

***IN SILICO* STUDIES OF TNF- ALPHA INHIBITION THROUGH QSAR MODELLING AND MOLECULAR DOCKING**

Submitted in partial fulfilment of the requirements for the award of
Master of Science Degree in Bioinformatics and Data Science

By

MEENAKSHI K (42739016)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A++” by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119

APRIL – 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Meenakshi K (42739016)**, who carried out the project entitled "***In Silico* studies of TNF- Alpha inhibition through QSAR Modelling and Molecular Docking**" under our supervision from **January 2024** to **April 2024**.

Internal Guide

Dr. Swetha Sunkar

External Guide

Dr. Swetha Sunkar

Head of the Department

Dr. Jemmy Christy

Submitted for Viva voce Examination held on 20/04/2024

Internal Examiner

External Examiner

DECLARATION

I, **MEENAKSHI.K (42739016)**, hereby declare that the Project Report entitled "***In Silico* studies of TNF- Alpha inhibition through QSAR Modelling and Molecular Docking**" done by me under the guidance of **Dr.Swetha Sunkar** at **Sathyabama Institute of Science and Technology, Chennai** and is submitted in partial fulfilment of the requirements for the award of Master of Science degree in Bioinformatics and Data Science.

DATE: 20/4/2024
PLACE: CHENNAI

K. Meenakshi
SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **the Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. REKHA CHAKRAVATHY**, Dean of Science and Humanities, I am deeply grateful for her mentorship and leadership, which have been a constant source of inspiration. Her dedication to fostering a conducive academic environment has been instrumental in the successful completion of this thesis.

I convey my thanks to **R. JEMMY CHRISTY**, Head of the Bioinformatics Department, for providing me the necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. SWETHA SUNKAR** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Bioinformatics who were helpful in many ways for the completion of the project.

ABSTRACT

Tumor necrosis factor alpha (TNF- α) is a cytokine that exhibits pleiotropic effects on several cell types. TNF- α is a 157 amino acid homotrimer protein that is primarily produced by T lymphocytes, activated macrophages, and natural killer cells. TNF inhibition has shown effective for patients with rheumatoid arthritis and other inflammatory conditions such as psoriasis, psoriatic arthritis, ankylosing spondylitis, and inflammatory bowel disease. Five FDA-approved medications, including Infliximab, Etanercept, Adalimumab, and Golimumab, inhibit TNF- α . In order to learn more about the origins of the bioactivity of 954 compounds, a large non-redundant data collection with published IC₅₀ values against TNF- α was acquired from ChEMBL and used in a quantitative structure-activity relationship (QSAR) analysis. Twelve fingerprint descriptors were used to characterize TNF- α inhibitors, and random forest was used to build prediction models from 100 distinct data splits. For the training set and external set, the generated models yielded R², RMSE, and MAE values in the ranges of 0.004–0.2, 0.004–8.4, and 0.009–1.6, respectively. The OECD standards were followed in selecting the best model created using the substructure count, which produced R², RMSE, and MAE values of 0.202354, 0.00487, and 0.11143, respectively. The substructure fingerprint count was then further analyzed to offer insightful information on the inhibitory action of TNF- α inhibitors. Furthermore, in order to provide structural insights into the origin of TNF- α inhibition, 20 different compounds were chosen for additional molecular docking experiments using Kennard-Stone sampling of the active ingredients. Ten of these compounds have low binding energies against human TNF- α , which are influenced by hydrophobic interaction, π – π stacking, and hydrogen bonding inside the binding pocket, according to molecular docking analysis of these compounds. Phytochemical molecular docking binding interactions are consistent with both molecular docking of chemical molecules and QSAR. The design of innovative and reliable TNF- α inhibitors may be guided by the information provided here.

Keywords: TNF- α , Inhibitors, QSAR, Molecular Docking, *Euphorbia hirta*, Machine

TABLE OF CONTENTS

S.NO	TOPICS	PAGE NO
	ABSTRACT	i
	LIST OF CONTENTS	ii
	TABLES	iv
	FIGURES	v
	ABBREVIATIONS	vii
1	INTRODUCTION	1
1.1	TNF-alpha	1
1.2	TNF-alpha inhibitors	1
1.3	TNF- α Signaling in Autoimmune Diseases	2
1.3.1	Rheumatoid Arthritis	2
1.3.2	Psoriatic Arthritis	3
1.3.3	Inflammatory Bowel Disease	3
1.3.4	Psoriasis	4
1.4	TNF alpha Inhibitors as Therapeutic Drugs	4
1.4.1	Infliximab	4
1.4.2	Etanercept	5
1.4.3	Adalimumab	5
1.4.4	Certolizumab Pegol	5
1.4.5	Golimumab	5
1.5	ChEMBL Database	6
1.6	QSAR	6
1.7	Molecular Docking	6
1.8	Discovery Studio and DockThor	6
1.9	Lipinski Descriptors	7
1.10	Fingerprint Descriptors	7
2	LITERATURE SURVEY	9

3	AIM AND SCOPE	12
4	MATERIALS AND METHODOLOGY	13
	4.1 Data Collection	13
	4.2 Description of inhibitors	13
	4.3 Data Splitting	13
	4.4 Multivariate analysis	13
	4.5 Validation of QSAR models	14
	4.6 Applicability domain analysis	14
	4.7 Molecular Docking	15
	WORKFLOW	16
5	RESULTS & DISCUSSION	17
	5.1 Data Collection	17
	5.2 Chemical space of AChE inhibitors	17
	5.3 QSAR model for predicting AChE inhibitory activity	20
	5.4 Applicability domain	22
	5.5 Mechanistic interpretation of feature importance	23
	5.6 Molecular Docking of TNF- alpha Inhibitors	26
	5.7 Molecular Docking of Phytochemical compounds	31
6	SUMMARY & CONCLUSION	37
7	REFERENCES	38
8	APPENDIX	45

LIST OF TABLES

S.NO.	TABLES	PAGE NO.
1.1	Summary of 12 sets of fingerprint descriptors employed in this study	7
5.1	10- fold Cross Validation (Performance summary of QSAR models for predicting pIC50.	22
5.2	List of top substructure fingerprints and their corresponding description.	25
5.3	Binding affinity and total energy of top 10 compounds using Dockthor	27
5.4	8 Phytochemicals that were docked with TNF-alpha	32

LIST OF FIGURES

S.NO.	FIGURES	PAGE NO.
1.1	Pathway of Psoriatic and Rheumatoid arthritis	2
1.2	Pathway of Inflammatory Bowel Disease	3
1.3	Pathway of Psoriasis	4
5.1	Chemical Space of TNF-alpha inhibitors	18
5.2	Box plot of TNF-alpha inhibitors using Lipinski's rule-of-five descriptors	19
5.3	Plot of experimental versus predicted pIC50 values for models constructed with 12 different fingerprint descriptors	21
5.4	Williams Plot	23
5.5	Feature importance Plot	24
5.6	Distribution of active TNF - alpha	26
5.7	Ligand Interactions of Compound 1	27
5.8	Ligand Interactions of Compound 2	28
5.9	Ligand Interactions of Compound 3	28
5.10	Ligand Interactions of Compound 4	29
5.11	Ligand Interactions of Compound 5	29
5.12	Ligand Interactions of Compound 6	30
5.13	Ligand Interactions of Compound 7	30
5.14	Ligand Interactions of Compound 8	31
5.15	Ligand Interactions of Compound 9	31
5.16	Ligand Interactions of Palmitic acid	32
5.17	Ligand Interactions of Oleic acid	33
5.18	Ligand Interactions of Linoleic acid	33
5.19	Ligand Interactions of Myristic acid	34
5.20	Ligand Interactions of Caffeic acid	34
5.21	Ligand Interactions of Chlorogenic acid	35
5.22	Ligand Interactions of Myricetin	35
5.23	Ligand Interactions of Quercetin	36

LIST OF ABBREVIATIONS

TNF-α	Tumor Necrosis Factor - alpha
FDA	Food and Drug Administration
RA	Rheumatoid Arthritis
PsA	Psoriatic Arthritis
DC	Dendritic cells
IBD	Inflammatory Bowel Disease
PS	Psoriasis
ChEMBL	Chemogenomic European Molecular Biology Laboratory
QSAR	Quantitative structure-activity relationships
IMPPAT	Indian Medicinal Plants, Phytochemistry And Therapeutics
PDB	Protein Data Bank
IC50	Half-maximal inhibitory concentration
PaDEL	Pharmaceutical Data Exploration Laboratory
MW	Molecular Weight
LogP	Partition Coefficient
nHBD_{on}	No. of Hydrogen Bond Donors
nHBA_{cc}	No. of Hydrogen Bond Acceptors
MolMR	Molecular Refractivity
TPSA	Topological polar surface area
RFR	Random Forest Regressor
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
SubFPC	Substructure Fingerprint Count
Tyr	Tyrosine
Lys	Lysine
Pro	Proline
Gly	Glycine
Leu	Leucine

Chapter 1

INTRODUCTION

1.1 TNF-alpha:

The cytokine tumor necrosis factor alpha (TNF- α) affects different types of cells in pleiotropic ways. It is known to play a role in the etiology of some inflammatory and autoimmune disorders and has been identified as a key regulator of inflammatory responses (Bradley, 2008). TNF- α is a homotrimer protein of 157 amino acids that is mostly produced by natural killer cells, T lymphocytes, and activated macrophages (Horiuchi et al., 2010). It is functionally known to initiate a cascade of other inflammatory chemicals, such as chemokines and additional cytokines. There are two forms of TNF- α : transmembrane and soluble. To release the soluble TNF- α (s TNF- α), membrane-bound disintegrin metalloproteinase TNF- α -converting enzyme (TACE) must process the originally generated precursor form, transmembrane TNF- α (tmTNF- α) (Wang, et al., 2017). Patients with rheumatoid arthritis and other inflammatory diseases, such as psoriasis, psoriatic arthritis, ankylosing spondylitis, and inflammatory bowel disease, have found success with TNF inhibition (Cuzzocrea, et al., 2009).

1.2 TNF-alpha inhibitors:

Developments in DNA hybrid technology have resulted in the creation of multiple biologicals that target TNF- α selectively. Currently on the market are three anti-TNF- α medications: adalimumab, infliximab, and etanercept. In 2000, the FDA authorized etanercept for rheumatoid arthritis (RA). It was subsequently approved for psoriatic arthritis, ankylosing spondylitis, and psoriasis. The FDA authorized infliximab and adalimumab for RA in 2002. Adalimumab is licensed for psoriatic arthritis and ankylosing spondylitis, while infliximab is also approved for psoriasis, psoriatic arthritis, ulcerative colitis, and Crohn's disease. Potential targets for anti-TNF- α therapy include other conditions such as bronchial asthma, diabetes mellitus, cancers, septic shock, Behcet's disease, bullous dermatitis, neutrophilic dermatitis, toxic epidermal necrolysis, systemic vasculitis, pyoderma gangrenosum, pustular dermatitis, alcoholic hepatitis, cerebral malaria, hemolytic uremic syndrome, pre-eclampsia, allograft rejection, uveitis, otitis

media, snakebite, erythema nodosum, myelodysplastic syndromes, graft versus host disease, dermatomyositis.(Lal Chaman et al., 2007)

1.3 TNF- α Signaling in Autoimmune Diseases:

1.3.1 Rheumatoid Arthritis:

The main symptoms of rheumatoid arthritis (RA), a chronic autoimmune disease that mostly affects the joints, are redness, swelling, and arthralgia. It restricts range of motion in severe situations (Wang, et al., 2018). About 1% of people have RA, which is defined by inflammation of the synovial tissue that causes progressive damage and can result in persistent impairment (Scutellari et al., 1998). TNF- α is well recognized as the primary inflammatory cytokine implicated in the development of RA and is detected at elevated levels in individuals afflicted with the condition. (Choy et al. 2001), Inflammation is linked to the build-up of inflammatory cells, primarily macrophages and type 1 helper T cells (Th1), but also B cells, plasma cells, and dendritic cells (DCs). In RA, synovial fibroblasts are activated, epidermal hyperplasia is promoted, and inflammatory cells are recruited by TNF- α released from Th1 cells and macrophages (Koch et al., 1994). Following activation by many cytokines, such as TNF- α , IL-1, and IL-6, synovial fibroblasts overexpress matrix metalloproteinases (MMPs) and cathepsins, which is followed by the breakdown of collagen and proteoglycans. Consequently, bone and cartilage are lost, leading to joint erosion (Feldmann et al., 1996).

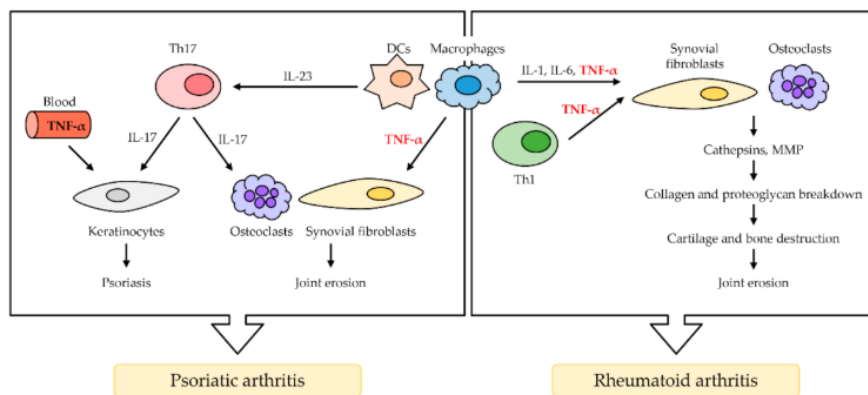


Figure 1.1 Pathway of Psoriatic and Rheumatoid arthritis (Goldring et al, 2002)

1.3.2 Psoriatic Arthritis:

Different from RA, psoriatic arthritis (PsA) is a chronic autoimmune arthropathy (Merola et al., 2018). It usually affects those who have psoriasis on the skin. Like RA, PsA affects about 1% of the population and is considered to be rather frequent (Al-Awadhi et al., 2007). Type 17 helper T cells (Th17), keratinocytes, macrophages, and DCs are the main cells involved in the pathogenesis of PsA. Activated DCs and macrophages stimulated by TNF- α overproduce TNF- α and IL-23. Naïve T cells undergo differentiation into Th17 cells, which overproduce IL-17, in response to IL-23. Subsequently, DCs and IL-17 stimulate keratinocytes, which encourage epidermal hyperplasia and draw in inflammatory cells. Through the NF- κ B signal pathway, TNF- α causes keratinocyte proliferation and anti-apoptosis, which ultimately results in the creation of micro abscess by increasing the recruitment of inflammatory cells in psoriasis (Ritchlin, et al, 2017).

1.3.3 Inflammatory Bowel Disease:

One class of autoimmune illnesses that affect the gastrointestinal tract is called inflammatory bowel disease (c). The two main forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD), which are differentiated by the site of the illness. While CD typically affects the colon and the rectum, it can also affect the mouth, esophagus, stomach, and anus in addition to the small and large intestine (Baumgart et al., 2007). The intestine's accumulated macrophages release TNF- α , IL-1, and IL-6, which ultimately cause vascular disruption, endothelial activation, intestinal matrix breakdown, and epithelial injury (Levin et al., 2016).

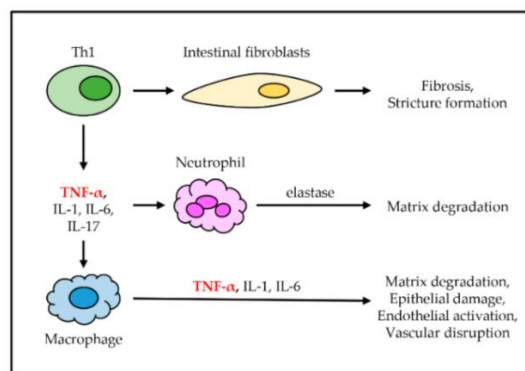


Figure 1.2 Pathway of Inflammatory Bowel Disease (Adegbola et al, 2018)

1.3.4 Psoriasis:

A particular area of aberrant skin characterizes psoriasis (PS), a chronic inflammatory illness with a significant familial diathesis and autoimmune. Plaque makes up over 90% of PS, with the other four forms being pustular, erythrodermic, guttate, and inverse PS. According to Boehncke et al. (2015), plaque-type PS typically manifests as pruritic plaques with white scales on top. The dysregulation of skin immune responses results in epidermal hyperplasia, angiogenesis, and inflammatory leukocyte infiltration into the dermis, which are reflected in the lesions of PS (Griffiths et al, 2007). $\text{TNF-}\alpha$, IL-1, and IL-6 were all abundantly produced by stressed keratinocytes. DCs are activated by certain cytokines. Naïve T cells are differentiated into Th1 cells by IL-12 from activated DCs and into Th17 cells by IL-23. DCs are constantly activated by $\text{TNF-}\alpha$ due to this aberrant immunological response. In the end, keratinocyte hyperproliferation and epidermal changes including acanthosis, parakeratosis, and hypogranulosis are caused by $\text{TNF-}\alpha$, $\text{IFN-}\gamma$, and IL-17 (Ogawa et al., 2018).

