Cheminformatics-Driven Exploration of Chemical Space for InhA Inhibitors in *Mycobacterium tuberculosis* through Virtual Screening and Molecular Modeling

Thesis submitted to the University of Calicut
in partial fulfillment of the requirements for
the award of the degree of

DOCTOR OF PHILOSOPHY IN CHEMISTRY

by

JALALA V.K.

DEPARTMENT OF CHEMISTRY

UNIVERSITY OF CALICUT

KERALA-673635

MARCH 2026

CERTIFICATE

This is to certify that the thesis entitled "**Cheminformatics-Driven Exploration of Chemical Space for InhA Inhibitors in *Mycobacterium tuberculosis* through Virtual Screening and Molecular Modeling**", submitted to the University of Calicut in partial fulfilment of the requirements for the award of **Degree of Doctor of Philosophy in Chemistry,** is a bonafide record of the research work carried out by **Ms. Jalala V.K**. during the period 2019-2025 in the Department of Chemistry, University of Calicut, under my supervision and guidance. The thesis has not formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, or other similar title to any candidate of any University. The contents of this thesis have been checked for plagiarism using the software 'iThenticate' and the similarity index falls within the permissible limit. The corrections recommended by the adjudicators have been incorporated into the thesis, and the contents of the thesis and the soft copy are the same.

University of Calicut
Date : 03/06/25

DEPARTMENT OF CHEMISTRY UNIVERSITY OF CALICUT

Dr. K. Muraleedharan

Professor (Retired)

Department of Chemistry

University of Calicut

Kerala

DECLARATION

I hereby declare that the research work embodied in the thesis entitled **"Cheminformatics-Driven Exploration of Chemical Space for InhA Inhibitors in *Mycobacterium tuberculosis* through Virtual Screening and Molecular Modeling"**, submitted to the University of Calicut, is a bonafide record of research work carried out by me during the period 2019–2025 under the supervision and guidance of Dr. K. Muraleedharan, Professor (Retd.), Department of Chemistry, University of Calicut. This work has not been submitted elsewhere, either in part or full, for the award of any degree or diploma. I also declare that I am solely responsible for the authenticity and originality of the findings and observations presented in this thesis and that the thesis is free from Artificial intelligence-generated content.

University of Calicut
Date: 03/06/25

Jalala V.K.

Acknowledgments

With deep humility and profound gratitude, I take this opportunity to express my heartfelt thanks to everyone who supported and guided me throughout my PhD journey. This academic milestone has been a transformative and rewarding experience, made possible by the encouragement, help, and prayers of many wonderful individuals.

First and foremost, I express my sincere and deepest gratitude to my research supervisor, Prof. (Dr.) K. Muraleedharan, Professor (Retd.), Department of Chemistry, University of Calicut, for his invaluable guidance, patience, and continuous support throughout my research work. His profound knowledge, encouragement, and mentorship have been the cornerstone of my academic growth and the successful completion of this thesis.

I extend my special thanks to Prof. (Dr) Binitha N. N., Head of the Department of Chemistry, University of Calicut, for providing a supportive academic environment. I also wish to acknowledge the former Heads of the Department, Prof. (Dr.) Rajeev S. Menon, Prof. (Dr.) P. Raveendran, and Dr. A. I. Yahya for their contributions to maintaining the academic excellence of the department and facilitating my research work.

I would also like to extend my heartfelt thanks to the faculty members of the Department: Prof. (Dr.) Abraham Joseph, Prof. (Dr.) N. K. Renuka, Prof. (Dr.) M. T. Ramesan, Dr. Susmita De, Dr. Roymon Joseph, Dr. Fazalurahman K., Dr. Suja T. D., Dr. Derry Holaday M. G., Dr. P. Pradeepan (former Associate Professor), and Dr. Suresh Babu (former Assistant Professor, UGC FRP), for their invaluable

support in creating a stimulating and competitive academic environment within the Department.

I also extend my sincere thanks to Prof. V. M. Abdul Mujeeb, Professor (Retd.), for his kind help and encouragement that enabled me to join the PhD programme. I am especially thankful to Dr. U.C.A. Jaleel, Principal Scientist, OSPF-NIAS Drug Discovery Lab, NIAS IISc Campus, Bangalore, and his team members, for their generous support and guidance in introducing me to machine learning and artificial intelligence. Their insights and training greatly enriched my research and helped broaden my interdisciplinary skills.

I want to express my gratitude to the technical and administrative staff of the Department for their invaluable assistance throughout this journey, including Mr. Satheesan K., Mr. Haridasan K. K., Mr. Jinson Antony, Ms. Shaini E., Ms. Beena K. Raghavan, Ms. Rajani E. P., Ms. Sangeetha T. G., Mr. Premarajan, and others.

I am also thankful to the University of Calicut for providing the necessary infrastructure and to the University Fellowship scheme for financially supporting my research.

I am also deeply indebted to all past and present research scholars from the Computational Chemistry group, including Dr. Shameera Ahamed, Dr. Vijisha K. Rajan, Dr. Sabira K., Dr. Sumayya P.C., Dr. Vinduja P., Ms. Jaseela M.A., Ms. Neenu Krishna P.U., Ms. Ragi C., Ms. Gopika Krishnan G.S., and Ms. Swathi Krishna. My sincere thanks also go to the research scholars from other groups, particularly Dr. Radhika Narayanan Nair, Dr. Lijin Rajan, Dr.

Nidhisha V, Mr. Sivakrishna Prakash, and., for their invaluable support.

I extend my deepest gratitude to my family, whose support has been the foundation of this journey. My heartfelt thanks to my beloved parents, Mr. Abdul Salam and Mrs. Haseena V. K., for their unconditional love, lifelong sacrifices, and endless prayers. I am grateful to my sister, Dr. Jaseela V. K., and my brother, Jaseel V. K., for their constant encouragement. I would also like to thank my husband, Shameem Muhammad, for his patience, support, and understanding throughout this journey, and my dear children, Yahyaa Shameem and Dihyah Shameem, for filling my life with love and purpose. Special thanks to my father-in-law, Dr. T. A. Mohammad, my mother-in-law, Mrs. Sakeena T. P., and Ummama, Mrs. Jameela K. V., for their constant support, kindness, and encouragement. My sincere thanks go to all my in-laws, nieces, and nephews for their warmth, affection, and encouragement throughout this journey. I am especially grateful for their kind support, prayers, and for bearing with me through all the hardships during the course of my PhD.

Above all, I offer my sincere gratitude to Almighty Allah for His infinite mercy and blessings that guided me through every challenge and enabled me to complete this work. All praise and thanks are due to Him alone.

Jalala V.K.

To my parents,

for their endless love and support

ABSTRACT

The research work presented in this thesis comprises a computational investigation into the identification and optimization of novel inhibitors targeting the enoyl-acyl carrier protein reductase (InhA) enzyme of *Mycobacterium tuberculosis*, a key enzyme in mycolic acid biosynthesis and a validated anti-tubercular drug target. The study integrates cheminformatics, QSAR modeling, and structure-based drug design methodologies to explore the chemical space of known InhA inhibitors and to identify new bioactive scaffolds. A comprehensive activity landscape analysis was conducted to gain insights into structure–activity relationships (SAR) and detect activity cliffs among known inhibitors. A predictive QSAR model was developed using molecular descriptors and machine learning algorithms to classify compounds as active or inactive. This model was applied to screen selected phytochemicals derived from traditional Indian medicinal plants, and active compounds were subjected to molecular docking for further screening and to evaluate their binding potential. Structure-based virtual screening of the ZINC database was performed using molecular docking, followed by MD simulations, MM-PBSA free energy calculations, and DFT analysis to evaluate binding affinity, stability, and electronic properties. ADMET profiling was also carried out to assess drug-likeness. The findings highlight several potential lead compounds that exhibit strong inhibitory potential against InhA and warrant further experimental validation.

Keywords: InhA inhibitors; Chemical space analysis; Activity landscape modeling; QSAR modeling; Structure-based virtual screening.

സംഗ്രഹം

ഈ പ്രബന്ധത്തിൽ അവതരിപ്പിച്ചിരിക്കുന്ന ഗവേഷണ പ്രവർത്തനങ്ങൾ മൈക്കോബാക്റ്റീരിയം ട്യൂബർക്കുലോസിസ് എന്ന രോഗാണുവിൽ മൈകോലിക് ആസിഡ് ജൈവസംശ്ലേഷണത്തിൽ നിർണ്ണായകമായ പങ്ക് വഹിക്കുന്ന ഒരു എൻസൈം ആയ ഇനോയിൽ-ആസിൽ കേരിയർ പ്രൊട്ടീൻ റിഡക്ടേസ് (ഐ.എൻ.എച്ച് .എ) നെ ലക്ഷ്യമാക്കി അതിനെ തടയാനുള്ള പുതിയ സംയുക്തങ്ങളെ കണ്ടെത്തുകയും അവരുടെ ഫലപ്രാപ്തി മെച്ചപ്പെടുത്തുകയും ചെയ്യുന്ന കംപ്യൂട്ടർ സഹായിയുള്ള ഒരു ശാസ്ത്രീയ പരിശ്രമമാണ്. ഈ പഠനം രാസവിവരശാസ്ത്രം, ക്യുഎസ്എആർ മാതൃക നിർമ്മാണം, ഘടന അടിസ്ഥാനമാക്കിയുള്ള ഔഷധ രൂപകല്പന എന്നീ കംപ്യൂട്ടർ ആശ്രിത രീതികളെ ഒരുമിച്ച് ഉപയോഗിച്ച്, ഐ.എൻ.എച്ച് .എ നെ തടയുന്ന നിലവിലുള്ള രാസസംയുക്തങ്ങളുടെ രാസ ഇടം വിശകലനം ചെയ്യുകയും, പുതിയ ജൈവ സജീവ ഘടനകൾ കണ്ടെത്തുകയും ചെയ്യുന്നു. പ്രവർത്തനഭൂപ്രകൃതിയുടെ വിശകലനം (ആക്ടിവിറ്റി ലാൻഡ്സ്കേപ്പ് അനാലിസിസ്) വഴി ഘടനയും പ്രവർത്തനഫലവും തമ്മിലുള്ള ബന്ധം വിശദമായി പരിശോധിക്കുകയും, ചെറിയ ഘടനാത്മക മാറ്റങ്ങൾ മൂലം ഫലത്തിൽ ഉണ്ടാകുന്ന വൻ വ്യത്യാസങ്ങൾ (ആക്ടിവിറ്റി ക്ലിഫുകൾ) തിരിച്ചറിയുകയും ചെയ്തു. രാസവിവരണം പ്രതിനിധീകരിക്കുന്ന അളവുകൾ ഉപയോഗിച്ചും യന്ത്രശിക്ഷണ ആൽഗോരിതങ്ങൾ വഴി ക്യുഎസ്എആർ തരംതിരിക്കൽ മാതൃക വികസിപ്പിക്കുകയും അതിന്റെ സഹായത്തോടെ സംയുക്തങ്ങൾ സജീവം അല്ലെങ്കിൽ അസജീവം എന്ന് തിരിച്ചറിയുകയും ചെയ്തു. ഇന്ത്യൻ പരമ്പരാഗത ഔഷധ സസ്യങ്ങളിൽ നിന്നെടുത്ത സസ്യ രാസസംയുക്തങ്ങൾ (ഫൈറ്റോകെമിക്കലുകൾ) ഈ മാതൃക ഉപയോഗിച്ച് തിരഞ്ഞെടുത്തു, അതിനുശേഷം അവയുടെ ബന്ധശേഷി വിലയിരുത്തുന്നതിനായി മോളിക്യുലാർ ഡോക്കിംഗ് നടത്തപ്പെട്ടു. അടിസ്ഥാന ഘടനയെ അടിസ്ഥാനമാക്കിയുള്ള വിർച്വൽ സ്ക്രീനിംഗ് എന്ന പദ്ധതിയിലൂടെ സിങ്ക് ഡാറ്റാബേസിൽ നിന്നുള്ള നിരവധി സംയുക്തങ്ങൾ തിരഞ്ഞെടുക്കപ്പെട്ടു. ഇതിന് പിന്നാലെ മോളിക്യുലാർ ഡൈനാമിക്സ് അനുസരണങ്ങൾ, ബന്ധസ്വതന്ത്ര ഊർജശാസ്ത്രം (എം.എം-പി.ബി.എസ്.എ), ഘനത പ്രവർത്തന സിദ്ധാന്തം (ഡി.എഫ്.ടി) അടിസ്ഥാനമാക്കിയുള്ള ഇലക്ട്രോണിക് വിശകലനങ്ങൾ എന്നിവ നടത്തിയിരുന്നു. സംയുക്തങ്ങളുടെ ഔഷധയോഗ്യത വിലയിരുത്തുന്നതിനായി ആബ്സോർപ്ഷൻ, ഡിസ്ട്രിബ്യൂഷൻ, മെറ്റബോളിസം, എക്സ്ക്രിഷൻ, ടോക്സിസിറ്റി (എ.ഡി.എം.ഇ.ടി.) വിവരങ്ങൾ പരിശോധിച്ചു. ഈ പഠനത്തിലൂടെ ഐ.എൻ.എച്ച് .എ എൻസൈമിനെ ശക്തമായി തടയാൻ കഴിവുള്ള പ്രധാന മുൻഗാമി സംയുക്തങ്ങൾ തിരിച്ചറിയുകയും, അവയുടെ പ്രായോഗിക പരിശോധനയ്ക്ക് സാധ്യതയുള്ളതാണെന്ന് വ്യക്തമാക്കുകയും ചെയ്യുന്നു.

സൂചനപദങ്ങൾ: ഐ.എൻ.എച്ച്.എ തടയുന്ന സംയുക്തങ്ങൾ; രാസ സ്പേസ് വിശകലനം; പ്രവർത്തന വിന്യാസ മോഡലിംഗ്; ക്യുഎസ്എആർ മാതൃക; ഘടനാ അടിസ്ഥാനമായ സ്ക്രീനിംഗ്.

PREFACE

The convergence of chemistry, biology, and medicine has profoundly impacted modern drug discovery. In recent decades, the integration of computational chemistry with cheminformatics and computer-aided drug design (CADD) has enabled the rapid and systematic identification of drug candidates through data-driven approaches. With the increasing availability of structural and biological data for thousands of target proteins in public databases, discovering new structure–activity relationships (SAR) and predicting biological activity using in silico methods has become a fundamental aspect of rational drug design.

In this context, tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains a global health challenge due to the rise of multidrug-resistant strains. One of the validated drug targets in TB is the enoyl-acyl carrier protein reductase (InhA), an essential enzyme involved in the biosynthesis of mycolic acids, which are critical for bacterial cell wall integrity. Given the availability of structural and activity data for InhA inhibitors, cheminformatics approaches offer great promise for exploring chemical space, understanding SAR, and identifying new inhibitors through ligand- and structure-based screening techniques.

This thesis, titled *"**Cheminformatics-Driven Exploration of Chemical Space for InhA Inhibitors in Mycobacterium tuberculosis through Virtual Screening and Molecular Modeling**"*, is organized into seven chapters. ***Chapter 1*** provides a brief introduction to chemical space as a source for new drugs, various computational methods for

rational drug design, and a detailed description of the target enoyl-acyl carrier protein reductase (InhA) and its therapeutic relevance**.** Besides, this chapter presents the motivation and objectives of the current research problem. ***Chapter 2*** describes the software tools, datasets, and computational methodologies used in each phase of the research.

The core research is presented across Chapters 3 to 6. ***Chapter 3*** presents a detailed cheminformatic analysis of Mycobacterium tuberculosis inhibitors, focusing on the exploration of chemical space and structural diversity. The study highlights scaffold diversity, molecular property distribution, and key structural features of known InhA inhibitors to gain insights into their chemical behavior and design potential**.** ***Chapter 4*** presents an activity landscape modeling study of InhA inhibitors, aimed at characterizing structure–activity relationships through the identification of activity cliffs. The analysis highlights subtle structural changes responsible for significant potency variations, offering insights into key molecular features that influence bioactivity. ***Chapter 5*** highlights the development of a machine learning-based QSAR classification model for predicting InhA inhibitory activity. The model was built using molecular descriptors and validated algorithms, and was later employed to identify potential InhA inhibitors from a library of phytochemicals derived from traditional Indian medicinal plants. ***Chapter 6*** describes the structure-based virtual screening of the ZINC database through combined molecular docking, molecular dynamics (MD) simulations, MM-PBSA binding free energy calculations, DFT-based electronic property analysis, and ADMET profiling. These integrated computational approaches were employed to evaluate the binding stability, electronic features, and pharmacokinetic behavior of selected InhA inhibitors**.**

Finally, ***Chapter 7*** summarizes the key findings of the study and outlines potential directions for future research, including experimental validation and scaffold optimization. This thesis demonstrates the strength of computational approaches in accelerating early-stage drug discovery, particularly for infectious diseases such as tuberculosis, and presents promising candidates for further exploration as InhA-targeted anti-tubercular agents.

CONTENTS

List of Tables

List of Figures

List of Abbreviations

TB	: Tuberculosis
WHO	: World Health Organization
MTBC	: Mycobacterium tuberculosis Complex
Mtb	: Mycobacterium tuberculosis
INH	: Isoniazid
RIF	: Rifampin
PZA	: Pyrazinamide
EMB	: Ethambutol
SM	: Streptomycin
OFX	: Ofloxacin
LEV	: Levofloxacin
MOX	: Moxifloxacin
CIP	: Ciprofloxacin
KAN	: Kanamycin
AMK	: Amikacin
CAP	: Capreomycin
ETH	: Ethionamide
PTH	: Prothionamide
CS	: Cycloserine
PAS	: p-Aminosalicylic Acid
InhA	: Enoyl-acyl carrier protein reductase
DprE1/DprE2	: Decaprenylphosphoryl-β-D-ribose epimerase
EmbCAB	: Ethambutol target enzyme complex
MmpL3	: Mycobacterial membrane protein Large 3
DdlA	: D-alanine–D-alanine ligase A
LdtM1/Ldts	: L,D-transpeptidases
MDR	: Multidrug-Resistant
XDR	: Extensively Drug-Resistant
RR	: Rifampicin Resistance
CADD	: Computer-Aided Drug Design
SBDD	: Structure-Based Drug Design
LBDD	: Ligand-Based Drug Design
QSAR	: Quantitative Structure–Activity Relationship
SBVS	: Structure-Based Virtual Screening
LBVS	: Ligand-Based Virtual Screening
MD	: Molecular Dynamics
AI	: Artificial Intelligence
ML	: Machine Learning

DL	: Deep Learning
ADMET	: Absorption, Distribution, Metabolism, Excretion, and Toxicity
SMILES	: Simplified Molecular Input Line Entry System
InChI	: International Chemical Identifier
TPSA	: Topological Polar Surface Area
Rg	: Radius of Gyration
SASA	: Solvent Accessible Surface Area
dPSA	: Dynamic Polar Surface Area
PDB	: Protein Data Bank
NCBI	: National Center for Biotechnology Information
IMPPAT	: Indian Medicinal Plants, Phytochemistry And Therapeutics
MM	: Molecular Mechanics
QM	: Quantum Mechanics
UFF	: Universal Force Field
CFF/CVFF	: Consistent Force Field / Consistent Valence Force Field
ECEPP	: Empirical Conformational Energy Program for Peptides
BMS	: Biopolymer Molecular Simulation
WHIM	Weighted Holistic Invariant Molecular descriptors
GETAWAY	: Geometry, Topology, and Atom-Weights AssemblY descriptors
CoMFA	: Comparative Molecular Field Analysis
GRID	: Grid-Independent Descriptors
ISIDA	: In Silico Design and Data Analysis descriptors
SMIFP	: SMILES Fingerprint
MFP	: Mini Fingerprint
BCI	: Barnard Chemistry Information fingerprints
MACCS	: Molecular ACCess System keys
Molprint2D/3D	: Molecular printing (2D and 3D variants)
ECFP	: Extended-Connectivity Fingerprints
FCFP	: Functional-Class Fingerprints
PP	: Pharmacophore Fingerprint
PCA	: Principal Component Analysis
SVD	: Singular Value Decomposition
SOM	: Self-Organizing Map
SAS	: Structure–Activity Similarity
AC	: Activity Cliff
T	: Tanimoto Coefficient
HTS	: High-Throughput Screening

SVM	: Support Vector Machine
RMSD	: Root Mean Square Deviation
RMSF	: Root Mean Square Fluctuation
PUMA	: Platform for Unified Molecular Analysis
RSF	: Rubberbanding Scaling Forcefield
MW	: Molecular Weight
RTB	: Rotatable Bonds
HBA	: Hydrogen Bond Acceptors
HBD	: Hydrogen Bond Donors
ALogP	: Partition Coefficient (octanol/water)
CDF	: Cumulative Distribution Function
FDA	: U.S. Food and Drug Administration
PCP	: Pharmaceutically relevant Physicochemical Properties
SALI	: Structure–Activity Landscape Index
SAS Map	: Structure–Activity Similarity Map
OPLS/AA	Optimized Potentials for Liquid Simulations / All Atom
AMBERTOOLS	: Assisted Model Building with Energy Refinement Tools
ACPYPE	: AnteChamber PYthon Parser Interface
SPC	: Simple Point Charge
RO5	: Rule of Five
SMOTE	: Synthetic Minority Over-sampling Technique
TPU	: Tensor Processing Unit
GPU	: Graphics Processing Unit
IMPs	: Indian Medicinal Plants
LGBM	: Light Gradient Boosting Machine
MIC	: Minimum Inhibitory Concentration
TPR	: True Positive Rate
TNR	: True Negative Rate
PubChemFP	: PubChem Fingerprint Descriptors
ZID	: Isonicotinic-acetyl-nicotinamide-adenine dinucleotide
MGL Tools	: Molecular Graphics Laboratory Tools
pdbqt	: PDB with Partial Charges and Torsions
ESP	: Electrostatic Surface Potential
HOMO	: Highest Occupied Molecular Orbital
LUMO	: Lowest Unoccupied Molecular Orbital
IP	: Ionization Potential
EA	: Electron Affinity
DR-TB	Drug-Resistant Tuberculosis

Chapter 1

Introduction and Review of Literature

1.1. Introduction

There is a dread disease which so prepares its victim, as it were, for death; which so refines it of its grosser aspect, and throws around familiar looks unearthly indications of the coming change; a dread disease, in which the struggle between soul and body is so gradual, quiet, and solemn, and the result so sure, that day by day, and grain by grain, the mortal part wastes, and withers away, so that the spirit grows light and sanguine with its lightening load, and, feeling immortality at hand, deems it but a new term of mortal life (Nicholas Nickleby, 1838, chapter 49). To date, Charles Dickens' expressions remain relevant [1]. Even in the twenty-first century, tuberculosis (TB) still affects around one-third of the world's population, making it one of the major infectious diseases [2].

According to the latest WHO reports, 8.2 million new cases and 1.25 million deaths are expected in 2023, and in 2024 around 8.3 million new TB cases were reported with approximately 1.23 million deaths globally [3]. TB is caused by a group of closely related bacteria known as the Mycobacterium tuberculosis complex (MTBC) [4]. The MTBC is a genetically related group of Mycobacteria that includes Mycobacterium tuberculosis (M. tuberculosis), Mycobacterium africanum (M. africanum), Mycobacterium bovis (M. bovis), Mycobacterium canettii (M. canettii), Mycobacterium microti (M. microti), Mycobacterium pinnipedii (M. pinnipedii) and Mycobacterium caprae (M. caparae) [5]. From those species, *M. tuberculosis* is the most significant member and primary pathogen causing TB in humans [6].

1.1.1. Mycobacterium tuberculosis

Mycobacterium tuberculosis (Mtb) also known as Koch's bacillus, plays a vital role in the history of microbiology. In 1882, the German scientist Robert Koch was the first microbiologist to declare the effective isolation of the causal agent of TB, which was later termed *Mycobacterium tuberculosis* [7]. Mtb primarily affects the lungs, but it can also affect other organs and tissues including the lymph nodes, brain, kidneys, and spine [8]. To tackle this successful human pathogen, we must have a deeper understanding of the fundamental biology of mycobacterial pathogenesis [9]. The complexity of the cell envelope, which is abundant in lipids and polysaccharides with unique chemical structures, is one of the specific features that allow mycobacteria to survive in the human body. As a result, the Mtb cell envelope is a promising target for vaccine and therapeutic development. Targeting mycobacterial cell wall is a successful and effective strategy for anti-TB drugs [10,11].

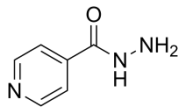
1.1.2. Important Drugs and drug targets

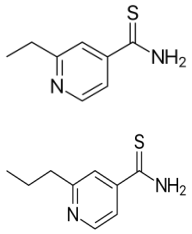
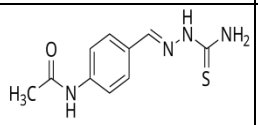
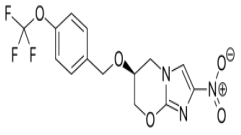
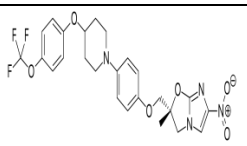
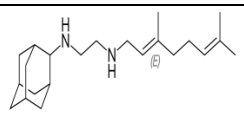
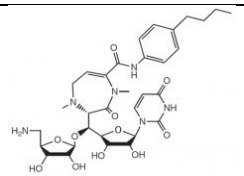
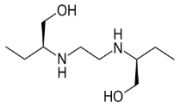
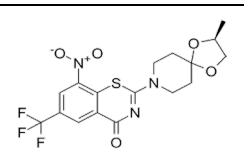
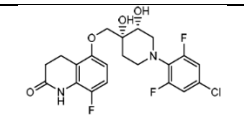
TB drugs are classified into first-line drugs and second-line drugs based on their function, use, and mode of action. The first-line therapeutic agents are the most efficient and least harmful when used to treat tuberculosis, but the second-line agents are less efficient, more costly, and more toxic [12]. In contrast, they are necessary for the treatment of drug-resistant bacterial strains. First-line drugs include isoniazid (INH), rifampin (RIF), pyrazinamide (PZA), ethambutol (EMB), and streptomycin (SM), while second-line drugs are classified

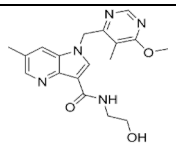
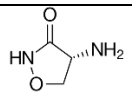
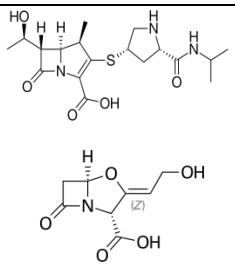
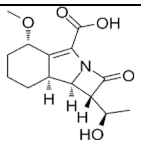
into three groups (i). Fluoroquinolones: Ofloxacin (OFX), Levofloxacin (LEV), Moxifloxacin (MOX), Ciprofloxacin (CIP) (ii) Injectable antituberculosis drugs- Kanamycin (KAN), amikacin (AMK) and capreomycin (CAP) (iii) Less-effective second-line antituberculosis drugs- Ethionamide (ETH)/Prothionamide (PTH), Cycloserine (CS)/Terizidone, P-aminosalicylic acid (PAS) [10,11].

The currently available drugs, including first- and second-line agents, mainly inhibit enzymes involved in the cell wall synthesis of Mtb [13]. The cell wall comprises three major components: the peptidoglycan layer, mycolic acid layer, and arabinogalactan polysaccharide [14]. A multitude of research exists that emphasizes the relevance of enzymes involved in the biosynthesis of the mycobacterial cell wall, responsible for its survival, proliferation, permeability, virulence, and resistance to drugs. As a result, the fundamental nature of cell wall synthesis and component assembly has made the mycobacterial cell wall the most commonly used target of anti-TB drugs [11,15,16]. Several inhibitors targeting enzymes involved in Mtb cell wall biosynthesis are summarized in **Table 1.1.**

Table 1.1. The lists of inhibitors and their structures that target cell wall biosynthesis

Name	**Structure**	**Cell wall component inhibited**	**Therapeutic Target**
INH		Mycolic acid	InhA

ETH / Prothionamide		Mycolic acid	InhA
Thiacetazone		Mycolic acid	HadAB
Pretomanid		Keto-mycolic acid	DprE2
Delamanid		Methox and Keto-mycolic acid	DprE2
SQ109		Mycolic acid transport	MmpL3
CPZEN-45		Arabinogalactan	WecA
Ethambutol		Arabinogalactan LAM	EmbCAB
BTZ043		Arabinogalactan LAM	DprE1
OPC167832		Arabinogalactan LAM	DprE1

TBA-7371		Arabinogalactan LAM	DprE1
Cycloserine		peptidoglycan	DdlA
Meropenem/ Clavulanate		peptidoglycan	LdtM1
Sanfetrinem		peptidoglycan	Ldts

1.1.3. Enoyl-acyl carrier protein reductase (InhA)

Enoyl acyl carrier protein reductase (InhA) is a major enzyme involved in fatty acid synthesis, particularly in mycolic acid biosynthesis. It is a member of the NADH-dependent acyl carrier protein reductase family [17,18]. Mycolic acids are a-branched, b-hydroxylated fatty acids with chains of 60-90 carbon atoms. They are crucial components of the mycobacterial cell wall. [19]. Mycolic acid biosynthesis is catalyzed by two enzyme systems: fatty acid synthase I (FAS I), which produces shorter chain fatty acids, and fatty acid synthase II (FAS II), which involves fatty acid chain elongation. In both systems, each elongation cycle consists of four successive steps: (1) Condensation of acyl and malonyl groups, (2) β-ketoacyl-ACP

reduction, (3) β-hydroxyacyl-ACP dehydration, and (4) trans-2-enoyl-ACP reduction, resulting in an extended acyl chain with two more carbon units [20,21]. InhA catalyzes the final step of the fatty acid elongation cycle, namely the NADH-dependent reduction of trans-2-enoyl-ACP [22]. The final step of mycolic acid biosynthesis catalyzed by InhA is illustrated in **Fig. 1.1.**

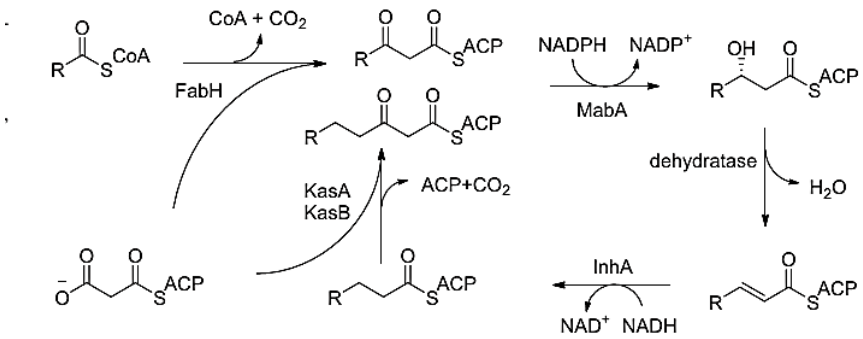


Fig. 1.1. The final step in the mycolic acid biosynthesis pathways in mycobacterium cell wall

InhA is one of the few clinically proven, attractive, and extensively studied targets for developing new TB treatments. InhA can be inhibited by both indirect and direct anti-TB drugs [23]. Indirect inhibitors include drugs such as INH, prothionamide, and ethionicotinamide, whereas direct inhibitors include triclosan/diphenyl ether, pyrrolidine formamide, pyrrole, pyrrolidine carboxamide, acetamide, thiadiazole, and triazole. Among these inhibitors, INH is one of the most effective and extensively used frontline antitubercular drugs, inhibiting TB [24,25]. INH is a prodrug that must be activated by KatG, the mycobacterial catalase-peroxidase, to generate an INH-NAD adduct **Fig. 1.2**, which inhibits the InhA enzyme of Mtb [26].

Nevertheless, the majority of INH-resistant clinical strains are identified largely due to the introduction of KatG mutants which do not generate an INH-NAD adduct [27]. Therefore, inhibitors that directly target InhA without requiring this activation might be good options for developing new anti-tubercular drugs. Consequently, researchers are working to discover InhA direct inhibitors that may aid in the global eradication of TB [28–31].

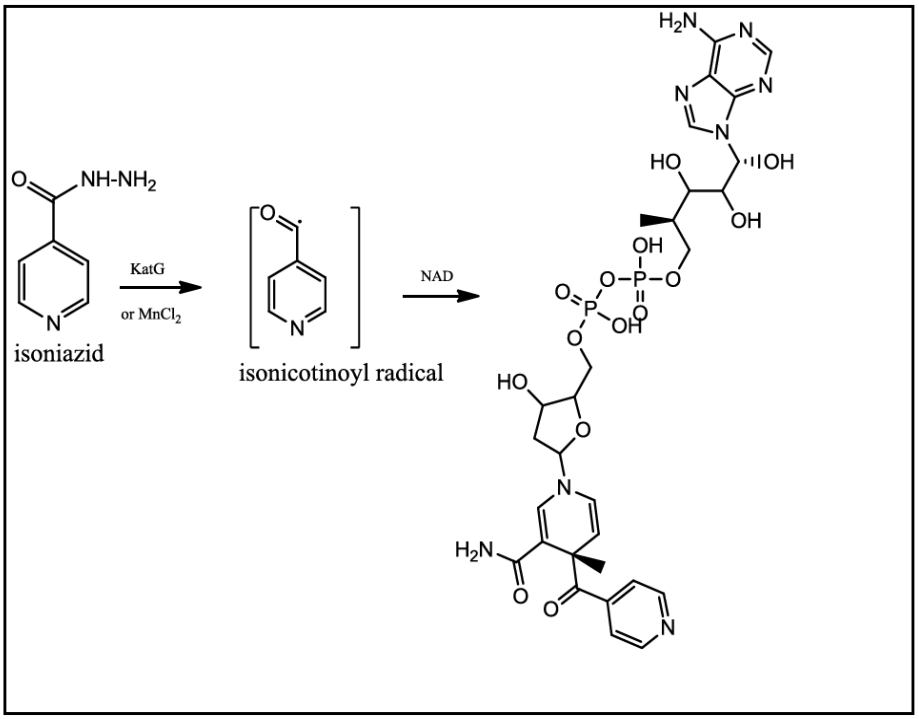


Fig. 1.2. Chemical structure of isoniazid and mechanism of formation of the INH-NADH adduct

1.2. Current challenges in TB drug discovery

Despite the availability of TB drugs, there are still many reasons why drug design and development are difficult. Therefore, developing

new, shorter, and more potent drugs to treat TB is essential in the field of drug discovery. The current challenges of developing a tuberculosis free world are the following.

1.2.1. Biological mechanism

One of the most challenging issues faced during the development of new drugs is Mtb's ability to survive in the microenvironment of the human host. Disrupting the host's innate and adaptive immunity, as well as its inherent persistence, complicates disease control. Our understanding of Mtb pathogenesis has advanced significantly during the last few decades. However, the whole complex biological mechanism of survival is still unclear [32,33].

1.2.2. Cell wall thickness

The Mtb cell wall is complicated and unique, consisting of a thick peptidoglycan layer and an outer membrane constituted mainly of lipopolysaccharides and fatty acids, with glycolipids and wax esters embedded. This lipid-rich cell wall is intrinsically resistant to many potential drugs, hindering drug penetration and efficacy. The rupture of the cell wall structure will promote the admission of TB drugs, resulting in mycobacterial cell death. As a result, many clinically utilized anti-tuberculosis drugs and inhibitors in development are targeting the cell wall [11,34].

1.2.3. Drug-resistant

Antibiotic resistance is one of the major reasons why TB remains difficult to eliminate, and it has become a struggle to conquer

[35]. The rise of drug-resistant strains has made disease control more challenging, emphasizing the vital need for ongoing research into innovative therapeutic treatments to address and alleviate the worldwide effect of tuberculosis [36,37]. The different types of drug-resistant TB are shown in **Table 1.2**.

Table 1.2. The different types of drug-resistant tuberculosis

Types of drug-resistant TB	**Drugs**
Mono-resistance	Resistance to one first-line anti-TB drug only
Poly-resistance	Resistance to more than one first-line anti-TB drug, other than both INH and RIF
Multidrug-resistance (MDR)	Resistance to both INH and RIF, the most effective first-line TB drugs
Extensively drug-resistance (XDR))	INH and RIF, a fluoroquinolone, and a second-line injectable (AMK, CAP, and KAN) Or INH, RIF, a fluoroquinolone, and bedaquiline or linezolid
Rifampicin resistance (RR)	resistance to RIF, with or without resistance to other anti-TB drugs

1.2.4. Co infection

Co-infections are a cause of increased morbidity and death, and the results of co-infections are often worse than those of single-pathogen infections. Mtb infection reduces the immunological response to HIV, hastening the progression from HIV infection to AIDS. People with diabetes are also two to three times more likely to get TB than those without diabetes [38,39].

1.2.5. Target identification

Along with other significant challenges, the lack of novel pharmacological targets, as well as the accompanying drop in the number of anti-TB drugs in the therapeutic development pipeline, have obstructed the development of new therapies [40]. These obstacles persist, with one of the fundamental causes likely to be the permeability of bacterial cell envelopes, and Mtb can produce alternate metabolic pathways that circumvent drug targets [41]. Identifying new Mtb targets that can be inhibited to effectively eliminate the current strains is, therefore, a global effort.

1.2.6. Insufficient funding

TB, which mostly affects the poorest among the poor, is an example of a condition that can significantly contribute to the sickness poverty trap. The lack of funding for TB prevention, diagnosis, treatment, and research is concerning. This is a major barrier to TB research [42]. The lack of funding for TB drug discovery and development has contributed to the epidemic's obstinate persistence. This lack has also fostered conditions where drug-resistant TB has grown and spread [43], emphasizing the relevance of academic-private sector collaboration in anti-TB drug development.

In the context of these rising challenges in TB drug discovery computer-aided drug design (CADD) offers a promising solution, enabling researchers to decode interactions between therapeutic targets and potential drugs.

1.3. Computer Aided Drug Design (CADD)

Advancements in computational methodologies have transformed the area of drug design, providing a formidable arsenal of tools known as in silico methods or computer-aided drug design and discovery (CADD). It is an interdisciplinary area that combines computational techniques with bioinformatics, cheminformatics, and other theoretical disciplines to find and optimize potential drug candidates. CADD methods include computational recognition of promising drug targets, virtual screening of extensive chemical libraries for potential drug candidates, additional optimization of candidate compounds, and in silico evaluation of their potential toxicity and bioavailability [44,45].

The conventional drug discovery and development process is lengthy and difficult, with a high failure rate among new drugs in clinical trials. To address these challenges, the pharmaceutical sector frequently employs computer-aided drug design (CADD) methods to enhance the drug discovery process. The general workflow of CADD in drug discovery is illustrated in **Fig. 1.3**. CADD uses computer algorithms and models to assist researchers in screening prospective drug compounds, optimizing drug design, and understanding molecular-level interactions, therefore expediting the new drug development cycle and lowering costs and time [46]. As CADD advances, researchers are investigating new trends, such as Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) technologies for dealing with massive volumes of biological data generated by combinatorial chemistry [47]. These strategies improve

the potency and precision of drug discovery, indicating a hopeful future in the quest for transformative drugs.

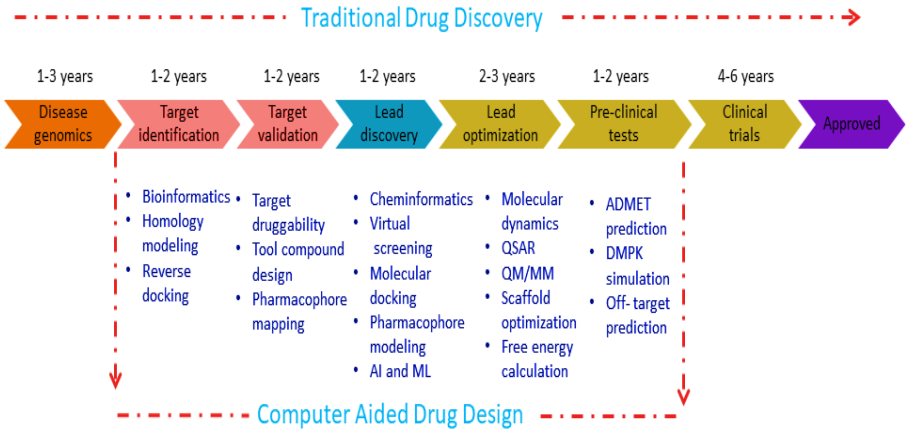


Fig. 1.3. Schematic representation of traditional drug discovery vs. CADD process

1.3.1. Types of CADD Techniques

CADD approaches are divided into two categories based on the availability of the target protein's 3D structure: structure-based drug design (SBDD) and ligand-based drug design (LBDD) [61]. SBDD uses computational chemistry methods to find or develop novel chemical compounds that potentially bind to the target, resulting in the inhibition of the target protein. SBDD has been made possible by the availability of 3D structures of therapeutically significant proteins, which favour the discovery of binding cavities. The most prevalent computational approaches used in SBDD are structure-based virtual screening (SBVS), molecular docking, and molecular dynamics (MD) simulation [62]. When 3D receptor information is unavailable, LBDD is employed, which is based on knowledge of compounds that bind to

the desired biological target. Some of the most useful LBDD approaches are ligand-based virtual screening (LBVS), similarity search, quantitative structure-activity relationship (QSAR) modelling, and pharmacophore generation [63]. **Fig. 1.4** shows the different types of SBDD and LBDD methods.

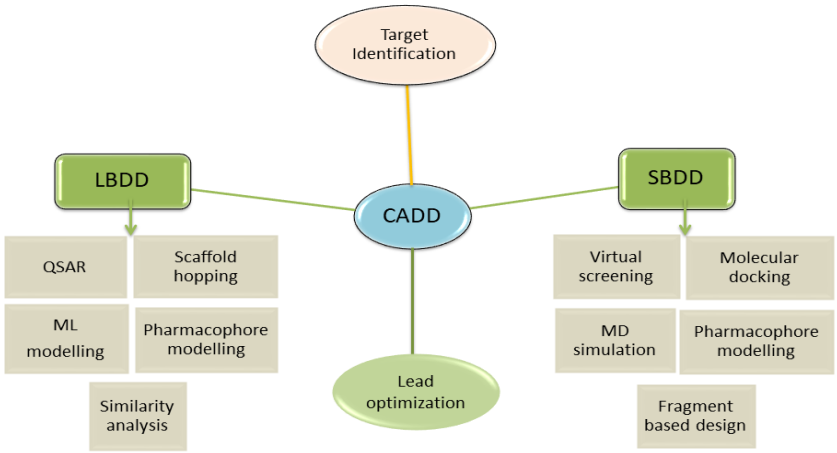


Fig. 1.4. Schematic representation of CADD techniques

The CADD methods have been successful in identifying numerous significant and innovative compounds since the 1970s, including antibiotics and protease inhibitors. These computational techniques were used in the development of several FDA-approved drugs. This comprises but is not restricted to, Aliskiren (Renin inhibitor), Boceprevir (protease inhibitor), Captopril (Angiotensin-converting enzyme inhibitor), Dorzolamide (Carbonic anhydrase inhibitor), Nelfinavir (HIV-1 protease inhibitors), Nolatrexed (thymidylate synthase inhibitor), Oseltamivir (neuraminidase inhibitor), Rupintrivir (rhinovirus 3C protease inhibitor), Saquinavir (HIV protease inhibitor), Imatinib (Tyrosine-kinase inhibitor) and Zanamivir

(Neuraminidase inhibitor) [64]. These fruitful findings unequivocally show that CADD offers useful and realistic methods to assist medicinal chemists and biologists in achieving their objectives of finding new active and lead-like compounds while removing inactive, reactive, and/or toxic ones.

1.3.2. Cheminformatics in Drug Discovery

Chemoinformatics is one of the growing and relevant fields of drug discovery by which one may quickly solve chemical problems with the assistance of computer techniques [48]. In 1998, Brown coined the term "cheminformatics", introducing it as “The use of information technology and management has become a critical part of the drug discovery process. Cheminformatics is the mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the area of drug lead identification and optimization” [49]. Chemoinformatics includes a wide range of computer tools for organizing, mining, visualizing, and analyzing the variety and coverage of the chemical space of compound collections. Because it considers molecules as graphs and their descriptors with related features (mostly physicochemical properties and biological activity), the ensemble of graphs (set of molecules) and its descriptors constitutes a chemical space in which the relationship between each compound must be established. This is the fundamental notion of chemoinformatics [50].

Chemical descriptors utilised in compound diversity analysis, compound activity prediction, and molecular data mining. Assessing

the structural novelty of compound libraries is aided by diversity analysis. Suppose the focus of a hit identification effort is to discover novel compounds. In that case, selecting collections with chemically diverse structures enhances the chance to identify new scaffolds that could serve as leads or prototypes for a specific biological target [51].

Cheminformatics uses ligand resources such as pharmacophore modelling, QSAR, molecular docking, and molecular dynamics (MD) simulations to investigate and understand the function of chemical systems. Pharmacophore modelling and QSAR study are mostly utilized to develop new ligands based on descriptor calculations and functional group substitutions [52]. Furthermore, the field of cheminformatics is anticipated to become even more significant in the future due to the growing emphasis on "Big Data", machine learning, and artificial intelligence in both drug discovery and society at large [53].

1.3.3. Molecular representations

Understanding the structure of molecules is one of the fundamental aspects of cheminformatics. Chemical structures provide critical information about the properties, behavior, and interactions of molecules, making them essential for drug discovery and other chemical applications. There are four ways to represent a molecule structure in a machine-readable format: as a string, a connection table, a collection of features like a fingerprint or a list of physical descriptors, or, most recently, a machine learning-based representation [54]. Among these string-based molecular representations, such as Simplified

Molecular Input Line Entry Systems (SMILES), and International Chemical Identifier (InChI) are commonly employed representations for the exchange and storage of chemical structures [55]. Different classes of molecular representations are shown in **Fig. 1.5.**

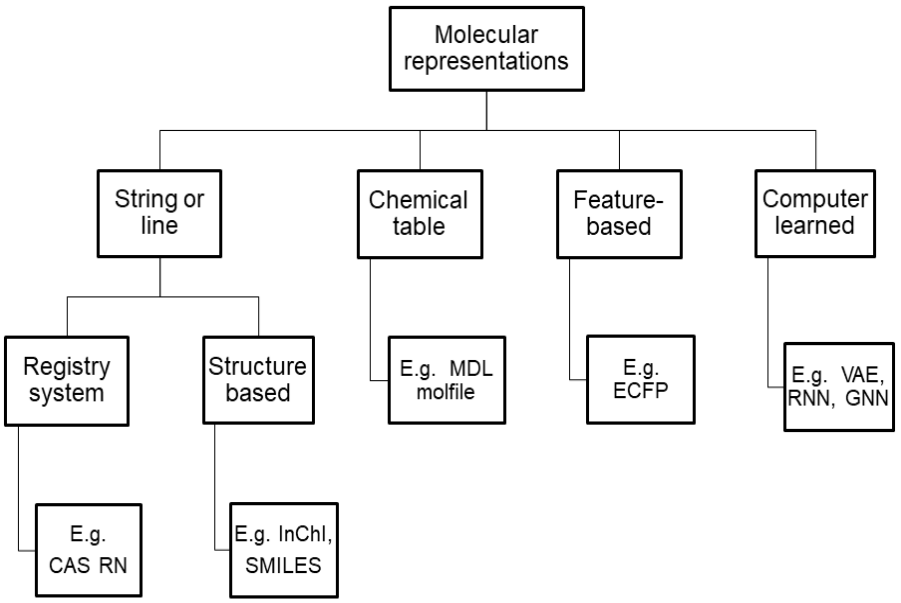


Fig. 1.5. Summary of different types of molecular representation

1.3.4. Molecular descriptors

Molecular descriptors are mathematical representations of structural or physicochemical properties of a molecule or a subset of a molecule generated by well-defined computer algorithms [56]. Numerous current applications in CADD and chemoinformatics rely on descriptor-based representations of molecules. These descriptors assist various cheminformatics applications including QSAR generation for ADMET (absorption, distribution, metabolism, excretion, and toxicity) predictions, similarity and diversity analysis, drug-like and lead-like

selection, substructure searching, chemical space navigation for drug subspace selection [57,58]. Molecular descriptors are mainly classified as 0D (zero-dimensional), 1D, 2D, 3D, and 4D descriptors as shown in **Table 1.3.** Commonly used software for the generation of molecular descriptors is also shown in **Table 1.4**.

Table 1.3. Molecular descriptor classifications and examples

Descriptor Dimension	**Descriptor Type**	**Example**
0D	Atom number, atom type, bonds and functional groups counts	Molecular weight (MW), Logp (partition coefficient)
1D	Molecular properties and physicochemical parameters	SMILES, Molecular formula
2D	Topological polar surface area (TPSA)	Molecular fingerprints (e.g. Morgan Fingerprint (ECFP4)) and constitutional descriptors (types and counts of atoms and bonds)
3D	Special properties of the molecule	Molecular shape descriptors (e.g. volume, surface area) Pharmacophore features
4D	Electrostatic potential descriptors with spatiotemporal aspects	Molecular dynamics descriptors, radius of gyration (Rg), solvent accessible surface area (SASA), Time-dependent properties (e.g., dynamic polar surface area (dPSA), time-dependent dipole moment

Table 1.4. Commonly used software for the generation of molecular descriptors

Software	**Description**	**Availability**
PaDEL	It is used to calculate molecular descriptors and fingerprints. The software currently calculates 1444 1D, 2D descriptors and 431 3D descriptors) and 12 types of fingerprints (total 16092 bits)	Free software
Dragon	Dragon 6.0 is the latest version is used for the calculation of molecular descriptors. It calculates 4885 molecular descriptors	Commercial
PowerMV	It is Windows based software for statistical analysis, molecular viewing, descriptor generation, and similarity search	Free software
MOE	It is a web-based tool for performing SAR analysis and visualization. It calculates over 300 topological, physical properties, and structural descriptors	Commercial
CDK	CDK is a java based descriptor calculation tool that calculates topological, geometrical, charge-based, and constitutional descriptors	Free software

1.3.5. Databases in Drug Discovery

Chemical and molecular target databases are the cornerstone of drug discovery, catalyzing the development of CADD tools to save time and cost and to develop a hypothesis for the identification and design of novel therapeutic compounds. In addition to the power of computing, resources such as open-access and commercial organic compounds, phytochemicals, experimental, investigational, and approved drugs, compound targets, peptides, and metabolomic databases have risen dramatically [54]. Utilizing these existing

chemical and drug databases for CADD techniques such as virtual screening speeds up the methods compared to designing a novel drug because the substances in the databases have already been synthesized (generally) and characterized [55]. Numerous databases are available that are helpful for SBDD and LBDD techniques and can support the most prevalent developments and applications, including virtual screening, Small-molecule docking, target prediction, and ADMET prediction. Databases are divided into various categories, such as chemical and biological databases, based on the data they store [56].

