Chapter 6

Chapter 6

Artificial Intelligence Assisted Identification of Potential Tau Aggregation Inhibitors

6.1. Introduction

AI has emerged as a powerful tool in several domains, and its application in the pharmaceutical industry has been pivotal for enhancing and streamlining drug discovery processes. In particular, AI-based technologies, such as ML and DL, have been employed to improve the efficiency of various stages involved in drug discovery, making it a cost-effective and time-efficient process [1, 2]. These technologies have enabled researchers to undertake complex tasks, including quantum mechanical calculations for identifying the properties of compounds, predicting protein structures, evaluating proteins, and designing retro-synthetic pathways for synthesis planning [3, 4]. The significance of AI in drug discovery is summarized in **Chapter 1**, **Section 1.10.3.**

6.2. Objectives

In traditional drug discovery methodologies, the processes are characterized by their extensive duration, significant financial implications, and intensive labor requirements. Considering the protracted nature of hit identification through the conventional design-make-test-analyze approach, this study aims to harness the capabilities of artificial intelligence to pinpoint potential inhibitors of tau aggregation.

The objectives of this study are as follows

- **Ligand based virtual screening**: To utilize the AI-assisted PyRMD ligand-based virtual screening tool to pinpoint potential active tau inhibitors, leveraging a comprehensive screening of expansive databases using training set molecules derived from the CheMBL database.

- **REMD Simulation**: To explore the three-dimensional (3D) structure of tau through replica exchange molecular dynamics simulations. This investigation is particularly challenging due to tau's intrinsic disordered nature, which lacks well-defined conformations.
- **Active site Identification**: To identify the active binding site within the three-dimensional conformation of tau generated through REMD simulations. This binding pocket will serve as the target for molecular docking with active ligands obtained from PyRMD screening.
- **Structure based Virtual screening:** To perform Molecular docking of active ligands against different clusters of tau to evaluate the binding mode of the compound with the targets.
- **In silico ADME:** To evaluate the physicochemical properties of prominent compounds that bind to various clusters of tau, with the aim of determining their optimal drug-like profiles.
- **Molecular Dynamics and Binding Free energy calculation:** To conduct molecular dynamics simulations and compute the binding free energy of the top compounds, aiming to assess the dynamic interaction between the protein and the compounds as well as the stability of the resultant complex.

6.3. Results and discussion

6.3.1. Artificial Intelligence Based Virtual Screening

PyRMD uses the Random matrix discriminant (RMD) algorithm implemented in Python as an ML-based tool for screening large databases of small molecules, by making use of the associated biological activity data [3]. In this study, we used PyRMD for performing ligand-based virtual screening to identify molecules that could be active against microtubule-associated tau protein (MAPT) aggregation. The activity data of small molecules against

MAPT downloaded from the ChEMBL database was used to train a model for the RMD-based algorithm [5]. The molecules were classified as active and inactive based on the given threshold values and later these molecules underwent a featurization process in which the molecular fingerprints were extracted from the smile structures of the molecules. This was followed by the actual screening, using the RMD fitting process, of 12 million molecules containing library downloaded from the ZINC database [6]. The screening process was completed with the algorithm giving an output of the SMILES structures and RMD scores of the active molecules. A higher RMD score indicates that the molecule is likely to be active.

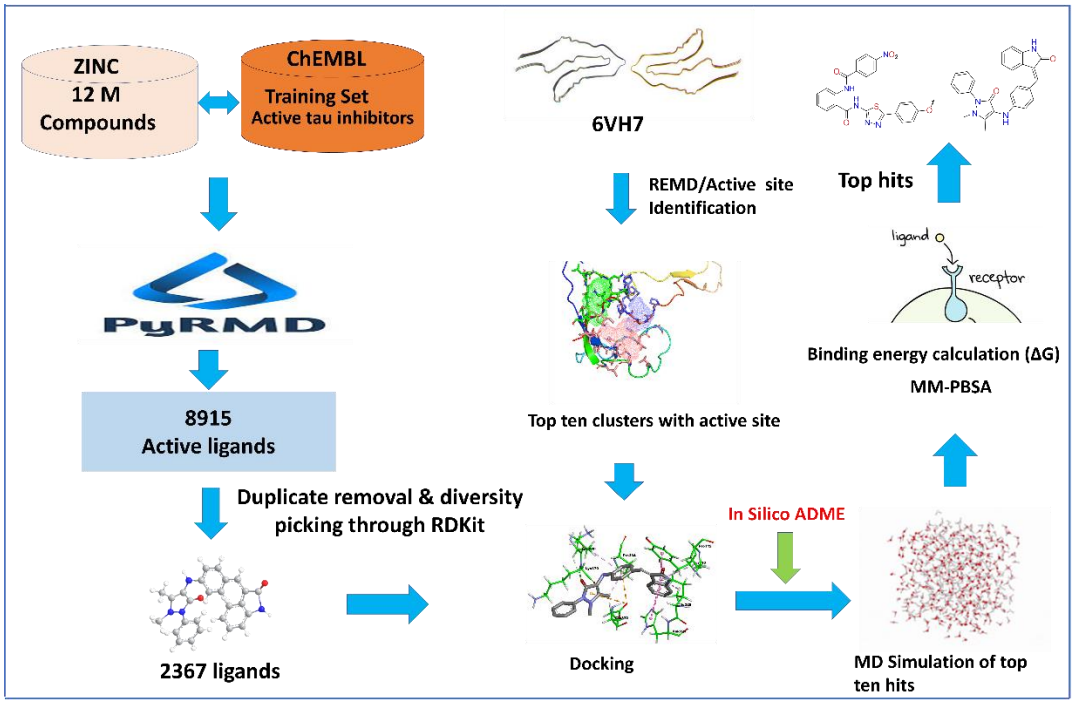

Figure 6.1. Schematic representation of virtual screening workflow showing identification of active ligands through AI-based virtual screening tool PyRMD. PyRMD screened 12 million compounds from the ZINC database and provided 8915 as active ligands based on the active tau inhibitors as training set molecules obtained from the ChEMBL database. After diversity picking and duplicate removal, the 2367 compounds were docked into the binding pocket of ten tau protein conformations obtained from REMD simulation. Further, the top ten compounds selected based on their docking score and optimal ADME parameters

were subjected to molecular dynamics simulation as well as binding free energy calculation through MMPBSA to determine the stability of the complex.

Our screening of 12 million molecules for activity against MAPT aggregation resulted in the identification of 15397 active molecules along with their corresponding RMD score. The output of the PyRMD screening also tags the active molecules as either PAINS active or PAINS inactive. Those which are PAINS active are molecules having false activity against most of the biological targets due to the existence of certain reactive functional groups [7]. Hence those molecules which were found to be PAINS active were removed and we were left with 8915 active molecules. Later in order to remove molecules that are the same or have similar chemical features, we performed a diversity-picking procedure for the 8915 active molecules. This ensured that we have sufficiently diverse molecules and those which have similar chemical features or overlapping structures were removed from further analysis. After the diversity picking using RDKit, implemented in KNIME software, we had 2367 active molecules representing a diverse chemical space [8, 9]. The entire workflow is depicted in **Figure 6.1**

6.3.2. Replica Exchange Molecular Dynamics (REMD) simulation and clustering of tau structures

Tau being an intrinsically disordered protein doesn't have a well-defined tertiary structure under physiological conditions [10]. Most of the secondary structural elements of this protein is identified in the microtubule-binding region, formed when the protein interacts with the cytoskeletal fragments [11]. It has been found by various experimental studies that the microtubule-binding region of the protein is involved in misfolding and aggregation of tau [12]. Certain regions of the microtubule-binding region have been studied in detail via experimental techniques and have been found to play role in the misfolding process. Most of the available structures of the tau microtubule-binding region are a short stretch of the peptide

sequence. Owning to the highly dynamic nature of the region of the protein it is not a reliable strategy to perform structure-based drug design approaches on the available tau fragments. Hence it is important to study the dynamic changes in tau to develop drug molecules which can inhibit the aggregation process. For this purpose, one of the most useful techniques is to use MD simulation. Various studies have been conducted which make use of MD simulations to study the folding patterns of tau under various conditions of post-translational modifications like phosphorylation, O-Glyc, N-Acylation, N-glycosylation, etc. [13-16]. But the use of classical all-atom MD simulations might not be enough to capture all different conformations of tau. Hence in this study, we used Replica Exchange Molecular Dynamics **(**REMD) simulations to model and simulate the tau microtubule-binding region. REMD being an enhanced sampling technique helps us to capture almost all the different conformations that the peptide sequence can take up [17, 18]. Here we simulated the tau in different replicas each corresponding to different temperature and the peptide structure is exchanged between different temperatures, making sure that the peptide can now sample higher energy states, which were forbidden due to large energy differences between some of the conformations. We used the temperature scale from 310 K to 400K using thirty-one replicas between them. After 100 ns REMD simulation, we examined the exchange probabilities, and the average exchange probability was ~24%, which was well within the accepted range making sure that each replica had sufficient overlap with one another.

We analysed the evolution of temperature across the replicas during the simulation and found that each replica has explored all temperature scales. The replicas are distributed widely across all temperatures. We also analysed the potential energy of each of the replicas along the simulation time and found that the potential energy has sufficient overlap with one another (**Figure 6.2B**). Sufficient overlap between potential energy indicates that the simulation was

able to capture most of the conformations of the tau peptide by transversing across higher temperatures and energy states which were forbidden initially due to high energy barriers.

The 310 K temperature replica was used for analysing the REMD simulation as it closely resembles the human body temperature. The raw simulation data were concatenated and conformational changes of the protein along the 100 ns were mapped. In order to check for any major structural deformities during the course of the simulation, we calculated the root mean square deviation (RMSD) for the protein backbone (**Figure 6.3A**) [19]. And, we found that the system gets stabilized after the initial 25 ns, after which the RMSD fluctuates between 2.2 and 2.7 nm. The RMSD plot suggests that after the equilibration, the protein gets stabilized and remains so for the rest of the trajectory.

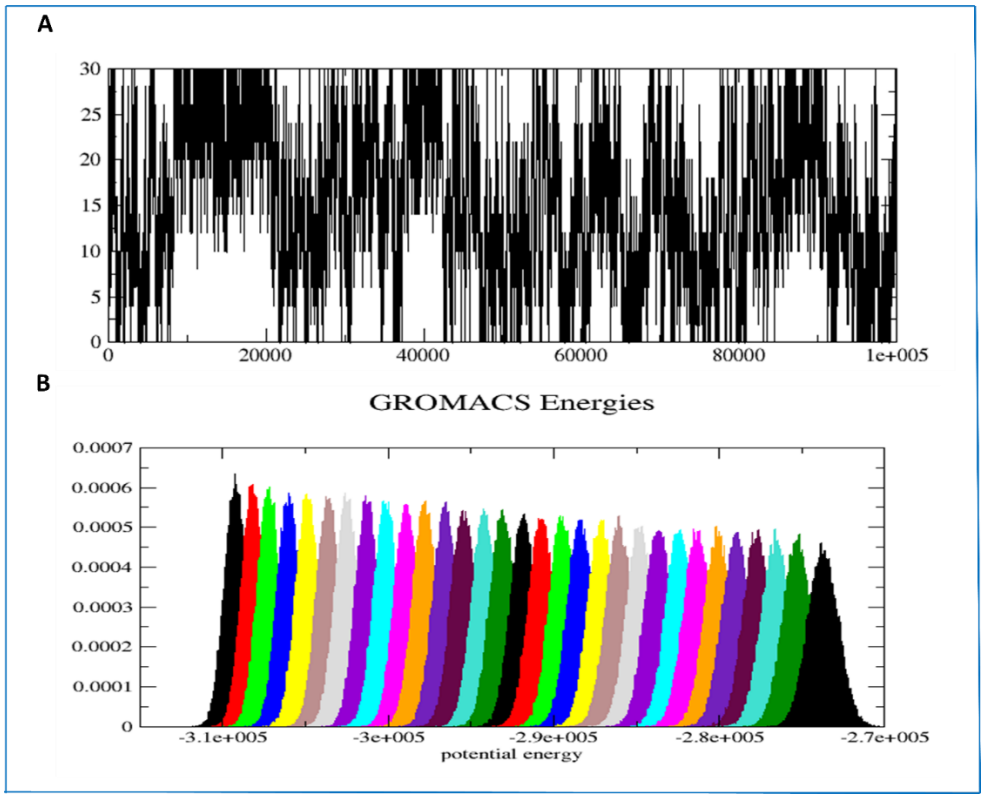

Figure 6.2. The diagram for the evolution of temperature across the replicas and the potential energy of each replica during the simulation time. (A) Shows the evolution of temperature across the replicas during the simulation time scale. Here we can see that each

replica spends sufficient time across all thirty-one different temperatures. **(B)** The potential energy of each replica across the simulation time**.** We can observe the overlap of potential energy between adjacent replicas, indicating the overlap of conformation of the peptide between temperatures during the simulation.

A slightly high RMSD value (2.2-2.7 nm) is due to the highly flexible nature of the tau protein. Later we analyzed the protein compactness throughout the trajectory using the radius of gyration. The radius of gyration helps us gain insights into the conformational preference of the protein, with a high value indicating an extended conformation, and a low value indicating a more compact protein conformation [20].

Figure 6.3B indicates that initially, the protein was in an extended conformation, with a high radius of gyration, but later reached a lower stable value and remained in a more compact conformation. **Figure 6.3C** shows the probability distribution of the radius of gyration, again indicating a single major peak corresponding to the most populated conformation having a radius of gyration 1.7 nm and another small peak at 1.5 nm. Analysis of representative structures indicates that these conformations are rather compact and have certain folded elements. We also analysed the end-to-end distance of the protein throughout the trajectory (**Figure 6.3D and Figure 6.3E**).

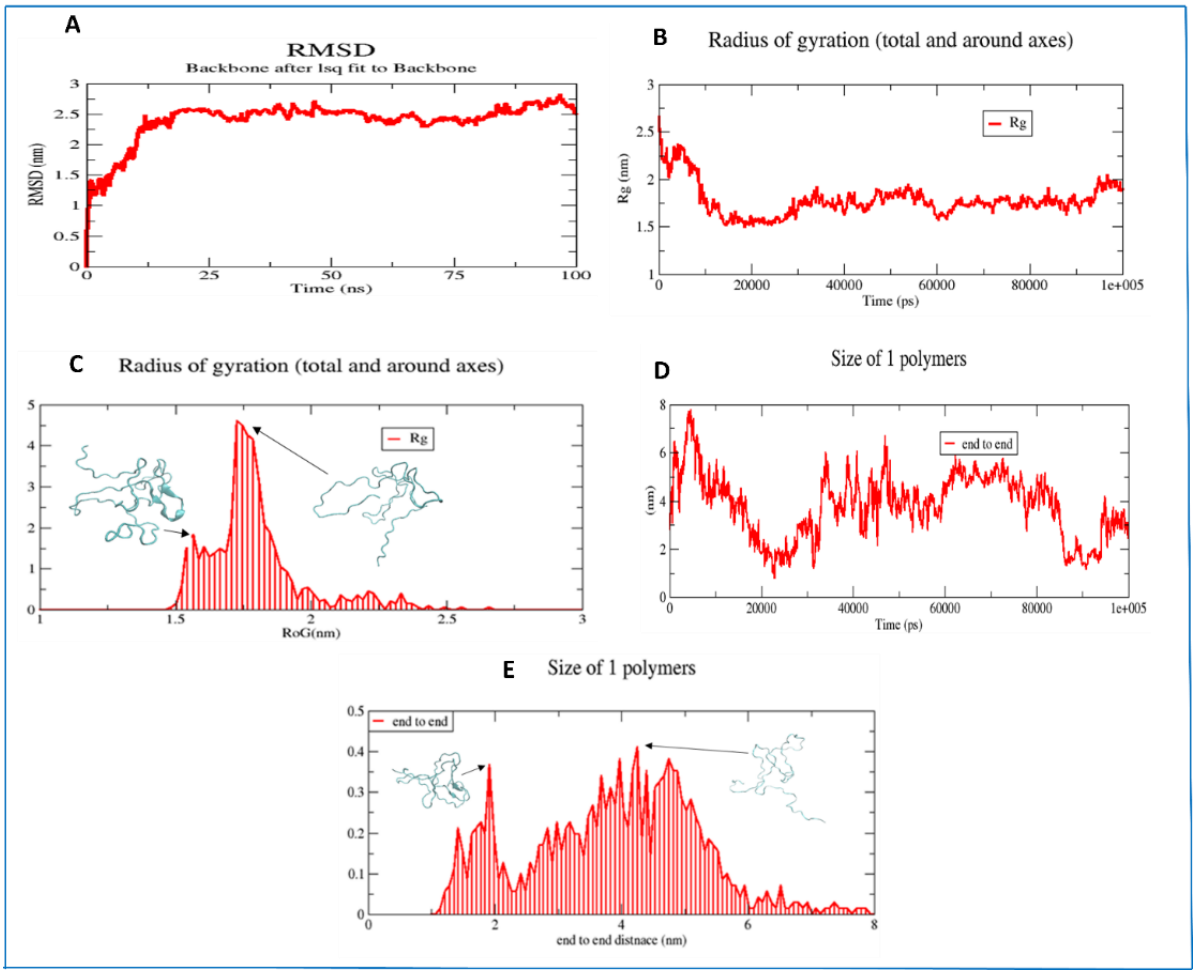

Figure 6.3. Tau peptide REMD trajectory Analysis. (A) Root mean square deviation (RMSD) for the protein backbone, **(B)** radius of gyration**, (C)** probability distribution of the radius of gyration, **(D)** end-to-end distance fluctuation throughout the trajectory, **(E)** probability distribution of end-to-end distance.