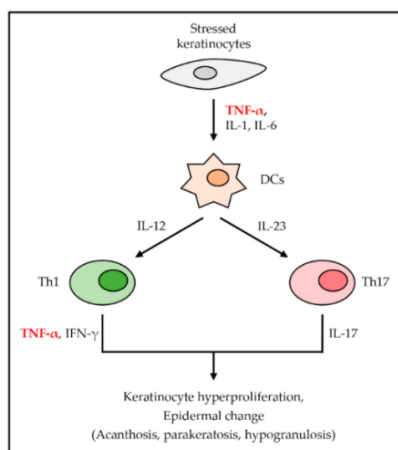


Figure 1.3 Pathway of Psoriasis (Ogawa et al., 2018)

1.4 TNF alpha Inhibitors as Therapeutic Drugs:

1.4.1 Infliximab (Remicade®):

Recombinant chimeric monoclonal antibody infliximab has a human IgG1 constant region and a mouse variable region. It efficiently prevents $\text{TNF-}\alpha$ from binding to its soluble and transmembrane receptors and is specific for all human variants of $\text{TNF-}\alpha$ (Monaco et al., 2015). Following infliximab treatment for inflammatory bowel disease (IBD), complement-

dependent and antibody-dependent cytotoxicity boosted and decreased inflammatory tissue in cell lines expressing TNF- α (Tracey et al., 2008). In 1998, the FDA first approved infliximab for the treatment of CD, and later on, it was also approved for RA. Additionally, it was approved to treat PS, CD, PsA, and ankylosing spondylitis (AS) (Melsheimer et al., 2019).

1.4.2 Etanercept (Enbrel®):

Two identical TNFR2 extracellular regions are fused to the Fc segment of human IgG1 to form the fusion protein etanercept (Melsheimer et al., 2019). After binding to either sTNF- α or tmTNF- α , etanercept prevents the receptors from interacting with them, rendering them inactive (Tracey et al., 2008). In plaque PS, etanercept also seems to regulate proinflammatory genes like NF- κ B, which significantly reduces TNF- α production. Large numbers of DCs in plaque also undergo apoptosis as a result of it, which prevents positive feedback linked to TNF- α by causing early apoptotic cell death before DCs are activated and mature (Mease et al., 2007). According to Moreland et al., the FDA approved it in 1998 for the treatment of juvenile idiopathic arthritis (JIA), PsA, RA, PS, and AS (Moreland et al., 1999).

1.4.3 Adalimumab (Humira®):

Since adalimumab's structure and function are the same as those of naturally occurring human IgG1, it is a totally human IgG1 monoclonal antibody that may selectively prevent human TNF- α from binding to receptors (Rau et al. 2012). Following approval for the treatment of RA, adalimumab was also licensed by the FDA in 2002 to treat PsA, PS, AS, JIA, CD, and uveitis (Rivkin et al., 2009).

1.4.4 Certolizumab Pegol (Cimzia®):

Humanized monoclonal antibody certolizumab pegol (CDP87) lacks the Fc region and has a polyethylene glycosylated (PEG) Fab fragment (Desai et al., 2012). Certolizumab pegol's superior performance over other TNF- α inhibitors may be attributed to its unique structural makeup (Curtis et al., 2019). By 2018, the FDA had approved certolizumab pegol for the treatment of CD, RA, PsA, AS, and plaque PS (Shealy et al., 2007).

1.4.5 Golimumab (Simponi®we):

A fully humanized IgG1 monoclonal antibody with functional specificity for human TNF- α is called golimumab (Cai, A et al., 2010). Golimumab is more potent and has a higher affinity for neutralizing sTNF- α and tmTNF- α than infliximab and adalimumab. As a result, it can effectively block TNF- α 's biological action (Mazumdar et al, 2009). When administered in combination with methotrexate, golimumab was given FDA approval in 2009 to treat RA. Treatment for UC, AS, and PsA was also authorized for it (Gaulton et al., 2012).

1.5 ChEMBL Database:

A vast variety of drug-like bioactive chemicals' binding, functional, and ADMET data are available in the Open Data database ChEMBL. Regularly extracted by hand from the primary published literature, these data are then further standardized and curated to optimize their quality and usefulness for a broad spectrum of chemical biology and drug-discovery research issues. As of right now, the database has 5200 protein targets and 5.4 million bioactivity values for over a million chemicals (Naenna et al, 2009).

1.6 QSAR:

With the application of statistical or machine learning techniques, the quantitative structure–activity relationship, or QSAR, paradigm makes it possible to predict biological activities for molecules of interest based on their characteristics. In addition to their capacity to forecast the activity, QSAR models have been helpful in providing insight into the genesis of these biological activities through the interpretation of the descriptors utilized in their construction (Pagadala et al., 2017).

1.7 Molecular Docking:

The molecular docking methodology investigates the actions of tiny compounds within a target protein's binding region. Molecular docking is becoming a more popular drug discovery strategy as more protein structures are discovered experimentally using techniques like nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography. For unknown protein structures, docking against homology-modeled

targets also becomes feasible. The compounds' druggability and specificity against a certain target can be computed using docking techniques, which are then useful for subsequent lead optimization procedures. Molecular docking algorithms employ a search technique that recursively evaluates the ligand's conformation until it converges to the minimal energy (Steinbeck et al., 2003).

1.8 Discovery Studio and DockThor:

A program called Discovery Studio is used to model and analyze molecular structures. The fields covered by this software include ligand design, structure-based design, simulations, pharmacophore modeling, QSAR, ADME, and prediction toxicity. It also includes macromolecular design and engineering. Discovery Studio's intuitive interface and potent computational algorithms enable the investigation of intricate biological systems and support the creation of novel materials and treatments (Jejurikar et al., 2021).

Since its initial release in 2013, the DockThor-VS web server has been a free virtual screening platform for the scientific community, with ongoing development. For posture and affinity predictions, the DockThor suite outperforms cutting-edge protein-ligand docking techniques in terms of accuracy (Guedes et al., 2017).

1.9 Lipinski Descriptors:

A heuristic method for predicting drug-likeness, Lipinski's "rule of five" states that compounds with molecular weights of 500 or less, log P values of 5, hydrogen bond donors of 5 or less, and hydrogen bond acceptors of 10 or less have poor absorption or penetration. According to the rule, a drug's poor absorption or penetration is more likely when its chemical structure satisfies two or more of the following requirements: (Lipinski, 2000)

1. The molecular weight (MW) exceeds 500.
2. The determined log P value is more than five.
3. The number of hydrogen bond donors (-NH-, -OH) exceeds five.
4. There are more than 10 hydrogen bond acceptors (the -O- and -N6-point double bond)

1.10 Fingerprint Descriptors:

Table 1.1 – Summary of 12 sets of fingerprint descriptors employed in this study.

No.	Fingerprint	Description	Reference
1	CDK	Fingerprint of length 1024 and search depth of 8	(Hall et al., 1995)
2	CDK extended	Extends the fingerprint with additional bits describing ring features.	(Hall et al., 1995)
3	CDK Graph only	A special version that considers only the connectivity and not bond order	(Hall et al., 1995)
4	E-State	Electro topological state atom types	(Durant et al., 2002)
5	MACCS	Binary representation of chemical features defined by MACCS keys	(Wang et al., 2009)
6	Pubchem	Binary representation of substructures defined by PubChem	(Salmina et al., 2015)
7	Substructure	Presence of SMARTS patterns for functional groups	(Justin Klekota et al., 2008)
8	Substructure Count	Count of SMARTS patterns for functional groups	(Justin Klekota et al., 2008)
9	Klekota–Roth	Presence of chemical substructures	(Carhart et al., 1985)
10	Klekota–Roth Count	Count of chemical substructures	(Carhart et al., 1985)
11	2D atom pairs	Presence of atom pairs at various topological distances	(Jang et al., 2021)
12	2D atom pairs Count	Count of atom pairs at various topological distances	(Jang et al., 2021)

Chapter 2

LITERATURE SURVEY

- ❖ Simeon et al., (2016) explains that tumor necrosis factor alpha (TNF- α) was first identified as a factor that induces tumor necrosis, but more recently, it was found to have crucial roles as a pathogenic component of autoimmune disorders.
- ❖ Boulaamane et al., (2023) using molecular docking, it was shown that compounds 13, 5, and 28 had the lowest binding energies against human AChE, measuring -12.2, -12.0, and -12.0 kcal/mol, respectively. These binding energies are influenced by hydrophobic interaction, π - π stacking, and hydrogen bonding within the binding pocket. These data could serve as a basis for creating innovative and reliable AChE inhibitors.
- ❖ Roy et al.,(2010) shows that the LOTUS natural products database was screened for possible α -synuclein inhibitors using the best-performing QSAR model, and the representative compounds picked from each cluster in the ChEMBL dataset were used to create a pharmacophore design. Molecular dynamics simulations ultimately proved that LTS0078917 is more stable than Anle138b, the clinical contender. We discovered that by selectively attaching to the α -synuclein monomer's hairpin-like coil in the N-terminal region, LTS0078917 exhibits promise in stabilizing it.
- ❖ Umar et al., (2020) reveals that a few Jurs parameters, structural parameters, topological branching index, and E-state indices of various fragments were found to be significant by the proposed QSAR models. The internal and external validation parameters indicated that every produced QSAR model was statistically significant.
- ❖ Esposito et al., (2009) demonstrates how a QSAR combined with molecular docking simulation techniques was used to examine a number of seventy-two anti-cancer compounds, determine the essential characteristics needed to create new candidates, and investigate the ways in which the chemicals and V600E-BRAF protein interacted. For both

the internal and exterior datasets, the model's regression coefficients (R^2 (0.864), R^2 adjusted (0.845), Q^2 cv (0.799), and R^2 pred (0.706)) demonstrated an equally strong model with adequate statistical validation keys.

- ❖ Abechi et al., (2023) demonstrates that TNF inhibition is a successful treatment for rheumatoid arthritis patients as well as those with psoriasis, psoriatic arthritis, ankylosing spondylitis, and inflammatory bowel disease. Furthermore, new research has called into doubt the effectiveness of treating AIDS with septic shock.
- ❖ Dermeche et al., (2015) demonstrated that the activity of these newly identified compounds has little to do with their lipophilicity, while longer substituents and less long N single bonds could result in molecules that are highly bioactive. Molecule 74 with PubChem ID 2998055 has superior qualities with a drug grading (ds) of 0.76 when compared to the reference drug grading (ds) of 0.38. Taken together, the identified compounds exhibit excellent pharmacokinetic, pharmacodynamic, and drug interaction characteristics that point to possible TNF-inhibition and absence of toxicity.
- ❖ Bai, et al., (2021) states that thalidomide is a tumor necrosis factor alpha (TNF- α) inhibitor which has been found to have abilities against tumor growth, inflammation. QSAR studies have been performed on twenty-three molecules of thalidomide derivatives. QSAR models have been evaluated for Thalidomide derivatives activities against four types of inflammatory cytokine TNF α .
- ❖ Halder et al., (2023) sought to use network pharmacology and molecular docking to explore the molecular mechanism of Radix Paeoniae Alba (white peony, WP) in treating immune inflammatory disorders of rheumatoid arthritis (RA) and tumor necrosis factor-alpha (TNF- α) inhibitors (TNFis). After screening 71 target genes and 13 active compounds from WP, it was shown that 49 of the target genes overlapped with inflammatory genes targeted by RA and were therefore regarded viable therapeutic targets.

- ❖ Zia et al., (2020) reveals that TNF α is a key target and that medications with strong toxicity, such as adalimumab and infliximab, have pharmacological potency. First, Protein Preparation Wizard and LigPrep were used to import and construct the protein TNF α (PDB ID: 2AZ5) with small molecule inhibitor and the US-FDA-approved therapeutic compounds (from Zinc database), respectively. This was followed by molecular docking, ADMET analysis, and prime MMGBSA. Lastly, these compounds' ADMET characteristics were enhanced by the application of bioisosteric replacement.
- ❖ Jabeen et al., (2020) demonstrates how affordable small-molecule medications that control TNF- α activity or levels could be a substitute for antibody therapies. Our goal is to find new small molecules from our internal database that adhere to the pharmacophoric characteristics of TNF- α inhibitors. Five compounds with good docking scores were then found, and they showed important interactions similar to the reference drug. Three of these compounds have been shown by experimental validation to have high inhibitory capability against TNF- α , with IC₅₀ values of 32.5 ± 4.5 μ M, 6.5 ± 0.8 μ M, and 27.4 ± 1.7 μ M, respectively. The discovered inhibitors may function as anti-inflammatory agents and provide a foundation for the creation of new medications.

Chapter 3

AIM AND SCOPE

Aim:

The aim of this study is to screen various TNF-alpha inhibitors and to identify the potential regions that are critical for TNF-alpha inhibition which helps to design the novel TNF – alpha inhibitors.

Scope:

The project focuses on leveraging in silico methodologies, specifically Quantitative Structure-Activity Relationship (QSAR) analysis and molecular docking, to explore the inhibition of Tumor Necrosis Factor Alpha (TNF- α) as a therapeutic target. The primary objective is to identify and characterize small molecule inhibitors with potential for TNF- α inhibition, offering insights into their binding mechanisms and structure-activity relationships. Novel Therapeutic Agents - Successful identification and validation of novel TNF alpha inhibitors through QSAR and docking approaches can lead to the development of new therapeutic agents for the treatment of inflammatory and autoimmune diseases. These compounds may offer improved efficacy, safety, and specificity compared to existing treatments, potentially transforming patient care. Personalized Medicine- By elucidating the structure-activity relationships and molecular mechanisms underlying TNF alpha inhibition, the project can contribute to the development of personalized medicine approaches. Tailoring treatment regimens based on individual patient characteristics, such as genetic makeup and disease subtype, could optimize therapeutic outcomes and minimize adverse effects. Drug Repurposing and Combination Therapies: The insights gained from the project may identify existing drugs or drug candidates with TNF alpha inhibitory properties, facilitating drug repurposing efforts. Additionally, the project findings can inform the rational design of combination therapies targeting multiple pathways involved in inflammation and immune regulation, potentially synergizing therapeutic effects and reducing the risk of drug resistance.

Chapter 4

MATERIALS AND METHODOLOGY

4.1 Data Collection:

- The ChEMBL 20 database was used to build a data collection of inhibitors against human TNF-alpha (Target ID ChEMBL1825), which included 1451 compounds in total (Gaulton et al., 2012).
- The original data set was put together using a variety of measurement units, such as IC50, Ki, % activity, % inhibition, MIC, EC50, etc., in decreasing order of data size. IC50 was chosen for more research (Simeon et al., 2016).

4.2 Description of inhibitors:

- The vector of fingerprint descriptors accounting for its molecular elements is encoded using the PaDEL-Descriptor program (Yap, 2011). RDKit were used to assess the performance of predictive models from 2D fingerprint descriptors. In this study the 12 different fingerprint descriptors for predicting the bioactivity of TNF-alpha inhibitors was investigated.
- The effectiveness of prediction models derived from 2D fingerprint descriptors was evaluated using RD Kit (Riniker & Landrum 2013). The 12 distinct fingerprint descriptors for predicting TNF-alpha inhibitor bioactivity were examined in this study.
- In addition, the PaDEL-Descriptor software computed the seven molecular descriptors that make up Lipinski's rule-of-five: molecular weight (MW), logarithm of the octanol/water partition coefficient (LogP), number of hydrogen bond donors (nHBDON), number of hydrogen bond acceptors (nHBACC), molar refractivity (MolMR), polar surface area (TPSA), and rotatable bonds (NumRotbonds).

4.3 Data Splitting:

The data set was split into internal and external sets in which the former comprises 80% whereas the latter constitutes 20% of the initial data set (Puzyn et al., 2011).

4.4 Multivariate analysis:

- Supervised learning involves building a model from labeled training data that can be applied to predict future or unseen data. Regression models are built in this work to allow for the prediction of the continuous response variable (pIC50) as a function of predictors (fingerprint descriptors) (Table 1.1) (James et al., 2013).
- The Random Forest (RF) ensemble classifier is made up of many decision trees. In summary, the basic principle of reinforcement learning is to produce many decision trees in order to decrease variance rather than maximize accuracy, rather than creating a deep decision tree with an infinite number of nodes that could lead to overfitting and overtraining of the data. Therefore, compared to a well-trained judgment, the outcomes will have more noise. As such, the results will be more noise when compared to a well-trained decision tree, yet these results are usually reliable and robust (Breiman, 2001).

