Machine learning and Tanimoto similarity based virtual screening of Chemical Libraries to Develop Novel PD Therapeutics

by

Divyanshi Singhal

Under the supervision of Dr. N Arul Murugan

Indraprastha Institute of Information Technology Delhi
May, 2024

Machine learning and Tanimoto similarity based virtual screening of Chemical Libraries to Develop Novel PD Therapeutics

by

Divyanshi Singhal

Submitted
in partial fulfilment of the requirements for the degree of
Master of Technology

to

Indraprastha Institute of Information Technology Delhi
May, 2024

Certificate

This is to certify that the thesis titled "**Machine learning and Tanimoto similarity based virtual screening of Chemical Libraries to Develop Novel PD Therapeutics**" being submitted by **Divyanshi Singhal** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May, 2024

Dr. N Arul Murugan
Department of Computational Biology
Indraprastha Institute of Information Technology
Delhi, New Delhi 110 020

Acknowledgements

I extend my sincerest appreciation to Dr. N Arul Murugan, my thesis supervisor, for his valuable mentorship and constant support throughout the work. The completion of this thesis would not have been possible without his expert knowledge and encouragement.

I would like to thank all the faculty and staff at IIIT Delhi, for their consistent assistance throughout the process. I am thankful to my fellow batchmates for their encouragement and support. I also thank to Prateek Paul from Ph.D. 2022 batch for his valuable contribution to this work.

Finally, I would like to thank my family, who have been my constant source of strength and inspiration, and without them, this academic journey would not have been possible

Abstract

Parkinson's disease is a prevalent neurodegenerative disorder that affects a significant portion of the aging population worldwide. The etiology of Parkinson's disease involves the accumulation of misfolded proteins, particularly alpha-synuclein, leading to the formation of toxic aggregates in the brain. These aggregates are believed to contribute to the degeneration of dopamine-producing neurons in the substantia nigra, resulting in motor and cognitive impairments characteristic of the disease. Research in cheminformatics and computational chemistry offers a promising avenue for identifying small molecule compounds capable of modulating the aggregation of alpha-synuclein or mitigating its neurotoxic effects. By leveraging computational tools and databases of chemical compounds, researchers can screen and prioritize potential drug candidates with the desired properties for targeting Parkinson's disease pathology. Protein-ligand binding affinities play a crucial role in drug discovery and development. Accurately predicting these affinities can significantly streamline identifying potential drug candidates. In recent years, machine learning models have emerged as powerful tools for predicting protein-ligand binding affinities. The objective is to identify potential selective MAO-B and LRRK2 inhibitors among natural products and repurpose existing drugs for Parkinson’s disease treatment, employing LBVS and validating through docking and ADMET properties prediction. The use of high-quality benchmarking datasets, including a novel selectivity oriented dataset for MAO-B and LRRK2, is essential for both retrospective and prospective validation of virtual screening methodologies in this context. One such model is SVR, which has shown promising results in this domain. However, it is essential to thoroughly validate and benchmark to assess its reliability and performance compared to existing methods. Furthermore, the use of computational simulations enables the exploration of the interactions between small molecules and their target proteins at the molecular level, providing insights into the structural determinants of binding and potential mechanisms of action. We used DrugBank,PubChem, Real Enamine, and the Zinc Database to offer unique attributes conducive to the drug discovery process. This knowledge can inform the rational design and optimization of novel compounds with enhanced efficacy and safety profiles, offering new possibilities for therapeutic intervention in Parkinson's disease. Machine learning models trained on the training set achieved r squared value of MAO-B 69.0 and LRRK2 60.0. In the pursuit of discovering novel compounds targeting Parkinson's disease, the integration of cheminformatics and computational chemistry holds the potential to accelerate the identification and development of innovative treatments that address the underlying molecular mechanisms of the disorder.

Keywords: Parkinson Disease, MAOB, LRRK2, Machine Learning, AutoDock vina, Drug Discovery, 1GOS, 7LI3

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Parkinson Disease

Parkinson's disease (PD) is a progressive neurodegenerative disorder that causes motor symptoms like tremors, bradykinesia (slowness of movement), rigidity (stiffness), and postural instability. Patients with PD may also experience non-motor symptoms such as sleep disturbances, cognitive changes, and autonomic dysfunction. The diagnosis of PD is based on clinical assessment, which includes medical history, physical examination, and evaluation of specific motor symptoms [16]. There are no specific diagnostic tests for Parkinson's disease, so the diagnosis is often challenging and may require the expertise of a neurologist. The molecular mechanisms underlying PD involve the dysfunction and degeneration of dopaminergic neurons in the "substantia nigra" region of the brain that leads to a decrease in dopamine levels, contributing to the disease's motor symptoms. Additionally, the accumulation of abnormal protein aggregates, such as alpha-synuclein, plays a role in the pathogenesis of the disease.

Treatment options for PD aim to progress the patient's quality of life and alleviate symptoms. It may involve dopaminergic medications, deep brain stimulation, physical therapy, and lifestyle modifications. While these treatments can provide relief, they do not halt the progression of the disease [4].

In the recent years, there has been ongoing research into new treatment strategies to modify the disease process. These include neuroprotective therapies, gene therapy, and stem cell therapy, which hold promise for potentially slowing or halting the progression of PD. Managing PD requires a multidisciplinary approach involving neurologists, movement disorder specialists, physical therapists, occupational therapists, and other healthcare professionals. This comprehensive approach aims to address both the motor and non-motor symptoms of the disease and provide individualized care for patients with Parkinson's disease

Recent research has shed light on the interconnected pathways of MAO-B inhibition and LRRK2 inhibition, offering a promising avenue for the rational design of novel compounds. By targeting both MAO-B and LRRK2 inhibition, these compounds hold the potential for multifaceted therapeutic benefits in treating neurological disorders, such as Parkinson's disease.

Understanding the intricate mechanisms underlying MAO-B and LRRK2 activity and its role in oxidative stress is crucial for designing effective compounds. A deeper exploration into the structural and functional aspects of MAO-B and LRRK2 and its interactions with oxidative stress pathways is necessary to identify critical targets for intervention [21].

Furthermore, developing novel compounds requires an in-depth analysis of their pharmacokinetic and pharmacodynamic properties to ensure optimal efficacy and safety profiles. By leveraging advanced computational modeling and structure-based drug design, it is possible to tailor compounds with enhanced selectivity and potency against MAO-B and LRRK2 while mitigating oxidative stress.

Incorporating these more profound insights into the rational design of novel compounds will expand the therapeutic arsenal for neurological disorders and pave the way for innovative drug discovery and development strategies. Monoamine oxidases (MAOs) playing a crucial role in the metabolism of monoamine neurotransmitters like serotonin, norepinephrine, and dopamine. These enzymes, existing as MAO-A and MAO-B isoforms, are implicated in various neuropsychiatric disorders [8]. Despite the therapeutic potential of MAO inhibitors, their usage is hindered by side effects and safety concerns. Selective irreversible inhibitors, such as selegiline and rasagiline, and the reversible MAO-B inhibitor, safinamide (SAG), have been introduced for Parkinson’s disease treatment. However, overcoming drawbacks necessitates continuous identification of new inhibitors through rational design strategies.

1.2 Drug Discovery

The drug discovery process is an intricate and resource-intensive endeavor involving the identification of potential disease targets and the subsequent search for suitable drug candidates within a vast chemical space. Critical considerations include high binding affinity and specificity to the target alongside optimal properties of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) [2]. Leveraging advancements in computational methods, particularly within machine learning and deep learning, have substantially mitigated the time and cost associated with drug prediction. Computational drug discovery amalgamates principles from chemistry, biology, and computer science, expediting the identification and development of novel therapeutics across various diseases.

Chemoinformatics tools, integral in contemporary drug development, facilitate the identification of novel drug candidates and ligand–target interactions. Ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) methods, utilizing similarity-based approaches and molecular docking, respectively, play pivotal roles in hit/lead identification and optimization [7]. This analysis focuses on uncovering shape patterns, physicochemical features, and specific ligand–receptor interactions correlated with selective and nonselective MAO-B inhibition. The objective is to identify potential selective MAO-B and LRRK2 inhibitors among natural products and repurpose existing drugs for Parkinson's disease treatment, employing LBVS and validating through docking and ADMET properties prediction. The use of high-quality benchmarking datasets, including a novel selectivity oriented dataset for MAO-B and LRRK2, is essential for both retrospective and prospective validation of virtual screening methodologies in this context.

1.3 Drug-like Properties

Drug design is the optimization of ligand properties, encompassing binding affinity, specificity, ADMET characteristics, solubility, and bioavailability. Binding affinity quantifies the effectiveness of a ligand in binding to its target. In contrast, binding specificity denotes its ability to bind to the intended target while minimizing off-target interactions selectively. ADMET plays a pivotal role in determining a drug's efficacy and safety profile [20].

1.4 Virtual screening

In drug discovery, computational methods such as virtual screening play a pivotal role. Virtual screening facilitates the identification of potential drug candidates by screening vast chemical libraries to pinpoint molecules with high binding potential to specific targets. This method, notably faster and more efficient than traditional High Throughput Screening (HTS), encompasses ligand-based and structure-based approaches [9].

1.4.1 Ligand/ Structure based virtual screening

LBVS on the similarity principle, wherein compounds with similar properties to a reference compound are identified based on descriptors such as 1D, 2D, or 3D properties. Conversely, structure-based virtual screening utilizes the 3D structure of the target receptor, obtained through techniques like X-ray Crystallography or NMR spectroscopy. By docking ligands to the receptor structure, the most favorable binding poses can be determined, aided by various scoring functions ranging from physics-based to machine learning-based approaches [3].

1.4.2 Machine Learning based virtual screening

Machine learning based scoring functions, including deep learning models, have emerged as powerful tools in predicting protein-ligand binding affinity with high accuracy. These models leverage complex data relationships to enhance predictive capabilities, often outperforming traditional empirical scoring functions. However, such models' reliability and accuracy depend on the training data's quality and diversity, necessitating thorough validation and benchmarking against experimental datasets to ensure robustness and consistency [5].

1.5 Scope of the current study

Even with the availability of numerous publications, patents, and proceedings related to the pathogenesis of Parkinson's disease, technological advancement, structural information, and molecular modeling for the development of MAOB and LRRK2 inhibitors. There are still unmet clinical needs in the treatment of PD.

To understand the binding properties of MAOB and LRRK2 inhibitors, we used machine learning and virtual screening approaches in this study. Our goal is to compare the results and determine the best method for predicting the pIC50 of MAOB and LRRK2 active compounds.

1.6 Molecular mechanisms and pathogenesis of Parkinson's disease

The pathogenesis of the disease involves a complex interplay of genetic and environmental factors, leading to the accumulation of abnormal protein aggregates, such as alpha-synuclein, in the brain. These protein aggregates contribute to the formation of Lewy bodies, which are a hallmark pathological feature of PD.

Furthermore, this work will delve into the role of mitochondrial dysfunction, oxidative stress, and neuroinflammation in the pathogenesis of PD. It will explore the intricate network of molecular pathways involved in the degeneration of dopaminergic neurons and the subsequent motor and non-motor symptoms that manifest in affected individuals. Additionally, this work will discuss emerging therapeutic strategies targeting these molecular mechanisms, aiming to provide a deeper understanding of the disease and potential avenues for developing novel treatment approaches [23].

Finally, this work will highlight the ongoing research efforts to unravel the intricate molecular mechanisms underlying Parkinson's disease, with the ultimate goal of identifying disease-modifying interventions that could slow or halt disease progression. Parkinson's disease is a complex and multifaceted condition that continues to present challenges in both understanding its underlying mechanisms and developing effective treatments. One active research area involves investigating the potential links between gut microbiota and Parkinson's disease. Recent studies have suggested that changes in the gut microbiome may influence the development and progression of Parkinson's disease, possibly through inflammatory and neuroimmune mechanisms.

Moreover, the emerging field of precision medicine offers promise in tailoring treatments to individual patients based on their unique genetic and molecular profiles. By understanding the specific molecular pathways involved in an individual's Parkinson's disease, researchers and healthcare professionals aim to develop targeted therapies that address the condition's underlying causes.

As research in this field progresses, it becomes increasingly clear that a multidisciplinary approach is crucial for advancing our understanding of PD and developing more effective treatments [1]. Collaboration between researchers, clinicians, and industry partners is essential for translating scientific discoveries into meaningful clinical interventions. By pooling expertise and resources, the scientific community can work towards improved patient care and the eventual conquest of Parkinson's disease.

Fig. 1 The evolution of Parkinson's disease research and drug development

Chapter 2

Materials and Methods

2.1 Dependencies Installation

This section delineates the system requisites and setup for our study. Ubuntu 18.04 served as the operating system for data processing, while the HPC server at IIIT Delhi was utilized for modeling purposes. Although Ubuntu 18.04 was employed for data preprocessing, the codes are compatible with Windows systems utilizing a Python compiler. Python, a versatile and user-friendly programming language, was selected for its robustness and extensive community support. Python 3, the latest version, was chosen for its enhanced security features and ongoing development efforts. Installation and configuration of Python entailed administrative access to a machine with Ubuntu 18.04 and an internet connection.

The Python environment setup commenced with command-line interaction, facilitated through the terminal application on Ubuntu. Both Python 2 and Python 3 were already pre-installed in Ubuntu 18.04, with system upgrades performed via apt-get to ensure the latest Python version. Pip, a package management tool for Python, was installed to facilitate the management of Python software packages. Additional Python packages were installed using pip3, enabling seamless integration of essential libraries such as Pandas, NumPy, and Biopython.

Furthermore, Ubuntu provides the option to create virtual environments, offering isolated spaces for individual Python projects. This approach allows distinct sets of dependencies and Python package versions for each project, minimizing interference between projects and affording greater project control. The venv module, inherent within the standard Python 3 library, was installed to enable virtual environment creation. Following directory creation and virtual environment setup using the venv module, the environment was activated through the source command, granting access to install required packages and libraries specific to the project. In this study, we used AutoDock Vina to perform molecular docking.

The following commands were used to install all of these libraries in the same environment.

- pip install pandas
- pip install numPy
- pip install biopython
- pip install perl

Pandas is a Python library specialized in manipulating and analyzing data facilitating tasks like cleaning, transforming, and aggregating structured data. NumPy, another Python library, is fundamental for numerical computing, offering support for multi-dimensional arrays and mathematical functions essential for scientific computing. Biopython is tailored for bioinformatics, providing modules for parsing and analyzing biological data. At the same time, Perl, a versatile programming language, complements Python in bioinformatics workflows, particularly for text processing and system tasks. These libraries form a comprehensive toolkit for researchers, enabling efficient data analysis, numerical computations, and bioinformatics research in scientific endeavors [11].