The end-to-end distance measures the average distance between the two ends of the protein and gives an indication about the protein conformation as a high value is indicative of an extended non-folded conformation while that of a smaller value might be indicative of compact and folded form [21]. Here we can see that the end-to-end distance fluctuates throughout the trajectory and the probability distribution indicates a narrow peak near 2 nm and a very broad cluster of peaks in the range of 3.5 nm -5 nm.

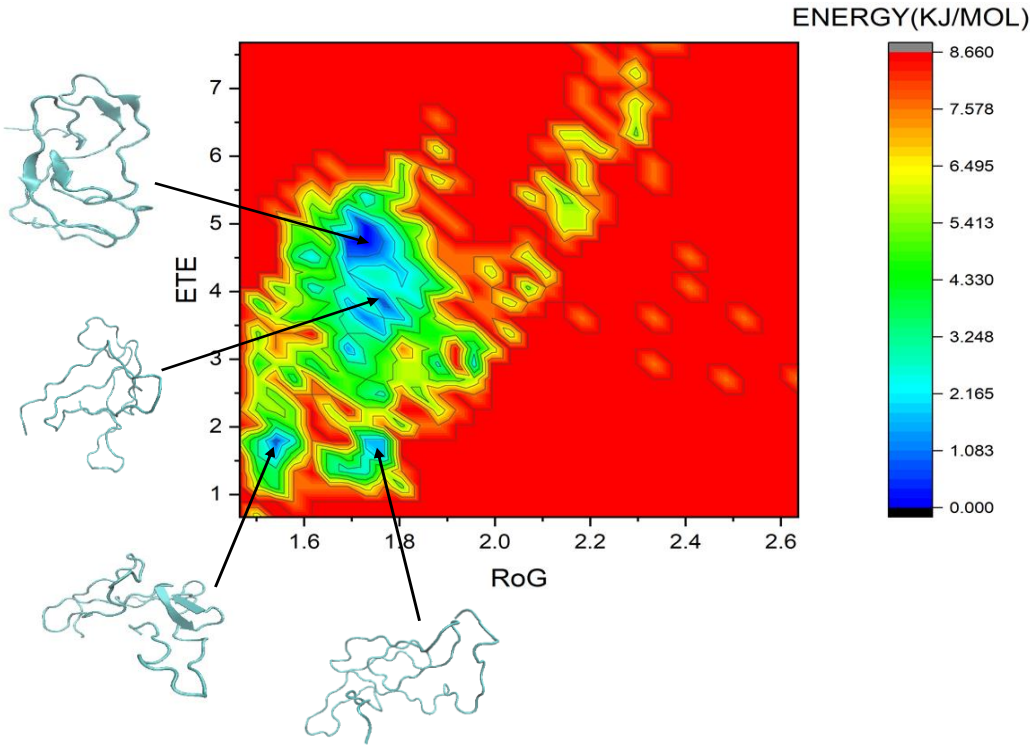

Figure 6.4. 2D free energy landscape diagram of tau peptide. Here we can see the rugged folding funnel of the peptide which mimics that of an IDP having multiple local energy minima, separated by low energy barriers separating them.

The representative structures show a more folded and compact structure near the 2 nm peak while a more open and extended conformation near the 4 nm peak. Free energy funnels are an important aspect of folding process of proteins, especially proteins having well defined structures. But with the case of intrinsically disordered proteins like tau, the folding funnels is usually rugged, having multiple local minima peaks which are separated from each other by a very small energy barrier such that each of those local minima can overcome those small energy barriers. This suggest that multiple conformations of the same protein can exist together as they are highly inter-convertible because of the small energy barrier [22-24]. Here we tried to analyse the free energy landscapes of the REMD simulation using the gmx sham module of GROMACS. This technique uses any two reaction coordinates from the simulation data, in this case we have used the end-to-end distance and radius of gyration, and makes a

multidimensional histogram. By Boltzmann inverting this multidimensional histogram, the free energy landscapes are plotted. From the 2D representation of the free energy landscape of the tau (**Figure 6.4**), we can see that there are multiple local energy minima i.e. rugged free energy landscapes, which are separated by small energy barriers. We can also find a few local energy minima that are well separated from others by a comparatively high energy barrier. The energy barrier between the local energy minima is within the range of 2-4 kJ/mol. This is not a very large barrier for the small tau peptide and suggests that the conformations get interconverted and multiple lower energy conformations exist for the protein. Analysis of some of the local energy minima suggests the presence of secondary structural features like beta sheets in some of the energy minima structures. The free energy landscapes again indicate the folding pattern of an intrinsically disordered protein existing in multiple highly interconvertible energy conformations.

6.3.3. Active site Identification

After obtaining the tau monomer structures from the REMD simulation study, active sites of each cluster were determined by using FTmap and FTSite web server [25]. FTmap is a well-known computational mapping server that identifies hotspots of the macromolecule that are domains of the surface with important contributions to the ligand-binding free energy [26, 27]. FT site web server determines the ligand binding site in a macromolecule. So, by using these web servers we determined the active sites for each tau cluster and each cluster has been found to contain three numbers binding sites that are represented in **Figure 6.5**.

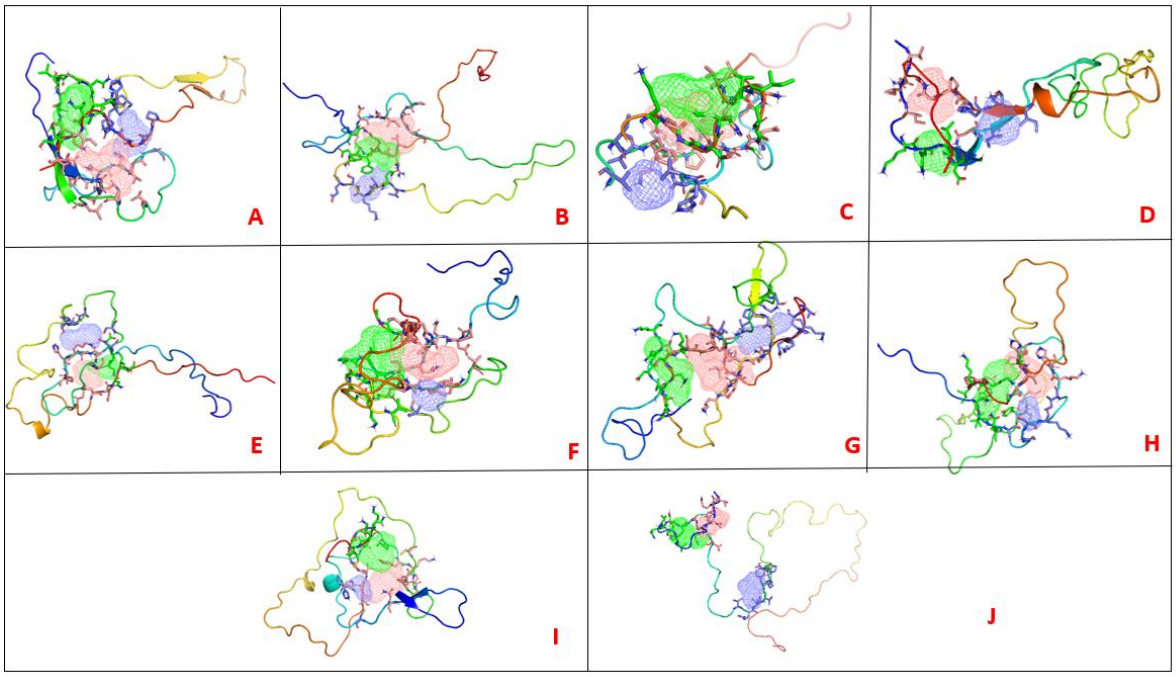

Figure 6.5. Ten clusters of tau (A to J), showing active sites represented in the red, blue, and green mesh-like spheres.

6.3.4. Structure-Based Virtual screening

The remaining 2367 molecules after diverse-picking were further docked to the binding pockets of different clusters (CA-CJ) by using the Autodock Vina tool. Then we picked up the top 100 ligands with their first pose from each cluster (total of 1000 poses) and further we checked the number of ligands from 1000 poses that showed a binding affinity for more than one cluster. The number of molecules was calculated binding to more than one cluster summarized in the **Table 6.1**. We selected the top thirty-three ligands binding to six or more. Then in silico pharmacokinetic screening was performed for thirty-three ligands and the top ten molecules were selected for further study. The two-dimensional structure of the top ten compounds is given in **Figure 6.6**

Table 6.1. The number of compounds from 1000 ligands showed binding interaction with more than one cluster.

Number of Compounds	**Number of clusters**
210	≥ 2
137	≥ 3
80	≥ 4
54	≥ 5
33	≥ 6
21	≥ 7
13	≥ 8
2	≥ 9

Among the top scoring molecules, **UNK_1172** showed the highest binding affinity to cluster F and the remaining compounds **UNK_175, UNK_298, UNK_1027, UNK_1172, UNK_1173, UNK_1179, UNK_1518, UNK_1778,** and **UNK_2181** showed the highest binding affinity to cluster H. **Table 6.2** shows the binding values of the top ten ligands to cluster A to cluster J.

Table 6.2. List of top ten ligands with their highest docking binding affinity to ten clusters.

Compound	**Docking score (kcal/mol) for clusters A-J**									
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
UNQ_175	-6.6	-7.8	-7.8	-6.3	-3.8	-9.8	-8.1	-10.1	-7.7	-7.9
UNQ_298	-6.9	-8.2	-8.1	-6.9	-3.7	-9.0	-8.1	-9.8	-7.7	-7.9
UNQ_1027	-6.9	-7.4	-7.8	-6.2	-3.8	-8.4	-8.2	-9.5	-7.6	-6.8
UNQ_1172	-6.8	-7.5	-6.9	-6.4	-4.0	-8.9	-8.1	-8.2	-7.3	-7.4
UNQ_1173	-6.2	-7.7	-6.5	-6.7	-4.1	-9.1	-8.1	-9.4	-7.9	-6.8
UNQ1179	-6.6	-7.5	-7.1	-6.4	-4.0	-8.7	-7.9	-10.1	-7.2	-7.2
UNQ_1237	-6.5	-7.9	-7.3	-6.6	-3.8	-9.9	-8.3	-10.3	-7.6	-7.4
UNQ_1518	-6.5	-8.1	-7.8	-6.2	-3.6	-9.2	-7.7	-9.7	-7.0	-6.9
UNQ_1778	-6.6	-7.3	-7.2	-6.0	-3.7	-9.0	-8.0	-9.5	-7.2	-6.8
UNQ_2181	-6.5	-7.1	-9.6	-6.2	-3.7	-8.4	-7.9	-9.6	-7.2	-6.6

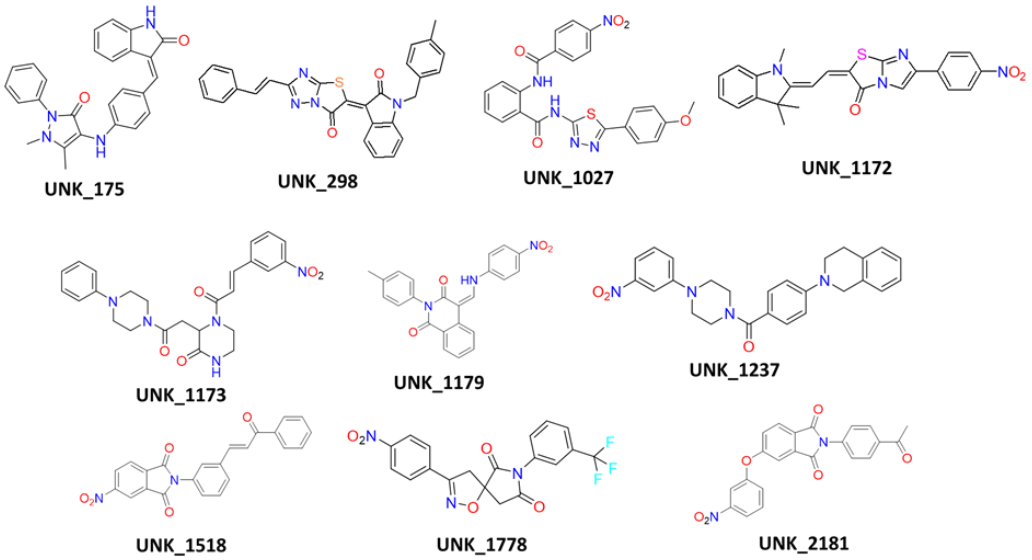

Figure 6.6. Two-dimensional structure of the top ten hit compounds identified by ligand and structure-based virtual screening.

The 2D and 3D interaction of **UNK_175** with cluster H **(Figure 6.7A)** revealed the indoline ketone undergoes π-donor hydrogen bond interaction with HSD329 as well as showed π-π stacking with TYR310 and alkyl interaction with Pro312 residue respectively. The carbonyl oxygen of pyrazole undergoes conventional hydrogen bond interaction and π-donor hydrogen bond interaction with LYS370 and LYS369 residues respectively. The π-electrons of the pyrazole ring undergo π-anion interaction with GLU372 residue and the substituted methyl group of pyrazoles showed alkyl interaction with Pro364 residue.

The binding interaction of ligand **UNK_298 (Figure 6.7B)** showed that indoline undergoes interaction π-π stacking as well as π-π T-shaped interaction with HSD329, GLU372, and TYR310 residues respectively. The 4-methyl benzyl group also formed different alkyl and π-alkyl interactions with the residues like ILE328, VAL287, and ILE371. It also undergoes different π-π stacking with the residues like HSD329 and GLU372. The ligand **UNK_1027 (Figure 6.7C)** showed a wide range of interaction containing thiadiazole ring which showed π-anion interaction with GLU372 and the loan pair on nitrogen atom undergoes conventional hydrogen bond interaction with GLY365 residue respectively. The methyl group of the 4-

methoxy phenyl group undergoes different alkyl and π-alkyl interactions with the residues ILE328, PRO312, VAL287, and TYR310. The π-electrons of three phenyl groups undergo π-anion, π-π stacked interaction with various residues like TYR310, GLU372, HSD362, and VAL300. The nitro group showed hydrogen bond interaction with VAL300 residue.

