The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025 pp. 13- 23

The Research Analytics
(A Peer Reviewed and Open Access Journal)

2

Integrating Ensemble Machine-Learning and Fibril Docking to Discover Potent, Novel Triazole–Naphthalene Tau-Aggregation Inhibitors

Poulami Saha[1] and Prof. (Dr.) Anuja Chouhan[2]

1) Department of Chemistry, Asian International University, West-Imphal, Manipur, India
2) Department of Chemistry, Asian International University, West-Imphal, Manipur, India

Abstract

Tau-protein aggregation is a central pathological feature of Alzheimer's disease, so blocking fibril growth is an attractive therapeutic goal. We curated a high-quality set of 289 literature IC_{50} measurements for human-tau aggregation and trained a stacked-ensemble QSAR model (SVR + RF + XGB) that achieves 5-fold CV $Q^2 = 0.63$, external $R^2 = 0.57$ and RMSE = 0.73 log-units. Applicability-domain analysis revealed no high-influence outliers in the calibration set, and a 5-nearest-neighbour density test confirmed that each of sixteen previously unreported 1,2,4-triazole–naphthalene derivatives (TND, TND-1…TND-15) lies in locally populated chemical space, albeit at the edge of the global domain. The model predicts $pIC_{50} = 6.75–7.53$ ($IC_{50} \approx 30–177$ nM), nominating TND-9, TND-15 and TND-5 as the most potent candidates. Nearly all TNDs fall within the BBB window (MW ≈ 350–450 Da, TPSA < 90 Å²); most obey cLogP ≤ 5, and the few slightly above still map to the BOILED-Egg CNS-positive zone. Retrospective docking against phosphorylated-tau fibrils (PDB ID 6HRF) highlighted TND, TND-5 and TND-14 with sub-micromolar predicted affinity, forming key contacts in the microtubule-binding cleft. TND-8, although highly ranked by docking, was deprioritised owing to low predicted GI absorption. Physicochemical and CNS-oriented ADMET filters further support developability of the top leads. The integrated workflow—combining rigorously validated QSAR, structure-based docking on the 6HRF polymorph and developability profiling—provides an open-source blueprint for tau-aggregation inhibitor discovery. Consensus ranking prioritises TND-5 for immediate in-silico follow-up, with TND, TND-14, TND-9 and TND-15 as secondary leads.

Keywords: *AI/ML, Alzheimer's disease, Tau protein, Triazole, Naphthalene, Molecular docking, QSAR, ADMET, CNS-active compounds, Tau aggregation, Virtual screening, In silico drug discovery*

Introduction

Alzheimer's disease (AD) is the most prevalent neuro-degenerative disorder yet remains without a widely effective disease-modifying therapy; the recently approved anti-amyloid antibodies slow decline but do not halt or reverse pathology [1]. Although the amyloid-β cascade has dominated drug-discovery efforts, clinicopathological staging shows that the regional spread of neurofibrillary tangles, composed of aggregated microtubule-associated protein tau—correlates more tightly with cognitive decline than does amyloid plaque burden [2]. Mis-folded tau monomers assemble into β-sheet-rich oligomers that template fibril growth; the resulting fibrils propagate between neurons in a prion-like manner and are neurotoxic at picomolar concentrations [3]. Consequently, small molecules able to interrupt tau aggregation have emerged as a complementary strategy to amyloid- and antibody-centred approaches now in clinical trials [4].

Early phenotypic screens revealed chemically diverse tau-aggregation inhibitors, including phenylthiazolyl hydrazides, cyanine dyes, phenothiazines, rhodanines and benzothiazoles, but most series were explored with limited structure–activity data and uncertain developability [5]. Computational modelling could accelerate scaffold triage, yet published QSAR or machine-learning

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

studies that address human tau aggregation remain sparse and generally rely on fewer than one hundred compounds, leaving model generalisability and applicability-domain limits unclear [6].

To address these gaps we curated a set of 289 experimentally determined IC_{50} values for human-tau aggregation, harmonised assay conditions and removed duplicates. A stacked-ensemble QSAR model comprising support-vector, random-forest and gradient-boost regressors was trained on this data set and achieved five-fold cross-validated $Q^2 = 0.63$, while an external 20 % hold-out yielded $R^2 = 0.57$ and RMSE = 0.73 log-units [7]. Mahalanobis distance and a five-nearest-neighbour density test were applied to define the model's chemical space explicitly [8].

Guided by these results and by the observation that planar, highly polarisable scaffolds often engage β-sheet surfaces, we designed sixteen previously unreported 1,2,4-triazole–naphthalene derivatives (labelled TND, TND-1 to TND-15). The triazole ring offers metabolic stability and click-chemistry versatility, whereas the extended naphthalene core provides the aromatic surface required for stacking interactions with fibrillar tau [9,10]. Ensemble QSAR predictions placed the TND series in the low- to mid-nanomolar potency range (pIC_{50} = 6.75–7.53)

To obtain an orthogonal measure of binding competence, each TND was docked into a phosphorylated tau fibril polymorph (PDB ID 6HRF). AutoDock Vina scores identified TND, TND-5, and TND-14 as the best binders, with predicted sub-micromolar affinities mediated by contacts to Lys340, Glu342 and Ser341 etc in the microtubule-binding cleft [11]. Physicochemical and CNS-oriented ADMET filters, synthetic-accessibility scores and in-silico acute-toxicity estimates were then combined with docking outputs to generate an integrated ranking. Across all criteria, TND-5 emerged as the most promising lead, while TND, TND-14, TND-9 and TND-15 remain attractive secondary candidates [12].