In the realm of bioinformatics, Biopython emerges as a vital resource for researchers grappling with biological data. Tailored specifically for genomics, proteomics, and molecular biology tasks, Biopython offers a wide range of modules and functionalities for sequence analysis, structure prediction, and data visualization. Its accessible interface and thorough documentation ensure that it is user-friendly, catering to both novice and seasoned bioinformaticians alike to novice and experienced bioinformaticians, facilitating the exploration of genetic sequences, the study of protein structures, and the analysis of evolutionary relationships [17]. With Biopython, researchers can unlock the mysteries of the biological world and advance our understanding of life's intricacies.

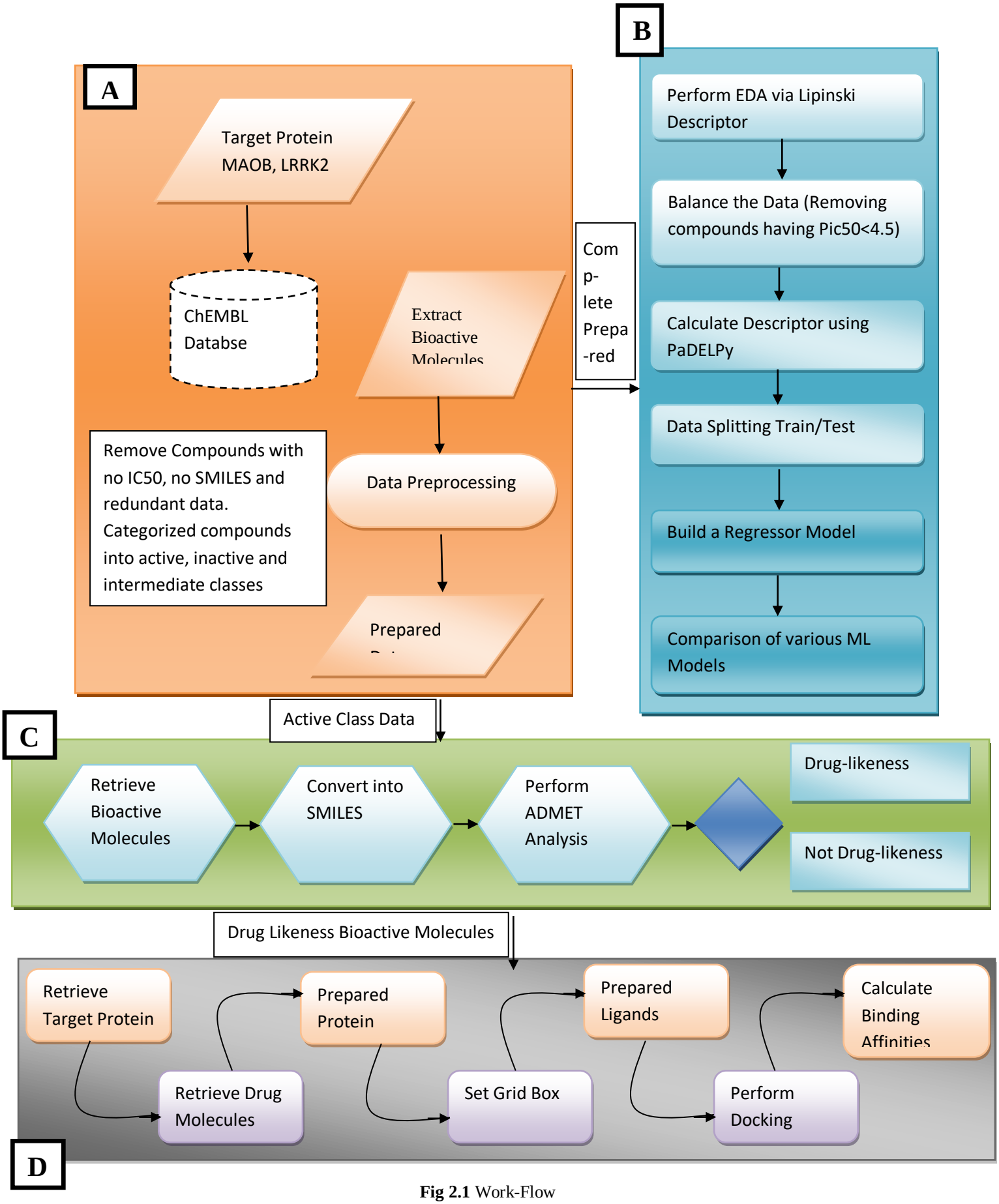


Fig 2.1 Work-Flow

2.2 Data Collection

Utilizing the open-access chEMBL database and its corresponding web interface, along with its available open-source code on GitHub, curated biological activities of over two billion compounds were accessed. Initial filtration involved selecting data entries with confidence scores equal to or exceeding 6, denoting the confidence level of target protein assignment to the compounds. Explicitly focusing on in-vitro binding assays (assay type: B), standard units of nanomolar (nm) were isolated, and p-activity values (pIC50) were derived from -log(IC50) measurements. Further refinement involved selecting IC50 values with associated activity comments that were neither "None" nor "Not Determined" while ensuring duplicate entries were removed. Subsequently, chemical compound structures were extracted in SMILES (simplified molecular-input line-entry system) format from chEMBL, with subsequent processing steps involving the removal of salts and solvents, neutralization, and conversion into canonical SMILES representations.

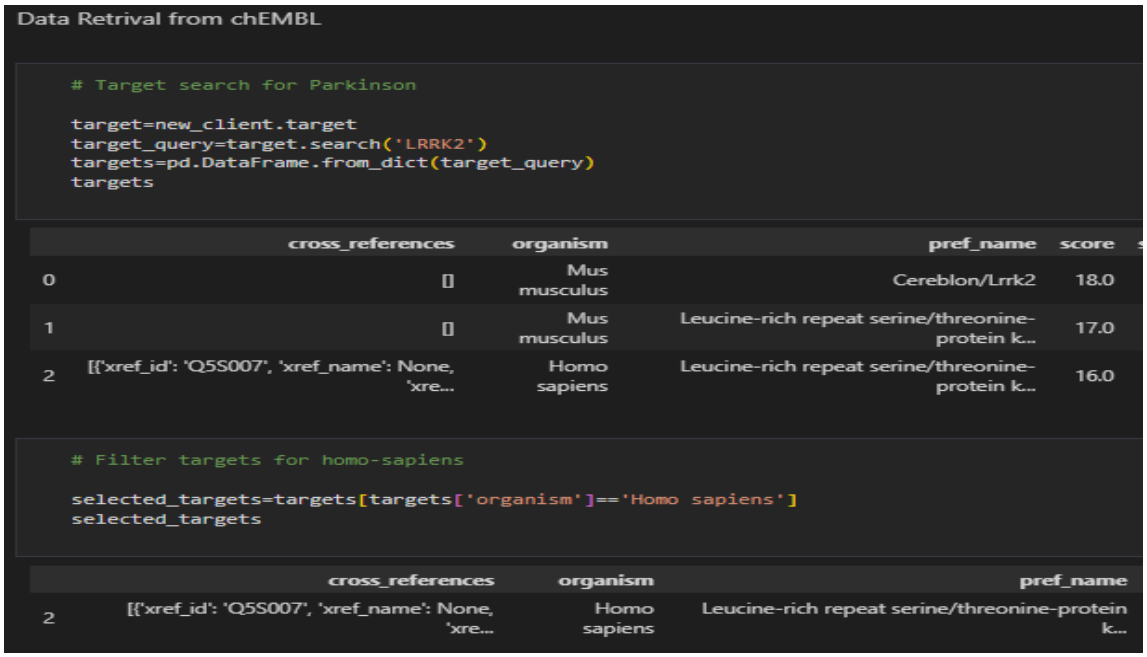


Fig 2.2.1 Data Retrieval from chEMBL

2.3 Targets

1) **MAO-B:** Monoamine oxidase-B (MAO-B) inhibitors represent a standard pharmacological approach for managing symptomatic manifestations of Parkinson's disease (PD). Monotherapy with MAO-B inhibitors demonstrates efficacy and safety in addressing early-stage PD, while their adjunctive use is prevalent in advanced disease stages. Notably, MAO-B inhibitors can ameliorate both the motor and the non-motor symptoms in PD patients and reduce "OFF" periods, potentially exerting a disease-modifying effect by delaying progression.

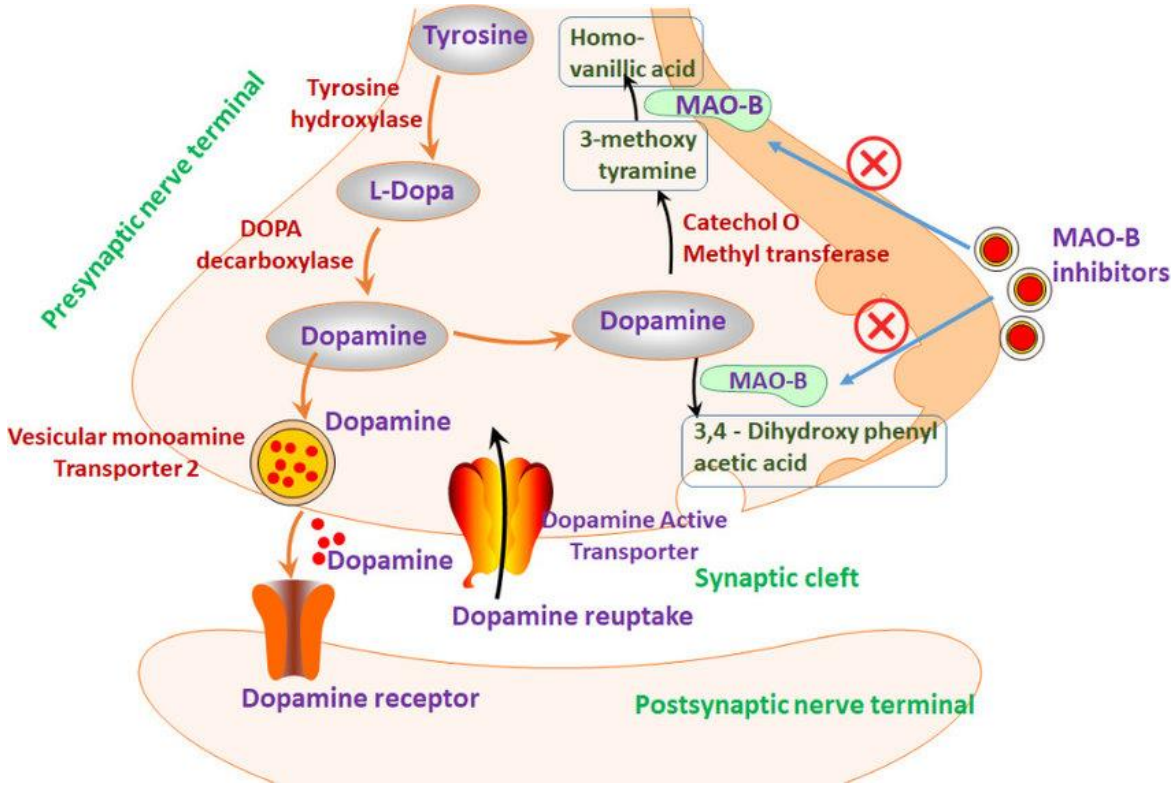


Fig 2.3.1.1 Mechanism of MAO-B-inhibitors-in-PD, DOI:10.3390/ijms21207648

2) **LRRK2 (Leucine-rich repeat kinase 2):** LRRK2 is a prominent member of the kinase protein family, characterized by multiple domains and sizeable molecular structure. It has emerged as a significant target of interest in the context of Parkinson's disease, a prevalent age-related neurodegenerative disorder. Mutations in the LRRK2 gene have been implicated in PD pathogenesis, with the G2019S variant (Gly2019 mutated to Ser) being the most common genetic alteration, accounting for 4–6% of hereditary PD cases and 1–2% of sporadic cases. This specific mutation, located in the kinase domain of LRRK2, is associated with heightened kinase activity compared to the wild-type form, offering a compelling rationale for exploring LRRK2 inhibition as a therapeutic strategy for PD.

   Moreover, studies using LRRK2 transgenic animal models have demonstrated Parkinsonian-like features, including dopaminergic neuron loss, behavioral abnormalities, and impaired locomotion. Notably, these effects are more pronounced in animals expressing mutant forms of LRRK2, underscoring the significance of LRRK2 dysfunction in PD pathophysiology and reinforcing the potential utility of LRRK2-targeted therapies in mitigating disease progression. LRRK2 phosphorylates a subset of RAB GTPases, which regulate vesicular trafficking and immune responses. Mutations that activate LRRK2 kinase are linked to PD.