4.5 Validation of QSAR models:

- Model validation is a crucial procedure that needs to be carried out to guarantee that a fitted model can predict answers for hypothetical or unidentified subjects with accuracy. The Root Mean Square Error (RMSE) and Pearson's correlation coefficient (r) were the two statistical measures utilized to assess the QSAR models' performance. One often used metric to show the strength of the link between two variables of interest is the r value. It can have a range of -1 to +1, where negative values indicate a negative correlation and vice versa between two variables. A metric for evaluating the prediction model's relative inaccuracy is RMSE. The Y-scrambling test, external validation, and 10-fold cross-validation were used to confirm the QSAR models' predictive performance (Eriksson et al., 2003).
- The complete data set is not used in the construction of the predictive model via the 10-fold cross-validation procedure. Rather, it divides the data into training and testing sets, enabling us to evaluate the model's performance on the testing sets using a model that was constructed using the training sets. Repeating the 10-fold validation and calculating average accuracies allows one to accurately evaluate the predictive model's performance.

4.6 Applicability domain analysis:

Based on how closely a compound resembles the compounds used to create the model,

the applicability domain (AD) (Gramatica, 2007) and (Tropsha, et al., 2003), calculates the probability of a trustworthy forecast for that compound. Therefore, substances outside of the AD may result in inaccurate forecasts. It is possible to determine whether new compounds will fall inside or outside of the domain using the leverage value. The calculation of leverage values for each chemical is done by adjusting X , which results in the hat matrix

$$H = X(X^T X)^{-1} X^T \quad (1)$$

where X is a two-dimensional matrix with n compounds and m descriptors, and X^T is X ... In the meantime, the i th diagonal element of

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (2)$$

where x_i is the descriptor row-vector of the i th compound. The warning leverage h^* is calculated by:

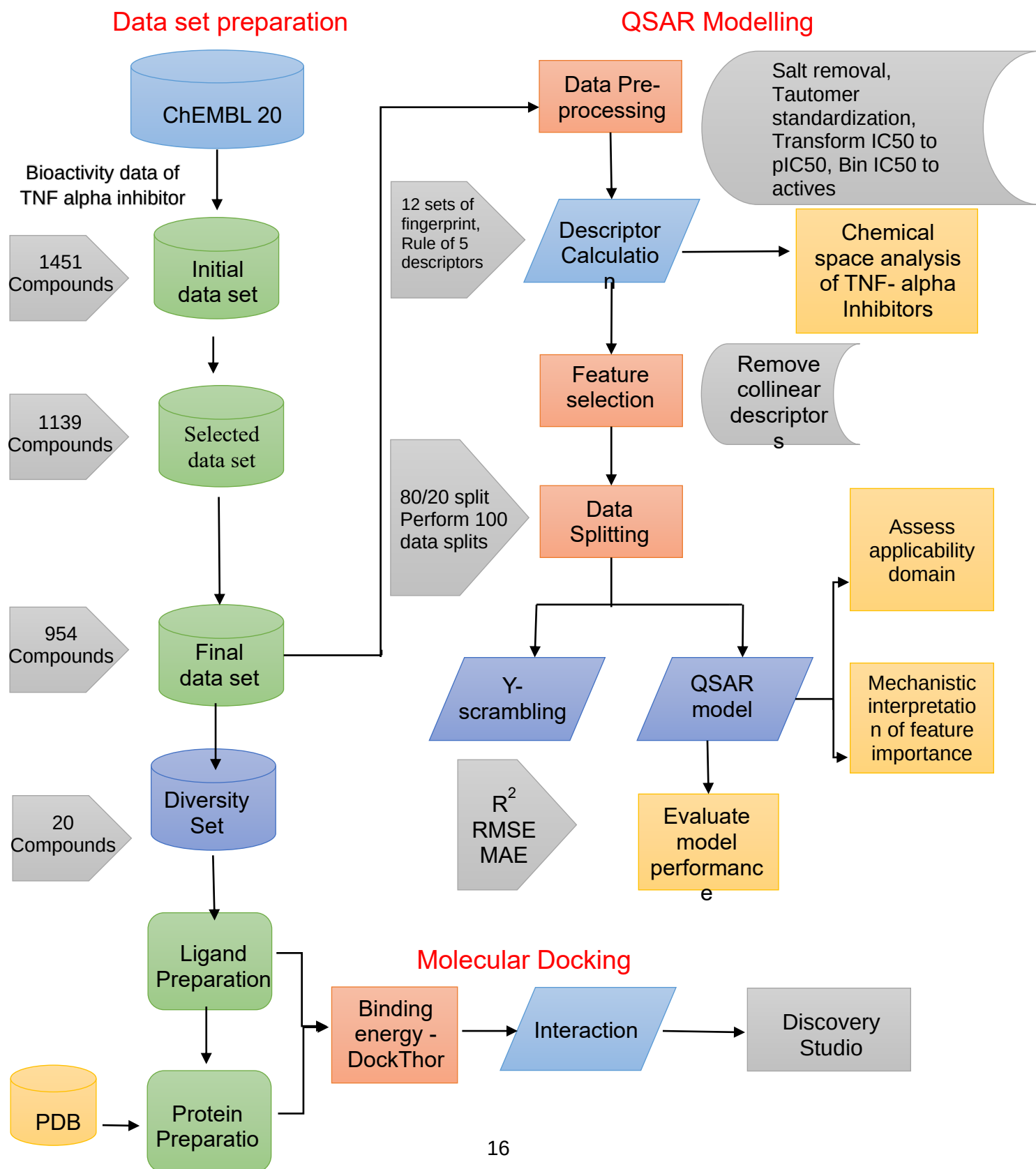
$$h^* = 3(p+1) / n \quad (3)$$

In practice, the AD of QSAR models is frequently evaluated using the William's plot and the leverage value. The standardized residuals vs the leverage value for each compound's h_i are shown to create the William's plot. Should the i th compound possess $h_i > h^*$ then it indicates that the i th component may not be included in the AD and has a significant impact on the QSAR model. Despite this, its standardized residual might be minimal, thus it doesn't seem to be an oddity.

4.7 Molecular docking:

Using the Kennard-Stone algorithm (Kennard & Stone, 1969), a collection of 20 representative and chemically varied compounds—referred to as the diversity set—were extracted from the entire set of active TNF-alpha inhibitors (i.e., $IC_{50} < 1 \mu M$). These substances served as ligands for the molecular docking process directed against human TNF-alpha. Using DockThor, the binding energy (kcal/mol) of TNF-alpha inhibitors was determined. Conformers with the lowest binding energy were chosen for additional binding mode investigation. Next, using Discovery Studio (DS BIOVIA) with the default parameters, the chosen ligands were docked with the human TNF-alpha structure (PDB ID: 2AZ5), and their ligand-protein interactions were produced and examined. Then phytochemicals from the plant *Euphorbia hirta* were also been docked with the TNF-alpha for analyzing the ligand – protein interactions.

WORKFLOW



Chapter 5

RESULTS AND DISCUSSION

5.1 Data Collection:

- The ChEMBL 20 database was used to build a data collection of inhibitors against human TNF-alpha (Target ID ChEMBL1825), which included 1451 compounds in total.
- Upon additional inspection, it was discovered that 312 compounds had no IC₅₀ values given or had indicators that were either more or less than what was required to be removed, leaving 1139 compounds.
- Additionally, in cases where the IC₅₀ standard deviation was less than 2, duplicate compounds with varying bioactivity values were retained, yielding a total of 954 compounds. 954 chemicals made up the total data set that was acquired.

5.2 Chemical space of AChE inhibitors:

By examining the Lipinski's rule-of-five descriptors, the chemical space of TNF-alpha inhibitors was navigated in order to obtain understanding of the structure–activity link. Chemical space analysis may offer crucial information about the overall characteristics of compounds that control their inhibitory qualities. The Lipinski's rule-of-five descriptors, which include MW, LogP, nHBDon, nHBAcc, MolMR, TPSA, and Num Rotatable bonds, were used to analyze exploratory data. The term molecular weight (MW) refers to a compound's size, which is often used since it is simple to compute and interpret. A compound's proper size is also crucial for its transit through lipid membranes. A popular metric for assessing a compound's lipophilicity is logP, which is also used to compute a compound's permeability and membrane penetration. The numbers nHBDon and nHBAcc, which are used to measure hydrogen bonding capacity, represent the number of hydrogen bond donors and acceptors, respectively. In order to explore the chemical space of TNF-alpha inhibitors, Fig. 5.2 displays a visualization of the LogP chemical space as a function of MW. In the MW space, inhibitors were found to be densely distributed, with a range of around 300–700 Da, and in the LogP space, a range of approximately - 2.5 to 7. Furthermore, Fig. 5.2 displays the Lipinski's descriptors box plot. Inactive inhibitors contain

compounds with negative LogP values that are generally closer to 0.0, while the majority of active inhibitors typically have lower average LogP values.

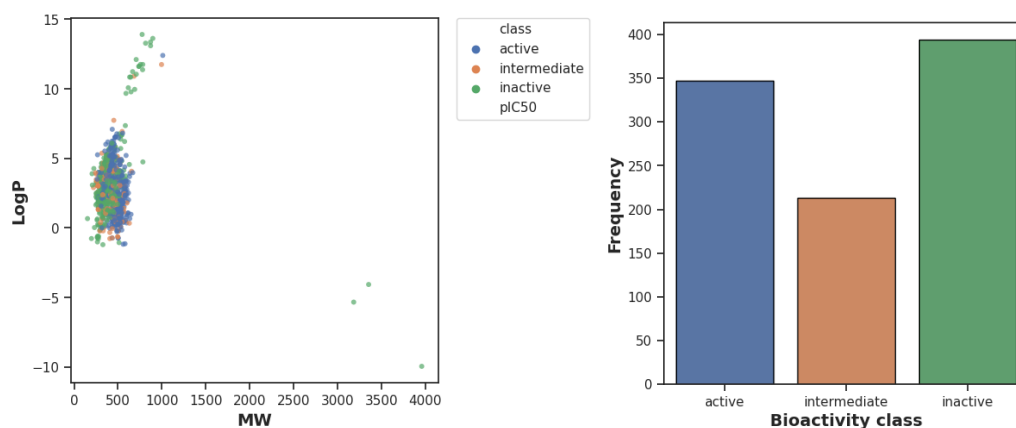


Figure 5.1 - Chemical Space of TNF-alpha inhibitors

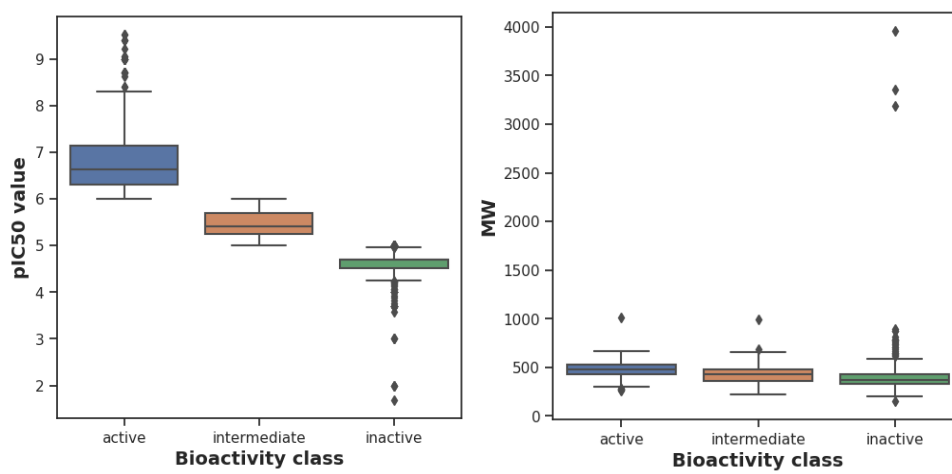
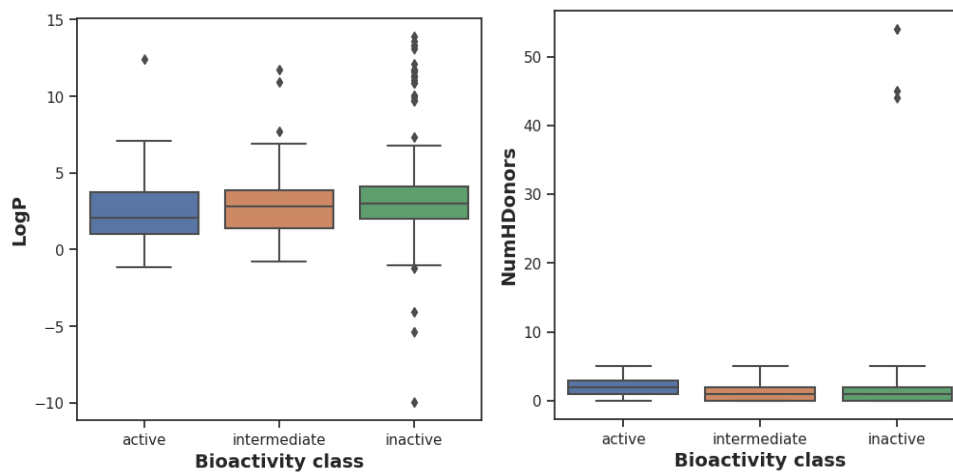


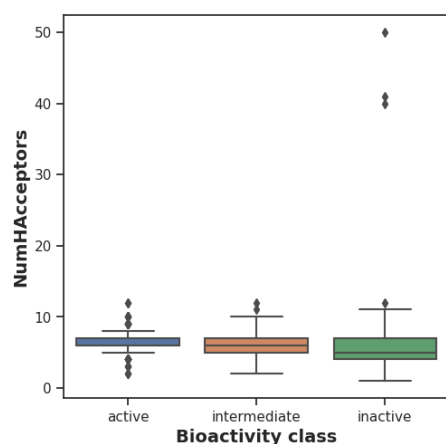
Figure – 5.2

5.2A

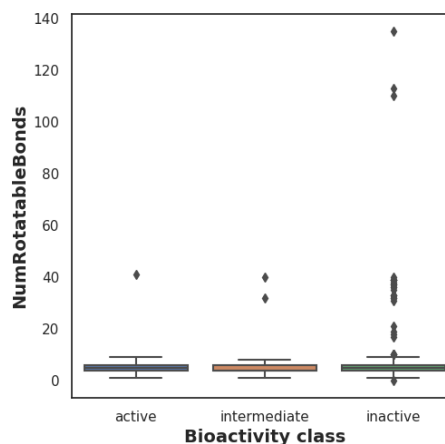


5.2B

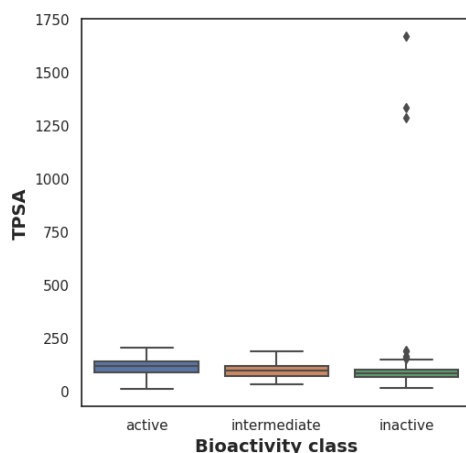
5.2C



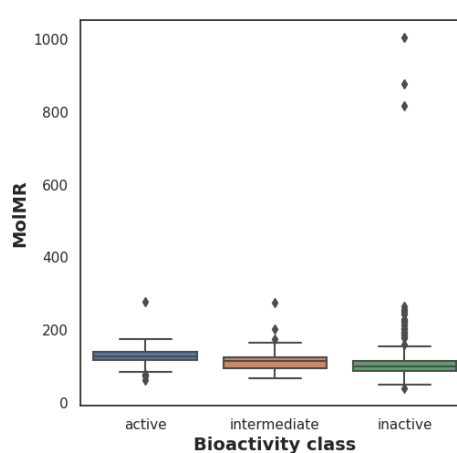
5.2 D



5.2 E



5.2 F



5.2 G

Figure 5.2 Box plot of TNF - alpha inhibitors using Lipinski's rule-of-five descriptors. Shown are 5.2 A- Molecular Weight (MW), 5.2B - Log P, 5.2C – Num H Bond Donors, 5.2D – Num H Bond Acceptors, 5.2E – Num Rotatable Bonds, 5.2F – Polar Surface Area, 5.2G – Molar Refractivity.