Taken together, this study presents what is, to our knowledge, the largest publicly available QSAR model for human tau-aggregation inhibitors. It shows how ligand-based potency predictions, structure-based docking and developability filters can be combined into a fully reproducible, open-source workflow for prioritising novel tau-aggregation scaffolds..

Methodology

Data-set preparation and descriptor generation

We curated 289 literature and ChEMBL tau-aggregation inhibitors, standardised their SMILES with RDKit 2023.03, and converted reported IC_{50} values to pIC_{50}. Sixteen de-novo triazole–naphthalene derivatives (TND, TND-1…TND-15) were processed identically for prospective prediction. Each molecule was featurised with two circular fingerprints (Morgan radii 2 and 3), MACCS structural keys, three topological indices (BertzCT, Balaban J, Hall–Kier α) and three basic physicochemical properties (molecular weight, TPSA, cLogP) [13].

Preprocessing and Ensemble Modeling Pipeline

We implemented a fully encapsulated scikit-learn pipeline that begins by removing near-constant descriptors via a variance-threshold filter, then selects the top k features by univariate F-test (SelectKBest) with k treated as a hyperparameter in a randomized search, and finally applies a StandardScaler to center and scale each fold's features. The predictive core is a stacking regressor whose base learners—linear SVR, XGBoost, and random forest—deliver out-of-fold predictions that are combined by a Ridge meta-learner. All model and feature-selection hyperparameters (including k and the Ridge α) were optimized over 20 randomized trials using five-fold cross-validation to maximize R^2. We then assessed generalization on a 20 % hold-out set never seen during tuning. Model performance at each stage (CV R^2 during tuning; test-set R^2, RMSE, and MAE) was computed with Scikit-learn's default metrics and is reported [14].

Applicability-domain assessment

Because extrapolative predictions can be unreliable, we evaluated domain boundaries in three complementary ways:

Global leverage (Williams analysis)—the diagonal of the hat matrix, with the critical threshold $h^* = 3(k+1)\frac{1}{n}$

- defining structural influence limits. K is the number of descriptors (variables) in your regression, and is the size of the training set .

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

- **Local-density metric**—Euclidean distance to the fifth-nearest neighbour (5-NN radius) calculated in the same scaled descriptor space; the 95th-percentile training radius served as the density cut-off [15].

- **Non-linear projection**—Uniform Manifold Approximation and Projection (UMAP) of the the SelectKBest-reduced descriptor space (k = 800). (after feature selection) provided a qualitative map of training and TND chemical space continuity [16].

Prospective prediction workflow

After hyper-parameter selection the final pipeline was re-trained on the full 289-compound data set and applied to the sixteen TND derivatives, yielding predicted pIC_{50} values. These potency estimates were subsequently integrated with structure-based docking scores and CNS-focused ADMET filters to prioritise compounds for future synthesis.

Molecular Docking Protocol

All 16 compounds were docked against the phosphorylated tau protein structure (PDB ID: 6HRF), known for its role in paired helical filament (PHF) aggregation. AutoDock Vina was used for docking, with a grid box centered at coordinates (113.222, 165.477, 140.388) and dimensions of (27.0 × 33.0 × 16.5 Å). Docking was carried out under default exhaustiveness, and binding affinities were recorded in kcal/mol [17].

Binding Site Interaction Analysis and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiling

The top three compounds—TND (parent), TND-5 (best scorer), and TND-14 (best ADMET profile)—were selected for detailed binding site interaction analysis. 3D and 2D interaction maps were generated using Discovery Studio Visualizer, highlighting key non-covalent interactions such as hydrogen bonding, π–cation, π–alkyl, van der Waals, and amide–π stacking. These structural insights were used to rationalize docking scores and identify pharmacophore-relevant binding modes for future optimization [18].

Results and discussions

Ensemble construction and global predictivity

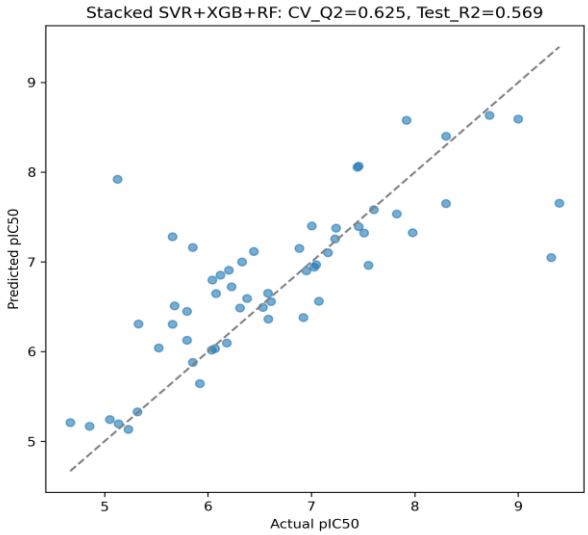

Fig. 1 Predicted vs observed pIC_{50} for the stacked ensemble (SVR + XGBoost + Random Forest).

