Proceedings of the 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI-2026)
DVD Part Number: CFP26US4-DVD; ISBN: 979-8-3315-5518-4

Computational Preprocessing Framework for Small-Molecule and α-synuclein Data towards Parkinson's Drug Repurposing

G. Angel,
Research Scholar,
Department of Computer Science,
VISTAS,
Pallavaram, Chennai.
aangel.8888@gmail.com

Dr.P.Sujatha,
Professor & Head,
Department of Computer Applications,
VISTAS,
Pallavaram, Chennai.
Sujatha.scs@vistas.ac.in

***Abstract*—Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by the loss of dopaminergic neurons and abnormal aggregation of the α-synuclein protein, resulting in motor and cognitive deficits. The complexity of PD pathology and the restricted blood-brain barrier (BBB) permeability of most compounds present major challenges for therapeutic discovery. This study presents an implemented computational preprocessing framework for curating small-molecule and α-synuclein datasets, supporting reliable drug-target interaction (DTI) prediction and drug repurposing. During drug preprocessing, a multi-step curation pipeline was executed to standardize and refine PubChem compounds through canonicalization, charge normalization, tautomer correction, and PAINS filtering, followed by Lipinski and Veber criteria with CNS multiparameter optimization to identify molecules with optimal drug-like neuroactive profiles. The protein preprocessing stage employed a confidence-guided refinement approach on the AlphaFold-Predicted α-synuclein model (AF-P37840-F1-v6), combining per-residue pLDTT scoring, masking of low-confidence regions, and intrinsic disorder was assessed with IUPred2A, and structural repair under physiological conditions to generate a confidence-weighted structure optimized for computational modelling. Together, these harmonized preprocessing stages establish a reproducible, domain-specific foundation for subsequent ensemble docking, feature extraction, and graph-based learning. This framework enhances data quality, consistency, and interpretability in computational neurotherapeutics, supporting the future discovery of small-molecule modulators that effectively target α-synuclein and potentially slow Parkinson's disease progression.**

Keywords—data preprocessing; machine learning; molecular data standardization; intrinsically disordered protein; Parkinson's disease

I. INTRODUCTION

Parkinson's disease (PD) ranks as the second most prevalent neurodegenerative disorder after Alzheimer's disease, impacting more than 10 million individuals globally [1]. The disease is primarily characterized by the progressive degeneration of dopaminergic neurons in the substantia nigra, along with the accumulation of Lewy bodies – intracellular inclusions composed mainly of misfolded α-synuclein protein [1-3]. The aggregation of α-synuclein into oligomeric and fibrillar species disrupts neuronal integrity, impairs synaptic signaling, and contributes to oxidative stress, mitochondrial dysfunction, and neuroinflammation [2,4]. Despite extensive research, no approved therapies currently exist that can halt or reverse α–synuclein–associated neurotoxicity, underscoring a critical demand for effective disease-modifying interventions [3,5].

Computational drug discovery, especially drug-target interaction (DTI) prediction and drug repurposing, presents a promising approach to expedite the identification of novel therapeutics for Parkinson's disease [6,7]. In contrast to traditional screening methods, computational pipelines leverage chemical, structural, and biological data to infer binding or modulatory relationships between small molecules and target proteins [6]. However, the reliability of these models is heavily dependent on the quality of input data – encompassing both molecular structures and protein conformations [7,8]. Noisy, incomplete, or chemically invalid datasets frequently lead to model overfitting, spurious correlations, and predictions [8]. Public chemical databases such as PubChem and ChEMBL provide access to millions of molecular entries; however, many records contain redundant, incomplete, or syntactically invalid SMILES representations, as well as nonstandard tautomer or charged fragments [9,10]. Likewise, predicted protein structures from tools like AlphaFold2 may not always be appropriate for computational docking or graph-based modelling, especially in the case of intrinsically disordered proteins (IDPs) like α-synuclein [11,12]. IDPs are characterized by dynamic conformations and variable confidence across residues, requiring additional curation to remove or mask structurally ambiguous regions [11]. As a result, preprocessing is often underemphasized in published DTI pipelines, but becomes a critical step to ensure data integrity and enable meaningful downstream analysis [8,13].

Although several studies have investigated graph-based and deep learning methods for DTI predictions, relatively few works have specifically addressed data preprocessing frameworks that unify both drug and protein refinement within a reproducible workflow [7,14]. Moreover, for neurodegenerative disease targets, there remains a notable lack of preprocessing protocols that incorporate blood-brain barrier (BBB) relevance and disorder-aware protein handling [15]. This gap significantly limits the interpretability and clinical translatability of computational predictions in Parkinson's disease research [14,15].