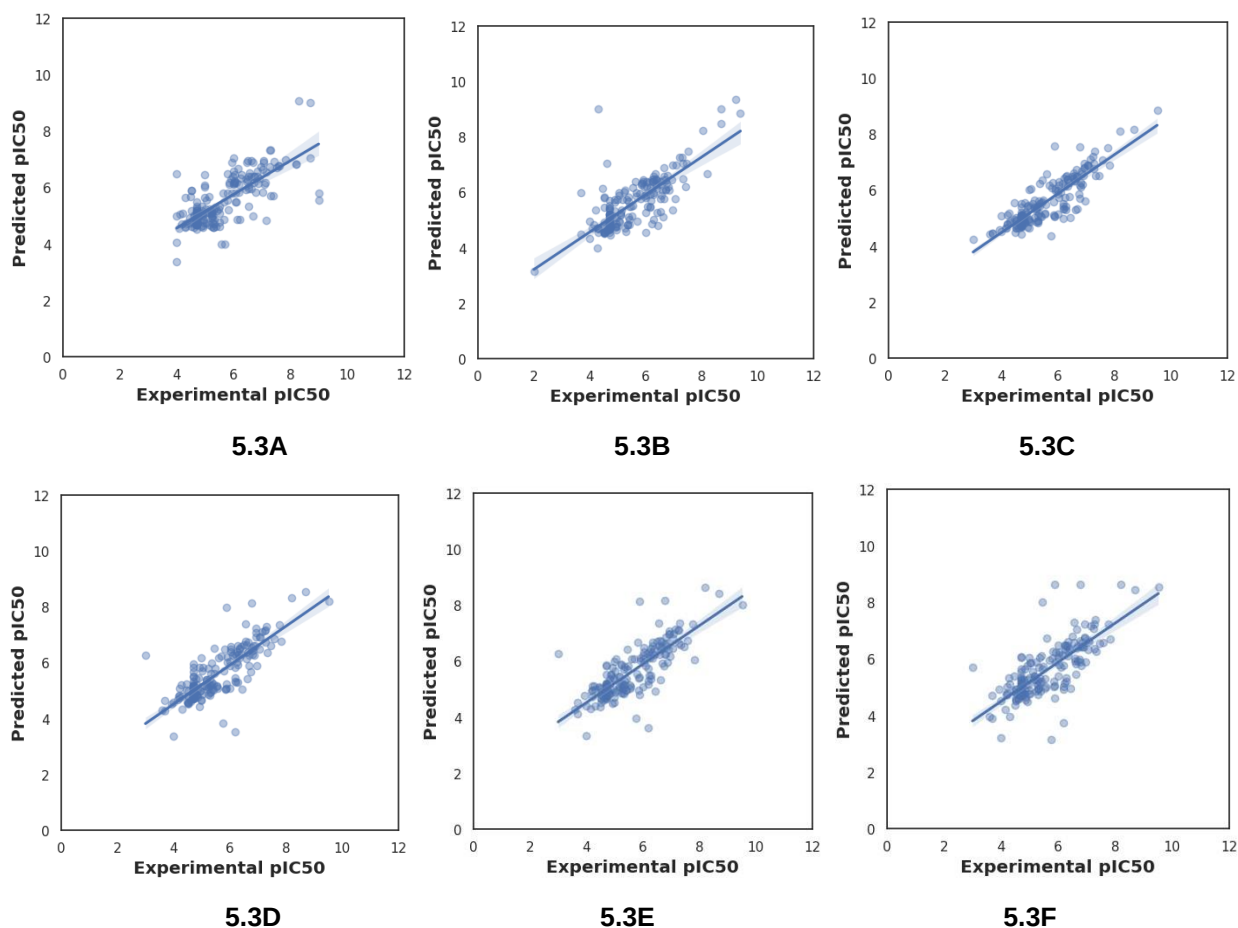
Figure 5.2 displays box plots that visually reflect the overall distribution of data values for Lipinski's descriptors. Specifically, Figs. 5A–G show MW, LogP, NumHDon, NumHAcc, NumRot Bonds, TPSA, and MolMR, respectively. Based on the edges of the boxes, which correspond to the first and third quartiles, an analysis of the box plots showed that there were no differences between the three bioactivity classes for NumRotatable Bonds.

It was discovered that there were variations in the bioactivity for LogP, MW, nHDon, nHBAcc, MolMR, and MolMR. In particular, the lowest LogP values were found for actives, with the remaining five classes showing little any difference. Moreover, of the three bioactivity classes, MW for actives was the biggest, which is followed by the intermediates

while inactives were smallest.

5.3 QSAR model for predicting TNF-alpha inhibitory activity:

954 chemicals made up the data set that was used to build the QSAR models. In particular, the optimal collection of fingerprint descriptors was determined by benchmarking twelve sets of the data. Feature selection was used before modeling in order to exclude collinear descriptors. After that, each of the twelve models was constructed with an 80/20 data split, meaning that 80 percent of the data set was used for the internal set and the remaining 20 percent for the exterior set.



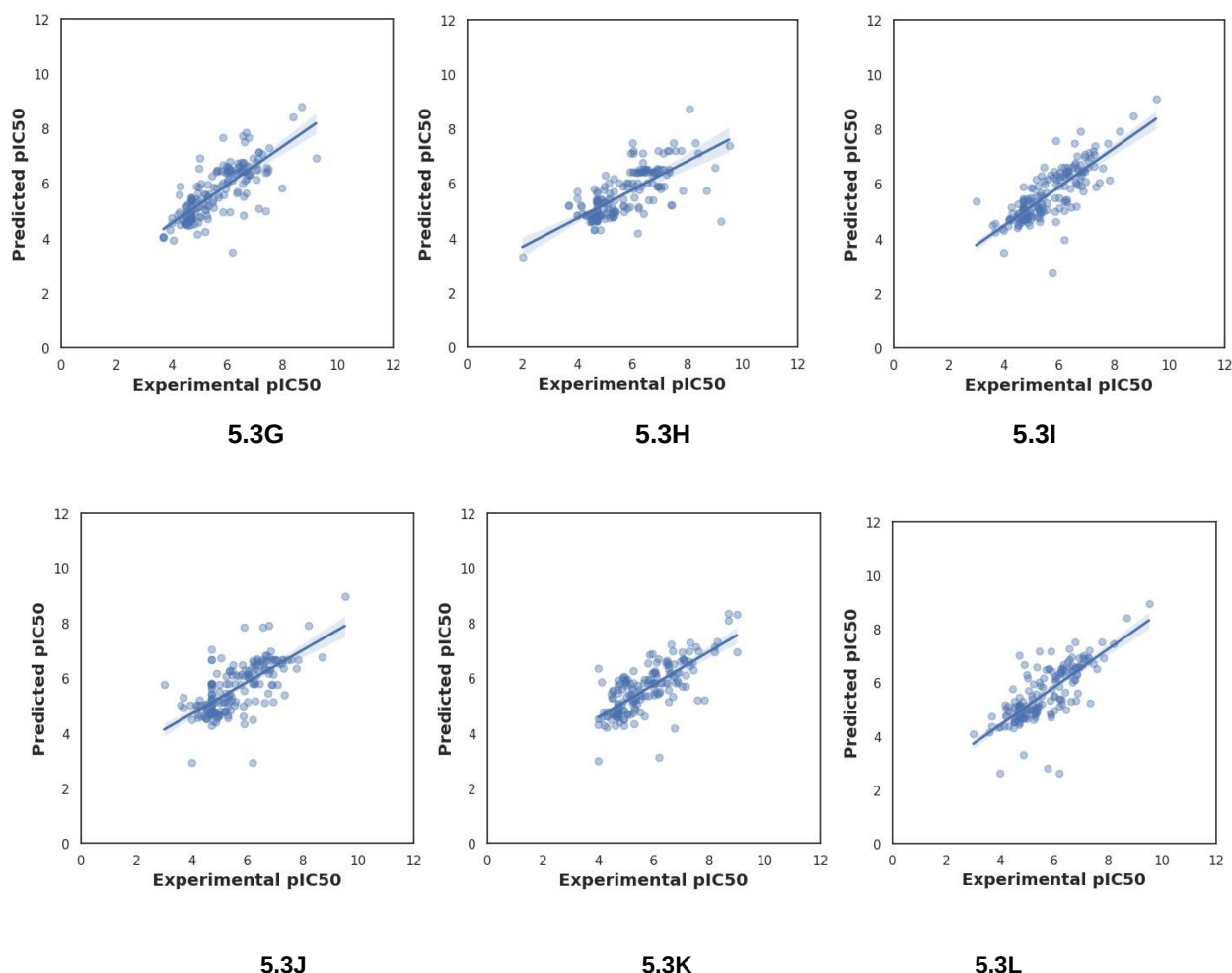


Figure 5.3 Plot of experimental versus predicted pIC50 values for models constructed with 12 different fingerprint descriptors. Shown are models built with Pubchem fingerprint (A), AtomPairs2D fingerprint (B), AtomPairs2D Count (C), CDK fingerprint (D), CDK extended fingerprint (E), CDK Graph Only fingerprint (F), E-state fingerprint (G), MACCS fingerprint (H), Substructure Count fingerprint (I), Substructure fingerprint (J), Klekota-Roth fingerprint (K) and Klekota-Roth Count fingerprint (L).

The model was constructed iteratively, using each of the 100 independent data splits. The mean and standard deviation values obtained from these runs are presented in Table 2 as performance results. Given that all twelve models produced R2, RMSE, and MAE (that is, both training sets and external sets) that were higher than the threshold values, it is evident that all twelve models are able to represent the inhibitory activity space of TNF-alpha inhibitors. Given that the top models have rather large descriptor sizes and are thus prone to overfitting, the best-performing model may not always be the best option. Substructure

count >2D Atom Pairs> Pubchem> Klekotaroth> Substructure> MACCS; E-state> Atom Pairs 2D count> CDK graph only; CDK extended> CDK> Klekotaroth Count> MACCS. This was validated based on their R² value (greater the value greater the effect), RMSE (lesser the value greater the effect) and MAE (lesser the value greater the effect).

Table 5.1 – 10-fold Cross Validation (Performance summary of QSAR models for predicting pIC50)

FingerPrint Descriptors	N	Training Set			External Set			Training Set – External set		
		RMSE	Rsquare	MAE	RMSE	R square	MAE	RMSEext- RMSE tr	R ² ext - Rsq tr	MAE ext - MAE tr
Pubchem	103	26.96157	0.184425	21.67343	26.25396	0.193215	21.211	0.70761	0.00879	0.46243
Klekotaroth	111	25.82823	0.13346	22.33771	24.38191	0.167269	20.72737	1.44632	0.033808	1.61034
MACCS	77	10.43905	0.118791	8.644944	2.006486	0.154033	1.666733	8.432564	0.035243	6.978211
Substructure Count	26	48.46268	0.061213	1.02786	48.45781	0.263566	1.91643	0.00487	0.202354	0.11143
Substructure	30	2.015392	0.138994	1.673514	1.993376	0.163963	1.663702	0.022016	0.024968	0.009812
Klekotaroth Count	72	1.127793	0.092988	0.900069	1.254999	0.080212	1.014327	0.127206	0.012776	0.11426
E state	21	1.055511	0.18905	0.783932	0.931743	0.237193	0.727544	0.123768	0.048142	0.056387
CDK Graph only	198	1.114221	0.100591	0.863468	1.001094	0.166832	0.808021	0.113127	0.066241	0.055446
CDK extended	948	1.118112	0.100054	0.885533	1.005781	0.137107	0.825449	0.112331	0.037053	0.060084
CDK	960	1.104911	0.109144	0.862852	0.985455	0.136872	0.806435	0.119456	0.027728	0.056417
AtomPairs 2D Count	38	1.207711	0.215381	0.937096	0.909893	0.305174	0.703626	0.297818	0.089793	0.233469
AtomPairs 2D	42	1.105714	0.109366	0.879848	0.997547	0.114194	0.828443	0.108167	0.004829	0.051405

5.4 Applicability domain:

The Williams figure in Figure 5.4 was used to define the AD of the suggested QSAR model. The 954 chemicals that made up the employed data set were divided into two subsets at random. Of the two subsets, the first, which made up 80% of the data set, was utilized as an internal set and the second, which made up the remaining 20%, was used as an external set. The Williams plot displays compounds that represent the internal set (blue dots) and external set (red dots). It is evident that nearly all of the 954 compounds were found within the applicability domain's boundaries, suggesting that the AD for our suggested QSAR model was well-defined.

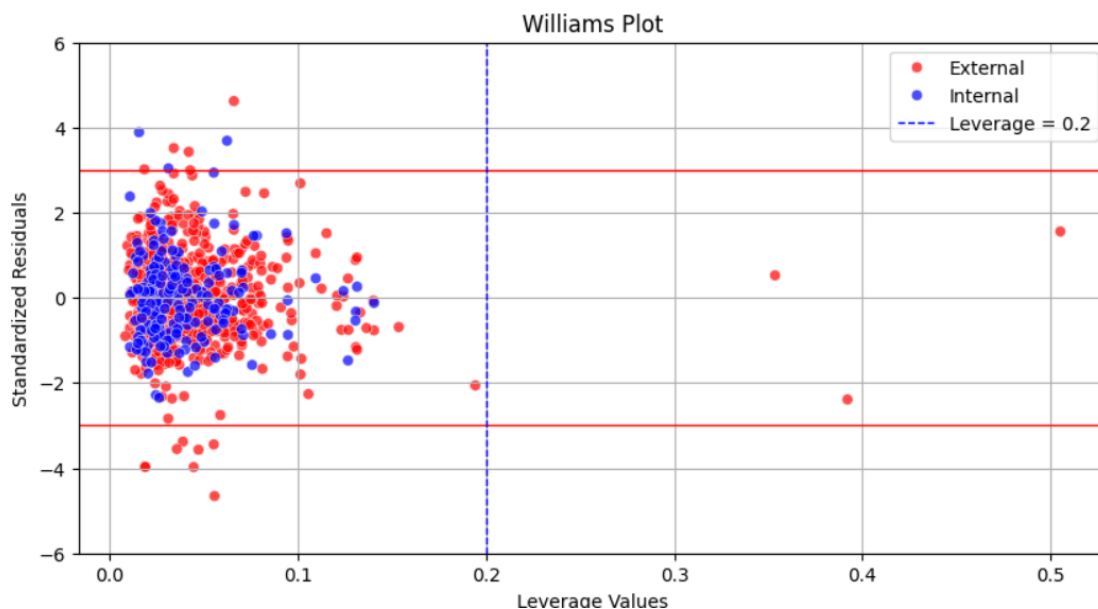


Figure – 5.4 Williams Plot -Evaluation of applicability domain using William plots for QSAR model built with substructure fingerprint count. Compounds in the internal and external sets are shown by blue and red dots, respectively. The solid and dashed lines correspond to the ± 3 standardized residual and the warning leverage value ($h^* = 0.025$), respectively.

In fact, relatively few chemicals fall outside of the ± 3 standardized residual range, as Fig. 5.4 illustrates. This was made up of three compounds (93,124,173) with standardized residuals greater than three from the external set and four compounds (1,212,241,344) from the internal set. Then, the standardized residual for seven compounds (6,152,275,317,393,488,581) in the internal set was less than -3. The leverage value of the three internal set compounds (401,411,705) was greater than 0.2.

5.5 Mechanistic interpretation of feature importance:

An examination of feature importance can help identify characteristics that are crucial for bioactivity. When utilizing the RF method to evaluate the relative relevance of features in models, there are essentially two parameters to consider: (i) accuracy and (ii) the variance of the responses, or the Gini index. In order to forecast the pIC50 of TNF-alpha inhibitors, the latter was used as a metric for prioritizing significant features (i.e., mean decrease of the Gini index).

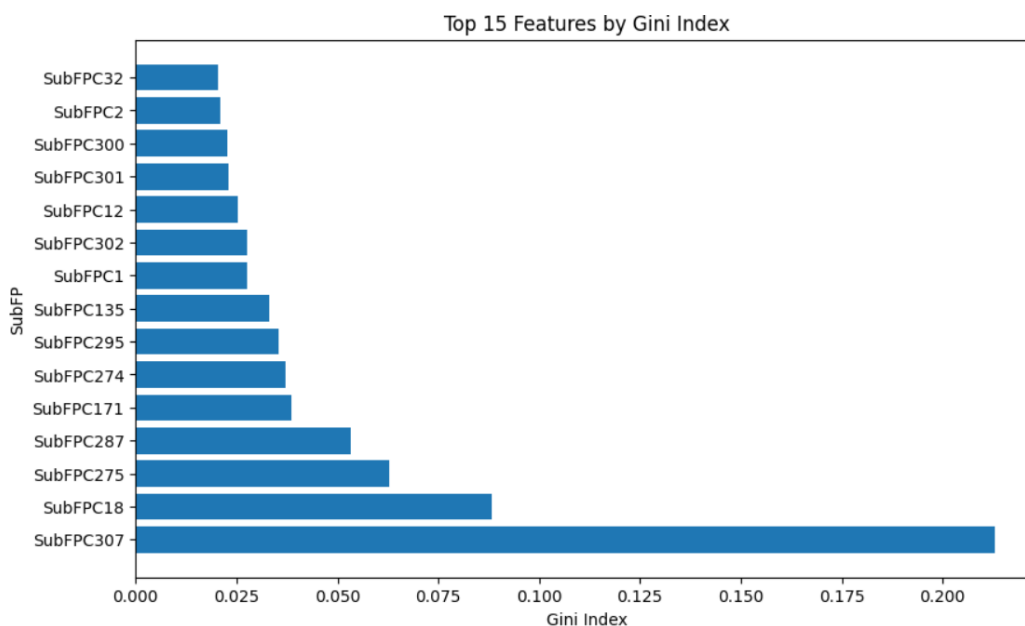


Figure – 5.5 Plot of feature importance as exemplified by the Gini index.