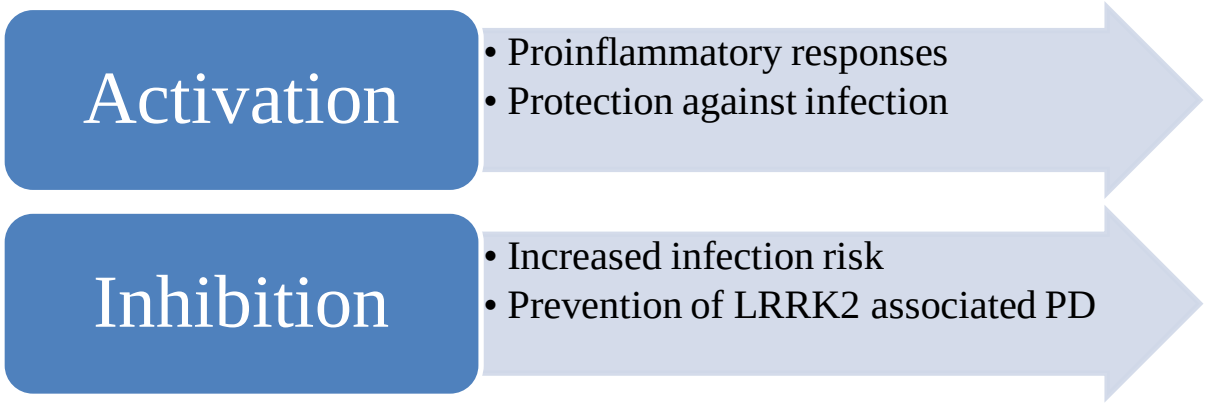


Fig 2.3.2.1 Predicted impacts of activating and inhibiting LRRK2

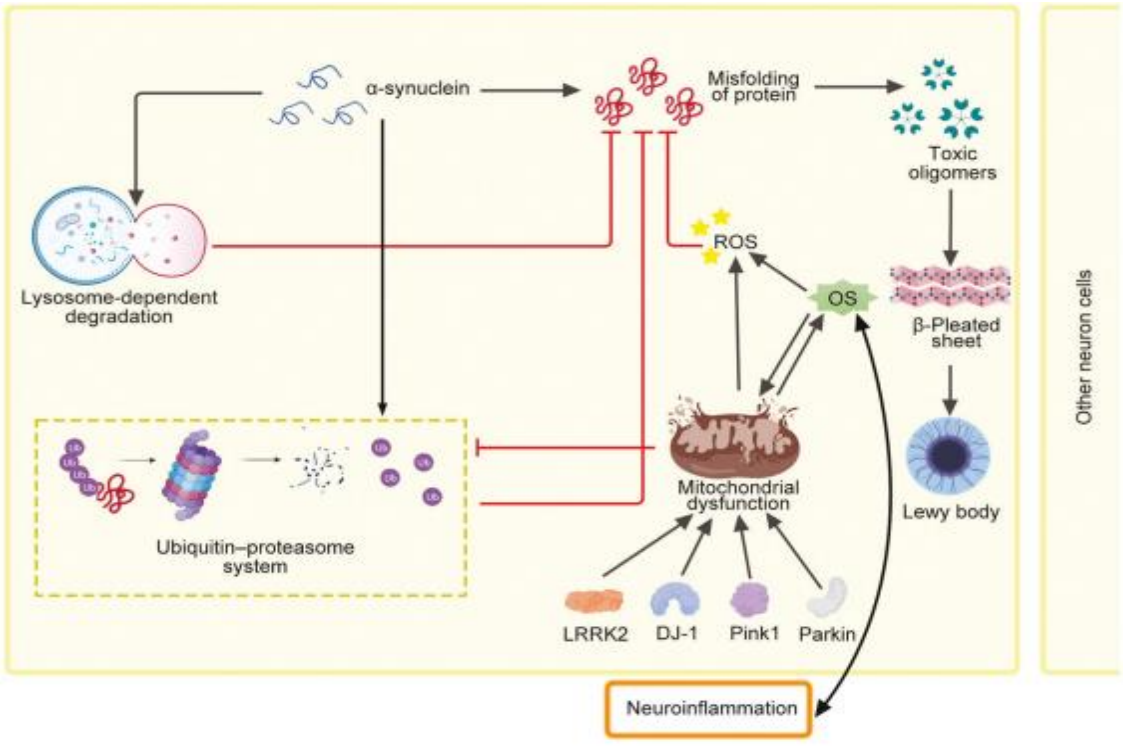


Fig 2.3.2.2 LRRK2 Pathway: The intracellular equilibrium of α-synuclein is upheld through the intricate interplay of ubiquitin-proteasome and lysosomal autophagy mechanisms. Dysfunction within these degradation pathways induced by oxidative stress (OS), mitochondrial impairment, or neuroinflammation may foster α-synuclein aggregation. Moreover, mutations affecting genes such as LRRK2, DJ-1, Parkin, and Pink1 are implicated in mitochondrial dysfunction, thereby exacerbating cellular demise. Additionally, a consequential association between oxidative stress and neuroinflammatory processes emerges, underscoring their intertwined roles in disease progression. https://doi.org/10.1038/s41392-023-01353-3

- Mutation in the LRRK2 locus, Gly^{2019}--->Ser is by far the most common.

2.4 Preparing Data for feature generation

SMILES strings serve as versatile inputs for molecular editors and cheminformatics software, facilitating the generation of 2D or 3D molecular structures. These structures are instrumental in diverse applications, ranging from machine learning modeling through data-driven approaches to virtual screening and molecular docking studies [19]. Conversion of SMILES strings to other file formats, such as mol2, using tools like Open Babel, an open-source chemical file format converter adept at handling various chemical formats.

SMILES offer a concise and easily interpretable means of representing molecular structures through character strings. This representation delineates heavy atoms with explicit atomic symbols, while hydrogen atoms are often implicit. For instance, CH4-C and NH3-N denote carbon and nitrogen atoms with associated hydrogens. Additionally, elements beyond the organic subset, denoted by symbols such as [Au], are accommodated. Bond types, including single, double, triple, or aromatic bonds, are symbolized by -, =, or #, respectively, while branching structures are encapsulated within parentheses (). Stereochemistry, denoted by @, further refines the depiction of molecular configurations.

The RDKit, a comprehensive software library implemented in Python, is a notable resource for cheminformatics and machine learning endeavors. Its functionalities encompass generating molecular structures from SMILES strings, structural searches, and computation of diverse chemical properties, enhancing research capabilities in the field. Furthermore, many tools and online platforms, including ChemAxon and the PubChem website, offer similar capabilities for converting SMILES to mol2 formats, thereby broadening the accessibility of such molecular modeling tools [6]. Notably, our study leveraged the capabilities of RDKit and Open Babel for comprehensive analyses.

2.4.1 Adding Features by PaDEL Descriptors

PaDEL is a robust tool capable of computing molecular descriptors and fingerprints. PaDEL currently offers a comprehensive set of 1875 descriptors, comprising 1444 for 1D and 2D properties, 431 for 3D properties, and 12 distinct fingerprints. The descriptors are generated utilizing The Chemistry Development Kit, augmented with additional descriptors such as atom type electron topological state descriptors, Crippen’s logP and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, and counts of chemical substructures [10]. Collectively, these features enable comprehensive characterization and analysis of small molecules in diverse research applications.

The Chemistry Development Kit (CDK) forms the basis for these calculations, with additional fingerprints and descriptors integrated into PaDEL's framework. This approach enhances the scope of feature extraction, facilitating a comprehensive analysis of molecular properties essential for diverse applications in cheminformatics and computational chemistry.

In drug discovery, the ability to extract and analyze molecular features with precision is paramount. PaDEL's diverse descriptor set empowers researchers to glean invaluable insights into small molecules' physicochemical properties, pharmacokinetics, and pharmacodynamics. By comprehensively characterizing molecular structures, PaDEL facilitates the identification of lead compounds, optimization of drug candidates, and prediction of biological activities [14]. Its versatility extends to virtual screening, quantitative structure-activity relationship (QSAR) modeling, and ligand-based drug design, accelerating the drug discovery pipeline.

PaDEL emerges as a pivotal tool in the arsenal of computational chemists and pharmaceutical researchers, offering an indispensable toolkit for elucidating the complex relationships between molecular structure and biological activity in the quest for novel therapeutics.

	nAcid	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond	nAtom	nHeavyAtom	nH
0	0	1.6008	2.562561	117.6211	58.249860	0	0	47	27	20
1	0	0.3997	0.159760	86.0519	44.152688	0	0	37	21	16
2	0	1.6624	2.763574	100.5407	50.481481	0	0	42	25	17
3	0	0.1536	0.023593	62.4704	33.760137	0	0	29	20	9
4	0	2.7572	7.602152	103.5727	49.485895	0	0	39	24	15
...	...	...	...	...	...	...	...	...	...	...
3371	0	1.6247	2.639650	85.7266	40.425930	0	0	31	21	10
3372	0	0.6734	0.453468	84.8359	41.938309	0	0	35	22	13
3373	0	1.5493	2.400330	91.2401	47.323481	0	0	40	23	17
3374	0	1.4545	2.115570	90.5119	42.725930	0	0	31	21	10

Fig 2.4.1.1 PaDEL Descriptor of MAOB (3776*1875)

	nAcid	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond	nAtom	nHeavyAtom	nH
0	0	1.1393	1.298004	92.7852	44.825895	0	0	37	22	15
1	0	0.5807	0.337212	86.0047	41.895102	0	0	34	20	14
2	0	-0.4409	0.194393	137.1293	73.967790	0	0	64	34	30
3	0	-2.6058	6.790194	166.1515	91.104134	0	0	80	42	38
4	0	-1.5875	2.520156	173.0941	94.657720	0	0	82	42	40
...	...	...	...	...	...	...	...	...	...	...
1756	0	-0.8291	0.687407	109.4646	54.164274	0	0	44	26	18
1757	0	-0.9900	0.980100	114.4431	56.357481	0	0	45	28	17
1758	0	-0.5339	0.285049	112.4294	55.555481	0	0	44	27	17
1759	0	-0.9121	0.831926	115.4688	58.539274	0	0	47	29	18
1760	0	0.0806	0.006496	120.5555	61.413274	0	0	50	32	18

Fig 2.4.1.2 PaDEL Descriptor of LRRK2 (1761*1875)

2.5 Protein Preparation

The Protein Data Bank (PDB) is a database that stores three-dimensional structural data of large biological molecules, such as proteins and nucleic acids, obtained through techniques including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. In our study, the structural information of the target protein, MAOB and LRRK2 (PDB id: 1GOS and 7LI3), was obtained from RCSB-PDB to gain insights into its functional and physical characteristics and to investigate amino acid interactions with ligands.

The three-dimensional structure of MAOB and LRRK2, comprising amino acid positions (1–520 and 1-2527), was retrieved in PDB format. This structure, resolved through X-ray diffraction and Electron Microscopy with a resolution of 3 and 3.8 Angstroms, was employed for further analysis. The protein preparation process was conducted utilizing AutoDock vina and the MGL tool to facilitate subsequent molecular docking studies.

Multiple steps were undertaken during the preparation of the MAOB and LRRK2 protein. Initially, the protein molecules were loaded into AutoDock, followed by removing water molecule heteroatoms to eliminate potential interference with ligand binding. Additionally, polar hydrogen atoms were added to the protein structure to enhance the detection of hydrogen bond interactions, thereby promoting favorable ligand-protein binding interactions. Kollman charges were assigned to the protein, and subsequently, the prepared protein structure was converted into the pdbqt format, which is compatible with AutoDock vina, facilitating subsequent molecular docking simulations.

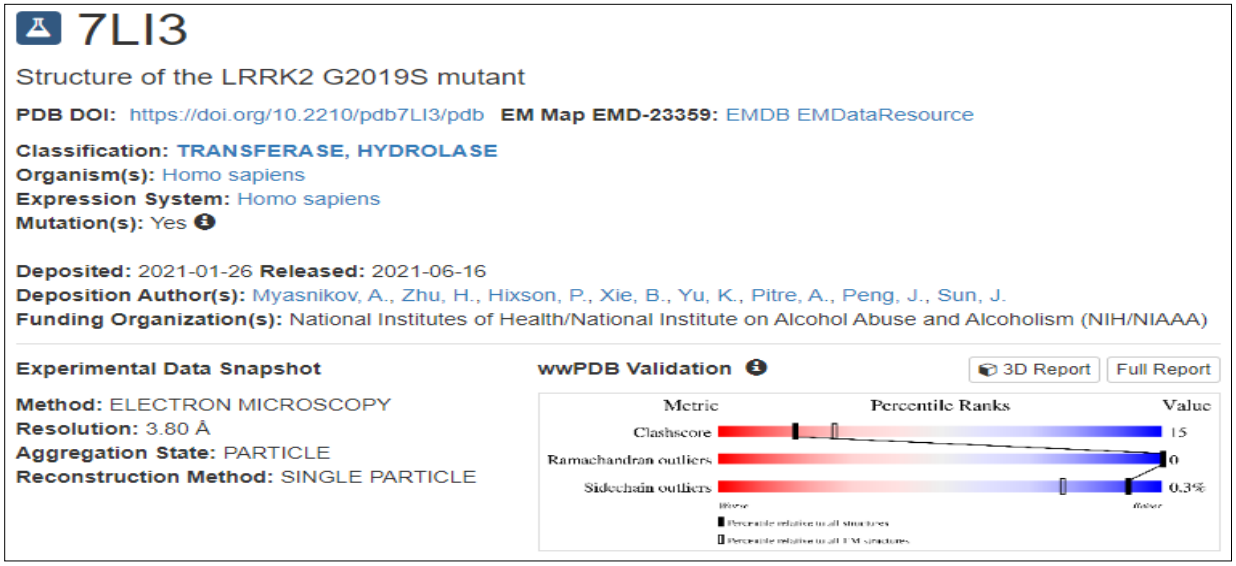


Fig 2.5.1 LRRK2 PDB ID information

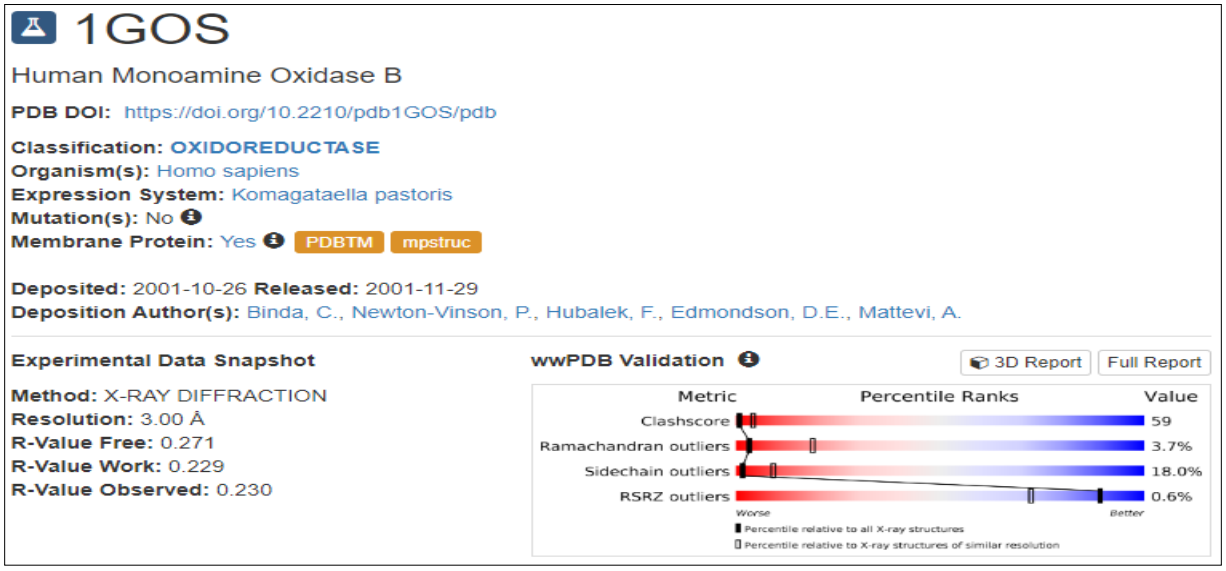


Fig 2.5.2 MAO-B PDB ID information

2.6 Ligand Preparation

Our study employed OpenBabel to facilitate the conversion of ligands represented in SMILES (smi) format to SDF (Structure Data File) format, followed by conversion from SDF to pdbqt format suitable for molecular docking simulations. This conversion process involved a standardization procedure aimed at refining the ligand structures.