In this work, we introduce an integrated preprocessing framework for small molecules and α-synuclein protein structures, specifically designed for

Parkinson's disease-focused drug discovery. The molecular preprocessing module includes canonicalization, charge normalization, tautomer correction, and removal of pan-assay interference compounds (PAINS), followed by physicochemical and CNS-related filtering [9,15]. The protein preprocessing module applies per-residue confidence scoring (pLDTT), disorder masking, and structural repair under physiological conditions to generate a reliable confidence-weighted model [11,12]. Collectively, these steps generate a standardized high-confidence dataset suitable for ensemble docking and deep learning-based DTI modelling. By ensuring both chemical validity and structural reliability, this work establishes a reproducible foundation for computational neurotherapeutics targeting α-synuclein aggregation. Therefore, the goal of this work is to provide a reproducible, Parkinson's-specific preprocessing framework that addresses the limitations of existing generic workflows. Unlike prior approaches that assume chemically valid inputs or structurally stable protein models [7,8,13], this work systematically evaluates how preprocessed filtering stages influence downstream computational suitability. Especially to determine the benefits of chemical preprocessing and CNS prioritization [9,15,20-26], assess confidence-guided trimming of α-synuclein structures [11,12,27-30], and show that harmonized drug-protein refinement generates model-ready datasets for docking and learning applications [33-36].

II. LITERATURE REVIEW

Cheminformatics preprocessing is widely adopted in virtual screening pipelines, routinely employed to ensure structural correctness and drug-likeness before predictive modelling. Established tools such as RDKit and Open Babel automate canonicalization, charge standardization, and filtering of unstable species, streamlining molecular curation workflows [17,18]. Similarly, PAINS detection algorithms, notably those introduced by Baell and Holloway [20], remain standard for excluding assay-interfering substructures that may contribute to false positives. However, the majority of existing cheminformatics workflows are optimized for general-purpose screening applications and do not routinely prioritize central nervous system pharmacokinetic considerations – such as blood-brain barrier permeability – which are increasingly vital for CNS-based drug discovery and development [23-26].

On the computational modelling side, large-scale drug-target interaction studies use public datasets such as ChEMBL, DrugBank, or PubChem to train advanced deep learning architectures like DeepDTA [7], GraphDTA [34], and transformer-based models [14]. These approaches have shown promising performance improvements by incorporating diverse molecular and protein sequence representations. Despite progress, many existing methods assume that input molecular data have already been thoroughly curated and high-quality, often overlooking structural noise and chemical invalidities. Such deficiencies can contribute to model overfitting, reduced generalizability, and decreased reproducibility in downstream predictions [8,13]. Therefore, incorporating rigorous preprocessing steps to ensure chemical and structural data integrity before deep learning model training remains critical for reliable and robust DTI prediction.

Protein preprocessing remains a less explored yet critical step in structure-based drug discovery for neurodegenerative diseases. Most structure-based studies rely on experimentally resolved PDB structures or AlphaFold2 predictions without applying residue-level confidence filtering. This approach often overlooks intrinsic disordered protein, especially in targets like α-synuclein, leading to unreliable binding-site geometries and misleading docking outcomes [11,27,29]. Recent research increasingly highlights the necessity of trimming low-confidence, low-pLDDT segments to enhance structural reliability and to avoid misleading docking conformations [28,30]. Such preprocessing safeguards against artefacts from flexible or disordered regions, improving the biological relevance and interpretability of computational predictions for neurotherapeutic development.

Prior works, therefore, lack a unified and computationally scalable preprocessing strategy that simultaneously standardizes chemical data, prioritizes CNS pharmacokinetic properties, and applies confidence-weighted refinement of intrinsically disordered protein structures. Existing workflows are typically optimized for general laboratory screening, lacking the disease-specific focus and model-ready output required for efficient machine learning and docking applications. The deficiency motivates the development of an integrated preprocessing framework designed to deliver harmonized, high-quality chemical and protein datasets specifically tailored for Parkinson's disease drug repurposing. Such a strategy improves computational efficiency and predictive accuracy, addressing critical gaps in the current process.