As shown in Fig. 5.5, the most important characteristic is the chiral center specified (SubFPC307), which is also referred to as a stereocenter or asymmetric center. An atom in a molecule can have four different substituents attached to it, forming enantiomers, which are non-superimposable mirror images (H Brooks et al, 2011). The Alkylaryether (SubFPC18) class of organic compounds, which consists of an aryl group (-Ar) and an alkyl group (-R) joined by an ether bond (-O-), is the second most significant characteristic. The heterocyclic ring (SubFPC275) is the third significant substructure. Because they enable π - π stacking contact with important amino acid residues in the TNF-alpha binding region, heterocycles are highly relevant in the creation of TNF-alpha inhibitors. It has been noted that TNF-alpha's binding sites are quite hydrophobic (Li, et al. 2013).

Conjugated Double Bond (SubFPC287), a sequence of alternating single and double

bonds in a molecule where the double bonds are separated by single bonds, is the fourth significant characteristic. The chemical known as Aryl Chloride (SubFPC171), which has a chlorine atom directly linked to an aromatic ring—also known as an aryl group—is the fifth significant substructure. The sixth significant substructure is Aromatic Ring (SubFPC274); the more aromatic rings a drug candidate has on its oral formulation, the more probable it is to be developed and exhibit favorable pharmacodynamics (Ritchie, 2009).

The seventh significant substructure is the C ONS Bond (SubFPC295), which is characterized as any carbon atom in a molecule that is joined to an oxygen, nitrogen, or sulfur atom. When compared to a carbon atom, these atoms are thought to have a high electron density because of their greater electronegativity. Uneven distribution of electron pairs in covalent bonds gives a molecule its polarity and dipole moment, which can produce a dipole–dipole attraction like a hydrogen bond between two polar molecules. Moreover, the presence of hydrogen bonds and increased polarity enhance a drug's water solubility, which is a crucial quality (Chanin et al., 2016). The eighth significant substructure is a carbonyl derivative (SubFPC135), which is necessary for all compounds to form strong covalent interactions with TNF – alpha inhibitors.

Primary carbon (SubFPC1), the dominant location for pi-pi interactions and hydrogen bonding with TNF-alpha inhibitors, is the ninth significant substructure. Rotatable bonds are the tenth significant substructure (SubFPC302). According to the rule of three, a compound can be classified as lead-like if it has rotatable bonds that total no more than three. However, Veber et al. (2002) pointed out that a drug's maximum oral bioavailability is seven rotatable bonds.

Table 5.2 – List of top substructure fingerprints and their corresponding description.

Substructure Count	Description
SubFPC1	Primary Carbon
SubFPC2	Secondary Carbon
SubFPC12	Alcohol
SubFPC18	Alkylaryether

SubFPC32	Secondary Mixed Amine
SubFPC135	Carbonyl derivative
SubFPC171	Aryl Chloride
SubFPC274	Aromatic Ring
SubFPC275	Heterocyclic Ring
SubFPC287	Conjugated Double Bond
SubFPC295	C ONS Bond
SubFPC300	1,3 Tautomerizable
SubFPC301	1,5 Tautomerizable
SubFPC302	Rotatable bond
SubFPC307	Chiral centered Specified

5.6 Molecular Docking of TNF- alpha Inhibitors:

Using the Kennard-Stone algorithm (Kennard & Stone, 1969), a chemically diverse set of 20 representative compounds with an IC₅₀ of less than 1 μ M were extracted from active TNF-alpha inhibitors in order to better understand the non-covalent interaction between TNF-alpha and its inhibitors. These compounds were then investigated for their binding interaction against the TNF-alpha active site. The distribution of the chosen subset of compounds is displayed in Figure 9 alongside the complete collection of actives, which was discovered to offer complete coverage of the initial chemical space. These twenty compounds were docked with the TNF-alpha structure, and DockThor was used to calculate and analyze their binding energy.

The docking finding for these inhibitors are correlated with the binding residues and interactions with the TNF-alpha (Shivaleela, et al. 2022) and (Zia, et al. 2020) such as TyrB:59, TyrB:119, TyrA:59, LeuA:57 etc,

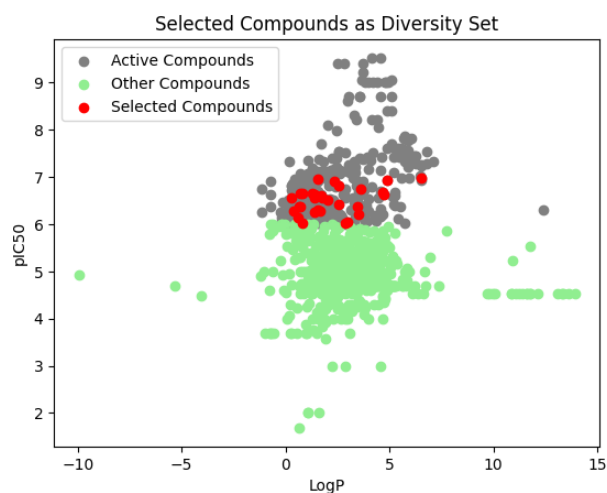


Figure 5.6: Plot showing the distribution of active TNF - alpha inhibitors (gray circles) and the diversity set (red circles) selected for molecular docking.

Table 5.3 – Binding affinity and total energy of top 10 compounds using Dockthor

Comp No.	Binding Affinity	Total energy	vDW energy	Elec. energy
1.	8.777	- 47.682	-102.467	- 155.156
2.	8.337	- 33.930	- 24.330	- 7.294
3.	8.559	- 28.577	- 17.561	- 16.224
4.	8.337	- 23.864	- 14.038	- 17.600
5.	8.149	- 16.789	- 11.862	- 5.149
6.	8.118	- 8.902	- 21.017	- 11.149
7.	7.363	- 3.706	- 23.560	- 7.116
8.	6.755	- 1.026	- 10.897	- 17.326
9.	6.644	1.209	110.801	- 171.710
10.	6.323	3.562	105.467	-169..452

Table 4 displays the total energy and binding affinity of the top 10 chemicals, together with DS BIOVIA analysis of their interactions. The chosen compounds' chemical structures are shown in figure 10. DS BIOVIA has been used to study the protein-ligand interaction, analyzing the type of interaction and identifying the interacting residues. Protein and ligand docking has been done using the LibDock / C Docker. The ligand interaction (5.7A) and the 2D schematic of the ligand-protein interaction (5.7B) of compound 1 are shown in

Figure 5.7.

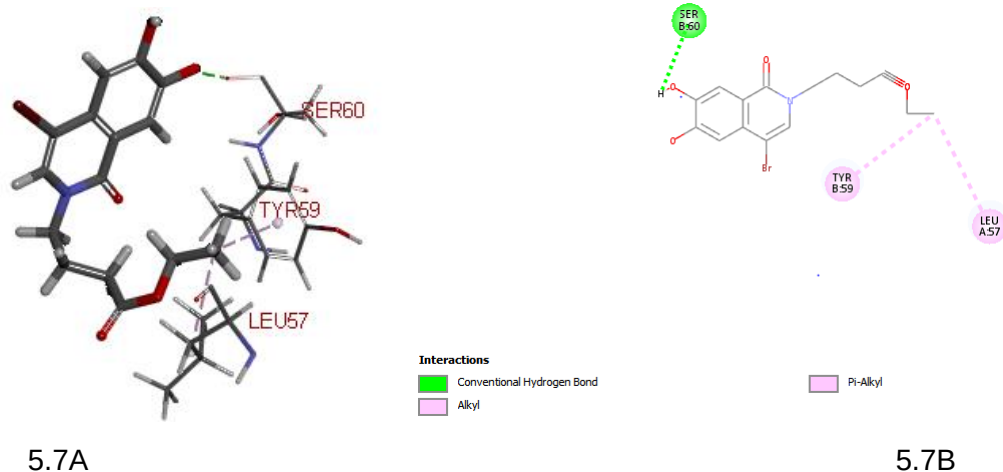


Figure 5.7 – Ligand Interaction of Compound 1 (5.7A) and 2D diagram of interaction (5.7B)

Figure 5.7B reveals three major interactions residues in compound 1 inside the binding pocket of human TNF – alpha. LeuA:57 and TyrB:59 both of which engages in pi-pi stacking interaction with the primary carbon of the ligand. SerB:60 forms a hydrogen bond with the C-O of heterocyclic ring.

The compound 2 showcases five interactions residues where LeuB:57, LeuA:57 and TyrB:151 pi – alkyl interaction with the alcohol and heterocyclic ring of the ligand. GlyA:121 carbon-hydrogen interaction with the alcohol, TyrB:59 pi-pi interaction with the heterocyclic ring of the ligand (Figure 5.8B).

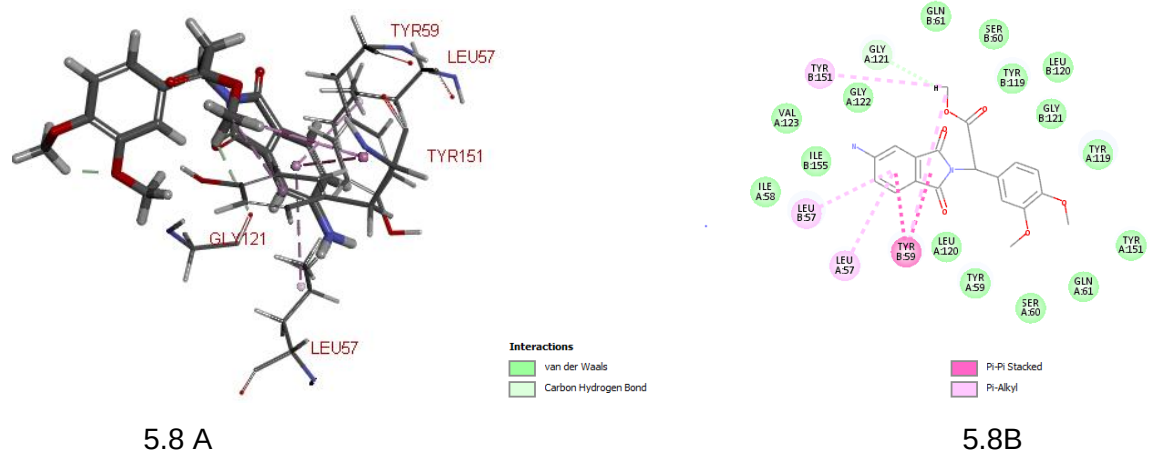
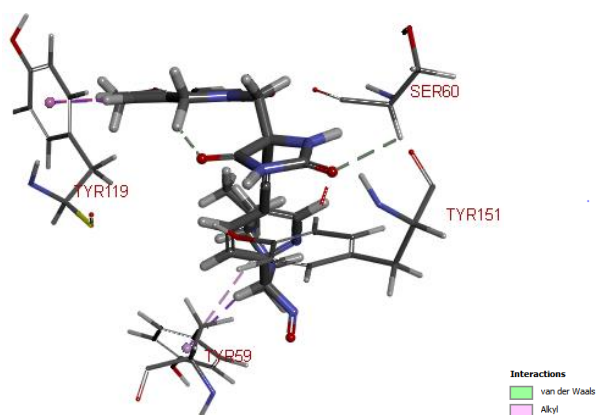
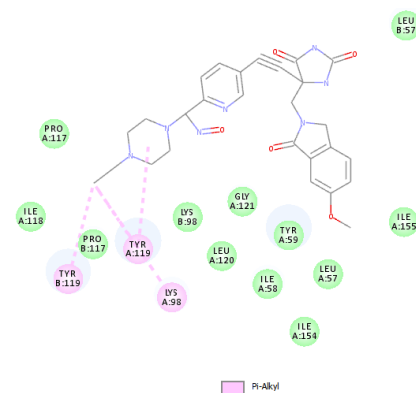


Figure 5.8 – Ligand Interactions of Compound 2 (5.8A) and 2D diagram of interaction (5.8B)

Figure 5.9B reveals three major interactions residues in compound 3 inside the binding pocket of human TNF – alpha. Tyr A:119, Tyr B:119 and Lys A:98 all of which engages in pi- alkyl interaction with the primary carbon and heterocyclic ring of the ligand.



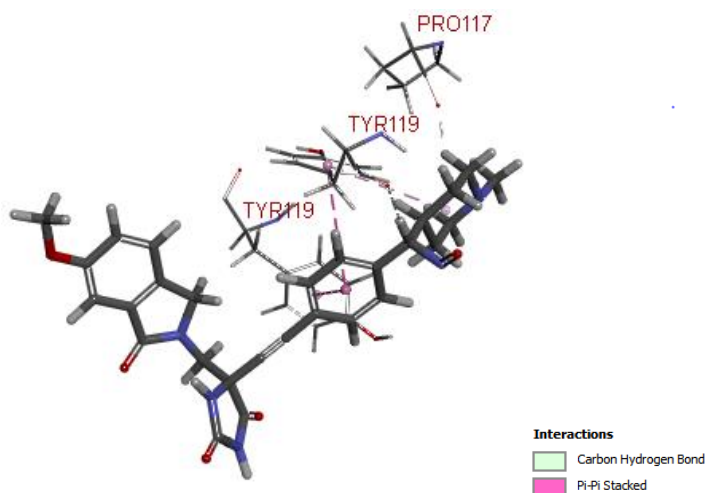
5.9A



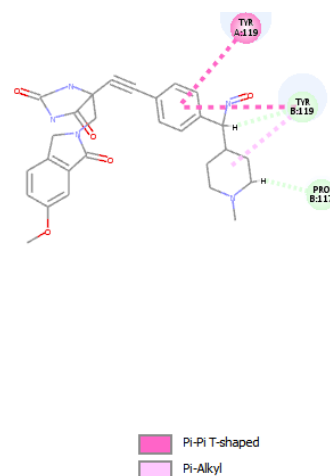
5.9B

Figure 5.9 – Ligand Interactions of Compound 3 (5.9A) and 2D diagram of interaction (5.9B)

Figure 5.10B reveals three major interactions residues in compound 4 inside the binding pocket of human TNF – alpha. ProB:119 and TyrB:119 both of which engages in carbon-hydrogen bond with the heterocyclic ring of the ligand. TyrA:119 forms a pi – pi stacking interaction with aromatic ring.



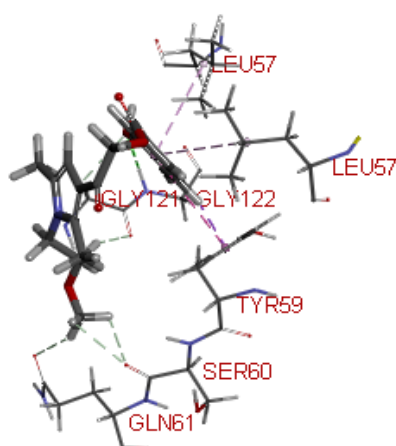
5.10A



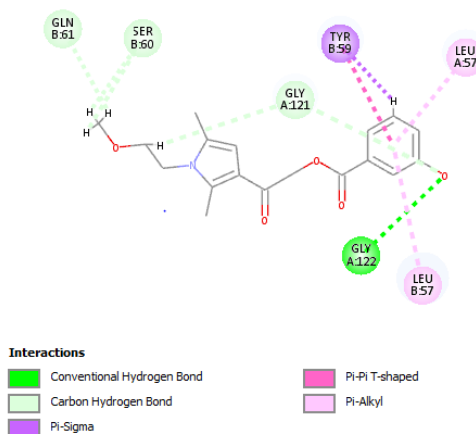
5.10B

Figure 5.10 – Ligand Interactions of Compound 4 (5.10A) and 2D diagram of interaction (5.10B)

The compound 5 showcases seven interactions residues where LeuB:57 and LeuA:57 pi – alkyl interaction with heterocyclic ring of the ligand. GlyA:121, SerB:60 and GlnB:61 shows carbon-hydrogen interaction with primary carbon, C-ONS bond. TyrB:59 pi- sigma interaction with the aromatic ring of the ligand and GlyA:122 shows hydrogen bonding with C-ONS bond of the ligand (Figure 5.11B).