1.3.5.1. Chemical Databases

Chemical databases are essential in cheminformatics because they offer a vast collection of information about chemicals that supports research, drug discovery, and other chemical-related endeavors [59]. Chemical databases include atomic and molecular information as well as encoded chemical structures. These databases enable scientists to quickly acquire information on the molecule of interest, as well as group and analyze various chemical compounds based on their physicochemical features, structural, and biological activities. Several chemical databases are available online, with a user-friendly interface. The major chemical databases that are easily accessible and frequently used are listed in **Table 1.5.**

Table 1.5. The list of major chemical databases that are freely accessible and widely used

Database	**Description**	**Size**
PubChem [60]	The world's largest open source repository that includes chemical structures and their biological test results	252M
ChEMBL [61]	ChEMBL is a manually curated database of bioactive compounds having drug-like properties	1.8M
ZINC-15 [62]	A free database of commercially accessible chemicals for virtual screening	230M
DrugBank [63]	Database providing comprehensive data about drugs, their mechanisms, interactions, and targets	0.0065 M
BindingDB [64]	It is a small-molecule binding affinity database	0.78 M
ChemDB [65]	The database provides data on compounds, including predicted or experimentally determined physicochemical attributes	5M
IMPPAT [66]	The most comprehensive database on the phytochemicals of Indian medicinal plants	0.017967M
Dr. Duke's Phytochemical and Ethnobotanical Databases [67]	One of the world's leading repository of ethnobotanical data	0.049788M
SuperNatural [68]	A freely accessible database of natural products and their derivatives	0.05 M

1.3.5.2. Protein Databases

Protein databases are a treasure trove of potential novel therapeutic targets. They provide researchers with access to structural and sequence information for tens of thousands of sequences from various organisms [69]. These databases work as a backbone for researchers, offering access to enormous amounts of protein-related data. They can learn about the biological functions of various proteins, their three-dimensional structures, and their amino acid sequences. RCSB Protein Data Bank (PDB) is an excellent example of a protein database since it is the principal database for 3D structures of biological macromolecules established by X-ray crystallography and NMR [70]. The list of important protein databases are shown in **Table 1.6.**

Table 1.6. The list of major protein databases that are accessible online

Database	**Description**	**Web address**
PDB	A database providing three-dimensional structural data of biological macromolecules	https://www.rcsb.org/
UniProt	The world's best high-quality, comprehensive, and publicly available database for protein sequence and functional information	https://www.uniprot.org/
NCBI	It is a compilation of sequences from various sources, including translations from annotated coding sections in RefSeq, GenBank, and TPA, as well as data from PIR, SwissProt, PDB, and PRF.	https://www.ncbi.nlm.nih.gov/protein

PDBsum	It is a pictorial database that offers an overview of the contents of each 3D structure deposited in the PDB	https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/
PDBbind	PDBbind is a comprehensive collection of experimentally determined binding affinity data (K_d, K_i, and IC_{50}) for protein-ligand complexes stored in the PDB	https://www.pdbbind-plus.org.cn/

The Role of the chemical database is very important in CADD. Data mining these massive chemical databases alongside biological information is crucial to the development of novel therapeutic drugs and offering insight into disease causes. Researchers can find novel chemical motifs, new chemical insights, or possible drug targets by examining patterns, trends, and relationships in the data.

1.3.6. Data mining of database

To identify unknown drugs that are similar to certain known drugs and investigate the relationship between them and their biological and chemical activities, the drug discovery methods employ data mining techniques and cheminformatics. Particularly in the area of chemical space mining, cheminformatics integrates the scientific disciplines of computer science and chemistry [71]. Chemical space refers to the collection of all feasible molecules or materials, including all chemical databases. Additionally, as chemical "big" data from combinatorial synthesis and high throughput screening (HTS) continues to grow rapidly, machine learning has emerged as a crucial tool for drug designers looking to mine chemical information from massive

compound databases to generate drugs with significant biological properties [51]. Data mining, also known as patterns analysis, is the process of analyzing large data sets using tools like association, clustering, segmentation, and classification to improve data manipulation and help pharmaceutical companies reduce costs while improving the quality of drug design and discovery methods [72].

Database mining approaches depend on the molecular descriptors used to characterize a structure database [73]. There are several data mining strategies accessible in cheminformatics. Traditional methods include descriptor calculations, chemical similarity searching, clustering (K-means and hierarchical), principal component analysis (PCA), association rule mining, scaffold analysis, substructure searching, and activity landscape modelling. Due to the rise of the chemical big data age, ML-based methods like as classification, regression, clustering, association rule mining, dimensionality reduction, and neural networks have been deployed. These approaches are utilized for popular cheminformatics applications in drug development, such as virtual screening, QSAR, chemical space exploration, activity cliff generation, and ADMET prediction [74].

1.3.7. Chemical Space and Its Relevance

A key idea in chemoinformatics, the concept of chemical space has extensive conceptual and practical applicability in many chemistry domains, including drug design and discovery. The multi-dimensional descriptor space, or "chemical space", is the enormous collection of all potential chemical compounds and their properties, including all known

drug molecules and those that have yet to be discovered. The total number of chemical space is predicted to be 10^{60}, although only a small portion of this chemical space has been explored yet [75,76]. The constant increase in the number of molecules available highlights the issue of how many compounds exist and which have the potential to become drugs. Given the extensive chemical space, identifying the regions most likely to contain biologically active compounds commonly referred to as biologically relevant chemical space (BRCS) poses a significant challenge for chemical biologists and drug discoverers [77]. In order to identify BRCS and develop novel drugs, we must effectively explore chemical space.

Various techniques, such as descriptor-based mapping, dimensionality reduction (visualization method), fragment and scaffold analysis, fingerprint comparison and similarity analysis, clustering, de novo design, virtual screening, and landscape modelling, can be used to explore chemical space [78]. Medicinal chemists are increasingly using visual representations of the chemical space to better understand chemical data. It is useful for examining the diversity of various data sets, exploring the relationships between collections, and evaluating the possibility of covering other regions in chemical space that have yet to be explored.

Dimensionality reduction techniques are used in visualization approaches to convert multi-dimensional data into 2D or 3D representations with minimal information loss. Principal Component Analysis (PCA), t-SNE, UMAP, and self-organizing map (SOM) are some of the most widely used dimensionality reduction approaches.

[79]. Chemical space visualization for drug discovery tasks using cheminformatics techniques allows for the systematic classification of approved drugs and databases annotated with biological activity to define chemical spaces that are relevant to biology and medicine, as well as the quantitative comparison of general screening collections. Visualizing chemical space allows medicinal chemists and researchers to explore the massive chemical space. Chemical space exploration aids researchers in a number of ways, including lead optimization, drug-like property optimization, understanding the structure-activity relationship (SAR), ML, and predictive modelling [80].

1.4. Role Of CADD in Screening and Developing Novel Anti-TB Drugs

Applying CADD has significantly increased the efficiency of finding new anti-TB compounds, helped to understand their mechanism of action, directed the structural optimization of lead compounds, and raised expectations that it will help produce better outcomes [81]. The use of SBDD in tuberculosis research has led to the discovery of several antimycobacterial drugs that were previously clinically evaluated, demonstrating their utility in the drug discovery and development process [82,83]. Furthermore, LBDD has a lengthy history, and despite the lack of protein structural information, many candidates have already been identified [84–86]. Finding strong-cum-druggable candidates in the current drug-development module is best accomplished by combining cheminformatics and bioinformatics. With the help of computational tools, we have an unparalleled chance to use the data at hand to better understand the disease, identify the most

promising drug and vaccine candidates, and create a clinically useful toolbox for tracking and managing antibiotic resistance [87].

A.K. Saxena *et al.,* developed a QSAR model based on molecular docking interactions to find new leads. With a strong correlation (r_{ext} = 0.851), this model predicts external datasets and can be applied to the discovery of novel chemical entities (NCEs) and repurposed medications for TB treatments [88]. The work of Gaurav Bhargava *et al*., included molecular dynamic simulation studies of putative InhA inhibitors that were virtually screened, docking studies, the development of a 3D-QSAR model, and the development of a ligand-based pharmacophore model. Future research into new InhA inhibitors to treat drug-resistant tuberculosis may benefit from this work [89]. Eugene Megnassan and colleagues designed new potent pyrrolidine carboxamide (PCAM) inhibitors of InhA using computational SBDD [90]. Numerous studies have already identified that PCAM directly inhibits InhA.

Vino Sundararajan *et al*., reviewed cutting-edge computational tools and ML techniques that have been effectively used in biomedical research and discussed their potential uses in the fight against TB [87]. Jiao Li *et al.,* developed and validated a successful virtual screening process using machine learning to repurpose existing drugs with anti-Mtb effectiveness. [91]. Shovonlal Bhowmick *et al.,* employed in-silico and ML methodologies to explore novel anti-TB compounds from the SelleckChem database against InhA [92]. Azeddine Ibrahimi *et al.,* employed three distinct CADD methods, mutation effect modelling,

virtual screening, and 3D-pharmacophore search to find direct InhA inhibitors [31].

1.5. Scope of the present study

CADD approaches have become a viable tool for identifying novel anti-TB drugs due to the challenges encountered in traditional TB drug development. Multidrug resistance TB is a serious issue worldwide it raises the urgent need for novel inhibitors with minimal resistance. Researchers use both SBDD and LBDD approaches to explore vast chemical space, which results in the discovery of potential anti-TB drugs. Virtual screening, docking approaches, molecular dynamics simulations, pharmacophore-based models, and sophisticated methodologies, such as ML-based classification algorithms, are applied in many areas of TB drug development. In this study, we attempted to analyze and visualize the chemical space of Mtb inhibitors to enhance our understanding of their diversity and extent of exploration. Furthermore, we identified activity cliff generators by modelling Mtb inhibitor activity landscapes. We have additionally screened the ZINC-15 database to find InhA inhibitors. Finally, we developed an ML-based QSAR model and screened molecules from Indian medicinal plants for their potential against TB.

1.6. Objectives of the present study

Even in the 21st century, TB still affects almost one-third of the world's population, making it one of the major infectious diseases. Identifying anti-TB drugs is essential because of the prevalence of tuberculosis worldwide and the rise of drug-resistant strains. Mtb is the

most important member and major pathogen causing tuberculosis in humans. Among Mtb drug targets, InhA is a clinically validated, attractive, and well-researched target for developing new TB drugs. Consequently, scientists primarily concentrate on this target in order to discover new TB inhibitors. However, the development of a tuberculosis free world is currently facing a few challenges including multidrug drug resistance TB. The widespread use of CADD techniques is facilitated in this context by the lack of funding for tropical neglected diseases like tuberculosis.

Cheminformatics and CADD technologies are developing as key tools, with many CADD methodologies being used in the rational drug design field. Researchers used virtual screening techniques to screen a large number of biologically active drugs against tuberculosis. Several studies developed QSAR classification models for the screening molecules against Mtb. Additionally, data mining and ML techniques aid in uncovering hidden patterns in the chemical space of Mtb inhibitors. Because MDR-TB has arisen as a serious threat among tuberculosis patients worldwide, and traditional drug discovery hurdles against tuberculosis drug development, it is vital to explore the chemical space of tuberculosis utilizing diverse cheminformatics methodologies. In this context we used CADD approaches with cheminformatics tools:

- For the diversity analysis and similarity search of Mtb inhibitors through the visualization of vast chemical space.

- To develop an activity landscape model for identifying activity cliff generators in order to develop effective QSAR models.

- To perform virtual screening of the ZINC-15 database to identify potential Mtb inhibitors using molecular docking analysis.

- To carry out molecular dynamics simulation and ADMET analysis on selected screened molecules.

- To build ML-based QSAR classification models with various algorithms.

- Using the best QSAR model, screen molecules from Indian medicinal plants that exhibit anti-tubercular activity.

References

[1] B.R. Bloom, A half-century of research on tuberculosis: Successes and challenges, J. Exp. Med. 220 (2023) e20230859. https://doi.org/10.1084/jem.20230859.

[2] S. Mousavi-Sagharchi, A. Ghorbani, M. Meskini, S.D. Siadat, Historical Examination of Tuberculosis; From Ancient Affliction to Modern Challenges, J. Infect. Public Health. 18 (2025) 102649. https://doi.org/10.1016/j.jiph.2024.102649.

[3] World Health Organization, Global Tuberculosis Report, Geneva, Switzerland, 2024. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024.

[4] G. Mancuso, A. Midiri, S. De Gaetano, E. Ponzo, C. Biondo, Tackling Drug-Resistant Tuberculosis: New Challenges from the Old Pathogen Mycobacterium tuberculosis, Microorganisms. 11 (2023). https://doi.org/10.3390/microorganisms11092277.

[5] R.D. Kanabalan, L.J. Lee, T.Y. Lee, P.P. Chong, L. Hassan, R. Ismail, V.K. Chin, Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery, Microbiol. Res. 246 (2021) 126674. https://doi.org/https://doi.org/10.1016/j.micres.2020.126674.

[6] M. Orgeur, C. Sous, J. Madacki, R. Brosch, Evolution and emergence of Mycobacterium tuberculosis, FEMS Microbiol. Rev. 48 (2024) fuae006. https://doi.org/10.1093/femsre/fuae006.

[7] E. Cambau, M. Drancourt, Steps towards the discovery of Mycobacterium tuberculosis by Robert Koch, 1882, Clin. Microbiol. Infect. 20 (2014) 196–201. https://doi.org/10.1111/1469-0691.12555.

[8] K. Sakamoto, The Pathology of Mycobacterium tuberculosis Infection, Vet. Pathol. 49 (2012) 423–439. https://doi.org/10.1177/0300985811429313.

[9] M.S. Glickman, W.R. Jacobs Jr., Microbial Pathogenesis of

Mycobacterium tuberculosis: Dawn of a Discipline, Cell. 104 (2001) 477–485. https://doi.org/10.1016/S0092-8674(01)00236-7.

[10] A. Diab, H. Dickerson, O. Al Musaimi, Targeting the Heart of Mycobacterium: Advances in Anti-Tubercular Agents Disrupting Cell Wall Biosynthesis, Pharmaceuticals. 18 (2025). https://doi.org/10.3390/ph18010070.

[11] C. Vilchèze, Mycobacterial Cell Wall: A Source of Successful Targets for Old and New Drugs, Appl. Sci. 10 (2020). https://doi.org/10.3390/app10072278.

[12] H.N. Jnawali, S. Ryoo, First– and Second–Line Drugs and Drug Resistance, in: B.H. Mahboub, M.G. Vats (Eds.), IntechOpen, Rijeka, 2013: p. Ch. 10. https://doi.org/10.5772/54960.

[13] M. Shaku, C. Ealand, B.D. Kana, Cell Surface Biosynthesis and Remodeling Pathways in Mycobacteria Reveal New Drug Targets, Front. Cell. Infect. Microbiol. 10 (2020). https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2020.603382.

[14] Y.M. Jacobo-Delgado, A. Rodríguez-Carlos, C.J. Serrano, B. Rivas-Santiago, Mycobacterium tuberculosis cell-wall and antimicrobial peptides: a mission impossible?, Front. Immunol. 14 (2023). https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1194923.

[15] M.J. Catalão, S.R. Filipe, M. Pimentel, Revisiting Anti-tuberculosis Therapeutic Strategies That Target the Peptidoglycan Structure and Synthesis, Front. Microbiol. 10 (2019). https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.00190.

[16] M. Chauhan, R. Barot, R. Yadav, K. Joshi, S. Mirza, R. Chikhale, V.K. Srivastava, M.R. Yadav, P.R. Murumkar, The Mycobacterium tuberculosis Cell Wall: An Alluring Drug Target for Developing Newer Anti-TB Drugs—A Perspective, Chem. Biol. Drug Des. 104 (2024) e14612.

https://doi.org/https://doi.org/10.1111/cbdd.14612.

[17] H. Marrakchi, G. Lanéelle, A.K. Quémard, InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II, Microbiology. 146 (Pt 2 (2000) 289–296. https://doi.org/10.1099/00221287-146-2-289.

[18] M. Toraskar, P. Kamble, Enoyl Acyl Carrier Protein Reductase Inhibitors: An Emerging Target, Int. J. ChemTech Res. 11 (2018) 123–133. https://doi.org/10.20902/IJCTR.2018.110715.

[19] X. Duan, X. Xiang, J. Xie, Crucial components of mycobacterium type II fatty acid biosynthesis (Fas-II) and their inhibitors, FEMS Microbiol. Lett. 360 (2014) 87–99. https://doi.org/10.1111/1574-6968.12597.

[20] G. Aner, H.J. Kalervo, K.A. J., Function of Heterologous Mycobacterium tuberculosis InhA, a Type 2 Fatty Acid Synthase Enzyme Involved in Extending C20 Fatty Acids to C60-to-C90 Mycolic Acids, during De Novo Lipoic Acid Synthesis in Saccharomyces cerevisiae, Appl. Environ. Microbiol. 74 (2008) 5078–5085. https://doi.org/10.1128/AEM.00655-08.

[21] R.P. Massengo-Tiassé, J.E. Cronan, Diversity in enoyl-acyl carrier protein reductases, Cell. Mol. Life Sci. 66 (2009) 1507–1517. https://doi.org/10.1007/s00018-009-8704-7.

[22] M. Ghattas, R. Mansour, N. Atatreh, R. Bryce, Analysis of Enoyl-Acyl Carrier Protein Reductase Structure and Interactions Yields an Efficient Virtual Screening Approach and Suggests a Potential Allosteric Site, Chem. Biol. Drug Des. 87 (2015). https://doi.org/10.1111/cbdd.12635.

[23] K. Rožman, I. Sosič, R. Fernandez, R. Young, A. Mendoza-Losana, S. Gobec, L. Encinas, A new 'golden age' for the antitubercular target InhA, Drug Discov. Today. 22 (2016). https://doi.org/10.1016/j.drudis.2016.09.009.

[24] M. Prasad, R. Bhole, P. Khedekar, R. Chikhale, Mycobacterium enoyl acyl carrier protein reductase (InhA): A key target for antitubercular drug discovery, Bioorg. Chem. 115 (2021) 105242. https://doi.org/10.1016/j.bioorg.2021.105242.

[25] S.R. Luckner, N. Liu, C.W. am Ende, P.J. Tonge, C. Kisker, A Slow, Tight Binding Inhibitor of InhA, the Enoyl-Acyl Carrier Protein Reductase from Mycobacterium tuberculosis, J. Biol. Chem. 285 (2010) 14330–14337. https://doi.org/https://doi.org/10.1074/jbc.M109.090373.

[26] B. Lei, C.-J. Wei, S.-C. Tu, Action Mechanism of Antitubercular Isoniazid: ACTIVATION BY MYCOBACTERIUM TUBERCULOSIS KatG, ISOLATION, AND CHARACTERIZATION OF InhA INHIBITOR, J. Biol. Chem. 275 (2000) 2520–2526. https://doi.org/https://doi.org/10.1074/jbc.275.4.2520.

[27] J.N. Torres, P. Lynthia V, R. Timothy C, V. Thomas C, A. Anu M, E. Afif, G. Amy P, R.-B. Sarah M, C. Ashu, Z. Victoria, S. Elizabeth M, S. Frederick A, C. Donald, R. Camilla, G. Maria Tarcela, C. Valeru, C. Antonino, F. and Valafar, Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates, Emerg. Microbes Infect. 4 (2015) 1–9. https://doi.org/10.1038/emi.2015.42.

[28] E. Lourdes, L. Si-Yang, R.-T. Joaquin, T. Rokeya, T. Sandeep, S. Heena, G.-P. Adolfo, L. Jin, G. del R. Rubén, D.M. Jaime, S. Verónica, S. Izidor, G. Stanislav, M.-L. Alfonso, C.P. J., M. Khisi, F. Nader, B.-A. David, N.E. L., Contribution of direct InhA inhibitors to novel drug regimens in a mouse model of tuberculosis, Antimicrob. Agents Chemother. 68 (2024) e00357-24. https://doi.org/10.1128/aac.00357-24.

[29] Q. Zhang, J. Han, Y. Zhu, F. Yu, X. Hu, H.H.Y. Tong, H. Liu, Discovery of novel and potent InhA direct inhibitors by ensemble docking-based virtual screening and biological assays, J. Comput. Aided. Mol. Des. 37 (2023) 695–706. https://doi.org/10.1007/s10822-023-00530-4.

[30] U.H. Manjunatha, S.P. S. Rao, R.R. Kondreddi, C.G. Noble, L.R. Camacho, B.H. Tan, S.H. Ng, P.S. Ng, N.L. Ma, S.B. Lakshminarayana, M. Herve, S.W. Barnes, W. Yu, K. Kuhen, F. Blasco, D. Beer, J.R. Walker, P.J. Tonge, R. Glynne, P.W. Smith, T.T. Diagana, Direct inhibitors of InhA are active against Mycobacterium tuberculosis, Sci. Transl. Med. 7 (2015) 269ra3-

269ra3. https://doi.org/10.1126/scitranslmed.3010597.

[31] G. el Haddoumi, M. Mansouri, B. Houda, E.M. Bouricha, I. Kandoussi, L. Belyamani, A. Ibrahimi, Facing Antitubercular Resistance: Identification of Potential Direct Inhibitors Targeting InhA Enzyme and Generation of 3D-pharmacophore Model by in silico Approach, Adv. Appl. Bioinforma. Chem. Volume 16 (2023) 49–59. https://doi.org/10.2147/AABC.S394535.

[32] J.D. Ernst, Mechanisms of M. tuberculosis Immune Evasion as Challenges to TB Vaccine Design, Cell Host Microbe. 24 (2018) 34–42. https://doi.org/https://doi.org/10.1016/j.chom.2018.06.004.

[33] P. Chandra, S.J. Grigsby, J.A. Philips, Immune evasion and provocation by Mycobacterium tuberculosis, Nat. Rev. Microbiol. 20 (2022) 750–766. https://doi.org/10.1038/s41579-022-00763-4.

[34] L. Zhang, Z. Rao, Structural biology and inhibition of the Mtb cell wall glycoconjugates biosynthesis on the membrane, Curr. Opin. Struct. Biol. 82 (2023) 102670. https://doi.org/https://doi.org/10.1016/j.sbi.2023.102670.

[35] J.-D. Pedelacq, M.C. Nguyen, T.C. Terwilliger, L. Mourey, A Comprehensive Review on Mycobacterium tuberculosis Targets and Drug Development from a Structural Perspective, in: Struct. Biol. Drug Discov., 2020: pp. 545–566. https://doi.org/https://doi.org/10.1002/9781118681121.ch23.

[36] A. Iacobino, L. Fattorini, F. Giannoni, Drug-Resistant Tuberculosis 2020: Where We Stand, Appl. Sci. 10 (2020). https://doi.org/10.3390/app10062153.

[37] S.M. Gygli, S. Borrell, A. Trauner, S. Gagneux, Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives, FEMS Microbiol. Rev. 41 (2017) 354–373. https://doi.org/10.1093/femsre/fux011.

[38] A.A. Boadu, M. Yeboah-Manu, S. Osei-Wusu, D. Yeboah-Manu, Tuberculosis and diabetes mellitus: The complexity of the comorbid interactions, Int. J. Infect. Dis. 146 (2024) 107140.

https://doi.org/https://doi.org/10.1016/j.ijid.2024.107140.

[39] B. Adeoye, L. Nakiyingi, Y. Moreau, E. Nankya, A.J. Olson, M. Zhang, K.R. Jacobson, A. Gupta, Y.C. Manabe, M.C. Hosseinipour, J. Kumwenda, M. Sagar, Mycobacterium tuberculosis disease associates with higher HIV-1-specific antibody responses, IScience. 26 (2023) 106631. https://doi.org/https://doi.org/10.1016/j.isci.2023.106631.

[40] G.S. Shetye, S.G. Franzblau, S. Cho, New tuberculosis drug targets, their inhibitors, and potential therapeutic impact, Transl. Res. 220 (2020) 68–97. https://doi.org/https://doi.org/10.1016/j.trsl.2020.03.007.

[41] A. Schami, M.N. Islam, J.T. Belisle, J.B. Torrelles, Drug-resistant strains of Mycobacterium tuberculosis: cell envelope profiles and interactions with the host, Front. Cell. Infect. Microbiol. 13 (2023). https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2023.1274175.

[42] T. Tanimura, E. Jaramillo, D. Weil, M. Raviglione, K. Lönnroth, Financial burden for tuberculosis patients in low- And middle-income countries: A systematic review, Eur. Respir. J. 43 (2014). https://doi.org/10.1183/09031936.00193413.

[43] G. Thakur, S. Thakur, H. Thakur, Status and challenges for tuberculosis control in India – Stakeholders' perspective, Indian J. Tuberc. 68 (2021) 334–339. https://doi.org/https://doi.org/10.1016/j.ijtb.2020.10.001.

[44] S.K. Niazi, Z. Mariam, Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis, Pharmaceuticals. 17 (2024). https://doi.org/10.3390/ph17010022.

[45] S. Brogi, T.C. Ramalho, K. Kuca, J.L. Medina-Franco, M. Valko, Editorial: In silico Methods for Drug Design and Discovery, Front. Chem. Volume 8- (2020). https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2020.00612.

[46] Z. Pei, Computer-aided drug discovery: From traditional

simulation methods to language models and quantum computing, Cell Reports Phys. Sci. 5 (2024) 102334. https://doi.org/https://doi.org/10.1016/j.xcrp.2024.102334.

[47] D.P. Mihai, G.M. Nitulescu, Computer-Aided Drug Design and Drug Discovery, Pharmaceuticals. 18 (2025). https://doi.org/10.3390/ph18030436.

[48] J. Xu, A. Hagler, Chemoinformatics in Drug Discovery, Molecules. 7 (2002). https://doi.org/10.3390/70800566.

[49] F.K. Brown, Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery., in: J.A.B.T.-A.R. in M.C. Bristol (Ed.), Academic Press, 1998: pp. 375–384. https://doi.org/https://doi.org/10.1016/S0065-7743(08)61100-8.

[50] J.C. Gómez-Verjan, K.D. Rodríguez-Hernández, R. Reyes-Chilpa, Chapter 8 - Bioactive Coumarins and Xanthones From Calophyllum Genus and Analysis of Their Druglikeness and Toxicological Properties, in: B.T.-S. in N.P.C. Atta-ur-Rahman (Ed.), Elsevier, 2017: pp. 277–307. https://doi.org/https://doi.org/10.1016/B978-0-444-63930-1.00008-9.

[51] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today. 23 (2018) 1538–1546. https://doi.org/https://doi.org/10.1016/j.drudis.2018.05.010.

[52] R. Kumar, A. Lathwal, G. Nagpal, V. Kumar, P.K. Raghav, Chapter 1 - Impact of chemoinformatics approaches and tools on current chemical research, in: N. Sharma, H. Ojha, P.K. Raghav, R. k. B.T.-C. and B. in the P.S. Goyal (Eds.), Academic Press, 2021: pp. 1–26. https://doi.org/https://doi.org/10.1016/B978-0-12-821748-1.00001-4.

[53] G. Restrepo, 7 - Approaching history of chemistry through big data on chemical reactions and compounds, in: S.C. Basak, M.B.T.-B.D.A. in C. and B. Vračko (Eds.), Elsevier, 2023: pp. 171–186. https://doi.org/https://doi.org/10.1016/B978-0-323-85713-0.00033-5.

[54] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, WIREs Comput. Mol. Sci. 12 (2022) e1603. https://doi.org/https://doi.org/10.1002/wcms.1603.

[55] A. Karthikeyan, U. Priyakumar, Artificial intelligence: machine learning for chemical sciences, J. Chem. Sci. 134 (2022). https://doi.org/10.1007/s12039-021-01995-2.

[56] R. Todeschini, V. Consonni, P. Gramatica, 4.05 - Chemometrics in QSAR, in: S.D. Brown, R. Tauler, B.B.T.-C.C. Walczak (Eds.), Elsevier, Oxford, 2009: pp. 129–172. https://doi.org/https://doi.org/10.1016/B978-044452701-1.00007-7.

[57] A. Mauri, V. Consonni, R. Todeschini, Molecular Descriptors, in: Handb. Comput. Chem., 2017: pp. 2065–2093. https://doi.org/10.1007/978-3-319-27282-5_51.

[58] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, in: Methods Princ. Med. Chem., 2009: p. 1252. https://doi.org/10.1002/9783527628766.

[59] A. Hersey, J. Chambers, L. Bellis, A. Patrícia Bento, A. Gaulton, J.P. Overington, Chemical databases: curation or integration by user-defined equivalence?, Drug Discov. Today Technol. 14 (2015) 17–24. https://doi.org/https://doi.org/10.1016/j.ddtec.2015.01.005.

[60] Q. Li, T. Cheng, Y. Wang, S.H. Bryant, PubChem as a public resource for drug discovery, Drug Discov. Today. 15 (2010) 1052–1057. https://doi.org/https://doi.org/10.1016/j.drudis.2010.10.003.

[61] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107. https://doi.org/10.1093/nar/gkr777.

[62] T. Sterling, J.J. Irwin, ZINC 15 – Ligand Discovery for Everyone, J. Chem. Inf. Model. 55 (2015) 2324–2337.

https://doi.org/10.1021/acs.jcim.5b00559.

[63] D. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901-D906, Nucleic Acids Res. 36 (2008) D901-6. https://doi.org/10.1093/nar/gkm958.

[64] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (2016) D1045–D1053. https://doi.org/10.1093/nar/gkv1072.

[65] J. Chen, S.J. Swamidass, Y. Dou, J. Bruand, P. Baldi, ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources, Bioinformatics. 21 (2005) 4133–4139. https://doi.org/10.1093/bioinformatics/bti683.

[66] K. Mohanraj, B.S. Karthikeyan, R.P. Vivek-Ananth, R.P.B. Chand, S.R. Aparna, P. Mangalapandi, A. Samal, IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics, Sci. Rep. 8 (2018) 4329. https://doi.org/10.1038/s41598-018-22631-z.

[67] S. Skoczen, R. Bussmann, ebDB - an International Ethnobotany Database, Lyonia. 11 (2006) 71–81.

[68] K. Gallo, E. Kemmler, A. Goede, F. Becker, M. Dunkel, R. Preissner, P. Banerjee, SuperNatural 3.0—a database of natural products and natural product-based derivatives, Nucleic Acids Res. 51 (2023) D654–D659. https://doi.org/10.1093/nar/gkac1008.

[69] C. Chichester, P. Gaudet, Target discovery from protein databases: challenges for curation, Drug Discov. Today Technol. 14 (2015) 11–16. https://doi.org/https://doi.org/10.1016/j.ddtec.2015.01.003.

[70] D. Zou, L. Ma, J. Yu, Z. Zhang, Biological Databases for Human Research, Genomics. Proteomics Bioinformatics. 13 (2015) 55–63. https://doi.org/https://doi.org/10.1016/j.gpb.2015.01.006.

[71] X. Yan, Mining of Chemical Data BT - Encyclopedia of Database Systems, in: L. LIU, M.T. ÖZSU (Eds.), Springer US, Boston, MA, 2009: pp. 1748–1750. https://doi.org/10.1007/978-0-387-39940-9_1299.

[72] F. Gullo, From Patterns in Data to Knowledge Discovery: What Data Mining Can Do, Phys. Procedia. 62 (2015) 18–22. https://doi.org/https://doi.org/10.1016/j.phpro.2015.02.005.

[73] G. Cruciani, M. Pastor, R. Mannhold, Suitability of Molecular Descriptors for Database Mining. A Comparative Analysis, J. Med. Chem. 45 (2002) 2685–2694. https://doi.org/10.1021/jm0011326.

[74] P. Willett, Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules, (2005).

[75] J.-L. Reymond, Chemical space as a unifying theme for chemistry, J. Cheminform. 17 (2025) 6. https://doi.org/10.1186/s13321-025-00954-0.

[76] J.-L. Reymond, The Chemical Space Project, Acc. Chem. Res. 48 (2015) 722–730. https://doi.org/10.1021/ar500432k.

[77] K.K. Duncan, D.D. Rudnicki, C.P. Austin, D.A. Tagle, Exploring Novel Biologically-Relevant Chemical Space Through Artificial Intelligence: The NCATS ASPIRE Program, Front. Robot. AI. 6 (2020). https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2019.00143.

[78] C. Lu, S. Liu, W. Shi, J. Yu, Z. Zhou, X. Zhang, X. Lu, F. Cai, N. Xia, Y. Wang, Systemic evolutionary chemical space exploration for drug discovery, J. Cheminform. 14 (2022) 19. https://doi.org/10.1186/s13321-022-00598-4.

[79] J. Medina-Franco, K. Martínez, M. Giulianotti, R. Houghten, C. Pinilla, Visualization of the Chemical Space in Drug Discovery, Curr. Comput. - Aided Drug Des. 4 (2008) 322–333. https://doi.org/10.2174/157340908786786010.

[80] J. de J. Naveja, F. Saldivar-Gonzalez, D. Prado-Romero, A. Ruiz-Moreno, M. Velasco-Velázquez, R. Miranda-Quintana, J. Medina-Franco, Visualization, Exploration, and Screening of Chemical Space in Drug Discovery, in: 2024: pp. 365–393. https://doi.org/10.1002/9783527840748.ch16.

[81] S.J. Macalino, J. Billones, V. Organo, M.C. Carrillo, In Silico Strategies in Tuberculosis Drug Discovery, Molecules. 25 (2020). https://doi.org/10.3390/molecules25030665.

[82] M.A. Ejalonibu, S.A. Ogundare, A.A. Elrashedy, M.A. Ejalonibu, M.M. Lawal, N.N. Mhlongo, H.M. Kumalo, Drug Discovery for Mycobacterium tuberculosis Using Structure-Based Computer-Aided Drug Design Approach, Int. J. Mol. Sci. 22 (2021). https://doi.org/10.3390/ijms222413259.

[83] E.M. Bruch, S. Petrella, M. Bellinzoni, Structure-Based Drug Design for Tuberculosis: Challenges Still Ahead, Appl. Sci. 10 (2020). https://doi.org/10.3390/app10124248.

[84] T. Fujita, Recent Success Stories Leading to Commercializable Bioactive Compounds with the Aid of Traditional QSAR Procedures, Quant. Struct. Relationships. 16 (1997) 107–112. https://doi.org/https://doi.org/10.1002/qsar.19970160202.

[85] Q. Gao, L. Yang, Y. Zhu, Pharmacophore Based Drug Design Approach as a Practical Process in Drug Discovery, Curr. Comput. Aided. Drug Des. 6 (2010) 37–49. https://doi.org/http://dx.doi.org/10.2174/157340910790980151.

[86] S. Sardari, M. Dezfulian, Cheminformatics in Anti-Infective Agents Discovery, Mini-Reviews Med. Chem. 7 (2007) 181–189. https://doi.org/http://dx.doi.org/10.2174/138955707779802633.

[87] A. Naidu, S.S. Nayak, S. Lulu S, V. Sundararajan, Advances in computational frameworks in the fight against TB: The way forward, Front. Pharmacol. 14 (2023). https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1152915.

[88] S. Ahmed, P. A.E., A.K. and Saxena, Molecular docking-based

interactions in QSAR studies on Mycobacterium tuberculosis ATP synthase inhibitors, SAR QSAR Environ. Res. 33 (2022) 289–305. https://doi.org/10.1080/1062936X.2022.2066175.

[89] S.K. Wahan, P. Shelly, C. Pooja A., G. and Bhargava, A Computational Study on the Structural Prediction of InhA Inhibitors as Antimycobacterial Agents, Polycycl. Aromat. Compd. 44 (2024) 4545–4565. https://doi.org/10.1080/10406638.2023.2254901.

[90] A.F. Kouassi, M. Kone, M. Keita, A. Esmel, E. Megnassan, Y.T. N'Guessan, V. Frecer, S. Miertus, Computer-Aided Design of Orally Bioavailable Pyrrolidine Carboxamide Inhibitors of Enoyl-Acyl Carrier Protein Reductase of Mycobacterium tuberculosis with Favorable Pharmacokinetic Profiles, Int. J. Mol. Sci. 16 (2015) 29744–29771. https://doi.org/10.3390/ijms161226196.

[91] S. Zheng, Y. Gu, Y. Gu, Y. Zhao, L. Li, M. Wang, R. Jiang, X. Yu, T. Chen, J. Li, Machine learning–enabled virtual screening indicates the anti-tuberculosis activity of aldoxorubicin and quarfloxin with verification by molecular docking, molecular dynamics simulations, and biological evaluations, Brief. Bioinform. 26 (2025) bbae696. https://doi.org/10.1093/bib/bbae696.

[92] R. V Chikhale, H.T.M. Abdelghani, H. Deka, A.D. Pawar, P.C. Patil, S. Bhowmick, Machine learning assisted methods for the identification of low toxicity inhibitors of Enoyl-Acyl Carrier Protein Reductase (InhA), Comput. Biol. Chem. 110 (2024) 108034. https://doi.org/https://doi.org/10.1016/j.compbiolchem.2024.108034.

Chapter 2

Theoretical and Methodological Overview

2.1. Introduction

Because it is far faster to sketch a molecule on a computer than to synthesise, purify, and characterise a molecule in a lab, computational chemistry allows scientists to study a vast, diverse range of chemical space. Computational chemistry is the use of computer modeling and simulation to explore the structures and characteristics of molecules and materials. This includes ab initio techniques based on quantum chemistry and empirical approaches. Also called cheminformatics, this field enables scientists to accurately analyze quantum data sets too complex to study without the aid of computational chemistry software [1]. Several computational chemistry techniques are employed in drug design and discovery to compute and predict events, including the drug's binding to its target and the chemical properties needed to design potential drug candidate.

2.2. Molecular representation

Everyone who works with molecules, whether they are chemists or not, must learn how to describe or represent molecules in a machine-readable manner, as this is essential for computational chemistry. These representations convert the chemical and structural data of molecules into machine-readable forms that computer programs can process effectively [2,3].

A molecular structure can be represented in one dimension (1D), two dimensions (2D), or three dimensions (3D). Simple 1D representation includes molecular formulas and nomenclature. However, it is not enough to define a molecule. Thus, a molecular

graph is a more effective 2D representation of a molecule. A graph is typically made up of nodes that are connected together by edges. Similarly, a 2D molecular graph describes a molecule's 2D topological structure, with nodes/atoms connected by edges/bonds. This representation is exemplified by MDL Information Systems, MDL SDF (structure data file), Mol2, and others.

A molecular graph can be represented and communicated using linear notation. At present, the Simplified Molecular-Input Line-Entry System (SMILES) and the IUPAC Chemical Identifier (InChI) are the most often used linear notations. The goal of the InChI is to give chemical structures a unique, or canonical, identifier. In contrast, SMILES strings are frequently used to store and exchange chemical structures, but there is currently no standard to generate a canonical SMILES string [4,5]. Cheminformatics programmers can utilise SMILES to program databases, chemical data entry systems for programs, and computerised mechanisms for the exchange of chemical research. For example, the SMILES and InChI notations of isoniazid are shown below.

SMILES - C1=CN=CC=C1C(=O)NN

InChI=1S/C6H7N3O/c7-9-6(10)5-1-3-8-4-2-5/h1-4H,7H2,(H,9,10)

As the last few decades have shown, understanding how molecules interact with their surroundings usually requires more than just graph-oriented representations. A molecule's steric and electronic characteristics are determined by the spatial arrangement of its atoms, which create its three-dimensional structures or conformations. A 3D depiction of a molecule illustrates how the atoms are geometrically arranged in space. 3D molecule geometry has a significant impact on

molecular binding, as evidenced by ligand-protein interactions and packing in crystal structures [6]. The conversion from 2D to 3D molecular representation entails capturing the structural complexity of molecules. Standard file formats such as PDB, CIF, and Mol2 include 3D information on the substance.

2.3. Molecular modeling

Molecular modeling is the process of using computers to simulate the behaviour of molecules in chemical or biological systems. The fundamental idea is to make as accurate predictions about the behaviour of chemical systems as possible by using approximate mathematical models. Molecular modeling enables scientists to visualise molecules using computers by numerically representing molecular structures and simulating their behaviour using quantum and classical physics equations. This allows them to find new lead compounds for drugs or improve existing drugs in silico. The two primary approaches to molecular modeling are molecular mechanics (MM)) and quantum mechanics (QM) [7].

Molecular mechanics is an empirical approach for calculating molecular parameters such as molecular geometry, strain energy, dipole moment, and vibrational frequencies. MM simulates molecular systems using classical mechanics. This approach is based on the atomic scale, generally stated as the ball and spring model, and it is based on a series of empirical equations that incorporate Hooke's and Coulomb's laws, as well as Newton's second law to derive velocities [8]. This approach works under the premise that the potential energy, which is dependent on molecular geometry, can be used to define a molecular set of atoms.

In molecular mechanics, force fields are used to compute the potential energy of every system. The following force fields are continuously applied: OPLS, SYBYL, TraPPE, UFF, CHARMM, CFF/CVFF, DREIDING, ECEPP, GROMOS, MM3, MM4, MMFF, MACROMODEL, AMBER, BMS, and others [9]. Bonded terms (Ebonded) and non-bonded terms (Enon-bonded) are included in the basic function form of force filed. The former comprises torsion angle term (Edihedral), bonding stretching term (Ebond), angle bending term (Eangle), and out-of-plane bending term (Eout-of-plane), while the latter contains Van der Waals (EVdW) and electrostatic forces (Eelectrostatic) [10].

The goal of quantum molecular modeling is to simulate molecule behaviour at the quantum level. In QM, the Schrödinger equation is solved to get the precise or approximate potential energy surface of the molecule. However, because of the enormous amount of computing involved, these methods are only applicable to very tiny numbers of atoms and electrons. This approach employs density functional theory, ab initio techniques, and semi-empirical techniques [11]. Molecular modeling quickly grabbed the interest of Medicinal Chemistry experts, given the relevance of molecular structure in understanding the mechanism of action and developing bioactive drugs. Even though molecular modeling is a broad field, the three most popular computational modeling components - molecular docking, MD simulation and ADMET modeling - have been essential in making it simple to identify leads for experimental in vitro and in vivo testing [12].

2.4. Molecular descriptors

The physical and chemical properties of the molecules are quantitatively described by the numerical values of molecular descriptors. Molecular descriptors can be defined as "the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." They play an important role in pharmaceutical science, chemistry, toxicology, environmental protection policy, health research, ecotoxicology, and quality control [13]. Several criteria can be used to classify molecular descriptors.

Classification based on the origin of descriptors

a) **Experimental descriptors**: experimental measurements, including polarizability, log P, molar refractivity, dipole moment, and additive physico-chemical properties in general.

b) **Calculated or theoretical descriptors:** These are obtained from a symbol of the molecule and can be further categorised based on the various kinds of molecular representations [14].

Classification based on the dimensionality of structure representation

a) **0D-descriptors:** constitutional descriptors, count descriptors.

b) **1D-descriptors:** list of structural fragments, fingerprints.

c) **2D-descriptors:** also known as topological descriptors, invariant properties are obtained from the molecular connectivity graph.

d) **3D-descriptors:** Quantum-chemical descriptors, WHIM descriptors, GETAWAY descriptors, sizes, steric, surface, and volume descriptors, and 3D-MoRSE descriptors. Also known as geometrical descriptors.

e) **4D-descriptors:** those obtained using the CoMFA or GRID techniques, Volsurf.

Classification based on described object

a) **Global descriptors:** explaining the whole molecule, including its topological indices, dipole moment, molecular volume, and molecular surface, etc.

b) **Local descriptors:** describing specific atoms or molecular fragments (including bond polarizabilities, atomic charges, CATS, and ISIDA descriptors, etc.)

c) **Field descriptors:** describing the molecular fields (electrostatic potential, COMFA descriptors, etc.) that surround the molecule.

Molecular structures are used to generate molecular descriptors. Chemical structures can be downloaded from various databases like ChEMBL, PubChem, Zinc, etc., or they can be drawn using programs like ChemSketch, Chemcraft, etc. After that, the molecules were

optimised with various computer programs. These optimized structures were then entered into various descriptor-generation programs, such as PaDEL, RDKit, Ochem, and others. By following these procedures, we can produce hundreds of descriptors.

2.5. Molecular fingerprints

Molecular fingerprints are used to represent chemical features (structural, physicochemical, etc.) of large-scale chemical sets at a minimal computing cost. Molecular fingerprints are often represented as bit strings, however in cheminformatics; any vector of numerical values can be used as a fingerprint. Using these fingerprints, molecules can be represented by noting their presence or absence, or by counting the frequency of specific features or substructures, like labelled paths in the 2D graph of bonds, being present [15]. In practice, the presence of subgraph features in a molecular graph produces integer vectors, most commonly binary (0,1) vectors, which combine to form the molecular fingerprint. Molecular fingerprints are a powerful and widely used method for similarity-based virtual screening [16]. Molecular fingerprints are classified into the following groups:

2.5.1. Types of molecular fingerprints

a) Topological fingerprints

The linear paths of molecular features, typically atoms, up to a specific number of connecting bonds are captured by topological fingerprints. Examples include Topological torsion, daylight fingerprints[17], and atom pairs [18].

b) Structural keys

The presence (1) or absence (0) of predefined functional groups, substructure motifs, or fragments is represented by each distinct bit position in structural key fingerprints. Mini Fingerprint (MFP), Barnard Chemistry Information (BCI) fingerprints, PubChem, Molecular ACCess System (MACCS), and SMIles Fingerprint (SMIFP) are examples [19].

c) Circular fingerprints

To generate circular fingerprints, the "circular" environment of each atom up to certain "radius" or "diameter" is determined. Common types of circular fingerprints include Molprint2D, Molprint3D, extended-connectivity fingerprints (ECFPs), and functional-class fingerprints (FCFPs) [16].

d) Pharmacophore fingerprints

Bit strings that encode the distances between sets of three (or four) pharmacophoric points in a ligand structure are known as pharmacophoric fingerprints. These distances are measured in distance-binning at the 3D level and in bonds at the 2D level, respectively. PharmPrint (3-point PP) and 4-point PP are two popular varieties of pharmacophore fingerprint (PP) [20].

e) Hybrid fingerprints

Hybrid fingerprints consistently improve search performance by combining preferred bit subsets from their parent fingerprints and highlighting compound class-specific features

during similarity searching. Unity 2D is an example of a hybrid fingerprint [21].

2.6. Feature Selection Method

Quantitative structure-activity relationship (QSAR) modeling is an alternate strategy for selecting lead molecules that uses information from reference active and inactive compounds. This method necessitates the use of high-quality molecular descriptors that accurately reflect the molecular properties causing the pertinent molecular activity. In modern QSAR analysis, feature selection plays a crucial role [22]. Additionally, feature selection is essential to building accurate and comprehensible prediction models. The risk of overfitting may be decreased, and significant features with significant property relationships in the data can be found. A model is said to be overfit if it learns too much from the training data, including irrelevant information (such as noise or outliers). Moreover, feature selection can lessen multicollinearity in feature sets, or linearly correlated features, which are frequently seen in molecular descriptors and can lead to unstable model coefficients [23]. Numerous factors influence a QSAR model's predictive ability. Various statistical techniques can assess whether a data set exhibit linear or nonlinear behavior. Conversely, feature selection methods are used to reduce the complexity of the model, lower the risk of overfitting or overtraining, and choose the most significant descriptors from the frequently over thousand computed [24].

Various algorithms are used for feature selection and are classified into three main categories: Filter, Wrapper, and Embedded

methods [25]. The filter method ranks features according to how well they correlate with the class in a variety of statistical tests. Features that score higher than a predetermined threshold are chosen, while those that score lower are eliminated. A subset of features can be provided as an input to the selected classifier algorithm after it has been chosen. Filter-based methods can be applied to a range of learning algorithms and are computationally efficient [26]. Wrapper approaches use greedy search algorithms to examine all potential feature combinations and pick the one that delivers the best result for a certain machine learning algorithm. However, because of the repeated learning processes and cross-validation, they are computationally more expensive than filter approaches [27]. Embedded approaches balance computational economy and performance by integrating feature selection with the model training procedure [28]. A list of examples from these three groups is provided in **Fig.2.1**.

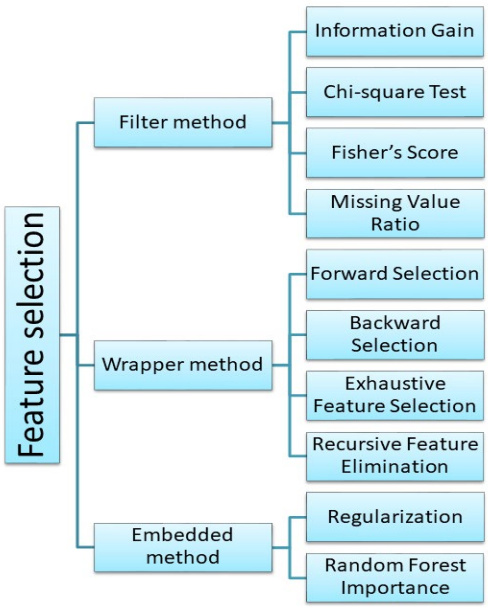


Fig. 2.1. Classification of feature selection methods

2.7. Molecular similarity

The term "similarity" in chemistry describes the common physicochemical properties, biological activity, composition, spatial arrangement, structures, and functionalities of various chemical compounds, biological systems, and macromolecular complexes, among other things [29]. It is predicated on the idea that molecules with similar structures frequently have similar physicochemical properties and biological activities. Molecular fingerprints or molecular descriptors represent the properties of small molecules and use bit string comparisons to compute how similar they are. The concept of molecular similarity is widely used in CADD. It can be used to replace undesirable functional groups as well as to identify new active compounds or compounds with particular properties (virtual screening) [17,30].