III. SYSTEM WORKFLOW AND PROCESSING

A unified preprocessing pipeline was developed to refine both chemical compounds and α-synuclein structured data before computational modelling. This workflow integrates drug and protein preprocessing modules within a reproducible data-curation framework, as illustrated in Fig.1. Formally, the resulting cleaned dataset can be defined as :

$$D_{clean} = f(D_{drug}, D_{protein}) \quad (1)$$

Where D_{drug} and $D_{protein}$ represent the processed chemical and protein datasets, respectively, and f denotes the integrated curation function.

A. Drug Preprocessing (Small-Molecule Data)

The initial chemical dataset was obtained from PubChem due to its extensive coverage of bioactive compounds, open accessibility, and inclusion of diverse chemical scaffolds suitable for data-driven drug discovery applications. PubChem raw drug molecules usually contain invalid atom valences, different aromaticity definitions, redundant and duplicated entries, corrupted SMILES strings, non-drug-like compounds, and errors resulting from data format conversion. In some cases, records do not contain a chemical structure. When it comes to disease specific some molecules are known to give false positives in screening assays. To remove these invalid drug molecules from the raw PubChem data, the model underwent a multi-stage refinement process to ensure structural integrity and pharmacological relevance [16]. The steps include:

- SMILES canonicalization to unify structural representation [17].
- Charge normalization to neutralize unstable ionized forms [18].
- Tautomer correction for consistent dominant structure [19].
- PAINS filtering to remove assay–interfering moieties [20,21].
- Fragment removal to eliminate disconnected atoms and non-druglike components [22].

Drug-likeness constraints were further evaluated using Lipinski's Rule-of-Five and Veber's permeability rules [23,24] to ensure suitable pharmacokinetic properties. Finally, CNS-active filtering was performed utilizing the multiparameter optimization score (CNS-MPO) [25]:

$$\text{CNS-MPO}=f(\text{cLogP, cLogS, MW, TPSA, HBD, pKa}) \quad (2)$$

Where cLogP describes lipophilicity, cLogS reflect solubility, MW denotes molecular weight, TPSA represents polarity, HBD indicates hydrogen-bond donor count, and pKa defines ionization propensity – collectively determining blood-brain barrier permeability.
Compounds fulfilling the threshold:

$$\text{CNS\_MPO} \geq 4 \quad (3)$$

are retained as candidates with a strong likelihood of CNS penetration for Parkinson's disease drug discovery [26]. This staged filtration produces a CNS-focused molecular library optimized for both structure-based docking and machine-learning prediction tasks.

B. *Protein Preprocessing (α-synuclein structural data)*

The predicted structure of α-synuclein from AlphaFold2 (Model ID: AF-P37840-F1-v6) is processed based on confidence and disorder characteristics relevant to interaction modelling. First, pLDDT-based residue quality filtering is applied to localize unreliable structural segments [27]. Intrinsic disorder is predicted using IUPred2A, a biophysics-based computational tool that predicts residue-level flexibility based on estimated inter-residue interaction energies, identifying regions unlikely to form stable tertiary structures [28-30]. Such flexible segments are biologically relevant in α-synuclein but may introduce uncertainty into docking and structural learning; therefore, they are masked while retained in the sequence. Structural repair is performed using PDBFixer for missing atoms and residue completion, followed by openMM for forcefield-based refinement and pH-dependent hydrogen assignment under physiological conditions (pH 7.4), ensuring proper protonation states, steric correctness, and stabilized electrostatic interactions required for reliable docking analysis [31,32].

Residue confidence assignment is defined as:

$$C(i) = \begin{cases} 1, \text{ if pLDDT}(i) \geq 70 \\ 0, \text{ if pLDDT}(i) < 70 \end{cases} \quad (4)$$

Where C(i) = 1 indicates a high-confidence residue retained for docking and structural learning, while C(i) = 0 represents a low-confidence residue that is masked to avoid unreliable spatial contributions. This selective filtration preserves α-synuclein's intrinsic disorder characteristics while ensuring downstream computational predictions remain focused on structurally reliable regions.

