsymbol{Kappa - 2}$	-2.374863	0.420752	5.64	<.0001*
$\boldsymbol{Kappa - 3\ indices\ alpha\ modified}$	2.7424539	0.482879	5.68	<.0001*
$\boldsymbol{Subgraph\ counts}\ (\boldsymbol{0 - 3})$	0.3890278	0.081594	4.77	<.0001*
$\boldsymbol{E - states\ keys\ sums}: \boldsymbol{S - ssCH2}$	-0.217393	0.044117	4.93	<.0001*
$\boldsymbol{E - states\ keys\ sums}: \boldsymbol{S - sSH}$	-0.148803	0.06979	2.13	0.0365*
$\boldsymbol{E - states\ keys\ sums}: \boldsymbol{S - sPH2}$	0.4030164	0.083699	4.82	<.0001*
$\boldsymbol{E - states\ keys\ sums}: \boldsymbol{S - sBr}$	0.040816	0.017443	2.34	0.0221*

By analysing Table IV.2.3, it can be seen that the most important significant parameters were chosen to be used in QSAR model construction. Out of the 130 descriptors considered for each compound in the dataset, only 11 proved to be significant: LUMO, ω, AlogP98, Weiner index, Kappa-**2** indices alpha modified, Kappa -3 indices alpha modified, subgraph counts (0-3), E-states keys sums: S-ssCH2, E-states keys sums: S-sSH, E-states keys sums: S-sPH2, and E-states keys sums: S-sBr showed a high correlation with pIC_{50} values of compounds. While the remaining 119 were eliminated from the model by setting their coefficient values to 0. The significant descriptors influence the inhibitory activity of compounds toward the NF-κB. The selection of these parameters was carried out according to the coefficient of determination and the P-value of the regression. If the parameters have a p-value less than 0.05, they are considered statistically significant. The obtained values of the parameters are represented in together with their statistical properties in Table IV.3. According to |t_ratio| values, it can be seen that *AlogP98* descriptor is the most statistically significant descriptor in the model, with the highest $\lvert t_{Ratio} \cdot \rvert$

IV.1.4. 2-D QSAR -MLR Model Generation

The best obtained QSAR-MLR model is given as follows:

$$\boldsymbol{pIC}\mathbf{50} = \mathbf{0.287} + \mathbf{0.45} * LUMO + \mathbf{0.197} * \omega + \mathbf{0.53} * \ AlogP98 - \mathbf{0.002} * \ Weiner\ index - \mathbf{2.37} * \ Kappa - 2 + \mathbf{2.74} * Kappa - 3\ indices\ alpha\ modifie + \mathbf{0.38} * \ Subgraph\ counts\ (0-3) - \mathbf{0.21} * \ E - states\ keys\ sums: S - ssCH2 - \mathbf{0.14} * E - states\ keys\ sums: S - sSH + \mathbf{0.40} * \ E - states\ keys\ sums: S - sPH2 - \mathbf{0.04} * \ E - states\ keys\ sums: S - sBr$$

IV.1.5. Model Evaluation

For the purpose of determining the mathematical model that describes the variation in inhibitory concentration as a function of the descriptors listed above, appropriate statistical methods were used to determine the best fitting of the model. According to the statistical significances indicated Table IV.4, the above equation shows strong correlation between activity and the eleven explanatory variables, with a correlation coefficient of 93% (therefore close to 100%). The relatively high values of R^2=0.879 and R_A^2= 0.860 and a low value of the root-mean-square deviation (RMSE=0.296) indicate that the proposed model is reliable and well adjusted. The regression equation has statistical significance, which is confirmed by the high value of F = 46.50 (the coefficients of the selected variables in the equation are not equal to zero). Regarding its robustness, the $\boldsymbol{Q^2}$ Coefficient shows that the model is robust and stable (due to the stability of its parameters with respect to the molecules of the training set) with good predictability. A low $\boldsymbol{R_{pred}^2}$ = 0.30 indicates an acceptable external predictive potential of this obtained QSAR model.

The contribution of each parameter to the investigated biological activity is determined by the value of the corresponding coefficient as well as the sign that precedes the coefficient value. The following significant descriptors: E-states keys sums: S-ssCH2, E-states keys sums: S-sSH and kappa-2, are preceded by a negative sign. As a result, these parameters will decrease the value of the dependent variable (pIC_{50}). while the remaining significant parameters: LUMO, ω, AlogP98, Weiner index, Kappa -3 indices alpha modified, subgraph counts (0-3), E-states keys sums: S-sPH2, and E-states keys sums: S-sBr are preceded by a positive sign; consequently, their influence will increase the pIIC_{50} value. Therefore, it can be concluded that the most active predicted compounds are characterized by E-states keys sums: S-ssCH2, E-states keys sums: S-sSH and kappa-2, which should not be elevated. The more the LUMO, ω, AlogP98, Weiner index, Kappa -3 indices alpha modified, subgraph counts (0-3), E-states keys sums: S-sPH2, and E-states keys sums: S-sBr values increase, the more the (pIC_{50}) value increases.

Table IV.2.4: Statistical parameters obtained by the MLR model and their threshold values.

Parameter	Obtained Values	Threshold
R^2	0.879	> 0.6
R	0.932	> 0.6
R_A^2	0.860	> 0.6
RMSE	0.296	Low value
F	46.507	High value
Q^2	0.880	> 0.5
R_{pred}^2	0.30	> 0.5

Figure IV.2.1 illustrates a linear relationship between predicted and experimental biological activities, demonstrating that the created model is internally predictable. Also, the MLR model has well-matched the experimental data, with the exception of a few scattered points. Therefore, it may be used to predict the pIC_{50} values with confidence. Furthermore, when the estimated residuals are plotted versus the experimental activity levels in Figure IV.2.2, the residuals are distributed around the zero line, and most of them are less than an absolute value of 0.5, indicating that the model is free of systematic errors.

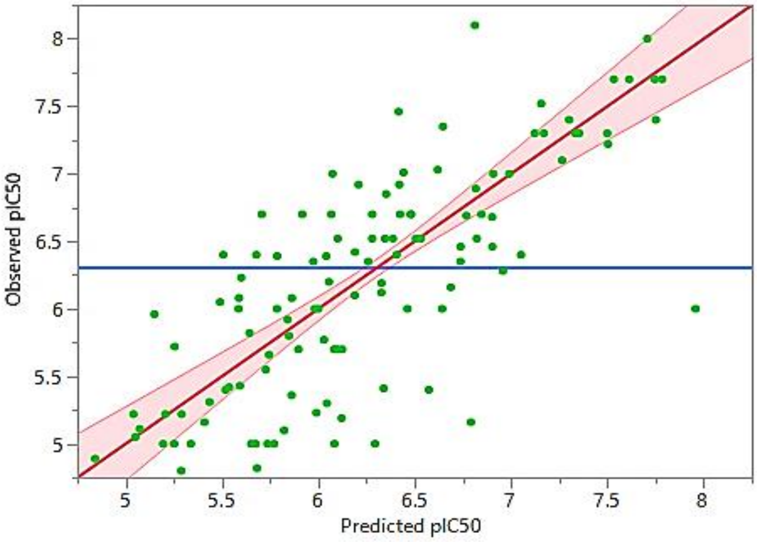


Figure IV.2.1: Experimental versus predicted pIC_{50}in MLR model

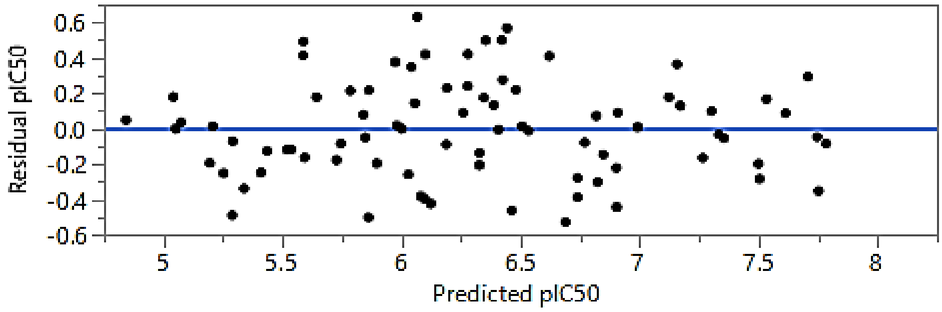


Figure IV.2.2: Residual pIC_{50} versus predicted pIC_{50} MLR model

For both the training and testing sets, the model was used to predict the pIC_{50} values. Table IV.2.5 represents the predicted and observed values of the investigated compounds as well as the residual. Since it can be seen that most of the predicted values are too much close to those obtained experimentally. The experimental pIC_{50} values of the investigated compounds range from 4.52 to 8.52. And the predicted pIC_{50} values range from 4.84 to 7.96.