2.7.1. Similarity metrics

In cheminformatics, quantifying the similarity between two molecules is a fundamental idea and a common task. The Tanimoto coefficient is the most widely used metric for comparing two compounds' chemical distances or similarity when binary vectors represent molecules with bits representing structural features' presence or absence [31]. The Tanimoto coefficient is defined as the number of bit sets in both individual bitstrings, normalized by the number of bit sets in common. This corresponds to a score ranging from 0.0 to 1.0 for no to perfect similarities. **Equation 2.1** provides the formula for calculating the Tanimoto similarity coefficient [32]:

$$T(a,b) = \frac{c}{a+b-c} \quad \textbf{(2.1)}$$

For the i^{th} and j^{th} molecules, the variables a and b stand for the number of fragment bits, respectively, and c for the number of fragment bits that are shared by both molecules. T=0, when there is no shared property between two substances. The T value rises in proportion to the amount of common properties between the two molecules. T=1 indicates that all characteristics between the two molecules are precisely the same when two sets of properties are identical. This means that every pair of molecules has a T similarity value that is always between 0 and 1.

2.8.Diversity analysis

Evaluation of chemical diversity is essential in drug discovery when it's important to explore new areas of the medicinally relevant chemical space or to maintain the harmony between novelty and diversity [33]. Molecular diversity and similarity metrics work in tandem in chemoinformatics: Clustering methods on chemical databases are based on molecular similarity, which offers a straightforward and widely used method for virtual screening. Likewise molecular diversity analysis underpins numerous strategies for compound selection and design by examining how molecules cover a specific structural space [34]. Physicochemical properties, structural fingerprints, and molecular scaffolds are commonly used to assess the diversity of the molecular libraries. Drug development and discovery benefit greatly from chemical diversity because various molecules can affect biological systems in different ways [35].

2.9. Chemical space analysis

Chemical space is a broad conceptual framework with several uses that deals with the diversity of molecules. Large and ultra-large multidimensional chemical spaces are linked to the substantial rise in the number of compounds that could be created and exist, as well as the growing number of experimental and computational descriptors that are emerging to encode the molecules' molecular structure and/or property aspects. The chemical space covered by commercial, in-house, virtual, and public compound collections has been analyzed and visualized for various purposes, including diversity analysis, data mining, virtual screening, library design, in silico property profiling, screening campaign prioritization, and compound collection acquisition. Chemical space visualisation techniques make it easy to project all compounds into a lower-dimensional space that can be visually analysed. Two important visualisation techniques are Principal Components Analysis (PCA) and Self-organising Maps (SOMs) [36].

2.9.1. Principal Components Analysis

PCA reduces the number of dimensions in vast data sets to principal components that preserve the majority of the original information. It accomplishes this by reducing potentially correlated variables to a smaller group of variables known as principal components (PCs). PCs are new variables created by combining or mixing the original variables linearly. PCA relies on linear algebraic ideas such as Singular Value Decomposition (SVD), covariance matrices, eigenvalues, and eigenvectors [37].

2.9.2. Self-Organizing Maps

SOMs, also known as Kohonen Map, is a sort of Artificial Neural Network that was motivated by biological models of systems in the brain from the 1970s. A Kohonen network may be utilised for analysing data in high-dimensional domains by projecting it onto a two-dimensional plane. This method, which was trained using an unsupervised machine learning approach, is capable of classifying various input types using data samples into cluster groups that share a number of features in a neurobiological-type approach. SOMs have a wide range of uses in cheminformatics research. SOMs cover a wide range of drug discovery topics, including scaffold-hopping, repurposing, and screening library design. Topology-preserving maps of small molecules can be efficiently generated using SOMs, allowing for the comparison of compounds and the evaluation of similarity. Chemical occurrences are projected by SOMs into a two-dimensional space from a multidimensional space (variables-molecular descriptors). The relationship between objects' similarity is preserved in this projection [38].

2.10. Activity landscape modeling

Activity landscape modeling is an effective technique for analysing structure-activity relationships quantitatively. The activity landscape is a structure-activity similarity (SAS) map that examines the structural similarity and activity difference of compounds to find the activity cliffs (ACs). In general, ACs are groups or pairs of structurally similar substances that exhibit significant potency differences but are

active against the same target [39]. ACs create discontinuities in the SAR landscape, which can have significant effects on the performance of lead optimization programs. In drug discovery, medicinal chemists can benefit from the discontinuities indicated by the activity cliffs. However, they also represent anomalies that could influence computational techniques like the building of predictive models and the choice of similarity search queries [40]. Details are provided in Chapter 4.

2.11. Virtual screening

Virtual screening is the most used approach in computer-aided drug discovery development. It has been demonstrated to be the most effective alternative for high-throughput screening. Virtual screening is a set of computational approaches that analyze vast databases or collections of substances to discover novel drug candidates against disease targets [41]. Virtual screening of molecular databases with high structural novelty could boost the probability of expanding the medicinal chemistry relevant space. Machine learning (ML) techniques have recently been investigated as ligand-based virtual screening tools to aid in the discovery of new drug leads. When many actives and inactives are known, which may occur after a high-throughput screening round, pattern recognition or ML techniques can be applied to develop a model that differentiates between the actives and inactives [42]. There are two main categories of virtual screening techniques: (a) ligand-based methods, which rely on the similarity of the compounds of interest with active compounds; and (b) structure-based methods, which

focus on the complementarity of the compounds of interest with the target protein's binding site [43]. **Fig.2.2** shows the different types of ligand based and structure based VS.

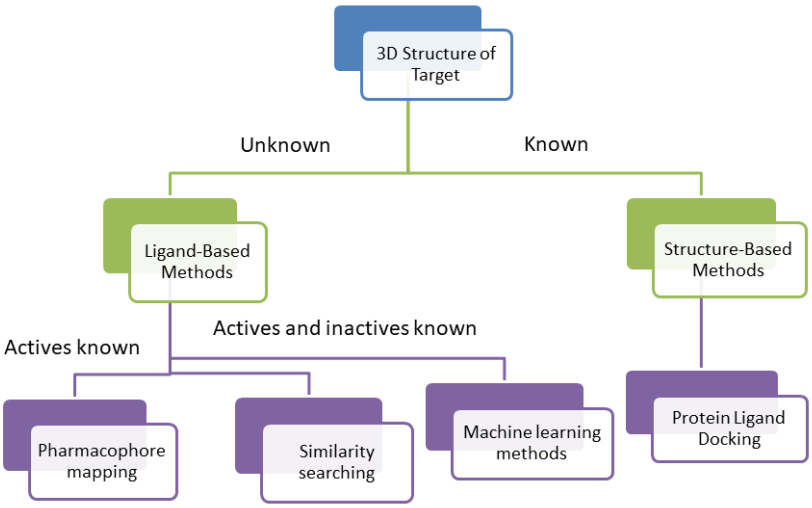


Fig.2.2. Classification of ligand based and structure based virtual screening

2.12. Quantitative Structure Activity Relationship

The most effective VS technique is quantitative structure–activity relationship (QSAR) analysis because of its quick and high throughput and good hit rate. A QSAR is a mathematical model that connects a quantitative numerical measure of chemical structure (physicochemical property) to a biological effect or a physical property (pharmacokinetic property) [44]. As an in silico technique, QSAR modeling aids in prioritising a large number of chemicals according to their desired biological activities, which greatly lowers the number of candidate chemicals that need to be tested in vivo experiments. **Equation 2.2** shows how biological activity in QSAR is represented as a function of these molecular descriptors [45].

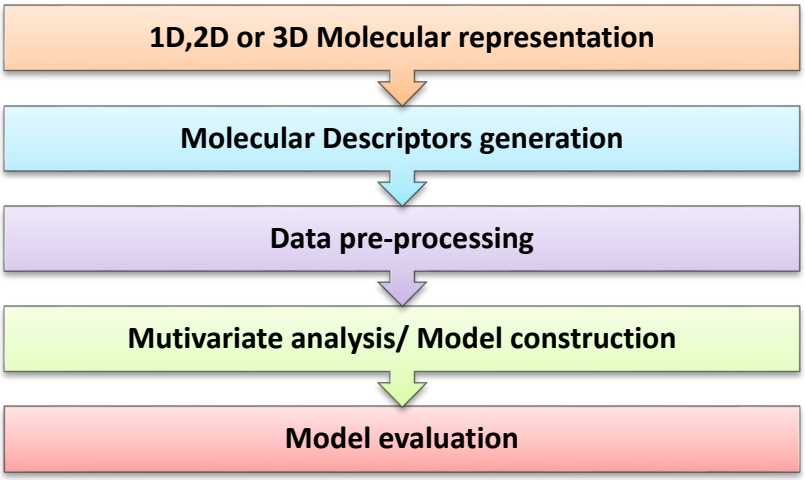


Fig. 2.3. A flowchart of the steps involved in developing a QSAR model

Biological activity = f (molecular descriptors) + error **(2.2)**

The model built based on the biological activities of known compounds is then used to predict the activity of unknown or novel compounds. QSAR model development consists of two major steps: (i) describing the molecular structure and (ii) performing multivariate analysis to correlate molecular descriptors with observed activities/properties [46]. **Fig. 2.3** displays a schematic representation of the QSAR process**.**

2.12.1 Types of QSAR

Based on predictor variables, QSAR are divided into six classes. [23]:

1) **1D-QSAR:** Molecular activity is correlated with molecular parameters such as pKa, log P, and so on.
2) **2D-QSAR**: Connecting structural 2D patterns, such as connectivity indices and 2D pharmacophores, with activity.

3) **3D-QSAR:** Activity is correlated with non-covalent interaction fields around the molecules.
4) **4D-QSAR:** In 3D-QSAR, an ensemble of ligand configurations is included as well.
5) **5D-QSAR:** 4D-QSAR explicitly represents several induced-fit models.
6) **6D-QSAR:** additionally, using various solvation models in 5D-QSAR.

2.13. Machine learning modeling

As chemical "big" data from combinatorial synthesis and HTS continues to grow rapidly, ML has emerged as a crucial tool for drug designers looking to mine chemical information from massive compound databases and discover drugs with significant biological properties [47]. ML methods are very useful in the field of QSAR to develop prediction models from datasets of chemical compounds and the corresponding biological activity. Three of the most popular forms of ML are supervised learning, unsupervised learning, and reinforcement learning. These are further subdivided into subcategories, which are shown in **Fig. 2.4**. Supervised learning predicts future outcomes based on previously labelled data, unsupervised learning finds patterns in data without labels to categorise future outcomes, and reinforcement learning improves decision-making over time through trial and error [48].

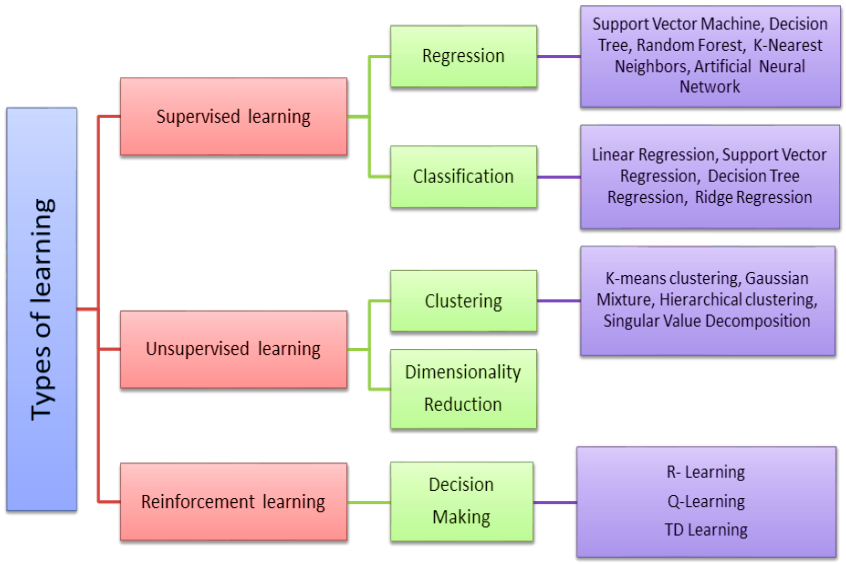


Fig. 2.4. Types of machine learning

QSAR uses labeled datasets to train supervised learning models, with the input features representing chemical structures and the output labels denoting corresponding biological activities, toxicity, or other properties. The development of QSAR prediction models is primarily based on traditional machine learning algorithms like support vector machines, neural networks, random forests, partial least squares regression, and decision trees, as well as molecular descriptors or fingerprints as characteristics of compounds [49,50]. We have employed classification algorithms for QSAR modeling in this study.

2.13.1. Model Performance Evaluation

In ML, performance evaluation quantifies how well a trained model performs on particular model evaluation metrics. We frequently use metrics like accuracy, precision, recall, F1 score, and confusion matrix to assess a classification model's performance. The percentage

of all correct classifications, whether positive or negative, is known as accuracy [51]. Precision indicates the proportion of predicted positive cases that are actually positive. Recall represents the proportion of true positives identified out of all positive instances. F1-score balances the trade-off between precision and recall by combining them into a single metric. They use the formulae provided in **Equations 2.3 - 2.6**. **Fig. 2.5** illustrates the confusion matrix, which compares the predicted and actual values for a dataset to evaluate the effectiveness of classification models in machine learning.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \textbf{(2.3)}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \textbf{(2.4)}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \textbf{(2.5)}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \textbf{(2.6)}$$

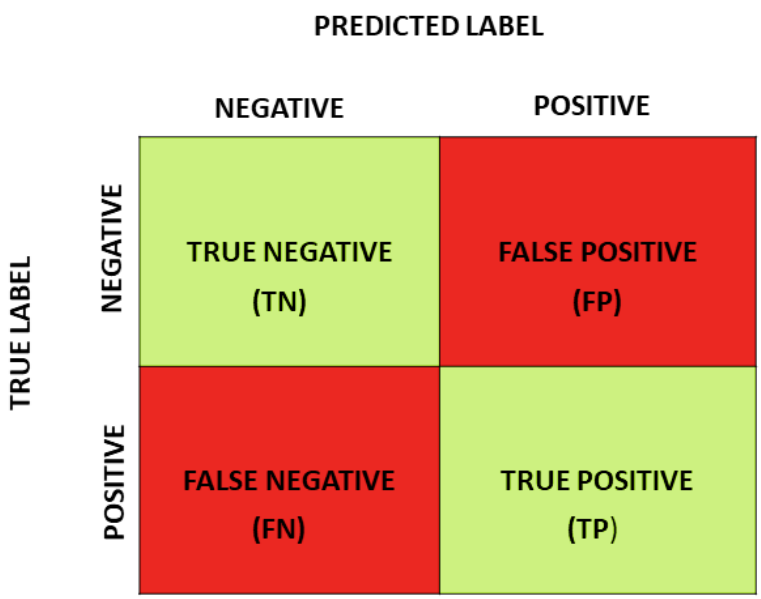


Fig. 2.5. Confusion matrix

2.14. Molecular docking

Molecular docking is a method for predicting the type of interactions and binding affinities that occur when a protein or enzyme interacts with ligands or small molecules or when two or more molecular structures fit together [52]. This structure-based method necessitates a high-resolution three-dimensional representation of the target protein, which can be acquired using methods such as nuclear magnetic resonance spectroscopy, cryo-electron microscopy, or X-ray crystallography [53]. Docking also needs that the ligand's 3D structure be tested in order to predict its binding affinity to the target. Selecting the most promising binding sites prior to molecular docking is crucial for improving docking efficiency. In the drug design and discovery process, docking technology is successfully used for three

primary purposes: (1) predicting the binding mode of a known active ligand, (2) predicting the binding affinities of allied compounds from a known active ligands and (3) screening novel ligands using VS.

2.14.1. Theory of Docking

Two steps are taken to implement this prediction: In the first step, the searching algorithm looks through the conformational space and displays potential locations for the molecule to be coupled to the target; in the second step, scoring determines the energy values required to execute the coupling in each of the potential linking sites. The most promising energy values for binding are believed to be the lower ones. The experimental binding mode should ideally be reproducible through sampling algorithms and should also be ranked highest among all generated conformations by the scoring function [54].

2.14.2. Search / Sampling Algorithms

The search algorithm in any docking tool should generate an optimal number of ways to combine two molecules. Typically, this is accomplished by looking for every possible orientation of ligand conformers that can "fit" one or more receptor-binding pockets. There are numerous popular techniques for searching conformational space in docking algorithms. Search techniques may be classified depending on the flexibility of the ligand or receptor/protein during docking analysis. Examples include Monte Carlo methods, systematic search methods, random or stochastic methods, and simulation approaches [55].

2.14.3. Scoring Functions

Scoring functions are mathematical calculations that estimate the binding affinity of two molecules following docking. The scoring function can be used to predict binding affinity, identify potential drug leads for a specific protein target, and determine the ligand's binding site and mode. The majority of docking software utilises one of three scoring functions: force-field, empirical, or knowledge-based [56].

2.15. Molecular dynamics simulation

The best-docked protein-ligand complexes are commonly evaluated by molecular dynamics (MD) simulations, which are crucial to drug development, to validate their binding. This method helps to reproduce real-time biological phenomena like macromolecular dynamics on a computer platform, allowing us to comprehend the fold and conformational changes in the protein-ligand complex [57]. In MD simulations, molecular mechanics biomolecular force fields define the forces between atoms and the system's potential energy. A force field is the result of combining a mathematical formula with related factors to explain a protein's energy as a function of its atomic coordinates. Several force fields have been constructed for MD simulations, including GROMOS, AMBER, OPLS-AA, and CHARMM [58].

The result of MD simulations called molecular dynamics trajectories. A careful investigation of simulation trajectories is the first step towards determining the parameters required for strong and stable interactions. Common post-simulation assessments include root mean square deviations (RMSD), root mean square fluctuations (RMSF),

radius of gyration (Rg), solvent-accessible surface area (SASA), and hydrogen bond [59]. Chapter 6 goes into further detail.

2.16. Software

2.16.1. Computational software

1. Gaussian 16

Gaussian is a computational chemistry software used for modeling and molecular analysis. It was released in 1970 as Gaussian 70 by John Pople and his Carnegie-Mellon University research team. Gaussian uses the fundamental laws of quantum mechanics to predict the energies, molecular structures, vibrational frequencies, and molecular properties of compounds and reactions in a variety of chemical environments. Apart from molecular mechanics, Gaussian provides a range of ab initio and semiempirical quantum chemistry methods to predict molecular energies, structures, spectroscopic information (such as NMR, IR, UV, and more), and much more [60].

2. Open Babel 2. 4.1

Open Babel is a chemical toolbox designed to communicate with chemical data in multiple languages. It's a collaborative, open project that lets anyone search, convert, analyse, or store data from solid-state materials, biochemistry, chemistry, molecular modeling, or related fields [61]. A solution to the proliferation of different chemical file formats is offered by Open Babel. Since Open Babel 2.3 was released, it has supported 111 different chemical file formats. This program has been used to convert file types throughout this study.

3. DataWarrior

DataWarrior is an open-source, interactive program for data analysis and visualization that combines cutting-edge and tried-and-true chemoinformatics methods in one setting [62]. DataWarrior has been utilised by research teams in academia, government, and business since it was made available to the public in 2014. DataWarrior enables interactive exploration of chemical space, activity landscapes, and activity cliffs**.**

4. PUMA

The Platform for Unified Molecular Analysis (PUMA) incorporates metrics such as chemical space visualisation, scaffold content, and chemical diversity analysis that are used to describe compound databases [63]. The platform connects two publicly available web resources: Activity Landscape Plotter, which examines structure-activity correlations, and Consensus Diversity Plots, which evaluates global diversity. In this work, we used PUMA to analyse the chemical space of Mtb inhibitors.

5. AutoDockTools

AutoDockTools (ADT) is the graphical user interface for configuring and running AutoDock. Researchers interested in computational docking can use the AutoDock Tool, which combines a free, open-source solution with accuracy in identifying the binding pose of a small chemical in a corresponding receptor pocket [64]. Additionally, it can be helpful for generating grids, calculating

hydrogen bonds, displaying secondary structure ribbons, and computing molecular surfaces.

6. CB-Dock2

CB-Dock uses the cutting-edge docking program AutoDock Vina to predict the binding regions of a given protein, computes the centres and sizes using a curvature-based cavity detection technique, and docks. The updated docking server, CB-Dock2, is a particularly effective and user-friendly tool for the cheminformatics and bioinformatics communities. It redesigned the input and output web interfaces and included a highly automatic docking pipeline [65]. The web server is freely available at https://cadd.labshare.cn/cb-dock2/.

7. SiBioLead

SiBioLead is a virtual biocomputing platform designed for molecular dynamic simulations and docking. SiBioLead uses the simulation software GROMACS. This web-based platform operationalizes a group of technical steps, including preprocessing, energy minimisation, equilibration, production dynamics, and trajectory analysis. The web-tool is available at (https://sibiolead.com/).

8. Anaconda python

Anaconda is an open-source data science and artificial intelligence distribution platform for Python and R programming languages. Anaconda includes important Python packages, tools such as Jupyter and RStudio, and the Conda package management. This makes package management and deployment easier for data scientists,

machine learning experts, and scientific computing professionals. We utilised Python and Jupyter Notebook for work analysis.

9. Google Colab

Google Colab, also known as Colaboratory, is a free cloud-based platform that lets users write and run Python code together in a Jupyter Notebook setting. Colab is an amazing platform for data scientists to carry out ML and Deep Learning projects with cloud storage. We used Google Colab to perform our ML-based QSAR modeling and screening.

10. PaDEL-Descriptor

PaDEL-Descriptor is a software tool for computing molecular descriptors and fingerprints [66]. The software presently calculates 1875 descriptors (1444 1D, 2D, and 431 3D descriptors) and 12 kinds of fingerprints (a total of 16092 bits). The Chemistry Development Kit is the primary tool used to calculate these fingerprints and descriptors (RDKit). A set of machine learning and cheminformatics tools written in Python and C++ is called the RDKit.

2.16.2. Visualization software

1. GaussView 6.1

Gaussview offers a graphical user interface for the Gaussian software. It facilitates the creation of Gaussian input files, removes the need for command line instructions by allowing the user to perform Gaussian computations from a graphical interface, and aids in the interpretation of Gaussian output.

1. LIGPLOT

The LIGPLOT tool converts standard Protein Data Bank file input into schematic 2-D representations of protein-ligand interactions [67]. It is operated via an easy Java interface that allows plots to be edited on-screen using mouse click-and-drag actions. The result is a colour or black-and-white PostScript file that depicts intermolecular interactions and their strengths, such as hydrogen bonds, hydrophobic interactions, and atom accessibilities. LIGPLOT was utilised in this investigation to visualise the docked ligand-protein complexes.

1. Discovery Studio Visualizer

BIOVIA Discovery Studio (BDS) Visualizer is a free molecular modeling program with several features for viewing, sharing, and analysing protein and small molecule data. The visualisation tools in BDS are intended to assist researchers in comprehending and communicating the results of molecular simulations, docking studies, and other molecular investigations. We utilised this program to visualize the protein-ligand complexes formed from the docking analysis.

2.17. Computer power

All calculations in this thesis were performed on the following computers: 1) Lenovo Thinkstation with processor Intel®Xeon®CPU E5-2660 v3 @2.60 GHz and 32.0 GB RAM, 2) FUJITSU with processor Intel®Xeon®CPU E5-2640 v4 @2.40 GHz and 32.0 GB RAM, and 3) ASUS VivoBook with Windows 11 Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz - 1.99 GHz, 64 bit operating system.

References

[1] R.C. Glen, Computational chemistry and cheminformatics: an essay on the future, J. Comput. Aided. Mol. Des. 26 (2012) 47–49. https://doi.org/10.1007/s10822-011-9501-6.

[2] T.-H. Nguyen-Vo, P. Teesdale-Spittle, J. Harvey, B. Nguyen, Molecular representations in bio-cheminformatics, Memetic Comput. 16 (2024) 519–536. https://doi.org/10.1007/s12293-024-00414-6.

[3] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, WIREs Comput. Mol. Sci. 12 (2022) e1603. https://doi.org/https://doi.org/10.1002/wcms.1603.

[4] N.M. O'Boyle, Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI, J. Cheminform. 4 (2012) 22. https://doi.org/10.1186/1758-2946-4-22.

[5] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, J. Cheminform. 7 (2015) 23. https://doi.org/10.1186/s13321-015-0068-4.

[6] E.L. Willighagen, Three-dimensional (3D) molecular representations, in: Handb. Chemoinformatics Algorithms, 2010: pp. 65–87. https://doi.org/10.1201/9781420082999.

[7] O. Silakari, P.K. Singh, Chapter 1 - Fundamentals of molecular modeling, in: O. Silakari, P.K.B.T.-C. and E.P. of M. and I. in D.D. Singh (Eds.), Academic Press, 2021: pp. 1–27. https://doi.org/https://doi.org/10.1016/B978-0-12-820546-4.00001-5.

[8] L.A. Silva, T.G. Garrot, A.M. Pereira, J.C.G. Correia, Historical perspective and bibliometric analysis of molecular modeling applied in mineral flotation systems, Miner. Eng. 170 (2021) 107062.

https://doi.org/https://doi.org/10.1016/j.mineng.2021.107062.

[9] G. Liu, X. Yang, H. Zhong, Molecular design of flotation collectors: A recent progress, Adv. Colloid Interface Sci. 246 (2017) 181–195. https://doi.org/https://doi.org/10.1016/j.cis.2017.05.008.

[10] A. Abdelrasoul, H. Zhang, C.-H. Cheng, H. Doan, Applications of molecular simulations for separation and adsorption in zeolites, Microporous Mesoporous Mater. 242 (2017) 294–348. https://doi.org/https://doi.org/10.1016/j.micromeso.2017.01.038.

[11] S. Vishwakarma, Molecular mechanics and quantum mechanics, 2021. https://doi.org/10.13140/RG.2.2.31742.72000.

[12] T.I. Adelusi, A.-Q.K. Oyedele, I.D. Boyenle, A.T. Ogunlana, R.O. Adeyemi, C.D. Ukachi, M.O. Idris, O.T. Olaoba, I.O. Adedotun, O.E. Kolawole, Y. Xiaoxing, M. Abdul-Hammed, Molecular modeling in drug discovery, Informatics Med. Unlocked. 29 (2022) 100880. https://doi.org/https://doi.org/10.1016/j.imu.2022.100880.

[13] D. Bajusz, A. Rácz, K. Héberger, 3.14 - Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching, in: S. Chackalamannil, D. Rotella, S.E.B.T.-C.M.C.I.I.I. Ward (Eds.), Elsevier, Oxford, 2017: pp. 329–378. https://doi.org/https://doi.org/10.1016/B978-0-12-409547-2.12345-5.

[14] A. Mauri, V. Consonni, R. Todeschini, Molecular Descriptors BT - Handbook of Computational Chemistry, in: J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, M. K. Shukla (Eds.), Springer International Publishing, Cham, 2017: pp. 2065–2093. https://doi.org/10.1007/978-3-319-27282-5_51.

[15] L. Pattanaik, C.W. Coley, Molecular Representation: Going Long on Fingerprints, Chem. 6 (2020) 1204–1207. https://doi.org/https://doi.org/10.1016/j.chempr.2020.05.002.

[16] J. Yang, Y. Cai, K. Zhao, H. Xie, X. Chen, Concepts and applications of chemical fingerprint for hit and lead screening, Drug Discov. Today. 27 (2022) 103356.

https://doi.org/https://doi.org/10.1016/j.drudis.2022.103356.

[17] I. Muegge, P. and Mukherjee, An overview of molecular fingerprint similarity search in virtual screening, Expert Opin. Drug Discov. 11 (2016) 137–148. https://doi.org/10.1517/17460441.2016.1117070.

[18] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, J. Chem. Inf. Comput. Sci. 25 (1985) 64–73. https://doi.org/10.1021/ci00046a002.

[19] K. López-Pérez, J.F. Avellaneda-Tamayo, L. Chen, E. López-López, K.E. Juárez-Mercado, J.L. Medina-Franco, R.A. Miranda-Quintana, Molecular similarity: Theory, applications, and perspectives, Artif. Intell. Chem. 2 (2024) 100077. https://doi.org/https://doi.org/10.1016/j.aichem.2024.100077.

[20] A.F.A. Moumbock, J. Li, P. Mishra, M. Gao, S. Günther, Current computational methods for predicting protein interactions of natural products, Comput. Struct. Biotechnol. J. 17 (2019) 1367–1376. https://doi.org/https://doi.org/10.1016/j.csbj.2019.08.008.

[21] B. Nisius, J. Bajorath, Molecular Fingerprint Recombination: Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types, ChemMedChem. 4 (2009) 1859–1863. https://doi.org/https://doi.org/10.1002/cmdc.200900243.

[22] A.E. Comesana, T.T. Huntington, C.D. Scown, K.E. Niemeyer, V.H. Rapp, A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties, Fuel. 321 (2022) 123836. https://doi.org/https://doi.org/10.1016/j.fuel.2022.123836.

[23] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, Drug Discov. Today. 21 (2016) 1291–1302. https://doi.org/https://doi.org/10.1016/j.drudis.2016.06.013.

[24] M. Goodarzi, B. Dejaegher, Y. Vander Heyden, Feature Selection Methods in QSAR Studies, J. AOAC Int. 95 (2012) 636–651. https://doi.org/10.5740/jaoacint.SGE_Goodarzi.

[25] X. Cheng, A Comprehensive Study of Feature Selection Techniques in Machine Learning Models, Insights Comput. Signals Syst. 1 (2024) 65–78. https://doi.org/10.70088/xpf2b276.

[26] N. Sánchez-Maroño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter Methods for Feature Selection – A Comparative Study, 2007. https://doi.org/10.1007/978-3-540-77226-2_19.

[27] N. El Aboudi, L. Benhlima, Review on wrapper feature selection approaches, in: 2016 Int. Conf. Eng. MIS, 2016: pp. 1–5. https://doi.org/10.1109/ICEMIS.2016.7745366.

[28] H. Hamla, K. Ghanem, Comparative Study of Embedded Feature Selection Methods on Microarray Data BT - Artificial Intelligence Applications and Innovations, in: I. Maglogiannis, J. Macintyre, L. Iliadis (Eds.), Springer International Publishing, Cham, 2021: pp. 69–77.

[29] A. Bender, R.C. Glen, Molecular similarity: a key technique in molecular informatics, Org. Biomol. Chem. 2 (2004) 3204–3218. https://doi.org/10.1039/B409813G.

[30] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening, Methods. 71 (2015) 58–63. https://doi.org/https://doi.org/10.1016/j.ymeth.2014.08.005.

[31] D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?, J. Cheminform. 7 (2015) 20. https://doi.org/10.1186/s13321-015-0069-3.

[32] M.A. Fligner, V. Joseph S, P.E. and Blower, A Modification of the Jaccard–Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings, Technometrics. 44 (2002) 110–119. https://doi.org/10.1198/004017002317375064.

[33] D.C. Spellmeyer, P.D.J. Grootenhuis, Chapter 28. Recent Developments in Molecular Diversity: Computational Approaches to Combinatorial Chemistry, in: A.M.B.T.-A.R. in M.C. Doherty (Ed.), Academic Press, 1999: pp. 287–296. https://doi.org/https://doi.org/10.1016/S0065-7743(08)60590-4.

[34] A.G. Maldonado, J.P. Doucet, M. Petitjean, B.-T. Fan, Molecular similarity and diversity in chemoinformatics: From theory to applications, Mol. Divers. 10 (2006) 39–79. https://doi.org/10.1007/s11030-006-8697-1.

[35] D. Gorse, A. Rees, M. Kaczorek, R. Lahana, Molecular diversity and its analysis, Drug Discov. Today. 4 (1999) 257–264. https://doi.org/https://doi.org/10.1016/S1359-6446(99)01334-3.

[36] G. Das, M. Chattopadhyay, S. Gupta, A Comparison of Self-organising Maps and Principal Components Analysis, Int. J. Mark. Res. 58 (2016). https://doi.org/10.2501/IJMR-2016-039.

[37] S. Mishra, U. Sarkar, S. Taraphder, S. Datta, D. Swain, R. Saikhom, S. Panda, M. Laishram, Principal Component Analysis, Int. J. Livest. Res. (2017) 1. https://doi.org/10.5455/ijlr.20170415115235.

[38] D. Digles, G.F. Ecker, Self-Organizing Maps for In Silico Screening and Data Visualization, Mol. Inform. 30 (2011) 838–846. https://doi.org/https://doi.org/10.1002/minf.201100082.

[39] D. Stumpfe, H. Hu, J. Bajorath, Evolving Concept of Activity Cliffs, ACS Omega. 4 (2019) 14360–14368. https://doi.org/10.1021/acsomega.9b02221.

[40] D. Stumpfe, J. Bajorath, Exploring Activity Cliffs in Medicinal Chemistry, J. Med. Chem. 55 (2012) 2932–2942. https://doi.org/10.1021/jm201706b.

[41] C.G. Wermuth, B. Villoutreix, S. Grisoni, A. Olivier, J.-P. Rocher, Chapter 4 - Strategies in the Search for New Lead Compounds or Original Working Hypotheses, in: C.G. Wermuth, D. Aldous, P. Raboisson, D.B.T.-T.P. of M.C. (Fourth E. Rognan (Eds.), Academic Press, San Diego, 2015: pp. 73–99. https://doi.org/https://doi.org/10.1016/B978-0-12-417205-0.00004-3.

[42] P.P. Parvatikar, S. Patil, K. Khaparkhuntikar, S. Patil, P.K. Singh, R. Sahana, R. V Kulkarni, A. V Raghu, Artificial intelligence: Machine learning approach for screening large database and drug discovery, Antiviral Res. 220 (2023) 105740. https://doi.org/https://doi.org/10.1016/j.antiviral.2023.105740.

[43] A. Gimeno, M.J. Ojeda-Montes, S. Tomás-Hernández, A. Cereto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas, S. Garcia-Vallvé, The Light and Dark Sides of Virtual Screening: What Is There to Know?, Int. J. Mol. Sci. 20 (2019). https://doi.org/10.3390/ijms20061375.

[44] B.J. Neves, R.C. Braga, C.C. Melo-Filho, J.T. Moreira-Filho, E.N. Muratov, C.H. Andrade, QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery, Front. Pharmacol. Volume 9- (2018). https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2018.01275.

[45] S. Clarancia, J.K. Dhanjal, V. Malik, N. Radhakrishnan, J. Mannu, D. Sundar, Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications, in: Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., 2018. https://doi.org/10.1016/B978-0-12-809633-8.20197-0.

[46] S. Kausar, A.O. Falcao, An automated framework for QSAR model building, J. Cheminform. 10 (2018) 1. https://doi.org/10.1186/s13321-017-0256-5.

[47] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today. 23 (2018) 1538–1546. https://doi.org/https://doi.org/10.1016/j.drudis.2018.05.010.

[48] I.H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Comput. Sci. 2 (2021) 160. https://doi.org/10.1007/s42979-021-00592-x.

[49] J. Mao, J. Akhtar, X. Zhang, L. Sun, S. Guan, X. Li, G. Chen, J. Liu, H.-N. Jeon, M.S. Kim, K.T. No, G. Wang, Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models, IScience. 24 (2021) 103052. https://doi.org/https://doi.org/10.1016/j.isci.2021.103052.

[50] T.A. Soares, A. Nunes-Alves, A. Mazzolari, F. Ruggiu, G.-W. Wei, K. Merz, The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods, J. Chem. Inf. Model. 62 (2022) 5317–5320. https://doi.org/10.1021/acs.jcim.2c01422.

[51] H. Dalianis, Evaluation Metrics and Evaluation BT - Clinical Text Mining: Secondary Use of Electronic Patient Records, in: H. Dalianis (Ed.), Springer International Publishing, Cham, 2018: pp. 45–53. https://doi.org/10.1007/978-3-319-78503-5_6.

[52] L. Pinzi, G. Rastelli, Molecular Docking: Shifting Paradigms in Drug Discovery, Int. J. Mol. Sci. 20 (2019). https://doi.org/10.3390/ijms20184331.

[53] P.C. Agu, C.A. Afiukwa, O.U. Orji, E.M. Ezeh, I.H. Ofoke, C.O. Ogbu, E.I. Ugwuja, P.M. Aja, Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management, Sci. Rep. 13 (2023) 13398. https://doi.org/10.1038/s41598-023-40160-2.

[54] N. Boggula, S. Katta, S. Mahaboobi, J. Megavath, R. Mukherjee, R. Tadikonda, Molecular Docking -An Overview, 8 (2023) 24–34.

[55] U. Yadava, Search algorithms and scoring methods in protein-ligand docking, Endocrinol. Int. J. 6 (2018). https://doi.org/10.15406/emij.2018.06.00212.

[56] J. Li, A. Fu, L. Zhang, An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking, Interdiscip. Sci. Comput. Life Sci. 11 (2019) 320–328. https://doi.org/10.1007/s12539-019-00327-w.

[57] S. Sasidharan, V. Gosu, T. Tripathi, P. Saudagar, Molecular Dynamics Simulation to Study Protein Conformation and Ligand Interaction BT - Protein Folding Dynamics and Stability: Experimental and Computational Methods, in: P. Saudagar, T. Tripathi (Eds.), Springer Nature Singapore, Singapore, 2023: pp. 107–127. https://doi.org/10.1007/978-981-99-2079-2_6.

[58] J. Mustali, I. Yasuda, Y. Hirano, K. Yasuoka, A. Gautieri, N. Arai, Unsupervised deep learning for molecular dynamics simulations: a novel analysis of protein–ligand interactions in SARS-CoV-2 Mpro, RSC Adv. 13 (2023) 34249–34261. https://doi.org/10.1039/D3RA06375E.

[59] S.K. Paul, M. Saddam, K.A. Rahaman, J.-G. Choi, S.-S. Lee, M. Hasan, Molecular modeling, molecular dynamics simulation, and

essential dynamics analysis of grancalcin: An upregulated biomarker in experimental autoimmune encephalomyelitis mice, Heliyon. 8 (2022) e11232. https://doi.org/https://doi.org/10.1016/j.heliyon.2022.e11232.

[60] B.J. Duke, B. O'Leary, The Gaussian programs as a teaching tool: A case study on molecular hydrogen calculations, J. Chem. Educ. 69 (1992) 529. https://doi.org/10.1021/ed069p529.

[61] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminform. 3 (2011) 33. https://doi.org/10.1186/1758-2946-3-33.

[62] E. López-López, N. J. Jesús, J.L. and Medina-Franco, DataWarrior: an evaluation of the open-source drug discovery tool, Expert Opin. Drug Discov. 14 (2019) 335–341. https://doi.org/10.1080/17460441.2019.1581170.

[63] M. González-Medina, J.L. Medina-Franco, Platform for Unified Molecular Analysis: PUMA, J. Chem. Inf. Model. 57 (2017) 1735–1740. https://doi.org/10.1021/acs.jcim.7b00253.

[64] N. El-Hachem, B. Haibe-Kains, A. Khalil, F.H. Kobeissy, G. Nemer, AutoDock and AutoDockTools for Protein-Ligand Docking: Beta-Site Amyloid Precursor Protein Cleaving Enzyme 1(BACE1) as a Case Study BT - Neuroproteomics: Methods and Protocols, in: F.H. Kobeissy, J. Stevens Stanley M. (Eds.), Springer New York, New York, NY, 2017: pp. 391–403. https://doi.org/10.1007/978-1-4939-6952-4_20.

[65] Y. Liu, X. Yang, J. Gan, S. Chen, Z.-X. Xiao, Y. Cao, CB-Dock2: improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting, Nucleic Acids Res. 50 (2022) W159–W164. https://doi.org/10.1093/nar/gkac394.

[66] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474. https://doi.org/https://doi.org/10.1002/jcc.21707.

[67] A.C. Wallace, R.A. Laskowski, J.M. Thornton, LIGPLOT: aprogram to generate schematic diagrams of protein-ligand interactions, Protein Eng. Des. Sel. 8 (1995) 127–134. https://doi.org/10.1093/protein/8.2.127.

Chapter 3

Unleashing the Potential of Cheminformatic Analysis for Mycobacterium Tuberculosis Inhibitors: Insights into Chemical Space and Structural Diversity

3.1. Introduction

Mycobacterium tuberculosis (Mtb), the pathogen responsible for tuberculosis (TB), is a major global health concern, causing significant morbidity and mortality [1]. Despite decades of research and efforts to combat TB, it remains a serious threat, with the emergence of drug-resistant strains making treatment even more challenging [2, 3]. The urgency for novel tuberculosis drugs is crucial due to the escalating drug resistance issue, requiring innovative treatments with distinct mechanisms [4, 5]. The intricate Mtb biology and the difficulty of finding effective drugs with minimal side effects complicate drug development. Hence, persistent research and investment in new TB treatments remain vital to confront this persistent worldwide health emergency [6].

Mycobacterium tuberculosis possesses a distinctive cell wall and a slow growth rate, rendering the development of effective drugs to combat this bacterium a formidable challenge [7]. Many of the current therapeutic agents for tuberculosis (TB) are designed to target specific elements of the Mtb cell wall, such as mycolic acids or the peptidoglycan layer. Their purpose is to disrupt the cell wall's integrity, thereby impeding bacterial growth [8]. Alternatively, some drugs focus on vital enzymes and pathways within Mtb's metabolism. *e.g.,* they may inhibit the folate biosynthesis pathway, the electron transport chain, or the ATP synthase complex, consequently arresting energy production and cell division [9–11]. Ongoing research actively explores the creation of novel TB drugs, encompassing various compounds with distinct mechanisms of action. This includes exploring inhibitors of

essential enzymes, RNA polymerase, protein synthesis, and bacterial respiration. Significantly, employing combination therapy involving multiple drugs that target diverse elements is crucial for preventing the emergence of drug-resistant strains and enhancing treatment outcomes [12, 13].

Treating TB involves employing a combination of various drugs to avert the emergence of drug-resistant strains [14, 15]. The primary drugs used for TB treatment, referred to as first-line drugs, consist of isoniazid, rifampin, ethambutol, and pyrazinamide. These drugs target different aspects of Mtb's cell wall or metabolism. In instances of ineffectiveness or resistance, second-line drugs are employed. These drugs possess diverse mechanisms of action, enabling them to target the bacterium through distinct pathways [16, 17]. Managing drug-resistant TB (DR-TB) is notably more intricate, necessitating a blend of second-line drugs. However, these drugs often exhibit reduced efficacy and increased toxicity compared to the first-line counterparts [18, 19]. The pursuit of novel and more potent TB treatment options constitutes an active realm of research, with several innovative compounds under exploration to confront the challenge of drug resistance.

The drug discovery process for new antitubercular treatments has evolved with advancements in technology and the availability of biological and chemical information. The recognition and validation of targets, coupled with the use of high-throughput screening technology and software innovations, have made computer-aided drug design (CADD) an essential part of TB pharmaceutical research [20]. With the rapid development of computational power, scientists are now

exploring the vast chemical space, estimated to be around 10^{60} organic molecules smaller than 500 Da, in search of scaffolds with medicinal promise [21–23]. Moreover, the utilization of data mining techniques within this chemical space may aid in the creation of predictive models that are useful for predicting the activity of novel compounds. The significance of analyzing chemical space and employing cheminformatic modeling is growing as the amount and complexity of structural data for Mtb inhibitors rise, since these techniques help to understand and anticipate how inhibitors and Mtb proteins interact. In pursuit of this goal, medicinal chemists employ various computational techniques for data mining and visualization, as well as machine learning algorithms, which were initially developed for computer science [24]. To arrange datasets and evaluate the chemical vicinity of effective hits, conventional statistical and classification methods are also applied [25–28]. Using these methodologies, researchers have effectively characterized the chemical space of inhibitors targeting numerous diseases or biological targets. An obvious example is the work of Jose L. Medina-Franco and colleagues, who have recorded the cheminformatic study of inhibitor chemical space across several decades[29–36]. In the past few years, various computational analyses of Mtb inhibitors have emerged. Nevertheless, comprehensive cheminformatic investigations on the structural diversity and distribution within the chemical space of Mtb inhibitors have been lacking. Therefore, it is of utmost importance to Explore and navigate the chemical space of clinically relevant inhibitors targeting Mtb. It may allow for the examination and listing of compounds in the chemical space of Mtb.

In the current study, we conducted a thorough chemoinformatics analysis of small molecules to characterize the chemical space of Mtb-targeting compounds. To achieve this, we utilized two types of Mtb inhibitors datasets. The first dataset was obtained from the ChEMBL database [37], while the second was gathered from the literature, specifically on the anti-tuberculosis effectiveness of Indian medicinal herbs. Given the rising prevalence of MDR-TB, we focused on investigating the chemical space of compounds derived from Indian medicinal plants, which has shown significant promise in the development of Mtb inhibitors. After collecting the molecules from the literature, the structures were downloaded from the PubChem database [38]. The study includes determining physicochemical properties, structural fingerprint analysis, molecular scaffold analysis, and ligand structural similarity analysis. Since the FDA-approved drug database serves as a common reference compound source in drug discovery campaigns, we conducted a comparison between the chemical space of Mtb inhibitors with the dataset comprising FDA-approved drugs and the Nutraceutical drugs dataset sourced from "DrugBank". Insights gained from this chemical space analysis will help to identify small molecule inhibitors for the treatment of Mycobacterium tuberculosis.

3.2 Materials and Methods

3.2.1 Datasets and Data Curation

The compound datasets used in this study were given in **Table 3.1.** DataWarrior software [39] was used to retrieve the chemical

structure and biological activity of the first dataset of Mtb inhibitors from the ChEMBL_32 database. ChEMBL queries are possible with DataWarrior. This data set named Mtbs, which included 2359 chemicals, was tested for Mtb inhibitory potency. In ChEMBL, bioactivities are commonly listed using units such as Ki, Kd, IC_{50}, and EC_{50}, along with assay details like cell line, organism, or tissue. Nevertheless, in our study, we concentrated solely on Human Mtb protein inhibitors.

The rising global prevalence of MDR-TB has drawn more attention to the development of novel drugs that may effectively treat MDR-TB and shorten the length of therapy. Consequently, there is an urgent need to discover novel anti-tuberculosis drugs that are not only safe and effective but also economically feasible. Medicinal plants have provided significant hope in meeting these demands because they are a proven template for the production of novel drugs [40, 41]. India is one of the countries with a vast abundance of medicinal plants and extensive traditional knowledge regarding the application of herbal medicine to treat a variety of diseases [42, 43]. So, the second data set consisted of data sourced from primary literature, published concerning the anti-tuberculosis effectiveness of Indian medicinal herbs, which had been meticulously extracted and compiled into a comprehensive list through manual effort. The structures of molecules were subsequently downloaded from the PubChem version 1.7.2 beta database. This dataset we called Phytochemicals dataset which included 916 molecules.

To carry out cheminformatic analysis on compound datasets, it is crucial to curate the data properly. We utilized the Microsoft Office

Excel filter tool to accomplish this task. In chemical space analysis, each compound must be unique and distinct. Therefore, we obtained datasets from reliable sources such as ChEMBL and PubChem, which contained SMILES representations of the chemical structures of the compounds. We performed data curation by removing duplicate compounds, compounds lacking activity information, and so on, based on their SMILES. In cases where compounds had the same chemical structure but conflicting bioactivity data, we retained the one with the lowest activity value. Following data curation, we obtained datasets comprising 1223 distinct molecules for Mtbs and 769 for Phytochemicals.

Table 3.1. Data sets analyzed in this study.

Data sets	source	No. unique compounds
Approved drugs (Drugs)	DrugBank	1657
Nutraceuticals	DrugBank	103
Mtb inhibitors I (Mtbs)	ChEMBL	1223
Mtb inhibitors II (Phytochemicals)	PubChem	769

3.2.1.1. *Reference Datasets*

The datasets of Mtb inhibitors were compared to two reference collections: 1657 Approved drugs and 103 Nutraceuticals obtained from DrugBank version 5.1.10 [44]. Nutraceuticals are products that contain bioactive compounds, such as vitamins, minerals, and plant-based substances, that are derived from food sources and are thought to

provide health benefits beyond basic nutrition. some studies suggest that certain plant-based compounds, such as garlic, honey, and propolis, may have antibacterial properties and could potentially be used as adjuvant therapy in combination with antibiotics to enhance their effectiveness and reduce the risk of resistance [45,46]. Exploring the chemical space of the Mtb dataset and comparing it with the nutraceutical dataset is beneficial for a few key reasons. It can help identify nutraceuticals that may have anti-tuberculosis properties, guiding the design of new drugs. Additionally, it may shed light on how these nutraceuticals work against tuberculosis. This can open up avenues for incorporating nutritional strategies into tuberculosis control and help identify combinations of drugs and nutraceuticals that work better together. In light of this, we selected the nutraceuticals as a reference dataset. The study utilized four datasets of compounds: (target in inhibitors set), Phytochemicals (target inhibitors set), Nutraceuticals (reference dataset), and Drugs (reference dataset).

3.2.2. Molecular Representation

To accurately represent the vast chemical space, various criteria have been developed to capture the unique structural features of molecules. These include pharmaceutical-relevant PCP, different designs of structural fingerprints, and molecular scaffolds. By using these criteria, researchers can better understand and manipulate the properties and behavior of molecules, allowing for the development of new drugs.

3.2.2.1 *Physicochemical Properties*

Physicochemical properties are important in chemical space analysis because they provide a way to quantify and compare the properties of different compounds [47]. By analyzing the physicochemical properties of a set of compounds, researchers can identify regions of chemical space where there may be clusters of compounds with similar properties or where there may be gaps that represent unexplored regions of chemical diversity. Our study focused on six important molecular properties for drug discovery: hydrogen bond donors (HBD), hydrogen-bond acceptors (HBA), partition coefficient octanol/water (AlogP), molecular weight (MW), number of rotatable bonds (RTB), and topological polar surface area (TPSA). These properties were chosen because they capture the key aspects of size (MW), flexibility (RTB), and molecular polarity. To analyze how these properties are distributed, we used violin plots generated with the Python data visualization library Seaborn version 0.11.2 [48] and also calculated summary statistics. Additionally, we performed a principal components analysis (PCA) using Python data analysis and the modeling library Scikit-learn version 0.24.2 [49] to visualize the chemical space based on these six physicochemical properties. This allowed us to obtain a better knowledge of the connection between the molecular properties and to compare compound collections for drug discovery.