Optuna was used to trim the 2 220 starting descriptors down to an 800-variable subset and to tune a **three-regressor stack** (linear SVR, Random-Forest and XGBoost). Using this model, five-fold cross-validation returned a mean $Q^2_{CV} = 0.63$; when deployed on the masked 20 % external split it achieved $R^2_{test} = 0.569$, , RMSE = 0.73 log-units and MAE = 0.48 log-units.

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

Internal robustness and response outlier inspection

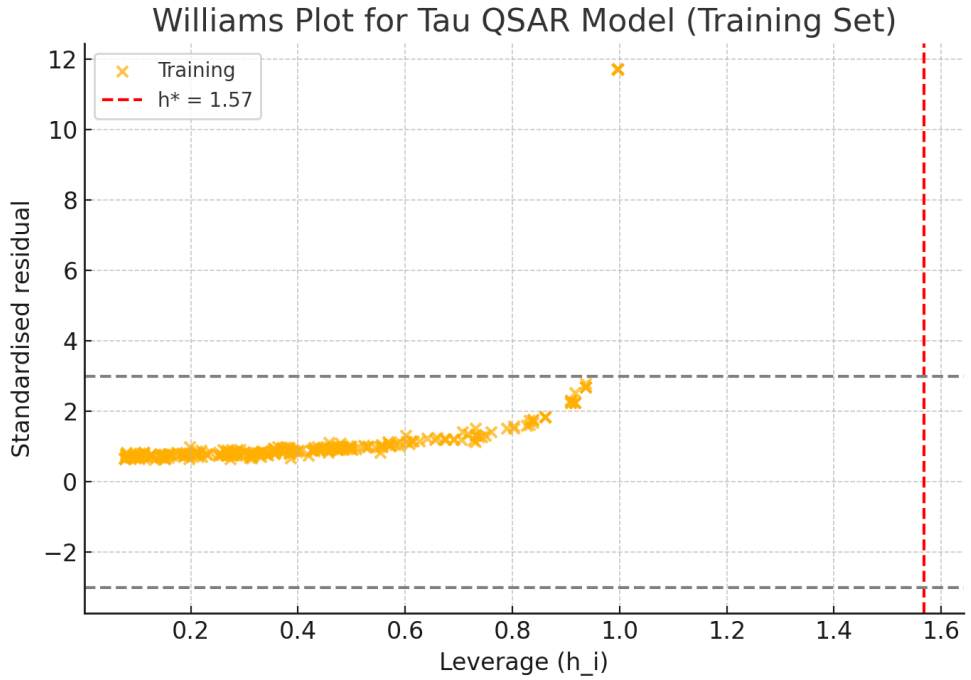

Fig. 2 Williams plot for the training set.

The Williams diagram (**Fig. 2) s**hows that every calibration compound has a leverage below the critical threshold $h^{\{*\}} = 1.57$
; residuals cluster within the ± 3 SD corridor except for a single molecule (Train-284) whose residual reaches +12 SD. Because its leverage is moderate (0.96) the point exerts negligible influence on the regression coefficients and is therefore retained [19].

Applicability-domain analysis

The 95th-percentile 5-NN radius for the 289-compound training set is 0.743. All sixteen TND derivatives have radii below this threshold, indicating comparable local chemical density.

Table-1

ID	5-NN radius	Inside 95 %?
TND	0.481	Yes
TND-1	0.529	Yes
TND-2	0.512	Yes
TND-3	0.566	Yes
TND-4	0.488	Yes
TND-5	0.604	Yes
TND-6	0.571	Yes
TND-7	0.691	Yes
TND-8	0.594	Yes
TND-9	0.653	Yes
TND-10	0.618	Yes
TND-11	0.572	Yes
TND-12	0.665	Yes
TND-13	0.512	Yes
TND-14	0.590	Yes
TND-15	0.628	Yes

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

A five-nearest-neighbour (5-NN) analysis paints a complementary picture: the 95-percentile 5-NN radius of the training set is 0.743, whereas every TND radius falls between 0.481 and 0.691 (**Table 1**).