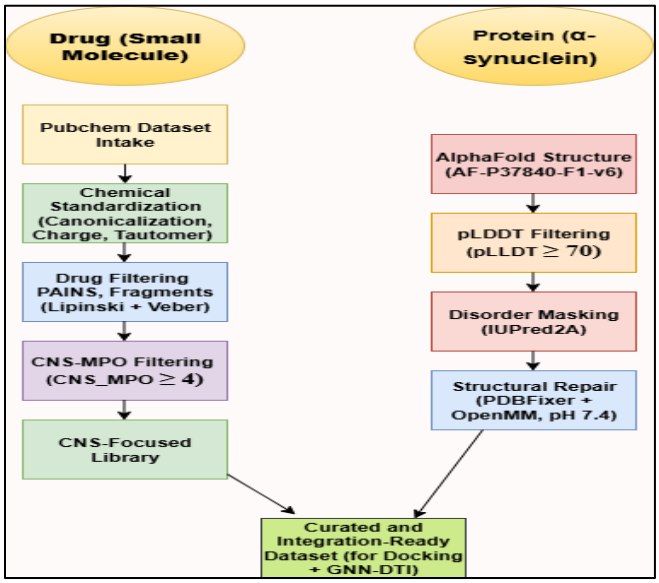

Figure 1. System Workflow for integrated drug and α-synuclein data preprocessing

C. *Preprocessed Dataset Generation and Integration Readiness*

The proposed preprocessing workflow produces two high-quality datasets:

(1) a chemically standardized and CNS-permeability-prioritized small-molecule library, and
(2) a confidence-weighted α-synuclein structural model optimized for computational analysis.

These curated datasets are designed for direct integration into ensemble molecular docking pipelines, which evaluate ligand binding across α-synuclein conformations to capture its intrinsic flexibility [33]. They also facilitate graph neural network (GNN)-based drug-target interaction models, enabling relational learning from both molecular graph structures and residue connectivity [34]. Furthermore, the enhanced data consistency supports robust feature extraction and interpretability, advancing computational neurotherapeutics that prioritize compounds modulating α-synuclein aggregation in Parkinson's disease [35,36].

IV. RESULTS AND ANALYSIS

The proposed preprocessing workflow was evaluated on both the chemical dataset and the α-synuclein structural model to assess improvements in data quality and computational suitability.

A. *Drug Dataset Refinement*

The chemical preprocessing pipeline demonstrated significant enhancement in dataset quality and domain relevance. As shown in Table 1, an initial set of 70,959 molecules sourced from PubChem was reduced to 62,982 compounds (88.76%) following chemical standardization and PAINS filtering. The subsequent application of Lipinski's Rule of Five and Veber's drug-likeness criteria further refined the dataset to 52,753 molecules (74.32%). The

most stringent filtering stage was based on CNS-MPO scoring (CNS_MPO ≥ 4), which prioritized blood-brain barrier permeability and yielded a final curated subset of 6,852 molecules (9.65%).

TABLE I. DRUG PREPROCESSING STATISTICS

Stage	Molecules Remaining	Removed	% Retained
Raw PubChem dataset	70,959	-	100%
Standardization + PAINS filtering	62,982	7.977	88.76%
Lipinski + Veber filtering	52,753	10,229	74.32%
CNS-MPO ≥ 4	6,852	45,901	9.65%

This stepwise refinement process, illustrated in Fig. 2, shows the sequential reduction of PubChem molecules during drug molecule preprocessing. Standardization converts the molecular representation of SMILES into a single, consistent format, ensuring that all software interprets the same molecule in the same way during drug screening. PAINS removal eliminates the molecules that tend to yield false positive results across various high-throughput screening (HTS) assays. Lipinski and Veber are used to filter the drug-likeness, such as absorption, distribution, metabolism, excretion, and toxicity. The final filter is based on the CNS-relevance scoring, which filters the drug molecule based on the ability to cross the blood-brain barrier (BBB) to avoid CNS-related side effects. This combined filter of drug molecules gives us a high-quality, curated dataset of chemical structures that are suitable for novel CNS-focused drug candidates.

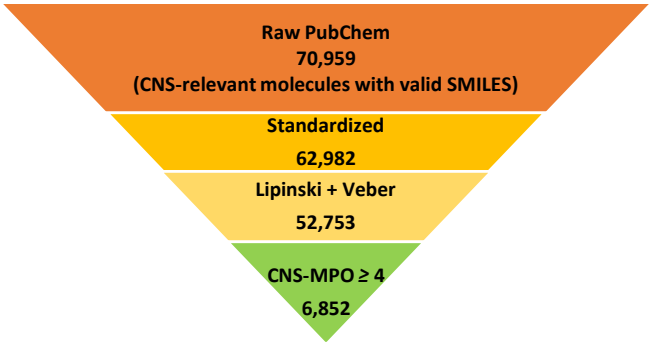

Figure 2. Drug preprocessing funnel showing progressive molecule reduction

Although filtering reduces the dataset size by over 90%, structural space remains preserved, suggesting that the removed molecules were largely non-CNS relevant or chemically invalid rather than valuable candidates. This is shown in Fig. 3 using PCA projection molecular fingerprints. The CNS-filtered subset (orange) continues to occupy the same broad chemical space as the full cleaned dataset (blue). Even though many drug molecules were removed based on CNS developability constraints, the retained molecules remain distributed across the full diversity landscape, with notable enrichment in regions associated with CNS-favoured properties. This proves that the filtering strategy pruned unsuitable molecules while preserving meaningful structural variability for downstream screening and modelling.