3.2.2.2 *Structural Fingerprints*

Molecular fingerprints have gained extensive utilization in drug development and virtual screening to compare the similarity of

compounds and identify potential drug candidates. They are easy to compute, store, and manipulate, making them a convenient way to represent large chemical databases [50]. There are several types of molecular fingerprints available, each with its own strengths and weaknesses. Three different molecular fingerprints were used to calculate the intra-library similarity for all compound pairs within the dataset: Molecular Access System (MACCS) keys (166-bit), Extended Connectivity Fingerprints (ECFP), and PubChem fingerprints (881-bit). The similarity coefficient for fingerprint comparison was the Tanimoto index [51,52]. Cumulative distribution functions (CDF) were used to analyze the distribution of the similarity values.

The Tanimoto similarity coefficient is calculated using the following formula (Equation 1):

$$T(a,b) = \frac{c}{a+b-c} \tag{1}$$

The variables a and b represent the number of fragment bits for the i^{th} and j^{th} molecules, respectively, while c denotes the count of shared fragment bits between both molecules. T=0, when two compounds don't share a single feature. As the number of shared features between the two molecules increases, so does the corresponding T value. When two sets of features are identical, as denoted by T=1, it means that all features match exactly between the two molecules. As a result, the T similarity value for every pair of molecules is in the range 0 and 1.

3.2.2.3 *Molecular Scaffolds*

A molecular scaffold is the core structure of a molecule that remains after the removal of its side chains and functional groups. It represents the fundamental structural framework of a molecule and is often used in drug discovery and chemical diversity analysis. Molecular scaffolds can be used to identify structurally similar compounds that share a common core structure, regardless of their functional groups and side chains. This is useful in the design of new drug candidates, where scaffolds with known biological activity can be modified to optimize their pharmacological properties. The scaffold framework was created by eliminating side chain terminals connected to the ring structure. Murcko and Skeleton scaffolds were used for the analysis. In the case of Murcko scaffolds, the approach involved the removal of the exocyclic double bonds and α-attached atom led to the scaffold generation [53]. Furthermore, the Murcko scaffold was utilized to make a skeleton scaffold wherein the analysis only included the ring, and the heteroatoms were replaced with carbon atoms.

3.2.3 Similarity Analysis

Similarity analysis is a typical approach used in chemistry to analyze the structures of two or more molecules to assess their degree of similarity. It is especially significant in drug discovery, as discovering compounds with similar structures and features can lead to the discovery of novel drugs. The comparison of similarity between the two molecules was calculated by matching SkelSpheres descriptors obtained from the molecular structures. This process was carried out

using DataWarrior Version 5.5.0. This entailed developing a diverse set of conformers.

3.2.4 Chemical Space Analysis

Chemical space analysis is an important part of drug development that involves the exploration and characterization of the wide chemical landscape of possible therapeutic compounds. The chemical space is the multidimensional space of all conceivable chemical compounds, which is expected to include billions of potential therapeutic candidates. Chemical space analysis entails using computer tools and methodologies to systematically explore, visualize, and analyze this space in order to discover molecules with the necessary pharmacological characteristics while minimizing the risk of hazardous consequences. This method enables drug discovery scientists to build and optimize therapeutic candidates by exploring new regions of chemical space, predicting the features of molecules, and prioritizing the most promising leads for further development. Chemical space analysis is an essential component of modern drug discovery and has substantially aided in the development of novel medicines for a wide range of ailments [54]. The evaluation of chemical space involved the analysis of scaffold, structural fingerprint, similarity, and various physicochemical properties. These analysis were performed using PUMA Version 1.0 [55], DataWarrior Version 5.5.0, Seaborn version 0.11.2, and Scikit-learn version 0.24.2 on a Windows 11 Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99 GHz, 64 bit OS system.

3.3. Result and Discussion

3.3.1 Physicochemical Properties

Fig. 3.1., shows the distribution of the six pharmaceutically relevant physicochemical properties described in Methods as violin plots implemented in Seaborn 0.11.2. A violin plot is a hybrid of a box plot and a kernel density plot that shows data peaks. It's used to display the distribution of numerical data. The white dot represents the median. The interquartile range is represented by the broad grey bar in the middle. The thin grey line shows the remainder of the distribution, except for points deemed to be "outliers" using an interquartile range-based method. To represent the data distribution shape, a kernel density estimation is given on either side of the grey line. The wider areas of the violin plot indicate a greater possibility that individuals in the population will take on the given value, while the skinnier sections indicate a lower probability. The statistical distribution of the six estimated physicochemical properties is given in **Table 3.2**.

According to **Fig. 3.1.**, the Phytochemicals dataset contains more HBA, while Mtbs has HBA distribution similar to Drugs and Nutraceuticals (i.e., a median of 4). The median and mean values of Mtbs also show a distribution of HBD that is comparable to that of Phytochemicals, but with a narrower spread, as evidenced by the smaller standard deviation. In comparison to the Drugs, which displayed a median of 2, the Nutraceuticals dataset has slightly more HBD (i.e., a median of 3). Similarly, while the TPSA values for the Drugs and Mtbs datasets are comparable, they are marginally higher and lower than the Drugs for the Nutraceuticals and Phytochemicals

datasets, respectively. These findings show that Mtbs and Phytochemicals are generally less or comparable polar than approved drugs.

Regarding ALogP as a measure of hydrophobicity, the Mtbs and Phytochemicals datasets have comparable values with Drugs while the Nutraceuticals dataset has a slightly lower value than other datasets. As per compound flexibility, measured by RTB, all three datasets were comparable to the Drugs. This finding suggests that the substances in these datasets are as flexible as the drugs in the drug dataset. The Mtbs MW violin plot is relatively short, indicating that compounds in this dataset have a limited MW range. Short boxes show that their data points tend to hover towards the middle values.

As calculated by Seaborn, none of the distributions had a normal distribution. Furthermore, two quantitative statistical tests derived from seaborn, namely skewness and excess kurtosis, can be employed to assess the alterations in normality of the distribution within each dataset. Skewness is a measure of asymmetry or the deviation of the distribution of a given random variable from a symmetric distribution. The peakedness of the distribution is measured by kurtosis. A normal distribution has skewness and excess kurtosis of zero, thus if the distribution is close to those values, it is probably normal. Negative skewness values denote a left-skewed distribution, whereas positive values indicate a right-skewed distribution of the data. Negative kurtosis indicates a "light-tailed" distribution with fewer outliers, whereas a positive kurtosis indicates a distribution that is "heavy-tailed" and has more outliers. According to West et al. (1996), a significant deviation from normality is defined as an absolute skew

value greater than 2.1 and an absolute kurtosis (proper) value greater than 7 [56]. Subtraction of three from the proper kurtosis value yields the "excess" kurtosis. In this study, the majority of the dataset displays moderate skewness. That is, the skewness ranges between -1 and 2. In the majority of the instances, a kurtosis value of less than 5 implies that the distribution has shorter tails and is thinner than a normal distribution, which could indicate a lack of outliers. In light of **Table 3.2** and **Fig. 3.1**, we can conclude that none of the datasets adhere to the assumption of normality. The Drugs and Nutraceuticals datasets display a more pronounced deviation from normality as evidenced by their skewness and kurtosis value. This deviation is even more significant than the range suggested by West et al.

Table 3.2. Statistical summaries of the distribution of each dataset

	MW				**RTB**			
Statistics	Drugs	Mtbs	Nutraceuticals	Phytochemicals	Drugs	Mtbs	Nutraceuticals	Phytochemicals
Mean	359.30	382.58	310.64	292.44	5.75	5.60	6.49	4.69
Median	331.06	374.20	241.13	272.18	5.00	5.00	4.00	4.00
Variance	27884.53	6279.50	51112.29	16760.39	20.86	5.22	48.35	11.79
Std.dev	166.99	79.24	226.08	129.46	4.57	2.29	6.95	3.43
min	46.04	134.05	75.03	58.04	0.00	0.00	0.00	0.00
max	1549.71	822.41	1354.57	709.16	40.00	13.00	44.00	17.00
Range	1503.67	688.35	1279.54	651.12	40.00	13.00	44.00	17.00
Q1	257.11	329.19	151.59	188.09	3.00	4.00	2.00	2.00
Q3	424.20	438.14	400.33	386.14	8.00	7.00	8.50	6.00
Int.Range	167.09	108.95	248.74	198.04	5.00	3.00	6.50	4.00
Skewness	1.74	0.47	2.16	0.60	2.13	0.65	2.66	1.40
Kurtosis	5.56	1.37	5.91	-0.39	8.85	0.42	9.44	1.64

Statistics	**HBA**				**HBD**			
Mean	5.16	4.74	6.18	2.63	2.12	1.37	3.16	1.86
Median	4.00	4.00	4.00	2.00	2.00	1.00	3.00	1.00
Variance	14.37	4.28	27.94	7.92	4.64	0.98	7.03	5.79
Std.dev	3.79	2.07	5.29	2.81	2.15	0.99	2.65	2.41
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
max	30.00	12.00	28.00	20.00	17.00	6.00	15.00	12.00
Range	30.00	12.00	28.00	20.00	17.00	6.00	15.00	12.00
Q1	3.00	3.00	3.00	1.00	1.00	1.00	2.00	0.00
Q3	7.00	6.00	8.00	4.00	3.00	2.00	4.00	2.00
Int.Range	4.00	3.00	5.00	3.00	2.00	1.00	2.00	2.00
Skewness	1.98	0.63	1.95	1.65	2.63	0.69	1.97	1.92
Kurtosis	6.13	0.26	4.19	3.34	10.75	0.78	5.58	3.70
Statistics	**TPSA**				**ALogP**			
Mean	89.96	86.06	110.98	66.08	-0.12	0.32	-1.00	0.30
Median	75.15	79.90	88.77	50.60	0.01	0.22	-1.35	0.36
Variance	4399.02	1117.16	7643.70	3781.64	3.70	2.92	9.35	4.32
Std.dev	66.33	33.42	87.43	61.50	1.92	1.71	3.06	2.08
min	0.00	6.48	0.00	0.00	-10.90	-4.35	-10.06	-6.25
max	563.45	225.60	485.44	321.66	9.87	4.87	12.63	9.61
Range	563.45	219.12	485.44	321.66	20.77	9.23	22.69	15.86
Q1	46.25	62.99	51.48	20.23	-1.13	-1.01	-2.45	-0.98
Q3	113.43	107.04	137.945	86.99	1.10	1.64	-0.27	1.67
Int.Range	67.18	44.05	86.47	66.76	2.24	2.65	2.18	2.65
Skewness	2.20	0.71	1.92	1.66	-0.81	0.02	1.34	0.15
Kurtosis	8.11	0.72	4.52	3.15	4.03	-0.72	5.20	1.43

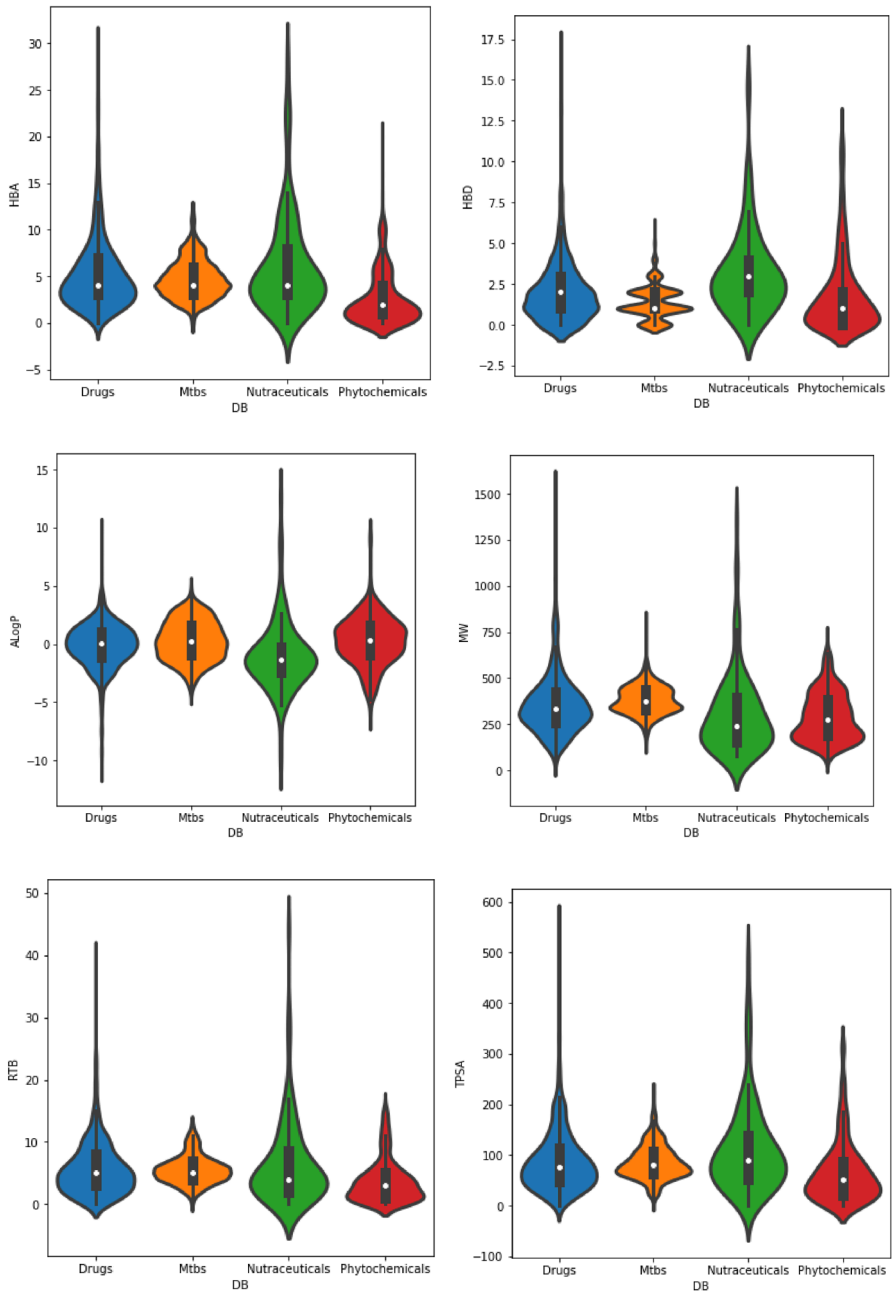


Fig. 3.1. Violin plots of the distribution of six physicochemical properties for the datasets.

Six molecular properties were used to calculate a pairwise comparison of the inter- and intra-distances of dataset compounds using the Euclidean distance. The concept behind this is that when two objects are separated by a greater distance, they are said to be dissimilar, and when they are separated by a smaller distance, they are said to be similar or closer [57]. The most commonly used distance measure is the Euclidean distance, which is the square root of the total of all squared distances between related data points. **Fig. 3.2** shows a distance matrix obtained from the mean inter and intra-database Euclidean distance measurements. The matrix is coloured from grey (low values) to red (high values).

According to these values, the highest inter-distance was found between Nutraceuticals and Phytochemicals datasets, followed by Nutraceuticals and Drugs, and finally between Nutraceuticals and Mtbs. Furthermore, based on the calculated molecular properties, the Nutraceuticals, and Mtbs datasets had the highest and lowest intra-distance diversity, respectively. This result suggests that the Mtbs dataset is less diverse than other datasets. It's interesting to observe that the Drugs dataset and the Mtbs dataset have the smallest inter-dataset distance, indicating that the two datasets share some database properties.

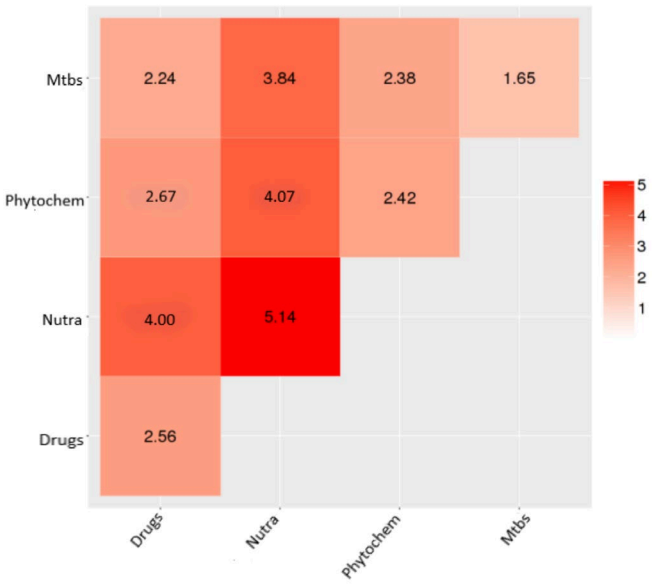


Fig. 3.2. Euclidean distance correlation matrix of datasets created based on six physicochemical properties of pharmaceutical relevance.

3.3.1.1 *Visual representation of the property space*

A principal component analysis (PCA) was performed on the six physicochemical properties. PCA is a statistical method that reduces the dimensionality of a dataset while retaining most of the variability in the data by identifying patterns and correlations among variables [58]. The summary of the covariance percentages for the first four principal components (PCs) are given in **Table 3.3**. These findings reveal that 2D and 3D PCA plots are suitable visual representations of the property space created for these libraries, with the first two PCs retrieving 80.12% of the variation and the first three PCs capturing 89.71% of the variation. In addition, **Table 3.3,** summarizes the loading values of each property for the first four PCs of the four datasets. For the first PC, ALogP had the highest positive loading and TPSA had the highest

negative loading. HBD was the property with the highest positive contribution to the second PC and ALogP had the highest negative contribution, and RTB was the property with the highest positive contribution to the third PC while ALogP had the highest negative contribution.

The 3D and 2D representations of the chemical space scatter plot of the distribution of the six PCP descriptors computed for two sets of Mtb inhibitors (Mtbs and Phytochemicals), as well as two reference datasets (Drugs and Nutraceuticals are shown in **Figs. 3.3** and **3.4**. As anticipated, the visual representation of the chemical space illustrates that the Drugs (**Fig. 3.4.(A)**), and Nutraceuticals (**Fig. 3.4.(C)**) occupy a broad expanse within the property space. This could be a result of two datasets being taken from DrugBank and also having different varieties of structures. Comparable findings have been derived from the aforementioned studies when comparing the property space of Nutraceuticals to that of currently approved drugs. In contrast, the Mtbs covers a more limited area of the space that exists within the property space of the remaining three databases. The compounds in the Phytochemicals dataset also cover a broad area of the property space. Besides that, most of the compounds of this dataset adhere to the property space of the drug database, although a few fall outside of this range. Furthermore, it has been noted that some of the compounds in the Phytochemicals and Nutraceuticals datasets inhabit portions of the chemical space covered by the Mtb dataset. This finding suggests that some compounds derived from nutraceuticals and phytochemicals may have mtb inhibitory activity. The region with higher population density is also occupied by the Drugs and Mtbs dataset. Within the Mtbs

dataset, molecules (data points) expand a wide range of scores along PC1 (x-axis). As noted previously, ALogP is the property that contributes the most to PC1. This aligns with the observations of the distinct distributions of these properties among Mtbs compounds (**Fig. 3.1**). However keep in mind that the visual representation of the chemical space offers only an approximate representation, which is comprehensively represented by the six PCP.

Combining the distribution of the physicochemical properties with the visual representation of the property space, it is possible to draw the general conclusion that all the datasets examined in this study share a resemblance to the property space occupied by drugs. Consequently, this similarity opens the door for further research into these datasets to discover novel compounds with potential therapeutic applications. Moreover, the analysis reveals that certain compounds within the Phytochemical dataset and Nutraceutical dataset occupy areas of the chemical space that remain unexplored by existing drugs. These compounds hold promise for application in virtual screening for therapeutic targets that remain unexplored in terms of biologically active molecules.

Table 3.3. Loadings for the first four principal components (PCs) of the property space of the four datasets.

	PC1	PC2	PC3	PC4
Cumulative eigenvalue (%)	62.59	80.12	89.71	95.13
MW	-0.393	-0.493	-0.273	0.35
TPSA	-0.479	0.109	-0.28	-0.058
RTB	-0.317	-0.562	0.645	-0.382
HBD	-0.424	0.326	-0.252	-0.678
HBA	-0.478	-0.008	-0.107	0.398
ALogP	0.326	-0.569	-0.597	-0.332

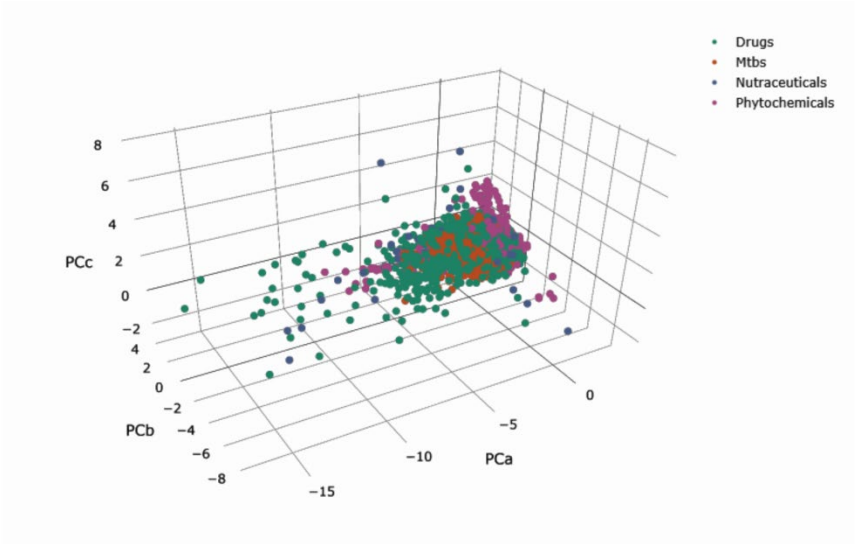


Fig. 3.3. The 3D visual representation of the property space of four datasets produced by principal component analysis (PCA) of six PCP.

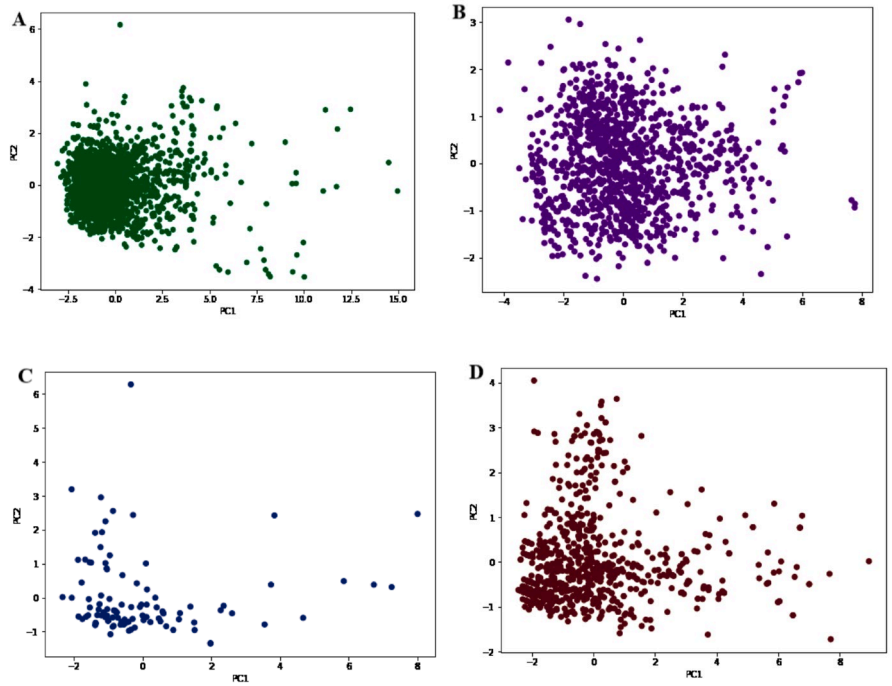


Fig. 3.4. The 2D visual representations of the property space (A) Drugs, (B) Mtbs, (C) Nutraceuticals, and (D) Phytochemicals in the form of PCA plot produced using six PCP.

The examination of the PCP distribution and a visual representation of the chemical space through pharmaceutically relevant features are both significant, but they do not directly yield insights into the specific molecular structure of Mtb inhibitors. The following section discusses the structural properties of Mtb inhibitors.

3.3.2 Fingerprint Diversity

Structural fingerprints allowed for the comparison of datasets based on atom connectivity and chemical structures. It is accomplished by transforming these molecules into a series of binary bits, enabling convenient comparisons between molecules. To assess the extent of structural variability, encompassing side chains, within the datasets, three types of binary molecular fingerprints were used, including extended connectivity fingerprints (ECFP), 166-bit MACCS keys, and 881-bit PubChem. The Tanimoto similarity coefficient was used to calculate structural similarity and generate a similarity matrix. Random samples of 5000 similarity values off the diagonal were extracted for each similarity matrix to compute various statistical metrics including the mean, median, interquartile range, and standard deviation. After that, a CDF analysis was carried out. The CDF of the pairwise intraset similarity is shown in **Fig. 3.5**. These values were computed with the Tanimoto coefficient and three different fingerprints, and summarizes representative statistics of the distributions in **Table 3.4**.

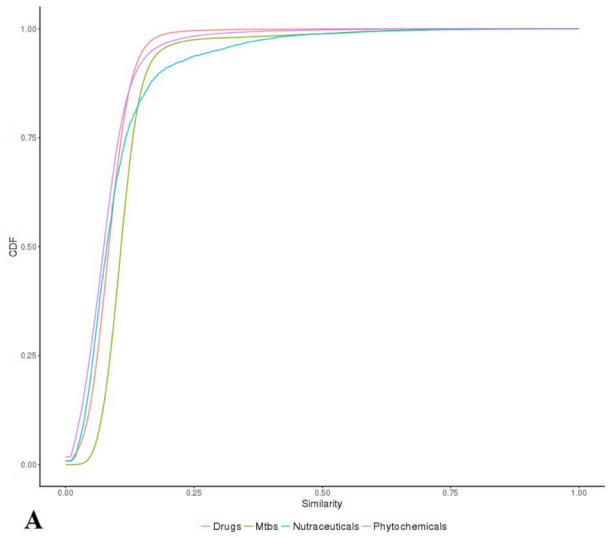
The findings reveal variations in the degree of similarity values across distinct fingerprint representations. In terms of mean similarity

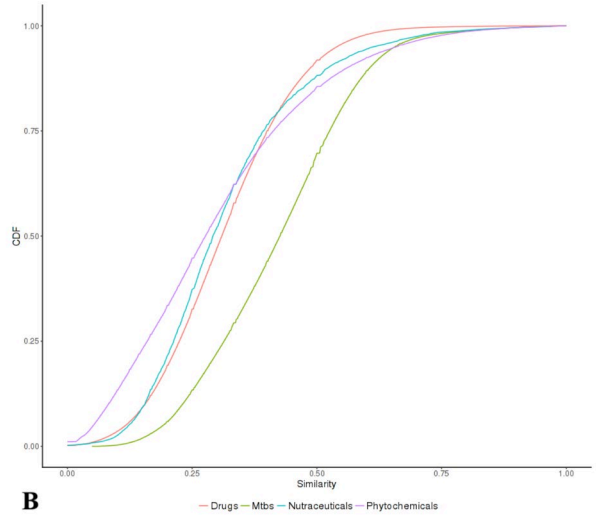
values, PubChem fingerprints exhibited the highest similarity values ranging from 0.28 to 0.47 across various compound databases studied. Following this, MACCS keys exhibited mean similarity ranging from 0.30 to 0.42, and ECFP4 showed the lowest values with mean similarity ranging from 0.08 to 0.12. In other words, the relative ranking of the similarity values for a given dataset followed the order PubChem > MACCS > ECFP4. This outcome underscores that ECFP4 exhibits the lowest similarity values among the three evaluated fingerprints, indicating their heightened resolution. Nevertheless, due to its consistent similarity values across datasets, ECFP4 is ineffective for identifying and classifying datasets based on structural diversity.

According to the mean of the PubChem keys/Tanimoto, the Nutraceuticals data set exhibited the highest level of intra-set similarity within the set, with a mean similarity value of 0.286, followed by Drugs with a mean similarity value of 0.344. This observation could arise from the diverse molecular targets covered by the Nutraceuticals and Drugs datasets collected from DrugBank, where each target possesses its distinct mechanism of action, leading to potential variations in ligand structures. Importantly, there was no apparent correlation noted between the size of the data sets and their structural diversity. Even though the Nutraceuticals dataset has fewer molecules (103) than other data sets, it is more diverse.

Remarkably, the Mtbs data set exhibited the highest intra-set similarity value across all types of fingerprints. The median similarity values were recorded as 0.108, 0.468, and 0.426 for ECFP4, PubChem, and MAACS key fingerprints, respectively. In other words, when compared to all other datasets, this dataset had the least structural

diversity. Moreover, the similarity values distributions indicated that Phytochemical compounds generally possess greater structural diversity than compounds within the Mtbs dataset. The distinct lack of diversity of Mtbs is evident from the CDF versus similarity plot generated from fingerprints, where its curve is notably separate from the others by being distinctly shifted to the right. These findings imply the necessity to generate novel chemical structures as inhibitors of Mtb by exploring a broader chemical space, thus presenting ample opportunities for enhancing novelty. This result also shows that the Phytochemicals dataset could offer a more innovative scaffold with the potential for favorable Mtb inhibitory potency in in-vitro testing.





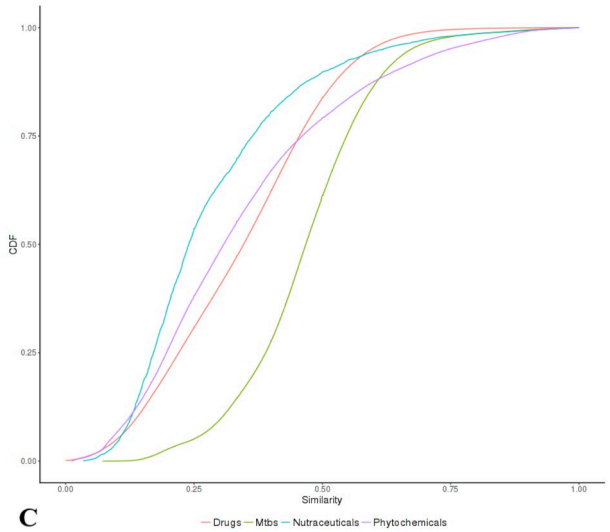


Fig. 3.5. CDF of pairwise Tanimoto similarity values calculated for all datasets (A) ECFP4, (B) MAACS, and (C) PubChem key fingerprints.

Table 3.4. Pairwise similarity distribution statistics calculated using three fingerprints and the Tanimoto coefficient.

fingerprint	dataset	Min	1st Qu	Median	Mean	3rd Qu	Max	Std.Dev
ECFP4	Drugs	0.00	0.06	0.08	0.08	0.10	1.00	0.04
	Mtbs	0.00	0.08	0.10	0.11	0.13	1.00	0.07
	Nutra	0.00	0.05	0.08	0.10	0.11	1.00	0.09
	Phyto chem	0.00	0.04	0.07	0.08	0.10	1.00	0.06
PubChem	Drugs	0.00	0.22	0.34	0.34	0.45	1.00	0.15
	Mtbs	0.06	0.39	0.46	0.46	0.54	1.00	0.13
	Nutra	0.03	0.17	0.24	0.28	0.36	1.00	0.16
	Phyto chem	0.01	0.19	0.30	0.34	0.46	1.00	0.19
MACCS	Drugs	0.00	0.22	0.31	0.31	0.40	1.00	0.13
	Mtbs	0.04	0.31	0.42	0.42	0.52	1.00	0.14
	Nutra	0.00	0.21	0.29	0.31	0.39	1.00	0.15
	Phyto chem	0.00	0.16	0.27	0.30	0.41	1.00	0.18

3.3.3 Scaffold Diversity

A common method for evaluating and comparing the structural diversity of chemical libraries is the use of scaffolds, which capture the core molecular framework of a molecule. In this study, we used Bemis Murcko's scaffold definition to calculate the molecular scaffolds within the Mtbs and Phytochemicals datasets. Here the scaffold is represented by all the ring systems and their connecting linkers. The 1223 known structures from the Mtbs dataset were grouped into 543 unique scaffolds by the Murcko scaffold analysis and the 769 known structures from the phytochemical dataset were grouped into 350 distinct

scaffolds. The scaffolds and their associated frequencies are depicted in Fig. 6A, B. According to the Mtbs and phytochemicals datasets' Murcko scaffold analyses, the scaffolds with the greatest frequencies are 47 and 59, respectively. The Murcko skeletal scaffold was created afterward, yielding 225 datasets for phytochemicals and 425 datasets for Mtbs. Within the Mtbs and phytochemical datasets, respectively, the skeleton scaffold analysis identified scaffolds with the highest frequency at 55 and 114. The scaffold diversity is determined by dividing the number of scaffolds by the total number of molecules [59]. A limited representation is suggested by the diversity analysis of the Murcko, Skeleton, and Singleton scaffold [60, 61] in Mtb inhibitors space (**Table 3.5.**).

Table 3.5. Comparative Analysis of the Scaffold Diversity of Mtbs and Phytochemicals Libraries.

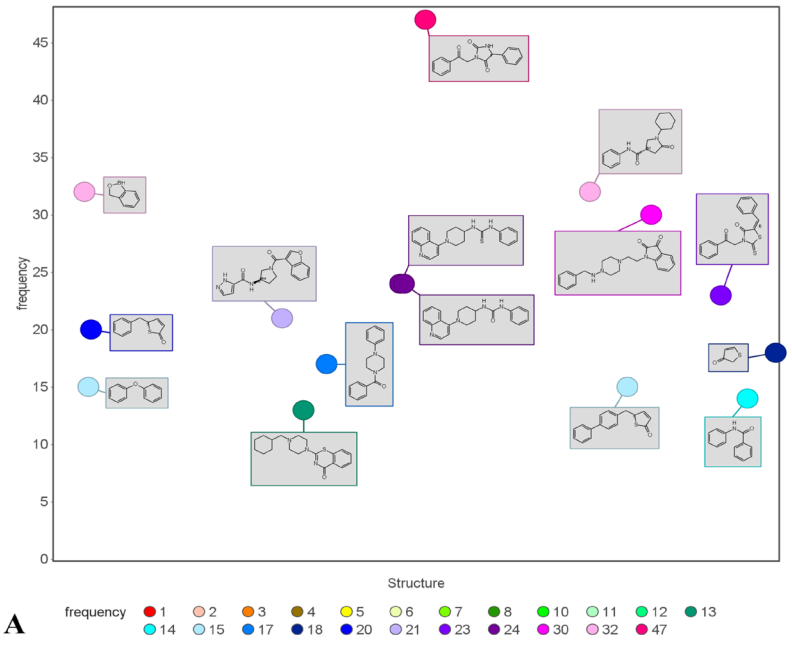
dataset	N	N_c	N_s	N_{sc}	N_{sc}/N	N_c/N	N_s/N	N_s/N_c
Mtbs	1223	543	420	425	0.35	0.44	0.34	0.77
Phytochemicals	769	350	259	225	0.29	0.45	0.33	0.74

N: total number of molecules in Mtbs and Phytochemical datasets; N_c: murcko scaffold; N_s: singleton scaffold, N_{sc}: skeleton scaffold; N_{sc}/N: the ratio of skeleton scaffolds (N_{sc}) to that of total number of molecules (N); N_C/N: ratio of Murcko scaffolds (N_c) and total number of molecules (N); (N_s/N): the ratio of singleton Murcko scaffolds and total number molecules (N); and (N_s/N_c): the ratio of singleton murcko scaffolds (N_s) to murcko scaffolds (N_c)

3.3.3.1 Scaffolds content in the databases

The 3-phenacyl-5-phenylimidazolidine-2,4-dione system is the most common scaffold in the Mtb inhibitor's chemical space, with 47 compounds, followed by the 1-cyclohexyl-5-oxo-N-phenylpyrrolidine-3-carboxamide system, which has 32 compounds. As evidenced by the enormous number of bioactive compounds presently available, the use of heterocyclic scaffolds, many of which incorporate nitrogen, is becoming increasingly significant in the production of therapeutically active pharmaceuticals. Several recent studies showed that imidazolidine and pyrrolidine derivatives have a variety of interesting pharmacological effects on antibacterial strains[62–65]. Due to the fast development of drug resistance, the search for novel antituberculosis drugs is still continuing. In this context, imidazolidine and pyrrolidine derivatives may emerge as a novel class of antimicrobial medicines with low resistance. The benzene ring is the most common scaffold in the Phytochemicals chemical space, accounting for 59 molecules, followed by the cyclohexene system with 17 compounds. However, a basic cyclic system such as benzene plays no role here. Following cyclohexene, the most common scaffold is flavone with frequency 16. Several studies have shown that flavone was the class of flavonoids with substantial antimycobacterial action. Flavones are gaining importance due to their broad biological value and possible therapeutic applications. These flavones are found in several Indian medicinal plants, like Acalypha indica, Allium cepa, Allium sativum, Adhatoda vasica, and Aloe vera. These plants have been tested for antimycobacterial activity. All had effective antimycobacterial action

[66,67]. Furthermore, literature review suggests that more than 100 flavonoids extracted from diverse plants display antimycobacterial, mainly antitubercular, activities. Among them, quercetin, rutin, apigenin, and catechin show significant anti-tuberculosis activity, making them promising candidates for future in vivo studies. Recent research has investigated the feasibility of synthesizing hybrid systems with antibacterial properties using the sol-gel route, which includes silica, polyethylene glycol, and quercetin. Similarly, caffeic acid, a major nutritional antioxidant from the flavonoid family, exhibits comparable potential[68–70]. As a result, in this era of rising drug resistance, compounds derived from Indian plants provide hope for the development of novel drugs with minimal resistance. By doing additional computational and experimental methods, the favored scaffolds discovered in both datasets from this study might be used as a starting point for the rational design of novel Mtb inhibitors.



A

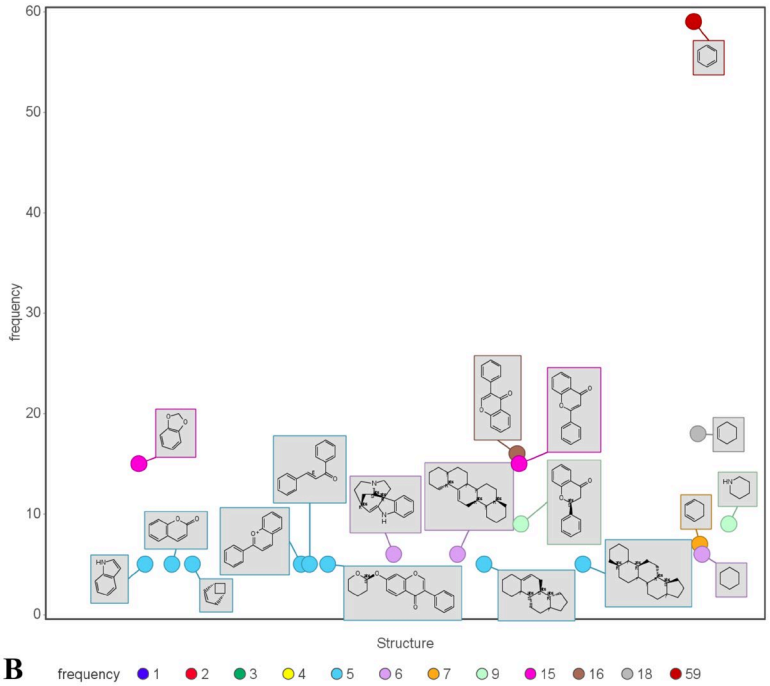


Fig. 3.6. Most favourable scaffold identified in the datasets (A) Mtbs (B) Phytochemicals. For each scaffold, the frequency is indicated in different colours.

3.3.4 similarity analysis

Drug discovery makes extensive use of the concept of similarity, where molecules may be classified based on their biological effects or physical characteristics [61]. Here we analyzed the similarity between the molecules in Mtbs and Phytochemicals datasets. The Rubberbanding Scaling Forcefield (RSF) approach to structural similarity calculation analysis is useful for understanding the chemical space of Mtb inhibitors. The chemical bonds are stretched, twisted, and snapped back to their original position using this approach. Compared to conventional PCA-based techniques, the RSF methodology enhances

similarity analysis. The structural descriptors produced from the three-dimensional structure of the molecules are employed to quantify the similarity of molecules. In our study, we created a SkelSpheres descriptor and used it to find similar molecules in Mtbs and Phytochemical datasets. Taking stereochemistry into account, the SkelSpheres descriptor counted the duplicate fragments and encoded the heteroatoms by computing similarity using this particular descriptor.

The Mtbs dataset generated 2587 pairs with a 99% similarity thr eshold, while the Phytochemicals dataset generated 1098 pairs with a 9 9% similarity threshold. The structure similarity chart (SkelSpheres) is shown in **Fig. 3.7.**, where molecules with a high degree of similarity are linked together by lines and colored according to how similar their structures are. According to their structural characteristics, the majority of the chemical scaffolds in the Mtbs dataset are grouped together, which suggests that there is less chemical diversity in Mtbs than there is in the Phytochemicals dataset. Some structurally similar molecules in both datasets are shown in **Fig. 3.8**. A careful investigation of the similarity space of structurally similar compounds may reveal significant redundancy in the collection of chemical molecules against Mtb targets. In addition, spontaneous mutations may give resistance to several structurally related drugs in M. tuberculosis strains that are rapidly developing drug-resistant. Therefore, it is advisable to include representative molecules from diverse scaffolds in order to expand the existing chemical space of Mtb-targeting drugs.

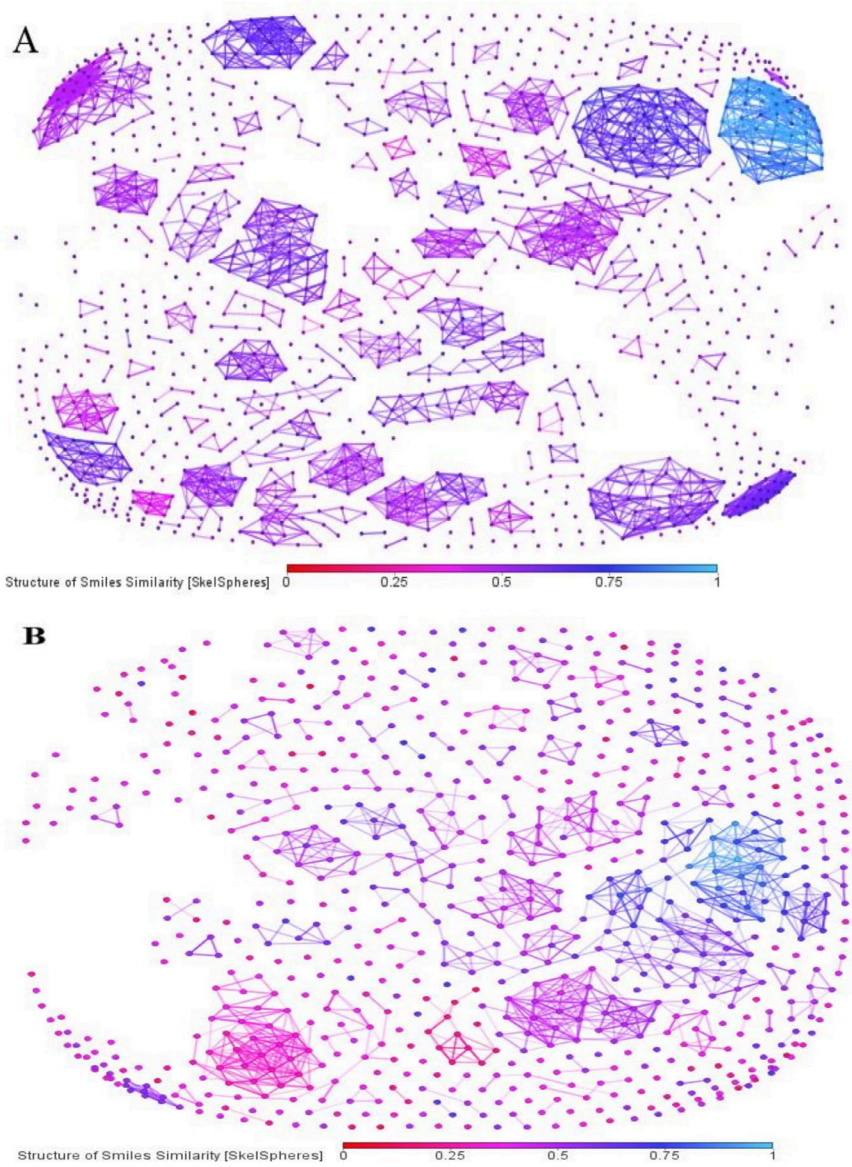


Fig. 3.7. Similarity Skelspheres visualization; similarity is indicated by color. (**A**) Mtbs dataset (**B**) Phytochemicals dataset.

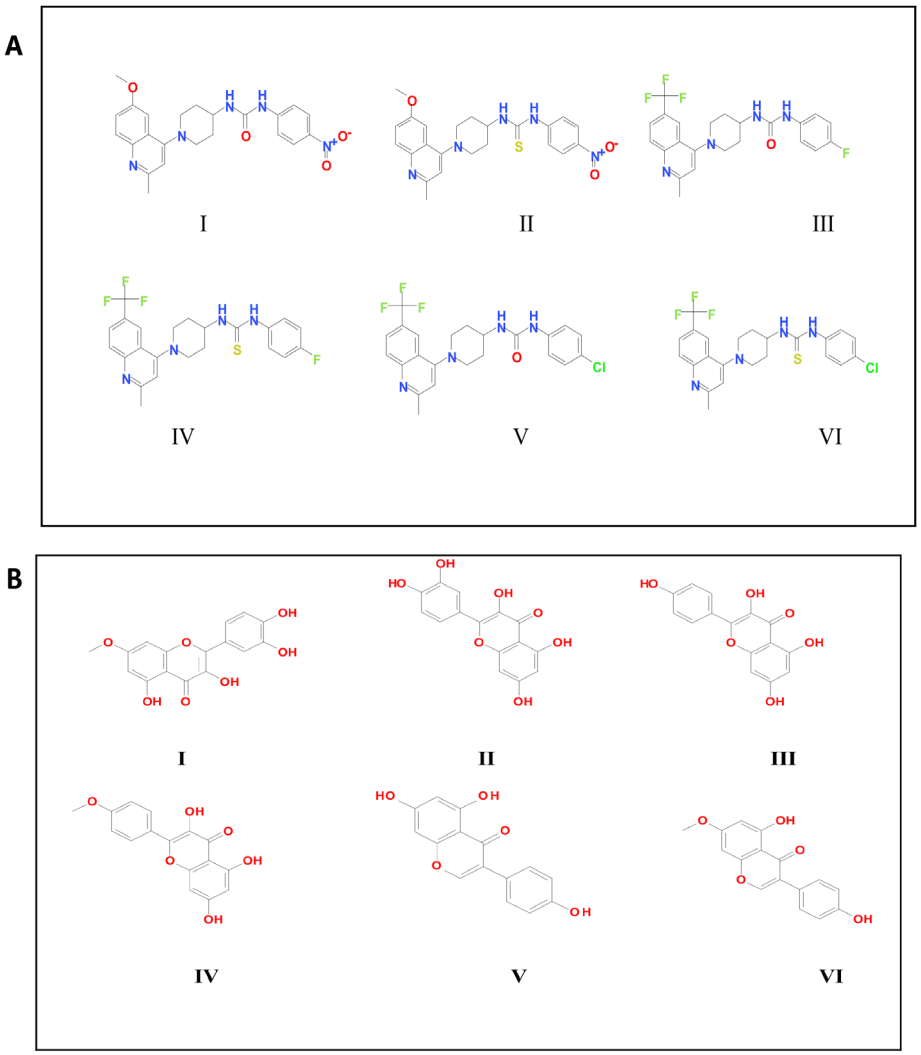


Fig. 3.8. Structurally similar molecules are shown in the figure. (A) Mtb dataset (B) phytochemicals dataset.

3.4. Conclusion

Chemical space analysis is a valuable tool for exploring the vast and diverse chemical space to identify potential new drug candidates against Mtb, especially in light of increasing drug resistance cases. Due to the complex nature of Mtb and its ability to mutate and develop drug

resistance, traditional drug discovery approaches have become less effective. Chemical space analysis allows for the exploration of diverse and novel chemical compounds that may have unique structural and physicochemical properties that can overcome drug resistance. In this study we present a comprehensive assessment of two datasets of Mtb inhibitors, namely Mtbs and Phytochemicals, using diverse criteria including physicochemical properties, scaffolds, fingerprints, and structural similarities. These datasets were compared to approved drugs and Nutraceuticals datasets from the DrugBank database. The analysis of physicochemical property distributions showed that Mtb inhibitors are generally less polar or equally polar compared to Approved drugs, as indicated by the distributions of hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), and total polar surface area (TPSA). Mtb inhibitors were also found to have similar flexibility as measured by rotatable bonds (RTB). Visual representation of the property space revealed that the Mtbs dataset covers a limited area of the space compared to the property space occupied by all other three datasets. Additionally, most of the compounds in the Phytochemicals dataset fall within the property space of the Drugs dataset, but some of the compounds occupy areas of the chemical space that are not yet covered by existing drugs. Some of the compounds in the Phytochemicals and Nutraceuticals datasets inhabit portions of the chemical space covered by the Mtb dataset. This indicates that some compounds derived from nutraceuticals and phytochemicals may have mtb inhibitory activity. The structural diversity of the datasets computed with the PCP descriptors and fingerprint representations revealed that the compounds in the Mtbs dataset are structurally less diverse than all other datasets.