Table IV.2.5: Experimental and predicted values of the studied compounds based on the MLR model

Compounds	**Exp(pIC_{50})**	**Prd(pIC_{50})**	**Residual**	**Compounds**	**Exp(pIC_{50})**	**Prd(pIC_{50})**	**Residual**
Compound 1	6.00	5.98	0.00	Compound 62	5.22	5.29	0.00
Compound 2	4.80	5.29	0.24	Compound 63	6.00	5.78	0.05
Compound 3	5.00	5.65	0.42	Compound 64	6.85	6.35	0.25
Compound 4	6.00	6.46	0.21	Compound 65	6.70	6.07	0.40
Compound 5	6.70	6.48	0.05	Compound 66	7.35	6.65	0.49
Compound 6	5.22	5.04	0.03	Compound 67	6.28	6.96	0.46
Compound 7	5.70	5.90	0.04	Compound 68	6.10	6.19	0.01
Compound 8	5.22	5.21	0.00	Compound 69	5.66	5.74	0.01
Compound 9	5.10	5.82	0.52	Compound 70	5.82	5.64	0.03
Compound 10	5.00	5.68	0.46	Compound 71	6.68	6.90	0.05
Compound 11	6.05	5.49	0.31	Compound 72	4.52	5.68	1.35
Compound 12	6.52	6.10	0.18	Compound 73	4.52	5.26	0.55
Compound 13	6.40	5.51	0.80	Compound 74	4.52	5.62	1.22
Compound 14	6.16	6.69	0.28	Compound 75	5.36	5.86	0.25
Compound 15	5.16	6.79	2.66	Compound76	7.01	6.44	0.32
Compound 16	7.00	6.07	0.86	Compound 77	5.19	6.12	0.87
Compound 17	6.69	6.77	0.01	Compound 78	6.19	6.33	0.02
Compound 18	6.52	6.82	0.09	Compound 79	6.35	5.97	0.14
Compound 19	5.40	6.57	1.38	Compound 80	6.92	6.21	0.51
Compound 20	6.70	5.71	0.99	Compound 81	5.31	5.43	0.02
Compound 21	6.00	5.59	0.17	Compound 82	5.96	5.15	0.66
Compound 22	5.7	6.10	0.16	Compound 83	5.11	5.07	0.00

Compound 23	6.52	6.51	0.00	Compound 84	5.00	5.77	0.59
Compound 24	5.16	5.41	0.06	Compound 85	5.00	5.25	0.06
Compound 25	4.82	5.68	0.74	Compound 86	6.08	5.86	0.05
Compound 26	6.00	6.00	0.00	Compound 87	5.00	5.19	0.04
Compound 27	5.00	5.34	0.11	Compound 88	5.55	5.73	0.03
Compound 28	5.70	6.12	0.18	Compound 89	6.12	6.33	0.04
Compound 29	5.40	5.52	0.01	Compound 90	7.30	7.17	0.02
Compound 30	6.00	6.64	0.41	Compound 91	6.35	6.74	0.15
Compound 31	5.23	5.99	0.58	Compound 92	7.46	6.42	1.09
Compound 32	5.00	6.08	1.17	Compound 93	6.89	6.82	0.01
Compound 33	5.00	6.29	1.67	Compound 94	5.77	6.03	0.07
Compound 34	4.70	5.67	0.94	Compound 95	6.39	6.04	0.12
Compound 35	5.00	5.74	0.54	Compound 96	6.46	6.74	0.08
Compound 36	6.52	6.53	0.00	Compound 97	7.03	6.62	0.17
Compound 37	5.30	6.04	0.55	Compound 98	6.92	6.42	0.25
Compound 38	6.70	5.92	0.61	Compound 99	6.46	6.90	0.20
Compound 39	6.70	6.42	0.08	Compound100	6.20	6.06	0.02
Compound 40	6.52	6.28	0.06	Compound 101	7.22	7.50	0.08
Compound 41	6.70	6.28	0.18	Compound 102	7.00	6.99	0.00
Compound 42	7.30	7.12	0.03	Compound 103	8.10	6.81	1.66
Compound 43	6.52	6.35	0.03	Compound 104	8.52	7.53	0.99
Compound 44	6.40	6.41	0.00	Compound 105	7.70	7.61	0.01
Compound 45	5.70	6.08	0.15	Compound 106	7.30	7.33	0.00
Compound 46	5.80	5.85	0.00	Compound 107	7.30	7.50	0.04
Compound 47	4.52	5.01	0.24	Compound 108	7.40	7.30	0.01
Compound 48	5.43	5.59	0.03	Compound 109	7.30	7.35	0.00
Compound 49	5.41	6.34	0.86	Compound 110	7.70	7.53	0.03
Compound 50	6.52	6.39	0.02	Compound 111	7.70	7.75	0.00
Compound 51	6.35	6.26	0.01	Compound 112	7.40	7.75	0.12
Compound 52	6.39	5.79	0.36	Compound 113	8.00	7.71	0.09
Compound 53	6.08	5.59	0.24	Compound 114	7.52	7.16	0.13
Compound 54	6.23	5.60	0.40	Compound 115	6.40	7.05	0.42
Compound 55	6.42	6.19	0.05	Compound 116	7.10	7.26	0.03
Compound 56	4.89	4.84	0.00	Compound 117	6.00	7.96	3.83
Compound 57	5.72	5.25	0.22	Compound 118	5.05	5.05	0.00
Compound 58	6.40	5.68	0.52	Compound 119	7.00	6.91	0.01
Compound 59	6.70	6.48	0.05	Compound 120	6.70	6.85	0.02
Compound 60	5.92	5.84	0.01	Compound 121	7.70	7.78	0.01
Compound 61	5.42	5.54	0.01				

IV.1.6. ANN models Set up

After ensuring the robustness of the developed MLR-QSAR model with good statistical qualities, the selected eleven molecular descriptors have been employed in the development of four ANN models as the input layer to build artificial neural networks with four layers (ANN). These models are characterized by the same input and output layers, having two hidden layers, and they are different in the number of nodes which is a critical parameter that has a significant impact on the correctness and complexity of the produced model [22.23]. The network consists of several neuron nodes. Direct communication activation functions that include the information necessary to create the output connect the neurons together. The hyperbolic tangent sigmoid activation function of each hidden neuron (H_k)can be obtained as follows:

$$H_k = tngh(\frac{1}{2}Y_k) \qquad \text{Eq3}$$

This function allows the transformation of Y_k values to be comprised between -1 and 1[24]. $Y_{k:}$ represents the linear combination of the inputs associated with hidden neuron which can be obtained as follows:

$$Y_k = \sum_{k=1}^{M}(W_{k,iput})\,(D_i)+b_k \qquad \text{Eq4}$$

$W_{k,iput}$
is the weight coefficient of the link between each input and hidden neuron k
b_k *represents the intercept bias of hidden neuron k*

The ANN models have been generated using the same molecular descriptors as those used in the MLR model. The statistical parameters are shown in Table IV.2.6. By analysing the obtained results, it was found that the models present small values of R^2_{Tr} *and* $R^2_{validation}$Indicating the non-existence of a non-linear regression between the 11 selected descriptors and the biological activities of the used compounds. Therefore, the ANN models are not reliable, and they cannot be predictive.

Table IV.2.6: statistical parameters in several neural network architectures based on the number of hidden neurons

	Layer 1	Layer 2	R^2_{Tr}	$R^2_{validation}$	$RMSE_{Tr}$	$RMSE_{validation}$
Model 1	5 nodes	5 nodes	0.301	0.653	0.841	0.462
Model 2	5 nodes	7 nodes	0.520	0.691	0.701	0.434
Model 3	5 nodes	9 nodes	0.462	0.712	0.732	0.43
Model 4	5 nodes	11 nodes	0.586	0.691	0.588	0.433

The desired ANN should have high generalizability and a low probability of overfitting. For that reason, A novel ANN model has been generated, constituted of four layers. Eight new selected significant descriptors constitute the first layer. (The selection of the significant descriptors has been performed based on the stepwise approach, including 70 descriptors). Tow hidden layers and output layer represent the biological activity $\mathrm{p}IC_{50}$. Before the training process, the database was divided into two subsets: training set 66 %, and test set 34 % of studied compounds.

Various ANN models were constructed, trained, and evaluated to determine the optimum ANN architecture with the highest prediction ability. The selected ANN model is the one with the smallest root mean square error (RMSE) value and a high value of the coefficient of determination R^2.