The standardization procedure included several vital steps:

1. Remove salts, mixtures, and metal ions from the ligand structures to isolate the desired chemical entities for subsequent analysis.

2. Correction of the chemical structure's geometry to ensure accurate representation and alignment with the desired conformation for molecular docking studies.

By leveraging OpenBabel's capabilities, we could effectively preprocess and refine the ligand structures, preparing them for downstream computational analyses such as molecular docking against the target protein. This standardized format of ligands in pdbqt enabled seamless integration into the docking software, facilitating the exploration of ligand-protein interactions and binding affinities essential for drug discovery and molecular modeling studies.

2.7 Molecular Visualization of Interacting Molecules

Visualizing interacting molecules at the molecular level provides crucial insights into the behavior and properties of these compounds. Understanding molecules' spatial arrangement and interactions is essential for various fields such as drug design, material science, and biochemistry. Utilizing advanced molecular visualization techniques enables researchers to observe the intricate details of molecular interactions, helping to elucidate the mechanisms underlying chemical reactions and biological processes.

One of the essential techniques for visualizing interacting molecules is molecular modeling, which involves using computer software to create three-dimensional representations of molecular structures. These models allow scientists to analyze the spatial relationships between individual atoms and molecules and the specific bonds and interactions between them. Through the simulation of molecular interactions, researchers can glean significant insights into the intricate dynamics of complex systems, thereby enabling the anticipation and prognostication of outcomes about chemical and biological phenomena.

In addition to molecular modeling, X-ray crystallography and nuclear magnetic resonance spectroscopy play a vital role in visualizing molecule interactions. These experimental methods provide detailed information about the three-dimensional arrangement of atoms within a molecule and the spatial orientation of interacting molecules in a complex. By integrating experimental data with computational modeling, researchers can develop comprehensive molecular visualizations illuminating essential chemistry and biology principles [22]. These visual representations of molecular interactions serve as invaluable tools for deciphering the structure and function of biomolecules. Furthermore, they are pivotal in facilitating the design and optimization of novel drugs, catalysts, and materials endowed with specific and tailored properties. This synergistic approach bridges the gap between experimental observations and theoretical predictions, enabling more profound insights into molecular behavior and fostering advancements in diverse scientific domains [12]. As we delve deeper into molecular visualization, the integration of experimental and computational approaches continues to push the boundaries of our understanding, unlocking the potential for groundbreaking discoveries and innovations.

2.8 Chemical Libraries

DrugBank, PubChem, Real Enamine, and the Zinc Database offer unique attributes conducive to the drug discovery process. DrugBank is a comprehensive repository of drug and drug target information, providing researchers with vital data on pharmacological properties, drug interactions, and therapeutic indications. Complementing this, PubChem, a component of the NCBI-National Center for Biotechnology Information, offers an extensive collection of chemical compounds and associated biological activities, aiding in the exploration of structure-activity relationships (SAR) and the identification of lead compounds [24]. Real Enamine, renowned for its vast library of commercially available compounds, provides researchers with ready access to diverse chemical scaffolds, facilitating high-throughput screening campaigns and hit identification endeavors. Similarly, the Zinc Database is a valuable resource, offering a curated collection of purchasable compounds enriched with structural and physicochemical properties, streamlining virtual screening efforts, and leading optimization initiatives. Through the synergistic utilization of these databases, researchers are empowered to expedite the drug discovery pipeline, from target identification to clinical candidate selection, thereby advancing therapeutic interventions to address unmet medical needs.

Table 2.8.1 Chemical Database and their counts

DataBase	**Compounds Count**
DrugBank	10182
PubChem	0.115 Billion
Real-Enamine	6 Billion
Zinc	0.8 Billion

2.9 Resources

- chEMBL (https://www.ebi.ac.uk/chembl/)
- OpenBabel (https://openbabel.org/)
- PaDEL-Descriptor (http://www.yapcwsoft.com/dd/padeldescriptor/)
- Scikit-learn https://scikit-learn.org/stable/index.html
- RCSB-PDB https://www.rcsb.org/
- AutoDock Vina https://vina.scripps.edu/download.html
- PyMOL https://pymol.org/2/
- RDKit (https://www.rdkit.org/)

Chapter 3

Machine Learning in Chemoinformatics: A Deep Dive

3.1 Introduction to Machine Learning

Machine learning has emerged as a powerful tool in chemoinformatics, revolutionizing how chemical data is analyzed and interpreted. Machine learning in chemoinformatics, shedding light on its impact on drug discovery, chemical synthesis, and predictive modeling.

Integrating machine learning algorithms with chemical databases has enabled rapid screening of chemical compounds, identification of potential drug candidates, and optimization of molecular structures. Furthermore, utilizing neural networks, SVM-Support Vector Machines, and deep learning techniques has facilitated the development of predictive models for pharmacokinetics, toxicity prediction, and bioactivity profiling.

This study delves into the underlying principles of machine learning in chemoinformatics, highlighting the challenges and opportunities in this interdisciplinary field. By examining the latest advancements and prospects, machine learning plays a pivotal role in shaping the landscape of chemoinformatics. The comprehensive study discusses the wide-ranging applications of machine learning in chemoinformatics, emphasizing its impact on drug discovery, chemical synthesis, and predictive modeling of chemical properties and behaviors [13]. Integrating machine learning algorithms with chemical databases has not only accelerated the screening process of chemical compounds but has also facilitated the identification of potential drug candidates and the optimization of molecular structures. In addition, incorporating neural networks, SVM-Support Vector Machines, and deep learning techniques has significantly contributed to developing predictive models for pharmacokinetics, toxicity prediction, and bioactivity profiling.

Moreover, the review delves into the fundamental principles of machine learning in chemoinformatics, emphasizing the inherent challenges and the promising opportunities in this interdisciplinary domain. By elucidating the latest advancements and prospects, it explores how machine learning can transform the field by enabling more efficient and accurate analysis, interpretation, and utilization of chemical data, revolutionizing critical processes in drug discovery, chemical synthesis, and molecular design.

3.2 SVMs:

Support Vector Machine in virtual screening for drug discovery. SVMs have gained attention in the pharmaceutical industry due to their ability to handle high-dimensional data and classify compounds based on their potential for drug-likeness and biological activity. This study will investigate how SVMs can be used to predict the activity of small molecules against specific drug targets, ultimately aiding in identifying new lead compounds for drug development [25]. Furthermore, we will analyze the advantages and limitations of using SVMs in virtual screening compared to other machine learning algorithms commonly employed in this field. Support Vector Machines have shown great promise in virtual screening for drug discovery due to their ability to handle high-dimensional data and effectively classify compounds based on their potential for drug-likeness and biological activity. This study will involve training SVM models on a diverse set of small molecules and using them to predict the activity of these molecules against specific drug targets. By doing so, this research aims to aid in identifying new lead compounds for drug development.

One of the advantages of using SVMs in virtual screening is their ability to handle non-linear data and complex relationships between molecular features and bioactivity. Additionally, SVMs can effectively handle high-dimensional data often encountered in chemical informatics. However, it is essential to carefully consider the selection of kernel functions and regularization parameters to optimize the performance of SVM models in virtual screening [15]. Comparative analysis with other machine learning algorithms commonly employed in virtual

screening, such as random forests and neural networks, will further highlight the strengths and limitations of SVMs in this context. SVM is a powerful tool for predicting the biological activity of small molecules and enhancing the drug discovery process.

3.3 Advanced Techniques in Machine Learning: Hyperparameter Optimization and Model Training

In the context of qualitative modeling, the dataset was stratified into two classes: active (defined by" IC50" ≤ 100 nM) and inactive (defined by" IC50"> 100 nM), with a predetermined experimental cutoff for activity levels. Machine learning models trained on the training set achieved r squared value of MAOB 69.0 and LRRK2 60.0. The performance metrics on the validation set guided the selection of Support Vector Regressor (SVR) hyperparameters.

For quantitative modeling, SVR demonstrated superior performance compared to other approaches.The models exhibited reasonable prediction errors despite incorporating experimental data from diverse sources and locations. The research underscores the efficacy of 1D, 2D, and 3D-based SVR methods over purely 3D-based techniques, demonstrating that robust models can be developed using conventional descriptors and machine learning methodologies.

In summary, quantitative modeling of the dataset yielded crucial prediction accuracy and holds promise for successful implementation in lead optimization scenarios. The study's observations from both qualitative and quantitative approaches highlight the potential impact of dataset size on modeling outcomes, suggesting that reasonable and adequate models can be developed even with relatively diverse and limited datasets.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to assess the drug-likeness of bioactive compounds using Lipinski's descriptors, which Pfizer scientist Chris Lipinski proposed initially. These descriptors serve as rules-of-thumb to evaluate potential drugs' pharmacokinetic properties ADMET. Lipinski's rule of five suggests that orally active drugs typically adhere to specific criteria, including-

- Molecular weight (MW) less than 500 Dalton,
- LogP (octanol-water partition coefficient) less than 5,
- Hydrogen bond donors (NumHDonors) less than 5, and
- Hydrogen-bond acceptors (NumHAcceptors) less than 10

The examination of Lipinski's rule of five descriptors facilitated the exploration of the chemical space of inhibitors, providing insights into the structure-activity relationship. This analysis is instrumental in elucidating the fundamental characteristics of substances that influence their inhibitory properties. Additionally, EDA were performed using Lipinski's descriptors. Frequency plots depicting the distribution of active and inactive classes and scatter charts comparing MW with LogP were generated. Notably, these visualizations revealed that both bioactivity classes occupy similar chemical regions, underscoring the importance of further investigation into the structure-activity relationship of the compounds.

One of the primary objectives of EDA in drug discovery is to assess the drug likeness of compounds, as exemplified by Lipinski descriptors. These descriptors, encapsulating fundamental pharmacokinetic properties, offer a framework for evaluating the potential of a molecule to become an orally active drug. Through EDA, researchers can scrutinize how various chemical attributes, such as molecular weight, lipophilicity, and hydrogen bonding capacity, influence a compound's bioactivity and pharmacokinetic profile [18].

Moreover, EDA facilitates the exploration of structure-activity relationships (SAR), a cornerstone of drug development. Researchers can uncover essential features that govern a compound's efficacy, potency, and selectivity by analyzing the relationship between molecular structures and biological activities. Through techniques like

clustering, principal component analysis (PCA), and correlation analysis, EDA enables the identification of chemical motifs associated with desired pharmacological effects, paving the way for rational drug design.

In the context of inhibitor discovery, EDA holds particular significance, as it sheds light on the chemical space inhabited by active and inactive compounds. By visualizing the distribution of molecular properties and bioactivity classes, researchers can discern similarities and differences, guiding the design of new inhibitors with enhanced potency and selectivity [2]. Furthermore, EDA empowers scientists to prioritize compounds for further experimentation, focusing resources on those most likely to succeed in preclinical and clinical studies.

In conclusion, EDA is a cornerstone of modern drug discovery, empowering researchers to navigate the complexities of chemical space and unlock the potential of novel therapeutics. By harnessing the power of data analysis and visualization, scientists can expedite identifying and optimizing drug candidates, ultimately advancing the frontiers of medicine and improving patient outcomes.