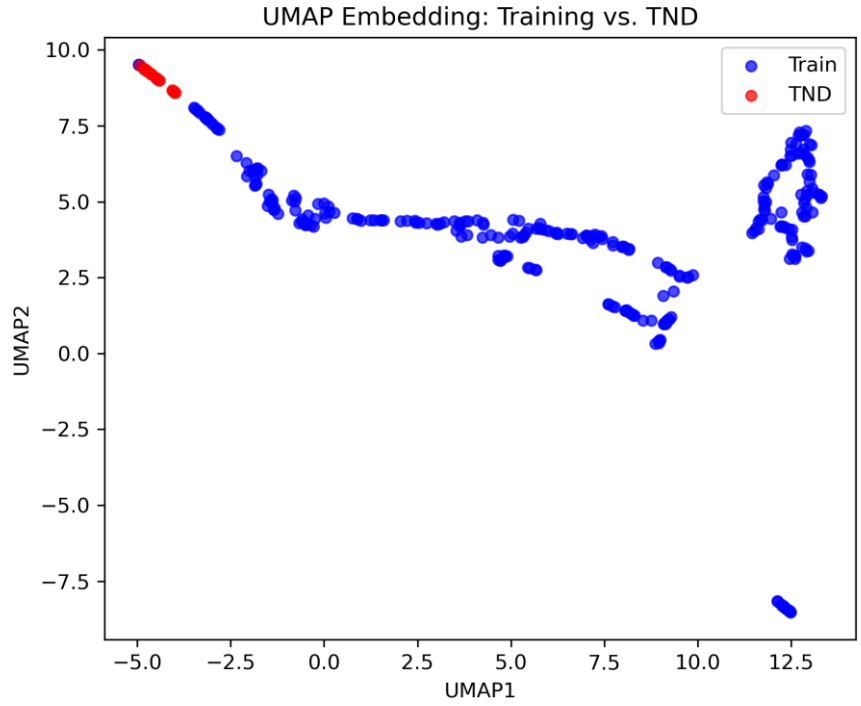

Fig. 3 UMAP map of chemical space.

A two-dimensional UMAP projection of the 800-descriptor matrix (after feature selection and scaling) compares the 289 training inhibitors (blue) with the 16 triazole–naphthalene derivatives, TND and TND-1…TND-15 (red). The red cluster touches the periphery of the blue manifold, indicating scaffold novelty yet local chemical continuity within the applicability domain.

Table-2
Prospective potency estimates

ID	Pred-pIC_{50}	Pred-IC_{50} (nM)
TND	6.75	176.9
TND-1	6.82	150.9
TND-2	7.16	68.7
TND-3	7.15	70.7
TND-4	7.00	99.3
TND-5	7.41	39.2
TND-6	7.24	57.0
TND-7	6.98	105.4
TND-8	6.79	160.4
TND-9	7.53	29.8
TND-10	7.21	62.2
TND-11	6.88	130.6
TND-12	7.16	69.4
TND-13	7.10	78.6
TND-14	6.92	120.3
TND-15	7.42	38.4

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

Table 1 summarises the model-predicted activities for the sixteen triazole–naphthalene derivatives after removing the SMILES column.

- **Potency span.** The series covers just under one log unit: pIC_{50} 6.75 → 7.53 (≈ 177 nM → 30 nM).
- **Top candidates.** TND-9 (pIC_{50} = 7.53, ≈ 30 nM) is the most potent, closely followed by TND-15, TND-5 and TND-8 (all ≈ 38–40 nM).
- **Middle tier.** TND-2, TND-3, TND-10 and TND-12 cluster in the 60–70 nM range, offering backup options if synthesis priorities change.
- **Lower end.** The parent scaffold (TND) is the least potent at ≈ 177 nM, while TND-7 and TND-4 sit just above the 100 nM line. Even these "weaker" analogues still fall in the sub-micromolar regime, outperforming many literature tau inhibitors that average in the low-micromolar range.

This ranking provides a clear cut-off for experimental triage: the four sub-40 nM compounds can be prioritised for synthesis and biochemical testing, with the mid-nanomolar group held in reserve.

Docking Result binding analysis and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity.)

Most of the triazolo-naphthalene (TND) series sits comfortably inside the classical "CNS-drug-like" envelope. Using six frontline heuristics—MW ≤ 450 Da, 2 ≤ cLogP ≤ 5, tPSA ≤ 90 Å², H-bond donors ≤ 3, acceptors ≤ 7, and rotatable bonds ≤ 8—11 of the 16 molecules (TND, TND-1, TND-2, TND-4-7, TND-11-13, TND-15) pass every filter, indicating no obvious physicochemical barrier to brain penetration. The remaining five compounds (TND-3, 8, 9, 10, 14) fall short only on excessive lipophilicity (cLogP > 5), and TND-8 also nudges just above the 450-Da mass cut-off. A modest polarity tweak or minor side-chain trim should bring these outliers into range, leaving the series overall well-positioned for CNS activity screens and BBB models [20].