The best ANN network architecture found after wide training was [8-11-11-1] (**Figure IV.2.3**) with two hidden layers having 11 and 11 neurons. The obtained ANN model can be expressed as follows:

$$pIC50 = 5.88 + 0.90 * H1_1 - 0.67 * H1_{10} - 0.90 * H1_{11} + 0.08 * H1_2 + 0.35 * H1_3 + 0.94 * H1_4 - 0.19 * H1_5 + 0.73 * H1_6 + 2.01 * H1_7 + 2.1 * H1_8 - 1.48 * H1_9$$

Where the hidden neurons: $H1_1$,$H1_2$, $H1_3$, $H1_4$,$H1_5$,$H1_6$, $H1_7$, $H1_8$,$H1_9$,$H1_{10}$,$H1_{11}$ are given in Table **IV.2.7**

Table IV.2.7: Hidden neurons equations of the obtained [8.11.11.1] ANN

Hidden neurons equations of layer 1
H1_1= TanH (-0.51- 0.21* H2_1 -0.25* H2_10 +0.20 * H2_11 + -0.97 * H2_2 + 0.69 *H2_3 + -0.24 * H2_4 + 0.12 * H2_5 -1.06*H2_6 + 0.238323151885082 * H2_7 + -1.37* H2_8 + -0.82 * H2_9)
H1_2= TanH (0.23 - 0.53 * H2_1 + 0.33 * H2_10 +0.46 * H2_11 + 0.71 * H2_2 -0.15 * H2_3 + 0.23 * H2_4 - 0.12* H2_5 -0.19 * H2_6 - 0.93 * H2_7 + 0.09* H2_8 + 0.33 * H2_9)
H1_3 = TanH (0.28 -1.34* H2_1 + 0.26 * H2_10 +0.06 * H2_11 - 0.08 * H2_2 - 0.79 * H2_3 + 0.21* H2_4 + 0.06 * H2_5 + 0.18 * H2_6 + 0.43 * H2_7 - 0.87 * H2_8 -0.20 * H2_9)
H1_4= TanH (0.08 - 0.64* H2_1 - 0.01 * H2_10 +0.46 * H2_11 + 0.30 * H2_2 - 0.79 * H2_3 + 0.07* H2_4 - 0.13 * H2_5 + 0.92* H2_6 + 0.32 * H2_7 + 0.11* H2_8 + 0.33 * H2_9)
H1_5 = TanH (-0.80 + 0.43 * H2_1 + 0.47 * H2_10 +0.39 * H2_11 -0.16 * H2_2 + 0.70 *H2_3 + 0.07* H2_4 - 0.30 * H2_5 + 1.36 * H2_6 + 0.13 * H2_7 - 0.08 * H2_8 +0.14 * H2_9)
H1_6 = TanH (-0.37 + 0.94 * H2_1 - 0.61 * H2_10 +0.007 * H2_11 + 0.42 * H2_2 + 0.43*H2_3 + - 0.35 * H2_4 -0.21 * H2_5 -0.00007 * H2_6 + 0.37 * H2_7 -0.63 * H2_8 -0.04 * H2_9)
H1_7 = TanH (-0.39+ 0.16* H2_1 - 0.29 * H2_10 -0.90 * H2_11 + 0.25 * H2_2 + 1.11 *H2_3 + -0.24 * H2_4 - 0.50 * H2_5 - 0.19 * H2_6 + 0.65 * H2_7 - 0.78 * H2_8 + 1.0* H2_9)

H1_8= TanH (0.96 - 0.41 * H2_1 + -0.28 * H2_10 +0.53 * H2_11 + 0.05 * H2_2 + 0.23 * H2_3 - 0.38* H2_4 + 0.98 * H2_5 + -0.04 * H2_6 - 0.32 * H2_7 + 1.10 * H2_8 + 0.009* H2_9)
H1_9 = TanH (0.47 -0.31 * H2_1 + 0.67 * H2_10 +0.49 * H2_11 + 0.14 * H2_2 + 0.52 * H2_3 + 0.06 * H2_4 - 0.27 * H2_5 + 0.21*H2_6 + 0.14 * H2_7 -0.25 * H2_8 -0.01 * H2_9)
H1_10 = TanH (-0.73 -0.005 * H2_1 + 0.17 * H2_10 -0.06 * H2_11 -0.28 * H2_2 -0.04 * H2_3 -0.46 * H2_4 -1.24 * H2_5 +0.62 * H2_6 + 0.37 * H2_7 - 0.25 * H2_8 - 0.13 * H2_9)
H1_11= TanH (-0.50-0.12 * H2_1 + 0.81 * H2_10 +0.589 * H2_11 + 0.24 * H2_2 - 0.42 * H2_3 + -0.65 * H2_4 -0.16 * H2_5 - 0.90 *H2_6 + 0.96 * H2_7 + 0.53 * H2_8 + 0.20* H2_9)
Hidden neurons equations of layer 2
H2_1 = TanH (5.46 -0.61 * HBA - 0.21 * HBD -1.41 * HOMO + 0.86 * LUMO -0.20 * RB +0.75* η + 0.35 * μ -0.41 * ω)
H2_2 = TanH (7 -0.18* HBA -0.44 * HBD +0.19 * HOMO + -0.05* LUMO + 0.167 * RB -3.06 * η + 0.55 * μ + -0.001 * ω)
H2_3= TanH (-17.53 -1.005 * HBA + 0.51 * HBD -1.86 * HOMO + 1.19 * LUMO + -0.50 * RB -0.14 * η + -3.79 * μ + 0.28 * ω)
H2_4 = TanH (11.36 -0.009 * HBA + 0.29* HBD -0.05 * HOMO + 1.35 * LUMO + 0.12 * RB +0.33 * η + 0.73 * μ -0.37 * ω)
H2_5= TanH (13.93 + 0.32 * HBA -1.83 * HBD +1.89 * HOMO + 1.90 * LUMO + 0.68 * RB+0.63* η + 0.04 * μ -0.040 * ω)
H2_6 = TanH (37.85 -0.35 * HBA + 0.63* HBD +0.38 * HOMO + 3.97 * LUMO + 0.17 * RB +1.57 * η + 3.24* μ -0.59 * ω)
H2_7= TanH (-15.03 - 0.50 * HBA -0.22* HBD+0.33* HOMO + 0.28 * LUMO + -0.06* RB + 0.47 * η -2.94 * μ + 0.68 * ω)
H2_8 = TanH (33.92+ 0.13 * HBA -0.40* HBD+1.70 * HOMO + 1.79 * LUMO -0.10 * RB +2.12 * η + 3.03 * μ -0.48 * ω)
H2_9= TanH (-20.29 -0.10 * HBA + 1.64* HBD -1.81 * HOMO + 0.14 * LUMO - 0.039 * RB +1.57 * η + -1.45 * μ + 0.056 * ω)
H2_10= TanH (-10.87+ 0.38* HBA - 0.07 * HBD -1.86 * HOMO + 2.51* LUMO + 0.31 * RB -0.92 * η -1.31 * μ -0.02 * ω)
H2_11= TanH (15.87 + 0.08 * HBA + -0.67 * HBD +2.38* HOMO + 0.78 * LUMO + 0.22 * RB -4.44 * η -0.80 * μ + 0.11 * ω)

This network is characterized by an RMSE value = 0.37 and an R^2value = 0.92, indicating that the (pIC_{50}) is significantly correlated with the eight variables adopted in this work. The ANN model's external predictability was then evaluated using a series of 39 testing set. As can be seen from the table, the model had strong predictive capability, with an R^2 external of 0.60 (Table IV.8). The high values of R^2 and low value of the root-mean-square deviation (RMSE$_{validation}$=0.61) indicates that the proposed model is robust and reliable. From the results presented in Table IV.2.8, the [8-11-11-1] ANN architecture was reasonably adequate for all datasets with good predictive ability. Thus, the ANN model can be used quite satisfactorily for the prediction and screening of a new series of nuclear factor-κB (NF-κB) inhibitors.