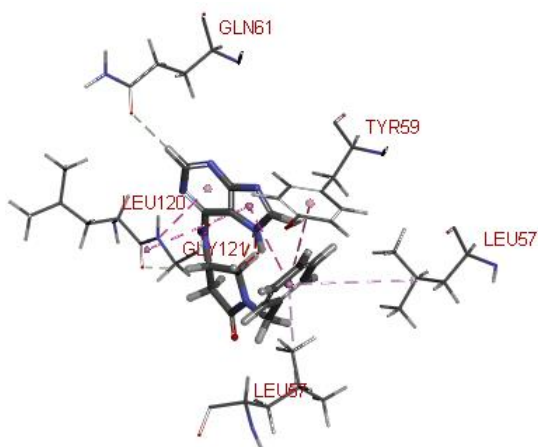
5.11A



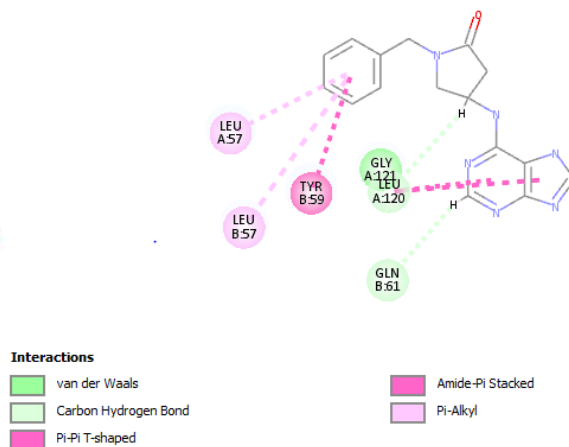
5.11B

Figure 5.11 – Ligand Interactions of Compound 5 (5.11A) and 2D diagram of interaction (5.11B)

Figure 5.12B reveals six major interactions residues in compound 6 inside the binding pocket of human TNF – alpha. LeuB:57 and LeuA:57 both of which engages pi-alkyl in with aromatic ring of the ligand. TyrB:59 forms a pi – pi stacking interaction with aromatic ring. GlyA:121, LeuA:120 and GlnB:61 forms carbon-hydrogen bond with heterocyclic ring.



5.12A



5.12B

Figure 5.12 – Ligand Interactions of Compound 6 (5.12A) and 2D diagram of interaction (5.12B)

The compound 7 showcases four interactions residues where LeuB:57 and LeuA:57 pi – alkyl interaction with aromatic ring of the ligand. TyrB:119, TyrB:151 shows pi – sulfur interaction with C-ONS bond. (Figure 5.13B).

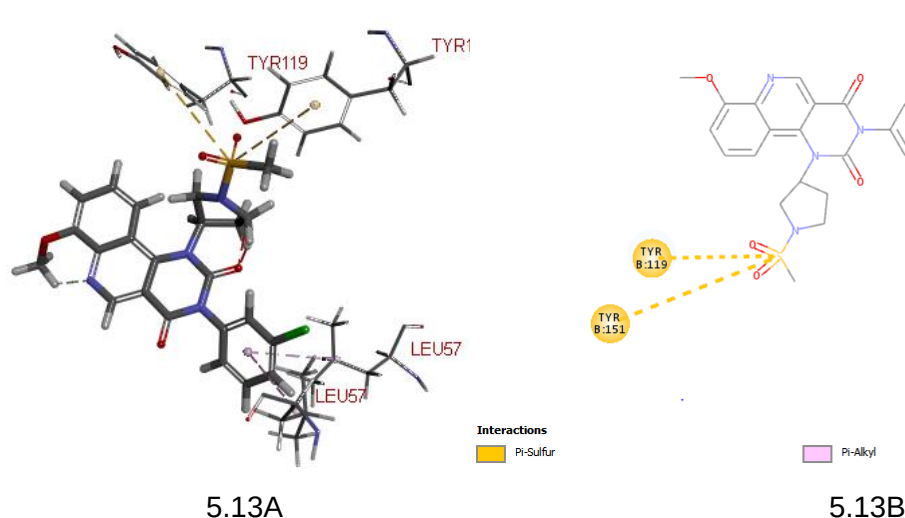


Figure 5.13 – Ligand Interactions of Compound 7 (5.13A) and 2D diagram of interaction (5.13B)

Figure 5.14B reveals seven major interactions residues in compound 8 inside the binding pocket of human TNF – alpha. TyrA:119 engages pi-sigma in with aryl aromatic ring of the ligand. LeuD:55 forms a pi – alkyl interaction with heterocyclic ring. GlyA:121, GlyB:121, TyrB:119 and GlnD:125 forms carbon-hydrogen bond with primary carbon and alcohol of the ligand.

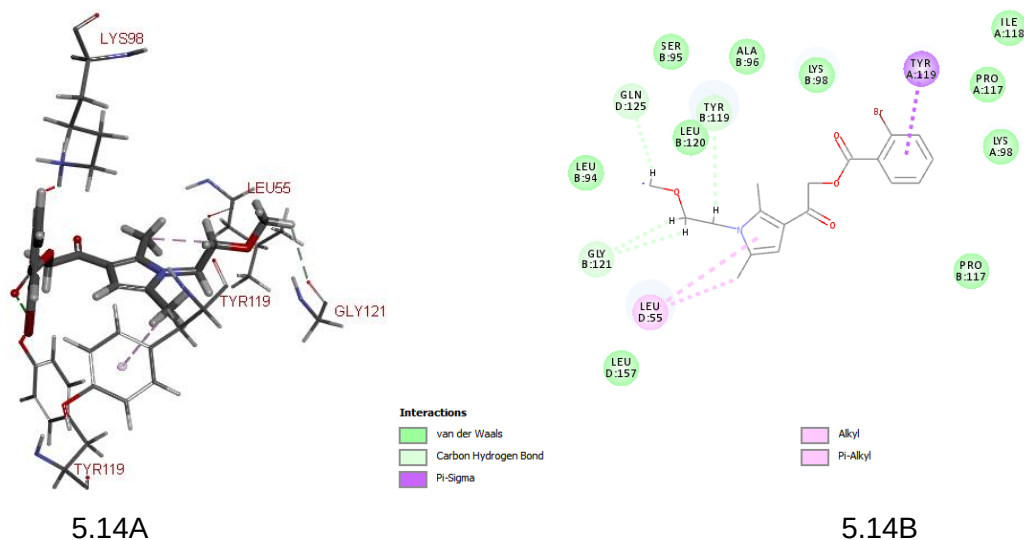


Figure 5.14 – Ligand Interactions of Compound 8 (5.14A) and 2D diagram of interaction (5.14B)

The compound 9 showcases three interactions residues where LeuB:57 shows pi – alkyl interaction with heterocyclic ring of the ligand. TyrA:119 shows pi -alkyl interaction with Aromatic ring. TyrB:59 pi- sigma interaction with the heterocyclic ring of the ligand (Figure 5.15B).

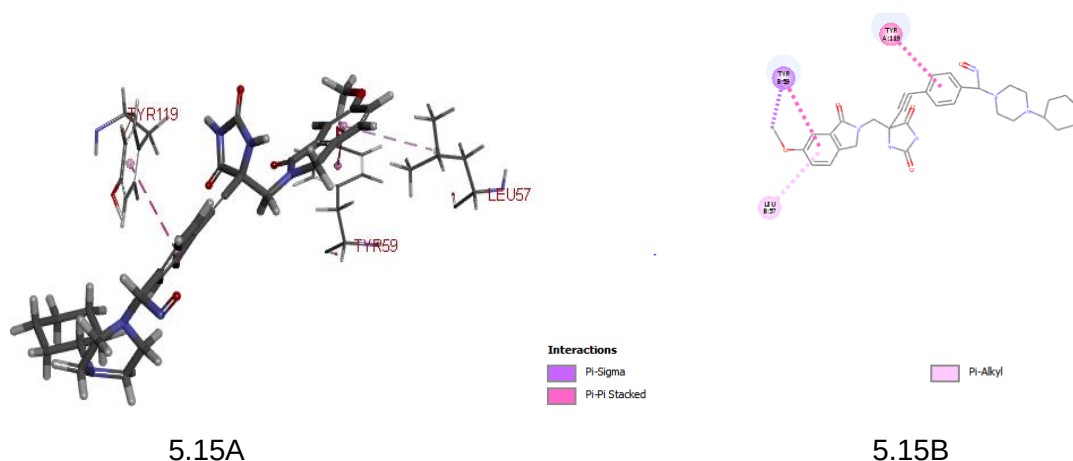


Figure 5.15 – Ligand Interactions of Compound 9 (5.15A) and 2D diagram of interaction (5.15B)

5.7 Molecular Docking of Phytochemical compounds:

Phytochemicals from the plant named *Euphorbia hirta* has been retrieved from IMPATT databases and these phytochemicals were docked with human TNF-alpha to identify their interactions with the target compound. Totally 8 phytochemicals docked with the TNF – alpha and their interactions were been recorded (Table 5)

Table 5.4 – 8 Phytochemicals that were docked with TNF-alpha

S.No.	Name	Docking Affinity
1.	Palmitic acid	-36.1343
2.	Oleic acid	-13.7307
3.	Linoleic acid	-7.38343
4.	Myristic acid	-31.8168
5.	Caffeic acid	-25.3855
6.	Cholorogenic acid	-38.9267
7.	Myricetin	-30.136
8.	Quercetin	-27.9749

The protein - ligand interaction has been carried out using DS BIOVIA and their interacting residues and the type of interaction has been analyzed. The LibDock / C Docker has been used to carry out docking between protein and ligand. Figure 5.16 exhibit the ligand interaction (5.16A), 2D diagram of ligand – protein interaction (5.16B) of palmitic acid.

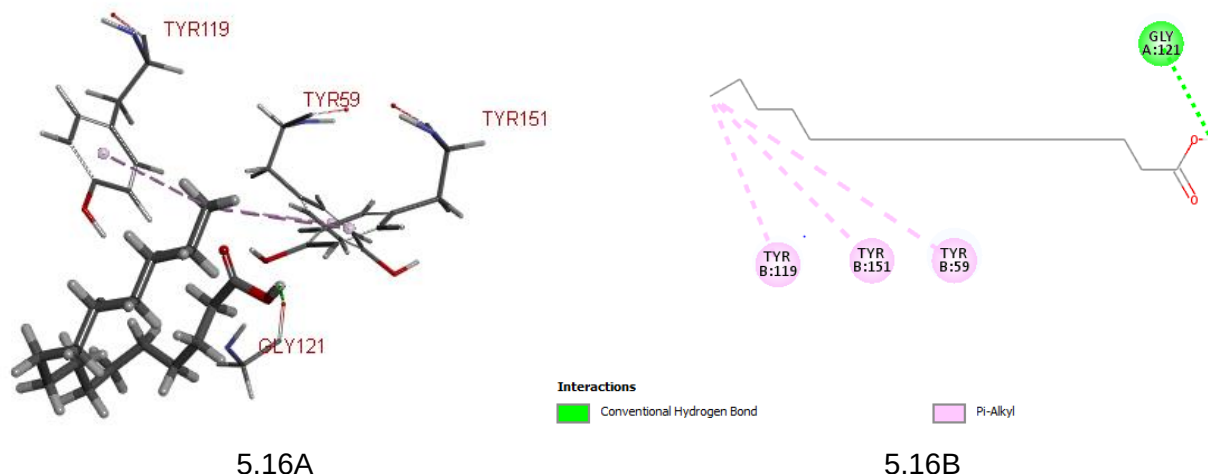


Figure 5.16 – Ligand Interactions of Palmitic acid (5.16A) and 2D diagram of interaction (5.16B)

Figure 5.16B reveals three major interactions residues in palmitic acid inside the binding pocket of human TNF – alpha. TyrB:119, TyrB:151 and TyrB:59 all of which engages in pi-alkyl interaction with the primary carbon of the ligand. The oleic acid showcases four interactions residues where LeuB:94, LeuB:120 both shows pi – alkyl interaction with the primary carbon of the ligand. GlyA:121 hydrogen bond interaction with the carbonyl derivative, LeuA:120 shows carbon-hydrogen interaction with the carbonyl derivative of the ligand (Figure 5.17B).

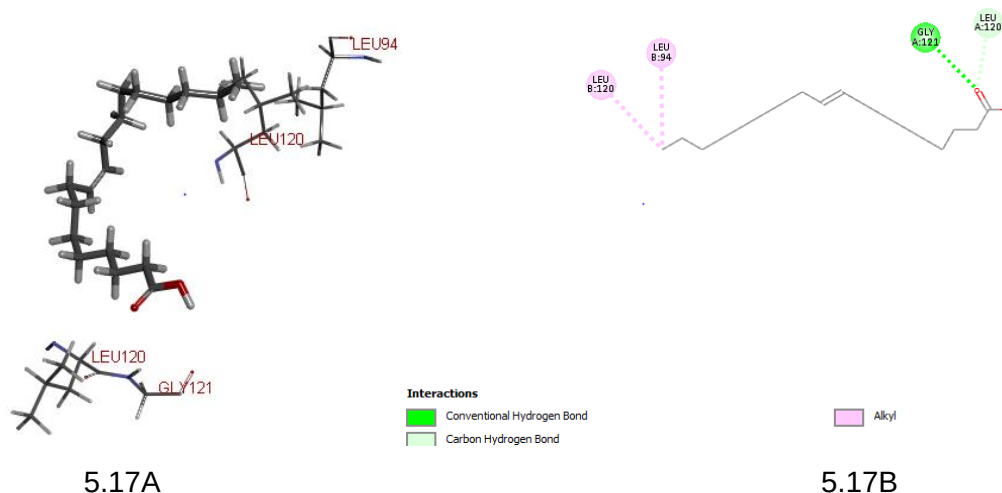


Figure 5.17 – Ligand Interactions of Oleic acid (5.17A) and 2D diagram of interaction (5.17B)

Figure 5.18 B reveals three major interactions residues in Linoleic acid inside the binding pocket of human TNF – alpha. Gly A:122 and Tyr B:59 engages in Carbon-hydrogen interaction with the alcohol and carbonyl derivative, GlyA:121 forms HBond with Alcohol of

the ligand.

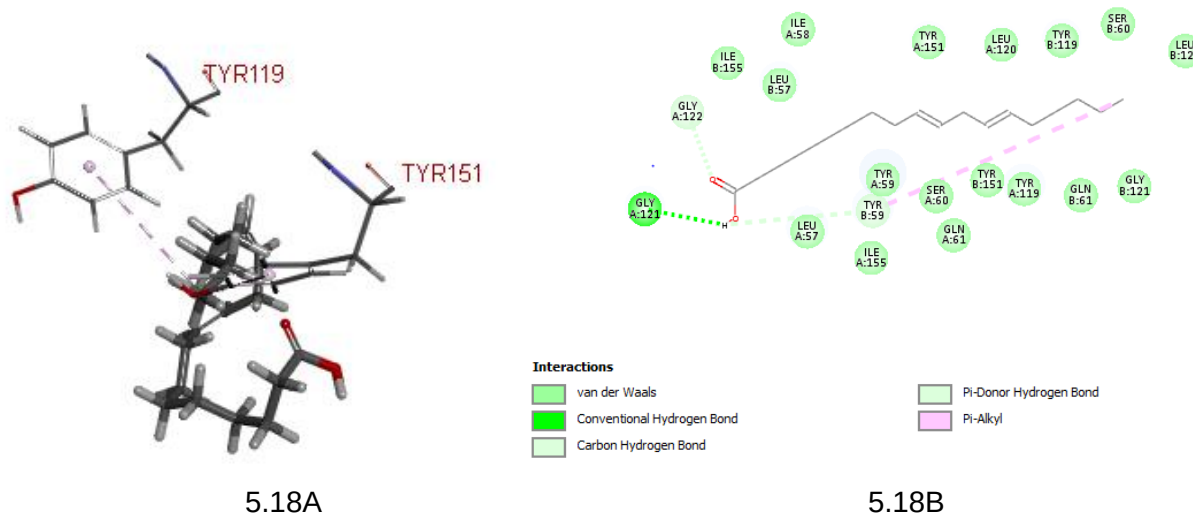


Figure 5.18 – Ligand Interactions of Linoleic acid (5.18A) and 2D diagram of interaction (5.18B)

Figure 5.19B reveals four major interactions residues in Myristic acid inside the binding pocket of human TNF – alpha. LeuA:57 and TyrB:59 both of which engages in pi-alkyl interaction with the primary carbon of the ligand. Then GlyA:121 forms hydrogen bond with Carbonyl derivative and Alcohol and LeuA:120 forms C-H bond with Carbonyl derivative of the ligand.