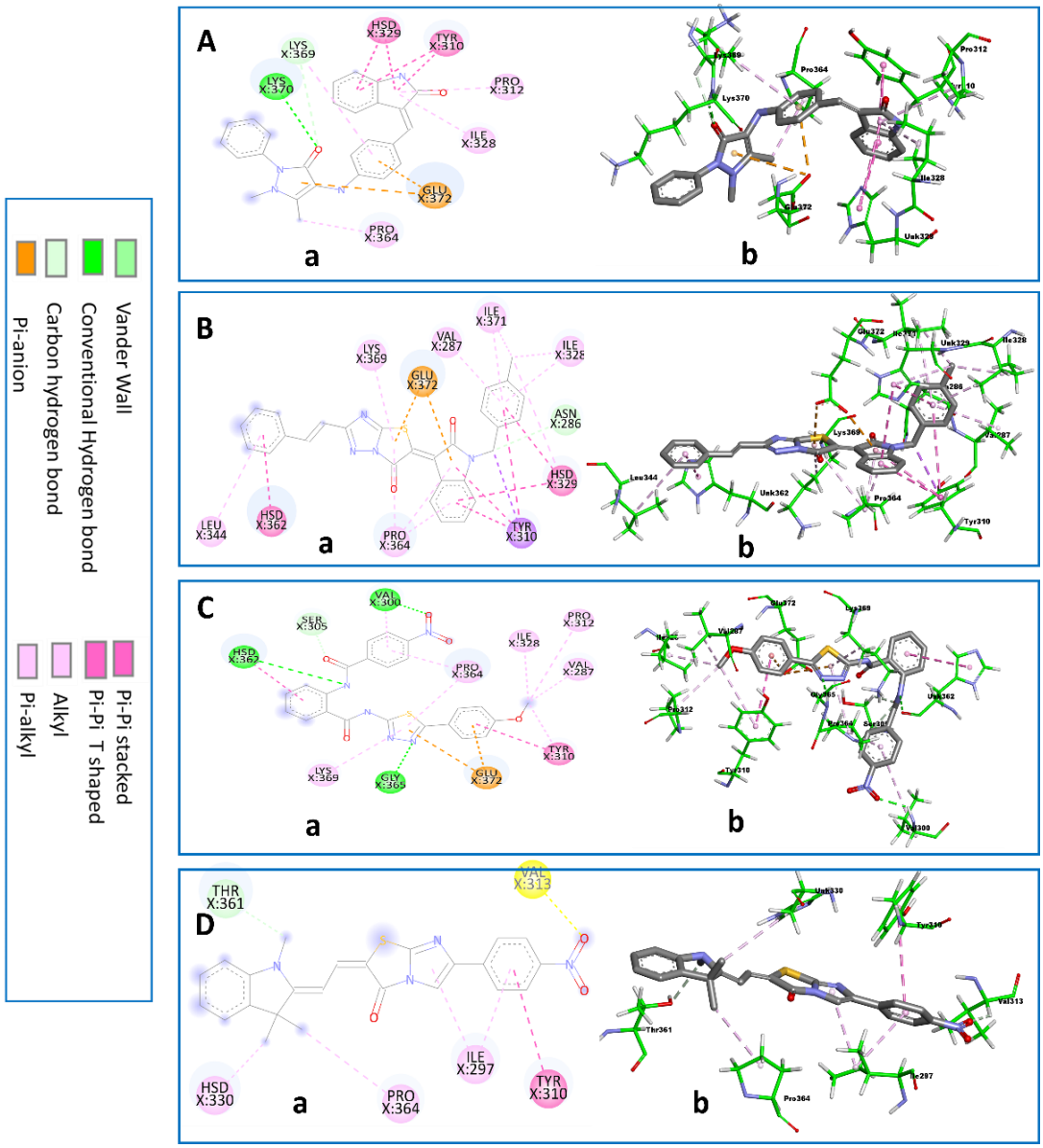

Figure 6.7. 2D (a) and 3D (b) interaction view of the molecules UNK_175 (A), UNK_298 (B), UNK_1027 (C) with clsuter H, UNK_1172 (D) with clsuter F.

The binding interaction of **UNK_1172** with cluster F (CF) **(Figure 6.7D)** showed that the π-electrons of pyrazole and phenyl ring undergo different π-π and π-alkyl interactions with

residues like TYR310, and ILE297. The methyl groups of the indoline moiety formed alkyl interaction with PRO364 and HSD330 and van der Waals interaction with THR361 respectively. The isoquinoline showed pi-anion interaction with GLU372 and pi-pi stacking with TYR310.

The binding prediction of the ligand **UNK_1173** (**Figure 6.8E**) showed that the π-electrons of the phenyl group in the phenyl piperazine moiety undergo π-π stacking and π-π T-shaped interaction with the residues VAL300, SER305, PRO364 residues. The two piperazine groups showed different alkyl and π-alkyl interactions with residues like Pro364, LYS369, and LYS370. The π-electrons of the nitrophenyl group showed π-π stacking with TYR330.

The binding pattern of compound **UNK_1179** (**Figure 6.8F**) revealed that the π-electrons of the phenyl group undergo π-π stacking with TYR310 and VAL287. Similarly, the π-electrons of the nitrophenyl group undergo π-π stacking and π-π T-shaped interaction with residues PRO364, ILE 308, and VAL306. The binding prediction of the ligand **UNK_1237** (**Figure 6.8G**) revealed that the π-electrons of isoquinoline moiety showed π-π stacking and π-π T-shaped interaction with VAL287, HSD329, and TYR310 respectively. The π-electron of phenyl groups showed different π-π and π-anion interactions with the residues like PRO364, LYS369, and GLU372. The nitro group of isoindoline for compound **UNK_1518** (**Figure 6.8H**) showed conventional hydrogen bond interaction with the residues GLY304 and van der Waals interaction with LEU344, and HSD299 respectively. The π-electrons of the phenyl group attached to the isoindoline showed pi-anion interaction with GLU372 and π-alkyl interaction with LYS369 respectively. The nitro group of the nitrophenyl group for compound **UNK_1778** (**Figure 6.9I**) undergoes conventional hydrogen bonding with GLN307 and HSD329 residues. The π-electrons of nitrophenyl groups undergo π-anion interaction with and π- π stacking with GLU372 and TYR310 respectively.

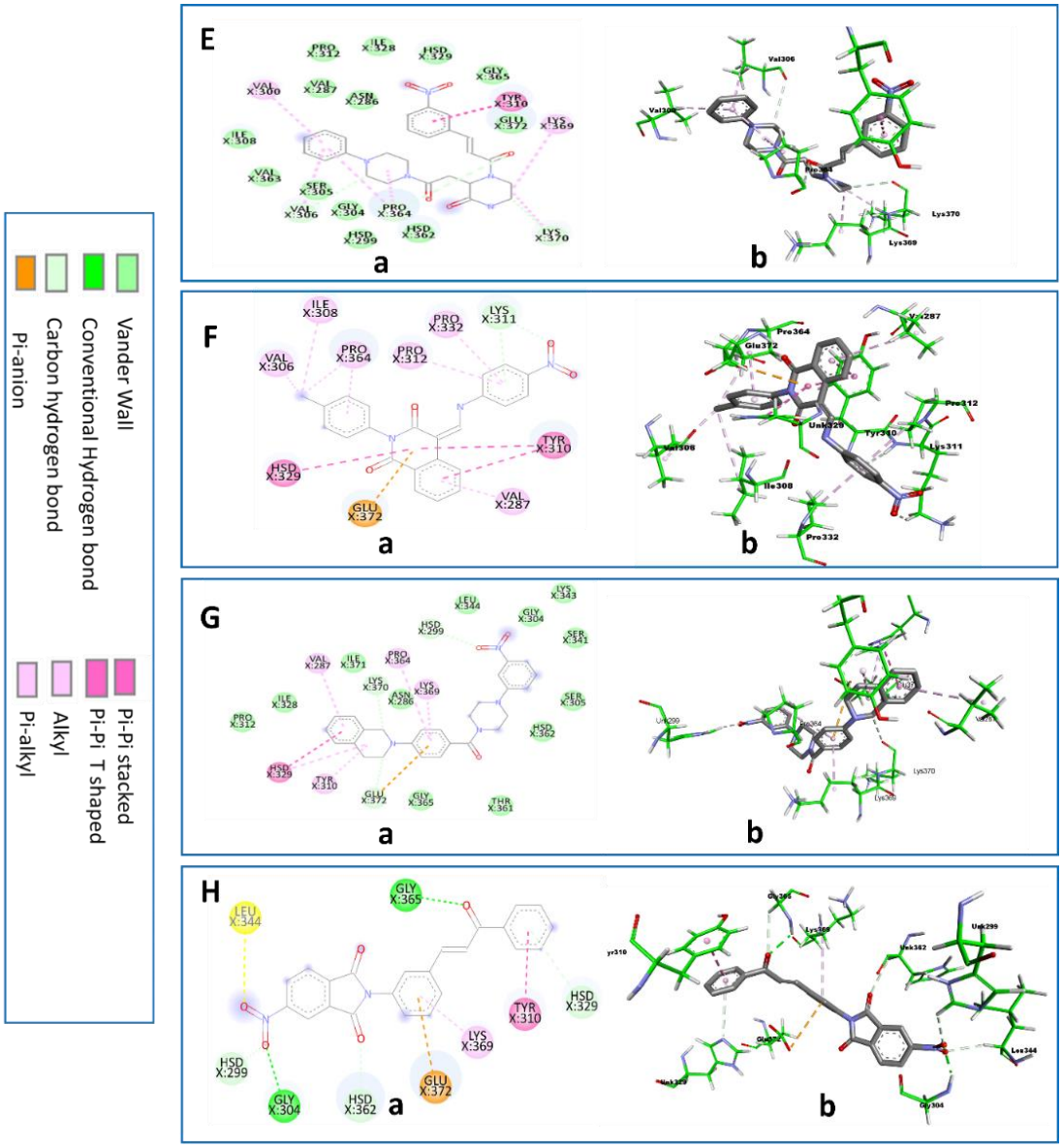

Figure 6.8. 2D (a) and 3D (b) interaction view of the molecules UNK_1173 (E), UNK_1179 (F), UNK_1237 (G), UNK_1518 (H) with Cluster H.

The oxygen atom of oxazoline undergoes conventional hydrogen bonding with LYS370 residue. The isoindoline group of the ligand **UNK_2181** (**Figure 6.9J**) undergoes π-anion interaction with GLU372 and π-alkyl interaction with LYS369 and PRO364 respectively. The oxygen atom of the nitro, phenoxy as well as acetyl groups formed conventional hydrogen bonds with VAL287 and GLY365 residues respectively.

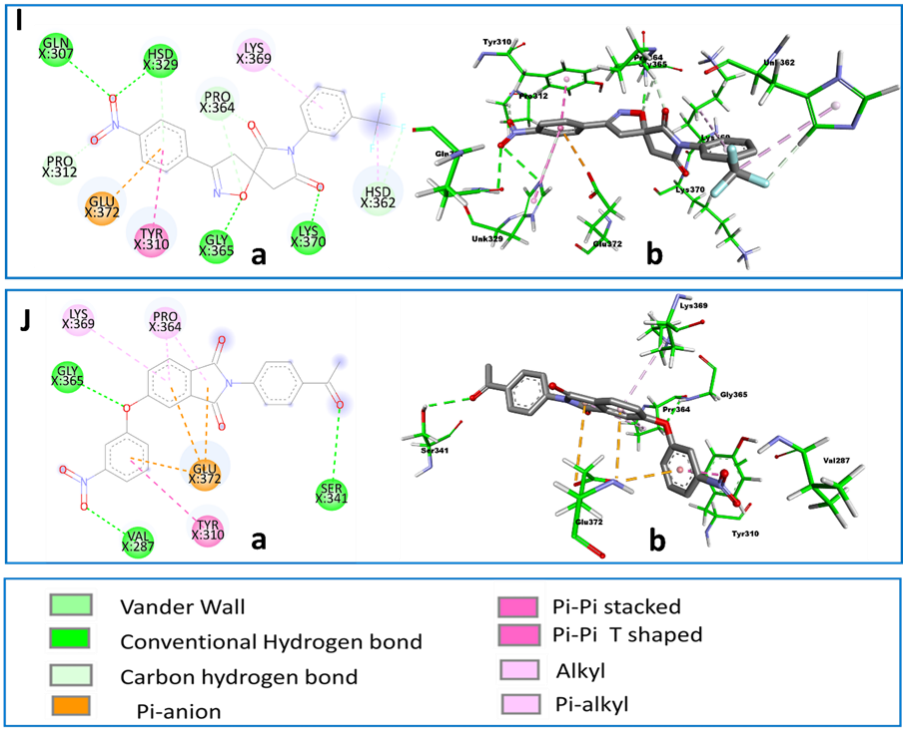

Figure 6.9. 2D (a) and 3D (b) interaction view of the molecules unk_1778 (I), unk_2181 (J) with cluster H.

6.3.5. In Silico ADME prediction

In order to achieve good bioavailability, every molecule must have good solubility and optimal pharmacokinetic properties [28]. So, an in silico pharmacokinetic prediction was performed for the top thirty-three ligands using DruMAP **(**Drug Metabolism and Pharmacokinetics Analysis Platform) webserver. Different parameters and descriptors were calculated for the top thirty-three poses of tau inhibitors [29]. From the **ADME** parameter, the top ten molecules are selected based on their high fraction of absorption (Fa), human Caco2 permeability (Papp), high solubility, and moderate to high metabolic stability (**Table 6.3**) [30-32]. Further, the ADME characteristics of the ten molecules obtained from the SwissADME web server described in **Table 6.4**, include the rule of five (molecular weight (MW), lipophilicity (iLOGP), Hydrogen bond acceptor (HBA), hydrogen bond donors (HBD) and several other parameters that includes TPSA (molecular polar surface area), number of rotatable bonds.

Table 6.3. In silico ADME properties of top thirty-three ligands caculated using DruMAP Server.

Compound	Fa_human	Papp_human_caco2	d_sol 7.4	fe_human	Clint_human
UNQ175	High	High	High	Low	Moderate
UNQ209	Moderate	Low	High	Low	Stable
UNQ298	High	High	High	High/Medium	Moderate
UNQ470	Moderate	Low	High	Low	Moderate
UNQ499	Moderate	Low	Low	Low	Moderate
UNQ542	High	Low	Low	Low	Stable
UNQ605	High	High	Low	Low	Moderate
UNQ842	High	High	Low	Low	Moderate
UNQ863	High	Low	High	Low	Stable
UNQ885	High	Low	Low	Low	Stable
UNQ926	High	High	Low	Low	Moderate
UNQ1006	Moderate	Low	High	Low	Stable
UNQ1027	High	High	High	Low	Moderate
UNQ1140	Moderate	Low	High	Low	Moderate
UNQ1172	High	High	High	Low	Moderate
UNQ1173	High	High	High	Low	Stable
UNQ1179	High	High	High	Low	Moderate
UNQ1237	High	High	High	Low	Moderate
UNQ1388	High	Low	Low	Low	Moderate
UNQ1478	High	High	Low	Low	Stable
UNQ1486	High	Low	High	Low	Moderate
UNQ1518	High	High	High	Low	Stable
UNQ1614	High	High	Low	Low	Moderate
UNQ1716	High	Low	Low	Low	Stable
UNQ1759	High	High	Low	Low	Moderate
UNQ1778	High	High	High	Low	Stable
UNQ1792	Moderate	High	High	Low	Moderate
UNQ2020	Moderate	Low	High	Low	Stable
UNQ2150	High	Low	High	Low	Moderate
UNQ2181	High	High	High	Low	Moderate
UNQ2223	Moderate	Low	High	Low	Moderate
UNQ2241	High	High	High	Low	Moderate
UNQ2310	High	High	Low	Low	Stable

Table: 6.4. In silico ADME properties of top ten hits caculated using SwissADME webserver.