Table IV.2.8: Statistical parameters obtained by the [8.11.11.1] ANN model

	Training	Validation
R^2	0.92	0.60
RMSE	0.37	0.61

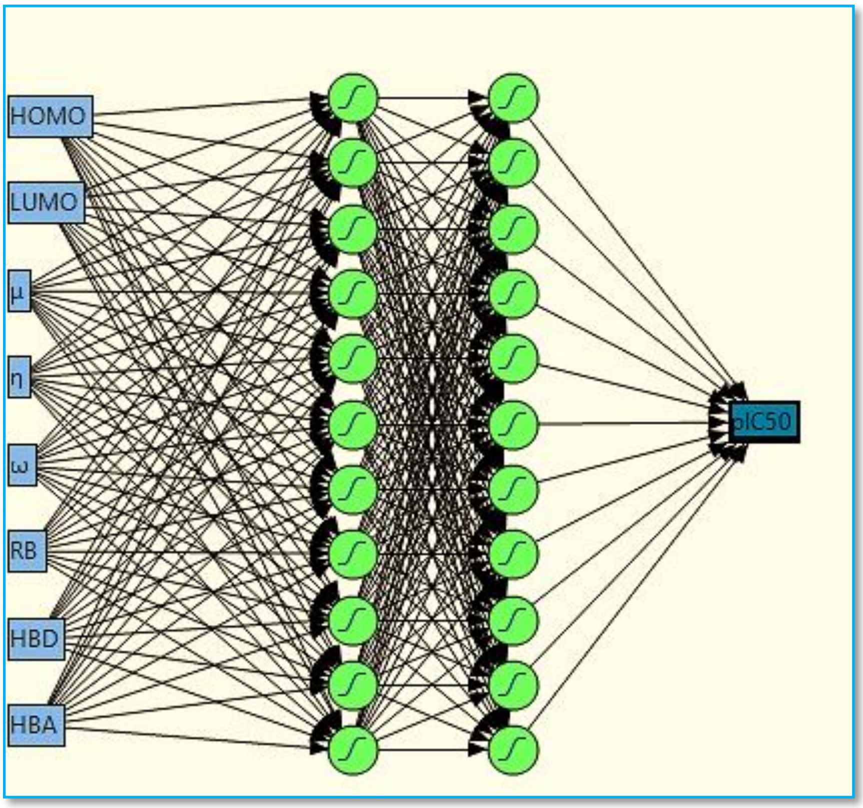


Figure IV.2.3: [8.11.11.1] architecture configuration of the artificial neural network for predicting the pIC_{50}.

Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Rotatable Bonds (RB), the energy of the highest occupied molecular orbitals (HOMO), energy of the lowest unoccupied molecular orbitals (LUMO), Global hardness (η), Chemical potential (μ), Electrophilicity index (ω)

Figures IV.2.4 and IV.2.5 illustrate a linear relationship between predicted and experimental biological activities for the training set and the test set, respectively, demonstrating that the created model is internally and externally predictable. Also, the ANN model has been well matched to the experimental data, with the exception of a few scattered points. Therefore, it may be used to predict the pIC_{50} values with confidence. Furthermore, when the estimated residuals for both the training and test sets are plotted versus the experimental activity levels in Figures IV.2.6 and IV.2.7, the residuals are distributed around the zero line, and most of them are less than an absolute value of 0.5, indicating that the model is free of systematic errors.

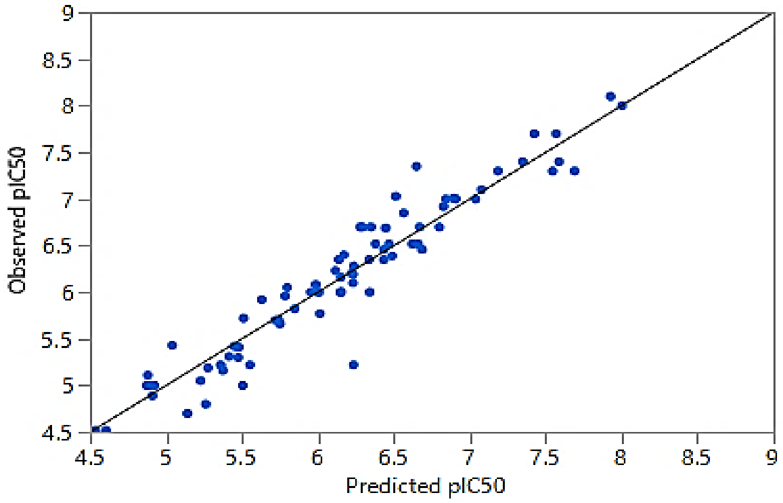


Figure IV.2.4: Experimental versus predicted pIC_{50} of training set compounds in ANN model

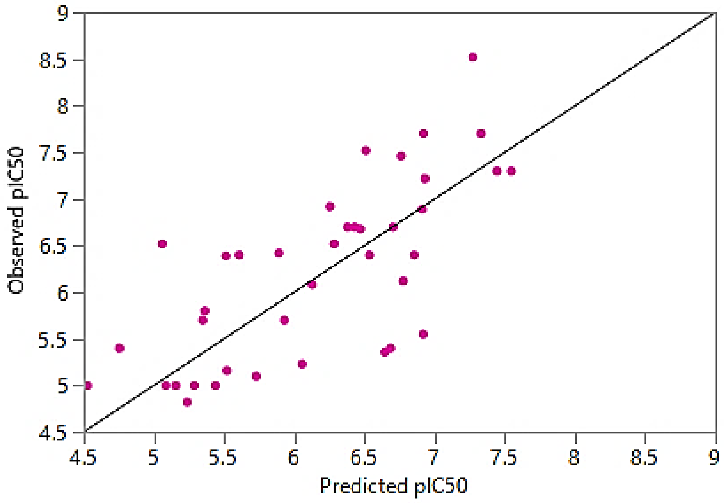


Figure IV.2.5: Experimental versus predicted pIC_{50} of test set compounds in ANN model

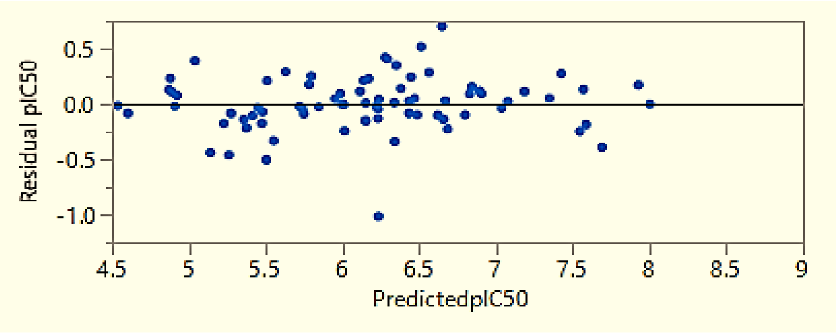


Figure IV.2.6: Residual pIC_{50} versus predicted pIC_{50} for the training set (ANN model)

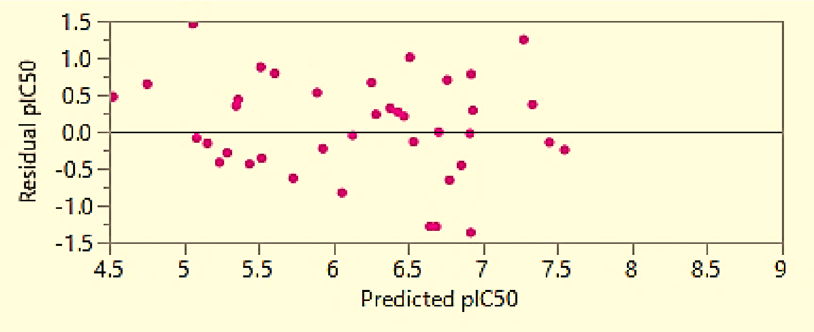


Figure IV.2.7: Residual pIC_{50} versus predicted pIC_{50} for the test set (ANN model).

For both the training and testing sets, the model ANN was used to predict the pIC_{50} values. Table IV.2.9 represents the predicted and observed values of the investigated compounds as well as the residual.