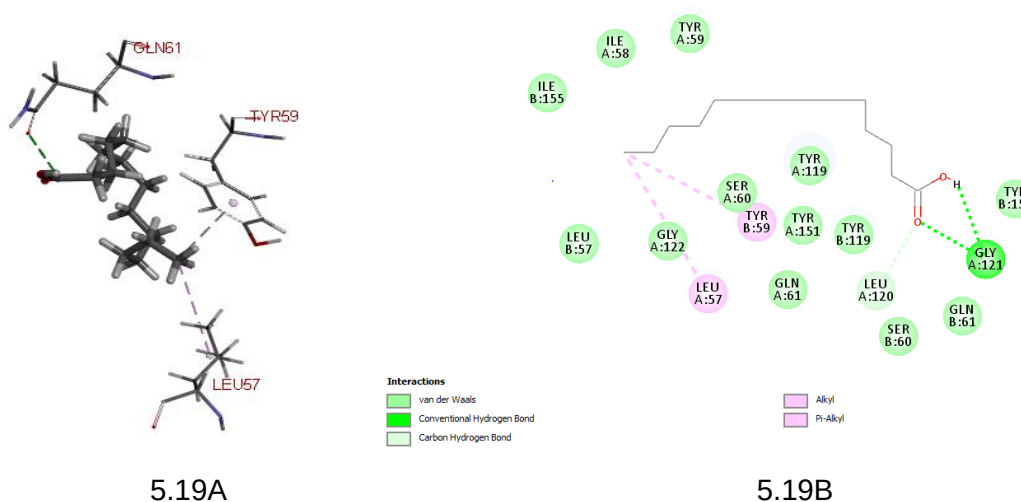
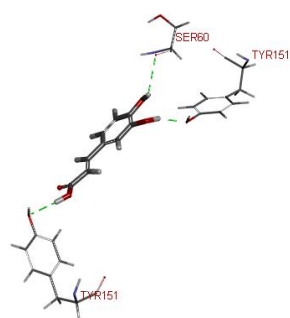
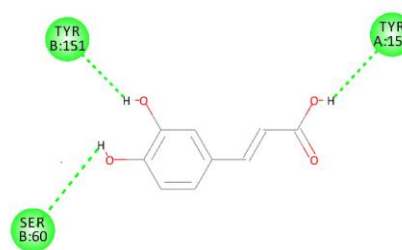


Figure 5.19 – Ligand Interactions of Myristic acid (5.19A) and 2D diagram of interaction (5.19B)

The caffeic acid showcases three interactions residues where TyrA:151, TyrB:151 and SerB:60 forms hydrogen bond interaction with the alcohol, of the ligand (Figure 5.20B).



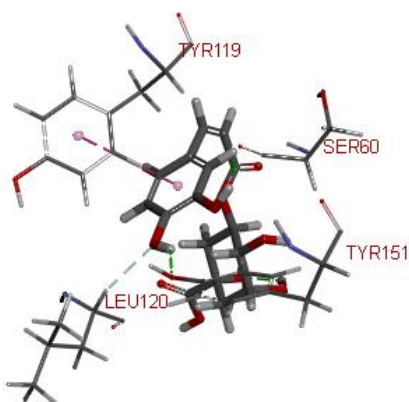
5.20A



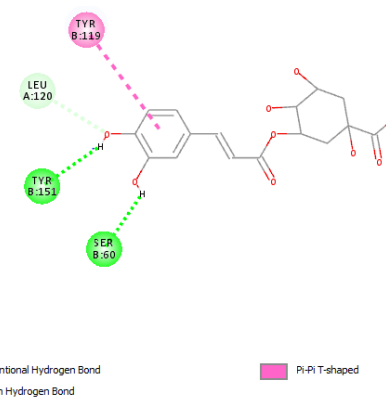
5.20B

Figure 5.20 – Ligand Interactions of Caffeic acid (5.20A) and 2D diagram of interaction (5.20B)

Figure 5.21B reveals four major interactions residues in chlorogenic acid inside the binding pocket of human TNF – alpha. TyrB:119 engages in pi-pi interaction with the aromatic ring of the ligand, LeuA:120 forms C-H bond and TyrB:151, SerB:60 forms H-Bond with alcohol of the ligand.



5.21A



5.21B

Figure 5.21 – Ligand Interactions of Chlorogenic acid (5.21A), 2D diagram of interaction (5.21B)

The myricetin showcases six interactions residues where TyrB:119, TyrA:119 both shows pi – alkyl interaction with the aromatic and heterocyclic ring of the ligand respectively. LeuB:120, SerB:60 and TyrB:151 hydrogen bond interaction with the alcohol, LeuA:120 shows carbon-hydrogen interaction with the alcohol of the ligand (Figure 5.22B).

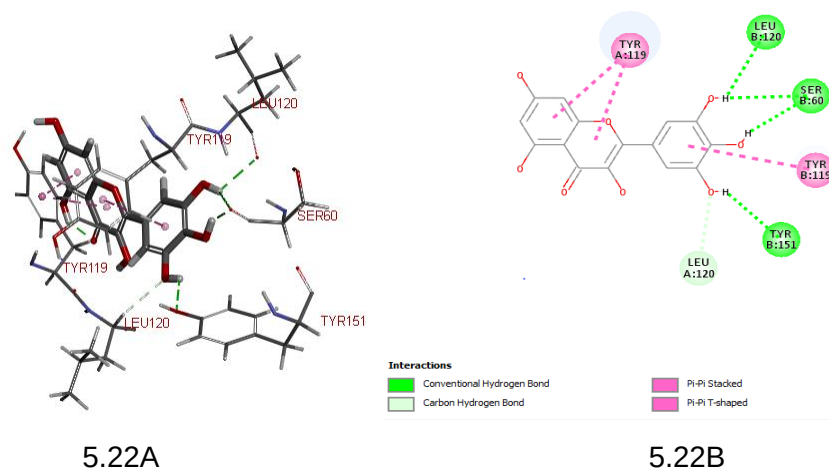


Figure 5.22 – Ligand Interactions of Myricetin (5.22A) and 2D diagram of interaction (5.22B)

Figure 5.23 B reveals seven major interactions residues in Quercetin inside the binding pocket of human TNF – alpha. SerB:60, GlyA:121, TyrB:151 and TyrA:151 all of which engages in H-bond interaction with the alcohol of the ligand, LeuA:120 forms C-H bond with alcohol and TyrB:119, TyrA:119 forms pi-alkyl interaction with the Aromatic and heterocyclic ring of the ligand respectively.

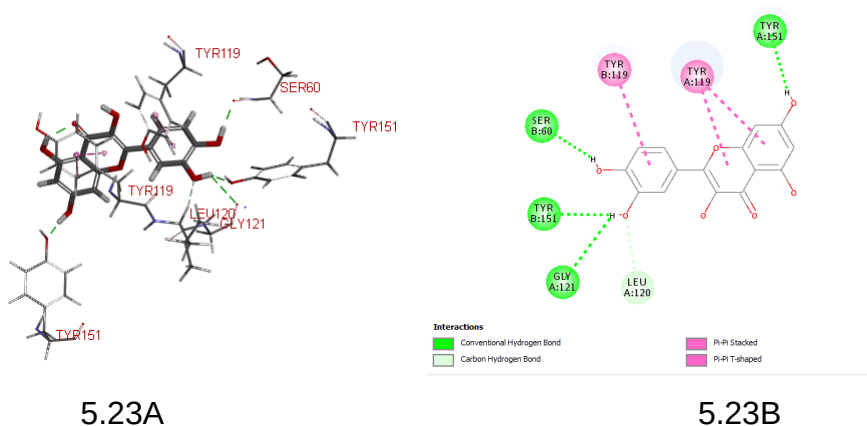


Figure 5.23 – Ligand Interactions of Quercetin (5.23A) and 2D diagram of interaction (5.23B)

Interestingly, the QSAR model's feature importance results, as presented in Table 5.2, remarkably support this finding. Among the top ten crucial substructures required for the bioactivity of TNF-alpha inhibitor are the aromatic moiety, alcohol, heterocyclic ring, C ONS bond, primary carbon, and carbonyl derivative moiety. Furthermore, the results of the QSAR (Table 5.2) in which the aromatic moiety, heterocyclic ring, primary carbon, alcohol, and carbonyl derivative moiety for the bioactivity of TNF-alpha inhibitor are listed, verified the findings of the phytochemical substances.

Chapter 6

SUMMARY & CONCLUSION

In conclusion, a comparative analysis of the performances of the twelve sets of fingerprint descriptors that were utilized to build QSAR models was conducted. A number of fingerprint descriptors were found to provide the built models with good performance, suggesting that they could capture the feature space of TNF-alpha inhibitors. The following significant features that are essential for TNF-alpha inhibition were found by utilizing the built-in feature importance estimator from RF known as the Gini index: chiral centered specified (SubFPC307), heterocyclic ring (SubFPC275), aryl chloride (SubFPC171), aromatic ring (SubFPC274), alcohol (SubFPC12), and primary carbon (SubFPC1). The above-mentioned conclusions from the QSAR models, according to which the aromatic, alcohol, heteroaromatic, and heterocyclic rings were preferred moieties for interacting with the hydrophobic pocket of TNF-alpha. The information gathered from this study is expected to be used as general guidelines for the design of novel TNF-alpha inhibitors.

Chapter 7

REFERENCES

1. Adegbola, S.O.; Sahnan, K.; Warusavitarne, J.; Hart, A.; Tozer, P. Angiotensin-Converting enzyme gene polymorphism in patients with psoriatic arthritis. *Rheumatol. Int.* 2007, 27, 1119–1123. DOI: <https://doi.org/10.1007/s00296-007-0349-y>
2. Baumgart, D.C.; Carding, S.R. Gastroenterology 1—Inflammatory bowel disease: Cause and immunobiology. *Lancet* 2007, 369, 1627–1640. DOI: [10.1016/S0140-6736\(07\)60750-8](https://doi.org/10.1016/S0140-6736(07)60750-8)
3. Brooks, H. W., Guida, W. C., & Daniel, K. G. (2011). The significance of chirality in drug design and development. *Current Topics in Medicinal Chemistry*, 11(7), 760-770. DOI: [10.2174/156802611795429235](https://doi.org/10.2174/156802611795429235)
4. Chaman Lal (2007). *Indian Journal of Critical Care Medicine*, 11(3), 139-148. DOI: [10.4103/0972-5229.37820](https://doi.org/10.4103/0972-5229.37820)
5. Choy, E. H. S., & Panayi, G. S. (2001). Mechanisms of disease: Cytokine pathways and joint inflammation in rheumatoid arthritis. *New England Journal of Medicine*, 344, 907–916. DOI: [10.1056/nejm200103223441207](https://doi.org/10.1056/nejm200103223441207)
6. Curtis, J.R.; Mariette, X.; Gaujoux-Viala, C.; Blauvelt, A.; Kvien, T.K.; Sandborn, W.J.; Winthrop, K.; de Longueville, M.; Huybrechts, I.; Bykerk, V.P. Long-Term safety of certolizumab pegol in rheumatoid arthritis, axial spondyloarthritis, psoriatic arthritis, psoriasis and Crohn's disease: A pooled analysis of 11 317 patients across clinical trials. *RMD Open* 2019, DOI: [10.1136/rmdopen-2018-000940](https://doi.org/10.1136/rmdopen-2018-000940)
7. Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. DOI: [10.1021/ci010132r](https://doi.org/10.1021/ci010132r)
8. Esposito, E., & Cuzzocrea, S. (2009). TNF-alpha as a therapeutic target in inflammatory diseases, ischemia-reperfusion injury and trauma. *Current Medicinal Chemistry*, 16(24), 3152–3167. DOI: [10.2174/092986709788803032](https://doi.org/10.2174/092986709788803032)
9. Feldmann, M., Brennan, F. M., & Maini, R. N. (1996). Role of cytokines in rheumatoid arthritis. *Annual Review of Immunology*, 14, 397–440. DOI: [10.1146/annurev.immunol.14.1.397](https://doi.org/10.1146/annurev.immunol.14.1.397)
10. Goldring, S. R. (2002). Pathogenesis of bone erosions in rheumatoid arthritis. *Current*

Opinion in Rheumatology, 14, 406–410. DOI: [10.1097/00002281-200207000-00011](https://doi.org/10.1097/00002281-200207000-00011)

11. Guedes, I. A., Krempser, E., & Dardenne, L. E. (2017). DockThor 2.0: a free web server for protein-ligand virtual screening. XIX SBQT–Simpósio Brasileiro de Química Teórica, 2013, 2013-2014. DOI: [10.13140/RG.2.2.27424.58882](https://doi.org/10.13140/RG.2.2.27424.58882)
12. Guo, Q., Wang, Y., Xu, D., Nossent, J., Pavlos, N. J., & Xu, J. (2018). Rheumatoid arthritis: Pathological mechanisms and modern pharmacologic therapies. Bone Research, 6, 15. DOI: [10.1038/s41413-018-0016-y](https://doi.org/10.1038/s41413-018-0016-y)
13. Hall, L. H., & Kier, L. B. (1995). Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. DOI: [10.1021/ci00027a005](https://doi.org/10.1021/ci00027a005)
14. Isarankura-Na-Ayudhya, C., Naenna, T., Nantasenamat, C., & Prachayasittikul, V. (2009). A practical overview of quantitative structure-activity relationship. DOI: [10.1002/qsar.200810151](https://doi.org/10.1002/qsar.200810151)
15. Jejurikar, B. L., & Rohane, S. H. (2021). Drug designing in discovery studio. DOI: [10.5772/intechopen.98145](https://doi.org/10.5772/intechopen.98145)
16. Jiang, Y., Yu, M., Hu, X., Han, L., Yang, K., Ba, H., Zhang, Z., Yin, B., Yang, X. P., Li, Z., & Wang, J. (2017). STAT1 mediates transmembrane TNF-alpha-induced formation of death-inducing signaling complex and apoptotic signaling via TNFR1. Cell Death and Differentiation, 24(4), 660–671. DOI: [10.1038/cdd.2016.158](https://doi.org/10.1038/cdd.2016.158)
17. Kim, S., Bolton, E.E. & Bryant, S.H. PubChem3D: conformer ensemble accuracy. J Cheminform 5, 1 (2013). DOI: <https://doi.org/10.1186/1758-2946-5-1>
18. Koch, A. E., Harlow, L. A., Haines, G. K., Amento, E. P., Unemori, E. N., Wong, W. L., Pope, R. M., & Ferrara, N. (1994). Vascular Endothelial Growth-Factor—A Cytokine Modulating Endothelial Function in Rheumatoid-Arthritis. Journal of Immunology, 152, 4149–4156. DOI: [10.1002/path.2287](https://doi.org/10.1002/path.2287)
19. Klekota, J., & Roth, F. P. (2009). Chemical substructures that enrich for biological activity. DOI: [10.1093/bioinformatics/btn486](https://doi.org/10.1093/bioinformatics/btn486)
20. Levin, A.D.; Wildenberg, M.E.; van den Brink, G.R. Mechanism of Action of Anti-TNF Therapy in Inflammatory Bowel Disease. J. Crohns Colitis 2016, 10, 989–997. DOI: [10.1093/ecco-icc/jjw053](https://doi.org/10.1093/ecco-icc/jjw053)
20. Li, J. J. (Ed.). (2013). Heterocyclic chemistry in drug discovery. John Wiley & Sons.