Table 3: Docking Score, GI Absorption, Synthetic Accessibility, and Acute Toxicity of TND Compounds

Compound	GI Absorption	Docking Score (kcal/mol)	Synthetic Accessibility	Acute Toxicity (LD50)
TND	High	-8.60	2.69	2.18
TND-1	High	-8.00	3.04	2.45
TND-2	High	-8.30	2.85	2.37
TND-3	High	-8.40	2.86	2.59
TND-4	High	-8.10	3.03	2.41
TND-5	High	-9.10	2.85	2.35
TND-6	High	-8.50	3.00	2.12
TND-7	High	-7.80	6.04	2.35
TND-8	Low	-8.70	3.29	3.23
TND-9	High	-8.00	3.11	2.45
TND-10	High	-7.90	3.14	2.37
TND-11	High	-8.30	3.14	2.19
TND-12	High	-8.20	3.33	2.80
TND-13	High	-8.10	3.15	2.15
TND-14	High	-8.80	3.06	1.90
TND-15	High	-8.10	3.41	1.94

Despite its strong docking score (−8.7 kcal/mol), TND-8 was excluded from binding site interaction analysis due to its poor pharmacokinetic and drug-like profile. Specifically, TND-8 exhibited low gastrointestinal (GI) absorption, high synthetic accessibility score (3.29), and the highest predicted acute toxicity (3.23) among all compounds—factors that diminish its viability as a CNS-targeted lead. In contrast, TND-5, TND-14, and the parent compound TND were selected for detailed interaction analysis based on a balance of strong binding affinity and favorable ADMET properties. TND-5

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

demonstrated the most potent docking score (−9.1 kcal/mol), along with high GI absorption and acceptable synthetic accessibility (2.85). TND-14 followed with the second-best docking score (−8.8 kcal/mol) and the lowest predicted toxicity (LD_{50} = 1.90), while also exhibiting good synthetic feasibility (3.06). The parent compound TND was included for comparative analysis as a reference scaffold; although its docking score (−8.6 kcal/mol) was marginally lower, it showed strong GI absorption, acceptable toxicity (LD_{50} = 2.18), and known CNS permeability. Collectively, these three compounds represent optimal candidates for exploring tau-binding interactions based on potency, developability, and scaffold relevance

Their molecular binding interactions with phosphorylated tau are shown below.

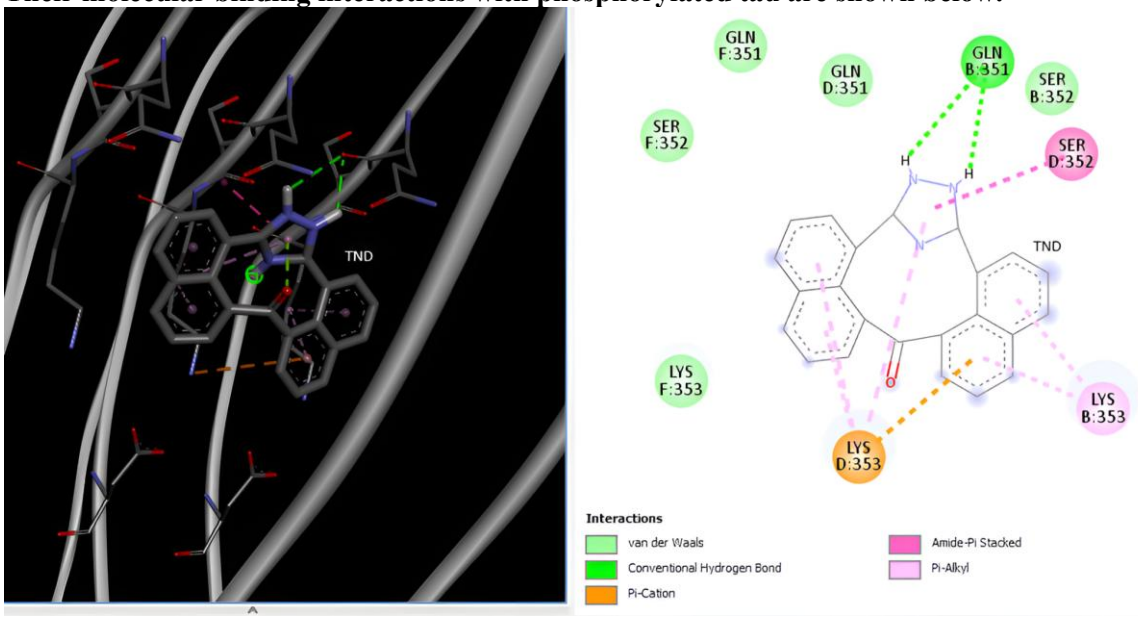

Figure 4: 2D and 3D docking interaction of TND with phosphorylated tau from Discovery Studio Visualizer

(A) 3D representation of the ligand–protein complex showing TND (gray sticks) bound within the active site pocket. Key interactions such as hydrogen bonds, π–cation, and hydrophobic contacts are highlighted with dashed lines.

(B) 2D interaction map generated in Discovery Studio Visualizer showing conventional hydrogen bonding, π–cation, π–alkyl, and van der Waals interactions between TND and the surrounding residues in the binding pocket.

According to the data from Fig 4 , TND exhibits strong binding interactions with key residues in the active site of phosphorylated tau protein. Notably, conventional hydrogen bonds are formed between the ligand and GLN B:351. In addition to hydrogen bonding, TND engages in several π-based interactions. LYS D:353 participates in a π–cation interaction, while LYS B:353 and SER D:352 are involved in π–alkyl and amide–π stacked interactions. These contacts are facilitated by the ligand's aromatic core, promoting enhanced electronic complementarity and structural stability.