This is evident in the lower intra-dataset distance and higher similarity value for all fingerprints. However, the scaffold analysis identified several promising candidates in the Mtbs and Phytochemical datasets that could serve as a starting point for the rational design of novel Mtb inhibitors. Moreover, similarity analysis revealed significant redundancy in the collection of chemical molecules targeting Mtb. This redundancy can contribute to the development of drug resistance in Mtb strains due to spontaneous mutations, especially for structurally related drugs. Therefore, there is a need to expand the existing chemical space of Mtb-targeting molecules by including representative molecules from various scaffolds to overcome the challenge of drug resistance.

References

[1]. WHO (2022) Global Tuberculosis Report 2022. Geneva

[2]. Zumla A, Maeurer M, Consortium for the H-DTN (HDT-N, et al (2015) Host-Directed Therapies for Tackling Multi-Drug Resistant Tuberculosis: Learning From the Pasteur-Bechamp Debates. Clin Infect Dis 61:1432–1438. https://doi.org/10.1093/cid/civ631

[3]. Palomino J, Martin A (2014) Drug Resistance Mechanisms in Mycobacterium tuberculosis. Antibiotics 3:317–340. https://doi.org/10.3390/antibiotics3030317

[4]. Gler M, Skripconoka V, Sanchez-Garavito E, et al (2012) Delamanid for Multidrug-Resistant Pulmonary Tuberculosis. N Engl J Med 366:2151–2160. https://doi.org/10.1056/NEJMoa1112433

[5]. Khawbung J, Nath D, Chakraborty S (2020) Drug Resistant Tuberculosis: A Review. Comp Immunol Microbiol Infect Dis 74:101574. https://doi.org/10.1016/j.cimid.2020.101574

[6]. Zumla A, Nahid P, Cole S (2013) Advances in the development of new tuberculosis drugs and treatment regimens. Nat Rev Drug Discov 12:388–404. https://doi.org/10.1038/nrd4001

[7]. Brennan PJ, Nikaido H (1995) THE ENVELOPE OF MYCOBACTERIA. Annu Rev Biochem 64:29–63. https://doi.org/10.1146/annurev.bi.64.070195.000333

[8]. Barry C, Lee R, Mdluli K, et al (1998) Mycolic acids: Structure, biosynthesis and physiological functions. Prog Lipid Res 37:143–179. https://doi.org/10.1016/S0163-7827(98)00008-3

[9]. Mashabela G, de Wet T, Warner D (2019) Mycobacterium tuberculosis Metabolism. Microbiol Spectr 7:. https://doi.org/10.1128/microbiolspec.GPP3-0067-2019

[10]. Wheeler PR, Ratledge C (1994) Metabolism of Mycobacterium tuberculosis. In: Tuberculosis. pp 353–385

[11]. Ehrt S, Schnappinger D, Rhee KY (2018) Metabolic principles of persistence and pathogenicity in Mycobacterium tuberculosis. Nat Rev Microbiol 16:496–507. https://doi.org/10.1038/s41579-018-0013-4

[12]. Marimani M (2020) Chapter 2 - Combination therapy against multidrug resistance. In: Wani MY, Ahmad ABT-CTAMR (eds). Academic Press, pp 39–64

[13]. Mondoni M, Saderi L, Sotgiu G (2021) Novel treatments in multidrug-resistant tuberculosis. Curr Opin Pharmacol 59:103–115. https://doi.org/https://doi.org/10.1016/j.coph.2021.05.007

[14]. Singh V, Chibale K (2021) Strategies to Combat Multi-Drug Resistance in Tuberculosis. Acc Chem Res 54:2361–2376. https://doi.org/10.1021/acs.accounts.0c00878

[15]. Dheda K, Mirzayev F, Cirillo DM, et al (2024) Multidrug-resistant tuberculosis. Nat Rev Dis Prim 10:22. https://doi.org/10.1038/s41572-024-00504-2

[16]. Jnawali HN, Ryoo S (2013) First– and Second–Line Drugs and Drug Resistance. In: Mahboub BH, Vats MG (eds). IntechOpen, Rijeka, p Ch. 10

[17]. Rendon A, Tiberi S, Scardigli A, et al (2016) Classification of drugs to treat multidrug-resistant tuberculosis (MDR-TB): evidence and perspectives. J Thorac Dis Vol 8, No 10 (October 01, 2016) J Thorac Dis

[18]. Yang T-W, Park HO, Jang H, et al (2017) Side effects associated with the treatment of multidrug-resistant tuberculosis at a tuberculosis referral hospital in South Korea: A retrospective study. Medicine (Baltimore) 96:e7482. https://doi.org/10.1097/MD.0000000000007482

[19]. Prasad R, Singh A, Gupta N (2019) Adverse drug reactions in tuberculosis and management. Indian J Tuberc 66:520–532. https://doi.org/10.1016/j.ijtb.2019.11.005

[20]. Ekins S, Freundlich JS, Choi I, et al (2011) Computational databases, pathway and cheminformatics tools for tuberculosis

drug discovery. Trends Microbiol 19:65–74. https://doi.org/https://doi.org/10.1016/j.tim.2010.10.005

[21]. Reymond J-L, van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Medchemcomm 1:30–38. https://doi.org/10.1039/C0MD00020E

[22]. Medina-Franco JL, Sánchez-Cruz N, López-López E, Díaz-Eufracio BI (2022) Progress on open chemoinformatic tools for expanding and exploring the chemical space. J Comput Aided Mol Des 36:341–354. https://doi.org/10.1007/s10822-021-00399-1

[23]. Goel M, Aggarwal R, Sridharan B, et al (2022) Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. Wiley Interdiscip Rev Comput Mol Sci 13:. https://doi.org/10.1002/wcms.1637

[24]. Macalino SJ, Billones J, Organo V, Carrillo MC (2020) In Silico Strategies in Tuberculosis Drug Discovery. Molecules 25:. https://doi.org/10.3390/molecules25030665

[25]. Wawer M, Lounkine E, Wassermann A, Bajorath J (2010) Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov Today 15:630–639. https://doi.org/10.1016/j.drudis.2010.06.004

[26]. Golbraikh A, Wang X, Zhu H, Tropsha A (2016) Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment. 1–38. https://doi.org/10.1007/978-94-007-6169-8_37-2

[27]. Pirhadi S, Shiri F, Ghasemi JB (2015) Multivariate statistical analysis methods in QSAR. RSC Adv 5:104635–104665. https://doi.org/10.1039/C5RA10729F

[28]. Naveja J de J, Saldivar-Gonzalez F, Sánchez Cruz N, Medina-Franco J (2018) Cheminformatics Approaches to Study Drug Polypharmacology. In: Methods in Pharmacology and Toxicology. pp 3–25

[29]. Fernandez-de Gortari E, Medina-Franco J (2015) Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. RSC Adv 5:87465. https://doi.org/10.1039/c5ra19611f

[30]. González-Medina M, Medina-Franco JL (2019) Chemical Diversity of Cyanobacterial Compounds: A Chemoinformatics Analysis. ACS Omega 4:6229–6237. https://doi.org/10.1021/acsomega.9b00532

[31]. Saldívar-González FI, Lenci E, Trabocchi A, Medina-Franco JL (2019) Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study. RSC Adv 9:27105–27116. https://doi.org/10.1039/C9RA04841C

[32]. Saldivar-Gonzalez F, Valli M, Andricopulo A, et al (2019) Chemical Space and Diversity of NuBBE Database: A Chemoinformatic Characterization. J Chem Inf Model 59:74–85. https://doi.org/10.1021/acs.jcim.8b00619

[33]. Prieto-Martínez F, Peña-Castillo A, Méndez-Lucio O, et al (2016) Molecular Modeling and Chemoinformatics to Advance the Development of Modulators of Epigenetic Targets: A Focus on DNA Methyltransferases. Adv Protein Chem Struct Biol 105:. https://doi.org/10.1016/bs.apcsb.2016.05.001

[34]. Prieto-Martínez FD, Gortari EF, Méndez-Lucio O, Medina-Franco JL (2016) A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. RSC Adv 6:56225–56239. https://doi.org/10.1039/C6RA07224K

[35]. Naveja JJ, Norinder U, Mucs D, et al (2018) Chemical space, diversity and activity landscape analysis of estrogen receptor binders. RSC Adv 8:38229–38237. https://doi.org/10.1039/C8RA07604A

[36]. Ahamed S, Muraleedharan K (2020) A Cheminformatic Study on Chemical Space Characterization and DiversityAnalysisof 5-LOX Inhibitors. J Mol Graph Model 100:107699. https://doi.org/10.1016/j.jmgm.2020.107699

[37]. Gaulton A, Bellis LJ, Bento AP, et al (2012) ChEMBL: a large-

scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1093/nar/gkr777

[38]. Kim S, Thiessen P, Bolton E, et al (2015) PubChem Substance and Compound databases. Nucleic Acids Res 44:. https://doi.org/10.1093/nar/gkv951

[39]. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. J Chem Inf Model 55:460–473. https://doi.org/10.1021/ci500588j

[40]. Gautam R, Saklani A, Jachak S (2007) Indian medicinal plants as a source of antimycobacterial agents. J Ethnopharmacol 110:200–234. https://doi.org/10.1016/j.jep.2006.12.031

[41]. Ramalingam G, S S, Manickan A (2016) Anti-Mycobacterial Activity of Acalypha indica and Andrographis paniculata, the Indian Medicinal Plants against Mycobacterium tuberculosis (MTB). Int J Pharm Pharm Res 7:33–43

[42]. Heinrich M, Gibbons S (2001) Ethnopharmacology in drug discovery: An analysis of its role and potential contribution. J Pharm Pharmacol 53:425–432. https://doi.org/10.1211/0022357011775712

[43]. Grange JM, Snell NJC (1996) Activity of bromhexine and ambroxol, semi-synthetic derivatives of vasicine from the Indian shrub Adhatoda vasica, against Mycobacterium tuberculosis in vitro. J Ethnopharmacol 50:49–53. https://doi.org/https://doi.org/10.1016/0378-8741(95)01331-8

[44]. Wishart D, Knox C, Guo AC, et al (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901-D906. Nucleic Acids Res 36:D901-6. https://doi.org/10.1093/nar/gkm958

[45]. Nasri H, Baradaran A, Shirzad H, Rafieian-kopaei M (2014) New Concepts in Nutraceuticals as Alternative for Pharmaceuticals. Int J Prev Med 5:1487–1499

[46]. Gruber K (2015) Access sought to tuberculosis drug from

nutraceutical company. Nat Med 21:. https://doi.org/10.1038/nm.3805

[47]. Van de Waterbeemd H (2003) Physicochemical concepts in drug design. EXS 243–257. https://doi.org/10.1007/978-3-0348-7997-2_12

48. Waskom M (2021) seaborn: statistical data visualization. J Open Source Softw 6:3021. https://doi.org/10.21105/joss.03021

[49]. Hao J, Ho T (2019) Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. J Educ Behav Stat 44:107699861983224. https://doi.org/10.3102/1076998619832248

[50]. Muegge I, Mukherjee P (2015) An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 11:. https://doi.org/10.1517/17460441.2016.1117070

[51]. Willett P, Barnard JM, Downs GM (1998) Chemical Similarity Searching. J Chem Inf Comput Sci 38:983–996. https://doi.org/10.1021/ci9800211

[52]. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular Similarity in Medicinal Chemistry. J Med Chem 57:3186–3204. https://doi.org/10.1021/jm401411z

[53]. Schuffenhauer A, Ertl P, Roggo S, et al (2007) The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J Chem Inf Model 47:47–58. https://doi.org/10.1021/ci600338x

[54]. Medina-Franco J, Martínez K, Giulianotti M, et al (2008) Visualization of the Chemical Space in Drug Discovery. Curr Comput - Aided Drug Des 4:322–333. https://doi.org/10.2174/157340908786786010

[55]. González-Medina M, Medina-Franco JL (2017) Platform for Unified Molecular Analysis: PUMA. J Chem Inf Model 57:1735–1740. https://doi.org/10.1021/acs.jcim.7b00253

[56]. Gao S, Mokhtarian P, Johnston R (2008) Nonnormality of Data

in Structural Equation Models. Transp Res Rec 2082:116–124. https://doi.org/10.3141/2082-14

[57]. Nikolova N, Jaworska J (2003) Approaches to Measure Chemical Similarity - A Review. QSAR Comb Sci 22:. https://doi.org/10.1002/qsar.200330831

[58]. Ringnér M (2008) What is principal component analysis? Nat Biotechnol 26:303–304. https://doi.org/10.1038/nbt0308-303

[59]. Langdon SR, Brown N, Blagg J (2011) Scaffold Diversity of Exemplified Medicinal Chemistry Space. J Chem Inf Model 51:2174–2185. https://doi.org/10.1021/ci2001428

[60]. Lipkus AH, Yuan Q, Lucas KA, et al (2008) Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. J Org Chem 73:4443–4451. https://doi.org/10.1021/jo8001276

[61]. Chhabra S, Kumar S, Parkesh R (2021) Chemical Space Exploration of DprE1 Inhibitors Using Chemoinformatics and Artificial Intelligence. ACS Omega 6:14430–14441. https://doi.org/10.1021/acsomega.1c01314

[62]. Savjani JK, Gajjar AK (2011) Pharmaceutical importance and synthetic strategies for imidazolidine-2-thione and imidazole-2-thione derivatives. Pakistan J Biol Sci PJBS 14:1076–1089. https://doi.org/10.3923/pjbs.2011.1076.1089

[63]. Cho S, Kim S-H, Shin D (2019) Recent applications of hydantoin and thiohydantoin in medicinal chemistry. Eur J Med Chem 164:517–545. https://doi.org/https://doi.org/10.1016/j.ejmech.2018.12.066

[64]. Li Petri G, Raimondi MV, Spanò V, et al (2021) Pyrrolidine in Drug Discovery: A Versatile Scaffold for Novel Biologically Active Compounds. Top Curr Chem 379:34. https://doi.org/10.1007/s41061-021-00347-5

[65]. Pohlmann J, Lampe T, Shimada M, et al (2005) Pyrrolidinedione derivatives as antibacterial agents with a novel mode of action. Bioorg Med Chem Lett 15:1189–1192.

https://doi.org/https://doi.org/10.1016/j.bmcl.2004.12.002

[66]. Bollikolla HB, Tyagi R, Gokada MR, et al (2022) Flavones as Important Scaffolds for Anticancer, Antioxidant and Anti-Tubercular Activities: An Overview of Reports 2015–2020. Moscow Univ Chem Bull 77:269–285. https://doi.org/10.3103/S0027131422050042

[67] Rabaan A, Alhumaid S, Albayat H, et al (2022) Promising Antimycobacterial Activities of Flavonoids against Mycobacterium sp. Drug Targets: A Comprehensive Review. Molecules. https://doi.org/10.3390/molecules27165335

[68]. Catauro M, D'Angelo A, Fiorentino M, et al (2023) Thermal, spectroscopic characterization and evaluation of antibacterial and cytotoxicity properties of quercetin-PEG-silica hybrid materials. Ceram Int 49:14855–14863. https://doi.org/https://doi.org/10.1016/j.ceramint.2022.07.256

[69]. Catauro M, D'Errico Y, D'Angelo A, et al (2021) Antibacterial Activity and Iron Release of Organic-Inorganic Hybrid Biomaterials Synthesized via the Sol-Gel Route. Appl. Sci. 11

[70]. Vertuccio L, Guadagno L, D'Angelo A, et al (2023) Sol-Gel Synthesis of Caffeic Acid Entrapped in Silica/Polyethylene Glycol Based Organic-Inorganic Hybrids: Drug Delivery and Biological Properties. Appl. Sci. 13

Chapter 4

Activity Landscape Modeling of InhA Inhibitors: Characterizing Potency Variations through Structural Similarity

4.1. Introduction

Over the past century, the convergence of chemistry, biology, and medicine has resulted in enormous advances. This advancement has resulted in domains such as cheminformatics and computer-aided drug design (CADD), both of which play critical roles in current drug discovery. For more than three decades, CADD techniques have played a crucial role in the discovery of clinically useful small compounds. Many proteins have been discovered as therapeutic targets for a variety of diseases, and online databases provide thorough information on the structure and function of inhibitors for these proteins. Researchers use computational approaches to discover trends in structure-activity data, allowing for better prediction of the biological activity of a compound on a given protein, an important component of rational drug development [1,2].

The rise of multi-drug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis (TB) has necessitated the discovery of novel anti-TB compounds, particularly those that shorten treatment time and limit drug resistance. In *Mycobacterium tuberculosis* (Mtb), the enzyme InhA functions as an enoyl-[acyl-carrier-protein] reductase, playing a crucial role in the biosynthesis of mycolic acids. It has emerged as an interesting target for the development of new TB medicines. Furthermore, InhA is targeted by isoniazid (INH), the first-line treatment and prevention of TB. However, the increase in treatment resistance stains emphasises the significance of the development of new therapeutic drugs to treat TB. In the context of increasing MDR-Mtb strains, InhA remains a prominent target for TB

drug discovery. Both experimental and computational methods have been used to identify and design a diverse range of InhA inhibitors. Various screening methods, such as combinatorial chemistry and high-throughput screening (HTS), have identified several inhibitors of InhA, as well as structure-activity relationship (SAR) [3–5]. Moreover, numerous InhA protein crystal structures have been identified using X-ray crystallography. The inhibitors developed through different screening techniques, together with their SAR data, are published in scientific literature and stored in publicly accessible databases such as BindingDB [6], PubChem [7], and ChEMBL [8]. Protein structures of InhA are also accessible in the Protein Data Bank (PDB) [9]. This extensive data collection is critical for the development of new inhibitors against InhA. Studying the SAR of these data is a key step towards virtual or experimental testing to identify new biologically active compounds [10,11]. Furthermore, data mining techniques can facilitate the creation of predictive models for assessing the activity of novel inhibitors [12]. Consequently, various predictive models, SAR and QSAR for InhA inhibitors have been developed to aid drug design [13–15].

While various approaches can produce reliable and predictive models, dealing with datasets containing discontinuous SARs remains challenging. This difficulty arises because QSAR and machine learning techniques frequently produce predictions with uncertain reliability, which can lead to misleading or ineffective results. To overcome this challenge, activity landscape modeling provides an effective framework for systematically evaluating and structuring large datasets,

allowing for early detection of activity cliffs, which is essential before quantitative predictive analysis. In this context, activity cliffs are commonly defined using the activity landscape concept, a valuable tool for SAR research. Specifically, an activity landscape can be represented as a hypersurface within a biologically relevant chemical space, similar to geographical maps, with compound potency added as a third dimension in a 2D projection of the chemical space. As a result, the SAR of the dataset can be viewed as an activity landscape, where potency introduces an additional dimension to the chemical space [16–18]. This landscape includes areas of smooth SAR, where similar compounds exhibit comparable activity, as well as zones of sharp discontinuity, where minor structural changes result in substantial variations in biological potency. This zone is called activity cliff region. It is crucial for medicinal chemists to visualize and understand the activity cliff regions within SAR data because they help highlight molecular features that are important for drug development [11,19]. However, activity cliffs pose a significant challenge when developing predictive models based on structural similarity. These models rely on the assumption that compounds with similar structures tend to exhibit comparable properties. However, a notable exception to this assumption provides valuable insights into the relationship between molecular structure and biological activity. This exception is represented by activity cliffs instances where two closely related molecules display significant differences in their biological activity. Consequently, Activity cliffs can introduce substantial errors in machine learning models, leading to incorrect activity predictions for certain compounds despite the model's overall strong performance [20].

In recent years, a number of quantitative and visual methodologies have emerged for mapping a compound's activity landscape across multiple biological endpoints. Techniques like statistical analysis and classification have been used to organize datasets and explore the chemical environment of potential hits [21–23]. Researchers, including the group led by José L. Medina-Franco, have used these techniques to analyze the SAR of inhibitors that target different diseases [24–28]. Despite this, no extensive research has been conducted on the structural diversity and activity landscape of InhA inhibitors. In this study, we extracted SAR from the dataset and visualized it using landscape modeling, which is a systematic pairwise comparison of structural and activity similarities. In addition, we sought to identify activity cliffs within the structure-activity landscape. These compounds were further investigated using protein-ligand interactions to provide a structural interpretation of the features that contribute to activity cliffs.

4.2.Materials and Methods

4.2.1 Dataset and Data Curation

The chemical structures of InhA inhibitors were obtained from the public database ChEMBL, based on their bioactivity data. We imported inhibitor structures and activities from the ChEMBL database into DataWarrior software [29], selecting the target as 'enoyl-[acyl-carrier-protein] reductase'. DataWarrior allows for ChEMBL queries. Initially, this dataset included 1329 InhA inhibitors.

In ChEMBL, bioactivities are often provided in units such as K_i, K_d, IC_{50}, and EC_{50}, as well as assay data such as cell line, organism, or tissue. In this work, we focused on compounds with reported IC_{50} values in the enzymatic inhibitory assay. The IC_{50} refers to the concentration required for 50% inhibition of the enzyme. Then the bioactivity data were transformed into logarithmic values as pIC_{50} (-log IC_{50}).

Data curation involved removing duplicate compounds and those lacking specific bioactivity information. In cases where compounds had the same chemical structure but different bioactivity endpoints, the compound with the lowest recorded activity value was included. After curation, the dataset comprised 290 unique InhA inhibitor molecules, with pIC_{50} values ranging from 3.54 to 8.7.

4.2.2. **Activity landscape modeling**

Activity landscapes are visual representations that integrate activity data and compound similarity [30]. It allows for a direct assessment of the similarity principle, which posits that structurally similar compounds in a data set should exhibit similar biological activities. SAR found in compound data sets can be accessed graphically through activity landscapes, which combine structural and potency information about active compounds [31]. For this study, we used Structure-Activity Similarity (SAS) maps and the Structure-Activity Landscape Index (SALI) as key methodologies from the range of available activity landscape analysis techniques.

4.2.2.1 *SAS maps*

SAS maps are one of the first techniques introduced to characterise SAR utilising the notion of activity landscape modeling. A schematic representation of a SAS map is shown in **Fig. 4.1**. Typically, The X-axis represents structural similarity, while the Y-axis plots the activity difference or activity similarity. SAS maps can be broadly divided into four major regions. The upper-right region of the figure is the most important zone examined in this study because it represent pair of compounds with greater activity differences and a greater molecular distance (activity cliff region). This zone indicates activity cliffs on SAS maps. The bottom-left area indicate pairs of compounds with low activity difference and low molecular distance (also known as similarity cliff or scaffold hopping region). On the other hand, the lower-right zone depict a pair of molecules with a minimal activity differences and a larger molecular distance (smooth SAR region). The top-left section represent pair of compounds with a discontinuous SAR in which compound pairs exhibit significant differences in activity despite having little structural similarity (Non-descriptive region). SAS maps were created in this study using an online tool called the 'Activity Landscape Plotter' [32].

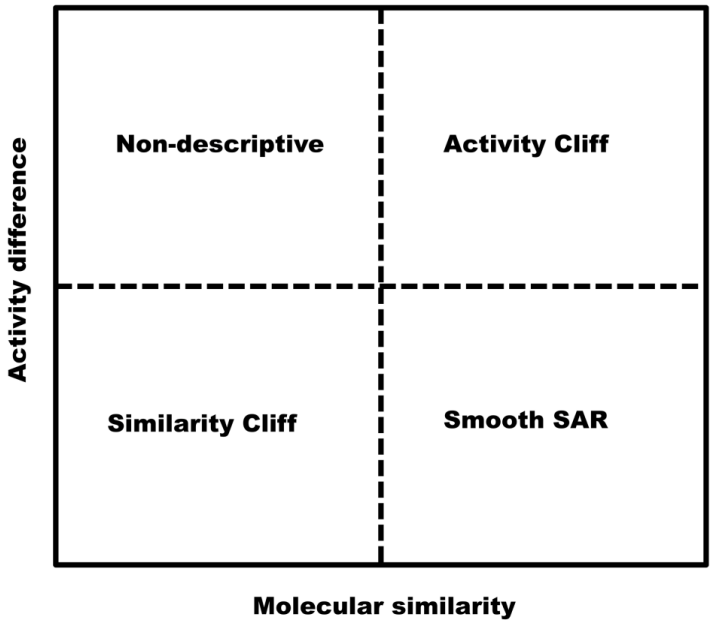


Fig. 4.1. Four major regions in a SAS map.

4.2.2.2 *Structure–Activity Landscape Index*

The Structure-Activity Landscape Index (SALI) was developed to measure activity cliffs by evaluating the relationship between molecular similarity and activity difference. In the Activity Landscape Plotter, SALI is calculated using the formula introduced by Guha and Van Drie [33] in **Equation (4.1)**:

$$SALI\ i,j = \frac{|A_i - A_j|}{1 - \mathrm{sim}(i,j)} \quad \textbf{(4.1)}$$

In this context, A_i and A_j represent the activities of the i^{th} and j^{th} molecules, respectively, while sim(i,j) denotes the similarity coefficient between the two molecules, which in this study is calculated using the ECFP4 fingerprint and the Tanimoto coefficient [34]. The SALI values

are then visualized on the Structure-Activity Similarity (SAS) maps with a continuous color scale ranging from green for the most structurally similar pairs to red for the least similar pairs.

4.2.2.3 *Activity Cliff Generators*

Activity landscape methods are widely used to identify activity cliffs, which are compounds with similar structures but unexpectedly different biological activities [35,36]. The presence of activity cliffs in datasets has the potential to significantly impact medicinal chemistry and computational projects [18]. Activity cliffs, for example, impede the development of effective predictive methods like quantitative structure-activity relationships (QSAR) [37,38]. Furthermore, compounds that frequently produce activity cliffs are unsuitable as query molecules in similarity-based virtual screening. In medicinal chemistry, activity cliffs can be useful because they highlight important pharmacophoric regions. Small structural changes that cause significant biological response modifications can be used to optimize leads.

4.2.3 **Molecular Docking**

A structure-focused strategy was used to assess the impact of specific molecular features on activity cliff formation and investigate their contribution to ligand-enzyme interactions. This study primarily utilized a ligand-based methodology, with in-depth structural analyses conducted on the most significant compounds. The activity cliff generators were docked into the InhA (PDB ID: 4COD), which is bound to a small ligand N-((3R,5S)-1-(benzofuran-3-carbonyl)-5-

(ethylcarbamoyl)pyrrolidin-3-yl)-3-ethyl-1-methyl-1H-pyrazole-5-carboxamide (KV1), a compound closely related to the activity cliff compound ChEMBL3125275 using CB-DOCK 2 [39,40]. 2D diagrams describing the interactions between the activity cliff and the receptor were also predicted via LigPlot+ software [41].

4.2.4 **Molecular dynamics simulation and MM-PBSA calculation**

To further investigate the interactions between the activity cliffs and their target protein, Molecular dynamics (MD) simulations were performed on the activity cliff generators using SiBioLead, an automated online platform [42]. The OPLS/AA force field was used in the simulation, with ligand parameters generated by AMBERTOOLS and the ACPYPE package [43]. The protein-ligand complex was immersed in a triclinic box filled with SPC water molecules and neutralized with NaCl counter ions. An additional 0.15 M of NaCl was added to simulate physiological conditions. Using NVT and NPT ensembles, the system was equilibrated over 300 ps. Every 20 ps, a trajectory snapshot was taken, yielding 5000 images. GROMACS' integrated tools were used to analyze these trajectories, and the results were displayed in xmgrace. In addition, the binding free energy ($\Delta G_{binding}$) of protein-ligand complexes was calculated. The molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method was used to estimate binding free energies, with an automated plugin available on the SiBioLead server [44,45]. The following **equation (4.2)** was used to calculate the binding free energy for each frame:

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{receptor} + \Delta G_{ligand}) \quad \textbf{(4.2)}$$

4.3. Results and Discussion

4.3.1 SAS Map

The SAS map, which contains 41,744 pairs derived from the InhA inhibitor dataset is shown in **Fig.4.2**. The map is divided into four primary regions (I-IV), each denoted by dotted lines. Various methods can be employed to determine thresholds for dividing the plot. In this study, the threshold for structural similarity along the X-axis was calculated by taking the dataset's median pairwise similarity value and increasing it by two standard deviations. This calculation resulted in a threshold value of 0.5. Similarly, the Y-axis potency difference threshold was set to 2 log units.

The SALI-SAS map (**Fig. 4.2. A**) shows compound pairs using a color-coded scheme based on their SALI values. Green dots represent pairs with the lowest SALI values, orange to yellow dots indicate moderate SALI values, and the red dot represents the pair with the highest SALI values. The majority of the map's points are green and yellow, indicating a predominantly continuous structure-activity relationship (SAR), in which structurally similar compounds exhibit comparable activities. In other words, minor changes in structure only result in minor activity changes. This pattern most likely reflects the dataset's origin in a lead optimization process [46–48]. **Fig.4.2.A** demonstrates that a significant portion of pairs of compounds in the dataset exhibit substantial differences in activity (greater than 1 or 2 log units), which contributes to the rugged activity landscape. However, activity cliffs form in a small portion of upper right zone of the map.

Even though, they provide vital information about SAR of InhA inhibitors. Furthermore, their presence assists medicinal chemists in identifying structural modifications that can improve potency or other desirable properties during the development of inhibitors against Mtb. Also, helps to make effective machine learning models with improved accuracy.

Similarly, **Fig.4.2B,** shows the distribution of activity levels for compounds in the InhA dataset, with different colors indicating the relative activity of each compound in a similarity pair. Red dots indicate the most potent compounds, yellow-to-orange dots indicate those with moderate activity, and green dots identify the least active compounds in each pair. Notably, the red circles on the SAS map are distributed across the potency spectrum. Red dots at the top of the plot represent pairs of compounds with significant potency differences, such as one highly active and one inactive compound. In contrast, the red dots near the bottom of the plot, where potency differences are minimal, represent pairs in which both compounds are equally active.

The 'Density SAS Map' (**Fig.4.2.C**) uses a continuous color gradient to represent the concentration of data points in different areas, with red indicating a high concentration (higher density data points) and grey indicating a low concentration (lower density data points). The region identified as the similarity cliff on the SAS map displays a dense cluster of data points, implying that the majority of the compounds in the dataset, despite having distinct chemical structures, exhibit comparable levels of activity.

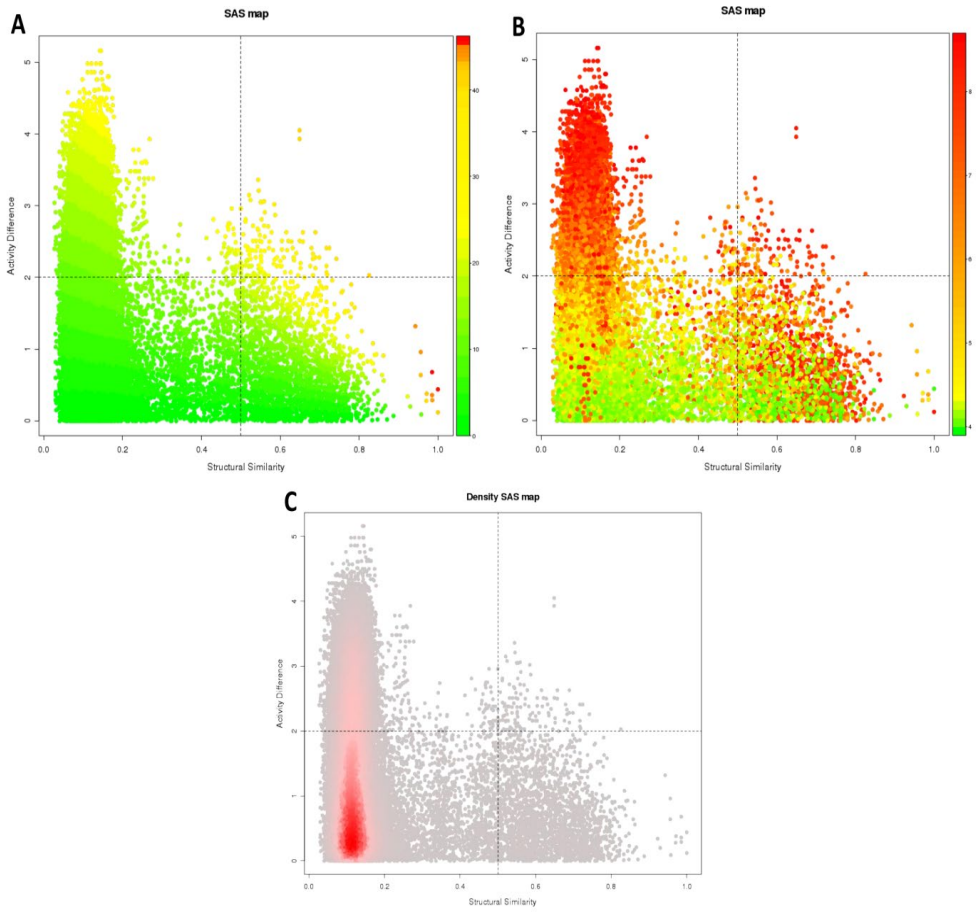


Fig.4.2. SAS maps of the global activity landscapes of InhA inhibitors: (**A)** SALI SAS map, (**B)** maximum activity SAS map, (**C)** density SAS map.

4.3.1.1. Quantitative Analysis of SAS Maps

As described in the methods section and illustrated in **Fig. 4.2**, the SAS maps were divided into four distinct regions to allow for systematic and quantitative analysis. Each region was assigned specific thresholds for structural similarity and activity difference. **Table 4.1.**, summarises the number of similarity pairs within each of the SAS

map's four areas (I-IV), demonstrating the dataset's diverse SAR, which includes both continuous and discontinuous regions. The quantitative analysis shows that activity cliffs constitute approximately 0.3% of the dataset. Although uncommon, these cliffs are important for understanding SAR because they represent instances where minor structural variations result in significant activity changes, posing potential challenges for predictive modeling. It's worth noting that the scaffold hop/similarity cliff region has the highest concentration of data points, accounting for 67.3% of the InhA inhibitor dataset. This implies that nearly 70% of the compound pairs have significantly different chemical structures but comparable activity levels. Furthermore, 5.2% of the compound pairs are located in a smooth SAR region, making them suitable for predictive modeling. These proportions are subject to change with the release of more activity data, as they are based on the current content of the ChEMBL.

Table 4.1. Quantitative Analysis Summary of four SAS Maps region

Quadrant	Region	No. of pairs	Percentage
1	Non-Descriptive	11390	27.3
2	Similarity cliff	28076	67.3
3	Smooth SAR	2164	5.2
4	Activity cliffs	113	0.3
	Total	41743	

4.3.2. Activity Cliff Generators and SAR Interpretation

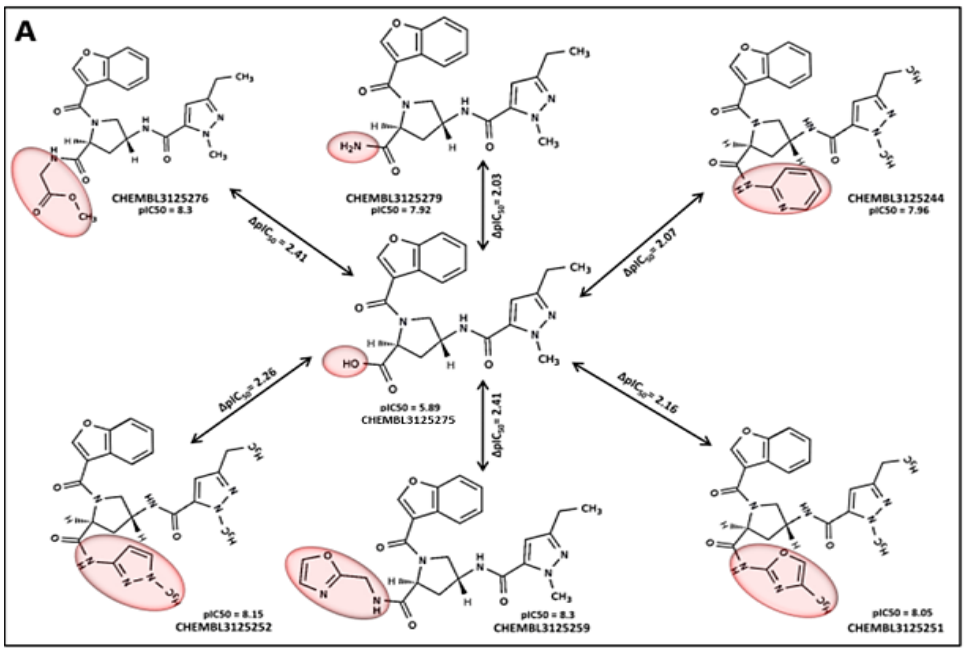
Examining and interpreting these activity cliff generators is anticipated to provide critical insights into the structural factors that influence compounds' inhibitory potency against InhA. For this study, an activity generator was defined as a compound that appeared in at least five distinct activity cliff pairs. Using this definition, ten compounds were recognized as activity cliff generators.

Fig.4.3, illustrates the chemical structures of two exemplary activity cliff generators and the associated compound pairs found in the dataset. The figure also shows ChEMBL IDs and their corresponding activity values (pIC_{50}). All compound pairs have a potency difference (ΔpIC_{50}) greater than 2 compared to their respective cliff generators. The activity cliffs related to (2S,4R)-1-(1-benzofuran-3-carbonyl)-4-[(5-ethyl-2-methylpyrazole-3-carbonyl)amino]pyrrolidine-2-carboxylic acid (ChEMBL ID 3125275, pIC_{50} = 5.89; IC_{50} = 1288 nM), as well as the chemical structures of six cliff-forming compounds, are shown in **Fig.4.3.A**. Upon analyzing the structural differences, it is evident that substituting a hydroxyl group in pyrrolidine-2-carboxylic acid leads to a sharp increase in binding affinity. Replacing pyrrolidine-2-carboxylic acid with (diethylcarbamoyl)pyrrolidin-3-yl further enhances the binding affinity (**Fig.4.3.B**). However, the molecule N-[(3R,5S)-1-(1-benzofuran-3-carbonyl)-5-(diethylcarbamoyl)pyrrolidin-3-yl]-5-ethyl-2-methylpyrazole-3-carboxamide (ChEMBL ID 3125280, pIC_{50} = 6.02; IC_{50} = 955 nM) can itself be considered an activity cliff generator.

By examining the activity cliffs and their corresponding compound pairs, it is evident that all of these compounds are pyrrolidine carboxamide derivatives. Recent research has highlighted pyrrolidine carboxamides as a new class of InhA inhibitors [49–51]. A notable observation is that the IC_{50} value decreases when a nitrogen-containing group, particularly a heterocyclic one, replaces the substituent in the carboxamide moiety. Current studies suggest that nitrogen-containing heterocyclic compounds are being widely explored for their potential as anti-TB agents. These nitrogen-based compounds have various targets, with InhA being a primary focus [52]. Therefore, structure-activity relationship (SAR) studies of this compound class are crucial for drug development.

From the analysis, it is apparent that all cliff compound pairs exhibit IC_{50} values that are at least two hundred times lower than their corresponding cliff generator compounds. This suggests that the abrupt change in activity between the cliff pairs leads to a discontinuous SAR. This discontinuity makes it challenging to perform predictive modeling, such as QSAR, using pyrrolidine carboxamide compounds, as SAR continuity is essential for QSAR analysis. Alternatively, using insights from activity cliff generators to modify the common structure could improve the efficacy of lead InhA inhibitors. This approach may improve potency, selectivity, and pharmacokinetic properties, potentially positioning these compounds as lead candidates in drug development. In summary, recognizing activity cliffs within compound datasets is crucial for guiding the development of accurate predictive models. By eliminating activity cliffs, the performance of models that

rely on the similarity principle, such as QSAR approaches, can be improved. To assess their predictive performance, it is critical to construct and test various InhA datasets, comparing those with and without activity cliffs.



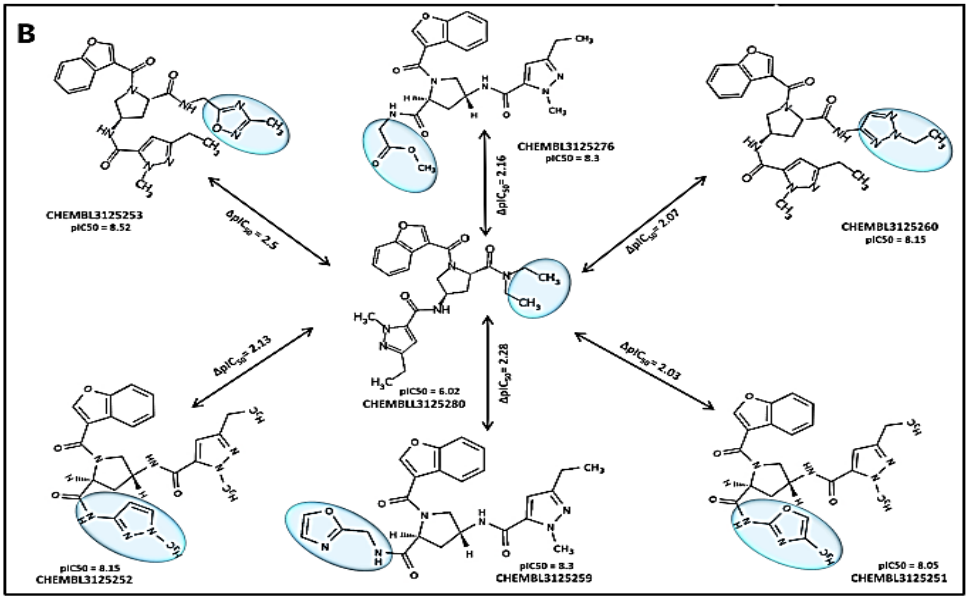


Fig.4.3. Representative activity cliff generators and selected pairs of compounds formed with the generators (A) (2S,4R)-1-(1-benzofuran-3-carbonyl)-4-[(5-ethyl-2-methylpyrazole-3-carbonyl)amino]pyrrolidine-2-carboxylic acid, (B) N-[(3R,5S)-1-(1-benzofuran-3-carbonyl)-5-(diethylcarbamoyl)pyrrolidin-3-yl]-5-ethyl-2-methylpyrazole-3-carboxamide in the activity landscape of InhA dataset. The structural changes resposible for the activity cliff are highlighted in color.

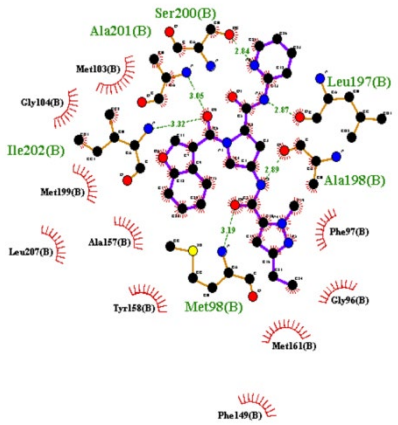
4.3.3. Molecular docking

We used molecular docking simulations of the previously identified activity cliff generators to investigate ligand-receptor interactions. Notably, functional activity is not always correlated with binding affinity, which is the main object of this analysis. Despite its shortcomings in determining exact binding energies, molecular docking is a popular technique for predicting possible binding modes. [53]. Docking was carried out using CB DOCK2, as described in the materials and methods section.

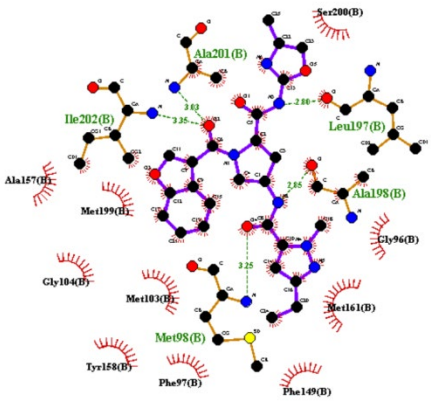
Fig.4.4, shows the predicted binding modes of selective compounds, identified as activity cliffs generators, within the binding site of InhA. The predicted binding modes of molecules containing the pyrrolidine carboxamide moiety are characterized by the formation of hydrogen bonds and hydrophobic interactions between the ligands and the InhA binding site. The cliff compounds form hydrogen bonds with key residues such as Ser200, Leu197, Ala198, Met198, Ala201, and Ile202. The compound CHEMBL3125275 interacts through the oxygen (O) present in carboxylic acid, whereas the others interact through the nitrogen (N) in the carboxamide moiety, which can change electronic distribution and affinity. As a result, subtle differences in hydrogen bond distances have been observed, which may influence binding affinity. Residues Ser200 and Leu197, which are involved in the interaction with the heterocyclic substituent, could be examples of activity cliff hot spots [54]. This study validates that the nitrogen-containing heterocyclic substituent in carboxamide moiety aids molecules in more easily fitting into protein binding pockets. Moreover, these substituents, which include pyridine, oxazol, pyrazol, and others, may improve the conformational changes or a better fit in the binding site, which would boost activity. These activity cliff compounds also have hydrophobic interactions with the protein via residues such as Met103, Gly104, Met199, Ala157, Leu207, Try158, Phe149, Met161, Gly96, and Phe97.

This analysis reveals that small structural modifications in the ligands strengthen the interactions between the compounds and the protein. The increased interaction is attributed to the nitrogen-

containing substituent present in compounds. These subtle structural differences are responsible for the activity cliffs between the compounds, directly influencing the IC_{50} values against InhA. As previously discussed, the IC_{50} values of the activity cliffs and their pairs show a significant change due to these structural variations, leading to increased biological activity of the compounds. In addition, we carried out MD simulations and MM-PBSA free energy calculations to investigate and validate the stability and binding affinity of these complexes. This analysis provided additional insights into binding free energies, as well as a quantitative assessment of the ligand-receptor complexes' interaction strength and stability.



A



B

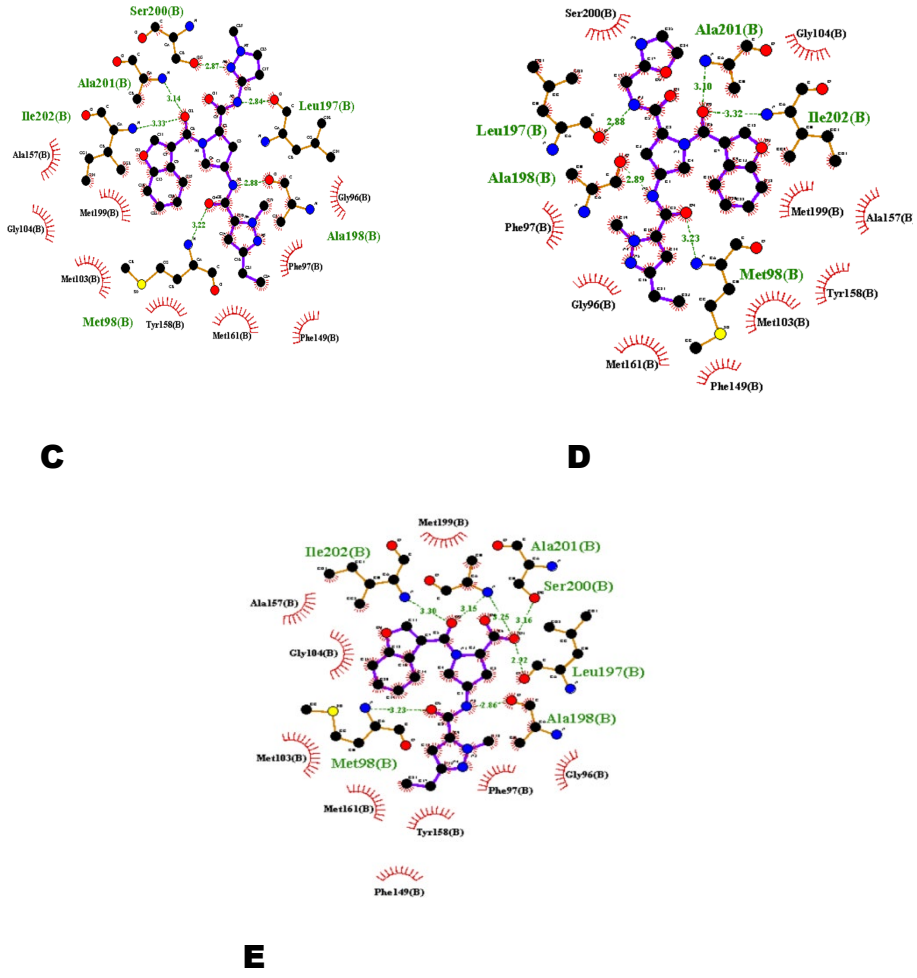


Fig.4.4. Predicted binding poses of selective compounds identified as activity cliffs generators, within the binding site of InhA. (A) CHEMBL3125244 (B) CHEMBL3125251 (C) CHEMBL3125252 (D) CHEMBL3125259 (E) CHEMBL3125275

4.3.4. Molecular dynamics simulation and free energy calculations

Fig.4.5, shows the RMSD plots for CHEMBL3125244, CHEMBL3125251, CHEMBL3125252, and CHEMBL3125259. CHEMBL3125244 exhibits a steady rise before stabilizing at 0.12 and

0.16 nm. This low RMSD indicates that the ligand is relatively stable inside the binding site, with just minimal variations. Compound CHEMBL312525 exhibits early stability, followed by variation between 0.08 and 0.15 nm, demonstrating structural flexibility before stabilization. CHEMBL3125252 exhibits a rise in value before stabilizing around 0.10-0.17 nm, indicating an initial adjustment before equilibrium. CHEMBL3125259 has the most stable RMSD value, constantly varying between 0.08 and 0.12 nm, indicating significant binding stability. Overall, our data indicate that all four compounds exhibit good binding stability in the binding pockets.

. Furthermore, the binding free energies of protein-ligand complexes were determined using MM-PBSA, which provides important information about the thermodynamic stability of these interactions. **Table 4.2**, shows the calculated Gibbs free energy values for the compounds CHEMBL3125244, CHEMBL3125251, CHEMBL3125252, and CHEMBL3125259. These values were -50.28, -48.51, -48.96, and -49.22, kcal/mol, respectively. Negative Gibbs free energy values indicate that ligand-protein interactions are thermodynamically favourable and stable. Lower binding energies for activity cliff pair compounds indicate better binding in the protein pocket.