Compound	MW	Rotatable bonds	H-bond acceptor	H-bond donors	MR	TPSA	I LOGP	X LOGP3	W LOGP	M LOGP	Bioavailability Score	PAINS alerts
UNK175	422.48	4	2	2	131.5	68.06	3.46	4.71	4.04	3.85	0.55	0
UNK298	476.55	4	4	0	142.41	95.81	3.66	5.71	3.34	4.19	0.55	0
UNK1027	475.48	9	7	2	129.09	167.27	2.74	4.47	4.24	1.92	0.55	0
UNK1172	444.51	3	4	0	132.49	111.67	3.46	5.52	3.76	3.11	0.55	0
UNK1173	477.51	8	5	1	146.8	118.78	2.97	1.81	0.04	1.51	0.55	0
UNK1179	399.4	4	4	1	119.28	95.23	2.78	4.41	3.97	2.6	0.55	0
UNK1237	442.51	5	3	0	140.91	72.61	3.25	4.45	2.83	3.77	0.55	0
UNK1518	398.37	5	5	0	115.59	100.27	2.47	4.05	3.8	3.75	0.55	0
UNK1778	419.31	4	9	0	105.43	104.79	2.21	2.93	3.83	3.24	0.55	0
UNK2181	402.36	5	6	0	112.49	109.5	2.76	3.5	4.01	1.87	0.55	0

MW: molecular weight; MR: molecular refractivity; TPSA**:** topological polar surface area; ILOGP, XLOGP3, WLOGP, and MLOGP are the predictive models of Swiss ADME and their values indicate lipophilicity of molecules.

6.3.6. Molecular Dynamics Simulation

On the basis of the docking score, ADME predictions, the top ten hits were subjected to MD simulations for 200 ns using the GROMACS v2020. All the top ten hits showed highest docking score with either **Cluster F (CF/ Cluster 06)** or **Cluster H (CH/Cluster 08)** of the tau protein. These high scoring complexes were used for the MD study. In the case of **CH_UNK175**, the RMSD of the tau protein (**Figure 6.10**) fluctuates between 0.6-1.25 nm, somewhat stable for 25-140 ns, and flapping in the rest simulation. RMSD of ligand **UNK_175** is lower than protein and is stable at 0.29 nm from 25-125 ns, fluctuated for the next 30 ns, and again stabilised at the end of the simulation at 0.28 nm. The RMSF of some of the ligand atoms is higher (0.2-0.35 nm) than that of the rest of the atoms, indicating their flexibility and exertions to interact with protein atoms. While the RMSF of some amino acids 325, 350, 355, and 357 is also higher (0.8-1.1 nm), signifying that they are considerably more flexible than others. The ligand **UNK_175** has formed 1-6 hydrogen bonds during the simulation timeline. Since tau is a fragmented protein, it lacks alpha helices and beta sheets that are responsible for compactness therefore the radius of gyration of this protein is observed to be fluctuating between 1.6-2 nm indicating the presence of flexible loops in the protein structure.

In the case of the **CH-UNK_298** complex (**Figure 6.11),** the RMSD of the protein (0.5-2 nm) is higher than that of the ligand (~0.24 nm). The RMSF of the protein fluctuates between 0.35-1.5 nm, while that of ligand remains between 0.04-0.28 nm. The ligand formed up to 3 hydrogen bonds during the simulation timeline. The radius of gyration is somewhat stable (1.6-1.85 nm) for 145 ns, then subsequently increased to a higher value at ~4 nm at the end of the simulation. In the **CH_UNK_1027** complex (**Figure 6.12**), the protein RMSD fluctuated for an initial 70 ns between 0.5-2.2 nm and stabilized for the rest of the simulation time with small fluctuations at the end between 1.65-1.8 nm. Furthermore, the RMSF of protein is swapping in between

0.75-2 nm, showing higher RMSF for some residues 20-33 and last three residues while the RMSF of the **UNK_1027** ligand fluctuated from 0.03-0.21 nm higher for atoms 1510-1523.

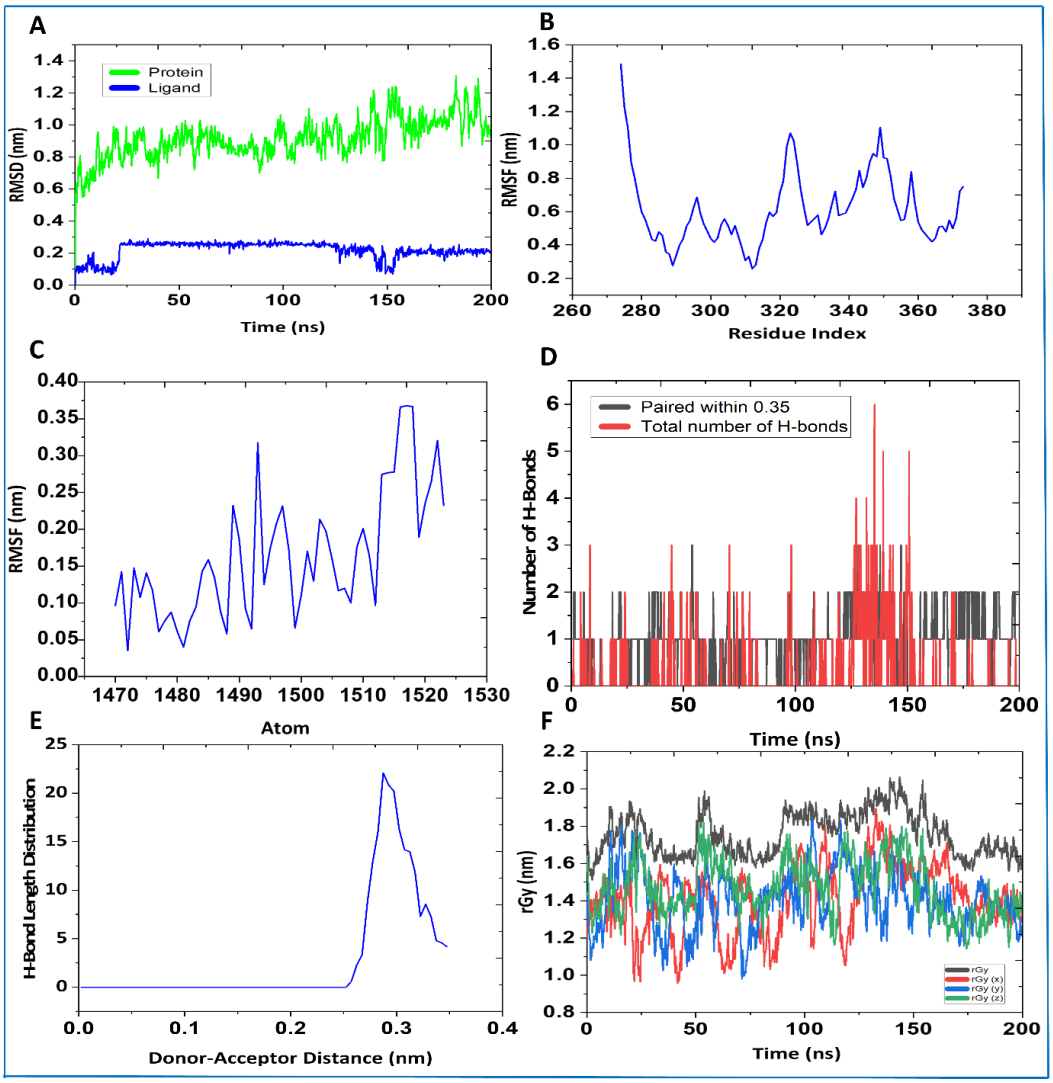

Figure 6.10. Molecular dynamics trajectory analysis of complex CH-UNK_175. (**A**) protein-ligand RMSD, (**B**) RMSF for protein, (**C**) RMSF for ligand, (D) Total counting of H-bonds for 200 ns simulation period, (E) Hydrogen bond distance, (F) Radius of gyration for the complex.

The hydrogen bond are formed throughout the simulation for this complex, which might be the reason for the stability of this complex and all the hydrogen bonds were within the cut off limit, which is 0.35 nm. Furthermore, the radius of the gyration value was unstable for ~70 ns but stabilised and remained at 1.6-1.8 nm for the rest of the simulation period.

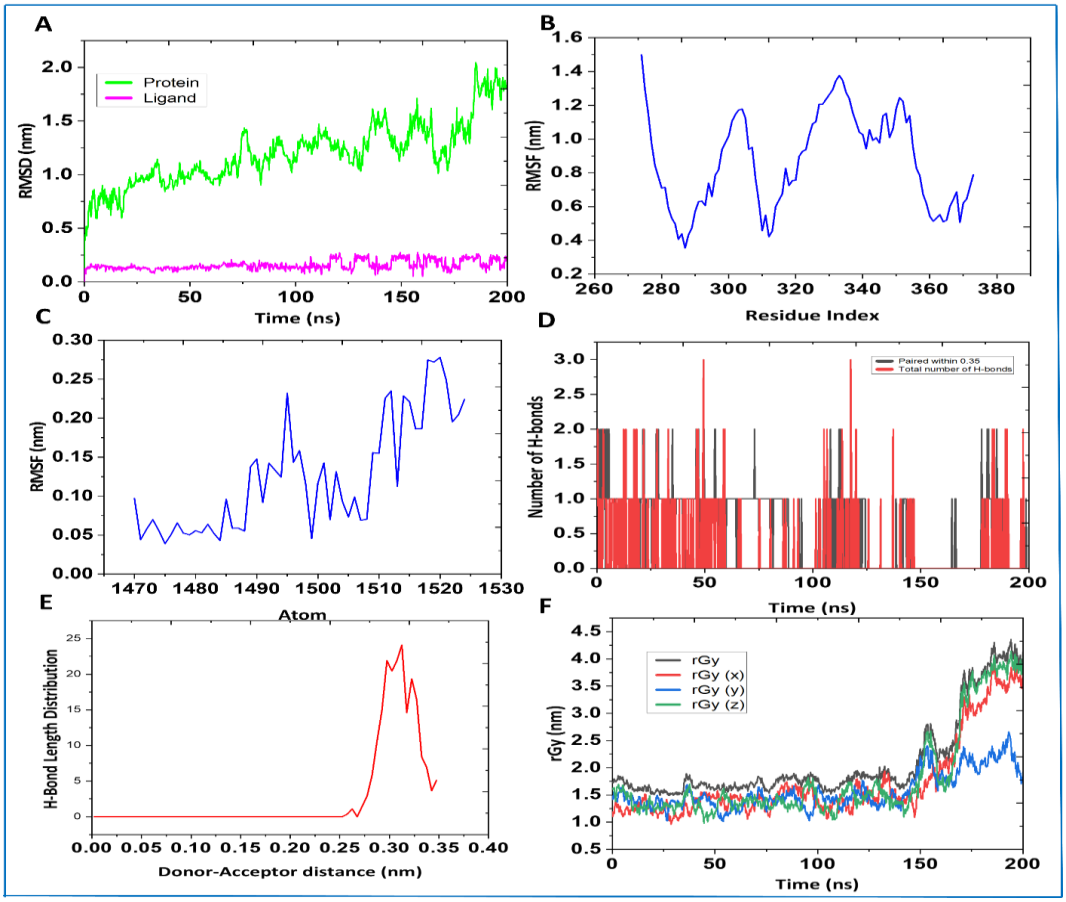

Figure 6.11. Molecular dynamic trajectory analysis of CH-UNK_298 complex. (**A**) protein-ligand RMSD, (**B**) RMSF for protein, (**C**) RMSF for ligand, (D) Total counting of H-bonds for 200 ns simulation period, (E) Hydrogen bond distance, (F) Radius of gyration for complex.

In the case of CF-UNK_1172 complex, the protein RMSD (**Figure 6.13**) fluctuates between 0.5-1.9 nm and was quite stable from 60 to 130 ns. While ligand **UNK_1172** is showing stable RMSD plot lower than that of the protein in the range of 0.1-0.2 nm. The RMSF of ligand UNK_1172 contains some atoms 1480-1483 & 1508-1515 that have higher fluctuations than other atoms, indicating their ease of movement and trying to stabilise by forming a bond with protein atoms. Moreover, RMSF of the protein ranging from 0.3-1.9 nm do have some residues like 282-295, 310-330, 340-350, 79-82, and 355-370 that are having more flexibility and hence more fluctuations than other residues of the protein. Up to 4 hydrogen bonds were formed between protein-ligand complex and all the hydrogen bonds were within the cut-off value of

0.35 nm. The compactness value of this complex fluctuates between 1.5-2.3 nm indicative of an unstructured protein.

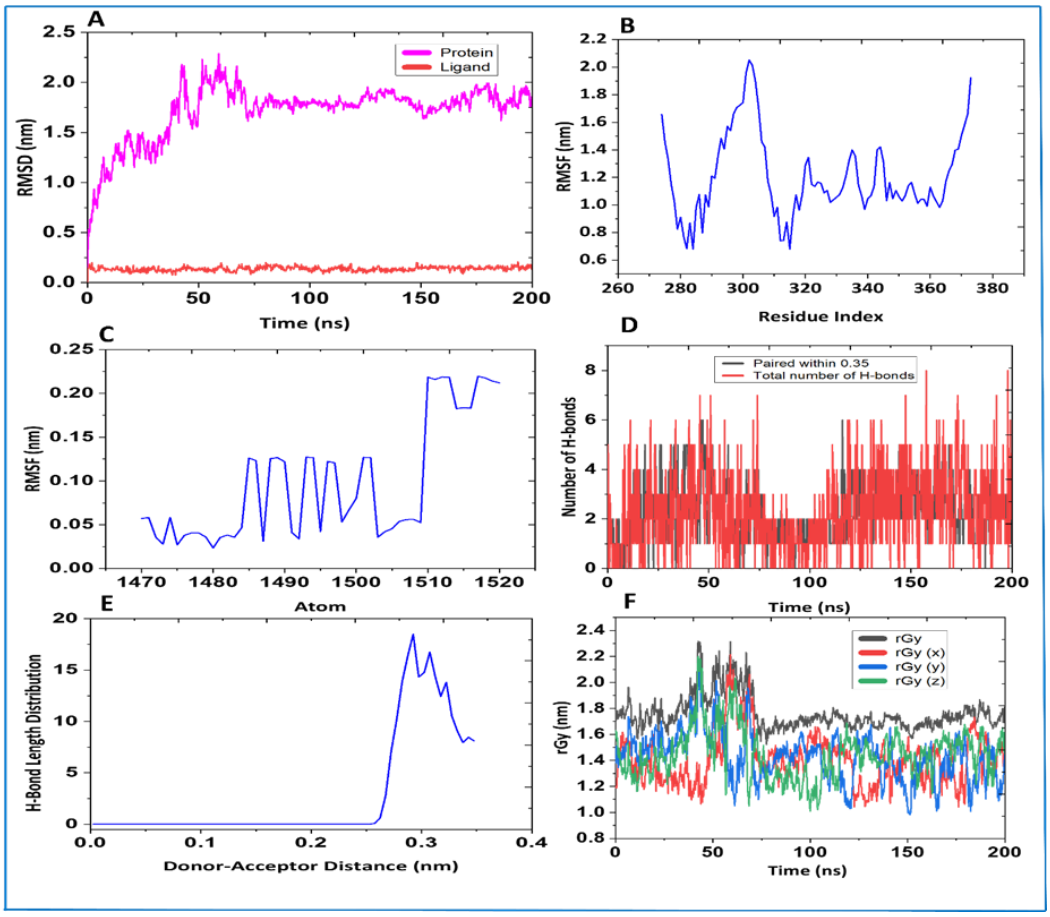

Figure 6.12. Molecular dynamic trajectory analysis of complex CH-UNK_1027 complex. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (D) Total counting of H-bonds for 200 ns simulation period, (E) Hydrogen bond distance, (F) Radius of gyration for complex.