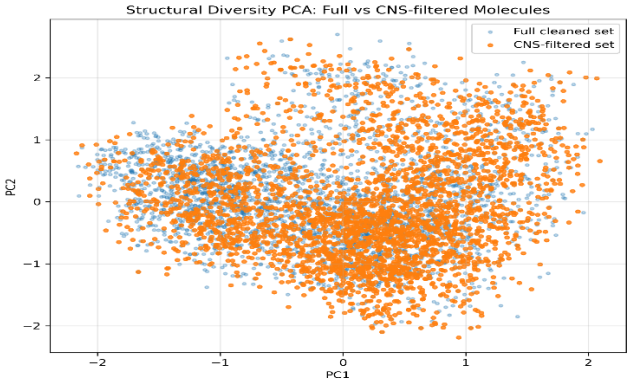

Figure 3. PCA-based structural diversity illustrating preservation of chemical space after CNS filtering

Next drug-only classification model using the cleaned molecular dataset was trained to track its validation behaviour. Each drug molecule was represented using Morgan circular fingerprints with radius 2 generated using RDKit and fed into a feed-forward neural network classifier trained on drug features. The model was executed for 40 epochs, during which validation accuracy and cross-entropy loss were logged per epoch to evaluate convergence, stability, and overfitting trends. Fig. 4 shows that accuracy quickly increased and stabilised around ~0.97, while loss steadily declined, indicating smooth optimisation and reliable generalisation. This confirms that the cleaned drug dataset, when encoded using Morgan fingerprints, is learnable and predictive.

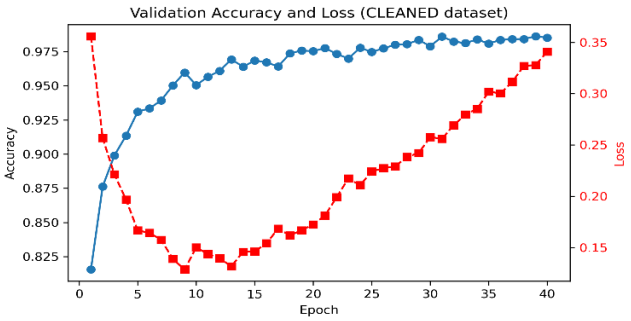

Figure 4. Training Stability Analysis: Validation accuracy gains and loss reduction observed for preprocessed drug representation

B. *Protein Structure Confidence Analysis*

The α-synuclein structural model exhibited heterogeneous residue stability, consistent with its intrinsically disordered nature. Applying a per-residue pLDDT threshold of 70 retained 88 high-confidence residues (62.9%), while masking 52 low-confidence residues (37.1%) to minimise unreliable contributions as summarised in Table II.

TABLE II. PROTEIN RESIDUE CONFIDENCE SUMMARY

Classification	Residue Count	Percentage
High-Confidence (pLDDT ≥ 70)	88	62.9%

Low-Confidence (pLDDT < 70)	52	37.1%

The spatial distribution of these residues, depicted in Fig. 5, highlights disorder-rich regions that, if unfiltered, may lead to mistakes in molecular docking and graph neural network-based structural models. This selective filtration enhances geometric fidelity while preserving biologically relevant flexible segments implicated in α-synuclein aggregation, thereby improving the accuracy and interpretability of downstream computational models. Furthermore, the improvement in structural reliability achieved through pLDTT-based filtering ensures stable and accurate downstream force field evaluations and contact map generation.