Table IV.2.9: Experimental and predicted values of the studied compounds based on the [8.11.11.1] ANN model

Compounds	**Exp(pIC_{50})**	**Prd(pIC_{50})**	**Residual**	**Compounds**	**Exp(pIC_{50})**	**Prd(pIC_{50})**	**Residual**
Compound 1	6.00	6.01	-0.01	Compound 62	5.22	5.36	-0.14
Compound 2	4.80	5.26	-0.46	Compound 63	6.00	6.34	-0.34
Compound 3	5.00	5.44	-0.44	Compound 64	6.85	6.56	0.29
Compound 4	6.00	6.00	0.00	Compound 65	6.70	6.29	0.41
Compound 5	6.70	6.35	0.35	Compound 66	7.35	6.65	0.70
Compound 6	5.22	5.55	-0.33	Compound 67	6.28	6.24	0.04
Compound 7	5.70	5.72	-0.02	Compound 68	6.10	6.23	-0.13
Compound 8	5.22	6.23	-1.01	Compound 69	5.66	5.75	-0.09
Compound 9	5.10	5.73	-0.63	Compound 70	5.82	5.85	-0.03
Compound 10	5.00	5.29	-0.29	Compound 71	6.68	6.47	0.21
Compound 11	6.05	5.80	0.25	Compound 72	4.52	4.24	0.28
Compound 12	6.52	5.06	1.46	Compound 73	4.52	4.54	-0.02
Compound 13	6.40	5.61	0.79	Compound 74	4.52	4.60	-0.08
Compound 14	6.16	6.15	0.01	Compound 75	5.36	6.64	-1.28
Compound 15	5.16	5.52	-0.36	Compound76	7.01	6.89	0.12
Compound 16	7.00	6.91	0.09	Compound 77	5.19	5.27	-0.08
Compound 17	6.69	6.45	0.24	Compound 78	6.19	6.23	-0.04
Compound 18	6.52	6.66	-0.14	Compound 79	6.35	6.14	0.21
Compound 19	5.40	6.69	-1.29	Compound 80	6.92	6.83	0.09
Compound 20	6.70	6.43	0.27	Compound 81	5.31	5.41	-0.10
Compound 21	6.00	6.01	-0.01	Compound 82	5.96	5.78	0.18
Compound 22	5.7	5.93	-0.23	Compound 83	5.11	4.88	0.23
Compound 23	6.52	6.28	0.24	Compound 84	5.00	5.50	-0.50
Compound 24	5.16	5.37	-0.21	Compound 85	5.00	4.92	0.08

Compound 25	4.82	5.24	-0.42	Compound 86	6.08	5.98	0.10
Compound 26	6.00	5.95	0.05	Compound 87	5.00	4.53	0.47
Compound 27	5.00	4.87	0.13	Compound 88	5.55	6.92	-1.37
Compound 28	5.70	5.74	-0.04	Compound 89	6.12	6.78	-0.66
Compound 29	5.40	4.75	0.65	Compound 90	7.30	7.55	-0.25
Compound 30	6.00	6.15	-0.15	Compound 91	6.35	6.43	-0.08
Compound 31	5.23	6.06	-0.83	Compound 92	7.46	6.76	0.70
Compound 32	5.00	5.16	-0.16	Compound 93	6.89	6.91	-0.02
Compound 33	5.00	5.08	-0.08	Compound 94	5.77	6.01	-0.24
Compound 34	4.70	5.14	-0.44	Compound 95	6.39	5.51	0.88
Compound 35	5.00	4.89	0.11	Compound 96	6.46	6.43	0.03
Compound 36	6.52	6.66	-0.14	Compound 97	7.03	6.51	0.52
Compound 37	5.30	5.47	-0.17	Compound 98	6.92	6.25	0.67
Compound 38	6.70	6.28	0.42	Compound 99	6.46	6.68	-0.22
Compound 39	6.70	6.80	-0.10	Compound100	6.20	6.22	-0.02
Compound 40	6.52	6.38	0.14	Compound 101	7.22	6.93	0.29
Compound 41	6.70	6.38	0.32	Compound 102	7.00	6.84	0.16
Compound 42	7.30	7.18	0.12	Compound 103	8.10	7.93	0.17
Compound 43	6.52	6.47	0.05	Compound 104	8.52	7.27	1.25
Compound 44	6.40	6.53	-0.13	Compound 105	7.70	7.33	0.37
Compound 45	5.70	5.35	0.35	Compound 106	7.30	7.69	-0.39
Compound 46	5.80	5.36	0.44	Compound 107	7.30	7.44	-0.14
Compound 47	4.52	4.49	0.03	Compound 108	7.40	7.59	-0.19
Compound 48	5.43	5.04	0.39	Compound 109	7.30	7.55	-0.25
Compound 49	5.41	5.48	-0.07	Compound 110	7.70	7.57	0.13
Compound 50	6.52	6.62	-0.10	Compound 111	7.70	7.42	0.28
Compound 51	6.35	6.34	0.01	Compound 112	7.40	7.35	0.05
Compound 52	6.39	6.49	-0.10	Compound 113	8.00	8.00	0.00
Compound 53	6.08	6.13	-0.05	Compound 114	7.52	6.51	1.01
Compound 54	6.23	6.11	0.12	Compound 115	6.40	6.85	-0.45
Compound 55	6.42	5.89	0.53	Compound 116	7.10	7.08	0.02
Compound 56	4.89	4.91	-0.02	Compound 117	6.00	6.15	-0.15
Compound 57	5.72	5.51	0.21	Compound 118	5.05	5.23	-0.18
Compound 58	6.40	6.17	0.23	Compound 119	7.00	7.04	-0.04
Compound 59	6.70	6.70	0.00	Compound 120	6.70	6.67	0.03
Compound 60	5.92	5.63	0.29	Compound 121	7.70	6.92	0.78
Compound 61	5.42	5.45	-0.03				

IV.1.7. Examination of outliers in a regression-based quantitative structure-activity relationship

In this work, the leverage method has been applied for the applicability domain of the obtained MLR and ANN model determination. The leverage method is a quantitative indicator used to determine structural outliers in the data. By calculating $the\ h_i$ value for each compound and making a comparison between it and a reference point which is the critical value (h*). h*represents the edge of the applicability domain. Where the occurrence of any value greater than or equal to this critical value (i.e., structural outliers) shows that the prediction is unreliable. Since compounds characterizing by h_i value higher than h* is structurally different to those used for the model construction. And it is possible that the prediction of these compounds may be judged relatively untrustworthy due to a large amount of extrapolation involved. In contrast, when h_i Value is lower than h*, which indicates the chemical similarity between the predicted compounds and the training set [23.24.25.25.26]. Leverage value can be obtained by the following equation:

$$h_i = v_i(V^tV)^{-1} \times v_i^{-1} \qquad \text{Eq5}$$

v_i: $represents\ a\ matrix\ consisting\ of\ the\ significant\ descriptors\ with\ dimensions\ of\ 1 \times d^*$

d^* : $the\ number\ of\ significant\ molecular\ descriptors$.

V: $is\ a\ p \times d^*\ matrix\ where\ p\ represents\ the\ number\ of\ experimental\ data\ points.$

$h_i\ the\ leverage\ value\ of\ each\ experimental\ data.$

$$h^* = 3\,(k+1)/n \qquad (20) \qquad \text{Eq6}$$

k is the number of model's descriptors, and n is the number of compounds included in the training set.

The leverage method also depends on standardized residual (SDR) values. This is given by the following equation:

$$SDR = \frac{\hat{Y}_i - Y_i}{\sqrt{\frac{\sum_{i-1}^{n}(\hat{Y}_i - Y)^2}{n}}} \qquad \text{Eq7}$$

$-3 < SDR < 3$

Where : $y_i\ and\ \hat{Y}_i$Are the measured and the predicted pIC_{50} values, respectively.

William plot is used to visualize a model's scope of application by plotting the standardized residuals (SDR) vs the leverage values. The extent to which the AD structural range is covered can be determined by comparing the number of compounds inside the AD to the outliers in a William plot. Via the following parameter:

$$AD_{coverage} = \frac{P_{inside}}{P_{total}} * 100 \qquad \text{Eq8}$$

P_{inside} : *represents the points contained inside the domain,*
P_{total}: *represents the total number of data points.*

After calculating the leverage of all studied compounds, the critical leverage h^* Which is found to be 0.44 and 0.32 for the MLR and ANN models, respectively and the standardized residual. William plot has been traced as it is shown in Figures IV.2.8 and IV.2.9. By analysing the AD of the obtained MLR and the [8.11.11.1] ANN models. It can be seen that both the training and test sets are comprised between the AD boundaries: the vertical line, blue dashed line $(h^* = 0.44\ \ MLR\ model)$ $(h^* = 0.32\ ANN\ model)$ and the two horizontal green lines (-3 <SDR<3). They presented leverage values less than the critical leverage h* value and standardized residuals values less than the thresholds. except for two compounds of the training set which are considered outliers because of their low prediction reliability. Since the majority of training set compounds and all compounds of test sets are located inside the AD. Both of the $MLR\ AD_{average}$ value and the ANN$AD_{average}$ value = 98 %. Therefore, in these ADs, the models are considered valid, and they can be used for pIC_{50} values prediction for a new compound with a high degree of confidence. Except for two compounds which were located outside the AD.