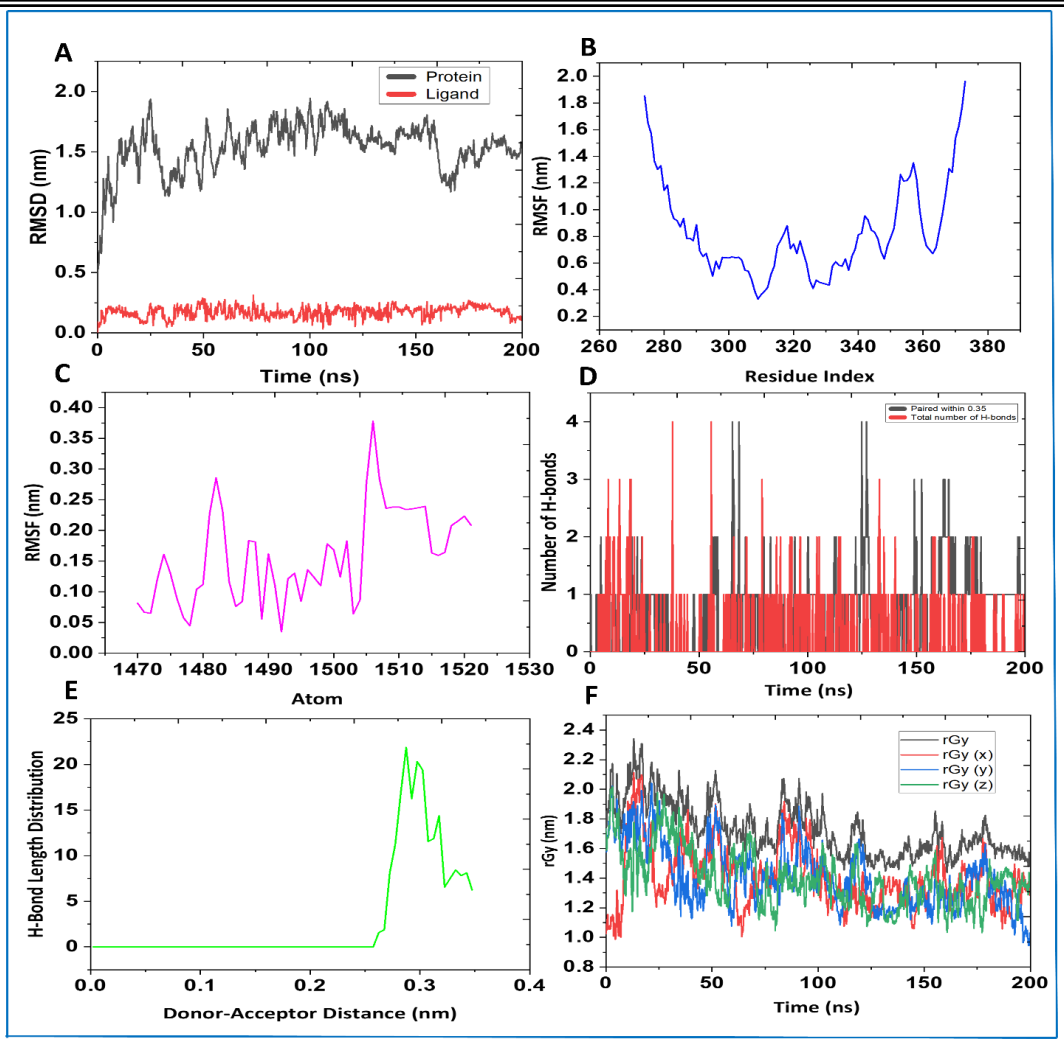

Figure 6.13. Molecular dynamic trajectory analysis of CF-UNK_1172 complex. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

In the case of **CH_UNK_1173** complex, the protein RMSD (**Figure 6.14**) fluctuates from 0.6-1.8 nm for 140 ns and becomes stable for the rest of the simulation period with fewer fluctuations and ligand UNK_1173 represents a stabilised RMSD within 1.2-1.6 nm throughout the simulation. The ligand RMSF depicts that some atoms flapping at higher values ranging in-between 0.05-0.3 nm while the RMSF of tau protein fluctuates between 0.2-2 nm with higher oscillations in some residues 300, 335-340, 350-360. The complex is stabilised by up to 5

hydrogen bonds which were within the cutoff value of 0.35 nm. The radius of gyration remains stable between 1.1 nm to 2.2 nm for 200 ns.

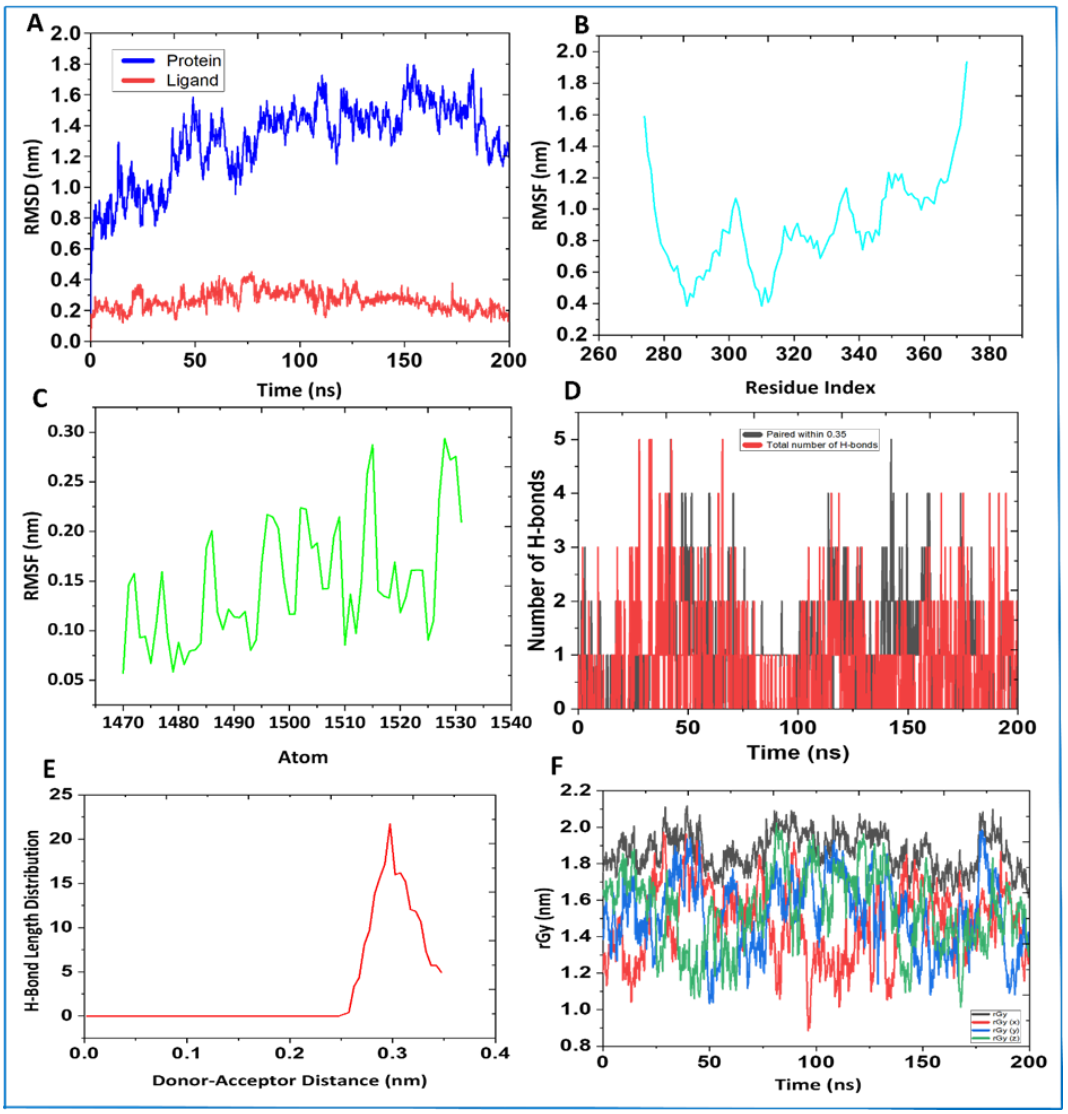

Figure 6.14. Molecular dynamic trajectory analysis of CH-UNK_1173 complex. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

In the case of **CH-UNK_1179** complex, the protein RMSD (**Figure 6.15**) fluctuates from 0.5-2.2 nm for 80 ns and becomes stable for the rest of simulation time period with less fluctuations

and ligand **UNK_1179** represents a stabilised RMSD within 0.18-0.25 nm throughout the simulation. The ligand RMSF depicts that some atoms flapping at higher value ranging in-between 0.08-0.3 nm while the RMSF of tau protein fluctuates in between 0.8-2.4 nm with higher oscillations in some residues. The complex is stabilised by up to 6 hydrogen bonds which were within the cut-off value 0.35 nm. The radius of gyration was stable initially at ~1.8 nm for 25 ns and increased to approximately 2.8, remains stable for about next 150 ns and lastly increases to 5.9 nm.

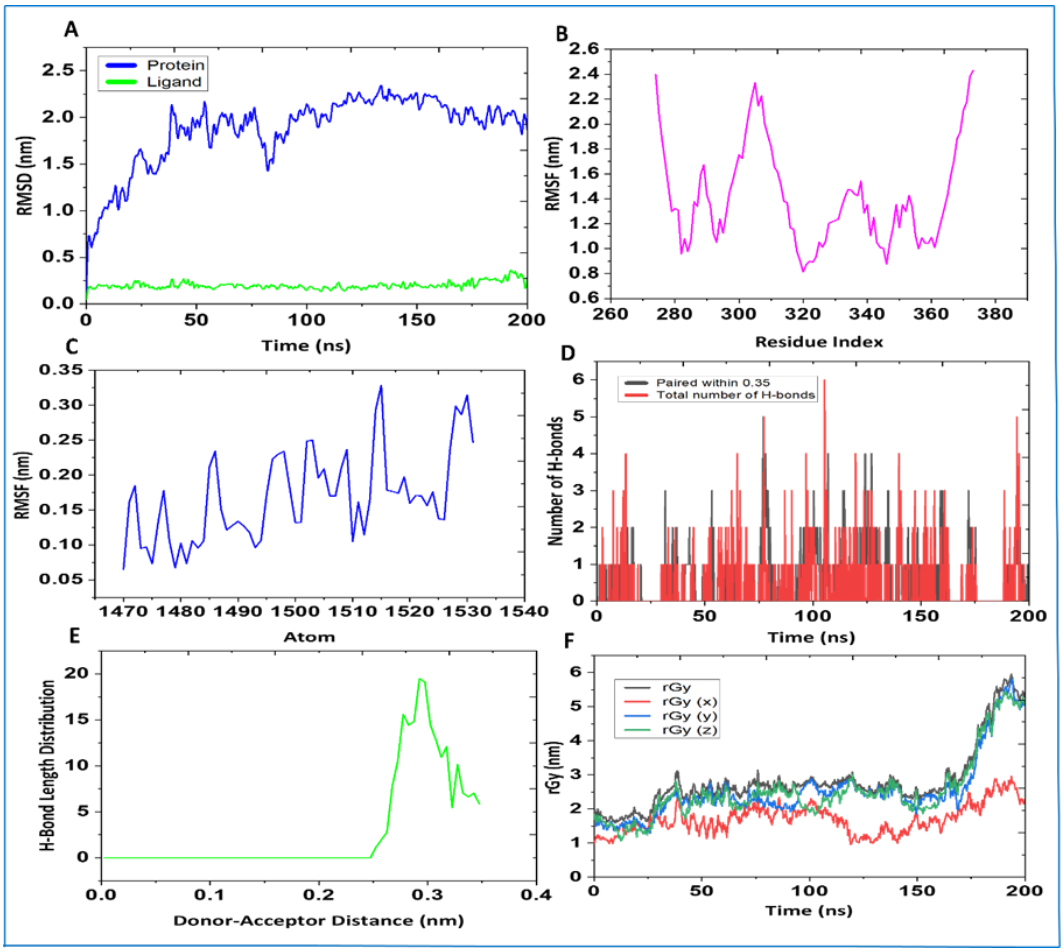

Figure 6.15. **Molecular dynamic trajectory analysis of CH-UNK_1179 complex.** (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

In the case of **CH-UNK_1237 complex,** the protein RMSD (**Figure 6.16**) is swapping in the range of 0.6-1.4 nm through the simulation with stabilization for the very short period from 20-50 ns & last 50 ns. Ligand **UNK_1237** has a stable and lower RMSD value ~0.25 nm than that of the protein. Ligand atoms have various atoms which have high fluctuations but all the atoms depict the RMSF value between range of 0.065-0.29 nm.

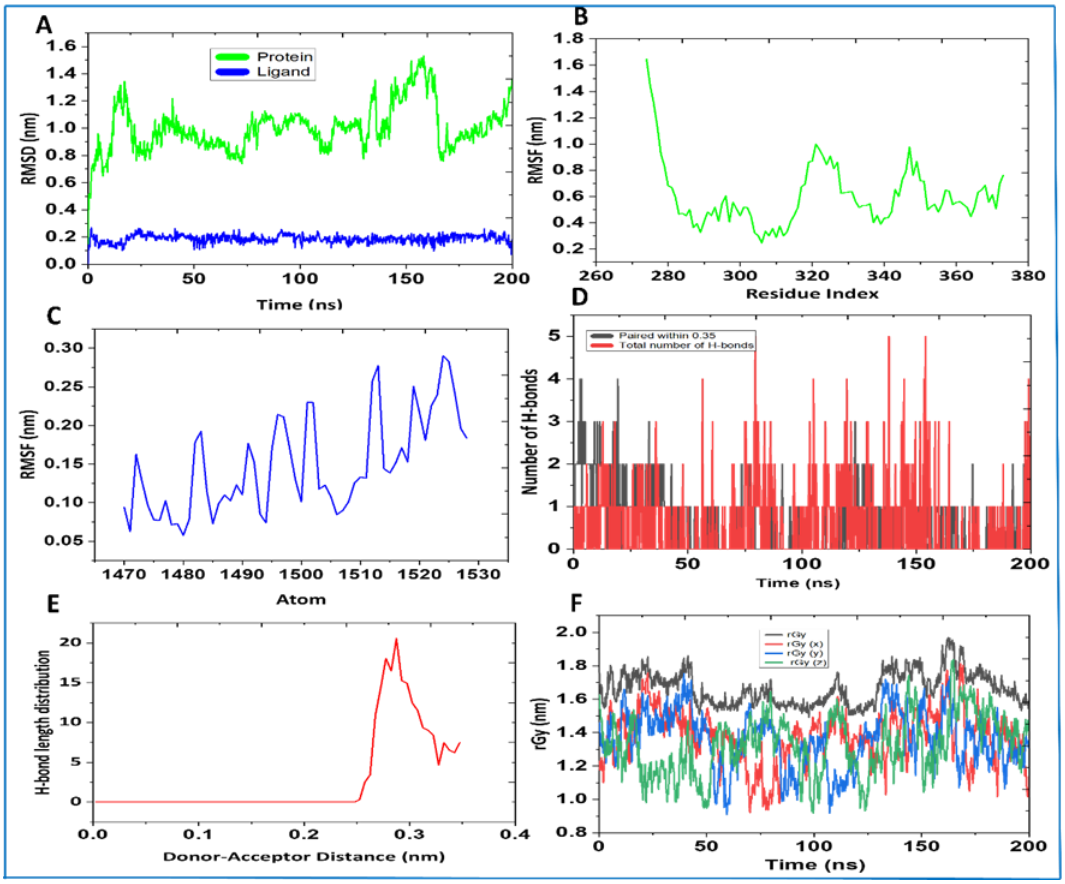

Figure 6.16. Molecular dynamic trajectory analysis of CH-UNK_1237 complex. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

On the other hand, the RMSF value for tau protein lies within 0.6-2.2 nm. Up to 5 hydrogen bonds were formed during the simulation timeline with a H-Bond distance below the cut-off

value of 0.35 nm. Consequently, the compactness value for the complex remains in the range of 0.8-2.0 nm and was steadiest from 50-180 ns.