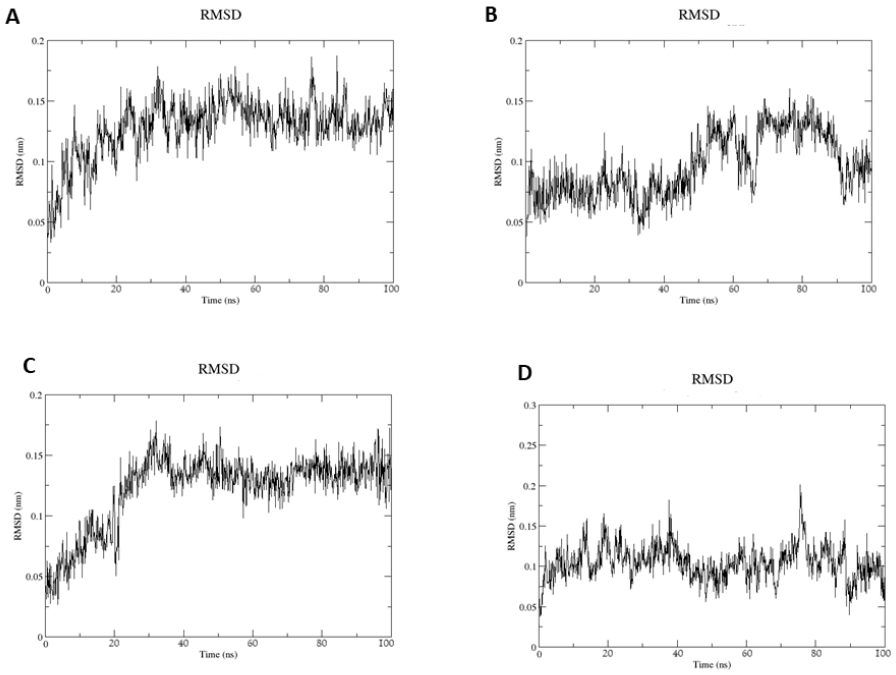


Fig.4.5. The RMSD plots of the activity cliff compounds (A) CHEMBL3125244, (B) CHEMBL3125251 (C) CHEMBL3125252 (D) CHEMBL3125259

Table 4.2. Binding free energy (Kcal/mol) for the selected activity cliff compounds

Compounds	**$\Delta G_{binding}$**
CHEMBL3125244	-50.28
CHEMBL3125251	-48.51
CHEMBL3125252	-48.96
CHEMBL3125259	-49.22

As previously stated, detecting activity cliffs in compound datasets is critical in shaping the development of prediction models. It

is suggested that excluding activity cliffs from these datasets may improve model accuracy, particularly for those that use the similarity principle, such as traditional QSAR techniques. It would still be necessary to develop and validate several prediction models with and without activity cliffs for the compound data set examined in this work and to evaluate the predictive potential quantitatively. Furthermore, in the case of many studies that identified pyrrolidine carboxamide as potent direct InhA inhibitors, the predicted activity cliff compounds can help highlight important molecular features, resulting in the development of more potent compounds against InhA.

4.4. Conclusion

This research work provides an in-depth cheminformatic analysis of the structure-activity relationship (SAR) of InhA inhibitors sourced from the CHEMBL database, which includes both activity landscape modeling and structure-based docking studies. The evaluation of the inhibitors' activity landscape revealed a largely heterogeneous SAR, with most compound pairs falling into similarity cliff regions, while some are in activity cliff regions. Quantitative analysis indicated that activity cliffs represent approximately 0.3% of the dataset, highlighting their rarity. However, these cliffs are crucial for understanding SAR, as they introduce sharp changes in activity with minimal structural alterations, which can complicate predictive modeling. We identified ten significant activity cliff generators within the InhA inhibitor dataset, all of which contain important pharmacophoric features that contribute to their potency. Analysis of

these cliffs and compound pairs showed that they are all pyrrolidine carboxamide derivatives, a novel class of InhA inhibitors. Notably, the IC_{50} value decreases when the carboxamide substituent is replaced by a nitrogen-containing group, particularly heterocyclic groups. Therefore, SAR studies of these compounds are critical for drug development. Docking analysis revealed that minor structural modifications in the ligands enhance their interactions with the protein. This increased interaction is attributed to the nitrogen-containing substituent forming various types of bonds, and these subtle structural changes are responsible for the activity cliffs between compounds, directly impacting their IC_{50} values against InhA. Additionally, molecular dynamics simulations and MM-PBSA calculations confirmed the docking results, showing lower RMSD and free energy values. These findings indicate that activity cliff compounds bind more effectively to the protein pocket. Identifying and addressing these activity cliffs could enhance the development of more efficient and reliable QSAR models for InhA inhibitors, as well as a better understanding of the key molecular features required for InhA drug development.

References

[1] A. Ece, Computer-aided drug design, BMC Chem. 17 (2023) 26. https://doi.org/10.1186/s13065-023-00939-w.

[2] S.K. Niazi, Z. Mariam, Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis, Pharmaceuticals. 17 (2024). https://doi.org/10.3390/ph17010022.

[3] U.H. Manjunatha, S.P. S. Rao, R.R. Kondreddi, C.G. Noble, L.R. Camacho, B.H. Tan, S.H. Ng, P.S. Ng, N.L. Ma, S.B. Lakshminarayana, M. Herve, S.W. Barnes, W. Yu, K. Kuhen, F. Blasco, D. Beer, J.R. Walker, P.J. Tonge, R. Glynne, P.W. Smith, T.T. Diagana, Direct inhibitors of InhA are active against Mycobacterium tuberculosis, Sci. Transl. Med. 7 (2015) 269ra3-269ra3. https://doi.org/10.1126/scitranslmed.3010597.

[4] M. Sabbah, V. Mendes, R.G. Vistal, D.M.G. Dias, M. Záhorszká, K. Mikušová, J. Korduláková, A.G. Coyne, T.L. Blundell, C. Abell, Fragment-Based Design of Mycobacterium tuberculosis InhA Inhibitors, J. Med. Chem. 63 (2020) 4749–4761. https://doi.org/10.1021/acs.jmedchem.0c00007.

[5] P. Pan, S.E. Knudson, G.R. Bommineni, H.-J. Li, C.-T. Lai, N. Liu, M. Garcia-Diaz, C. Simmerling, S.S. Patil, R.A. Slayden, P.J. Tonge, Time-Dependent Diaryl Ether Inhibitors of InhA: Structure–Activity Relationship Studies of Enzyme Inhibition, Antibacterial Activity, and in vivo Efficacy, ChemMedChem. 9 (2014) 776–791. https://doi.org/https://doi.org/10.1002/cmdc.201300429.

[6] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities, Nucleic Acids Res. 35 (2007) D198–D201. https://doi.org/10.1093/nar/gkl999.

[7] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem Substance and Compound databases, Nucleic Acids Res. 44 (2016) D1202–D1213. https://doi.org/10.1093/nar/gkv951.

[8] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107. https://doi.org/10.1093/nar/gkr777.

[9] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242. https://doi.org/10.1093/nar/28.1.235.

[10] P. Sidorov, B. Viira, E. Davioud-Charvet, U. Maran, G. Marcou, D. Horvath, A. Varnek, QSAR modeling and chemical space analysis of antimalarial compounds, J. Comput. Aided. Mol. Des. 31 (2017) 441–451. https://doi.org/10.1007/s10822-017-0019-4.

[11] J.-L. Reymond, The Chemical Space Project, Acc. Chem. Res. 48 (2015) 722–730. https://doi.org/10.1021/ar500432k.

[12] S. Ekins, A.M. Clark, K. Dole, K. Gregory, A.M. Mcnutt, A.C. Spektor, C. Weatherall, N.K. Litterman, B.A. Bunin, Data Mining and Computational Modeling of High-Throughput Screening Datasets BT - Reporter Gene Assays: Methods and Protocols, in: R. Damoiseaux, S. Hasson (Eds.), Springer New York, New York, NY, 2018: pp. 197–221. https://doi.org/10.1007/978-1-4939-7724-6_14.

[13] V. Bhaskar, S. Kumar, A. Sujathan Nair, S. Gokul, P. Rajappan Krishnendu, S. Benny, C.T. Amrutha, D.S. Manisha, V. Bhaskar, S. Mary Zachariah, T.P. Aneesh, M.A. Abdelgawad, M.M. Ghoneim, L.K. Pappachen, O. Nicolotti, B. Mathew, In silico development of potential InhA inhibitors through 3D-QSAR analysis, virtual screening and molecular dynamics, J. Biomol. Struct. Dyn. (n.d.) 1–23. https://doi.org/10.1080/07391102.2023.2291549.

[14] S.R. Prem Kumar, I.A. Shaikh, M.H. Mahnashi, M.A. Alshahrani, S.R. Dixit, V.H. Kulkarni, C. Lherbet, A.K. Gadad, T.M. Aminabhavi, S.D. Joshi, Design, synthesis and computational approach to study novel pyrrole scaffolds as active inhibitors of enoyl ACP reductase (InhA) and

Mycobacterium tuberculosis antagonists, J. Indian Chem. Soc. 99 (2022) 100674. https://doi.org/https://doi.org/10.1016/j.jics.2022.100674.

[15] S.D. Joshi, U.A. More, D. Koli, M.S. Kulkarni, M.N. Nadagouda, T.M. Aminabhavi, Synthesis, evaluation and in silico molecular modeling of pyrroyl-1,3,4-thiadiazole inhibitors of InhA, Bioorg. Chem. 59 (2015) 151–167. https://doi.org/https://doi.org/10.1016/j.bioorg.2015.03.001.

[16] J. Bajorath, Modeling of activity landscapes for drug discovery, Expert Opin. Drug Discov. 7 (2012) 463–473. https://doi.org/10.1517/17460441.2012.679616.

[17] J.L. Medina-Franco, Activity Cliffs: Facts or Artifacts?, Chem. Biol. Drug Des. 81 (2013) 553–556. https://doi.org/https://doi.org/10.1111/cbdd.12115.

[18] M. Cruz-Monteagudo, J.L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M.N.D.S. Cordeiro, F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?, Drug Discov. Today. 19 (2014) 1069–1080. https://doi.org/https://doi.org/10.1016/j.drudis.2014.02.003.

[19] J.L. Medina-Franco, Interrogating Novel Areas of Chemical Space for Drug Discovery using Chemoinformatics, Drug Dev. Res. 73 (2012) 430–438. https://doi.org/https://doi.org/10.1002/ddr.21034.

[20] D. Stumpfe, H. Hu, J. Bajorath, Evolving Concept of Activity Cliffs, ACS Omega. 4 (2019) 14360–14368. https://doi.org/10.1021/acsomega.9b02221.

[21] M. Wawer, E. Lounkine, A. Wassermann, J. Bajorath, Data structures and computational tools for the extraction of SAR information from large compound sets, Drug Discov. Today. 15 (2010) 630–639. https://doi.org/10.1016/j.drudis.2010.06.004.

[22] A. Golbraikh, X.S. Wang, H. Zhu, A. Tropsha, Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment BT - Handbook of Computational Chemistry, in: J. Leszczynski (Ed.), Springer Netherlands,

Dordrecht, 2016: pp. 1–38. https://doi.org/10.1007/978-94-007-6169-8_37-2.

[23] S. Pirhadi, F. Shiri, J.B. Ghasemi, Multivariate statistical analysis methods in QSAR, RSC Adv. 5 (2015) 104635–104665. https://doi.org/10.1039/C5RA10729F.

[24] S. Ahamed, K. Muraleedharan, Towards a systematic analysis of structure-activity relationships of 5-LOX inhibitors through activity landscape and chemotype enrichment, Chemom. Intell. Lab. Syst. 207 (2020) 104188. https://doi.org/10.1016/j.chemolab.2020.104188.

[25] E. López-López, O. Rabal, J. Oyarzabal, J.L. Medina-Franco, Towards the understanding of the activity of G9a inhibitors: an activity landscape and molecular modeling approach, J. Comput. Aided. Mol. Des. 34 (2020) 659–669. https://doi.org/10.1007/s10822-020-00298-x.

[26] E. López-López, F.D. Prieto-Martínez, J.L. Medina-Franco, Activity Landscape and Molecular Modeling to Explore the SAR of Dual Epigenetic Inhibitors: A Focus on G9a and DNMT1, Molecules. 23 (2018). https://doi.org/10.3390/molecules23123282.

[27] J.J. Naveja, U. Norinder, D. Mucs, E. López-López, J.L. Medina-Franco, Chemical space, diversity and activity landscape analysis of estrogen receptor binders, RSC Adv. 8 (2018) 38229–38237. https://doi.org/10.1039/C8RA07604A.

[28] O. Méndez-Lucio, J. Pérez-Villanueva, R. Castillo, J. Medina-Franco, Identifying activity cliff generators of PPAR ligands using SAS Maps, Mol. Inform. 31 (2012) 837–846. https://doi.org/10.1002/minf.201200078.

[29] T. Sander, J. Freyss, M. von Korff, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, J. Chem. Inf. Model. 55 (2015) 460–473. https://doi.org/10.1021/ci500588j.

[30] D. Stumpfe, H. Hu, J. Bajorath, Advances in exploring activity cliffs, J. Comput. Aided. Mol. Des. 34 (2020) 929–942.

https://doi.org/10.1007/s10822-020-00315-z.

[31] M. Vogt, Progress with modeling activity landscapes in drug discovery, Expert Opin. Drug Discov. 13 (2018) 605–615. https://doi.org/10.1080/17460441.2018.1465926.

[32] M. González-Medina, O. Méndez-Lucio, J.L. Medina-Franco, Activity Landscape Plotter: A Web-Based Application for the Analysis of Structure–Activity Relationships, J. Chem. Inf. Model. 57 (2017) 397–402. https://doi.org/10.1021/acs.jcim.6b00776.

[33] S. Chhabra, S. Kumar, R. Parkesh, Chemical Space Exploration of DprE1 Inhibitors Using Chemoinformatics and Artificial Intelligence, ACS Omega. 6 (2021) 14430–14441. https://doi.org/10.1021/acsomega.1c01314.

[34] D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?, J. Cheminform. 7 (2015) 20. https://doi.org/10.1186/s13321-015-0069-3.

[35] J.J. Naveja, J.L. Medina-Franco, Activity landscape of DNA methyltransferase inhibitors bridges chemoinformatics with epigenetic drug discovery, Expert Opin. Drug Discov. 10 (2015) 1059–1070. https://doi.org/10.1517/17460441.2015.1073257.

[36] D. Stumpfe, Y. Hu, D. Dimova, J. Bajorath, Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry, J. Med. Chem. 57 (2013). https://doi.org/10.1021/jm401120g.

[37] M. Dablander, T. Hanser, R. Lambiotte, G. Morris, Exploring QSAR Models for Activity-Cliff Prediction, 2023.

[38] A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, Data set modelability by QSAR, J. Chem. Inf. Model. 54 (2013). https://doi.org/10.1021/ci400572x.

[39] Y. Liu, X. Yang, J. Gan, S. Chen, Z.-X. Xiao, Y. Cao, CB-Dock2: improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting,

Nucleic Acids Res. 50 (2022) W159–W164. https://doi.org/10.1093/nar/gkac394.

[40] Y. Liu, M. Grimm, W. Dai, M. Hou, Z.-X. Xiao, Y. Cao, CB-Dock: a web server for cavity detection-guided protein–ligand blind docking, Acta Pharmacol. Sin. 41 (2020) 138–144. https://doi.org/10.1038/s41401-019-0228-6.

[41] A.C. Wallace, R.A. Laskowski, J.M. Thornton, LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions, Protein Eng. Des. Sel. 8 (1995) 127–134. https://doi.org/10.1093/protein/8.2.127.

[42] M. Al Shahrani, R.M. Gahtani, M. Makkawi, C-5401331 identified as a novel T-cell immunoglobulin and mucin domain-containing protein 3 (Tim-3) inhibitor to control acute myeloid leukemia (AML) cell proliferation, Med. Oncol. 41 (2024) 63. https://doi.org/10.1007/s12032-023-02296-z.

[43] H. Elsir Khair, B. Ahmed Mohamed, B. Yousef Nour, H. Ali Waggiallah, Prevalence of BCR-ABL T315I Mutation in Different Chronic Myeloid Leukemia patients Categories, Pakistan J. Biol. Sci. PJBS. 25 (2022) 175–181. https://doi.org/10.3923/pjbs.2022.175.181.

[44] E. Larocque, N. Naganna, C. Opoku-Temeng, A. Lambrecht, H. Sintim, Front Cover: Alkynylnicotinamide-Based Compounds as ABL1 Inhibitors with Potent Activities against Drug-Resistant CML Harboring ABL1(T315I) Mutant Kinase (ChemMedChem 12/2018), ChemMedChem. 13 (2018) 1159. https://doi.org/10.1002/cmdc.201800278.

[45] Q. Zhao, Z.E. Wu, B. Li, F. Li, Recent advances in metabolism and toxicity of tyrosine kinase inhibitors, Pharmacol. & Ther. 237 (2022) 108256. https://doi.org/10.1016/j.pharmthera.2022.108256.

[46] Y. Teneva, R. Simeonova, V. Valcheva, V.T. Angelova, Recent Advances in Anti-Tuberculosis Drug Discovery Based on Hydrazide–Hydrazone and Thiadiazole Derivatives Targeting InhA, Pharmaceuticals. 16 (2023). https://doi.org/10.3390/ph16040484.

[47] E.F. Khaleel, A. Sabt, M. Korycka-Machala, R.M. Badi, N.T. Son, N.X. Ha, M.F. Hamissa, A.E. Elsawi, E.B. Elkaeed, B. Dziadek, W.M. Eldehna, J. Dziadek, Identification of new anti-mycobacterial agents based on quinoline-isatin hybrids targeting enoyl acyl carrier protein reductase (InhA), Bioorg. Chem. 144 (2024) 107138. https://doi.org/https://doi.org/10.1016/j.bioorg.2024.107138.

[48] T. Matviiuk, J. Madacki, G. Mori, B.S. Orena, C. Menendez, A. Kysil, C. André-Barrès, F. Rodriguez, J. Korduláková, S. Mallet-Ladeira, Z. Voitenko, M.R. Pasca, C. Lherbet, M. Baltas, Pyrrolidinone and pyrrolidine derivatives: Evaluation as inhibitors of InhA and Mycobacterium tuberculosis, Eur. J. Med. Chem. 123 (2016) 462–475. https://doi.org/https://doi.org/10.1016/j.ejmech.2016.07.028.

[49] X. He, A. Alian, R. Stroud, P.R. Ortiz de Montellano, Pyrrolidine Carboxamides as a Novel Class of Inhibitors of Enoyl Acyl Carrier Protein Reductase from Mycobacterium tuberculosis, J. Med. Chem. 49 (2006) 6308–6323. https://doi.org/10.1021/jm060715y.

[50] M. Prasad, R. Bhole, P. Khedekar, R. Chikhale, Mycobacterium enoyl acyl carrier protein reductase (InhA): A key target for antitubercular drug discovery, Bioorg. Chem. 115 (2021) 105242. https://doi.org/10.1016/j.bioorg.2021.105242.

[51] L. Encinas, H. O'Keefe, M. Neu, M.J. Remuiñán, A.M. Patel, A. Guardia, C.P. Davie, N. Pérez-Macías, H. Yang, M.A. Convery, J.A. Messer, E. Pérez-Herrán, P.A. Centrella, D. Álvarez-Gómez, M.A. Clark, S. Huss, G.K. O'Donovan, F. Ortega-Muro, W. McDowell, P. Castañeda, C.C. Arico-Muendel, S. Pajk, J. Rullás, I. Angulo-Barturen, E. Álvarez-Ruíz, A. Mendoza-Losana, L. Ballell Pages, J. Castro-Pichel, G. Evindar, Encoded Library Technology as a Source of Hits for the Discovery and Lead Optimization of a Potent and Selective Class of Bactericidal Direct Inhibitors of Mycobacterium tuberculosis InhA, J. Med. Chem. 57 (2014) 1276–1288. https://doi.org/10.1021/jm401326j.

[52] P. Gupta, K. Tilekar, N. Upadhyay, R. C. S., Recent Discoveries

Of Nitrogen-Containing Heterocyclic Compounds as InhA Inhibitors Against Mycobacterium Tuberculosis: An Overview, Infect. Disord. Drug Targets. 22 (2022). https://doi.org/10.2174/1871526522666220420092618.

[53] X.-Y. Meng, M.-X. Song, M. Mezei, M. Cui, Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery, Curr. Comput. Aided. Drug Des. 7 (2011) 146–157. https://doi.org/10.2174/157340911795677602.

[54] N. Furtmann, Y. Hu, M. Gütschow, J. Bajorath, Identification of Interaction Hot Spots in Structures of Drug Targets on the Basis of Three-Dimensional Activity Cliff Information, Chem. Biol. Drug Des. 86 (2015) 1458–1465. https://doi.org/https://doi.org/10.1111/cbdd.12605.

Chapter 5

Machine learning-based QSAR classification modeling and screening of Indian medicinal plants against InhA

5.1. Introduction

With over a million fatalities per year, tuberculosis (TB), which is brought on by *Mycobacterium tuberculosis* (Mtb), continues to rank among the world's most deadly infectious diseases [1]. The development of extensively drug-resistant (XDR) and multi-drug-resistant (MDR) TB strains has significantly reduced the effectiveness of traditional antibiotics, despite the availability of treatment regimens. This necessitates the urgent development of new anti-tubercular drugs [2,3]. The enzyme known as InhA, or Enoyl-Acyl Carrier Protein (ACP) Reductase from Mtb, is involved in the biosynthesis of fatty acids, primarily mycolic acid, which is a significant constituent of mycobacterial cell walls [4]. InhA has proven to be one of the most reliable targets for developing drugs against TB or InhA inhibitors. Therefore, we decided to search for novel inhibitors of InhA, which is the target of isoniazid, a first-line TB drug (also known as isonicotinoic acid hydrazide [INH]) [5].

Antimicrobials and other therapeutic compounds have long been found in natural products, especially phytochemicals made from Indian medicinal plants [6]. Given its distinct biodiversity and abundance of medicinal plants, India offers a promising prospect for the development of anti-TB medications [7]. However, conventional discovery methods are time-consuming and prohibitively expensive due to the wide range of chemicals and the limited capacity for experimental screening. By identifying patterns in current bioactivity data, Quantitative Structure–Activity Relationship (QSAR) modeling and machine learning (ML) techniques provide a data-driven framework to speed up the

identification of possible bioactive compounds [8]. ML-based QSAR models can narrow down candidates for experimental validation by predicting the efficacy of untested compounds by correlating molecular descriptors with anti-TB activity [9]. The primary objective of QSAR analysis is to connect a collection of chemical descriptors (predictor variable, X) to activity (response variable, Y). Methods for connecting X and Y and chemical descriptors have drawn a lot of research interest [10].

In this work, we introduce an integrated ML-based QSAR pipeline that trains and assesses predictive models that can distinguish between active and inactive compounds against Mtb using the bioactivity data available in the ChEMBL database [11]. Following that, a library of phytochemicals found in Indian medicinal plants downloaded from plant-based databases is screened using the trained model. Finally, molecular docking was performed on screened molecules to further identify the best inhibitors of InhA, based on protein-ligand interaction. The goal of this strategy is to use the combination of contemporary computational methods and ethnopharmacological knowledge to find new leads in the fight against TB. To improve QSAR modeling and virtual screening and more effectively identify lead compounds against complex targets like Mtb, recent developments in computational drug discovery have increasingly depended on machine learning techniques [12–14]. The ligand and structure-based approaches used in this work are expected to be beneficial in the discovery and development of effective InhA inhibitors [15].

5.2. Materials and Methods

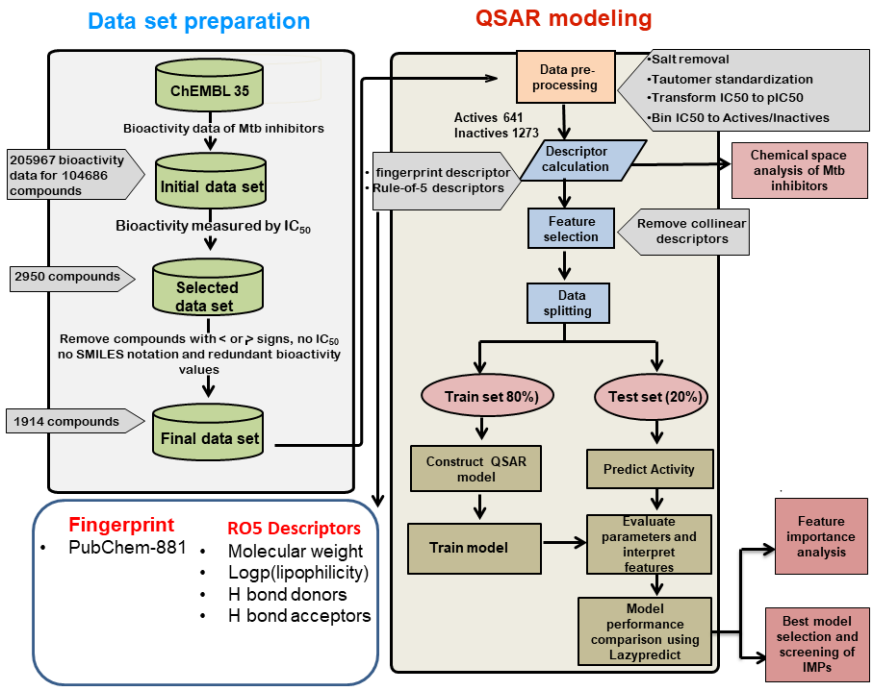


Fig.5.1. A flowchart of the QSAR modeling and evaluation process

5.2.1. Data Collection and Curation

The ChEMBL database, version 35, which is accessible to the public, provided the bioactivity information for Mycobacterium tuberculosis (Mtb) inhibitors (Target ID: CHEMBL 360) used in this investigation. The original dataset contained 205,967 bioactivity records, representing 104,686 distinct compounds, including MIC, IC_{50}, MIC50, % inhibition, % activity, and more. We selected 2950 compounds with IC_{50} values for our subsequent study. To maintain reliability and consistency in downstream modelling, compounds lacking IC_{50} measurements, having ambiguous inequality signs such as

'<' or '>', missing SMILES representations, or presenting redundant entries were carefully removed. This rigorous curation process resulted in a refined dataset containing 1,914 structurally and bio actively validated compounds suitable for quantitative structure–activity relationship (QSAR) classification modelling. A schematic representation of the workflow of this study is provided in **Fig. 5.1.**

5.2.2 Molecular Descriptors

Molecular descriptors are mathematical values that characterize part of a molecule or structural or physicochemical properties of a molecule. They are used in QSAR models to predict a compound's molecular properties or biological activity [16,17]. In this study, two types of descriptors were generated to characterize the chemical space of the curated compounds and their modelling. The first included PubChem-based fingerprint descriptors, which capture binary substructure patterns useful for chemical similarity and bioactivity profiling. PaDel-Discriptor software was used to generate 881 PubChem molecular descriptors from the SMILES formula using the Python code [18,19]. PubChem fingerprints show whether a particular set of chemical features is present in a compound or not. These fingerprints were further subjected to interpretation of the feature importance [20].

The second category included physicochemical descriptors based on Lipinski's Rule-of-Five (RO5), which included molecular weight (MW), LogP (a measure of lipophilicity), the number of hydrogen bond donors (HBD), and the number of hydrogen bond

acceptors (HBA). These descriptors were chosen based on their relevance in drug-likeness evaluation and interpretability in QSAR modelling [21]. We used RDKit to generate Lipinski descriptors. RO5 descriptors were used for the chemical space analysis of drug-like properties of inhibitors.

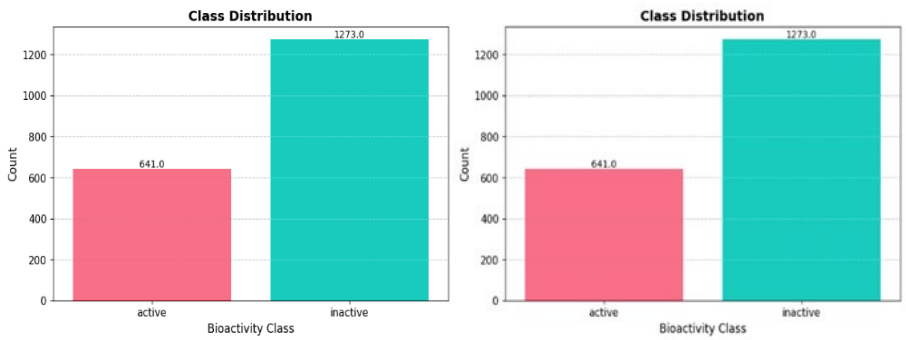
The selection of PubChem fingerprint descriptors was further justified based on the known mechanism of InhA inhibition. InhA (enoyl-acyl carrier protein reductase) plays a critical role in the fatty acid elongation cycle involved in mycolic acid biosynthesis in Mycobacterium tuberculosis. The active site of InhA contains a hydrophobic substrate-binding pocket that accommodates long-chain fatty acyl substrates and NADH-dependent inhibitors. Therefore, structural fingerprints capturing specific substructures and functional groups are important for identifying chemical patterns that may influence enzyme binding and inhibition.

5.2.3. Data Filtering and Preprocessing

Several preprocessing stages were carried out in order to get the data ready for analysis based on machine learning. To guarantee consistency among chemical representations, salts were first eliminated from compound structures, and tautomers were standardized. A negative base-10 logarithmic transformation was then used to convert the IC_{50} values to pIC_{50} values in order to provide a normalized, continuous range of bioactivity values. A threshold on pIC_{50} values <1000 and >1000 nM, respectively, was then used to classify the compounds as actives and inactives. There were 1,273 inactive

compounds and 641 active compounds in the final distribution. Then this dataset was subjected to the generation of molecular descriptors.

After the generation of descriptors Synthetic Minority Oversampling Technique (SMOTE) was used to enhance the minority class in the training dataset in order to counteract the impact of class imbalance, which might skew the model in favor of the majority class. This approach mitigates the overfitting issue caused by random oversampling. The component generates new instances using existing minority cases as input. SMOTE works by choosing samples that are near in the feature space, drawing a line between them, and drawing a new sample at a position along that line [22]. **Fig.5.2** shows the bar plots of active and inactive class distribution before and after resampling with SMOTE.



A **B**

Fig.5.2. Bar plots of class distribution. (A) Before Resampling (B) After Resampling with SMOTE

The balanced dataset was then subjected to feature selection. In order to reduce duplication and multicollinearity, highly collinear properties were removed from the descriptor dataset. Statistical

characteristics were chosen using the Chi-Square test, which made it possible to find and keep descriptors that were substantially linked to chemical activity. The chi-square test is a statistical approach for selecting features in machine learning [23]. It's employed to ascertain whether categorical variables significantly correlate with one another. The Chi-square statistic is calculated as **Equation 5.1**:

$$\chi 2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad \textbf{(5.1)}$$

O_i = observed frequency, E_i = expected frequency

5.2.4. Data Splitting and Test Selection

Following feature selection, with 80% of the final dataset going to training and 20% going to independent testing, the stratified train–test split ensured robust model validation by maintaining the relative distribution of active and inactive chemicals in both groups. The unseen test set was put aside for the external evaluation of model performance, while the training set was used for model development and hyperparameter tuning. GridSearchCV was used to tune the hyperparameters in the selected models. It is performed to improve the efficiency of the model [24].

5.2.5. QSAR Classification Model Building

An ML-based QSAR classification modeling framework was employed to classify compounds based on their potential Mtb inhibitory activity. ML Classification is the approach used to predict the class of data points. Model selection was initiated using the

LazyPredict classifier suite, which provided a comparative performance analysis across a wide range of supervised learning algorithms, including but not limited to Random Forest, Support Vector Machine, and Gradient Boosting classifiers [25]. The classifier demonstrating superior performance metrics on the training data was selected for further optimization.

5.2.6. Statistical Assessment for Model Validation and Performance

Accuracy, precision, recall, and F1-score, which are the standard classification metrics, were used to statistically evaluate the model's performance. These measures made it easier to conduct a thorough assessment of the QSAR model's external validity and internal consistency. The percentage of all correct classifications, whether positive or negative, is known as accuracy. Precision indicates the proportion of predicted positive cases that are actually positive. Recall represents the proportion of true positives identified out of all positive instances. F1-score balances the trade-off between precision and recall by combining them into a single metric. They use the formulae provided in **Equations 5.2 - 5.6.** A confusion matrix was also utilised to examine the classification results, displaying the number of correct (TP, TN) and incorrect (FP, FN) classifications.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \textbf{(5.2)}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \textbf{(5.3)}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \textbf{(5.4)}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \textbf{(5.5)}$$

TP, TN, FP, and FN stand for the number of true positive, true negative, false positive, and false negative cases, respectively.

Google Colab, a cloud-based Jupyter notebook environment that offers free access to GPUs and TPUs, was used to execute the modeling and screening. This makes it perfect for ML tasks. Python 3 with necessary libraries, including scikit-learn for machine learning algorithms, RDKit for calculating molecular descriptors, Pandas and NumPy for data manipulation, Matplotlib and Seaborn for visualization, and imbalanced-learn for SMOTE-based data balancing, was used for the modeling and analysis. During model training and evaluation, the environment made use of Colab's backend, which had a Tesla K80 GPU and 12 GB of RAM for effective computation.

5.2.7. Screening of Indian medicinal plants Molecules

After validation, the best machine learning model was used to screen phytochemicals in Indian medicinal plants (IMPs). We chose 15 traditional IMPs with anti-tubercular activity from the literature survey shown in **Fig. 5.3.** Several in vitro studies identified that these plants molecules show potential anti-tubercular activities. We downloaded a list of 769 phytochemicals found in IMPs from Dr. Duke's Phytochemical and Ethnobotanical databases [26]. Then, 3D structure molecules were downloaded from the PubChem and IMPPAT databases [27,28]. Then duplicates were removed, and molecular descriptors of phytochemicals were computed and fed into the trained

model, which predicted their biological activity. The predicted results were analysed to identify promising candidates using the molecular docking method.

Fig.5.3. Selected 15 Indian Medicinal plants for the screening of molecules against Mtb

5.2.8. Molecular docking

The crystal structure of InhA (PDB ID: 1ZID) at 2.70 Å resolution was downloaded from the Protein Data Bank database [29]. The protein was then prepared using the Discovery Studio software. All water molecules and heteroatoms were removed from the protein. Finally, kollman charges were added, and the protein was exported in pdbqt

format for docking analysis. Previously screened plant molecules for anti InhA activity converted to pdbqt format with Open Babel [30].

The AutoDock Vina software was then used to conduct the molecular docking analysis. The compounds were evaluated based on their docking scores and interaction profiles, and the best candidates were selected.

5.3. Result and discussion

5.3.1. Chemical space analysis of Mtb inhibitors

The training set of Mtb inhibitors was first subjected to chemical space analysis to gain a better understanding of the dataset before developing the classification model. The dataset's chemical space was examined using Lipinski RO5 (rule of five) descriptors, such as MW, LogP, HBD, and HBA, to comprehend the structure-activity relationship (SAR). In practice, RO5 is most commonly used to assess drug-likeness, which helps choose compounds with a better chance of success. This rule offers straightforward criteria for determining whether a chemical compound with a particular pharmacological or biological activity possesses characteristics that make it an orally active drug. In drug development, the RO5 predicts that poor absorption or penetration is more likely when there are more than 5 HBD, 10 HBA, and the MW is larger than 500 Dalton (Da), and the computed Log P is greater than 5.

Initially, the chemical space of the dataset was examined by plotting the distribution of actives and inactives as a scatter plot of MW vs. LogP. The active and inactive compounds are then compared based

on Lipinski descriptors. **Fig. 5.4** depicts the chemical space distribution of the training set in the form of a scatter plot. The figure indicated a nearly the same distribution, with the majority of active and inactive compounds having a MW between 200 and 600 Da and a LogP value between 0 and 7. This result suggests that actives and inactives in the training set shared the same chemical space.

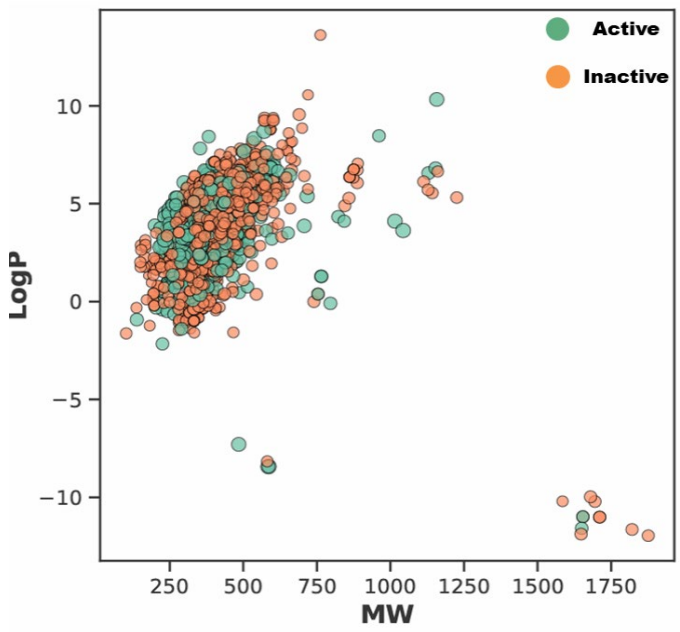


Fig. 5.4. Chemical space of the Mtb inhibitor dataset. Molecular weight on the X-axis, logP on the Y-axis

Fig. 5.5 depicts the total distribution of the dataset based on RO5. **Fig. 5.5A** demonstrates that active molecules have somewhat lower MW values than inactive ones. It may be deduced from the median values in the box plot. Similarly, inactive compounds showed somewhat higher logP values than active compounds (**Fig. 5.5B**). The distribution of HBA, as determined by the median, reveals that the

active and inactive drugs had almost identical HBA values (**Fig. 5.5C**). The distribution of HBD demonstrates that active compounds had a lower HBD value than active compounds (**Fig. 5.5D**). However, practically all compounds in the training set exhibit drug-like features, making it challenging to predict inhibitor activity using basic Lipinski molecular descriptors due to the same distribution of actives and inactives.

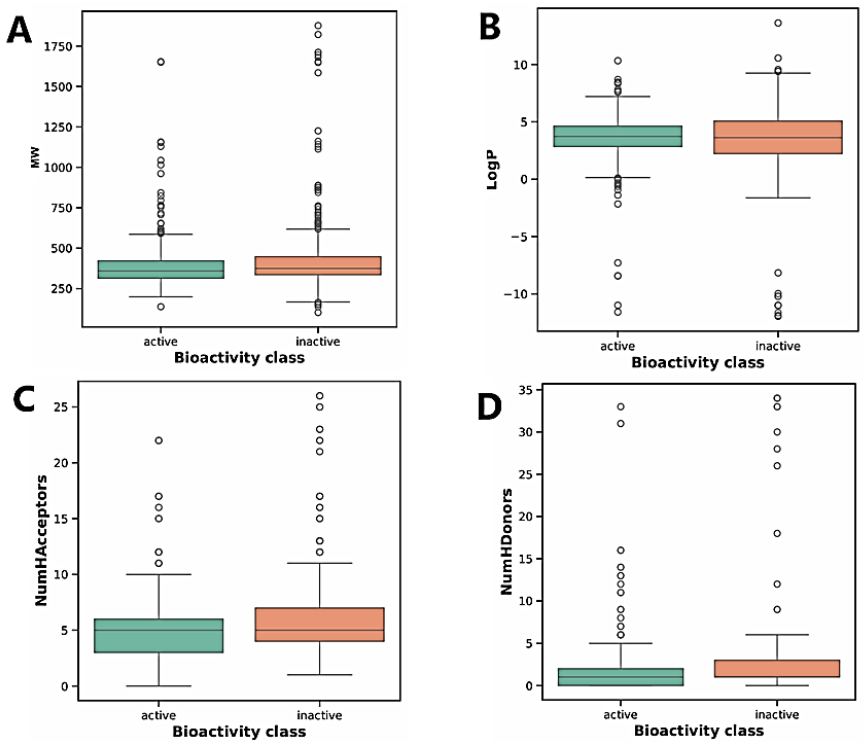


Fig 5.5. Box plots of drug likeness evaluation of Mtb inhibitors. **(A)** Molecular weight **(B)** LogP **(C)** Hydrogen bond acceptors **(D)** Hydrogen bond donors

5.3.2. Machine Learning based QSAR Screening Model

After the preprocessing and fingerprint descriptor calculation the LazyPredict package was used to acquire robust QSAR ML models, as described in methods. A comparison of many machine learning models assessed with the LazyPredict library is shown in **Table 5.1.** At 0.91 for accuracy, balanced accuracy, and F1 score, Light Gradient Boosting Machine Classifier (LGBM Classifier) outperformed the other evaluated classifiers in every important statistic. LGBM is a high-performance gradient boosting framework that uses decision tree algorithms. It is used for ranking, classification, and other ML tasks. With just slightly lower scores, ExtraTreesClassifier and BaggingClassifier likewise showed excellent performance. These findings demonstrate the effectiveness and resilience of ensemble-based techniques in Mtb inhibitors bioactivity prediction-related binary classification tasks.

Using GridSearchCV search techniques, hyperparameter tweaking was carried out to improve the predictive power and generalizability of the model. Using the training and testing datasets, the final optimized model referred to as the best model was put through a thorough review. While classifiers such as LabelPropagation and GaussianNB had notable disparities, the majority of the best-performing models maintained constant accuracy and F1 scores, as seen in **Fig.5.6**.

The training duration for each model is shown in **Fig.5.7**. Although CalibratedClassifierCV performed well, it took the longest to

train (42.3 seconds), which made it less useful than models like RidgeClassifierCV and LogisticRegression that were faster but still as accurate.

Table 5.1. Model Performance comparison results using LazyPredict

Sl. No	Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken (s)
1	LGBMClassifier	0.91	0.91	0.91	0.72
2	ExtraTreesClassifier	0.9	0.9	0.9	1.21
3	BaggingClassifier	0.89	0.89	0.89	4.28
4	RandomForestClassifier	0.89	0.89	0.89	1.03
5	CalibratedClassifierCV	0.88	0.88	0.88	64.95
6	LogisticRegression	0.88	0.88	0.88	0.53
7	SVC	0.87	0.87	0.87	1.95
8	NuSVC	0.87	0.87	0.87	3.22
9	KNeighborsClassifier	0.86	0.86	0.86	0.34
10	RidgeClassifierCV	0.86	0.86	0.86	1.22
11	LinearSVC	0.86	0.86	0.86	12.06
12	RidgeClassifier	0.86	0.86	0.86	0.47
13	LinearDiscriminantAnalysis	0.85	0.85	0.85	1.28
14	DecisionTreeClassifier	0.85	0.85	0.85	0.55
15	ExtraTreeClassifier	0.85	0.85	0.85	0.24
16	Passive Aggressive Classifier	0.85	0.85	0.85	0.63
17	SGDClassifier	0.83	0.83	0.83	0.49
18	Perceptron	0.82	0.82	0.82	0.67
19	QuadraticDiscriminantAnalysis	0.81	0.81	0.81	2.76
20	AdaBoostClassifier	0.75	0.75	0.75	5.31
21	BernoulliNB	0.72	0.72	0.72	1.37
22	NearestCentroid	0.71	0.71	0.71	0.25
23	LabelSpreading	0.65	0.66	0.61	0.8
24	LabelPropagation	0.65	0.66	0.61	0.7
25	GaussianNB	0.59	0.6	0.53	0.24
26	DummyClassifier	0.49	0.5	0.32	0.21

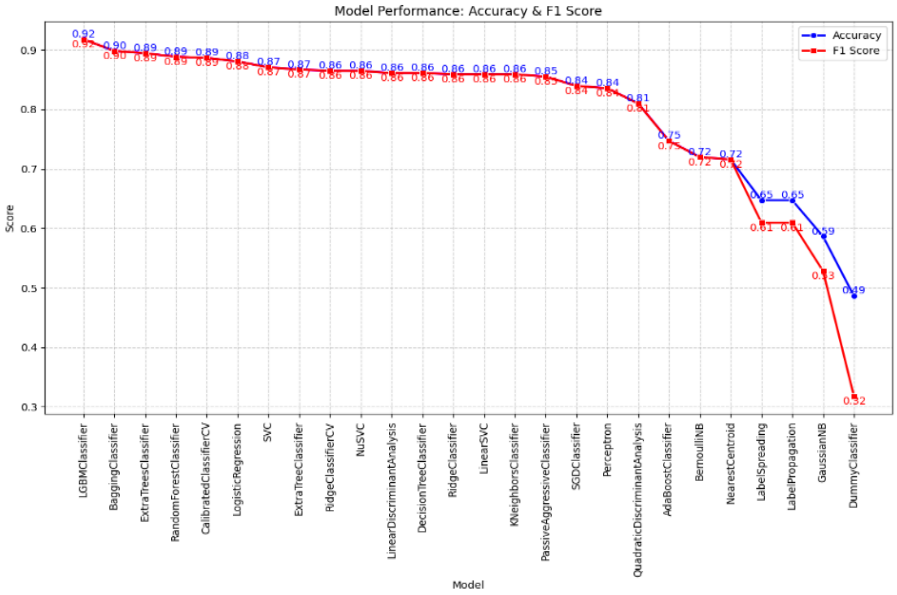


Fig.5.6. Accuracy and F1 Score comparison of various machine learning models generated using LazyPredict.

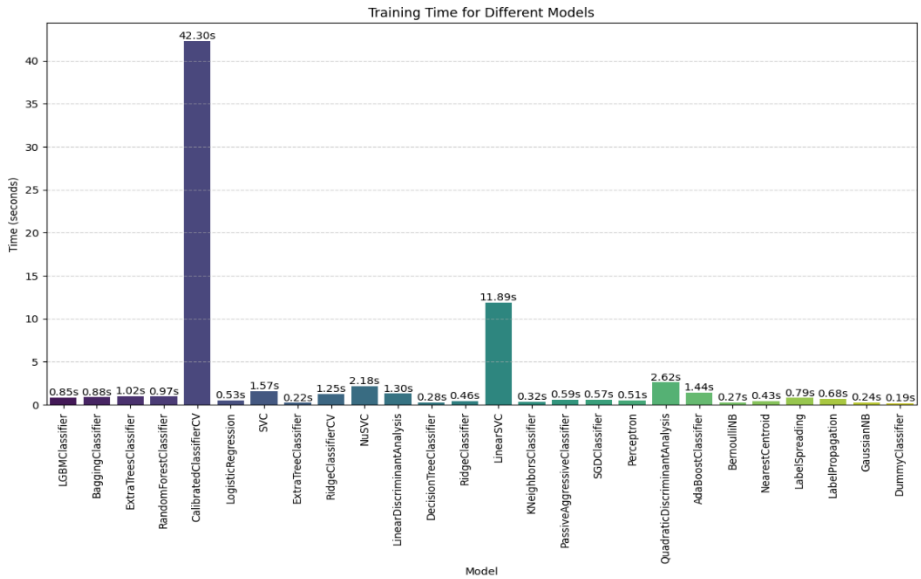


Fig.5.7. Training time (in seconds) for different classifiers evaluated using LazyPredict.

5.3.3. Statistical Assessment for Model Validation and Performance

The unseen test set was then used to evaluate the performance of the selected LGBM classifier QSAR model. Accuracy, precision, recall, and F1-score, which are the standard classification metrics were used to statistically evaluate the QSAR model's performance. These evaluations were performed using the equations described in method **5.2.6**.

The confusion matrix (**Fig.5.8**) shows strong classification performance, with 223 true negatives and 241 true positives, and only a small number of misclassified instances (25 false positives and 21 false negatives). This suggests the model is both sensitive and specific across the active and inactive classes. The classification report (**Table 5.2**) confirms this with balanced precision, recall, and F1-scores of 0.91 for both classes. Such alignment indicates consistent performance across class labels without significant bias toward one class. The macro and weighted averages also equal 0.91, further validating model stability across the dataset.

Expanded metrics in **Table 5.3** reveal an accuracy of 90.98%, a precision of 0.9060, and a sensitivity/recall of 0.9198. The specificity or true negative rate (TNR) of 0.8992 complements this, ensuring both types of classification errors are minimized. The true positive rate (TPR) also stands at 0.9198, emphasizing the model's strength in identifying actual active cases. These results collectively indicate that the LGBM classifier reliable and robust classifier with strong generalization capability. The statistical validation not only supports its

predictive utility but also demonstrates that overfitting is unlikely, given the balanced and high-performing metrics across all evaluation parameters. This LGBM classifier QSAR model was then used to predict the inhibitory activity of molecules of IMPs against Mtb.

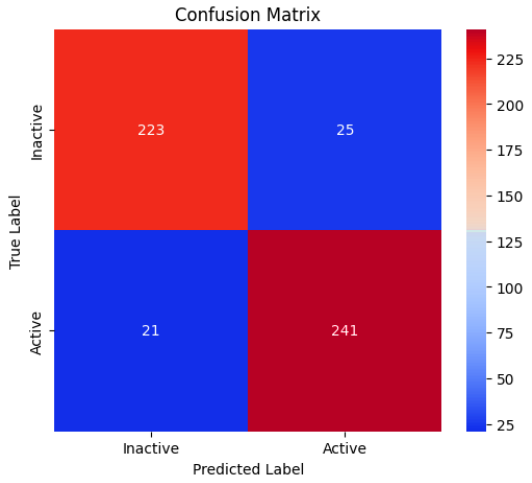


Fig 5.8. Confusion matrix summarizing the classifier's predictions

Table 5.2. The classification report of the actives and inactives

Classification report			
Class	**Precision**	**Recall**	**F1-Score**
active	0.91	0.90	0.91
inactive	0.91	0.92	0.91
accuracy			0.91
macro avg	0.91	0.91	0.91
weighted avg	0.91	0.91	0.91

Table 5.3. The expanded evaluation metrics of the ML-QSAR model

Metrics	Value
Accuracy	0.9098
Precision	0.9060
Recall (Sensitivity)	0.9198
F1 Score	0.9198
True Positive Rate (TPR)	0.9198

TNR = True Negative Rate, TPR = True Positive Rate

5.3.4. Mechanistic Analysis of Feature Importance

Features that are crucial for bioactivity can be found via feature importance analysis. LGBM has two types of feature importance scores: split and gain [31]. To compute the important feature, the default split feature importance method was used. The feature importance from the tweaked LGBM model is computed and visualized as a bar chart, which shows the top 20 most essential features. As demonstrated in **Fig.5.9**, the greatest contributing feature was PubchemFP374, with an importance score of 151. PubchemFP374 feature is a carbon with 3 neighboring hydrogens, regardless of bond type, i.e., CH_3 group. In the context of Mtb drug design, such groups may enhance the membrane permeability, lipophilicity, and proper hydrophobic interactions.