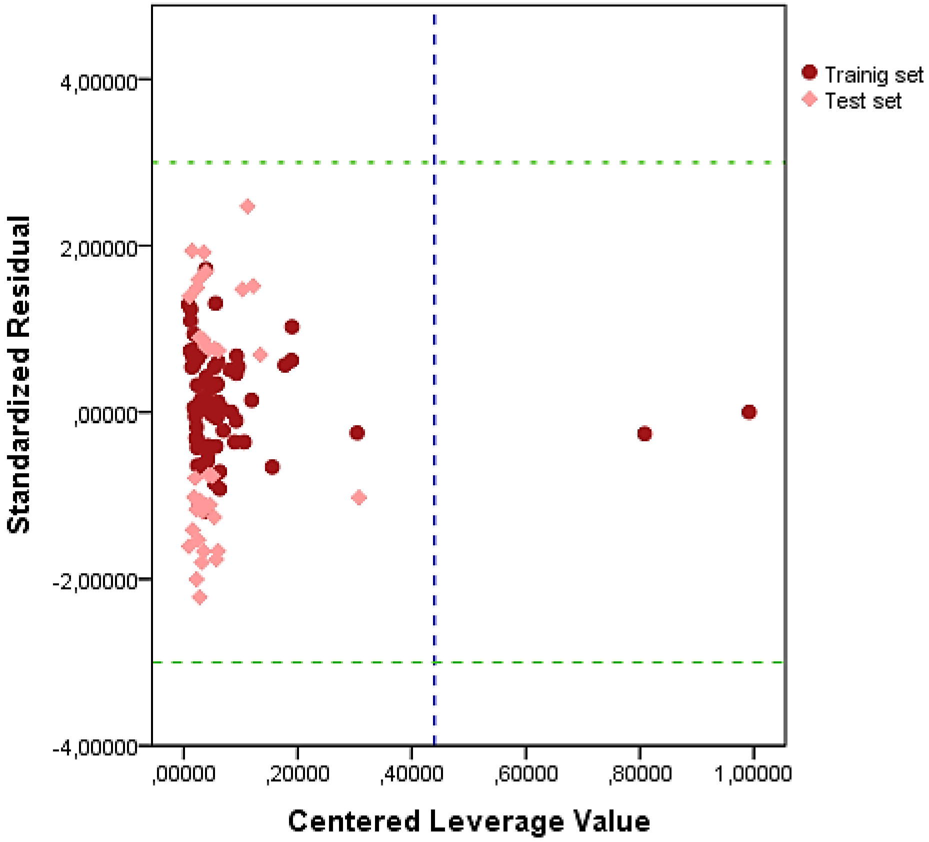


Figure IV.2.8: William plots for the MLR model

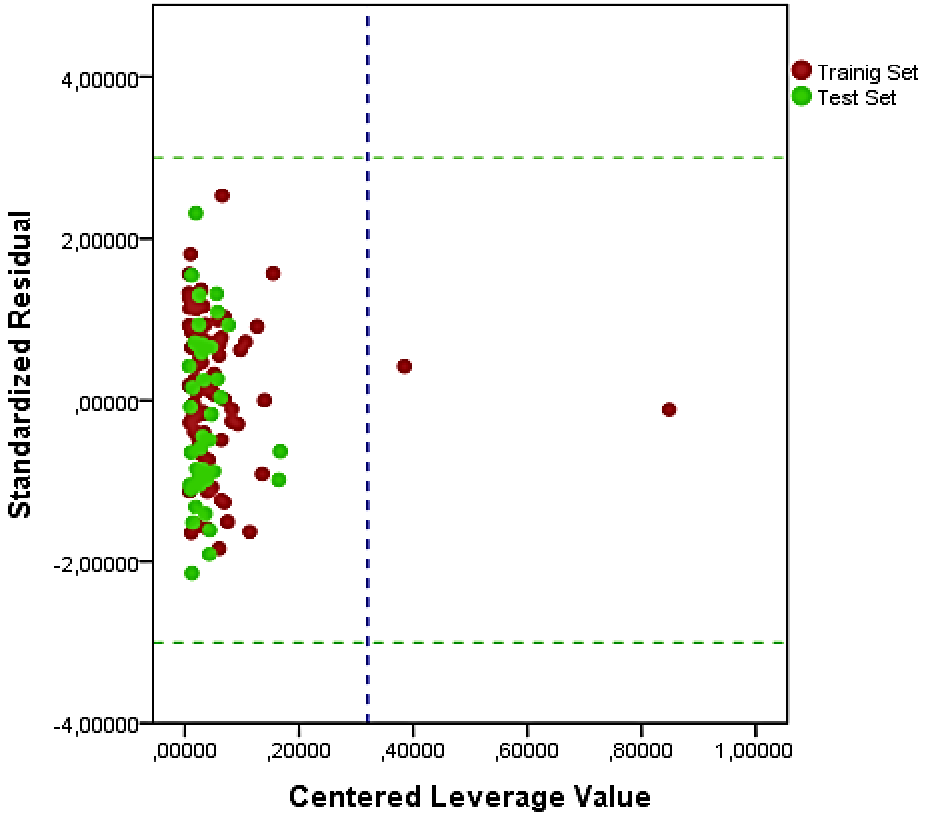


Figure IV.2.9: William plots for the ANN model

IV.1.8. Importance of significant molecular descriptors in MLR and [8.11.11.1] ANN QSAR models

One of the QSAR analysis objectives is to understand the molecular descriptor that governs the activity of a particular class of compounds and to aid in the design of new compounds that may be drugs. Therefore, the evaluation of the descriptors proved to be very interesting and useful to better understand the NF-κB inhibition. To achieve this analysis, a randomization strategy was employed to determine the relative value of each descriptor used to create the MLR and ANN models.

This also enables the identification of the optimal molecular descriptors. Immediately after the models were constructed, the first column corresponding to the first descriptor utilized in the model was eliminated, leaving the remaining descriptor matrix and Y-column in their original positions. the mean absolute deviations (Δ_{mi}) between the observed and predicted activities of compounds has been determined for the eleven descriptors for the MLR model and the eight descriptors for the ANN model. The descriptor's contribution is given as follows:

$$C_i\% = \frac{\Delta_{mi}}{\sum_{j=1}^{11} \Delta_{mj}} \qquad \text{Eq9}$$

By employing scrambled descriptor values, the increment in $C_i\%$, can calculate the relevance of the significant molecular descriptors used for model construction measurement. where the larger the increments, the greater the significance [27]. This procedure was applied to all eleven significant descriptors used for the MLR model and all eight significant descriptors used for the ANN model.

Results are presented in Figures IV.2.10 and IV.2.11. As it is depicted in the figures the lipophilicity Alogp98 and the electrophilicity index ω are of particular relevance in the MLR model. In contrast, the Global hardness (η), and electrophilicity index ω are of particular importance in the ANN model.