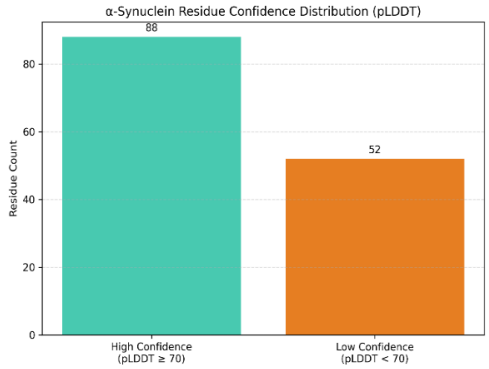

Figure 5. Residue-level pLDTT confidence distribution for the α-synuclein

To validate whether the pLDDT-filtered structure supports stable feature learning, a protein-only classification model was trained on residue-level embeddings, which were derived from the high-confidence regions. Fig. 6 shows validation accuracy increased sharply in early epochs and later stabilised around ~0.85-0.90, while loss consistently declined over training. This behaviour demonstrates that filtering low-confidence residues improves representational reliability and enables smoother optimisation compared to unfiltered structural inputs.

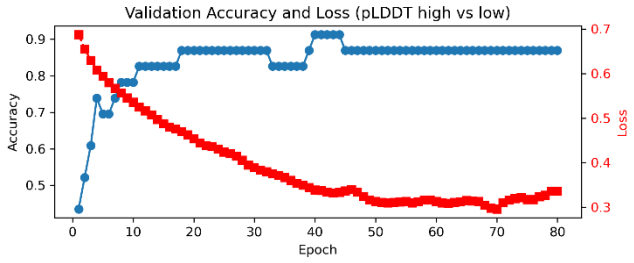

Figure 6. Residue-level pLDTT confidence distribution for the α-synuclein

C. *Effectiveness of Preprocessing for Modelling Integration and Limitations*

The final output demonstrates significant enhancement in both chemical and structural data quality. The curated drug library eliminates poorly defined and non-CNS-active molecules, thereby reducing false positives during virtual screening and improving physicochemical suitability for CNS-target machine learning models. Meanwhile, confidence-based trimming of the α-synuclein structure reduces geometric instability and minimises steric artefacts, yielding a model more appropriate for docking scoring functions and other neural network encodings. Due to these improvements, the dataset is now computationally efficient by reducing the number of docking simulations by over 90%, while retaining disorder-aware features critical for studying α-synuclein aggregation. Together, these refinements enhance data reliability, reduce the risk of model overfitting, and strengthen the interpretability of future binding predictions and neurotherapeutic screening outcomes. These learnability patterns confirm that preprocessing stabilises downstream optimisation, directly improving suitability for DTI modelling applications.

Compared with raw PubChem input and unfiltered AlphaFold structures, the generic preprocessed workflows reported in literature do not include CNS prioritisation or disorder-aware trimming. Our results show additional performance improvements and modelling stability, indicating that the proposed framework offers measurably improved computational behaviour and interpretability, showing that the framework overcomes the key limitations of unrefined pipelines. Although the proposed framework improves chemical validity and structural reliability, it remains computational in nature and does not incorporate experimental binding validation or in-vivo performance assessment. The current workflow also assumes AlphaFold-derived conformations as starting models and may benefit from ensemble or NMR-derived disorder representations in future studies. Also, toxicity and ADMET properties are not fully captured in this stage of workflow, which indicates an opportunity for further enhancement as downstream docking and machine learning-based affinity prediction.

D. *Comparative Analysis with Existing Preprocessing Techniques*

There are many preprocessing strategies in cheminformatics and protein modelling; however, most are designed for general-purpose virtual screening, and they do not address neurotherapeutic or α-synuclein-specific needs. Table III compares the commonly adopted preprocessing techniques and the proposed Parkinson's-based framework. Existing chemical preprocessing tools, such as RDKit and OpenBabel, mostly perform the basic cleaning and PAINS removal, but lack integrated CNS-relevant filtering and multi-stage refinement necessary for neurological drug discovery [17],[18],[20],[23],[25]. For protein, most of the studies use AlphaFold or PDB structures without residue-level confidence filtering, which is problematic for intrinsically disordered proteins like α-synuclein, where low-pLDDT regions compromise docking fidelity [11], [27-30]. The proposed framework combines CNS-aware chemical filtering with pLDTT-and disorder-guided protein refining. This addresses the limitation of existing methods and makes the dataset ready for modeling Parkinson's disease DTI.