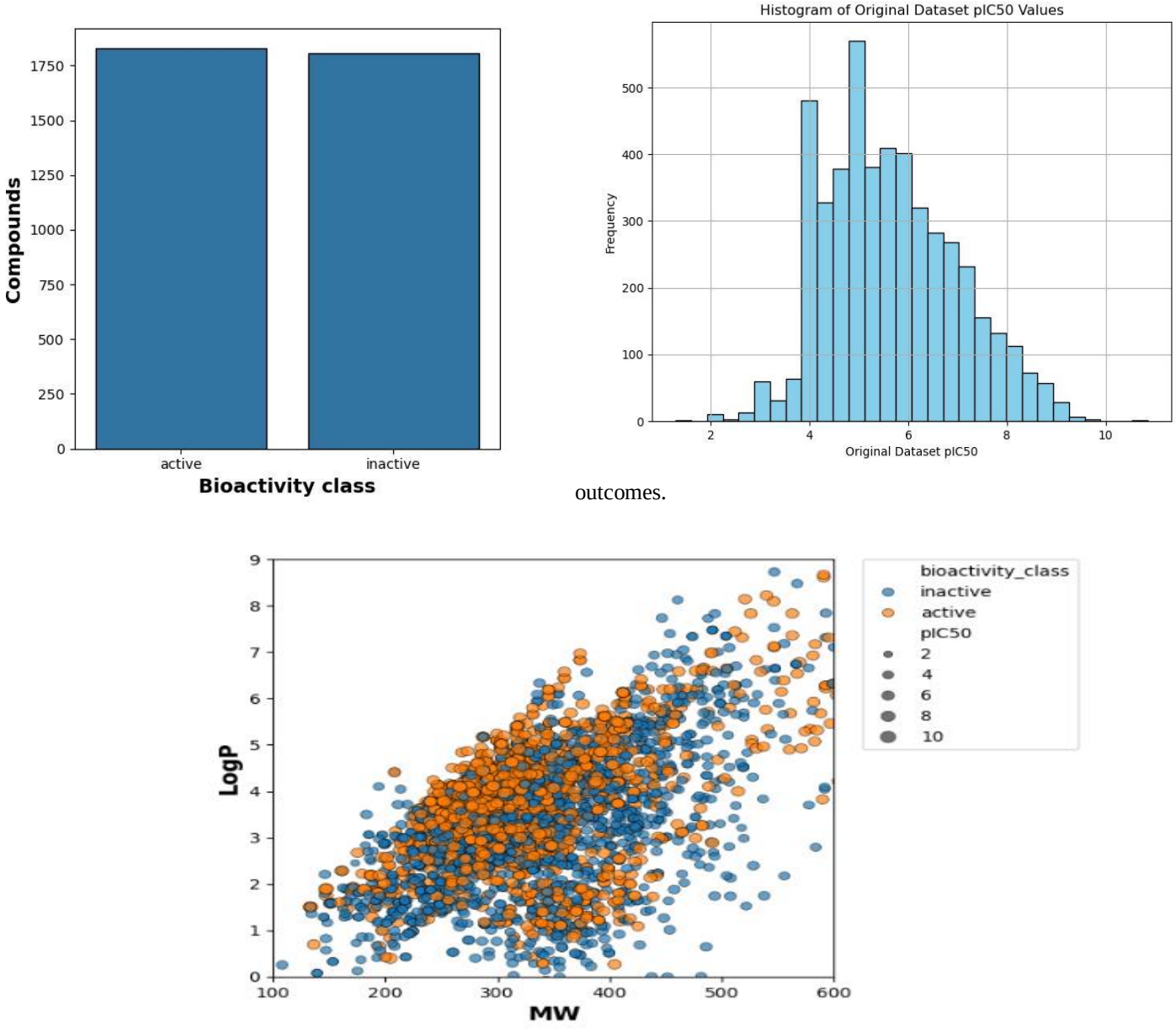


Fig 3.4.1 Compilation of MAO-B Dataset from chEMBL Source

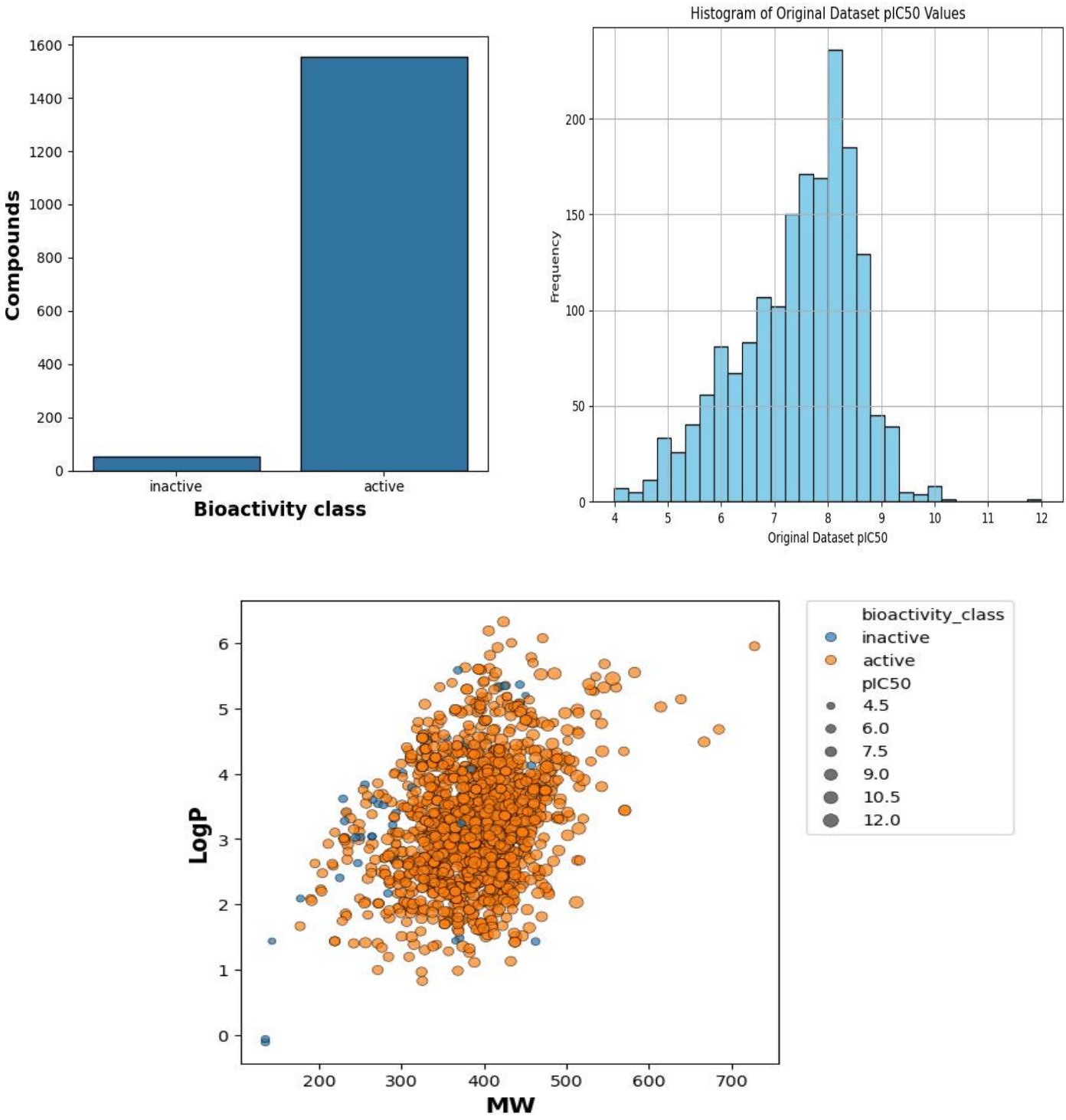


Fig 3.4.2 Compilation of LRRK2 Dataset from chEMBL Source

3.5 Support Vector Regression (SVR) and Extra Trees Regression (ETR) Models with Feature Selection and Evaluation

This code implements a robust pipeline for building and evaluating both Support Vector Regression (SVR) and Extra Trees Regression (ETR) models using Python's Scikit-learn library. The process begins with meticulous data preprocessing techniques to ensure data integrity and model reliability. Missing values in the dataset are handled by replacing them with column medians, ensuring that the dataset remains complete for further analysis. The dataset is then divided into training and testing sets to assess the models' performance on unseen data, employing the `train-test-split` function from Scikit-learn. Imputation is performed using the median strategy to address any remaining missing values, maintaining the dataset's consistency and integrity.

Feature selection is a critical step in model development, aiming to identify the most informative features for predicting the target variable. In this code, mutual information regression is employed to select the most relevant features, with a minimum of 1500 features chosen or the total number of features available, whichever is smaller, for SVR. Similarly, for ETR, the top 1200 features are selected based on their relevance to the target variable. Following feature selection, the selected features are standardized using StandardScaler to ensure uniformity in feature scales, facilitating better model convergence and performance.

Both SVR and ETR models are then initialized with specific hyperparameters and trained on the selected and scaled features of the training dataset. Subsequently, the model's performance is evaluated using various metrics, including-

- R-squared (R2),
- Mean Squared Error (MSE),
- Mean Absolute Error (MAE),
- Root Mean Squared Error (RMSE), and
- Adjusted R-squared (Adj R2).
  Based on the selected features and evaluation techniques, these metrics provide valuable insights into the models' predictive capabilities, accuracy, and goodness of fit.

The SVR model achieves an R-squared value of 0.686, indicating its ability to explain a substantial portion of the variance in the target variable. In contrast, for ETR, the R2 value, calculated to be 0.60, signifies the model's ability to explain the variance in the target variable. A higher R-squared value suggests better predictive performance.

3.5.1 Data Preprocessing:

- Import NumPy for numerical computations and Scikit-learn's modules for preprocessing and modeling.
- It calculates the median values for each column in the dataset, including NaN (missing) values.
- Then, it replaces the dataset's infinite values and NaN values with the respective column medians.

3.5.2 Data Splitting:

- After preprocessing, the dataset is split into training and testing sets using the train_test_split(80:20) function from Scikit-learn. This ensures that the model's performance can be evaluated on unseen data.

3.5.3 Missing Value Imputation:

- It handles any remaining missing values by imputing them with the median of their respective columns using Scikit-learn's SimpleImputer class.

3.5.4 Feature Selection:

- Feature selection is performed using mutual information regression to evaluate the dependency between variables. It selects the most informative features for predicting the target variable.
- The SelectKBest method selects the top features based on their scores. Here, it selects a minimum of 1500 or the total number of features, whichever is smaller.
- It then filters the training and testing data with the selected features.

3.5.5 Feature Scaling:

- Feature scaling is essential for many machine learning algorithms. Here, the features are standardized using StandardScaler mean of 0 and standard deviation of 1, which helps in better convergence and performance of the model.

3.5.6 Model Training:

- Support Vector Regression (SVR) is chosen as the modeling technique. SVR is a Support Vector Machine (SVM) used for regression tasks. It finds the hyperplane that best fits the data while minimizing the error.
- The SVR model is initialized with specific hyperparameters (C, gamma, kernel) and then fitted to the training data.

3.5.7 Model Evaluation:

After training, the model predicts the target variable on the testing data. Several evaluation metrics are calculated:

- R-squared (R2) represents the percentage of variation in the dependent variable that can be explained by the independent variables.
- Mean Squared Error (MSE) assesses the average of the errors squared between the predicted and actual values.
- Mean Absolute Error (MAE) measures the average of the absolute differences between predicted and actual values.
- Root Mean Squared Error (RMSE): Represents the square root of the MSE, providing a measure of the spread of the residuals.
- Adjusted R-squared (Adj R2): Adjusts the R2 value based on the number of predictors in the model, providing a more accurate assessment of the model's goodness of fit.

Table 3.5.1 The ExtraTreeRegressor(ETR) model prediction displays the predicted pIC50 values for LRRK2, based on the top 1200 features

Regressor Models (Trained on top 1200 features)	**R^2 Value**	**RMSE**	**MSE**	**MAE**
ETR	0.60	0.65	0.42	0.46
XGBoost Regressor	0.56	0.67	0.45	0.48

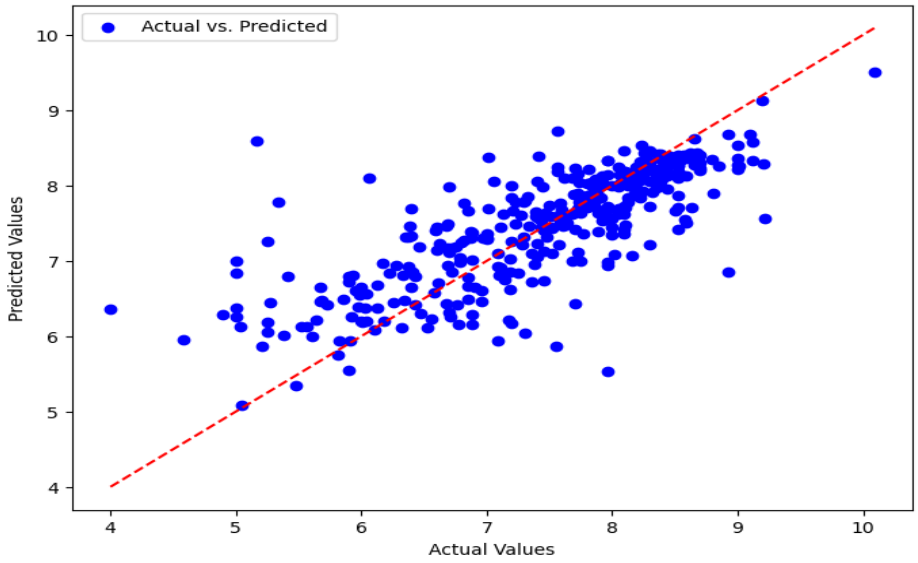


Fig 3.5.1 Predictive Performance Assessment: R-squared Analysis of MAO-B pIC50 Prediction Model

Table 3.5.2 The Suport Vector Regressor(SVR) model prediction displays the predicted pIC50 values for MAOB, based on the top 1500 features

Regressor Models(Trained on top 1500 features)	**R^2 Value**	**RMSE**	**MSE**	**MAE**
SVR	0.69	0.78	0.61	0.57
ETR	0.64	0.84	0.71	0.62
Gradient Boosting Regressor	0.62	0.86	0.74	0.64
XGBoost Regressor	0.62	0.86	0.75	0.63

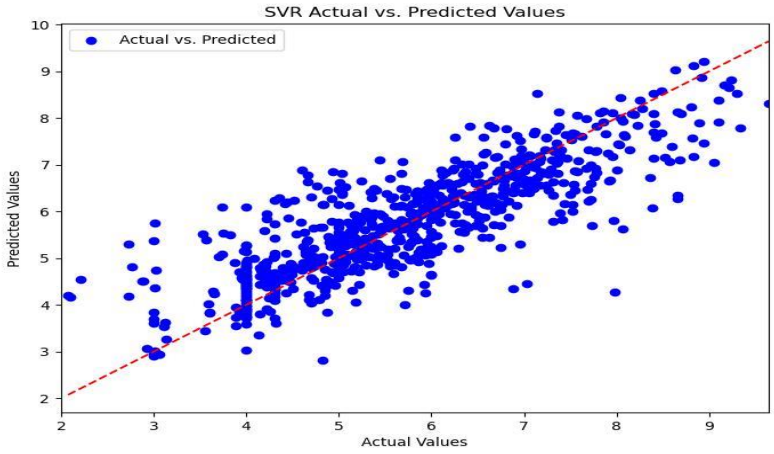


Fig 3.5.2 Predictive Performance Assessment: R-squared Analysis of LRRK2 pIC50 Prediction Model

Chapter 4

Virtual Screening for Compound Database Assembly and Screening Techniques

4.1 Introduction to Virtual Screening

Virtual screening and compound database assembly, focusing on the techniques and methods used in the process. The virtual screening process involves computational methods to prioritize and select potential drug compounds for further testing. This work will explore the different computational approaches, such as molecular docking, pharmacophore modeling, and quantitative structure-activity relationship analysis. Additionally, it will delve into the challenges and limitations of virtual screening and the potential applications and advancements in this field. Virtual screening has become an essential tool in drug discovery, offering a cost-effective and time-efficient method for identifying potential drug candidates [9]. One essential technique in virtual screening is molecular docking, which involves the computational analysis of the interaction between small molecules and a target protein. By simulating the binding of many small molecules to the target protein, molecular docking can help identify promising compounds for further study.

Pharmacophore modeling is another critical approach in virtual screening, focusing on identifying key molecular features essential for interacting with a biological target. This method allows for the rational design and selection of compounds based on their ability to match the specific pharmacophore features required for biological activity. Quantitative structure-activity relationship analysis is also commonly used in virtual screening to predict the biological activity of compounds based on their chemical structure. By establishing correlations between the chemical structure of compounds and their biological activity [11], QSAR models can be utilized to prioritize compounds for experimental testing.

Despite the advantages of virtual screening, it is essential to acknowledge the challenges and limitations of this approach. One of the main challenges is the accuracy of the computational models used in virtual screening and the need for experimental validation of the identified hits. Additionally, the availability and quality of the compound databases used for virtual screening can significantly impact the success of the process.

In recent years, advancements in computational power and algorithms have contributed to improving virtual screening techniques, allowing for the screening of more extensive compound libraries and the exploration of diverse chemical space. Furthermore, the integration of machine learning and artificial intelligence has shown promise in enhancing the predictive capabilities of virtual screening models.

As the field of virtual screening continues to evolve, it holds tremendous potential for accelerating the drug discovery process and identifying novel therapeutics for various diseases. By combining computational approaches with experimental validation, virtual screening can be crucial in expanding the scope of drug discovery research [20].