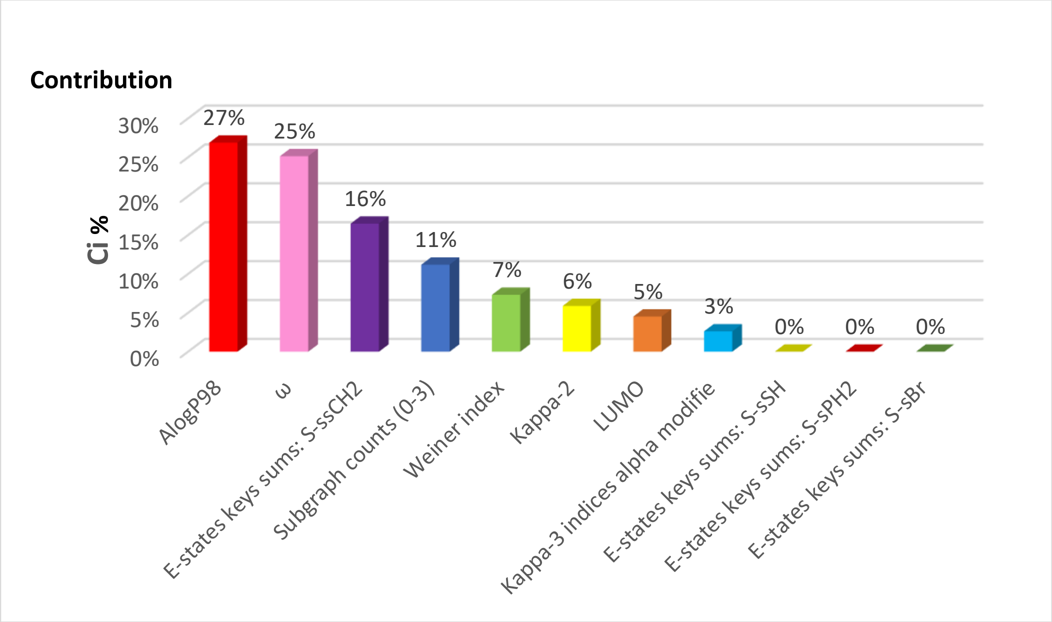


Figure IV.2.10 : descriptors contribution in the MLR model

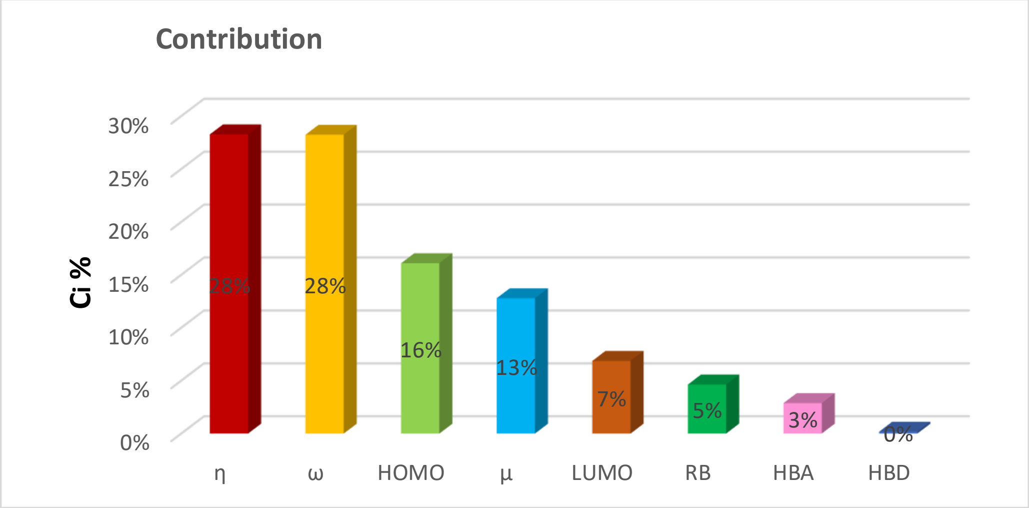


Figure IV.2.11 : descriptors' contribution in the ANN model

IV.1.9. Models Comparison

Based on the statistical results mentioned in Tables 7,9,10, a simple comparison between the obtained models' performances has been performed. when comparing the MLR model with the [11.5.5.1], [11.5.7.1], [11.5.9.1], [11.5.11.1] ANN models constructed with the same molecular. It can be concluded that the MLR can predict the pIC_{50} of compounds excellently. In contrast, the ANN models cannot be useful for prediction because of the non-existence of a non-linear correlation between the pIC_{50} and the molecular descriptors. Comparing the MLR model with the [8.11.11.1] ANN, it can be concluded that the ANN model is more performant and more reliable than the MLR model. Thus, the ability of the network to predict the NF-κB inhibition activity was estimated by internal and external validation, which includes the $R^2_{training\,,}$ $R^2_{validation}RMSE_{training}$and $RMSE_{validation}$, which meet the model validation criteria. While the MLR model, although it has presented a high value of R^2 and Q^2 , however its $R^2{}_{predicted}$ Value which indicates its external robustness was very low. This study allowed us to conclude that the [8-11-11-1] ANN architecture can establish a satisfactory non-linear relationship between the structural characteristics of the studied compounds and their nuclear factor-κB (NF-κB inhibition activity, and it can make a reliable prediction for new compounds.

IV.2 Conclusion

The quantitative structure-activity relationship models are very interesting tools for drug development and compound characteristics optimization. In this work, the most important steps for successful QSAR model elaboration have been well explained. Also, a series of nuclear factor-κB (NF-κB) inhibitors has been performed for MLR and several ANN model constructions. The models have been assessed and subjected to internal and external validation. The leverage method has been applied for the applicability domain of the obtained MLR and ANN models to the aim of the outliers examination. Based on the statistical analyses, it was found that the obtained [8.11.11.1] ANN model using the following descriptors: Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Rotatable Bonds (RB), energy of the highest occupied molecular orbitals (HOMO), energy of the lowest unoccupied molecular orbitals (LUMO), Global hardness (η), Chemical potential (μ), Electrophilicity index (ω), is reliable, robust and showed the better predictive ability compared to the MLR model. Thus, the [8.11.11.1] ANN model can be used quite satisfactorily for the prediction and screening of a new series of nuclear factor-κB (NF-κB) inhibitors.

REFERENCES

References

[1] Murata, T., Shimada, M., Sakakibara, S., Yoshino, T., Masuda, T., Shintani, T., ... & Yamauchi, M. (2004). Synthesis and structure–activity relationships of novel IKK-β inhibitors. Part 3: orally active anti-inflammatory agents. Bioorganic & medicinal chemistry letters, 14(15), 4019-4022.

[2] Ma, W., Wang, Y., Chu, D., & Yan, H. (2019). 4D-QSAR and MIA-QSAR study on the bruton's tyrosinekinase (Btk) inhibitors. Journal of molecular graphics and modelling, 92, 357-362.

[3] Nam, N. H. (2006). Naturally occurring NF-κB inhibitors. Mini reviews in medicinal chemistry, 6(8), 945-951. 7

[4] Musa, A. Y., Mohamad, A. B., Kadhum, A. A. H., Takriff, M. S., & Ahmoda, W. (2012). Quantum chemical studies on corrosion inhibition for series of thio compounds on mild steel in hydrochloric acid. Journal of industrial and engineering chemistry, 18(1), 551-555. doi.org/10.1016/j.jiec.2011.11.131

[5] Musa, A. Y., Jalgham, R. T., & Mohamad, A. B. (2012). Molecular dynamic and quantum chemical calculations for phthalazine derivatives as corrosion inhibitors of mild steel in 1 M HCl. Corrosion science, 56, 176-183. doi.org/10.1016/j.corsci.2011.12.005

[6] Vora, J., Patel, S., Sinha, S., Sharma, S., Srivastava, A., Chhabria, M., & Shrivastava, N. (2019). Molecular docking, QSAR and ADMET based mining of natural compounds against prime targets of HIV. Journal of biomolecular structure and dynamics, 37(1), 131-146.

[7] Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery, 3(11), 935-949.

[8] Bell, E. W., & Zhang, Y. (2019). DockRMSD: An open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. Journal of cheminformatics, 11(1),

[9] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. 1993, 98, 5648−5652

[10] Homeyer N, Horn AH, Lanig H, Sticht H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J Mol Model*. 2006;12(3):281-289. doi:10.1007/s00894-005-0028-4

[11] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79,

[12] Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. J. Phys. Chem. A 2001, 105, 9954−9960.

[13] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale´, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. J. Comput. Chem. 2005, 26, 1781− 1802

[14] Pang, X., Fu, H., Yang, S., Wang, L., Liu, A. L., Wu, S., & Du, G. H. (2017). Evaluation of novel dual acetyl-and butyrylcholinesterase inhibitors as potential Anti-Alzheimer's disease agents using pharmacophore, 3D-QSAR, and molecular docking approaches. Molecules, 22(8), 1254. doi.org/10.3390/molecules22081254

[15] Huang, S., Song, C., Wang, X., Zhang, G., Wang, Y., Jiang, X., ... & Li, L. (2017). Discovery of new SIRT2 inhibitors by utilizing a consensus docking/scoring strategy and structure–activity relationship analysis. Journal of chemical information and modeling, 57(4), 669-679. doi.org/10.1021/acs.jcim.6b00714

[16] 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. doi.org/10.1166/jbns.2018.1503

[17] Hou, T., Wang, J., Zhang, W., & Xu, X. (2007). ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? Journal of chemical information and modeling, 47(2), 460-463. doi.org/10.1021/ci6003515

[19] Medjahed, S., Belaidi, S., Djekhaba, S., Tchouar, N., & Kerassa, A. (2016). Computational study of molecular electrostatic potential, drug likeness screening and structure-activity/property relationships of thiazolidine-2, 4-dione derivatives. Journal of bionanoscience, 10(2), 118-126. doi.org/10.1166/jbns.2016.1358

REFERENCES

[20] Belaidi, S., Belaidi, H., & Bouzidi, D. (2015). Computational methods applied in physical-chemistry property relationships of thiophene derivatives. Journal of Computational and theoretical nanoscience, 12(8), 1737-1745. doi.org/10.1166/jctn.2015.3952

[21] Hammoudi, N. E. H., Benguerba, Y., Attoui, A., Hognon, C., Lemaoui, T., Sobhi, W., ... & Monari, A. (2022). In silico drug discovery of IKK-β inhibitors from 2-amino-3-cyano-4-alkyl-6-(2-hydroxyphenyl) pyridine derivatives based on QSAR, docking, molecular dynamics and drug-likeness evaluation studies. *Journal of Biomolecular Structure and Dynamics*, *40*(2), 886-902.