In the case of **CH-UNK_1518** complex, the protein RMSD (**Figure 6.17**) fluctuates between 0.5-1.5 nm and while the ligand **UNK_1518** is showing a stable RMSD plot lower than the protein in the range of 0.1-0.25 nm. RMSF of ligand **UNK_1518** contains various atoms having higher fluctuations than other atoms, indicating their ease of movement. Moreover, the RMSF of the protein ranging from 0.3-1.7 nm with some residues like 300-310, 315-338, and 345-350 that are having more flexibility and hence more fluctuations than other residues of the protein.

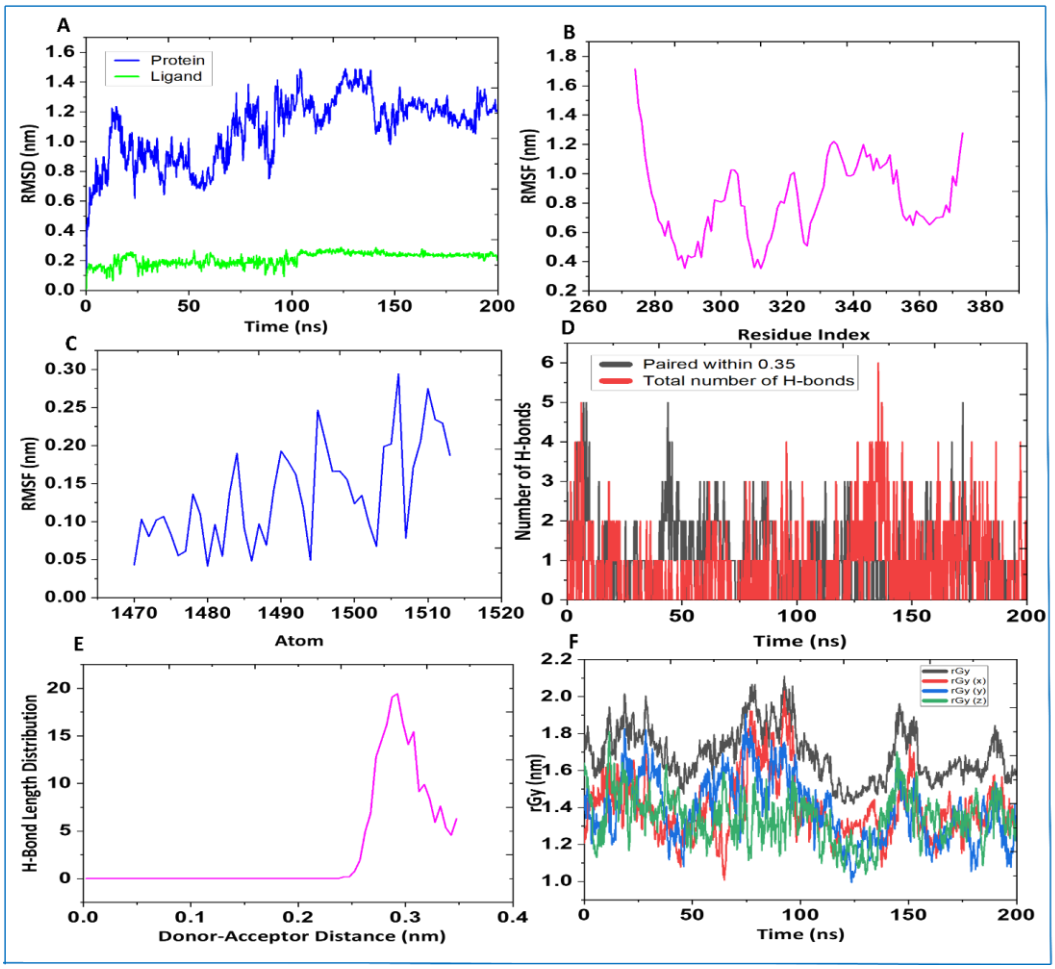

Figure 6.17. **Molecular dynamic trajectory analysis of CH-UNK_1518 complex**. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

Up to 6 hydrogen bonds contributes to the formation of the stable complex between **UNK_1518** and tau protein and all hydrogen bonds are formed within the cut-off value of 0.35 nm. The compactness value of this complex fluctuates between 1.5-2.3 nm indicative of an unstructured protein.

In the case of **CH-UNK_1778** complex, the RMSD of protein (**Figure 6.18**) was lower for the initial 25 ns between 0.6-1 nm, jumped to 1.2 nm and with small variations in the plot, it remains steady for the rest of the simulation time. The RMSF of protein is swapping in between 0.4-1.4 nm, showing higher RMSF for some residues, 290-300, 310-340 & 343-last residues while the RMSF of the **UNK_1778** ligand fluctuated from 0.03-0.35 nm higher for atoms 1482-1490 & 1495-1503. Number of hydrogen bonds were fluctuating between 0-7 and all the hydrogen bonds were within the cut-off limit, which is 0.35 nm. Furthermore, the radius of gyration value fluctuated in the range of 1.7-3.1 nm and was stable for a short period from70-125 ns.

In the case of **CH-UNK_2181** complex, the protein RMSD (**Figure 6.19**) was at lower value ranging from 0.6-1.1 for the initial 20 ns, then it rises to 1.2 and remains quite stable with little fluctuation in the range of 0.9-1.3 nm for rest of the simulation time. The RMSF value for the **UNK_2181** ligand atoms remain in between 0.03-0.27 nm, with some atoms 1496-1510 swapping with higher RMSD. The RMSF for the protein remains in the range of 0.12-1.5 nm, with some residues in higher RMSF values like 315-340 & 345-358. Up to 6 hydrogen bonds were formed during the simulation period and all hydrogen bonds were within the cut-off value of 0.35 nm. The radius of the gyration value remains in the range of 1.5-1.9 nm with lower variations over the time of the simulation.

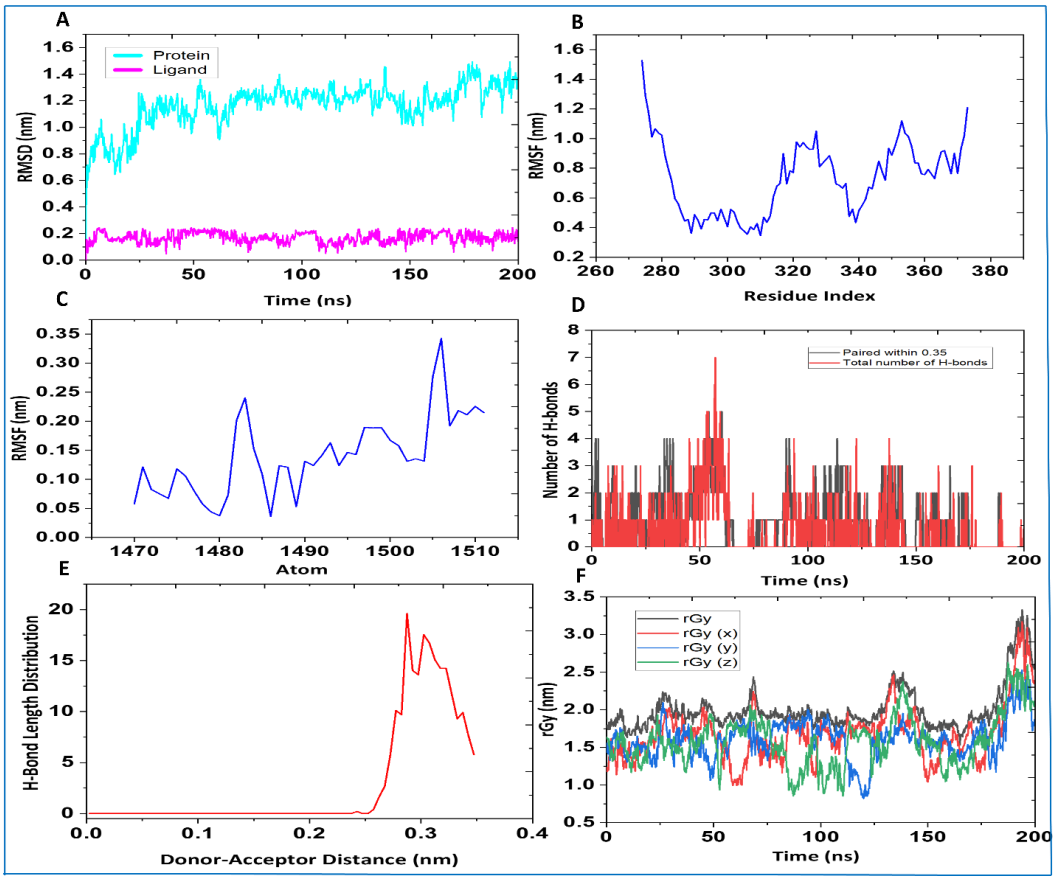

Figure 6.18. Molecular dynamic trajectory analysis of CH-UNK_1778 complex. (**A**) protein-ligand RMSD, (**B**) Root mean square fluctuation (RMSF) for protein, (**C**) RMSF for ligand, (**D**) Total counting of H-bonds for 200 ns simulation period, (**E**) Hydrogen bond distance, (**F**) Radius of gyration for complex.

Free-energy landscape (FEL) analysis was carried out which further explains the possible conformations taken by the tau protein complex during the simulation period together with the Gibbs free energy (ΔG). FEL can be representative of two variables that can directly reflect the distinctive properties of the system and measure the conformational changes. In order to obtain the Free energy minima landscape of ten complexes, the FEL diagram was obtained against RoG and RMSD as two reaction coordinates (**Figure 6.20, Figure 6.21, Figure 6.22**). FEL study revealed the changes in the Gibbs free energy (ΔG) value within the range of 0 and 11.80 kJ/mol for the complexes. The shape and size of the minima energy landscape was investigated.

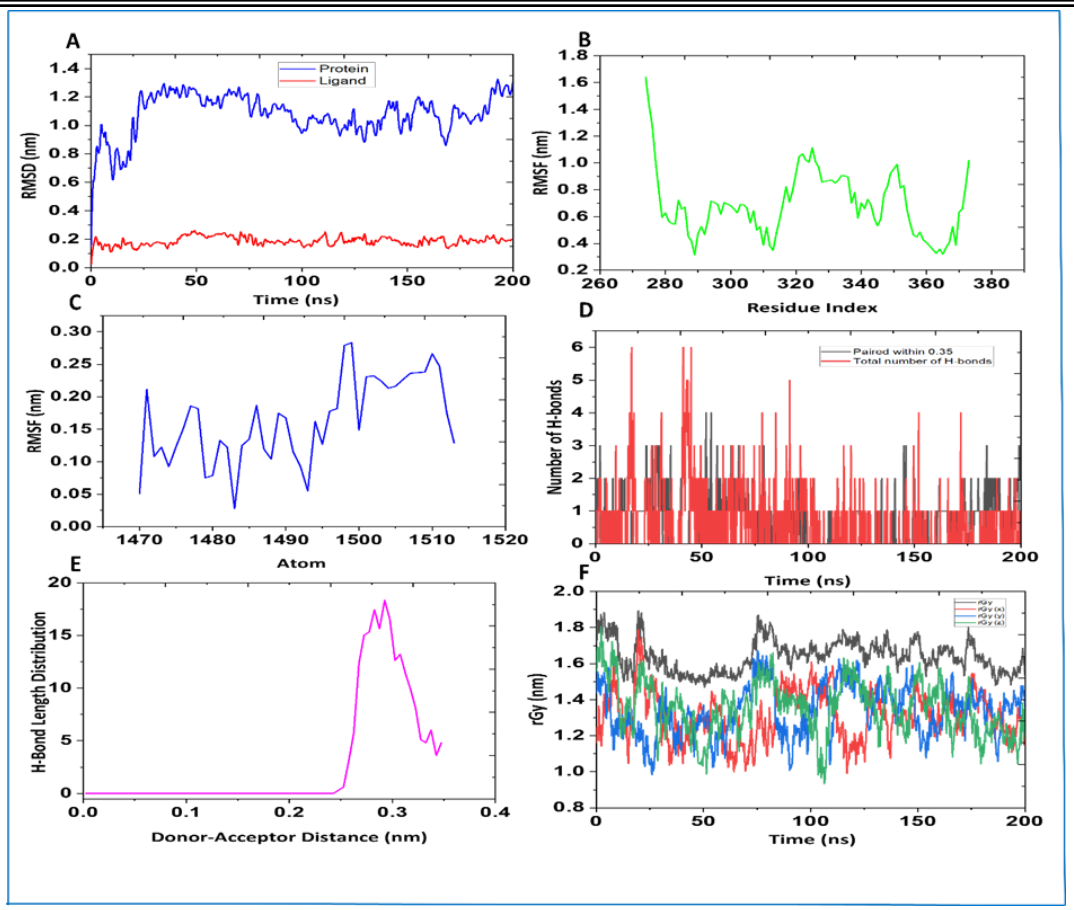

Figure 6.19. Molecular dynamic trajectory analysis of CH-UNK_2181 complex. (A) protein-ligand RMSD**, (B)** Root mean square fluctuation (RMSF) for protein, **(C)** RMSF for ligand, **(D)** Total counting of H-bonds for 200 ns simulation perio**d, (E)** Hydrogen bond distance, **(F)** Radius of gyration for complex.

Where minor and centralized blue areas indicate the complex within the cluster with the most stability and with the least energy. The narrow-shaped funnel formed in the 3D projections tells that the dynamic changes in conformations with respect to time in order for the system to reach a native structure with the least energy. The overall 3D plots demonstrate that the complexes of cluster H with the ligand **UNK_175, UNK_1027, UNK_1173, UNK_1237, UNK_1518,** and **UNK_2181** and cluster F with the ligand **UNK_1172** have generated a single funnel, which along with the 2D contour plot represents that the complexes have one local energy minima, thus having a stable folding process in the simulated system, except for the complex **CH-**

UNK_298, CH-UNK_1179, and **CH-UNK_1778** for which the 3D plot shows a minimal split thus having two local energy minima however they both lie very close within the three-dimensional space and is having a relatively unstable folding process.