4.2 ADMET Analysis

ADMET analysis extensively examines bioactive molecules within the active class across various essential categories, including physicochemical properties, lipophilicity, water solubility, pharmacokinetics, drug-likeness, and medicinal chemistry. A comprehensive assessment encompassed 37 properties related to absorption, distribution, metabolism, excretion, and toxicity. Physicochemical properties such as molecular weight (MW), topological polar surface area (TPSA), number of hydrogen bond acceptors (NHA), and number of hydrogen bond donors (NHD) were scrutinized, with a focus on polarity as indicated by NHA and NHD [4]. Lipophilicity was evaluated using consensus log Po/w, while water solubility was categorized into distinct classes. Pharmacokinetics properties were assessed based on gastrointestinal (GI) absorption, blood-brain barrier (BBB) permeability, and skin

permeation (log Kp). Drug-likeness criteria were determined through compliance with established rules, while medicinal chemistry properties were evaluated based on synthetic accessibility scores. Notably, adherence to drug-like chemical space limits, particularly lipophilicity, molecular size, and polarity, played a pivotal role in assessing the potential of candidate compounds [10]. Compounds that met ADMET criteria were considered potential drug candidates, warranting further validation through molecular docking studies.

4.3 Molecular Docking

Visualizing interacting molecules at the molecular level provides crucial insights into the behavior and properties of these compounds. Understanding molecules' spatial arrangement and interactions is essential for various fields such as drug design, material science, and biochemistry. Utilizing advanced molecular visualization techniques enables researchers to observe the intricate details of molecular interactions, helping to elucidate the mechanisms underlying chemical reactions and biological processes [19].

One of the essential techniques for visualizing interacting molecules is molecular modeling, which involves using computer software to create three-dimensional representations of molecular structures. These models allow scientists to analyze the spatial relationships between individual atoms and molecules and the specific bonds and interactions between them. By simulating molecular interactions, researchers can gain valuable insights into the behavior of complex systems and predict the outcome of chemical and biological processes.

In addition to molecular modeling, X-ray crystallography and nuclear magnetic resonance spectroscopy play a vital role in visualizing molecule interactions. These experimental methods provide detailed information about the three-dimensional arrangement of atoms within a molecule and the spatial orientation of interacting molecules in a complex [3]. By combining experimental data with computational modeling, researchers can create comprehensive visualizations of molecular interactions, shedding light on fundamental principles in chemistry and biology.

The visualization of interacting molecules not only aids in understanding the structure and function of biomolecules but also facilitates the design of new drugs, catalysts, and materials with tailored properties. As we delve deeper into molecular visualization, the integration of experimental and computational approaches continues to push the boundaries of our understanding, unlocking the potential for groundbreaking discoveries and innovations. ## Advanced Techniques in Molecular Visualization

In addition to the fundamental techniques mentioned above, advancements in molecular visualization have led to the development of more sophisticated methods for studying interacting molecules. One notable approach is molecular dynamics simulation, which employs computational algorithms to model the movement and behavior of molecules over time [24]. This technique allows researchers to observe the dynamic nature of molecular interactions, providing insights into how molecules change their conformations and interact under varying conditions.

Furthermore, the emergence of virtual and augmented reality technologies has revolutionized how scientists visualize and interact with molecular structures. By immersing themselves in a virtual environment, researchers can manipulate and explore complex molecular models in unprecedented detail, gaining a deeper understanding of the spatial relationships and interactions between molecules.

Another advanced technique that has significantly enhanced molecular visualization is single-molecule imaging. This cutting-edge method enables the direct observation of individual molecules in real-time, allowing researchers to study their behavior and interactions at the single-molecule level [7]. By capturing high-resolution images of these dynamic processes, scientists can unravel the intricacies of molecular interactions with exceptional precision.

Integrating these advanced techniques with traditional methods has opened new frontiers in molecular visualization, providing a more comprehensive and dynamic understanding of interacting molecules. As researchers continue to innovate and refine these approaches, the potential for breakthrough discoveries and transformative applications in drug design, materials science, and beyond becomes increasingly promising.

4.4 Tanimoto Similarity

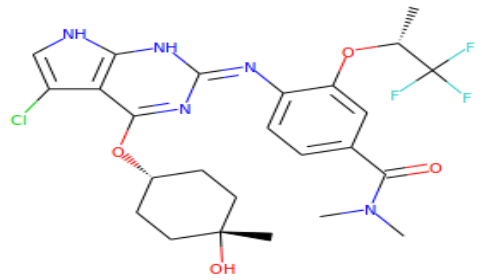
Tanimoto similarity is a crucial metric utilized extensively in drug discovery and cheminformatics to assess the similarity between chemical compounds. This similarity measure is pivotal in identifying potential drug candidates by comparing their structural features. The Tanimoto coefficient quantifies the degree of structural resemblance between two molecules based on their chemical fingerprints or descriptors. It is particularly significant in virtual screening, where large chemical databases are explored to pinpoint molecules with similar properties to known active compounds [16]. This approach aids in predicting potential drug targets and understanding structure-activity relationships (SAR) critical for drug design.

In drug discovery, the Tanimoto similarity coefficient serves several essential purposes. Firstly, it enables the efficient screening of vast chemical databases by rapidly identifying compounds with structural characteristics akin to known bioactive molecules. This computational approach accelerates the lead discovery process, narrowing potential candidates for further experimental validation. Secondly, Tanimoto similarity assists in exploring chemical space and identifying novel scaffolds or pharmacophores that exhibit biological activities similar to existing drugs. This method aids in uncovering new avenues for drug development and optimization. Additionally, Tanimoto similarity facilitates the elucidation of SAR by highlighting structural features crucial for a compound's biological activity [25]. This understanding is pivotal in guiding medicinal chemists toward designing analogs with improved potency, selectivity, or other desired properties.

The Tanimoto similarity coefficient is particularly valuable due to its ability to handle large-scale chemical data efficiently while providing meaningful insights into molecular relationships. Its application extends beyond virtual screening to encompass diverse areas such as compound clustering, diversity analysis, and lead optimization. By leveraging Tanimoto similarity, researchers can navigate the vast chemical space more intelligently, enhancing drug discovery campaigns' efficiency and success rate [25]. This metric thus stands as a cornerstone in cheminformatics, facilitating the translation of chemical knowledge into tangible therapeutic outcomes.

1) LRRK2

Reference Compound based on highest pIC50 value

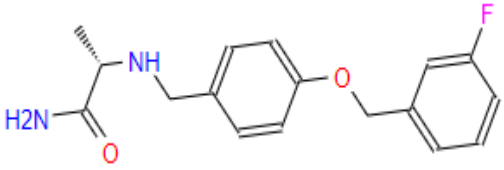


pIC50 value= 12.0

Binding Affinity= -8.9 (kcal/mol)

2) MAOB

Reference Compound Safinamide FDA approved Drug



Binding Affinity= -10.2 (kcal/mol)

4.5 Docking with their respective Binding Affinities

Table 4.5.1 LRRK2 DrugBank Docking Result

Smiles	Compound Name	Binding Affinity
O=C1NC(=O)c2c1c1c3ccccc3n3c1c1c2c2ccccc2n1[C@H]1CC[C@@H]3O1	Rel-(9R,12S)-9,10,11,12-tetrahydro-9,12-epoxy-1H-diindolo[1,2,3-FG:3',2',1'-KL]pyrrolo[3,4-I][1,6]benzodiazocine-1,3(2H)-dione	-11.9
O=C(c1cc(Cn2c(=O)[nH]c(=O)c3c(F)cccc32)ccc1F)N1CCN(c2ncccn2)CC1	Senaparib	-11.6
CC(=N)N1CCC(Oc2ccc3c4ccccc4n(Cc4ccc5ccc(C(=N)N)cc5c4)c3c2)CC1	7-[[2-[[1-(1-Iminoethyl)piperidin-4-YL]oxy]-9H-carbozol-9-YL]methyl]naphthalene-2-carboximidamid	-12.4
NC(N)=N[C@@H](C(=O)NCC(=O)N1CCC(c2cc(-c3cccc(Cl)c3Cl)n[nH]2)CC1)C1CCCCC1	2-Cyclohexyl-N-(2-{4-[5-(2,3-dichloro-phenyl)-2H-pyrazol-3-YL]-piperidin-1-YL}-2-oxo-ethyl)-2-guanidino-acetamide	-11.8
CC(C)(CNC(=O)c1ccc(Oc2ccccc2)cc1)c1ccc2c(N)nc(N)nc2n1	N-[2-(2,4-Diaminopyrido[2,3-D]pyrimidin-7-Yl)-2-Methylpropyl]-4-Phenoxybenzamide	-11.0
O=C(O)[C@H]1C[C@@]12C[C@@H](c1ccc(OCc3cc(-c4ccccc4)nc4ccccc34)cc1)NC2=O	(1s,3r,6s)-4-Oxo-6-{4-[(2-Phenylquinolin-4-Yl)methoxy]phenyl}-5-Azaspiro[2.4]heptane-1-Carboxylic Acid	-11.9
Nc1nc(Nc2ccc3c(c2)CC[C@@H](N2CCCC2)CC3)nn1-c1cc2c(nn1)-c1ccccc1CCC2	Bemcentinib	-11.1
CNC(=O)c1ccc(-c2cnc3ncc(Cc4ccc5ncccc5c4)n3n2)cc1F	Capmatinib	-11.5
O=C1CCc2c(Oc3ccc4c(c3)[C@@H]3[C@H](O4)[C@H]3c3nc4ccc(C(F)(F)F)cc4[nH]3)ccnc2N1	Beigene-283	-11.3
O=C(CN1C(=O)C2(CCCC2)NC[C@H]1c1cc(F)cc(F)c1)Nc1ccc2c(c1)C[C@@]1(C2)C(=O)Nc2ncccc21	MK-3207	-12.4
Cc1cc(C[C@@H](NC(=O)N2CCC(c3cc4ccccc4[nH]c3=O)CC2)C(=O)N2CCN(C3CCN(C)CC3)CC2)cc2cn[nH]c12	Zavegepant	-12.4
CN1C(=O)C(F)=C[C@]2(C)[C@H]3CC[C@]4(C)[C@@H](C(=O)NCc5nc6ncccc6[nH]5)CC[C@H]4[C@@H]3CC[C@@H]12	MK-0773	-11.6

Table 4.5.2 LRRK2 PubChem Docking Results

Smiles	Compound Name	Binding Affinity
C[C@]1(O)C[C@H](c2nc(-c3ccc4ccc(-c5ccccc5)nc4c3)c3c(N)nccn32)C1	Linsitinib	-11.6
Clc1ccc2c(N3CCC(CCC4CCN(c5ncnc6cc(Cl)ccc56)CC4)CC3)ncnc2c1	1,2-Bis(1-(7-chloroquinazolin-4-yl)piperidin-4-yl)ethane	-11.6
CC1=Nc2c(F)cc(-c3nc(Nc4ccc(C5CCN(C)CC5)cn4)ncc3F)cc2C12CCCC2	Crozbaciclib	-11.4
Cc1nc(N2CCN(C)CC2)cc(N(C)c2cccc(NC(=O)Nc3ccc(Cl)c(C(F)(F)F)c3)c2)n1	1-[4-Chloro-3-(trifluoromethyl)phenyl]-3-[3-[methyl-[2-methyl-6-(4-methylpiperazin-1-yl)pyrimidin-4-yl]amino]phenyl]urea	-11.5
Cc1nn(C)cc1-c1cc(C(=O)Nc2c(Cl)ccc(NS(=O)(=O)c3ccccc3F)c2F)c2ncnc(N)c2c1	4-amino-N-(6-chloro-2-fluoro-3-{[(2-fluorophenyl)sulfonyl]amino}phenyl)-6-(1,3-dimethyl-1H-pyrazol-4-yl)quinazoline-8-carboxamide	-11.5
FC(F)(F)c1cccc(Nc2nccc(-c3c(-c4ccc(Cl)cc4)nn4ncccc34)n2)c1	Pyrazolo[1,5-b]pyridazine deriv. 83	-11.0
CC(=O)N1CCN(c2ccc(Nc3ncc(Cl)c(Nc4ccc(C(F)(F)F)c4)n3)nc2)CC1	1-[4-[6-[[5-Chloro-4-[3-(trifluoromethyl)anilino]pyrimidin-2-yl]amino]pyridin-3-yl]piperazin-1-yl]ethanone	-11.7
Cc1cccc(Cl)c1NC(=O)c1cnc(Nc2ccc(N3CCN(C)CC3)cc2)nc1Nc1ccccc1NC(=O)CCl	4-[2-[(2-chloroacetyl)amino]anilino]-N-(2-chloro-6-methylphenyl)-2-[4-(4-methylpiperazin-1-yl)anilino]pyrimidine-5-carboxamide	-11.3
CN1CCN(c2ccc(Nc3ncc(Cl)c(Nc4ccc(C(F)(F)F)cc4)n3)nc2)CC1	5-chloro-2-N-[5-(4-methylpiperazin-1-yl)pyridin-2-yl]-4-N-[4-(trifluoromethyl)phenyl]pyrimidine-2,4-diamine	-11.8
Cc1nc(Nc2cccc(NC(=O)Nc3ccc(Cl)c(C(F)(F)F)c3)c2)cc(N2CCN(C)CC2)n1	1-[4-Chloro-3-(trifluoromethyl)phenyl]-3-[3-[[2-methyl-6-(4-methylpiperazin-1-yl)pyrimidin-4-yl]amino]phenyl]urea	-12.1
CC1(NC(=O)NC2CCN(c3ncnc4c3nc(-c3ccccc3Cl)n4-c3ccc(Cl)cc3)CC2)CCC(F)(F)CC1	1-[1-[8-(2-Chlorophenyl)-9-(4-chlorophenyl)purin-6-yl]piperidin-4-yl]-3-(4,4-difluoro-1-methylcyclohexyl)urea	-11.7
CC(C)n1cnnc1-c1cccc(NC(=O)Nc2ccc3ccccc3n2)n1	1-[6-(4-Propan-2-yl-1,2,4-triazol-3-yl)pyridin-2-yl]-3-quinolin-2-ylurea	-10.1