The second most important feature is the presence of two or more oxygen atoms (PubChemFP19) with an importance score of 142. The existence of such a feature in molecules implies that oxygen-rich substructures, such as nitro groups, C=O, and OH, are crucial for drug

inhibition activity. This could be related to binding affinity, solubility, and enzyme interaction potential.

And the third significant feature is PubChemFP866 with an importance score of 115 which is a linear alkyl chain with 8 carbon atoms. **Table 5.2** lists the top 20 feature importance with their respective descriptions. In the dataset, these high-ranking features are thought to be the most statistically significant for class difference, as actives and inactives. These identified important features may be critical for InhA inhibition. A large-scale study focused on compounds classes having these important features could be highly profitable.

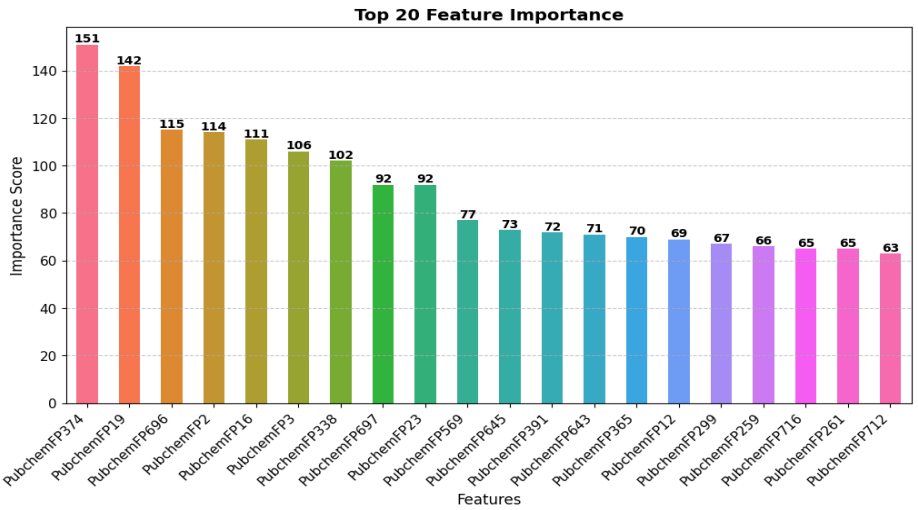


Fig 5.9. Bar Plots of Feature Importance using Chi-square Test

Table 5.4. The lists of the top 20 feature importance with their respective descriptions

Fingerprints	**Description**
PubChemFP374	C(~H)(~H)(~H)
PubChemFP19	>= 2 O
PubChemFP696	C-C-C-C-C-C-C-C
PubChemFP2	>= 16 H
PubChemFP16	>= 4 N
PubChemFP3	>= 32 H
PubChemFP338	C(~C)(~C)(~H)(~N)
PubChemFP697	C-C-C-C-C-C(C)-C
PubChemFP23	>= 1 F
PubChemFP569	N-C-C-N
PubChemFP645	O=C-N-C-C
PubChemFP391	N(~C)(~C)(~C)
PubChemFP643	[#1]-C-C-N-[#1]
PubChemFP365	C(~H)(~N)
PubChemFP12	>= 16 C
PubChemFP299	N-H
PubChemFP259	>= 3 aromatic rings
PubChemFP716	Cc1ccc(N)cc1
PubChemFP261	>= 4 aromatic rings
PubChemFP712	C-C(C)-C(C)-C

[~ indicate the presence of atom nearest neighbor patterns, regardless of bond order, "-", "=", and "#" matches a single bond, double bond, and triple bond]

5.3.5. ML-based screening of medicinal plant molecules

The best LGBM classifier QSAR model was then used to predict the inhibitory activity of phytochemicals present in IMPs against Mtb. **Figure 5.10** illustrates the distribution of predicted bioactivity classes for the screened medicinal plant-derived molecules using the trained machine learning classifier. The bar chart shows that out of the total predictions, 489 compounds were classified as “active” and 280 as “inactive”. This distribution highlights a dominant presence of potentially bioactive compounds in the dataset, suggesting that a significant portion of the phytochemicals may exhibit relevant biological effects against the target.

The slight imbalance in class predictions does not indicate bias, as model evaluation metrics such as precision, recall, and F1-score (refer to Section **5.3.3**) confirm robust and consistent performance across both classes. This plot supports the utility of the model in rapidly prioritizing candidates for further experimental validation and helps streamline downstream drug discovery workflows by filtering out less promising compounds.

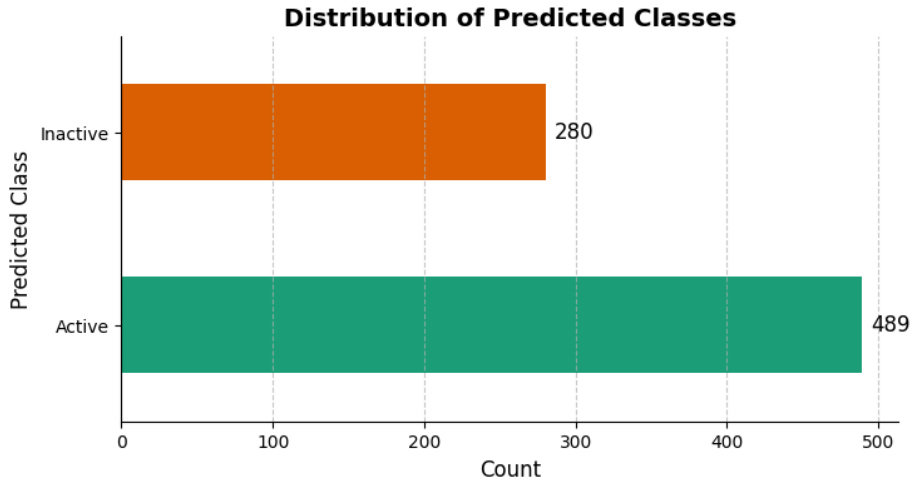


Fig 5.10. Predicted class distribution on unknown phytochemical dataset

5.3.6 Molecular docking based virtual screening

A machine learning-based screening of 769 phytochemicals from IMPs revealed 489 to be active against Mtb. These compounds were subsequently screened using the molecular docking approach. The top eight molecules were then chosen based on their binding scores, which are presented in **Table 5.5**.

The docking analysis revealed that somniferine had the highest binding affinity, with a score of -11.8 kcal/mol. Other notable compounds included sitoindoside IX (-11.5 kcal/mol), withanicandrine (-11.3 kcal/mol), alpha-amyrine (-11.1 kcal/mol), glabrolide (-11.0 kcal/mol), liquoric acid (-11 kcal/mol), withasomnilide (-10.9), and Withanolide D (-10.9). The majority of these high-scoring ligands are phytochemicals found in Withania somnifera, a herb known for its antimicrobial properties. These compounds could be useful adjuvants in conjunction with conventional anti-tuberculosis drugs. Furthermore,

Alstonia scholaris provides alpha-amyrine, while Glycyrrhiza glabra supplies glabrolide and liquoric acid. These medicinal plants are also said to have significant antimycobacterial properties. **Fig. 5.11** depicts the structures of some of the selected ligands.

Given the current challenges in tuberculosis drug discovery, these molecules require additional experimental validation to identify potential anti-TB drug candidates. Furthermore, the medicinal plants from which these compounds are derived may be useful sources of antimycobacterial agents, with potential for use as adjuvant therapies in tuberculosis treatment.

Table 5.5. Binding Affinity of Top 8 Compounds and Their Source Plants

Compound	**Docking score (kcal/mol)**	**Name of plant**
Somniferine	-11.8	Withania somnifera
Sitoindoside IX	-11.5	Withania somnifera
Withanicandrine	-11.3	Withania somnifera
Alpha amyrine	-11.3	Alstonia scholaris
Glabrolide	-11.1	Glycyrrhiza glabra
Llicoric acid	-11	Glycyrrhiza glabra
Withasomnilide	-10.9	Withania somnifera
Withanolide D	-10.9	Withania somnifera

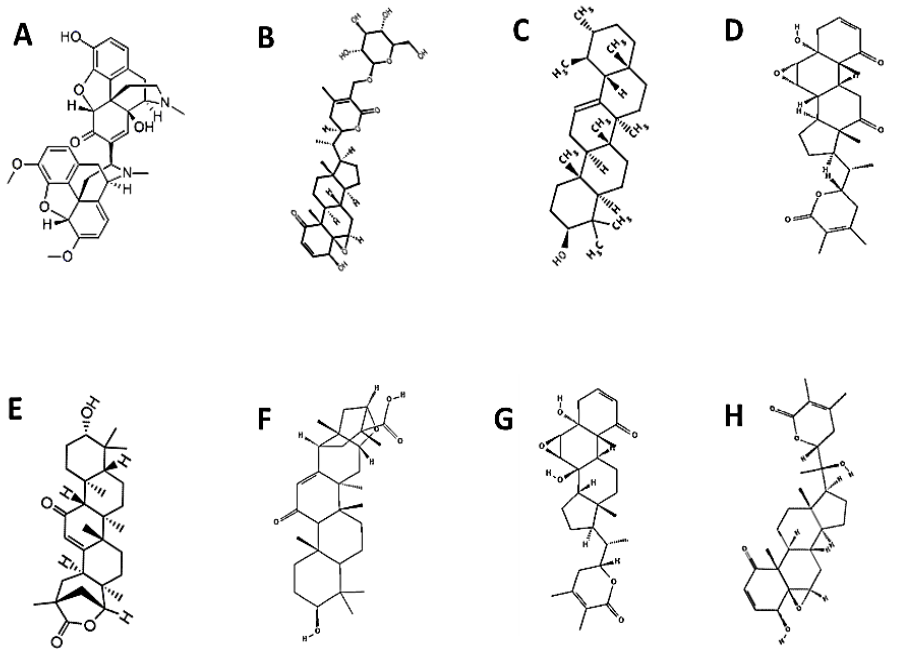


Fig. 5.11. The structures of selected molecules with the highest docking score (A) Somniferine (B) Sitoindoside IX (C) Withanicandrine (D) Alpha amyrine (E) Glabrolide (F) Llicoric acid (G) Withasomnilide (H) Withanolide D

5.4. Conclusion

A QSAR modeling framework based on machine learning was created for this study in order to forecast the anti-tubercular properties of Indian medicinal plant compounds. PubChem fingerprint descriptors were utilized to build QSAR models, and their performances were compared. It was discovered that numerous fingerprint descriptors performed well in the developed models, showing that they may capture the feature space of Mtb inhibitor. By quickly evaluating several classifiers, the LazyPredict tool made it possible to determine

which model performed best based on classification metrics like accuracy, F1-score, and specificity. The LGBM Classifier QSAR model performed well over other ML classifiers. The finished model showed strong predictive capabilities, underscoring machine learning's potential to speed up drug discovery based on natural products. The model's suitability for ranking promising candidates was further illustrated by the screening of bioactive phytochemicals using molecular docking analysis. We can infer from the docking score that Withania somnifera is one of the key medicinal plants inhibiting Mtb activity. According to our findings, our LGBM-QSAR classification model should be considered in the further development of Mtb inhibitors prediction, and traditional Indian medicinal plants may be used to treat tuberculosis. Also, the phytochemicals screened may be used to develop novel TB drugs. All things considered, this strategy offers a repeatable and scalable pipeline for incorporating computational techniques into early-stage anti-TB medication screening.

References

[1] M. Orgeur, C. Sous, J. Madacki, R. Brosch, Evolution and emergence of Mycobacterium tuberculosis, FEMS Microbiol. Rev. 48 (2024) fuae006. https://doi.org/10.1093/femsre/fuae006.

[2] K. Dheda, F. Mirzayev, D.M. Cirillo, Z. Udwadia, K.E. Dooley, K.-C. Chang, S.V. Omar, A. Reuter, T. Perumal, C.R. Horsburgh, M. Murray, C. Lange, Multidrug-resistant tuberculosis, Nat. Rev. Dis. Prim. 10 (2024) 22. https://doi.org/10.1038/s41572-024-00504-2.

[3] H.N. Jnawali, S. Ryoo, First– and Second–Line Drugs and Drug Resistance, in: B.H. Mahboub, M.G. Vats (Eds.), IntechOpen, Rijeka, 2013: p. Ch. 10. https://doi.org/10.5772/54960.

[4] M. Toraskar, P. Kamble, Enoyl Acyl Carrier Protein Reductase Inhibitors: An Emerging Target, Int. J. ChemTech Res. 11 (2018) 123–133. https://doi.org/10.20902/IJCTR.2018.110715.

[5] M. Prasad, R. Bhole, P. Khedekar, R. Chikhale, Mycobacterium enoyl acyl carrier protein reductase (InhA): A key target for antitubercular drug discovery, Bioorg. Chem. 115 (2021) 105242. https://doi.org/10.1016/j.bioorg.2021.105242.

[6] I. Ahmad, A.Z. Beg, Antimicrobial and phytochemical studies on 45 Indian medicinal plants against multi-drug resistant human pathogens, J. Ethnopharmacol. 74 (2001) 113–123. https://doi.org/https://doi.org/10.1016/S0378-8741(00)00335-4.

[7] R. Gautam, A. Saklani, S. Jachak, Indian medicinal plants as a source of antimycobacterial agents, J. Ethnopharmacol. 110 (2007) 200–234. https://doi.org/10.1016/j.jep.2006.12.031.

[8] S.K. Niazi, Z. Mariam, Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review, Int. J. Mol. Sci. 24 (2023). https://doi.org/10.3390/ijms241411488.

[9] D.A. Winkler, The impact of machine learning on future tuberculosis drug discovery, Expert Opin. Drug Discov. 17 (2022) 925–927. https://doi.org/10.1080/17460441.2022.2108785.

[10] S.C. Gad, QSAR, in: P.B.T.-E. of T. (Third E. Wexler (Ed.), Academic Press, Oxford, 2014: pp. 1–9. https://doi.org/https://doi.org/10.1016/B978-0-12-386454-3.00971-4.

[11] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107. https://doi.org/10.1093/nar/gkr777.

[12] S.P. Leelananda, S. Lindert, Computational methods in drug discovery, Beilstein J. Org. Chem. 12 (2016) 2694–2718. https://doi.org/10.3762/bjoc.12.267.

[13] A.N. Lima, E.A. Philot, G.H.G. Trossini, L.P.B. Scott, V.G. Maltarollo, K.M. Honório, Use of machine learning approaches for novel drug discovery, Expert Opin. Drug Discov. 11 (2016) 225–239. https://doi.org/10.1517/17460441.2016.1146250.

[14] M. Kumari, N. Tiwari, S. Chandra, N. Subbarao, Comparative analysis of machine learning based QSAR models and molecular docking studies to screen potential anti-tubercular inhibitors against InhA of mycobacterium tuberculosis, Int. J. Comput. Biol. Drug Des. 11 (2018) 209. https://doi.org/10.1504/IJCBDD.2018.094630.

[15] M.N. Drwal, R. Griffith, Combination of ligand- and structure-based methods in virtual screening, Drug Discov. Today Technol. 10 (2013) e395–e401. https://doi.org/https://doi.org/10.1016/j.ddtec.2013.02.002.

[16] S. Ahmadi, S. Ketabi, M. Jebeli Javan, Molecular Descriptors in QSPR/QSAR Modeling BT - QSPR/QSAR Analysis Using SMILES and Quasi-SMILES, in: A.P. Toropova, A.A. Toropov (Eds.), Springer International Publishing, Cham, 2023: pp. 25–56. https://doi.org/10.1007/978-3-031-28401-4_2.

[17] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, Drug Discov. Today. 21 (2016) 1291–1302. https://doi.org/https://doi.org/10.1016/j.drudis.2016.06.013.

[18] D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni, S.A. Sieber, Effectiveness of molecular fingerprints for exploring the chemical space of natural products, J. Cheminform. 16 (2024) 35. https://doi.org/10.1186/s13321-024-00830-3.

[19] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474. https://doi.org/https://doi.org/10.1002/jcc.21707.

[20] D. Rengasamy, J.M. Mase, A. Kumar, B. Rothwell, M.T. Torres, M.R. Alexander, D.A. Winkler, G.P. Figueredo, Feature importance in machine learning models: A fuzzy information fusion approach, Neurocomputing. 511 (2022) 163–174. https://doi.org/https://doi.org/10.1016/j.neucom.2022.09.053.

[21] O. Pillai, A.B. Dhanikula, R. Panchagnula, Drug delivery: an odyssey of 100 years, Curr. Opin. Chem. Biol. 5 (2001) 439–446. https://doi.org/https://doi.org/10.1016/S1367-5931(00)00226-X.

[22] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem BT - Advances in Knowledge Discovery and Data Mining, in: T. Theeramunkong, B. Kijsirikul, N. Cercone, T.-B. Ho (Eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, 2009: pp. 475–482.

[23] R. Singhal, R. Rana, Chi-square test and its application in hypothesis testing, J. Pract. Cardiovasc. Sci. 1 (2015). https://doi.org/10.4103/2395-5414.157577.

[24] T. Puślecki, K. Walkowiak, Hyperparameters Optimization Using GridSearchCV Method for TinyML Models BT - Progress on Pattern Classification, Image Processing and Communications, in: R. Burduk, M. Choraś, R. Kozik, P. Ksieniewicz, T. Marciniak, P. Trajdos (Eds.), Springer Nature Switzerland, Cham, 2023: pp. 63–69.

[25] L. Lydia, S. Althubiti, C. Anupama, V. Kollati, Prediction of

Sepsis Disease Using Random Search to Optimize Hyperparameter Tuning Based on Lazy Predict Model, in: 2023: pp. 351–367. https://doi.org/10.1007/978-981-99-6706-3_31.

[26] S. Skoczen, R. Bussmann, ebDB - an International Ethnobotany Database, Lyonia. 11 (2006) 71–81.

[27] Q. Li, T. Cheng, Y. Wang, S.H. Bryant, PubChem as a public resource for drug discovery, Drug Discov. Today. 15 (2010) 1052–1057. https://doi.org/https://doi.org/10.1016/j.drudis.2010.10.003.

[28] K. Mohanraj, B.S. Karthikeyan, R.P. Vivek-Ananth, R.P.B. Chand, S.R. Aparna, P. Mangalapandi, A. Samal, IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics, Sci. Rep. 8 (2018) 4329. https://doi.org/10.1038/s41598-018-22631-z.

[29] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242. https://doi.org/10.1093/nar/28.1.235.

[30] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminform. 3 (2011) 33. https://doi.org/10.1186/1758-2946-3-33.

[31] A.I. Adler, A. Painsky, Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection, Entropy. 24 (2022). https://doi.org/10.3390/e24050687.

Chapter 6

Integrating Virtual screening, MD Simulation, MM-PBSA, ADMET and DFT calculations for Identifying InhA inhibitors in Mycobacterium Tuberculosis

6.1. Introduction

Tuberculosis (TB), caused by the pathogen *Mycobacterium tuberculosis* (Mtb), remains a serious global health concern due to its high morbidity and mortality rates [1]. Despite extensive research over the years, tuberculosis remains a formidable threat, compounded by the emergence of drug-resistant strains [2]. Addressing this growing problem necessitates the development of new therapeutic approaches with distinct mechanisms of action [3]. The complex biology of Mtb, as well as the difficulty of developing drugs that are both effective and have minimal side effects, make this a daunting task. As a result, ongoing research and investment in innovative tuberculosis treatments are critical for addressing this global health crisis [4].

Mtb has a distinct cell wall composition and slow growth, making it difficult to develop effective treatments [5]. Existing TB drugs frequently target specific elements of the Mtb cell wall, such as mycolic acids or the peptidoglycan layer, in order to weaken the cell wall's structural integrity and inhibit bacterial growth [6]. Other drugs target critical metabolic enzymes and pathways, such as folate biosynthesis, the electron transport chain, and ATP synthase, disrupting energy production and preventing cell division [7]. Enoyl acyl carrier protein reductase (InhA), a key enzyme in the fatty acid biosynthesis pathway responsible for mycolic acid production, has emerged as a promising target for all Mycobacterium species. InhA's inhibition by various molecular scaffolds makes it an important target for developing new tuberculosis therapies [8]. In response to the growing prevalence of multidrug-resistant (MDR) and extensively drug-resistant (XDR)

tuberculosis (TB), InhA has been identified as a critical protein target for the development of potent *Mycobacterium tuberculosis* (Mtb). Recognising the importance of the situation, our research focuses on InhA as the chosen target, using a variety of computational approaches to tackle this pressing challenge.

The use of computational techniques has helped to accelerate the search for new lead compounds that target InhA. To find potential drug candidates that can bind to a biological target like InhA, virtual screening (VS) is a computational method used in drug development to evaluate large libraries of small molecules or compounds. The ligand-based (LBVS) and structure-based (SBVS) methods are the most common ways to carry out the VS protocol, and each has advantages and disadvantages of its own [9]. Investigating the chemical structure of potential bioactive drug candidates for a biological target is the foundation of the LBVS strategy, which is chosen for novel computational chemistry simulations, whereas SBVS is used to search for new bioactive compounds against a specific drug target in a chemical compound library during the early stages of a drug discovery campaign [10–13]. Over the years, significant progress has been made in the identification of InhA inhibitors. Huanxiang Liu and colleagues, for instance, used ensemble docking and biological assays to identify new and potent InhA inhibitors [14]. Similarly, Pornpan Pungpo et al. identified promising inhibitors using in silico screening and pharmacokinetic prediction, whereas Zeddine Ibrahimi et al. created a 3D-pharmacophore model and identified potential direct inhibitors using in silico methods [15,16]. To identify a lead compound targeting

InhA, this study used a structure-based virtual screening (SBVS) approach in conjunction with molecular docking, molecular dynamics simulations (MDs), and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations. The selected ligands were also evaluated for drug-likeness, physicochemical and pharmacokinetic properties, and electronic structure using density functional theory (DFT).

6.2. Materials and Methods

6.2.1. Virtual Screening and Molecular Docking

ZINC 15 is a publicly accessible repository of chemical compounds that can be purchased, with over 997 million molecules available in 3D configurations suitable for virtual screening [17–19]. For this study, we used a curated dataset of 2,310,097 "in-stock" lead-like molecules with neutral charges, formatted as pdbqt files and ready for docking studies with AutoDock Vina. These lead-like molecules were chosen using specific criteria: a molecular weight range of 250-350 g/mol, a predicted partition coefficient (xLogP) not greater than 3.5, and a maximum of 7 rotatable bonds.

The three-dimensional X-ray crystallographic structure of InhA in complex with isonicotinic-acetyl-nicotinamide-adenine dinucleotide (ZID) (PDB ID: 1ZID) at 2.70 Å resolution was retrieved from the Protein Data Bank (PDB) [20]. Prior to molecular docking, the protein was prepared using BIOVIA Discovery Studio. Polar hydrogen atoms were added to the protein structure to eliminate potential docking interference and water molecules and heteroatoms were removed. A

specific ligand conformation was chosen from multiple poses in the crystal structure, and the X, Y, and Z coordinates were recorded to determine the binding affinity. Ligands were assigned Gasteiger charges using MGL Tools (AutoDock Vina), which also made it easier to convert the protein file format from PDB to PDBQT (Protein Data Bank with Partial Charges and Atom Types), which is required for AutoDock Vina [21,22].

Molecular docking was performed using AutoDock Vina. The Vina scoring function, an empirical free-energy scoring function integrated within the software, was used to estimate binding affinities expressed in kcal/mol. The docking grid was centered on the active site of InhA based on the coordinates of the co-crystallized ligand in the 1ZID structure. The grid center was set at x = −4.346558, y = 34.669423, and z = 13.433750. The grid box dimensions were defined as 20 × 20 × 20 Å to adequately cover the substrate-binding pocket and surrounding active site residues. The exhaustiveness parameter was set to 8 to ensure sufficient conformational sampling while maintaining computational efficiency. The best binding pose for each ligand was selected based on the lowest predicted binding energy.

The compounds were evaluated using their docking scores and interaction profiles, and the top five candidates were chosen. The binding poses and interactions of these compounds were analysed and visualised using the Discovery Studio Client 2021.

6.2.2. Molecular dynamics simulation

An automated protein-ligand protocol developed at SiBioLead that makes use of the GROMACS simulation package was used to carry out the molecular dynamics (MD) simulations [23]. The simulation setup used the OPLS/AA force field, with ligand parameters generated using AMBERTOOLS and the ACPYPE package [24]. The protein-ligand complex was immersed in a triclinic box containing SPC water molecules and neutralised with NaCl counter ions. To simulate physiological conditions, an additional 0.15 M concentration of NaCl was introduced. The system was equilibrated using NVT and NPT ensembles over 300 ps. Trajectory snapshots were taken every 20 ps, resulting in 5000 frames. These trajectories were analysed using GROMACS' integrated tools, and the results were displayed using xmgrace.

6.2.3. MM-PBSA calculation

The binding free energy ($\Delta G_{binding}$) of protein-ligand complexes was calculated. The molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method was used to estimate binding free energies, with an automated plugin available on the SiBioLead server [25,26]. **Equation 6.1** was used to compute the binding free energy for each frame.

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{receptor} + \Delta G_{ligand}) \quad \textbf{(6.1)}$$

6.2.4. ADMET profile

A computational assessment was performed to assess the selected compounds' physicochemical properties and pharmacokinetic

profiles, with a focus on absorption, distribution, metabolism, excretion, and toxicity (ADMET). The web-based tools SwissADME and pkCSM were used to conduct the evaluations [27,28]. SwissADME provides an accessible platform with a variety of predictive models for analysing pharmacokinetics, drug-likeness, physicochemical properties, and compatibility with medicinal chemistry principles. On the other hand, pkCSM is a free web server that uses graph-based molecular signatures to predict ADMET properties. These signatures encode atomic distance patterns, which represent molecular structures and serve as training data for predictive algorithms.

6.2.5. DFT studies

The quantum chemical calculations were performed, and the geometries of all the structures were fully optimized with M06-2X/6-311++G (d, p) level of theory using the Gaussian 16W suite [29]. Density Functional Theory (DFT) has become a widely adopted computational tool due to its convenience and accuracy in predicting physical and chemical properties. This accuracy stems from its ability to determine electron density and energy properties effectively. According to the literature, within the Minnesota family of functionals, the M06-2X functional stands out for its top performance in quantum chemical calculations [30–32]. Its reliability makes it an excellent choice for a wide range of studies. The output verification files were analysed using GaussView 6.0 [33]. The electrostatic surface potential (ESP), the HOMO–LUMO energy, and thermochemical parameters (thermal energies, thermal enthalpies, thermal free energies, hardness,

softness, ionization energy, and electron affinity) of the ligands were obtained from the optimized geometry.

6.3. Result and Discussion

6.3.1. Virtual Screening and Molecular Docking

Molecular docking was used to determine the molecular interactions and binding affinities of a protein-ligand complex. The most promising ligands were chosen and prioritised based on their binding orientation and affinity to the protein's active site. A comprehensive virtual screening of the extensive ZINC database, which contains millions of known compounds, resulted in the identification of 158 candidates with good binding affinity. 25 compounds had binding energies ranging from -11.5 to -12.1 kcal/mol. Compounds with binding energies greater than -11.7 kcal/mol were chosen for further analysis. **Table 6.1,** summarises the docking results for the top-ranked ligands within InhA's active site. The binding energies of ZINC82139221, ZINC4090770, ZINC401340, ZINC49940, and ZINC35877800 were −12.1, −11.9, −11.9, −11.8, and −11.7 kcal/mol. Among these, ZINC000082139221 showed the highest binding affinity, outperforming all other compounds**. Tables 6.1** and **6.2** detail the interactions and critical residues within the InhA active site, which were used to identify the lead compound.

Table 6.1. The Docking score of the top ranked five compounds in the active site of InhA.

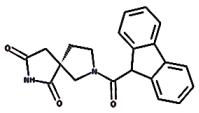
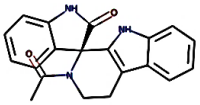
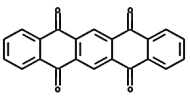
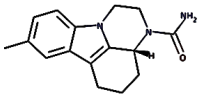
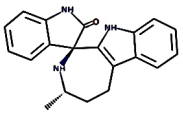
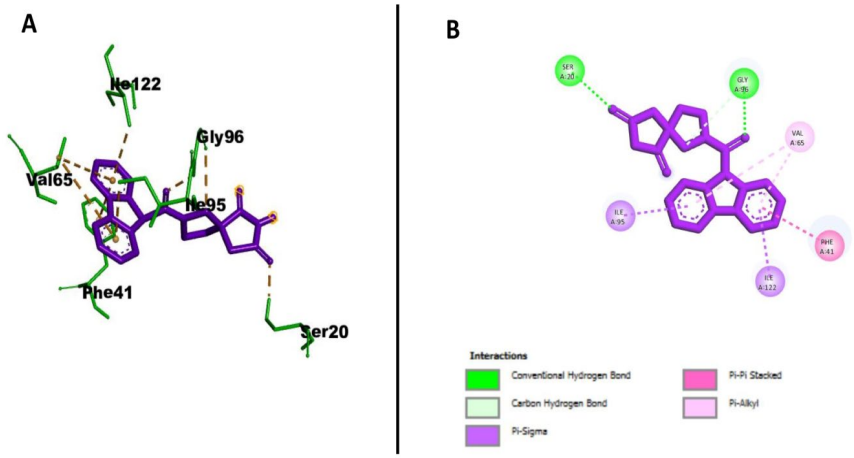
Compounds	**Structure**	**Residues involved in hydrogen bonding**	**Distance (Å)**	**Docking score (kcal/mol)**
ZINC82139221		Gly96, Ser20	2.20, 2.08	-12.1
ZINC4090770		Ile194, Ala191, Gly192	1.95, 2.86, 2.33	-11.9
ZINC401340		Val65, Gly14	2.03, 3.49	-11.9
ZINC49940		Thr39, Ile15	2.30, 2.10	-11.8
ZINC35877800		Ile194, Ala191, Gly192	1.87, 3.09, 2.47	-11.7

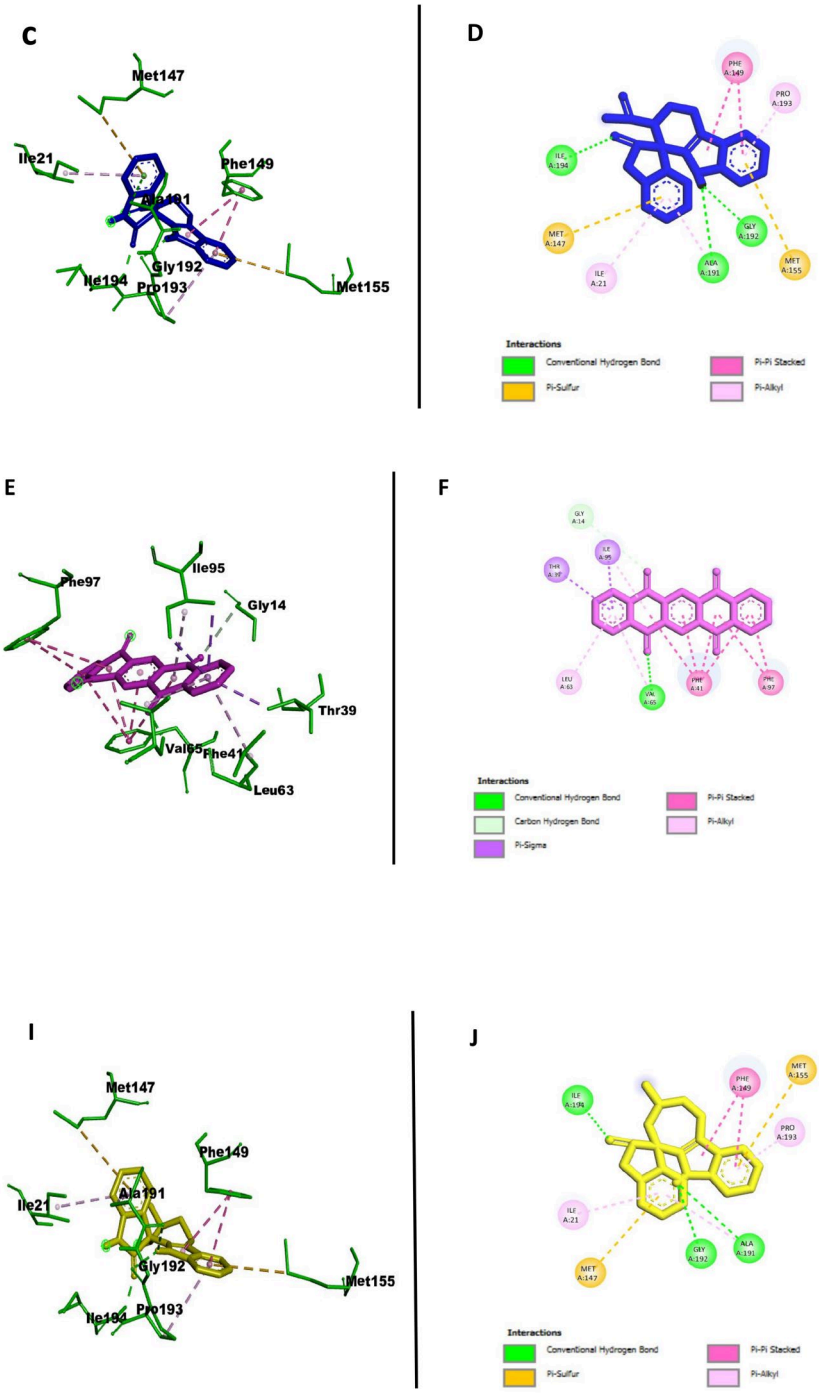
Table 6.2. The protein-ligand interactions for top-ranked compounds in the active site of InhA.

Compounds	**Residues involved in protein-ligand interaction**	**Type of interaction**
ZINC82139221	Gly96, Ile95	Pi-Sigma
	Ile22, Phe41	Pi-Pi stacked
	Val65	Pi-Alkyl
ZINC4090770	Met147, Met155	Pi-Sulfur
	Phe149	Pi-Pi stacked
	Pro193, Ile21, Ala191	Pi-Alkyl
ZINC401340	Thr39, Ile95	Pi-Sigma
	Phe41, Phe97	Pi-Pi stacked
	Ile95, Leu63,Val65	Pi-Alkyl
ZINC49940	Phe97	Pi-Sigma
	Phe41, Phe97	Pi-Pi stacked
	Val65, Ile95, Ile122	Alkyl
	Phe41, Ile122	Pi-Alkyl
ZINC35877800	Met147, Met155	Pi-Sulfur
	Phe149	Pi-Pi stacked
	Ile21, Ala191,Pro193	Pi-Alkyl

The compound ZINC82139221 formed a hydrogen bond with the residues of Gly96 and Ser20, π–π stacking interactions with Ile22 and Phe41, showed a π-sigma interaction with Gly96, Ile95 and π -alkyl interaction with Val65 residues. Also, ZINC4090770 formed a hydrogen bond with Ile194, Ala191 and Gly192, participated in the π–π stacking interactions with Phe149, π-alkyl interactions with Pro193,

Ile21, Ala191 and π-sulfur interactions with Met147, Met155 residues. The analysis of docking results for ZINC401340 displayed that this compound made the hydrogen bond interaction with the residues of Val65 and Gly14, π–π stacking with Phe41 and Phe97, π-sigma interactions with Thr39, Ile95 and Ile95, and Pi-Alkyl interactions with Leu63, Val65 residues. ZINC49940 showed two hydrogen bonds with Thr39, Ile15, two π-alkyl interactions with Phe41, Ile122, two π–π stacking interactions with Phe41, Phe97, one π-sigma interaction with Phe97 and three Alkyl interactions with Val65, Ile95 and Ile122 residues. ZINC35877800 showed a π-alkyl interaction with the Ile21, Ala191 and Pro193 residues of the InhA, respectively. In addition, this compound formed three hydrogen bonds with Ile194, Ala191, Gly192, π–π stacking interaction with Phe149 and two Pi-Sulfur interactions with Met147, Met155. **Fig. 6.1** shows the molecular interaction between the five top ranked hits and receptor.




C
Met147
Ile21
Phe149
Ala191
Gly192
Ile194
Pro193
Met155
D
Interactions
Conventional Hydrogen Bond
Pi-Sulfur
Pi-Pi Stacked
Pi-Alkyl
E
Phe97
Ile95
Gly14
Thr39
Val65
Phe41
Leu63
F
Interactions
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Pi-Sigma
Pi-Pi Stacked
Pi-Alkyl
I
Met147
Phe149
Ile21
Ala191
Gly192
Met155
Ile194
Pro193
J
Interactions
Conventional Hydrogen Bond
Pi-Sulfur
Pi-Pi Stacked
Pi-Alkyl

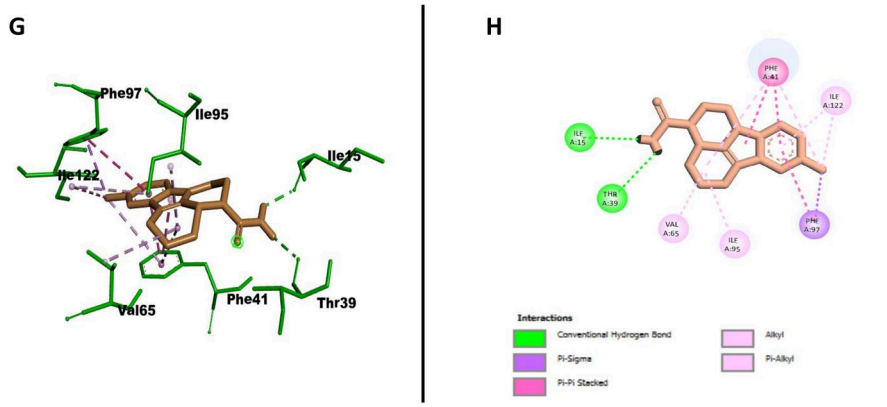


Fig. 6.1. 3D and 2D Interactions and orientations of ZINC82139221 (A, B), ZINC4090770 (C, D), ZINC401340 (E, F), ZINC49940, (G, H) and ZINC35877800 (I, J) in the binding pocket of InhA

6.3.2. Molecular dynamics simulation results

The dynamic behaviour of the studied systems was explored through molecular dynamics simulations, giving further confirmation for the molecular docking findings. The top-ranked five ligands from the docking experiments were used for the MD simulations. To better understand their behaviour, molecular dynamic simulations were used to calculate the Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). RMSD quantifies the average positional deviation between corresponding atoms in two aligned molecular structures, providing information about the system's overall stability. RMSF, on the other hand, measures the degree of fluctuation of specific atoms or groups of atoms relative to a reference structure, taking into account all relevant atoms. This metric aids in determining structural flexibility, with higher RMSF values indicating flexible regions such as loops or loosely bonded areas, whereas lower values indicate more rigid and stable regions, which frequently correspond to

secondary structural elements such as helices and sheets. Solvent Accessible Surface Area (SASA) refers to the portion of the protein-ligand complex that interacts directly with solvent molecules. Variations in SASA values reflect structural changes; increases indicate expansion, while decreases indicate greater compactness and stability. Hydrogen bonds, which play an important role in ligand binding, are essential for determining interaction specificity and influencing drug absorption and metabolism. Another stability indicator, the radius of gyration (Rg), was calculated. Rg measures the compactness of the protein-ligand complex, with lower values usually indicating a more stable and compact structure. These parameters, when combined, provide comprehensive insights into the dynamics and stability of the complexes under study.

The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg for the InhA in complexation with ZINC82139221 are shown in **Fig. 6.2.** In the receptor- ZINC82139221 complex, the RMSD value (**Fig. 6.2A**) remains below 0.1 nm (1Å) ranging from 0.01 to 0.05 nm. This low RMSD indicates that the ligand remains relatively stable within the binding site, with minimal fluctuations. The RMSF values (**Fig. 6.2B**), observed in the range of 0.05–0.3 nm, suggest that the protein maintains a relatively stable structure with localized flexibility. A maximum of four hydrogen is observed between the receptor and ligand (**Fig. 6.2C**). The low Rg values (**Fig.6.2D**) fluctuating between 1.84 and 1.87 nm, indicate the stability of the complex, which can be correlated to the compactness of the structure. The SASA plot (**Fig.**

6.2E) shows minimal fluctuations during ligand binding, with lower values suggesting greater stability and compactness.

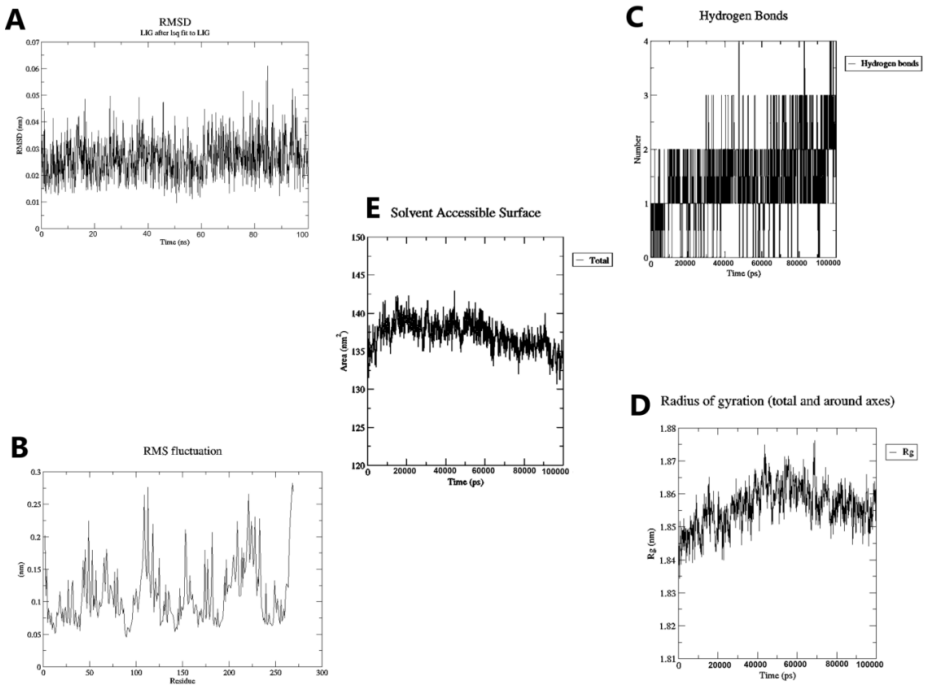


Fig. 6.2. The RMSD, RMSF, SASA, number of hydrogen bonds, and Rg plots of the InhA - ZINC82139221 complex.

The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg for the InhA in complexation with ZINC4090770 are given in **Fig. 6.3.** In the receptor- ZINC4090770 complex, the RMSD values (**Fig. 6.3A**) remain below the 0.1 nm (1Å) ranging from 0.01 to 0.08 nm with considerable fluctuations. This low RMSD Indicates that the ligand remains relatively stable within the binding site. The RMSF values (**Fig. 6.3B**) observed in the range of 0.05–0.3 nm indicate that the protein maintains a relatively stable structure with localized flexibility. A maximum of four hydrogen bonds are observed between the receptor

and ligand (**Fig. 6.3C**). The Rg values (**Fig. 6.3D**), ranging from 1.82 and 1.85 nm, indicate the stability of the complex, which can be correlated to the compactness of the structure, despite notable fluctuations. The SASA plot (**Fig. 6.3E**) shows multiple fluctuations during ligand binding with a slight increase though the values still suggest stability and compactness.

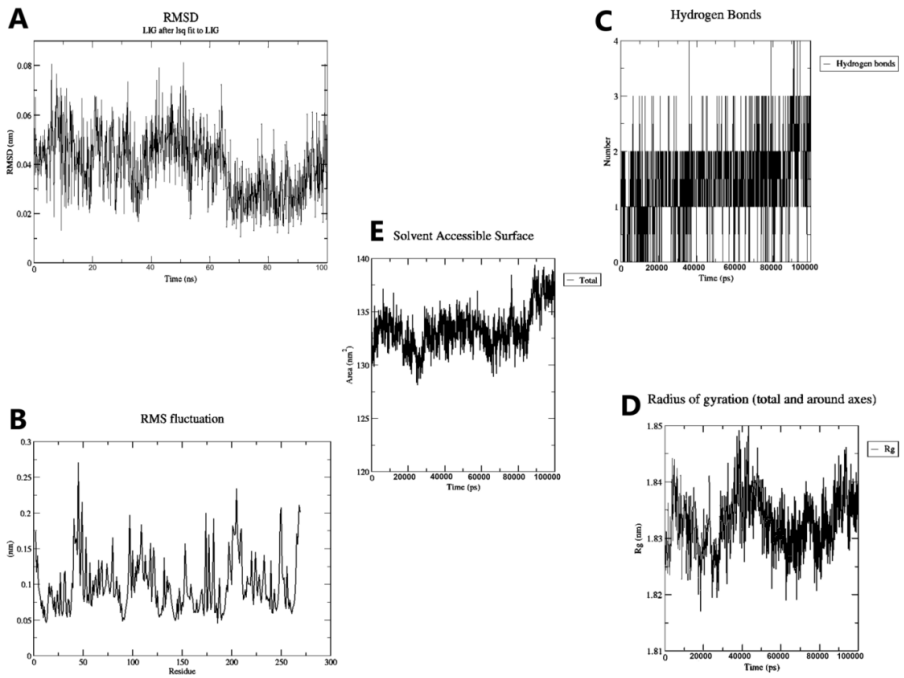


Fig. 6.3. The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg plots of the InhA-ZINC4090770 complex.

Fig. 6.4., shows the RMSD, RMSF, SASA, number of hydrogen bonds, and Rg for InhA complexed with ZINC401340. The RMSD values in this complex (**Fig. 6. 4A**) range from 0.01 to 0.05 nm, which is consistently less than 0.1 nm (1Å). Low RMSD values indicate that the ligand is relatively stable within the binding site, with

minimal fluctuations. The RMSF values (**Fig. 6.4B**) ranged from 0.05 to 0.4 nm, indicating that the protein maintains a stable structure with localised flexibility. A maximum of two hydrogen bonds are detected between the receptor and ligand (**Fig. 6.4C**), though this number is lower than that observed for other molecules. The Rg values (**Fig. 6.4D**), which range between 1.82 and 1.85 nm, reflect the complex's stability and compactness, despite noticeable variations. The SASA plot (**Fig. 6.4E**) shows some fluctuations during ligand binding, but the overall decrease in values indicates greater structural stability and compactness.

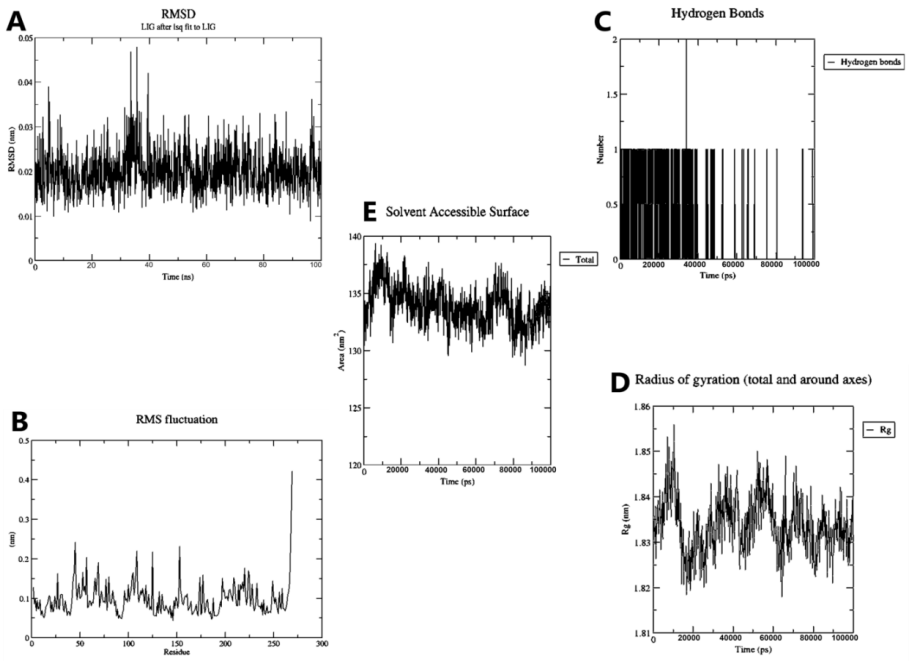


Fig. 6.4. The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg plots of the InhA - ZINC401340 complex.

Fig. 6.5 illustrates the RMSD, RMSF, SASA, number of hydrogen bonds, and Rg of InhA in complex with ZINC49940. The

RMSD values in this complex (**Fig. 6.5A**) are consistently less than 0.1 nm (1Å), ranging from 0.01 to 0.08 nm, with occasional fluctuations. These low RMSD values indicate that the ligand remains relatively stable within the binding site. The RMSF values (**Fig. 6.5B**) range from 0.05 to 0.3 nm, indicating that the protein structure is stable with localised flexibility. The receptor and ligand form up to three hydrogen bonds (**Fig. 6.5C**), indicating a moderate level of interaction strength. The Rg values (**Fig. 6.5D**) range from 1.82 to 1.87 nm, indicating the complex's stability and compactness. Similarly, the SASA plot (**Fig. 6.5E**) shows minimal variations during ligand binding, indicating that the complex is structurally stable and compact.