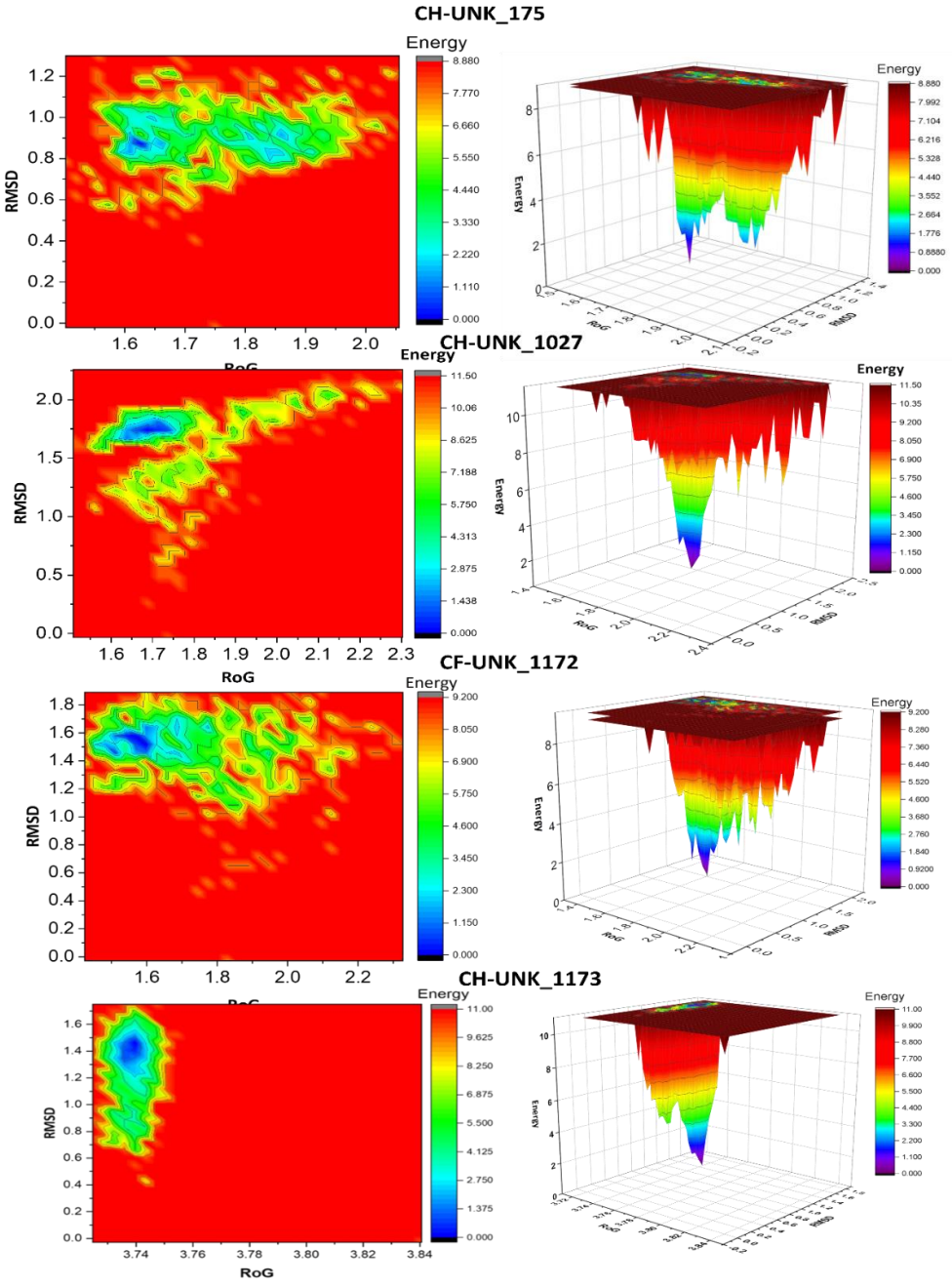

Figure 6.20. The 2D and 3D free energy landscape diagram as a function of RMSD and RoG as the two coordinates of tau cluster and ligand complex. The free energy is displayed in terms of kj/mol where the purple color indicates least energy and red the highest energy. The folding funnel formed by the complex **CH-UNK_175**, **CH-UNK_1027**, **CF-UNK_1172**, **CH-**

UNK_1173 shows stable folding.

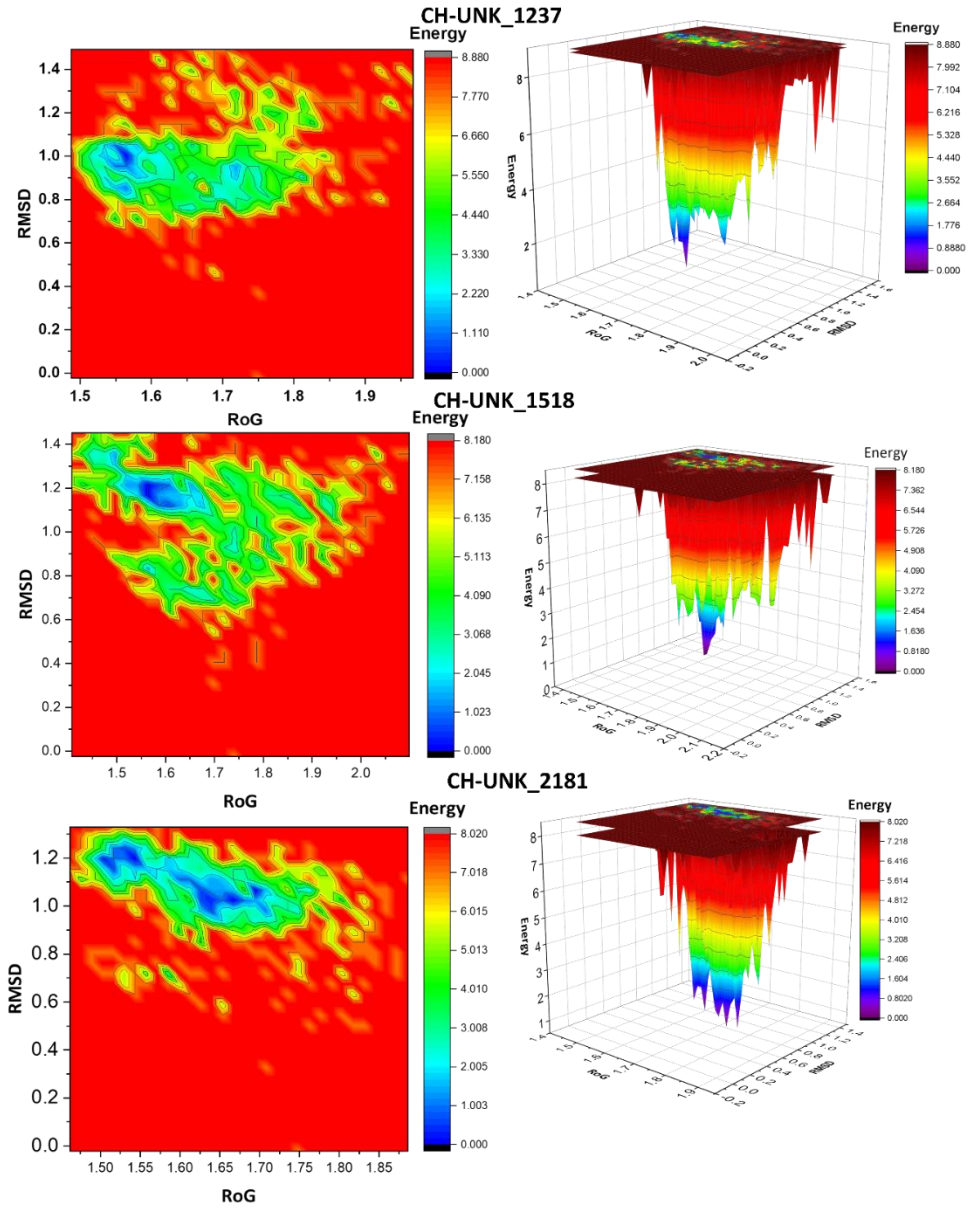

Figure 6.21. The 2D and 3D free energy landscape diagram as a function of RMSD and RoG as the two coordinates of tau cluster and ligand complex. The free energy is displayed in terms of kj/mol where the purple color indicates least energy and red the highest energy. The folding funnel formed by the complex **CH-UNK_1237**, **CH-UNK_1518**, **CH-UNK_2181** shows stable folding.

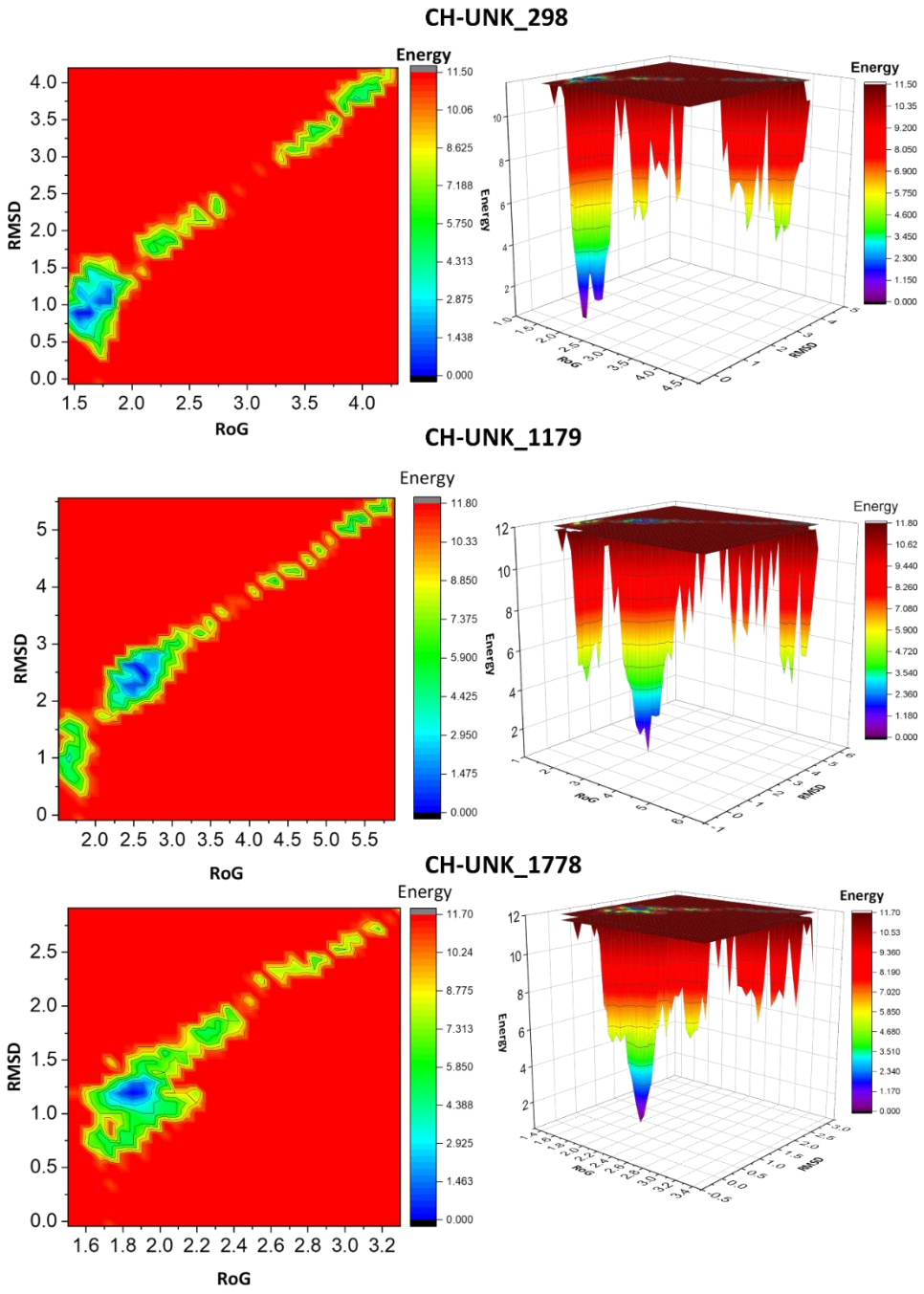

Figure 6.22. **The 2D and 3D free energy landscape diagram as a function of RMSD and RoG as the two coordinates of tau cluster and ligand complex**. The free energy is displayed in terms of kj/mol where the purple colour indicates least energy and red the highest energy. The folding funnel formed by the complex system into one narrow multiple funnels denoted an unstable folding towards energy minima. The folding funnels of complex system **CH-**

UNK_298, **CH-UNK_1179** and **CH-UNK_1778** show an unstable folding process with multiple funnels.

6.3.7. Binding Free Energy Calculation

To comprehend the molecular binding affinity of the top ten compounds at the active binding pocket of cluster F and cluster H, we calculated the binding free energy (ΔG) using MM-PBSA which includes the energetic terms that accounts for van der Waals contribution from MM (VDWAALS), electrostatic energy (EEL), the electrostatic contribution to the solvation free energy calculated by GB (EGB), nonpolar contribution to the solvation free energy calculated by an empirical model (ESURF). **Figure 6.23** demonstrates the violin plot for eight compounds showing good binding affinity. At first, we conducted MM-PBSA calculation to find the binding free energy, on the lowest energy minima frame of the simulation trajectory and two of its adjacent frames, the resultant ΔG was taken as an average of the three frames (**Table 6.5**). Finally, the compound **UNK_1027** was found to have the best-predicted binding with the cluster H having ΔG score of -42.02 kcal/mol, followed by **UNK_1518**, **UNK_1237**, **UNK_1173**, **UNK_2181,** and **UNK_175** respectively. **UNK_1172** also showed good binding affinity to cluster F. Though the ligand **UNK_298** showed a good binding score -32.01 but we have excluded the compound because the docked complex was not stable during the MD simulation. Similarly, the compounds **UNK_1179** and **UNK_1778** were also not forming a stable complex with the tau protein as revealed by MD simulation.

Table 6.5. MM-PBSA binding free energy (ΔG) calculations for top ligands with the respective cluster[a].

Compound	van der Waals	Electrostatic energy	EGB	ESURF	GGAS	GSOLV	Averaged ΔG (kcal/mol)
UNK_175	-33.27	-17.45	30.82	-3.97	-50.73	26.85	-23.73
UNK_298	-47.95	-31.7	52.97	-5.69	-79.29	47.27	-32.01
UNK_1027	-49.48	-32.08	45.71	-6.17	-81.56	39.54	-42.02
UNK_1172	-39.86	-16.93	35.91	-5.54	-56.79	30.37	-26.42
UNK_1173	-35.67	-22.61	36.96	-4.87	-58.28	32.09	-26.2
UNK_1179	0	-0.45	1.11	0	-1.11	0.45	0
UNK_1237	-49.02	-12.24	39.42	-6.48	-61.27	32.93	-28.33
UNK_1518	-49.02	-34.94	52.24	-5.56	-77.26	46.68	-30.57
UNK_1778	-1.97	-1.99	2.09	-0.33	-3.25	2.24	-0.14
UNK_2181	-31.09	-5.01	27.71	-4.33	-36.10	22.37	-26.05

[a]MM-PBSA: molecular mechanics Poisson–Boltzmann surface area; EGB: the electrostatic contribution to the solvation free energy calculated by PB or GB; ESURF: nonpolar contribution to the solvation free energy calculated by an empirical model; GGAS: Gibbs free energy into a gas-phase term; GSOLV: Gibbs free energy into a solvation term. All energies are expressed in kcal/mol.

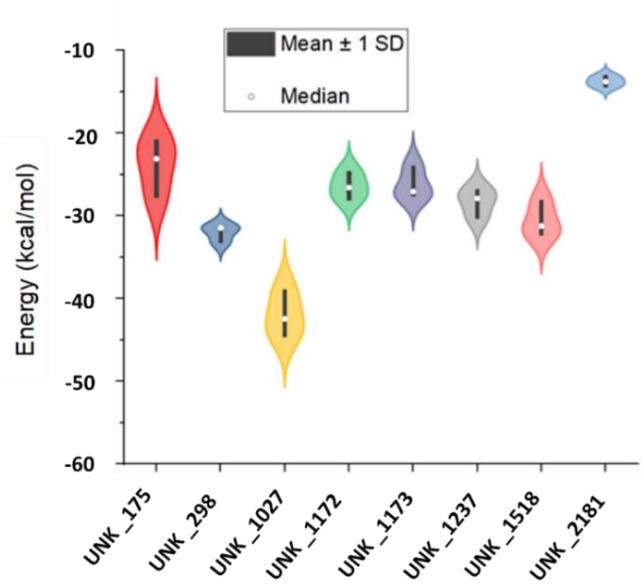

Figure 6.23. Violin plot of the ΔG (kcal/mol) calculated using MM-PBSA method for compounds UNK175, UNK1027, UNK1172, UNK_1173, UNK_1237, UNK_1518 and

UNK_2181. It is a hybrid plot comprising of the box plot and a kernel density plot that shows the peak in the data, it can depict the summary statistics and the density of each variable. The mean denotes the average of the total data with a standard deviation of ± 1 and the median denoted by the white circle is the exact middle value in the dataset.