Van der Waals interactions with residues such as GLN D:351, GLN F:351 , SER B:352 SER F:352 and LYS F:353 further stabilize the ligand conformation and support its retention within the active site. Collectively, these non-covalent interactions reflect the binding potential of TND and support its relevance as a candidate molecule in the development of tau aggregation inhibitors .

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

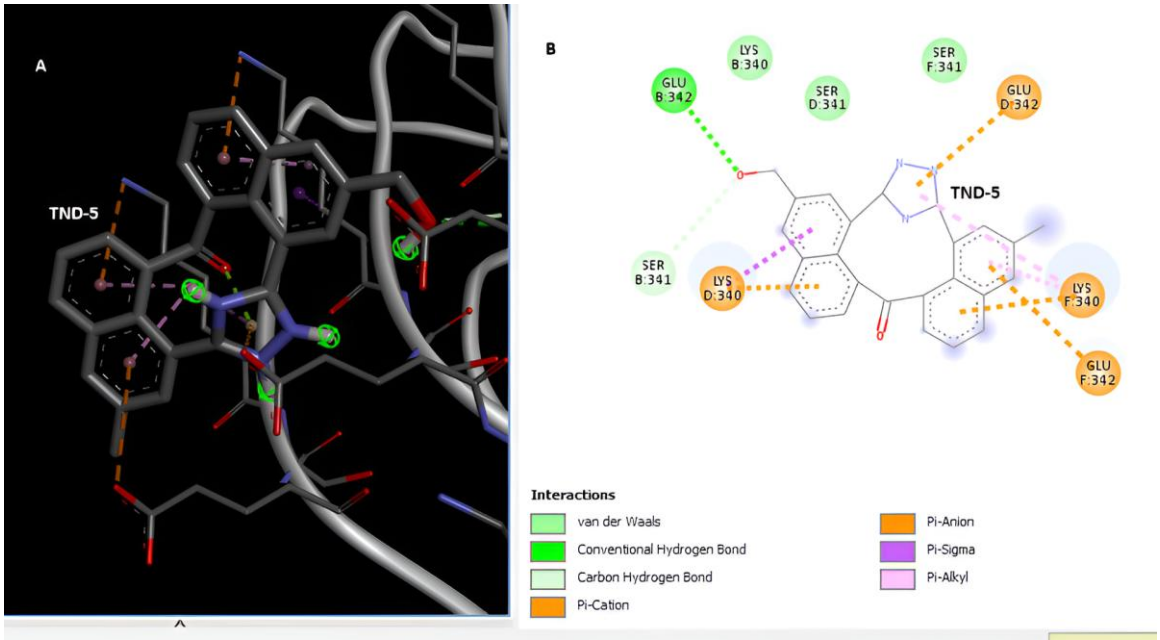

Figure 5: 2D and 3D docking interaction of TND-5 with phosphorylated tau from Discovery studio visualizer

(A) 3D representation of the ligand–protein complex showing TND-5 (gray sticks) bound within the active site pocket. Key interactions such as hydrogen bonds, π–cation, and hydrophobic contacts are highlighted with dashed lines.

(B) 2D interaction map generated in Discovery Studio Visualizer showing conventional hydrogen bonding, π–cation, π–alkyl, and van der Waals interactions between TND-5 and the surrounding residues in the binding pocket.

According to the data from Fig 5 TND-5 demonstrates strong binding interactions with key residues in the active site of phosphorylated tau protein. Notably, conventional hydrogen bonds and carbon hydrogen bonds are observed between the ligand and GLU B:342 and SER D:341, which play a significant role in anchoring the compound within the binding pocket.

Additionally, several electrostatic and π-based interactions are present. GLU D:342, GLU F:342, LYS F:340, and LYS D:340 engage in π–anion π–cation interactions, while π–alkyl and π–sigma stacking interactions are detected between the aromatic system of TND-5 and nearby residues. These interactions contribute to further stabilization of the ligand through favorable electronic and hydrophobic contacts.

Surrounding residues such as SER F:341, SER D:341 and LYS B:340 also participate in van der Waals interactions, reinforcing the ligand's positioning within the pocket. Collectively, these non-covalent forces support the high docking affinity and pharmacological relevance of TND-5 as a potential inhibitor of tau aggregation.