DOI: [10.1002/9781118742983](https://doi.org/10.1002/9781118742983)

21. Lipinski CA. J Pharmacol Toxicol Methods. 2000;44(1):235. DOI: [10.1016/s1056-8719\(00\)00107-6](https://doi.org/10.1016/s1056-8719(00)00107-6)
22. Mazumdar, S.; Greenwald, D. Golimumab. mAbs 2009, 1, 422–431, DOI: [10.4161/mabs.1.5.9571](https://doi.org/10.4161/mabs.1.5.9571)
23. Melsheimer, R.; Geldhof, A.; Apaolaza, I.; Schaible, T. Remicade (R) (infliximab): 20 years of contributions to science and medicine. Biol. Targets Ther. 2019, 13, 139–178. DOI: [10.2147/BTT.S207246](https://doi.org/10.2147/BTT.S207246)
24. Merola, J. F., Espinoza, L. R., & Fleischmann, R. (2018). Distinguishing rheumatoid arthritis from psoriatic arthritis. RMD Open. DOI: [10.1136/rmdopen-2018-000656](https://doi.org/10.1136/rmdopen-2018-000656)
25. Monaco, C.; Nanchahal, J.; Taylor, P.; Feldmann, M. Anti-TNF therapy: Past, present and future. Int. Immunol. 2015, 27, 55–62. DOI: [10.1093/intimm/dxu102](https://doi.org/10.1093/intimm/dxu102)
26. Ogawa, E.; Sato, Y.; Minagawa, A.; Okuyama, R. Pathogenesis of psoriasis and development of treatment. J. Dermatol. 2018, 45, 264–272. DOI: [10.1111/1346-8138.14139](https://doi.org/10.1111/1346-8138.14139)
27. Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: a review. DOI: [10.1007/s12551-017-0241-4](https://doi.org/10.1007/s12551-017-0241-4)
28. Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A. et al. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. Struct Chem 22, 795–804 (2011). DOI: <https://doi.org/10.1007/s11224-011-9757-4>
29. Roy, P. P., & Roy, K. (2010). Molecular docking and QSAR studies of aromatase inhibitor androstenedione derivatives. Journal of Pharmacy and Pharmacology. DOI: [10.1211/jpp.62.12.0004](https://doi.org/10.1211/jpp.62.12.0004)
30. Salmina, E. S., Haider, N., & Tetko, I. V. (2015). Extended functional groups (EFG): an efficient set for chemical characterization and structure-activity relationship studies of chemical compounds. DOI: [10.3390/molecules21010001](https://doi.org/10.3390/molecules21010001)
31. Scutellari, P. N., & Orzincolo, C. (1998). Rheumatoid arthritis: Sequences. European Journal of Radiology, 27(Suppl. 1), S31–S38. DOI: [10.1016/s0720-048x\(98\)00067-4](https://doi.org/10.1016/s0720-048x(98)00067-4)
32. Shealy, D.; Cai, A.; Lacy, E.; Nesspor, T.; Staquet, K.; Johns, L.; Baker, A.; Brigham-Burke, M.; Emmell, E.; Bugelski, P.; et al. Characterization of golimumab (CNTO 148), a novel fully human monoclonal antibody specific for human TNF α . Ann. Rheum. Dis. 2007, 66, 151. DOI: [10.1136/ard.2007.079426](https://doi.org/10.1136/ard.2007.079426)

33. Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A. A., Prachayasittikul, V., Wikberg, J. E. S., & Nantasenamat, C. (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. DOI: [10.7717/peerj.2322](https://doi.org/10.7717/peerj.2322)
34. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. DOI: [10.1021/ci025584y](https://doi.org/10.1021/ci025584y)
35. Tracey, D.; Klareskog, L.; Sasso, E.H.; Salfeld, J.G.; Tak, P.P. Tumor necrosis factor antagonist mechanisms of action: A comprehensive review. *Pharmacol. Ther.* 2008, 117, 244–279. DOI: [10.1016/j.pharmthera.2007.10.001](https://doi.org/10.1016/j.pharmthera.2007.10.001)
36. Umar, A. B., Uzairu, A., Shallangwa, G. A., & Uba, S. (2020). QSAR modelling and molecular docking studies for anti-cancer compounds against melanoma cell line SK-MEL-2. *Heliyon*. DOI: [10.1016/j.heliyon.2020.e03645](https://doi.org/10.1016/j.heliyon.2020.e03645)
37. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12), 2615-2623. DOI: [10.1021/jm020017n](https://doi.org/10.1021/jm020017n)
38. Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. DOI: [10.1093/nar/gkn290](https://doi.org/10.1093/nar/gkn290)
39. Zia, K., Ashraf, S., Jabeen, A., Saeed, M., Nur-e-Alam, M., Ahmed, S., ... & Ul-Haq, Z. (2020). Identification of potential TNF- α inhibitors: from in silico to in vitro studies. *Scientific Reports*, 10(1), 20974. DOI: [10.1038/s41598-020-77723-2](https://doi.org/10.1038/s41598-020-77723-2)

APPENDIX

PYTHON CODE

```
! pip install chembl_webresource_client
import pandas as pd
from chembl_webresource_client.new_client import new_client
```

TARGET SEARCH

```
target = new_client.target
target_query = target.search('TNF-alpha')
targets = pd.DataFrame.from_dict(target_query)
targets
```

SELECTING THE TARGET ID

```
selected_target = targets.target_chembl_id[0]
selected_target
activity = new_client.activity
res = activity.filter(target_chembl_id=selected_target).filter(standard_type="IC50")
df = pd.DataFrame.from_dict(res)
df
```

HANDLING THE MISSING VALUE

```
df2 = df[df.standard_value.notna()]
df2
len(df2.canonical_smiles.unique())
df2_nr = df2.drop_duplicates(['canonical_smiles'])
df2_nr
```

DATA PRE-PROCESSING

```
selection = ['molecule_chembl_id', 'canonical_smiles', 'standard_value']
df3 = df2_nr[selection]
df3
bioactivity_threshold = []
for i in df4.standard_value:
    if float(i) >= 10000:
```

```

    bioactivity_threshold.append("inactive")
elif float(i) <= 1000:
    bioactivity_threshold.append("active")
else:
    bioactivity_threshold.append("intermediate")
bioactivity_class = pd.Series(bioactivity_threshold, name='class')
df5 = pd.concat([df4, bioactivity_class], axis=1)
df5
!# wget https://repo.anaconda.com/miniconda/Miniconda3-py37_4.8.2-Linux-x86_64.sh
! chmod +x Miniconda3-py37_4.8.2-Linux-x86_64.sh
! bash ./Miniconda3-py37_4.8.2-Linux-x86_64.sh -b -f -p /usr/local
! conda install -c rdkit rdkit -y
import sys
sys.path.append('/usr/local/lib/python3.7/site-packages/')
df = pd.read_csv('TNF-alpha_03_bioactivity_data_curated.csv')
df
df_no_smiles = df.drop(columns='canonical_smiles')
smiles = []
for i in df.canonical_smiles.tolist():
    cpd = str(i).split('.')
    cpd_longest = max(cpd, key = len)
    smiles.append(cpd_longest)
smiles = pd.Series(smiles, name = 'canonical_smiles')
df_clean_smiles = pd.concat([df_no_smiles, smiles], axis=1)
df_clean_smiles
LIPINSKI RULE
!pip install rdkit-pypi
import numpy as np
from rdkit import Chem
from rdkit.Chem import Descriptors, Lipinski
def lipinski(smiles, verbose=False):

```

```

moldata= []
for elem in smiles:
    mol=Chem.MolFromSmiles(elem)
    moldata.append(mol)
baseData= np.arange(1,1)
i=0
for mol in moldata:
    desc_MolWt = Descriptors.MolWt(mol)
    desc_MolLogP = Descriptors.MolLogP(mol)
    desc_NumHDonors = Lipinski.NumHDonors(mol)
    desc_NumHAcceptors = Lipinski.NumHAcceptors(mol)

    row = np.array([desc_MolWt,
                    desc_MolLogP,
                    desc_NumHDonors,
                    desc_NumHAcceptors])
    if(i==0):
        baseData=row
    else:
        baseData=np.vstack([baseData, row])
    i=i+1
columnNames=["MW","LogP","NumHDonors","NumHAcceptors"]
descriptors = pd.DataFrame(data=baseData,columns=columnNames)
return descriptors

df_lipinski = lipinski(df.canonical_smiles,df.standard_value)
df_lipinski
df
df_combined = pd.concat([df,df_lipinski], axis=1)
df_combined
import numpy as np

```

```
def pIC50(input):
    pIC50 = []
    for i in input['standard_value_norm']:
        molar = i*(10**-9) # Converts nM to M
        pIC50.append(-np.log10(molar))
    input['pIC50'] = pIC50
    x = input.drop('standard_value_norm', 1)
    return x
```

```
df_combined.standard_value.describe()
```

```
def norm_value(input):
    norm = []
    for i in input['standard_value']:
        if i > 1000000000:
            i = 1000000000
        norm.append(i)
    input['standard_value_norm'] = norm
    x = input.drop('standard_value', 1)
    return x
```

```
df_norm = norm_value(df_combined)
```

```
df_norm
```

```
df_final = pIC50(df_norm)
```

```
df_final
```

DATA VISUALIZATION

```
import seaborn as sns
```

```
sns.set(style='ticks')
```

```
import matplotlib.pyplot as plt
```

```
plt.figure(figsize=(5.5, 5.5))
```

```
sns.countplot(x='class', data=df_2class, edgecolor='black')
```

```
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
```

```
plt.ylabel('Frequency', fontsize=14, fontweight='bold')
```

```
plt.savefig('plot_bioactivity_class.pdf')
```

```

plt.figure(figsize=(5.5, 5.5))
sns.scatterplot(x='MW', y='LogP', data=df_2class, hue='class', size='pIC50',
edgecolor='black', alpha=0.7)
plt.xlabel('MW', fontsize=14, fontweight='bold')
plt.ylabel('LogP', fontsize=14, fontweight='bold')
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0)
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'pIC50', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('pIC50 value', fontsize=14, fontweight='bold')
def mannwhitney(descriptor, verbose=False):
    from numpy.random import seed
    from numpy.random import randn
    from scipy.stats import mannwhitneyu
    seed(1)
    selection = [descriptor, 'class']
    df = df_2class[selection]
    active = df[df['class'] == 'active']
    active = active[descriptor]
    selection = [descriptor, 'class']
    df = df_2class[selection]
    inactive = df[df['class'] == 'inactive']
    inactive = inactive[descriptor]
    stat, p = mannwhitneyu(active, inactive)
    alpha = 0.05
    if p > alpha:
        interpretation = 'Same distribution (fail to reject H0)'
    else:
        interpretation = 'Different distribution (reject H0)'
    results = pd.DataFrame({'Descriptor':descriptor,
                           'Statistics':stat,

```

```

        'p':p,
        'alpha':alpha,
        'Interpretation':interpretation}, index=[0])
filename = 'mannwhitneyu_' + descriptor + '.csv'
results.to_csv(filename)
return results
mannwhitney('pIC50')
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'MW', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('MW', fontsize=14, fontweight='bold')
mannwhitney('LogP')
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'NumHDonors', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('NumHDonors', fontsize=14, fontweight='bold')
mannwhitney('NumHDonors')
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'NumHAcceptors', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('NumHAcceptors', fontsize=14, fontweight='bold')
mannwhitney('NumHAcceptors')
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'MolMR', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('MolMR', fontsize=14, fontweight='bold')
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'class', y = 'NumRotatableBonds', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('NumRotatableBonds', fontsize=14, fontweight='bold')
plt.figure(figsize=(5.5, 5.5))

```

```

sns.boxplot(x = 'class', y = 'TPSA', data = df_2class)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('TPSA', fontsize=14, fontweight='bold')
! pip install padelpy
import glob
xml_files = glob.glob("*.xml")
xml_files.sort()
xml_files
FP_list = ['AtomPairs2DCount',
'AtomPairs2D',
'EState',
'CDKextended',
'CDK',
'CDKgraphonly',
'KlekotaRothCount',
'KlekotaRoth',
'MACCS',
'PubChem',
'SubstructureCount',
'Substructure']
fp = dict(zip(FP_list, xml_files))
y import padelpy
fingerprint = 'Substructure'
fingerprint_output_file = ".join([fingerprint, '.csv']) #Substructure.csv
fingerprint_descriptortypes = fp[fingerprint]
padelpy(mol_dir='molecule.smi',
        d_file=fingerprint_output_file, #'Substructure.csv'
        #descriptortypes='SubstructureFingerprint.xml',
        descriptortypes= fingerprint_descriptortypes,
        detectaromaticity=True,

```

```

        standardizenitro=True,
        standardizetautomers=True,
        threads=2,
        removesalt=True,
        log=True,
        fingerprints=True)
descriptors = pd.read_csv(fingerprint_output_file)
descriptors

```

MACHINE LEARNING- RANDOM FOREST REGRESSOR

```

import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
X = descriptors.drop(columns = ["Name"])
Y = df3.pIC50
dataset3 = pd.concat([X,Y], axis=1)
dataset3
from sklearn.feature_selection import VarianceThreshold
selection = VarianceThreshold(threshold=(.8 * (1 - .8)))
X = selection.fit_transform(X)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
X_train.shape, Y_train.shape
X_test.shape, Y_test.shape
df = pd.DataFrame(X_train,Y_train)
df = pd.DataFrame(X_test,Y_test)
import numpy as np
from sklearn.ensemble import RandomForestRegressor
# Filter the samples based on Y_train and Y_test without changing the number of samples
max_value = np.finfo('float128').max
Y_train_filtered = Y_train[Y_train <= max_value]
Y_test_filtered = Y_test[Y_test <= max_value]

```

```

# Get the indices where the condition holds true
indices_train = np.where(Y_train <= max_value)[0]
indices_test = np.where(Y_test <= max_value)[0]
# Filter X_train and X_test using the obtained indices
X_train_filtered = X_train[indices_train]
X_test_filtered = X_test[indices_test]
# Create and fit the model using the filtered data
import numpy as np
np.random.seed(100)
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train_filtered, Y_train_filtered, sample_weight=np.ones(len(X_train_filtered)))
# Evaluate the model
r2 = model.score(X_test_filtered, Y_test_filtered)
Y_pred = model.predict(X_test_filtered)
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(color_codes=True)
sns.set_style("white")
ax = sns.regplot(x=Y_test,y=Y_pred, scatter_kws={'alpha':0.4})
ax.set_xlabel('Experimental pIC50', fontsize='large', fontweight='bold')
ax.set_ylabel('Predicted pIC50', fontsize='large', fontweight='bold')
ax.set_xlim(0, 12)
ax.set_ylim(0, 12)
ax.figure.set_size_inches(5, 5)
plt.show
from padelpy import padeldescriptor
fingerprint = 'SubstructureCount'
fingerprint_output_file = ".join([fingerprint, '.csv']) #Substructure.csv
fingerprint_descriptortypes = fp[fingerprint]
padeldescriptor(mol_dir='molecule.smi',
                 d_file=fingerprint_output_file, #'Substructure.csv'

```

```
#descriptortypes='SubstructureFingerprint.xml',
descriptortypes= fingerprint_descriptortypes,
detectaromaticity=True,
standardizenitro=True,
standardizetautomers=True,
threads=0,
removesalt=True,
log=True,
fingerprints=True)
```

APPLICABILITY DOMAIN

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df_external = pd.read_csv("Standardised_residuals.csv")
df_internal = pd.read_csv("Standardised_residuals_test1.csv")
df_external['Data_Source'] = 'External'
df_internal['Data_Source'] = 'Internal'
combined_df = pd.concat([df_external, df_internal])
plt.figure(figsize=(10, 5))
sns.scatterplot(x='Leverage _value', y='standardised_value', hue='Data_Source',
data=combined_df, palette=['red', 'blue'], alpha=0.7)
plt.title("Williams Plot")
plt.xlabel('Leverage Values')
plt.ylabel('Standardized Residuals')
plt.axhline(y=3, color='red', linestyle='-', linewidth=1)
plt.axhline(y=-3, color='red', linestyle='-', linewidth=1)
plt.axvline(x=0.2, color='blue', linestyle='--', linewidth=1, label='Leverage = 0.2')
plt.grid(True)
plt.legend()
plt.ylim(-6, 6)
plt.show()
```

```
plt.savefig('AD_final.png')
```

GINI INDEX

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_csv("dddd.csv")
X = df.drop(columns=["pIC50", "pIC50.1"]) # Assuming "pIC50" and "pIC50.1" are the target
variables
y = df["pIC50"]
X.replace([np.inf, -np.inf], 1e9, inplace=True)
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X, y)
feature_importances = rf_model.feature_importances_
importance_df = pd.DataFrame({"Feature": X.columns, "Importance":
feature_importances})
importance_df = importance_df.sort_values(by="Importance", ascending=False)
top_15_importance_df = importance_df.head(15)
print(top_15_importance_df)
plt.figure(figsize=(10, 6))
plt.barh(top_15_importance_df["Feature"], top_15_importance_df["Importance"])
plt.xlabel("Gini Index")
plt.ylabel("SubFP")
plt.title("Top 15 Features by Gini Index")
plt.show()
plt.savefig('gini_index_plot.png')
```

MOLECULAR DOCKING SELECTION

```
import numpy as np
import pandas as pd
data_df = pd.read_csv('TNF-alpha_08_bioactivity_data_3class_pIC50_MACCS_fp.csv')
```

```

filtered_data_df = data_df[data_df['pIC50'] < 6] # Assuming pIC50 values are in negative
logarithm
selected_compounds_df = filtered_data_df.sample(n=30, random_state=42) # Use a fixed
random state for reproducibility
selected_compounds_features = selected_compounds_df.drop(columns=['pIC50']).values
selected_pIC50_values = selected_compounds_df['pIC50'].values
print(selected_compounds_df)
import matplotlib.pyplot as plt
plt.scatter(data_df[data_df['class'] == 'active']['LogP'], data_df[data_df['class'] ==
'active']['pIC50'], color='gray', label='Active Compounds')
plt.scatter(data_df[data_df['class'] != 'active']['LogP'], data_df[data_df['class'] !=
'active']['pIC50'], color='lightgreen', label='Other Compounds')
plt.scatter(selected_compounds_df['LogP'], selected_compounds_df['pIC50'], color='red',
label='Selected Compounds')
plt.xlabel('LogP')
plt.ylabel('pIC50')
plt.title('Selected Compounds as Diversity Set')
plt.legend()
plt.show()

```

R CODE FOR VALIDATION

```

mydata<-read.csv ("Substructure.csv")
install.packages("caret")
library(caret)
install.packages("caret",dependencies=c("Depends","Suggests"))
mycontrol = trainControl(method = "cv", number = 10)
model1 = train (Total~p IC50,data = mydata, method = 'lm', trControl=myControl)
model1

```