Table 4.5.3 MAOB DrugBank Docking Results

Smiles	Compound Name	Binding Affinity
CC(C)(CNC(=O)c1ccc(Oc2ccccc2)cc1)c1ccc2c(N)nc(N)nc2n1	N-[2-(2,4-Diaminopyrido[2,3-D]pyrimidin-7-Yl)-2-Methylpropyl]-4-Phenoxybenzamide	-11.8
CC(=N)N1CCC(Oc2ccc3c4ccccc4n(Cc4ccc5ccc(C(=N[illegible]5c4)c3c2)CC1	7-[[2-[[1-(1-Iminoethyl)piperidin-4-YL]oxy]-9H-carbozol-9-YL]methyl]naphthalene-2-carboximidamid	-10.9
Cc1cc(Nc2ccc3cc4ccccc4cc3c2)n2ncnc2n1	N-anthracen-2-yl-5-methyl[1,2,4]triazolo[1,5-a]pyrimidin-7-amine	-12.1
O=C1NC(=O)c2c1c1c3ccccc3n3c1c1c2c2ccccc2n1[C@H]1CC[C@@H]3O1	Rel-(9R,12S)-9,10,11,12-tetrahydro-9,12-epoxy-1H-diindolo[1,2,3-FG:3',2',1'-KL]pyrrolo[3,4-I][1,6]benzodiazocine-1,3(2H)-dione	-11.4
Nc1nc(Nc2ccc3c(c2)CC[C@@H](N2CCCC2)CC3)nn1-c1cc2c(nn1)-c1ccccc1CCC2	Bemcentinib	-14.2
Cc1cc2ncc(C(=O)NCC(C)(C)NCC(=O)N3CCC[C@H]3C#N)cn2n1	Anagliptin	-11.2
FC(F)(F)c1ccc(/C=C/c2nc(COc3ccc(CCCCn4ccnn4)cc3)co2)cc1	Mubritinib	-11.7
O=C(O)C[C@H]1CC[C@@H](c2ccc(-c3ccc(Nc4ccc(C(F)(F)F)nc4)cn3)cc2)CC1	Pradigastat	-12.2
CC(C)(O)c1ccccc1-c1ccc2nc(/C=C/c3ccc(C(F)(F)F)cc3)[nH]c2c1	Mavatrep	-11.7
O=C1CCc2c(Oc3ccc4c(c3)[C@@H]3[C@H](O4)[C@H]3c3nc4ccc(C(F)(F)F)cc4[nH]3)ccnc2N1	Beigene-283	-12.3
O=C(Cc1cccc(OC(F)(F)F)c1)Nc1ccc(CCCCc2nnc(NC(=O)Cc3ccccn3)s2)nn1	Telaglenastat	-12.7
Cc1cc(C[C@@H](NC(=O)N2CCC(c3cc4ccccc4[nH]c3=O)CC2)C(=O)N2CCN(C3CCN(C)CC3)CC2)cc2cn[nH]c12	Zavegepant	-11.7

Table 4.5.4 MAOB PubChem Docking Results

Smiles	Compound Name	Binding Affinity
CN1CCc2ccc(OCCCCN3CCC(c4noc5cc(F)ccc45)CC3)cc2C1=C	7-[4-[4-(6-Fluoro-1,2-benzoxazol-3-yl)piperidin-1-yl]butoxy]-2-methyl-3,4-dihydroisoquinolin-1-one	-12.0
O=C(/C=C/c1ccccc1)c1ccc(OCc2cc(-c3ccc([N+](=O)[O-])cc3)no2)cc1	(E)-1-[4-[[3-(4-nitrophenyl)-1,2-oxazol-5-yl]methoxy]phenyl]-3-phenylprop-2-en-1-one	-12.0
CC(C)c1ccc(-c2ccc(Oc3cc(C(=O)N[C@@H]4CCN(C5CCOCC5)C[C@@H]4F)ccc3CN3C[C@@H](C)CC3=O)cc2)nn1	N-[(3S,4R)-3-fluoro-1-(oxan-4-yl)piperidin-4-yl]-4-[[(4S)-4-methyl-2-oxopyrrolidin-1-yl]methyl]-3-[4-(6-propan-2-ylpyridazin-3-yl)phenoxy]benzamide	-11.6
COc1ccc(-c2cccc(C(=O)N3CCC(NC(=O)c4ccc(C(F)(F)F)cc4)CC3)c2)nn1	N-[1-[3-(6-methoxypyridazin-3-yl)benzoyl]piperidin-4-yl]-4-(trifluoromethyl)benzamide	-11.9
O=C(/C=C/c1ccccc1)c1ccc(OCc2cc(-c3ccc4c(c3)OCCO4)no2)cc1	(E)-1-[4-[[3-(2,3-dihydro-1,4-benzodioxin-6-yl)-1,2-oxazol-5-yl]methoxy]phenyl]-3-phenylprop-2-en-1-one	-12.5
O=C1/C(=C/c2cc(-c3ccc4c(c3)OCO4)n[nH]2)Cc2ccccc21	(2E)-2-[[3-(1,3-benzodioxol-5-yl)-1H-pyrazol-5-yl]methylidene]-3H-inden-1-one	-12.0
O=C1c2ccc(OCCCC(F)(F)F)cc2-c2nnc(-c3cccc(C(F)(F)F)c3)cc21	8-(4,4,4-Trifluorobutoxy)-3-[3-(trifluoromethyl)phenyl]indeno[1,2-c]pyridazin-5-one	-11.1
O=C(c1ccc(F)c(F)c1)N1CCn2nc(COc3cccc(F)c3)cc2C1	(3,4-difluorophenyl)-[2-[(3-fluorophenoxy)methyl]-6,7-dihydro-4H-pyrazolo[1,5-a]pyrazin-5-yl]methanone	-11.5
O=C(CCCCCCc1ccccc1)c1ccc(-c2ccco2)nn1	1-(6-(Furan-2-yl)-pyridazin-3-yl)-7-phenylheptan-1-one	-10.5
O=C1N[C@@]2(CCc3cc(OCCCC(F)(F)F)ccc32)Cc2nn(CCC3CC3)c(-c3nn[nH]n3)c21	US11198695, Example II-84	-11.7
NC(=O)c1c(F)ccc(OCc2cc(-c3ccc(-c4cccc5ccccc45)cc3)no2)c1F	2,6-Difluoro-3-[[3-(4-naphthalen-1-ylphenyl)-1,2-oxazol-5-yl]methoxy]benzamide	-11.6
COc1cc2c(cc1OC)C(=O)[C@H](CC1CCN(Cc3cccc(Cl)n3)CC1)C2		-11.7

Table 4.5.5 MAOB Real-Enamine Docking Results

Smiles	Compound Name	Binding Affinity
C[C@H](N)C(=O)NCc1ccc(OCc2cccc(F)c2)cc1	(2S)-2-amino-N-[[4-[(3-fluorophenyl)methoxy]phenyl]methyl]propanamide	-9.8
C[C@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)O	(S)-2-((4-((3-Fluorobenzyl)oxy)benzyl)amino)propanoic acid	-10.0
C[C@H](O)[C@@H](C)NCc1ccc(OCc2cccc(F)c2)cc1		-9.7
C[C@H](O)[C@@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)O		-10.1
C[C@@H](NC(=O)[C@@H](C)NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)O		-10.6
C[C@@H](O)[C@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)O		-10.5
NC(=O)[C@@H](NCc1ccc(OCc2cccc(F)c2)cc1)c1ccc(F)cc1		-10.6
NC(=O)[C@H](Cc1ccccc1)NCc1ccc(OCc2cccc(F)c2)cc1		-10.6
C[C@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)NC(=O)NC(C)(C)C		-10.2
COc1ccc(NC(=O)[C@@H](C)NCc2ccc(OCc3cccc(F)c3)cc2)cc1		-10.4

Table 4.5.6 MAOB Zinc Docking Results

Smiles	Compound Name(Zinc ID)	Binding Affinity
CC(C)NC(=O)[C@H](C)NCc1ccc(OCc2cccc(F)c2)cc1	ZINC666467886	-9.7
CCOC(=O)[C@H](C)NCc1ccc(OCc2cccc(F)c2)cc1	ZINC734156037	-9.8
NC(=O)[C@H](Cc1ccc(F)cc1)NCc1ccc(OCc2cccc(F)c2)cc1	ZINC541304027	-9.6
CC[C@@H](C)NC(=O)[C@@H](C)NCc1ccc(OCc2cccc(F)c2)cc1	ZINC45811994	-9.5
C[C@@H](NCc1ccc(OCc2ccccc2)cc1)C(N)=O	ZINC53092456	-9.9
C[C@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)NC(C)(C)C	ZINC45811984	-10.2
C[C@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)Nc1ccccc1	ZINC642100971	-10.7
C[C@@H](NCc1ccc(OCc2cccc(F)c2)cc1)[C@@H](C)C(=O)O	ZINC901447295	-9.7
C[C@@H](NCc1ccc(Oc2cccc(F)c2)cc1)C(N)=O	ZINC683725095	-9.6
C[C@@H](NCc1ccc(OCc2cccc(F)c2)cc1)C(=O)NCc1ccccc1	ZINC642100983	-11.2

Chapter 5

Results and Discussion

5.1 Exploratory Data Analysis and Comparative Analysis of Bioactive Compounds: Results and Discussion

This section will delve deeper into the comparative analysis of bioactive compounds. Firstly, a detailed examination of each compound's chemical composition focuses on the presence and quantity of various bioactive compounds. This analysis will provide a clearer understanding of the potential health benefits associated with each compound.

Following the chemical analysis, a comparative exploration of the biological activities of these compounds will be undertaken. This will involve an in-depth investigation of the pharmacological properties, to assess the potential therapeutic value of the bioactive compounds.

Moreover, a comprehensive review of existing literature on the pharmacological effects of these bioactive compounds will be conducted to provide further insights into their potential applications and implications for human health. This in-depth examination will contribute to a more holistic understanding of the bioactive compounds and their potential pharmacological effects, ultimately contributing to the body of knowledge in natural product research and pharmaceutical sciences. After conducting the chemical analysis and comparative exploration of the biological activities of the bioactive compounds, the next step will involve presenting the findings visually engagingly. Data visualization techniques such as bar graphs, pie charts, and heat maps will enhance the accessibility and comprehension of the complex information gathered from the analysis.

Furthermore, a statistical analysis will be performed to determine any significant correlations between the presence of specific bioactive compounds and their observed pharmacological effects. This will provide quantitative evidence to support the qualitative findings and contribute to the robustness of the comparative analysis.

Subsequently, the implications of the findings for potential applications in the pharmaceutical industry, nutraceutical products, and traditional medicine will be discussed. This will involve critically evaluating the practical implications of the identified bioactive compounds and their pharmacological effects, laying the groundwork for future research and development.

The combined results and discussion will offer valuable insights into the bioactive compounds and provide a foundation for further exploration and exploitation of these natural resources to benefit human health and well-being.

5.2 ADMET Analysis

The discovery of novel drugs is a critical medical process. Recent advancements in drug discovery have focused on identifying potential drug candidates for oral bioavailability, metabolic stability, and pharmacokinetic profile.

Table 5.2.1 ADMET Analysis of MAO-B selected Compounds

Canonical SMILES	MW	BBB permeant	Leadlikeness #violations	GI absorption	Bioavailability Score
Cc1cc(Nc2ccc3c(c2)cc2c(c3)cccc2)n2c(n1)ncn2	325.37	Yes	1	High	0.55
O=c1[nH]c(=O)c2c1c1c3c4c2c2ccccc2n4[C@@H]2O[C@H](n3c3c1cccc3)CC2	393.39	Yes	2	High	0.55
Fc1ccc2c(c1)onc2C1CCN(CC1)CCCCOc1ccc2c(c1)C(=O)N(CC2)C	451.53	Yes	2	High	0.55
O=C1/C(=C/c2[nH]nc(c2)c2ccc3c(c2)OCO3)/Cc2c1cccc2	330.34	Yes	1	High	0.55
Fc1cccc(c1)OCc1cc2n(n1)CCN(C2)C(=O)c1ccc(c(c1)F)F	387.36	Yes	1	High	0.55
O=C(c1ccc(nn1)c1ccco1)CCCCCCc1ccccc1	334.41	Yes	2	High	0.55
COc1cc2c(cc1OC)C[C@H](C2=O)CC1CCN(CC1)Cc1cccc(n1)Cl	414.93	Yes	2	High	0.55
O=C([C@@H](N)C)NCc1ccc(cc1)OCc1cccc(c1)F	302.34	Yes	0	High	0.55
OC(=O)[C@@H](NCc1ccc(cc1)OCc1cccc(c1)F)C	303.33	Yes	0	High	0.55
C[C@@H]([C@H](NCc1ccc(cc1)OCc1cccc(c1)F)C)O	303.37	Yes	0	High	0.55
Fc1ccc(cc1)[C@@H](C(=O)N)NCc1ccc(cc1)OCc1cccc(c1)F	382.4	Yes	3	High	0.55
Fc1cccc(c1)COc1ccc(cc1)CN[C@H](C(=O)N)Cc1ccccc1	378.44	Yes	3	High	0.55
COc1ccc(cc1)NC(=O)[C@H](NCc1ccc(cc1)OCc1cccc(c1)F)C	408.47	Yes	3	High	0.55
CC(NC(=O)[C@@H](NCc1ccc(cc1)OCc1cccc(c1)F)C)C	344.42	Yes	1	High	0.55
CCOC(=O)[C@@H](NCc1ccc(cc1)OCc1cccc(c1)F)C	331.38	Yes	2	High	0.55
Fc1ccc(cc1)C[C@@H](C(=O)N)NCc1ccc(cc1)OCc1cccc(c1)F	396.43	Yes	3	High	0.55
CC[C@H](NC(=O)[C@H](NCc1ccc(cc1)OCc1cccc(c1)F)C)C	358.45	Yes	3	High	0.55
NC(=O)[C@H](NCc1ccc(cc1)OCc1ccccc1)C	284.35	Yes	0	High	0.55
O=C([C@@H](NCc1ccc(cc1)OCc1cccc(c1)F)C)NC(C)(C)C	358.45	Yes	3	High	0.55
Fc1cccc(c1)COc1ccc(cc1)CN[C@H](C(=O)Nc1ccccc1)C	378.44	Yes	3	High	0.55
Fc1cccc(c1)COc1ccc(cc1)CN[C@@H]([C@H](C(=O)O)C)C	331.38	Yes	1	High	0.55
Fc1cccc(c1)Oc1ccc(cc1)CN[C@@H](C(=O)N)C	288.32	Yes	0	High	0.55
Fc1cccc(c1)COc1ccc(cc1)CN[C@@H](C(=O)NCc1ccccc1)C	392.47	Yes	3	High	0.55

Table 5.2.2 ADMET Analysis of LRRK2 selected Compounds

Canonical SMILES	MW	BBB permeant	Leadlikeness #violations	GI absorption	Bioavailability Score
O=c1[nH]c(=O)c2c1c1c3c4c2c2ccccc2n4[C@@H]2O[C@H](n3c3c1cccc3)CC2	393.39	Yes	2	High	0.55

5.3 Drug Discovery for MAOB and LRRK2 Targets

MAOB and LRRK2 targets, which play crucial roles in neurodegenerative diseases such as Parkinson's disease. Identifying small molecules that can modulate these targets has become a significant area of interest in the pharmaceutical industry and academic research. Understanding the structure and function of MAOB and LRRK2 and their involvement in disease pathology is essential for developing effective therapeutics.