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

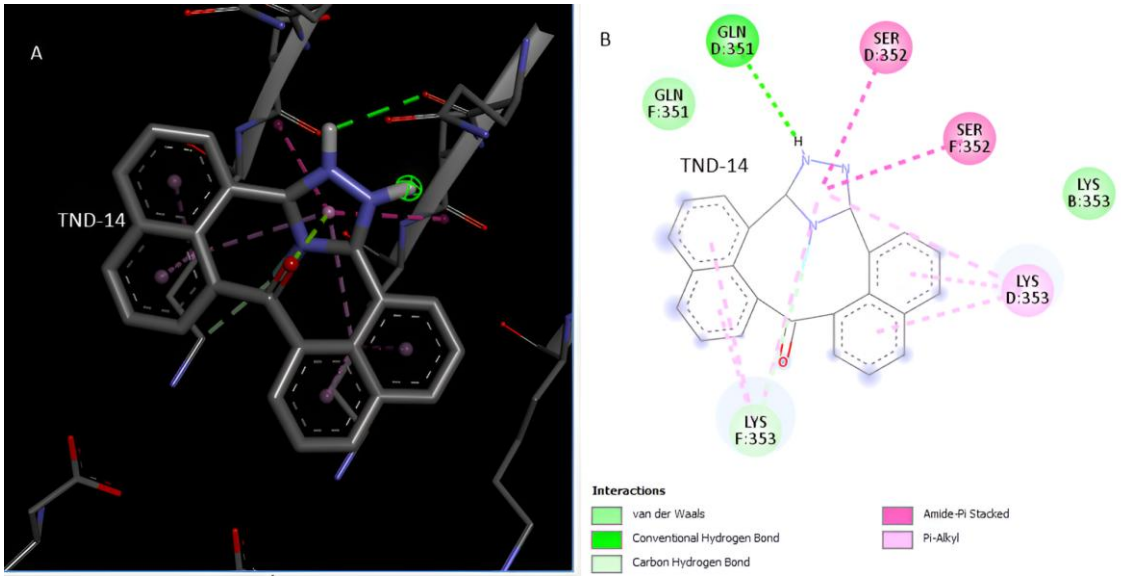

Figure 6: 2D and 3D docking interaction of TND-14 with phosphorylated tau from Discovery Studio Visualizer.
(A) 3D representation of the ligand–protein complex showing TND-14 (gray sticks) occupying the active site cleft of phosphorylated tau. Key interactions such as hydrogen bonds, π–alkyl, amide–π stacking, and van der Waals contacts are visualized using dashed lines.
(B) 2D interaction map generated in Discovery Studio Visualizer showing conventional hydrogen bonding, π–cation, π–alkyl, and van der Waals interactions between TND-14 and the surrounding residues in the binding pocket.

As illustrated in Figure 6, TND-14 forms a well-anchored complex within the phosphorylated tau binding pocket, supported by a network of diverse non-covalent interactions. Two strong conventional hydrogen bonds are observed between the ligand and GLN D:351 , contributing to directional stability within the binding site. In addition, LYS F:353 and LYS D:353 establish π–alkyl interactions with the aromatic rings of TND-14, enhancing hydrophobic engagement and reinforcing the ligand's conformational fit.

Amide–π stacking interactions with SER D:352 and SER F:352 provide further electronic stabilization, suggesting favorable alignment with key residues involved in tau self-assembly. Van der Waals contacts, particularly with LYS B:353, GLN F:351 support shape complementarity and fine-tune the binding orientation. Collectively, these interactions explain the strong docking score of TND-14 (−8.8 kcal/mol) and its pharmacological potential as a CNS-active tau aggregation inhibitor.

Limitations

QSAR extrapolation at the frontier. All sixteen triazole-naphthalene derivatives display leverages that exceed the global threshold and Mahalanobis distances beyond the 95 % χ^2 ellipsoid. Although a 5-NN density test indicates acceptable local support, the predicted pIC_{50} values must be regarded as exploratory and subject to larger error bars than those implied by the external-test RMSE.

Single-assay noise. The training IC_{50} data were compiled from multiple laboratories but ultimately derive from variations of the thioflavin-T aggregation assay; inter-lab variance is typically on the order of 0.5–0.6 log-units. Any systematic bias in those measurements propagates directly into the model.

Docking score uncertainty. Binding affinities were estimated with AutoDock Vina using a rigid tau-fibril receptor; absolute ΔG values can deviate by several kcal mol^{-1} from experiment, and protein flexibility is not accounted for. Docking results therefore serve only as a rank-ordering heuristic, not a quantitative predictor of IC_{50}.

In-silico ADMET filters. SwissADME outputs are statistical in nature; they cannot guarantee CNS penetration or metabolic stability in vivo.

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

No experimental confirmation (yet). The study is entirely computational. Wet-lab synthesis and biochemical testing of the top-ranked TNDs are currently underway; model re-training with those data will be required to refine predictivity.
Single receptor conformation. Docking utilised one cryo-EM tau fibril structure; alternative polymorphs may alter ligand ranking and binding modes.

Conclusion

In this study, we have demonstrated an integrated, fully reproducible in silico workflow for discovering novel tau-aggregation inhibitors. By curating the largest publicly available dataset of human-tau IC_{50} values and training a rigorously validated stacked-ensemble QSAR model (SVR + RF + XGBoost), we achieved strong internal ($Q^2 = 0.63$) and external ($R^2 = 0.57$, RMSE = 0.73 log-units) predictivity. Applicability-domain analyses (global leverage, 5-NN density, UMAP) confirmed that all sixteen designed 1,2,4-triazole–naphthalene derivatives lie within or at the periphery of the model's reliable space. Prospective predictions identified four sub-40 nM candidates (notably TND-9, TND-15, and TND-5), which also show favorable BBB-penetration profiles in BOILED-Egg and logBB analyses. Orthogonal fibril docking against the 6HRF polymorph further prioritized TND-5 for its sub-micromolar binding affinity and balanced ADMET properties. Together, these results highlight the power of combining ligand-based QSAR, structure-based docking, and developability filters into a cohesive pipeline—offering a blueprint for rapid lead prioritization in Alzheimer's drug discovery. Future work will focus on experimental validation of the top candidates and iterative model refinement using the resulting bioactivity data.