TABLE III. EXISTING PREPROCESSING TECHNIQUES VS. PROPOSED FRAMEWORK

Preprocessing Aspect	**Existing Methods**	**Limitation**	**Proposed Framework**
Chemical Sanitization	RDKit / OpenBabel canonicalization, valence check [17],[18]	Does not enforce CNS-related and may retain invalid SMILES	Complete validation + tautomer correction + charge normalization
Pains filtering	Standard PAINS screens [20],[21]	Often executed independently	Combined PAINS

		and doesn't combine with downstream filters	removal within the multistage process
Drug-likeness evaluation	Lipinski/Veber rules [23],[24]	General purpose, no neuro-related constraints	Ro5 + Veber + CNS-MPO for BBB relevance
Structural diversity preservation	Rarely evaluated	No analysis of chemical space retention	PCA diversity assessment (Fig. 3)
Protein structure validation	Direct use of AlphaFold2 or PDB [11], [27]	Ignores low-confidence regions; unsuitable for IDPs	pLDDT-based residue filtering + disorder masking
Docking / ML	Assumes stable geometry	Sensitive to noise → docking artefacts	Geometry – stabilized α-synuclein + cleaned chemical set improves scoring
Combined drug-protein preprocessing	Not present in existing pipelines	Drug and Protein are processed independently	Integrated, disease-specific workflow supporting DTI and repurposing

V. Conclusion

This work presented a novel, integrated preprocessing framework that concurrently refines chemical and protein data into harmonised, Parkinson's disease-tailored inputs. Unlike conventional approaches that process small-molecule and protein data independently, this framework incorporates disease-specific CNS permeability constraints alongside pLDDT-based confidence filtering for the intrinsically disordered α-synuclein. The resulting datasets exhibit enhanced structural reliability, reduced noise and artefact prevalence, and enhanced computational suitability, enabling more precise and interpretable downstream modelling. The key innovation of this work lies in unifying CNS-aware molecular refinement with disorder-aware α-synuclein trimming within a reproducible, Parkinson's-specific workflow, addressing gaps left by the general-purpose preprocessing strategies. Looking ahead, these curated datasets will be integrated into ensemble docking workflows targeting multiple α-synuclein conformations, supported by graph neural network-based drug-target interaction prediction to identify promising modulators of α-synuclein aggregation behavior. Further development will combine toxicity screening, experimental bioactivity data, and ADMET profiling to enhance translational confidence, while multi-omics and pathway-level analyses may also strengthen mechanistic insights and predictive accuracy. Together, this preprocessing foundation provides a robust and scalable platform to accelerate the discovery and prioritization of neurotherapeutics aimed at α-synuclein pathology in Parkinson's disease, driving progress towards more interpretability and disease-aligned computational pipelines that advance precision medicine interventions.

References

[1] Calabresi, P., Mechelli, A., Natale, G., Volpicelli-Daley, L., Di Lazzaro, G., & Ghiglieri, V. (2023). Alpha-synuclein in Parkinson's disease and other synucleinopathies: from overt neurodegeneration back to early synaptic dysfunction. *Cell Death and Disease*, *14*(3), 176. https://doi.org/10.1038/s41419-023-05672-9

[2] Du, X., Xie, X., & Liu, R. (2020). The role of A-Synuclein oligomers in Parkinson's disease. *International Journal of Molecular Sciences*, *21*(22), 8645. https://doi.org/10.3390/ijms21228645

[3] Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, *397*(10291), 2284–2303. https://doi.org/10.1016/s0140-6736(21)00218-x

[4] Bridi, J. C., & Hirth, F. (2018). Mechanisms of A-Synuclein induced synaptopathy in Parkinson's disease. *Frontiers in Neuroscience*, *12*, 80. https://doi.org/10.3389/fnins.2018.00080

[5] Schapira, A. H. (2004). Disease modification in Parkinson's disease. *The Lancet Neurology*, *3*(6), 362–368. https://doi.org/10.1016/s1474-4422(04)00769-0

[6] Ezzat, A., Wu, M., Li, X., & Kwoh, C. (2018). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, *20*(4), 1337–1357. https://doi.org/10.1093/bib/bby002

[7] Ru, X., Xu, L., Han, W., & Zou, Q. (2025). In silico methods for drug-target interaction prediction. *Cell Reports Methods*, *5*(10), 101184. https://doi.org/10.1016/j.crmeth.2025.101184

[8] Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/258814

[9] Landrum, Greg. "Rdkit: Open-source cheminformatics software." 2016

[10] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2015). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951

[11] Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, *433*(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

[12] Wilson, C. J., Choy, W., & Karttunen, M. (2022). AlphaFold2: A role for Disordered Protein/Region Prediction? *International Journal of Molecular Sciences*, *23*(9), 4591. https://doi.org/10.3390/ijms23094591