[22] Lemaoui, T., Abu Hatab, F., Darwish, A. S., Attoui, A., Hammoudi, N. E. H., Almustafa, G., ... & Alnashef, I. M. (2021). Molecular-based guide to predict the pH of eutectic solvents: promoting an efficient design approach for new green solvents. *ACS Sustainable Chemistry & Engineering*, *9*(17), 5783-5808. AP

[23] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225-1231.

[24] Shahbaz, K., Baroutian, S., Mjalli, F. S., Hashim, M. A., & AlNashef, I. M. (2012). Densities of ammonium and phosphonium based deep eutectic solvents: Prediction using artificial intelligence and group contribution techniques. *Thermochimica Acta*, *527*, 59-66.

[25] Melagraki, G., & Afantitis, A. (2013). Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. *Chemometrics and Intelligent Laboratory Systems*, *123*, 9-14.

[26] Oluwaseye, A., Uzairu, A., Shallangwa, G. A., & Abechi, S. E. (2020). Quantum chemical descriptors in the QSAR studies of compounds active in maxima electroshock seizure test. *Journal of King Saud University-Science*, *32*(1), 75-83.

[27] 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.

General Conclusion

General Conclusion

A wide variety of malignancies common in human exhibit aberrant NF-κB constitutive expression which results in tumorigenic processes and cancer survival in a variety of solid tumour, including pancreatic cancer, lung, cervical, prostate, breast and gastric carcinoma. Numerous evidences indicate that NF-κB signalling mechanism is mainly involved in the progression of several cancers which may intensify an enhanced knowledge on its role in disease particularly lung tumorigenesis. This has led to tremendous research in designing a variety of NF-κB antagonists with enhanced clinical applications through different approaches the most common being suppression of IκB kinase (IKK) beta activity. The work of this thesis id devoted into tow parts; The first consists of targeting the IKK-β enzyme for the NF-κB inhibition. The second part consist of the direct inhibition of NF-κB

The main objective of the first part is to design new active compounds for the IKK-β. To achieve this goal a QSAR model of thirty2-amino-3-cyano-4-piperidin-6-(2-hydroxyphenyl) pyridine derivatives against IKK-β enzyme was performed based on theoretical molecular descriptors which were selected by stepwise regression. This model was used to predict the biological activity of new designed compounds. The applicability domain of the elaborated QSAR model succeeded to screen 21 new designed compounds with higher biological activity, and most of compounds which were characterized by a substitution at para position showed a higher pIC50values than those were characterized by a substitution at meta position of the phenyl ring. Docking simulation was carried out to position the investigated compounds into the IKK-β enzyme active site.

The results revealed that all compounds had poses in the pocket. This result has been validated via the molecular dynamics study. Among the new designed compounds, four compounds exhibited better binding interaction toIKK-β which are A1m, A1p, A9m, A11, with better binding affinity to the active site. Their XP GScore are respectively -12.375, -10.713, -10.633, -9. 381. Docking study also showed that CYS 99 the GLU149 amino acid residues might be crucial for IKK-β inhibition and the hydroxylic groups of compounds may enhance the affinity of compounds. The compounds to be considered as drug like molecule were evaluated on the basis of Veber rule and Lipinski rule. most of the designed compounds showed no violations as per Veber and Lipinski rule.

As for the main objective of second part is the development of new predictive QSAR models for the pIC_{50} of of nuclear factor-κB (NF-κB) inhibitors prediction. To achieve this goal a series of nuclear factor-κB (NF-κB) inhibitors has been performed for MLR and several ANN model constructions. The models have been assessed and subjected to internal and external validation.

The leverage method has been applied for the applicability domain of the obtained MLR and ANN models to the aim of the outliers examination. Based on the statistical analyses, it was found that the obtained [8.11.11.1] ANN model using the following descriptors: Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Rotatable Bonds (RB), energy of the highest occupied molecular orbitals (HOMO), energy of the lowest unoccupied molecular orbitals (LUMO), Global hardness (η), Chemical potential (μ), Electrophilicity index (ω), is reliable, robust and showed the better predictive ability compared to the MLR model. Thus, the [8.11.11.1] ANN model can be used quite satisfactorily for the prediction and screening of a new series of nuclear factor-κB (NF-κB) inhibitors.

Abstract

The Inhibition of IKK-β (nuclear factor kappa B kinase subunit beta), a specific modulator of NF-κB (nuclear factor-κB), is considered a valid target to discover new active compounds for various cancers and rheumatoid arthritis treatment. The aim of this work is twofold: Firstly, to develop new designed compounds IKK-β inhibitors using different computer aided drug design tools (QSAR, Molecular Docking, Molecular Dynamics and drug likeness evaluation). The analysis of the results of QSAR model and molecular docking succeeded to screen 21 interesting compounds with better inhibitory concentration having a good affinity to IKK-β. Secondly, is to develop new predictive QSAR models for the pIC_{50} of of nuclear factor-κB (NF-κB) inhibitors prediction, based on machine learning methods. Based on the statistical analyses, it was found that the obtained [8.11.11.1] ANN model is reliable, robust and showed the better predictive ability compared to the MLR model. Thus, the [8.11.11.1] ANN model can be used quite satisfactorily for the prediction and screening of a new series of nuclear factor-κB (NF-κB) inhibitors.

Keywords: NF-κB, IKK-β, QSAR, ANN, MLR

Résumé

L'inhibition de IKK-β (nuclear factor kappa B kinase subunit beta), un modulateur spécifique de NF-κB (nuclear factor-κB), est considéré comme une cible valable pour découvrir de nouveaux composés actifs pour divers cancers et le traitement de la polyarthrite rhumatoïde. L'objectif de ce travail est double : Premièrement, développer de nouveaux composés inhibiteurs de l'IKK-β à l'aide de différents outils de conception de médicaments assistés par ordinateur (QSAR, l'amarrage moléculaire, la dynamique moléculaire et L'évaluation de la similarité des médicaments). L'analyse des résultats du modèle QSAR et de l'amarrage moléculaire a permis de cribler 21 composés intéressants avec une meilleure concentration inhibitrice ayant une bonne affinité pour IKK-β. Deuxièmement, il s'agit de développer de nouveaux modèles prédictifs QSAR pour la prédiction du pIC_{50}des inhibiteurs (NF-κB), basés sur des méthodes du machine Learning. Sur la base des analyses statistiques, il a été constaté que le modèle ANN [8.11.11.1] obtenu est fiable, robuste et a montré une meilleure capacité prédictive par rapport au modèle MLR. Ainsi, le modèle ANN [8.11.11.1] peut être utilisé de manière tout à fait satisfaisante pour la prédiction et le criblage d'une nouvelle série d'inhibiteurs du facteur nucléaire-κB (NF-κB).

Mots clés : NF-κB, IKK-β, QSAR, ANN, MLR

الملخص

يعتبر تثبيط إنزيم IKK- β ، وهو مُعدِّل محدد لـ NF- κB(العامل النووي) هدفًا صالحًا لاكتشاف مركبات نشطة جديدة لمختلف أنواع السرطان وعلاج التهاب المفاصل الروماتويدي. الهدف من هذا العمل ذو شقين: أولاً، تصميم مثبطات جديدة وفعالة باستخدام أدوات تصميم دوائية مختلفة بمساعدة الكمبيوتر (QSAR، الإرساء الجزيئي، الديناميكيات الجزيئية وتقييم تشابه الأدوية). نجح تحليل نتائج نموذج والالتحام الجزيئي في فحص 21 مركبًا مثيرًا للاهتمام بتركيز مثبط أفضل له صلة جيدة بـIKK-b ثانيًا، تطوير نماذج QSARتنبؤية جديدة تسمح بتنبؤ التركيز المثبط ل NF- κB بناءً على طرق التعلم الآلي. بناءً على التحليلات الإحصائية، وجد أن نموذج الذي تم الحصول عليه ANN[8.11.11.1] موثوق وقوي وأظهر قدرة تنبؤية أفضل مقارنة بنموذج MLRوبالتالي، يمكن استخدام نموذج بشكل مرضٍ تمامًا للتنبؤ وفحص سلسلة جديدة من مثبطاتNF- κB

الكلمات المفتاحية: العامل النوويNF- κB, إنزيم IKK- β, الانحدار الخطي المتعدد, شبكة اعصاب صناعية