Reference

1. van Dyck CH, Swanson CJ, Aisen P, Bateman RJ, Chen C, Gee M, Kanekiyo M, Li D, Reyderman L, Cohen S *et al.* (2023) Lecanemab in early Alzheimer's disease. *N Engl J Med* 388:9–21. https://doi.org/10.1056/NEJMoa2212948

2. Gulisano W, Maugeri D, Baltrons MA, Fà M, Amato A, Palmeri A, D'Adamio L, Grassi C, Devanand DP, Honig LS, Puzzo D, Arancio O (2018) Role of amyloid-β and tau proteins in Alzheimer's disease: confuting the amyloid cascade. *J Alzheimers Dis* 64(Suppl 1):S611–S631. https://doi.org/10.3233/JAD-179935

3. Majewski J, Jones EM, Vander Zanden CM, Biernat J, Mandelkow E, Chi EY (2020) Lipid membrane templated misfolding and self-assembly of intrinsically disordered tau protein. *Sci Rep* 10:13324. https://doi.org/10.1038/s41598-020-70208-6

4. Manglano-Artuñedo Z, Peña-Díaz S, Ventura S (2024) Small molecules to target tau amyloid aggregation. *Neural Regen Res* 19:509–511. https://doi.org/10.4103/1673-5374.380900

5. Bulic B, Pickhardt M, Schmidt B, Mandelkow E-M, Mandelkow E (2009) Development of tau aggregation inhibitors for Alzheimer's disease. *Angew Chem Int Ed* 48:1740–1752. https://doi.org/10.1002/anie.200802621

6. Gholampour M, Seradj H, Sakhteman A (2023) Structure–selectivity relationship prediction of tau imaging tracers using machine-learning-assisted QSAR models and interaction fingerprint map. *ACS Chem Neurosci* 14:1490–1502. https://doi.org/10.1021/acschemneuro.3c00038

7. Noviandy TR, Maulana A, Idroes G, Emran T, Tallei T, Helwani Z, Idroes R (2023) Ensemble machine learning approach for quantitative structure–activity relationship based drug discovery: a review. *Infolitika J Data Sci* 1:91. https://doi.org/10.60084/ijds.v1i1.91

8. Ghorbani H (2019) Mahalanobis distance and its application for detecting multivariate outliers. *Facta Univ Ser Math Inform* 34:583–595. https://doi.org/10.22190/FUMI1903583G

The Research Analytics ISSN ISSN (Online): 3107-6165
Volume 1, Issue 3, July-Sept 2025

9. Abdelli A, Azzouni S, Plais R, Gaucher A, Efrit ML, Prim D (2021) Recent advances in the chemistry of 1,2,4-triazoles: synthesis, reactivity and biological activities. *Tetrahedron Lett* 86:153518. https://doi.org/10.1016/j.tetlet.2021.153518

10. Saragi RT, Calabrese C, Juanes M, Pinacho R, Rubio JE, Pérez C, Lesarri A (2023) π-Stacking isomerism in polycyclic aromatic hydrocarbons: the 2-naphthalenethiol dimer. *J Phys Chem Lett* 14:207–213. https://doi.org/10.1021/acs.jpclett.2c03299

11. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461. https://doi.org/10.1002/jcc.21334

12. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* 7:42717. https://doi.org/10.1038/srep42717

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

14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V **et al.** (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830.

15. Sahigara F, Ballabio D, Todeschini R, Consonni V (2013) Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J Cheminform* 5:27. https://doi.org/10.1186/1758-2946-5-27

16. Healy J, McInnes L (2024) Uniform manifold approximation and projection. *Nat Rev Methods Primers* 4:82. https://doi.org/10.1038/s43586-024-00363-x

17. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 30:2785–2791. https://doi.org/10.1002/jcc.21256

18. Baroroh U, Muscifa ZS, Destiarani W, Rohmatullah FG, Yusuf M (2023) Molecular interaction analysis and visualization of protein–ligand docking using BIOVIA Discovery Studio Visualizer. *Indones J Comput Biol* 2(1):e46322. https://doi.org/10.24198/ijcb.v2i1.46322

19. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17:4791–4810. https://doi.org/10.3390/molecules17054791

20. Prasanna S, Doerksen RJ (2009) Topological polar surface area: a useful descriptor in 2D-QSAR. *Curr Med Chem* 16:21–41. https://doi.org/10.2174/092986709787002817

Data availability. All raw data, processed descriptor tables, scripts and supplementary files are freely available at **https://github.com/Rym174/tau**