[13] Zhang, Y., Hu, Y., Han, N., Yang, A., Liu, X., & Cai, H. (2023). A survey of drug-target interaction and affinity prediction methods via graph neural networks. *Computers in Biology and Medicine*, *163*, 107136. https://doi.org/10.1016/j.compbiomed.2023.107136

[14] Talukder, M. A., Kazi, M., & Alazab, A. (2025b). Predicting drug-target interactions using machine learning with improved data balancing and feature engineering. *Scientific Reports*, *15*(1), 19495. https://doi.org/10.1038/s41598-025-03932-6

[15] Wager, T. T., Hou, X., Verhoest, P. R., & Villalobos, A. (2010). Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach To Enable Alignment of Druglike Properties. *ACS Chemical Neuroscience*, *1*(6), 435–449. https://doi.org/10.1021/cn100008c

[16] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2015b). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951

[17] Rogers, D., & Hahn, M. (2010). Extended-Connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

[18] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*(1), 33. https://doi.org/10.1186/1758-2946-3-33

[19] Guasch, L., Sitzmann, M., & Nicklaus, M. C. (2014). Enumeration of Ring–Chain tautomers based on SMIRKS rules. *Journal of Chemical Information and Modeling*, *54*(9), 2423–2432. https://doi.org/10.1021/ci500363p

[20] Baell, J. B., & Holloway, G. A. (2010). New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, *53*(7), 2719–2740. https://doi.org/10.1021/jm901137j

[21] Baell, J., & Walters, M. A. (2014). Chemistry: Chemical con artists foil drug discovery. *Nature*, *513*(7519), 481–483. https://doi.org/10.1038/513481a

[22] Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, *39*(15), 2887–2893. https://doi.org/10.1021/jm9602928

[23] Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., & Andrade, C. H. (2018). QSAR-Based Virtual Screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, *9*, 1275. https://doi.org/10.3389/fphar.2018.01275

[24] Veber, D. F., Johnson, S. R., Cheng, H., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, *45*(12), 2615–2623. https://doi.org/10.1021/jm020017n

[25] Wager, T. T., Hou, X., Verhoest, P. R., & Villalobos, A. (2016). Central Nervous System Multiparameter Optimization Desirability: Application in drug Discovery. *ACS Chemical Neuroscience*, *7*(6), 767–775. https://doi.org/10.1021/acschemneuro.6b00029

[26] Hitchcock, S. A., & Pennington, L. D. (2006). Structure−Brain exposure relationships. *Journal of Medicinal Chemistry*, *49*(26), 7559–7583. https://doi.org/10.1021/jm060642i

[27] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[28] Brotzakis, Z. F., Zhang, S., Murtada, M. H., & Vendruscolo, M. (2025). AlphaFold prediction of structural ensembles of disordered proteins. *Nature Communications*, *16*(1), 1632. https://doi.org/10.1038/s41467-025-56572-9

[29] Grewal, A., Sheokand, D., Saini, V., & Kumar, A. (2024). Molecular docking analysis of α-Synuclein aggregation with Anle138b. *Bioinformation*, *20*(3), 217–222. https://doi.org/10.6026/973206300200217

[30] Venati, S. R., & Uversky, V. N. (2024). Exploring intrinsic disorder in human synucleins and associated proteins. *International Journal of Molecular Sciences*, *25*(15), 8399. https://doi.org/10.3390/ijms25158399

[31] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, *13*(7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659

[32] Talagayev, V., Chen, Y., Doering, N. P., Obendorf, L., Denzinger, K., Puls, K., Lam, K., Liu, S., Wolf, C. A., Noonan, T., Breznik, M., Knaus, P., & Wolber, G. (2025). OpenMMDL - Simplifying the complex: Building, Simulating, and analyzing Protein–Ligand systems in OpenMM. *Journal of Chemical Information and Modeling*, *65*(4), 1967–1978. https://doi.org/10.1021/acs.jcim.4c02158

[33] Trott, O., & Olson, A. J. (2009). Soleymani, F., Paquet, E., Viktor, H., Michalowski, W., & Spinello, D. (2022). Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, *20*, 5316–5341. https://doi.org/10.1016/j.csbj.2022.08.070

[34] *Journal of Computational Chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

[35] Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2020). GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, *37*(8), 1140–1147. https://doi.org/10.1093/bioinformatics/btaa921

[36] Baldassarre, F., Hurtado, D. M., Elofsson, A., & Azizpour, H. (2020). GraphQA: protein model quality assessment using graph convolutional networks. *Bioinformatics*, *37*(3), 360–366. https://doi.org/10.1093/bioinformatics/btaa714