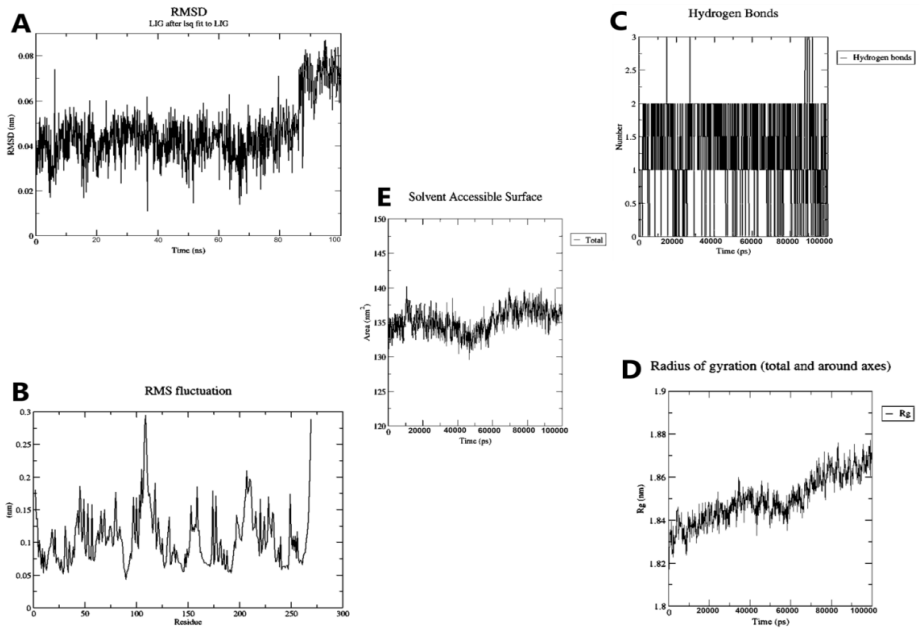


Fig. 6.5. The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg plots of the InhA - ZINC49940 complex.

Fig. 6.6., shows the RMSD, RMSF, SASA, number of hydrogen bonds, and Rg of InhA in complexation with ZINC35877800. In the receptor-ZINC35877800 complex, the RMSD values (**Fig. 6. 6A**) range from 0.01 to 0.04 nm. This low RMSD reflects the ligand's relative stability within the binding site, with only minor fluctuations. The RMSF values (**Fig. 6.6B**) range from 0.05 to 0.3 nm, indicating that the protein has a stable structure with localised flexibility. A maximum of three hydrogen bonds are observed between the receptor and the ligand (**Fig. 6.6C**). The Rg values (**Fig. 6.6D**) vary significantly between 1.81 and 1.85 nm, indicating the complex's stability, which is linked to structural compactness. Meanwhile, the SASA plot (**Fig. 6.6E**) shows noticeable fluctuations during ligand binding, but the overall values indicate that the complex is more stable and compact.

According to the MD simulation results, all of the complexes have favourable values and show overall stability. Among them, ZINC82139221 has values comparable or similar to the other complexes but stands out for its low fluctuations, indicating a relatively stable interaction within the binding site. We carried out MM-PBSA free energy calculations to investigate and validate the stability and binding affinity of these complexes. This analysis provided additional insights into binding free energies, as well as a quantitative assessment of the ligand-receptor complexes' interaction strength and stability.

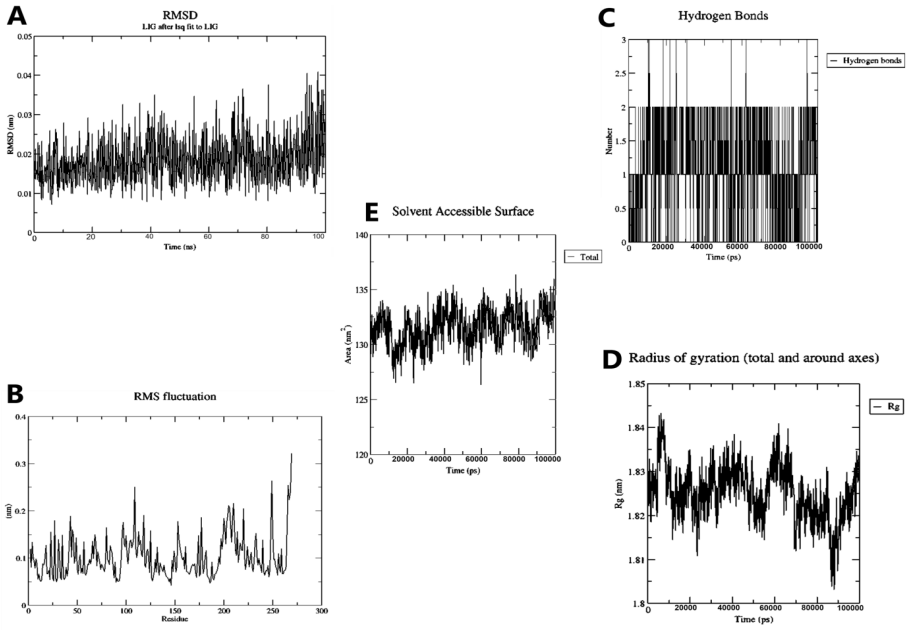


Fig. 6.6. The RMSD, RMSF, SASA, number of hydrogen bonding, and Rg plots of the InhA - ZINC35877800 complex.

6.3.3. Free energy calculations

The binding free energies of the target protein-ligand complexes with InhA were determined using MM-PBSA, which provides critical information about the thermodynamic stability of these interactions. **Table 6.3,** shows the calculated Gibbs free energy values for the compounds ZINC82139221, ZINC4090770, ZINC401340, ZINC401340, and ZINC5877800, which were -39.01, -38.02, -37.56, -36.27, and -32.66 kcal/mol, respectively. Negative Gibbs free energy values indicate ligand-protein interactions that are thermodynamically favourable and stable. These calculations took into account molecular mechanics, solvation effects, and entropy changes, confirming the complexes' stability and binding strength as observed in the simulations [34]. While the binding energy differences were minor, ZINC82139221

had a slightly higher binding affinity than the other compounds, consistent with docking and MD simulation results. In addition, as part of our comprehensive study, the pharmacokinetic properties of all compounds were examined to determine their potential suitability for future drug development.

Table 6.3. Binding free energy (Kcal/mol) for the selected compounds of InhA

compounds	**$\Delta G_{binding}$**
ZINC82139221	-39.01
ZINC4090770	-38.02
ZINC401340	-37.56
ZINC49940	-36.27
ZINC35877800	-32.66

6.3.4. Pharmacokinetic profile

The compounds' pharmacokinetic properties and toxicity profiles were determined using the pk-CSM and SwissADME webservers [35]. **Tables 6.4** and **6.5** gives a detailed summary of the ADME-Tox analysis carried out on these platforms. Key Lipinski Rule of Five (Ro5) parameters, such as molecular weight (MW), hydrogen bond acceptors (HBA), and hydrogen bond donors (HBD), were examined for their relevance to ligand interactions with the protein's active site. Unlike these parameters, the lipophilicity index (LogP) is experimentally determined and reflects the compound's biophysical properties. The topological polar surface area (TPSA) is an important

parameter that indicates a compound's ability to penetrate cell membranes [36]. Values above 140 Å indicate reduced penetration [37]. All of the studied compounds passed the Ro5 criteria as well as Ghose and Veber and were within acceptable limit.

Table 6.4. Summary of physicochemical properties of selected compounds against InhA determined using SwissADME web tool.

Compounds	MW	HBA	HBD	RB	TPSA	cLogp	Lipinski/ Ghose/ Veber violations
ZINC82139221	346.38	3	1	2	66.48	2.1	0
ZINC4090770	331.37	2	2	1	65.2	2.21	0
ZINC401340	338.41	4	0	0	68.28	3.28	0
ZINC49940	269.34	1	1	1	51.26	1.97	0
ZINC35877800	317.28	2	3	0	56.92	2.71	0

Drug interactions with the human body can be predicted using parameters like human intestinal absorption (HIA%), in vitro plasma protein binding, and blood-brain barrier (BBB) permeability, which reflect the drug's distribution profile. The HIA% values for the selected ligands ranged from 92.42 to 100%, indicating efficient intestinal absorption. Furthermore, these ligands demonstrated good cell membrane permeability, with in vitro Caco-2 cell permeability values falling within acceptable limits. The negative values for in vitro skin permeability imply that the compounds have minimal or no skin penetration. The majority of the compounds had BBB permeability

values that were below or near the crossing threshold, indicating a low risk of neurotoxicity [38]. Furthermore, all of the ligands had acceptable plasma protein binding (PPB) values, and all compounds performed well across all of the parameters tested.

Table 6.5. ADME Properties of Selected compounds Analysed Using the pk-CSM Web Server.

compounds	**Absorption**			**Distribution**	
	Caco-2 permeability	**HIA%**	**Skin Permeability**	**BBB**	**PPB**
ZINC82139221	0.94	97.82	-2.75	-0.28	0.19
ZINC4090770	0.72	95.55	-2.74	0.34	0.13
ZINC401340	1.04	100.00	-2.73	-0.30	-1.22
ZINC49940	1.29	92.42	-2.75	0.37	0.24
ZINC35877800	1.25	92.64	-2.73	0.27	0.74

Table 6.6, shows the toxicity profiles for the selected compounds. Two of the compounds tested were found to be Ames toxicity-free. Carcinogenicity tests revealed that four of the ligands were not carcinogenic, while one compound had mild toxicity. Given that blocking hERG K+ channels can cause QT interval prolongation and potentially fatal outcomes, this parameter was assessed. The results showed that one compound had moderate toxicity, while the other four were considered safe. Drug-induced liver injury, an important factor in

ADMET evaluations, varied from safe to moderately toxic across the compounds. Overall, ZINC82139221 had the best safety profile of any tested ligand and is regarded as a promising candidate for InhA enzyme inhibition. Nonetheless, other ligands were considered for additional DFT analysis due to their moderate risk levels.

Table 6.6. Toxicity profile for selected compound performed using pk-CSM webserver.

compounds	**AMES toxicity**	**Carcinogenesis**	**hERG Inhibitors**	**Liver Injury I (DILI)**
ZINC000082139221	safe	safe	safe	safe
ZINC000004090770	Medium toxic	safe	safe	Medium toxic
ZINC000000401340	safe	less toxic	safe	Medium toxic
ZINC000000049940	High toxic	Safe	safe	Medium toxic
ZINC000035877800	Medium toxic	safe	Medium toxic	low toxic

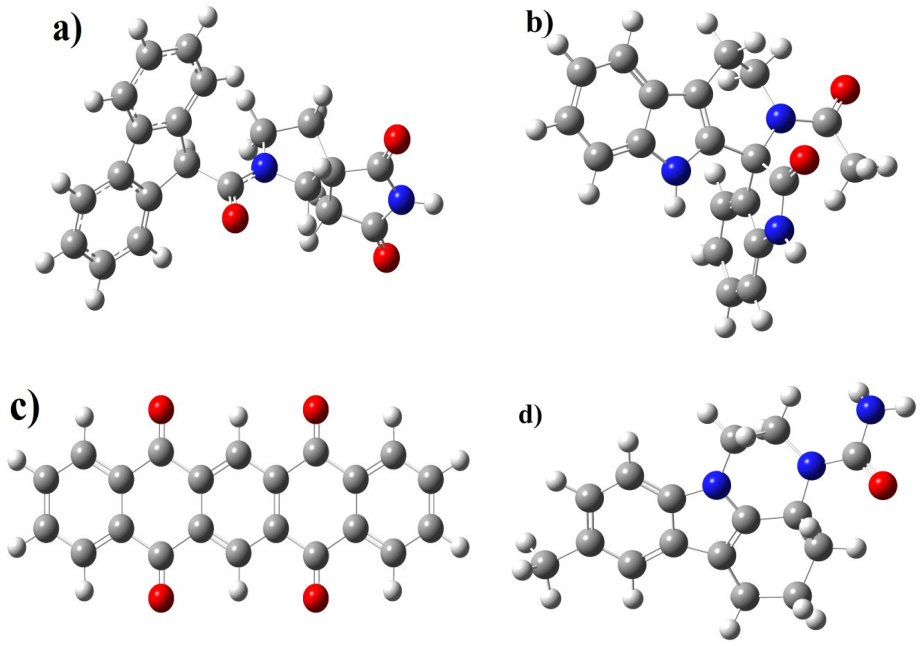
6.3.5. DFT analysis

The optimization of ligands was initially conducted using the M06-2X/6-311++G (d, p) level of theory in the gas phase, and the results obtained from this optimization are presented in **Table 6.7**. The vibrational frequencies were calculated at the same level of theory to

confirm that all stationary points correspond to minima, with zero imaginary frequencies. **Fig.6.7.,** represents the optimised image of the ligands.

Table 6.7. The geometric parameters of the ligands.

Compounds	Optimization energy (hartree)	Polarizability (α) (a.u.)	Dipole moment (Debye)
ZINC82139221	-1146.162624	255.110000	3.886856
ZINC4090770	-1086.968257	249.811667	6.254238
ZINC401340	-1145.151355	266.195333	0.001204
ZINC49940	-860.542533	208.717667	3.671247
ZINC35877800	-1012.933965	250.665333	3.812681



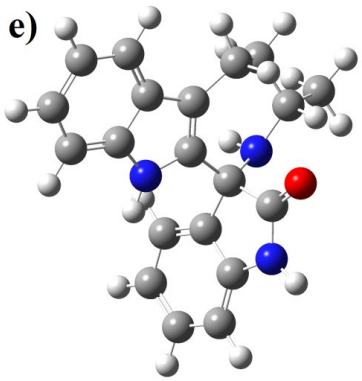


Fig. 6.7. Optimized structures of the Ligands a) ZINC82139221, b) ZINC4090770, c) ZINC401340, d) ZINC49940, and e) ZINC35877800 at M06-2X/6– 311++G(d,p) level of theory.

Locality in physical space is vital for understanding chemical reactivity and plays a key role in predicting molecular properties, particularly protein-ligand interactions. The concept of molecular orbitals (MO) in the quantum mechanical treatment of chemical systems has been highly successful over the past several decades. Since the Frontier molecular orbitals (FMOs), particularly the highest occupied molecular orbitals (HOMOs) and the lowest unoccupied molecular orbitals (LUMOs) pinpoint the locality of chemical bonds, also along with the HOMO-LUMO gap have been extensively utilized by chemists to analyse the reactivity and regioselectivity of diverse chemical systems. The results of HOMO, LUMO locations in the ligands and their energy gap are given in **Fig. 6.8**. The cyan and blue color distributions represent the positive and negative phases, respectively, in the MO wave function. It is noticed that all the five ligands possess almost similar energy gap, approximately of 6.3-6.9 eV, all the molecules are expected to exhibit comparable chemical reactivity and stability.

The energetic descriptors; ionisation potential (IP), electron affinity (EA), electronegativity (χ), hardness (η), chemical potential (μ), softness (S) and electrophilicity index (ω) values of all the five ligands were identified utilizing the HOMO and LUMO energies and tabulated in **Table 6.8**. The equations behind the calculation are given in **(Equations 6.2-6.8)**. These quantities represent the linear responses of the electron density to variations in the external potential and the number of electrons. Hardness reflects the overall stability of a system and fundamentally indicates its resistance to the deformation or polarization of the electron cloud when subjected to small perturbations during chemical reaction[39]. Softness is inversely proportional to chemical hardness. Low hardness value of ligands ZINC401340 and ZINC49940 suggests a heightened potential for chemical reactivity as evident from the smallest HOMO–LUMO energy gap value of 6.33 and 6.29 eV. Moreover, the comparatively high value of electron affinity (2.81 eV) and high value of IP (9.14 eV), reveals that it is difficult to remove an electron from the ligand ZINC401340 than to accept[40–42]. Moreover, the electronegativity of the ligand is also found to be relatively high, approximately 5.97 eV, allowing it to retain some charge. Since the chemical potential refers to the tendency of an electron to escape[43], and the electrophilic index indicates the strength of a species' electrophilicity [44], the relatively high values of both for the ligand ZINC401340 (chemical potential of -5.97 eV and electrophilic index of 5.64 eV), in comparison with other ligands, suggest a higher likelihood of electron transfer, where the ligand ZINC401340 will accept electrons during the process. Hence,

compared to ZINC401340, all other ligands are expected to act as electron donors during their interaction with the protein system.

Ionization potential (IP) = $-E_{HOMO}$ **(6.2)**

Electron Affinity (EA) = $-E_{LUMO}$ **(6.3)**

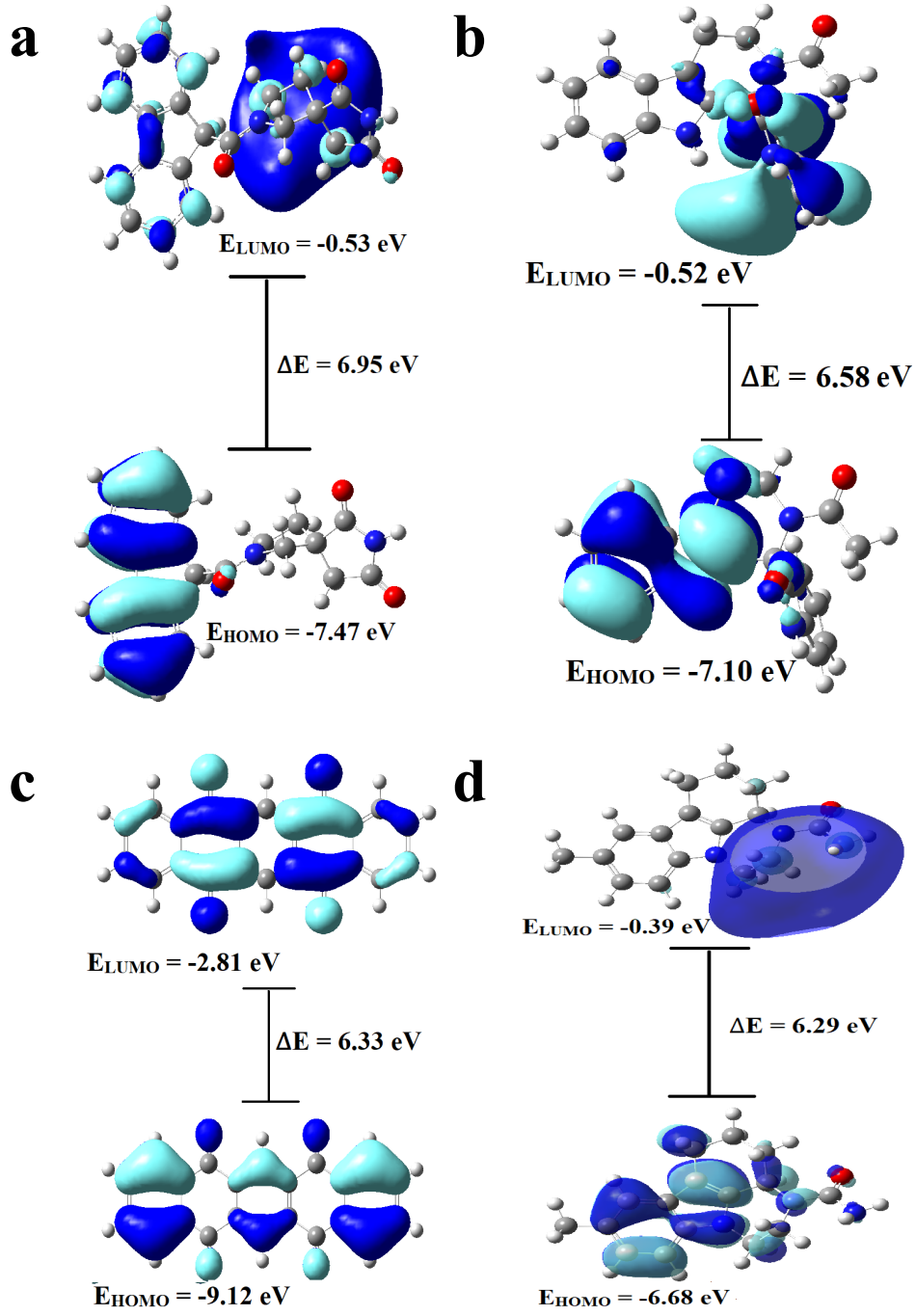
$$Electronegativity\left(\chi\right)=\frac{IP+EA}{2} \quad \textbf{(6.4)}$$

$$Chemical\ hardness\left(\eta\right)=\frac{IP-EA}{2} \quad \textbf{(6.5)}$$

$$Chemical\ potential\left(\mu\right)=-\chi \quad \textbf{(6.6)}$$

$$Chemical\ softness\left(S\right)=\frac{1}{2\eta} \quad \textbf{(6.7)}$$

$$Electrophilicity\ index\left(\omega\right)=\frac{\mu^2}{2\eta} \quad \textbf{(6.8)}$$


a
E_{LUMO} = -0.53 eV
ΔE = 6.95 eV
E_{HOMO} = -7.47 eV
b
E_{LUMO} = -0.52 eV
ΔE = 6.58 eV
E_{HOMO} = -7.10 eV
c
E_{LUMO} = -2.81 eV
ΔE = 6.33 eV
E_{HOMO} = -9.12 eV
d
E_{LUMO} = -0.39 eV
ΔE = 6.29 eV
E_{HOMO} = -6.68 eV

e

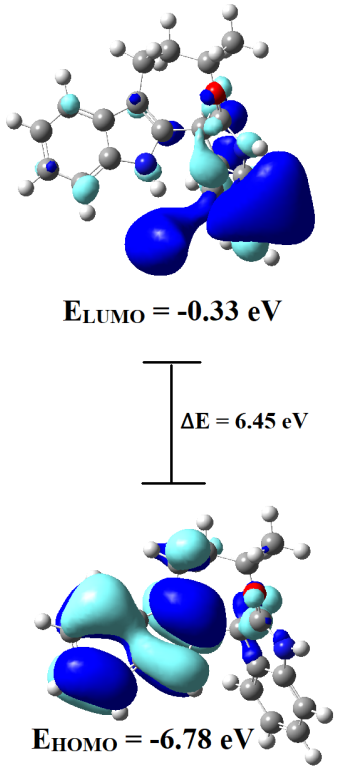
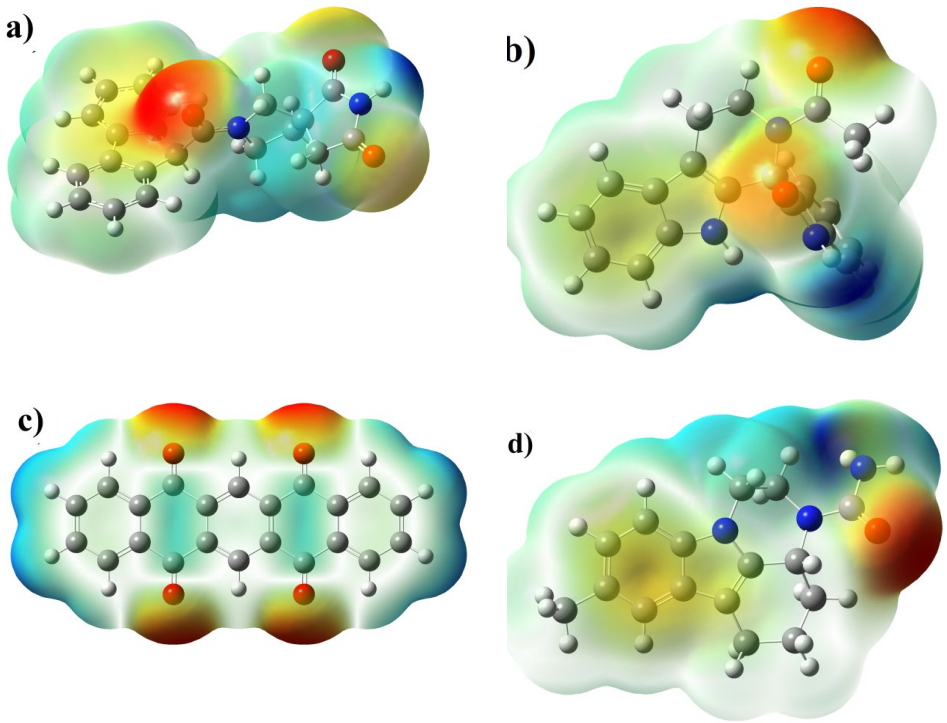


Fig. 6.8. DFT calculated HOMO(E_{HOMO}), LUMO (E_{LUMO}), and their energy gap (ΔE) for a) ZINC82139221, b) ZINC4090770, c) ZINC401340, d) ZINC49940, and e) ZINC35877800 at M06-2X/6–311++G (d,p) level of theory (Isovalue = 0.02).

Table 6.8. Energetic parameters of the ligands under investigation (Unit eV and for softness eV^{-1})

Compounds	**IP**	**EA**	**χ**	**η**	**μ**	**S**	**ω**
ZINC82139221	7.47	0.51	4.00	3.47	-4.00	0.14	2.30
ZINC4090770	7.10	0.52	3.81	3.29	-3.81	0.15	2.20
ZINC401340	9.14	2.81	5.97	3.16	-5.97	0.16	5.64
ZINC49940	6.68	0.39	3.53	3.14	-3.53	0.16	1.98
ZINC35877800	6.78	0.33	3.56	3.23	-3.56	0.16	1.96

The electrostatic potential offers a visual representation of the chemically active sites and atomic reactivity. A detailed analysis of the ESP of the title compound will provide valuable insights into the key interactions between the compound and DNA bases. The analysed electrostatic potential map of the ligands is shown in **Fig.6.9**. The red spheres on the electrostatic potential graph represent the negative charge sites. The map specified the balanced charge distribution in all the ligands which facilitates the binding of the compound to biological enzymes.



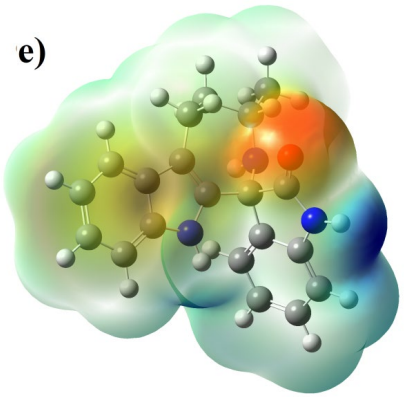


Fig. 6.9. The ESP of a) ZINC82139221, b) ZINC4090770, c) ZINC401340, d) ZINC49940, and e) ZINC35877800 at M06-2X/6–311++G (d,p) level of theory

6.4.Conclusion

The aim of this study was to identify potential InhA inhibitors, a critical target in combating drug-resistant tuberculosis, using an in silico structural-based virtual screening approach. A comprehensive workflow was utilised, including molecular docking, MD simulations, MM-PBSA binding energy calculations, drug-likeness evaluation, physicochemical and pharmacokinetic property assessments, and DFT analysis. Through virtual screening based molecular docking, five top-ranking compounds were selected from the ZINC 15 “in-stock” lead-like molecules library due to their high binding affinities. MD simulation analysis revealed that all selected compounds had similar RMSD, RMSF, hydrogen bond, Rg, and SASA values, indicating favourable interactions and overall stability of the ligand-protein complexes. The MM-PBSA calculations confirmed the thermodynamic stability of the complexes, as all tested compounds had negative Gibbs free energy values, indicating energetically favourable ligand-protein

interactions. Based on pharmacokinetic and ADMET profiling, ZINC82139221 was identified as the most promising compound in terms of safety and pharmacokinetics. Other candidates, such as ZINC4090770, ZINC401340, ZINC49940, and ZINC35877800, also showed strong inhibitory activity and favourable properties. DFT analysis showed the balanced charge distribution in all the ligands which facilitates the binding of the compound to biological enzymes. These findings indicate that ZINC82139221 is a highly promising lead compound, while the other compounds chosen are viable candidates for further experimental and clinical testing as potential InhA inhibitors.

References

[1] World Health Organization, Global Tuberculosis Report, Geneva, Switzerland, 2022. https://iris.who.int/bitstream/handle/10665/363752/9789240061729-eng.pdf?sequence=1.

[2] A. Zumla, M. Maeurer, for the H.-D.T.N. (HDT-N. Consortium, A. Zumla, J. Chakaya, M. Hoelscher, F. Ntoumi, R. Rustomjee, C. Vilaplana, D. Yeboah-Manu, V. Rasolof, P. Munderi, N. Singh, E. Aklillu, N. Padayatchi, E. Macete, N. Kapata, M. Mulenga, G. Kibiki, S. Mfinanga, T. Nyirenda, L. Mboko, A. Garcia-Basteiro, N. Rakotosamimanana, M. Bates, P. Mwaba, K. Reither, S. Gagneux, S. Edwards, E. Mfinanga, S. Abdulla, P.-J. Cardona, J.B.W. Russell, V. Gant, M. Noursadeghi, P. Elkington, M. Bonnet, C. Menendez, T.N. Dieye, B. Diarra, A. Maiga, A. Aseffa, S. Parida, C. Wejse, E. Petersen, P. Kaleebu, M. Oliver, G. Craig, T. Corrah, L. Tientcheu, M. Antonio, T.D. McHugh, A. Sheik, G. Ippolito, G. Ramjee, S.H.E. Kaufmann, G. Churchyard, A.J.C. Steyn, M.P. Grobusch, I. Sanne, N. Martinson, R. Mandansein, R.J. Wilkinson, R.S. Wallis, B. Mayosi, M. Schito, M. Maeurer, for the H.-D.T.N. (HDT-N. Consortium, Host-Directed Therapies for Tackling Multi-Drug Resistant Tuberculosis: Learning From the Pasteur-Bechamp Debates, Clin. Infect. Dis. 61 (2015) 1432–1438. https://doi.org/10.1093/cid/civ631.

[3] J. Khawbung, D. Nath, S. Chakraborty, Drug Resistant Tuberculosis: A Review, Comp. Immunol. Microbiol. Infect. Dis. 74 (2020) 101574. https://doi.org/10.1016/j.cimid.2020.101574.

[4] A. Zumla, P. Nahid, S. Cole, Advances in the development of new tuberculosis drugs and treatment regimens, Nat. Rev. Drug Discov. 12 (2013) 388–404. https://doi.org/10.1038/nrd4001.

[5] P.J. Brennan, H. Nikaido, THE ENVELOPE OF MYCOBACTERIA, Annu. Rev. Biochem. 64 (1995) 29–63. https://doi.org/10.1146/annurev.bi.64.070195.000333.

[6] C. Barry, R. Lee, K. Mdluli, A. Sampson, B. Schroeder, R.

Slayden, Y. Yuan, Mycolic acids: Structure, biosynthesis and physiological functions, Prog. Lipid Res. 37 (1998) 143–179. https://doi.org/10.1016/S0163-7827(98)00008-3.

[7] G. Mashabela, T. de Wet, D. Warner, Mycobacterium tuberculosis Metabolism, Microbiol. Spectr. 7 (2019). https://doi.org/10.1128/microbiolspec.GPP3-0067-2019.

[8] M. Prasad, R. Bhole, P. Khedekar, R. Chikhale, Mycobacterium enoyl acyl carrier protein reductase (InhA): A key target for antitubercular drug discovery, Bioorg. Chem. 115 (2021) 105242. https://doi.org/10.1016/j.bioorg.2021.105242.

[9] S.S. Bhunia, M. Saxena, A.K. Saxena, Ligand- and Structure-Based Virtual Screening in Drug Discovery BT - Biophysical and Computational Tools in Drug Discovery, in: A.K. Saxena (Ed.), Springer International Publishing, Cham, 2021: pp. 281–339. https://doi.org/10.1007/7355_2021_130.

[10] G. Lanka, D. Begum, S. Banerjee, N. Adhikari, Y. P, B. Ghosh, Pharmacophore-based virtual screening, 3D QSAR, Docking, ADMET, and MD simulation studies: An in silico perspective for the identification of new potential HDAC3 inhibitors, Comput. Biol. Med. 166 (2023) 107481. https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107481.

[11] J. Ricci-Lopez, S.A. Aguila, M.K. Gilson, C.A. Brizuela, Improving Structure-Based Virtual Screening with Ensemble Docking and Machine Learning, J. Chem. Inf. Model. 61 (2021) 5362–5376. https://doi.org/10.1021/acs.jcim.1c00511.

[12] F. Zare, E. Ataollahi, P. Mardaneh, A. Sakhteman, V. Keshavarz, A. Solhjoo, L. Emami, A combination of virtual screening, molecular dynamics simulation, MM/PBSA, ADMET, and DFT calculations to identify a potential DPP4 inhibitor, Sci. Rep. 14 (2024) 7749. https://doi.org/10.1038/s41598-024-58485-x.

[13] V. Jha, N. Kaur, K. Thakur, V. Dhamapurkar, P. Kapadia, S. Tiwari, A. Sahu, D. Nikumb, A. Kumar, S. Narvekar, Molecular Docking and Molecular Dynamic Simulation of Potential

Inhibitors of Integrase from Human Immunodeficiency Virus 1 (HIV-1) Using Phytochemicals, Comput. Biol. Bioinforma. 10 (2022) 34–48. https://doi.org/10.11648/j.cbb.20221001.16.

[14] Q. Zhang, J. Han, Y. Zhu, F. Yu, X. Hu, H.H.Y. Tong, H. Liu, Discovery of novel and potent InhA direct inhibitors by ensemble docking-based virtual screening and biological assays, J. Comput. Aided. Mol. Des. 37 (2023) 695–706. https://doi.org/10.1007/s10822-023-00530-4.

[15] C. Hanwarinroj, N. Phusi, B. Kamsri, P. Kamsri, A. Punkvang, S. Ketrat, P. Saparpakorn, S. Hannongbua, K. Suttisintong, P. Kittakoop, J. Spencer, A.J. Mulholland, P. Pungpo, Discovery of Novel and Potent InhA Inhibitors By an In Silico Screening and Pharmacokinetic Prediction, Future Med. Chem. 14 (2022) 717–729. https://doi.org/10.4155/fmc-2021-0348.

[16] G. el Haddoumi, M. Mansouri, B. Houda, E.M. Bouricha, I. Kandoussi, L. Belyamani, A. Ibrahimi, Facing Antitubercular Resistance: Identification of Potential Direct Inhibitors Targeting InhA Enzyme and Generation of 3D-pharmacophore Model by in silico Approach, Adv. Appl. Bioinforma. Chem. Volume 16 (2023) 49–59. https://doi.org/10.2147/AABC.S394535.

[17] T. Sterling, J.J. Irwin, ZINC 15 – Ligand Discovery for Everyone, J. Chem. Inf. Model. 55 (2015) 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

[18] M.T. Rehman, M.F. AlAjmi, A. Hussain, G.M. Rather, M.A. Khan, High-Throughput Virtual Screening, Molecular Dynamics Simulation, and Enzyme Kinetics Identified ZINC84525623 as a Potential Inhibitor of NDM-1, Int. J. Mol. Sci. 20 (2019). https://doi.org/10.3390/ijms20040819.

[19] T. Almeleebia, S. Ahamad, I. Ahmad, A. Alshehri, A. Alkhathami, Y. Alshahrani, M. Asiri, A. Saeed, J. Siddiqui, M. Saeed, Identification of PARP12 Inhibitors By Virtual Screening and Molecular Dynamics Simulations, Front. Pharmacol. 13 (2022). https://doi.org/10.3389/fphar.2022.847499.

[20] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B.

Yu, J. Zhang, S.H. Bryant, PubChem Substance and Compound databases, Nucleic Acids Res. 44 (2016) D1202–D1213. https://doi.org/10.1093/nar/gkv951.

[21] O. Trott, A.J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455–461. https://doi.org/https://doi.org/10.1002/jcc.21334.

[22] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, J. Chem. Inf. Model. 61 (2021) 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203.

[23] M. Al Shahrani, R.M. Gahtani, M. Makkawi, C-5401331 identified as a novel T-cell immunoglobulin and mucin domain-containing protein 3 (Tim-3) inhibitor to control acute myeloid leukemia (AML) cell proliferation, Med. Oncol. 41 (2024) 63. https://doi.org/10.1007/s12032-023-02296-z.

[24] H. Elsir Khair, B. Ahmed Mohamed, B. Yousef Nour, H. Ali Waggiallah, Prevalence of BCR-ABL T315I Mutation in Different Chronic Myeloid Leukemia patients Categories, Pakistan J. Biol. Sci. PJBS. 25 (2022) 175–181. https://doi.org/10.3923/pjbs.2022.175.181.

[25] E. Larocque, N. Naganna, C. Opoku-Temeng, A. Lambrecht, H. Sintim, Front Cover: Alkynylnicotinamide-Based Compounds as ABL1 Inhibitors with Potent Activities against Drug-Resistant CML Harboring ABL1(T315I) Mutant Kinase (ChemMedChem 12/2018), ChemMedChem. 13 (2018) 1159. https://doi.org/10.1002/cmdc.201800278.

[26] Q. Zhao, Z.E. Wu, B. Li, F. Li, Recent advances in metabolism and toxicity of tyrosine kinase inhibitors, Pharmacol. & Ther. 237 (2022) 108256. https://doi.org/10.1016/j.pharmthera.2022.108256.

[27] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (2017) 42717. https://doi.org/10.1038/srep42717.

[28] D.E. V Pires, T.L. Blundell, D.B. Ascher, pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures, J. Med. Chem. 58 (2015) 4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104.

[29] and D.J.F. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V., Gaussian 16, Revision C.01, (2016).

[30] P.C. Sumayya, G. Babu, K. Muraleedharan, Quantum chemical investigation of the antiradical property of avenanthramides, oat phenolics, Heliyon. 7 (2021) e06125. https://doi.org/10.1016/j.heliyon.2021.e06125.

[31] and D.G.T. Y. Zhao, N. E. Schultz, Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions Title, J. Chem. Theory Comput. 2 (2006) 364–82. https://doi.org/10.1021/ct0502763.

[32] Y.Z. and D.G. Truhlar, The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other function, Theor. Chem. Acc. 120 (2008) 215–41. https://doi.org/10.1007/s00214-007-0310-x.

[33] R. Dennington, T.A. Keith, J.M. Millam, GaussView {V}ersion {5.0.8}, (2008).

[34] T. Tuccinardi, What is the current value of MM/PBSA and MM/GBSA methods in drug discovery?, Expert Opin. Drug Discov. 16 (2021) 1233–1237. https://doi.org/10.1080/17460441.2021.1942836.

[35] S. Mahanta, T. Naiya, K. Biswas, L. Changkakoti, Y.K. Mohanta, B. Tanti, A.K. Mishra, T.K. Mohanta, N. Sharma, Plant Source Derived Compound Exhibited In Silico Inhibition of Membrane Glycoprotein In SARS-CoV-2: Paving the Way to Discover a New Class of Compound For Treatment of COVID-

19, Front. Pharmacol. 13 (2022) 805344.

[36] S. Ferraro, D. Biganzoli, R.S. Rossi, F. Palmisano, M. Bussetti, E. Verzotti, A. Gregori, F. Bianchi, M. Maggioni, F. Ceriotti, C. Cereda, G. Zuccotti, P. Kavsak, M. Plebani, G. Marano, E.M. Biganzoli, Individual risk prediction of high grade prostate cancer based on the combination between total prostate-specific antigen (PSA) and free to total PSA ratio, 61 (2023) 1327–1334. https://doi.org/doi:10.1515/cclm-2023-0008.

[37] J.M. Jayaraj, K. Muthusamy, Role of deleterious nsSNPs of klotho protein and their drug response: a computational mechanical insights, J. Biomol. Struct. Dyn. 42 (2024) 2886–2896. https://doi.org/10.1080/07391102.2023.2214230.

[38] D. Osmaniye, I. Ahmad, B.N. Sağlık, S. Levent, H.M. Patel, Y. Ozkay, Z.A. Kaplancıklı, Design, synthesis and molecular docking and ADME studies of novel hydrazone derivatives for AChE inhibitory, BBB permeability and antioxidant effects, J. Biomol. Struct. Dyn. 41 (2023) 9022–9038. https://doi.org/10.1080/07391102.2022.2139762.

[39] R.G. Parr, W. Yang, Density functional approach to the frontier-electron theory of chemical reactivity, J. Am. Chem. Soc. 106 (1984) 4049–4050. https://doi.org/10.1021/ja00326a036.

[40] I. Bâldea, A quantum chemical study from a molecular transport perspective: Ionization and electron attachment energies for species often used to fabricate single-molecule junctions, Faraday Discuss. 174 (2014) 37–56. https://doi.org/10.1039/c4fd00101j.

[41] P.C. Sumayya, K. Muraleedharan, Quenching of reactive species by Avenanthramides: theoretical insight to the thermodynamics of electron transfer, Theor. Chem. Acc. 143 (2024) 37. https://doi.org/10.1007/s00214-024-03111-2.

[42] P.C. Sumayya, V.M.A. Mujeeb, K. Muraleedharan, Radical scavenging capacity, UV activity, and molecular docking studies of 2′, 5′, 3, 4-Tetrahydroxychalcone: An insight into the photoprotection, Chem. Phys. Impact. 5 (2022) 100126. https://doi.org/https://doi.org/10.1016/j.chphi.2022.100126.

[43] D.C. Young, Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems, John Wiley & Sons, Inc, 2001. https://doi.org/DOI:10.1002/0471220655.

[44] D. Botten, G. Fugallo, F. Fraternali, C. Molteni, Structural Properties of Green Tea Catechins, J. Phys. Chem. B. 119 (2015) 12860–12867. https://doi.org/10.1021/acs.jpcb.5b08737.

Chapter 7

Conclusion and Future outlook

Mycobacterium tuberculosis, the bacteria that causes TB, is a global health problem, impacting millions of people worldwide. Effective treatment and control of TB are significantly hampered by drug-resistant TB (DR-TB). As DR-TB rates rise, new medications are being developed, and existing medications are being repurposed to treat DR-TB. The development of active compounds that can be utilised in future chemotherapeutic development to combat global tuberculosis resistance has been made possible by the advancement of computational techniques in drug design with a new generation of software. The contemporary method utilizes computational methods, including cheminformatics and bioinformatics, along with extensive data mining and curation, to propose biological targets and the small-molecule modulators that affect them. A variety of cheminformatics and computer-aided drug design techniques were employed in this study to aid in the identification of new and promising therapeutic candidates. These techniques provide quick and affordable ways to speed up antitubercular drug discovery and fight resistance by utilising both structure-based and ligand-based insights.

In the initial phase, we characterised the chemical space of Mtb-targeting compounds by performing a comprehensive chemoinformatics analysis of small molecules. We accomplished this by comparing two distinct types of Mtb inhibitor datasets chemical space with FDA-approved drugs and nutraceutical drugs datasets. Mtb inhibitors are typically less polar or equally polar when compared to approved drugs, according to the analysis of physicochemical property distributions. It was also discovered that Mtb inhibitors exhibited

comparable flexibility. A visual representation of the property space showed that, in comparison to the property space occupied by the other three datasets, the Mtbs dataset only occupies a small portion of the space. Furthermore, while the majority of the compounds in the Phytochemicals dataset are found in the Drugs dataset's property space, some of the compounds are found in regions of the chemical space that are not yet occupied by existing drugs. The Mtb dataset covers some of the same chemical space as some of the compounds found in the Phytochemicals and Nutraceuticals datasets. The compounds in the Mtbs dataset have less structural diversity than those in all other datasets, according to the structural diversity of the datasets calculated using the PCP descriptors and fingerprint representations. Nonetheless, the scaffold analysis found a number of promising candidates in the phytochemical and Mtbs datasets that might be used as a starting point for the rational discovery of new Mtb inhibitors. Significant redundancy was also found in the group of chemical molecules that target Mtb, according to similarity analysis

After exploring chemical space, we used landscape modeling - a methodical pairwise comparison of structural and activity similarities - to visualize the SAR that we had extracted from the InhA inhibitors dataset. We further searched for activity cliffs, and in order to provide a structural interpretation of the characteristics that lead to activity cliffs, these compounds were further examined using protein-ligand interactions. The analysis of the inhibitors' activity landscape showed a largely heterogeneous SAR, with some compound pairs located in activity cliff regions and the majority in similarity cliff regions. Within

the InhA inhibitor dataset, we found ten noteworthy activity cliff generators, each of which has crucial pharmacophoric characteristics that enhance its potency. A study of these cliffs and compound pairs revealed that they are all derivatives of pyrrolidine carboxamide, a new class of inhibitors of InhA. According to docking analysis, ligands' interactions with the protein are improved by small structural changes. Furthermore, the docking results were validated by MM-PBSA calculations and molecular dynamics simulations, which revealed reduced RMSD and free energy values. These results suggest that compounds with an activity cliff have a stronger affinity for the protein pocket. A better understanding of the essential molecular characteristics needed for InhA drug development, as well as the development of more effective and reliable QSAR models for InhA inhibitors, may result from discovering and addressing these activity cliffs.

Building on the earlier computational approaches, we then present an integrated machine learning (ML)-based QSAR pipeline that develops and evaluates predictive models capable of differentiating between compounds that are active and inactive against Mtb. After that, the trained model is used to screen a library of phytochemicals present in Indian medicinal plants. Lastly, molecular docking was used to find the best InhA inhibitors. Our results suggest that traditional Indian medicinal plants may be used to treat tuberculosis, and that our LGBM-QSAR classification model should be taken into account in the future development of Mtb inhibitors prediction. Additionally, the phytochemicals screened could be utilised to create new TB drugs. All things considered, this approach provides a scalable and repeatable

pipeline for integrating computational methods into early-stage screen for anti-TB drugs.

The last part of this study involved finding a lead compound that targets InhA using a structure-based virtual screening (SBVS) method in combination with molecular docking, molecular dynamics simulations (MDs), and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations. The chosen ligands were also assessed for drug-likeness, physicochemical properties, and pharmacokinetics. The electronic structure was then evaluated with DFT. Five top-ranking compounds were selected through virtual screening-based molecular docking. All chosen compounds had comparable RMSD, RMSF, hydrogen bond, Rg, and SASA values, according to MD simulation analysis, suggesting favourable interactions and general stability of the ligand-protein complexes. The thermodynamic stability of the complexes was validated by MM-PBSA calculations, as all compounds under test had negative Gibbs free energy values. The most promising compound in terms of safety and pharmacokinetics was determined to be ZINC82139221, based on pharmacokinetic and ADMET profiling. ZINC4090770, ZINC401340, ZINC49940, and ZINC35877800 were among the other candidates that demonstrated notable inhibitory activity and favourable characteristics. DFT analysis revealed that all of the ligands had a balanced charge distribution, which makes it easier for the compound to bind to biological enzymes.

Therefore, throughout the entire study, we have attempted to build QSAR models that can forecast the InhA inhibitory potency of

any given compound and to broaden the InhA chemical space by finding new InhA inhibitors through virtual screening. The current work may be extended in a number of ways in the future.

- The in vitro examination of possible InhA hits that are filtered through different cheminformatics tool.
- Perform additional research on cliff generators that are present in the structure-activity landscape of InhA inhibitors' activity cliff region.
- Utilising a smooth SAR region of the structure-activity landscape of InhA inhibitors, generate QSAR models of InhA inhibitors.
- Our ML-based QSAR classification model can be deployed as a web tool to identify InhA inhibitors in the future.
- Identify new InhA inhibitors using ligand structure-based drug design methods to broaden the chemical space.

DETAILS OF PUBLICATIONS

Details of published papers				
In International journals				
No.	Details of paper	Journal	Impact factor	ISSN/ ISBN
1.	Unleashing the potential of cheminformatic analysis for Mycobacterium tuberculosis inhibitors: Insights into chemical space and structural diversity V.K. Jalala, K. Muraleedharan Hybrid Advances, Volume 6,August 2024, 100235	Hybrid Advances, - Elsevier	0.00	2773-207X
2.	Activity landscape modeling of InhA inhibitors: Characterizing potency variations through structural similarity V.K. Jalala, K. Muraleedharan Volume 6, April 2026, 101391	Next Research - Elsevier	0.00	3050-4759
3.	Visualization of UV and ECD spectra of E&Z isomers of N-(4'-Hydroxycinnamoyl)-5-hydroxyanthanilic acid P.C.Sumayya, V.K.Jalala, T.K.Shameera Ahammed, K.Muraleedharan, Computational Biology and Chemistry, Volume 101, December 2022, 107777	Computational Biology and Chemistry-Elsevier	3.1	1476-9271

4.	Complexation behaviour of piceatannol ligand with Ti (IV) and Zr (IV) metal ions: a combined DFT and deep learning investigation, P.U. Neenu Krishna, V.K. Jalala, K. Muraleedharan, Struct. Chem. (2023).	Structural Chemistry-Springer	2.1	1572-9001

Seminar proceedings

1. Predictive Modeling of MAO-B Inhibitory Compounds: A Machine Learning-Enhanced QSAR Analysis, Jalala V.K., Muraleedharan K, International Conference on Emerging Frontiers in Chemical Sciences (EFCS-2023) ISBN:978-81-953059-9-5

Seminar presentations

International	
1.	Presented a poster on the topic 'Predictive Modeling of MAO-B Inhibitory Compounds: A Machine Learning-Enhanced QSAR Analysis' at the International Conference on Emerging Frontiers in Chemical Sciences, held at Farook College in 2023
2.	Presented a poster titled 'Molecular-Level Virtual Screening of Nutraceuticals Against COVID-19 Using Artificial Intelligence and Machine Learning' at the Virtual International Chemical Science Symposium 2020 on 'How Can Machine Learning and Autonomy Accelerate Chemistry?', organized by the Royal Society of Chemistry
3	Presented a poster at the International Conference on Advanced Materials for Sustainability (ICAMS 2023), organized by the School of Physical Sciences, University of Calicut, held December 2023
National	
1.	Oral presentation on the topic “Machine Learning-Based Prediction of Monoamine Oxidase B Inhibitors: A Comparative Analysis of Regression Models” at the two-day National seminar “Computational Chemistry Summit-2023” conducted by the Department of Chemistry, Government College, Malappuram.

Computational training programs attended

1. Two-day hands-on training program provided by Zamorin's Guruvayurappan College, Calicut in different tools of molecular docking.
2. Two-day workshop on "In Silico Drug Design using BIOVIA Discovery Studio" held at Amrita School of Pharmacy, Kochi, Kerala.
3. 60-hour training on "Artificial Intelligence-based Machine Learning in Drug Discovery" organized by Open Source Pharma Foundation, Bengaluru.