Several strategies have been employed in the search for potential drug candidates, including virtual screening, high-throughput screening, and structure-based drug design. These approaches have led to the identification of lead compounds with the potential to modulate the activity of MAOB and LRRK2. However, the challenge lies in optimizing these leads to enhance their potency, selectivity, and pharmacokinetic properties while minimizing off-target effects [14].

In addition to traditional small molecule approaches, there is growing interest in developing biologics such as monoclonal antibodies and gene therapies targeting MAOB and LRRK2. These approaches offer the potential for precise and potent modulation of these targets, opening up new avenues for drug discovery in neurodegenerative diseases.

The ongoing research in drug discovery for MAOB and LRRK2 targets holds great promise for developing innovative therapies for Parkinson's disease and other neurodegenerative disorders. By delving deeper into the molecular mechanisms underlying these targets and harnessing advanced drug discovery technologies, researchers are moving closer to addressing the unmet medical needs of patients suffering from these devastating conditions.

5.4 Integrated Computational Approach for Drug Candidate Selection Targeting MAOB and LRRK2

In the pursuit of identifying promising drug candidates targeting monoamine oxidase B (MAOB) and leucine-rich repeat kinase 2 (LRRK2), an integrated computational workflow was employed, combining molecular similarity screening, molecular docking, ADMET analysis, and filtering based on blood-brain barrier (BBB) permeability.

5.5 Molecular Similarity Screening and Compound Selection

The process commenced with the retrieval of compounds from diverse chemical libraries, including PubChem, ZINC, DrugBank, and real Enamine. Molecular similarity screening was conducted using Tanimoto similarity scores, focusing on identifying compounds with structural resemblance to known ligands or inhibitors of MAOB and LRRK2. Compounds exhibiting a Tanimoto similarity score ranging from 0.70 to 1.00 were retained for further analysis.

5.6 Molecular Docking Studies

The retained compounds underwent molecular docking simulations using AutoDock Vina to predict their binding affinities and interaction modes with the MAOB and LRRK2 proteins. This step enabled the identification of top-ranking compounds based on their binding energies and potential interactions with critical amino acid residues within the binding sites of the target proteins.

5.7 ADMET Analysis for Drug-Likeness Evaluation

Following molecular docking, the top 10 compounds from each chemical library for both MAOB and LRRK2 were subjected to ADMET analysis using SwissADME. This comprehensive analysis evaluated crucial drug-likeness properties based on established criteria, including Lipinski's rule of five (related to oral bioavailability), Ghose's rule (for drug-like molecular properties), Veber's rule (for pharmacokinetic properties), Egan's rule (for oral

bioavailability), and Muegge's rule (related to drug-like properties). Compounds were categorized as 'yes' or 'no' based on their adherence to at least 3 out of these 5 drug-likeness rules.

5.8 Blood-Brain Barrier (BBB) Permeability Filter

Furthermore, compounds that passed the drug-likeness assessment were screened for their ability to penetrate the blood-brain barrier (BBB). This step is crucial for central nervous system (CNS) drug development, particularly in targeting neurodegenerative diseases associated with MAOB and LRRK2.

5.9 Final Compound Selection

Finally, based on the results of ADMET analysis and BBB permeability filtering, the top 5 compounds from each library were selected as potential drug candidates for MAOB and LRRK2. These compounds demonstrated favorable binding affinities and exhibited desirable drug-like properties and potential CNS penetration, making them promising candidates for further preclinical and clinical investigations.

This integrated computational approach harnesses the power of molecular modeling, virtual screening, and pharmacoinformatics to streamline the drug discovery process, accelerating the identification of novel therapeutic agents for challenging biological targets like MAOB and LRRK2 implicated in neurodegenerative disorders.

5.10 Future Scope

1. Predicting unknown binding affinity and processing molecular dynamic simulations

2. Using protein sequences to implement these pipelines for MAO-B and LRRK2.

3. Wet lab validation.

Table 5.9.1 Top 5 Compounds ADMET Analysis

Canonical SMILES	MW	BBB permeant	Leadlikeness #violations	GI absorption	Bioavailability Score
COc1cc2c(cc1OC)C[C@H](C2=O)CC1CCN(CC1)Cc1cccc(n1)Cl	414.93	Yes	2	High	0.55
O=C1/C(=C/c2[nH]nc(c2)c2ccc3c(c2)OCO3)/Cc2c1cccc2	330.34	Yes	1	High	0.55
Fc1ccc2c(c1)onc2C1CCN(CC1)CCCCOc1ccc2c(c1)C(=O)N(CC2)C	451.53	Yes	2	High	0.55
O=c1[nH]c(=O)c2c1c1c3c4c2c2ccccc2n4[C@@H]2O[C@H](n3c3c1cccc3)CC2	393.39	Yes	2	High	0.55
Cc1cc(Nc2ccc3c(c2)cc2c(c3)cccc2)n2c(n1)ncn2	325.37	Yes	1	High	0.55

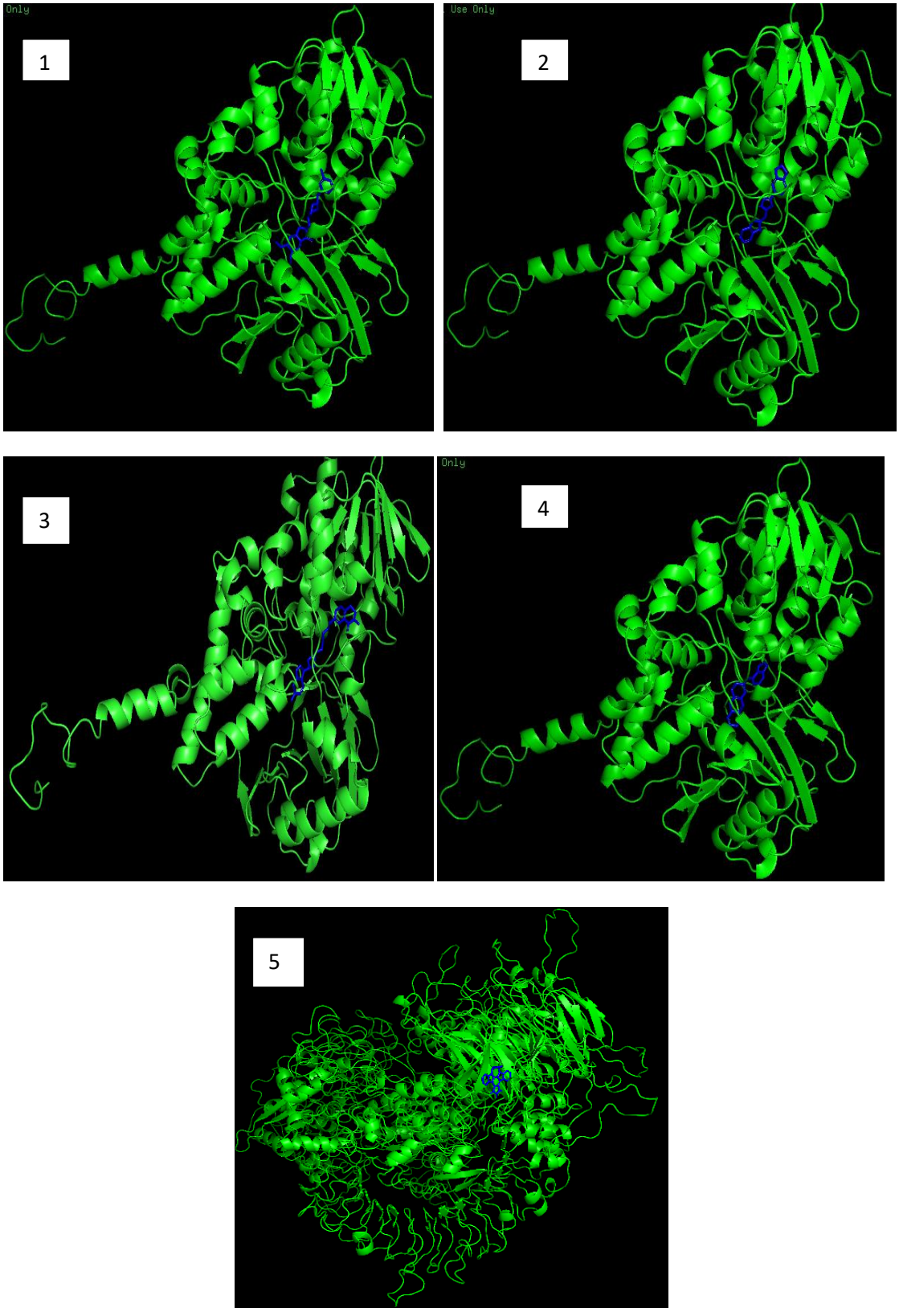


Fig 5.9.1 Top 5 Compounds and Their Binding Affinities in Docking Results 1) -11.7, 2) -12.0, 3) -12.0, 4) -12.1, 5) -11.9

References

1. A. Bohush, W. Leśniak, S. Weis and A. Filipek, *Int. J. Mol. Sci.*, 2021, **22**, 1–14.
2. B. Spencer, R. Potkar, M. Trejo, E. Rockenstein, C. Patrick, R. Gindi, A. Adame, T. Wyss-Coray and E. Masliah, *J. Neurosci.*, 2009, **29**, 13578–13588.
3. C. B. Lücking and A. Brice, *Cell. Mol. Life Sci.*, 2000, **57**, 1894–1908.
4. C. Mao, H. Wang, H. Luo, S. Zhang, H. Xu, S. Zhang, J. Rosenblum, Z. Wang, Q. Zhang, M. Tang, M. J. Shepard, X. Wang, Y. Wang, Z. Zhuang, C. Shi and Y. Xu, *Neurobiol. Aging*, 2019, **75**, 38–41.
5. D. O. Borroto-Escuela and K. Fuxe, *J. Neural Transm.*, 2019, **126**, 455–471.
6. E. Angelopoulou, Y. N. Paudel and C. Piperi, *Cell. Mol. Life Sci.*, 2021, **78**, 1445–1453.
7. E. E. Benarroch, *Neurology*, 2011, **76**, 660–667.
8. G. George, S. Singh, S. B. Lokappa and J. Varkey, *Genomics*, 2019, **111**, 819–830.
9. H. Deng, K. Gao and J. Jankovic, *Mov. Disord.*, 2013, **28**, 569–575.
10. K. Ebanks, P. A. Lewis and R. Bandopadhyay, *Front. Neurosci.*, 2020, **13**, 1–13.
11. K. M. Strauss, L. M. Martins, H. Plun-Favreau, F. P. Marx, S. Kautzmann, D. Berg, T. Gasser, Z. Wszolek, T. Müller, A. Bornemann, H. Wolburg, J. Downward, O. Riess, J. B. Schulz and R. Krüger, *Hum. Mol. Genet.*, 2005, **14**, 2099–2111.
12. K. Ogawa, T. Yamada, Y. Tsujioka, J. Taguchi, M. Takahashi, Y. Tsuboi, Y. Ujino, M. Nakajima, T. Yamamoto, H. Akatsu, S. Mitsui and N. Yamaguchi, *Psychiatry Clin. Neurosci.*, 2000, **54**, 419–426.
13. M. Repici and F. Giorgini, *J. Clin. Med.*, 2019, **8**, 1–11.
14. N. B. Chauhan, G. J. Siegel and J. M. Lee, *J. Chem. Neuroanat.*, 2001, **21**, 277–288.
15. N. Song, J. Wang, H. Jiang and J. Xie, *Free Radic. Biol. Med.*, 2010, **48**, 332–341.
16. P. Lei, S. Ayton, D. I. Finkelstein, P. A. Adlard, C. L. Masters and A. I. Bush, *Int. J. Biochem. Cell Biol.*, 2010, **42**, 1775–1778.
17. R. Von Coelln, V. L. Dawson and T. M. Dawson, *Cell Tissue Res.*, 2004, **318**, 175–184.
18. S. Avissar, Y. Nechamkin, G. Roitman and G. Schreiber, *Am. J. Psychiatry*, 1997, **154**, 211–217.
19. S. L. Rhodes, D. D. Buchanan, I. Ahmed, K. D. Taylor, M. A. Loriot, J. S. Sinsheimer, J. M. Bronstein, A. Elbaz, G. D. Mellick, J. I. Rotter and B. Ritz, *Neurobiol. Dis.*, 2014, **62**, 172–178.
20. S. Stern, S. Lau, A. Manole, I. Rosh, M. M. Percia, R. Ben Ezer, M. N. Shokhirev, F. Qiu, S. Schafer, A. A. F. Mansour, K. P. Mangan, T. Stern, P. Ofer, Y. Stern, A. P. Diniz Mendes, J. Djamus, L. R. Moore, R. Nayak, S. H. Laufer, A. Aicher, A. Rhee, T. L. Wong, T. Nguyen, S. B. Linker, B. Winner, B. C. Freitas, E. Jones, I. Sagi, C. Bardy, A. Brice, J. Winkler, M. C. Marchetto and F. H. Gage, *npj Park. Dis.*, , DOI:10.1038/s41531-022-00366-z.
21. S. T. Feng, Z. Z. Wang, Y. H. Yuan, X. Le Wang, H. M. Sun, N. H. Chen and Y. Zhang, *Pharmacol. Res.*, 2020, **151**, 104553.
22. T. Yamada, P. L. McGeer, K. G. Baimbridge and E. G. McGeer, *Brain Res.*, 1990, **526**, 303–307.
23. X. L. Kuang, F. Liu, H. Chen, Y. Li, Y. Liu, J. Xiao, G. Shan, M. Li, B. J. Snider, J. Qu, S. W. Barger and S. Wu, *J. Neurosci. Res.*, 2014, **92**, 1319–1329.
24. Y. C. Chee, J. B. Koprich, S. Endo and O. Isacson, *J. Neurosci.*, 2007, **27**, 8314–8323.
25. Y. Sun, X. Cheng, L. Zhang, J. Hu, Y. chen, L. Zhan and Z. Gao, *Mol. Neurobiol.*, 2017, **54**, 1008–1021.