6.4. **Materials and methods**

6.4.1. Artificial Intelligence Based Virtual Screening

For ligand-based virtual screening, we used PyRMD, which is an open-source, AI-assisted ligand-based virtual screening tool (https://github.com/cosconatilab/PyRMD) [3]. PyRMD can easily screen millions of compounds with highly efficient and accuracy and it uses a specific RMD algorithm for ligand-based virtual screening. It is highly customizable and uses a variety of benchmarking metrics to increase model performance. Additionally, PyRMD provided higher classification capabilities in comparison to other popularly used cheminformatics tools and it has been found PyRMD has more ability to identify inactive compounds accurately. In order to investigate the advantage of PyRMD performance over the other cheminformatics 2D methods, Amenodla and Cosconati performed several experiments by using the same data set and protocols associated with others like the random forest, gradient boosting, logistic regression, and naive Bayes, and observed that PyRMD, random forest, and gradient boosting produced better performance than the others [3].

PyRMD takes the bioactivity of compounds from a database like ChEMBL, and uses the data to classify compounds into active and inactive. Based on the data, it trains a model, and use the trained model to screen a large chemical library and predict biological active molecules from the dataset. PyRMD uses a random matrix discriminant algorithm for the identification of small molecules endowed with biological activity. A typical PyRMD workflow begins with the user-identified biological activity database of small molecules against a particular target of interest, which acts as the training data for the AI model. Later the database undergoes a preparation

process in which the duplicate molecules are removed, and the molecules are segregated as active, inactive, or discarded based on the threshold values for activity set by the user. The next step is the featurization process in which all the molecules are converted into their SMILES format and various user-defined molecular fingerprints of the ligands are calculated for active and inactive molecules. Some of the most widely used fingerprints are MHFP, ECFP, FCFP, etc. [33]. In the next step, the actual RMD algorithm-based classification of an unknown dataset of small molecules takes place. In this step based on the trained model, a new library of small molecules, with unknown activities is classified as either active or inactive in a two-step fitting and screening process. Finally, the output is produced having the active molecules and their RMD scores [3].

6.4.2. Screening database

For the ligand-based virtual screening process using PyRMD, we downloaded the biological data from the ChEMBL database [5]. The biological activity data of small molecules reported against human microtubule-associated tau protein were downloaded from the ChEMBL ID: ChEMBL 1293224 (https://www.ebi.ac.uk/chembl/g/#search_results/all/query=mapt) and was used as the training dataset. For the featurization process, we used the MHFP fingerprints with 2048 bits length [33]. The activity threshold was set at 1100 nM and an inactive threshold was kept at 33000 nM. The molecules having IC_{50}, EC_{50}, K_i, K_d, or potency inferior to the activity threshold were considered as actives, and above the inactive threshold were considered inactive and those that fall in between were discarded. Other training parameters like epsilon cut-off for actives and inactive were kept at default. Also, the compounds that fall outside the range of 200-500 molecular weight, -2 to 5 logP, 0-10 hydrogen bond donors, 0-15 hydrogen bond acceptors, 0-20 rotatable bonds, 10-50 heavy atoms were discarded. We used in-stock compounds library from the ZINC database which contained ~12 million compounds as a screening library [6].

6.4.3. **Diversity picking using RDKit Diversity picker**

After the initial screening process, the output obtained from PyRMD was further subjected to diversity-picking process. This was achieved using the Diversity picker module of RDKit, which is a cheminformatics tool capable of efficiently removing molecules with similar chemical structures or overlapping chemical features. The process was executed using KNIME software, which has been noted for its ability to provide efficient solutions for molecular format transformation and conformer generation [9, 34, 35]. Compared to other cheminformatics tools, RDKit has been identified as a valuable tool due to its high efficiency and capability to generate high-quality conformers in a relatively short amount of time [36, 37].

6.4.4. **REMD simulation**

Chain A of protein with PDB ID 6VH7 was used for building the system as it contains the paired helical filaments (PHF) region [38]. The amino acid sequence of tau protein from residue number 274 to 374 was used generate a random coil conformation by using the webserver (https://spin.niddk.nih.gov/bax/nmrserver/pdbutil/ext.html). The webserver takes the amino acid sequence and coordinates and generate a random coil conformation. The topology of the system was built using the CHARMM-GUI web server, using the random coil conformation of the tau peptide [39]. We used the CHARMM36m force field for the simulation as it is an improved force field specifically designed to define the motions associated with intrinsically disordered proteins like tau [40, 41]. The peptide was solvated in a rectangular box of dimension 90Å x 70Å x 40Å using the TIP3 water model [42]. By using the Monte Carlo ion placing method the system was neutralized for ionic charges by adding Na^+ and Cl^- such that the total ionic concentration was kept at 0.15 M. It was followed by energy minimization utilizing the steepest descent algorithm for 5000 steps with the LINCS algorithm used to constrain bonds having hydrogens [43].

For performing REMD simulation, we created thirty-one replicas with a temperature range from 310 K to 400 K. The temperatures were selected such that there would be a sufficient overlap of potential energy between adjacent replicas with acceptable exchange probabilities [44, 45]. We ran a short REMD simulation for the system to check the average exchange probabilities and it fell in the range of 20-30 %, and hence the temperature range was accepted for the production runs. Each of the energy-minimized replica was equilibrated at their corresponding temperature for 1000 ps using Nose –Hoover temperature coupling under a canonical ensemble [46]. For the REMD production simulation, we carried out 100 ns REMD simulation runs with replica exchange attempted periodically in every 100 steps and the total production simulation time was 3.1 microseconds. All the simulations were carried out using the GROMACS 2020.3 simulation package [47]. The temperatures used for each replica were 310.0, 312.5, 315.0, 318.0, 321.0, 324.0, 327.0, 330.0, 333.0, 336.0, 339.0, 342.0, 345.0, 348.0, 351.0, 354.0, 357.0, 360.0, 363.0, 366.0, 369.0, 372.0, 375.0, 378.0, 381.0, 384.0, 387.0, 390.0, 393.0, 396.0, 400.0 K and the average exchange probabilities were ~24%.

The REMD simulation trajectory corresponding to the 310 K temperature was used for the clustering process. The gmx cluster module of GROMACS also called the GROMOS clustering algorithm, was used to perform the RMSD clustering with a cut-off value of 0.5 nm [48].

6.4.5. **Active site identification**

In order to identify probable regions of the peptide having high ligand binding affinity, we used the FTsite webserver (https://ftsite.bu.edu/) [25]. FTsite uses a solvent mapping algorithm that computationally maps sixteen different small molecule probes on the protein grids and identifies empirical free energy functions. Based on the interaction of these sixteen probes, consensus clusters are identified based on their overlap. Later the clusters on the protein are ranked based on non-bonded interaction between probes and protein, with the cluster having the highest interaction getting the first rank. The representative structure from each of the top

ten clusters obtained after clustering of REMD simulation was submitted to FTsite webserver. The server calculated the favourable binding pockets in each of the ten structures and identified three favourable binding regions in each of them.

6.4.6. Molecular Docking

The present study employed the Autodock Vina for molecular docking [49-51], for the 2367 compounds that we obtained after diversity picking through RDKit. Prior to docking, all the ligands were converted into PDBQT files by using Open Babel software [52]. Protein structure was prepared by using AutoDock tools, water molecules were deleted, polar hydrogen and assign AD4 type atoms are added and finally, Kohlman charges were added to the protein. Grid box was generated acceding to the active site of the clusters and the exhaustiveness value was kept at eight [53]. The resulting molecular interactions were visualized using the BIOVIA Discovery Studio Visualizer 2021 [54, 55], and both 2D and 3D interaction images were generated.

6.4.7. In Silico ADME prediction

DruMap web server predicts pharmacokinetic parameters with more accuracy and efficiency [56]. We used the DruMap web server (https://drumap.nibiohn.go.jp/) in order to determine the ADME properties of thirty-three ligands. DruMap web server used different descriptors and parameters to determine the ADME properties of different molecules. We also used the Swiss ADME web server (http://www.swissadme.ch/) for the determination of properties like molecular weight (MW), lipophilicity (iLOGP), Hydrogen bond acceptor (HBA), hydrogen bond donors (HBD) and other parameters that includes TPSA (molecular polar surface area) [57, 58].

6.4.8. Molecular Dynamics simulation

Molecular dynamics simulation is a popular computational simulation method used to check the physical motion of atoms as well as molecules. So, in order to evaluate the stability of the docked molecules, the docked protein-ligand complex was simulated for 200 ns using GROMACS 2020 software [13]. The best-docked pose of the ligand was used to generate ligand topology files using topology files with the help of the CHARMM-GUI web server using the input generator for the GROMACS. The server utilizes the CHARMM 36m force field to the ligand [39, 59]. The TIP3 model was used to solvate the protein and the protein was neutralized by using the Monte Carlo ion placing method adding Na^+ and Cl^- ions. The temperature was kept constant at 303.15 K with a Noose Hoover thermostat. All the parameters like NVT and NPT files are generated according to the previous method [46]. At last, the system was subjected to production simulation for 200 ns. When the MD run is completed, all the trajectories were analyzed by using GROMACS utilities. RMSD, RMSF, no of hydrogen bonds, and radiation of gyration were analysed and plotted.

6.4.9. Binding Free Energy Calculation

The enthalpy and entropy contributions were calculated for the energy minima structure obtained from the Free Energy Landscape diagram along with its two adjacent frames of the trajectory. Before running the estimation with gmx_MMPBSA, PBC conditions were removed and the GROMACS output trajectory was accurately fitted [60], here we have used a single trajectory approach to calculate the binding free energy differences. gmx_MMPBSA is a tool written in Python 3.8 [61] that combines the functionality from GROMACS and Amber tools [62] in order to build input files in an accurate manner that can be reproducible to perform the calculations of end-state free energy. The work has been carried out in three crucial steps (i) preparation where the MMPBSA.py calculation engine provided in the gmx_MMPBSA was used to carry out the calculations, the MD simulation output topology files from GROMACS

was used for conversion into Amber topology format, (ii) multiple calculations are carried out for binding free energy with different solvation models (PB, GB, or 3D-RISM), and (iii) visualization and analysis of result was performed once the calculations are complete by the help of graphical user interface application (gmx_MMPBSA_ana).

6.5. Conclusion

In the present work, at first, we made use of an extensive modelling process i.e., REMD to capture all different possible stabilised conformations of tau protein. Then, by using an AI-assisted ligand-based virtual screening, we identified 8915 hit molecules out of 12 million compounds contained in the ZINC database. Then, various filters were applied like PAINS substructure, Lipinski rule of five, and diversity picker using the KNIME platform to obtain a dataset of 2367 ligands. Subsequently, with the help of Autodock Vina software, we docked these ligands against ten different conformations of tau protein which we got after REMD. Further based on binding affinity we chose the top 100 molecules after docking to find out such types of molecules that showed a higher binding affinity with more than five clusters. We found thirty-three candidates which were further prioritised by predicting their physicochemical parameters to yield seven molecules retaining high solubility, permeability, and absorption. Finally, by using molecular dynamics and binding free energy calculations we determined that these seven molecules i.e. UNK175, UNK1027, UNK1172, UNK_1173, UNK_1237, UNK1518, and UNK2181 potential tau aggregation inhibitors. We believe that an AI-assisted study will pave the way for the identification of potential therapeutic ligands against tauopathy.

6.6. References

1. S. Nag, A.T.K. Baidya, A. Mandal, A.T. Mathew, B. Das, B. Devi, R. Kumar, Deep learning tools for advancing drug discovery and development, 3 Biotech, 12 (2022) 110.
2. L. Patel, T. Shukla, X. Huang, D.W. Ussery, S. Wang, Machine learning methods in drug discovery, Molecules, 25 (2020) 5277.
3. G. Amendola, S. Cosconati, PyRMD: a new fully automated ai-powered ligand-based virtual screening tool, Journal of Chemical Information and Modeling, 61 (2021) 3835-3845.
4. N.A. Murugan, G.R. Priya, G.N. Sastry, S. Markidis, Artificial intelligence in virtual screening: models versus experiments, Drug Discovery Today, (2022).
5. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic acids research, 40 (2012) D1100-D1107.
6. J.J. Irwin, B.K. Shoichet, ZINC− a free database of commercially available compounds for virtual screening, Journal of chemical information and modeling, 45 (2005) 177-182.
7. J.B. Baell, J.W.M. Nissink, Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017. Utility and Limitations, ACS chemical biology, 13 (2018) 36-44.
8. P. Tosco, N. Stiefl, G. Landrum, The integration of Open3DTOOLS into the RDKit and KNIME, Journal of Cheminformatics, 6 (2014) 1-1.
9. M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, KNIME-the Konstanz information miner: version 2.0 and beyond, AcM SIGKDD explorations Newsletter, 11 (2009) 26-31.
10. J. Avila, J.J. Lucas, M. Perez, F. Hernandez, Role of tau protein in both physiological and pathological conditions, Physiological reviews, (2004).
11. M.G. Spillantini, M. Goedert, Tau protein pathology in neurodegenerative diseases, Trends in neurosciences, 21 (1998) 428-433.
12. G. Lee, R.L. Neve, K.S. Kosik, The microtubule binding domain of tau protein, Neuron, 2 (1989) 1615-1624.
13. A.T. Mathew, A.T.K. Baidya, B. Das, B. Devi, R. Kumar, N-glycosylation induced changes in tau protein dynamics reveal its role in tau misfolding and aggregation: A microsecond long molecular dynamics study, Proteins, (2022).

14. M.I. Khan, F. Hasan, K.A. Hasan Al Mahmud, A. Adnan, Domain focused and residue focused phosphorylation effect on tau protein: A molecular dynamics simulation study, Journal of the mechanical behavior of biomedical materials, 113 (2021) 104149.
15. L. Rani, S.S. Mallajosyula, Phosphorylation-Induced Structural Reorganization in Tau-Paired Helical Filaments, ACS Chem Neurosci, 12 (2021) 1621-1631.
16. L. Rani, S.S. Mallajosyula, Phosphorylation versus O-GlcNAcylation: Computational Insights into the Differential Influences of the Two Competitive Post-Translational Modifications, The journal of physical chemistry. B, 121 (2017) 10618-10638.
17. R.C. Bernardi, M.C. Melo, K. Schulten, Enhanced sampling techniques in molecular dynamics simulations of biological systems, Biochimica et Biophysica Acta (BBA)-General Subjects, 1850 (2015) 872-877.
18. W. Jiang, J.C. Phillips, L. Huang, M. Fajer, Y. Meng, J.C. Gumbart, Y. Luo, K. Schulten, B. Roux, Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD, Computer physics communications, 185 (2014) 908-916.
19. R. Brüschweiler, Efficient RMSD measures for the comparison of two molecular ensembles, Proteins: Structure, Function, and Bioinformatics, 50 (2003) 26-34.
20. M.Y. Lobanov, N. Bogatyreva, O. Galzitskaya, Radius of gyration as an indicator of protein structure compactness, Molecular Biology, 42 (2008) 623-628.
21. A. Möglich, K. Joder, T. Kiefhaber, End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation, Proceedings of the National Academy of Sciences, 103 (2006) 12394-12399.
22. G.G. Maisuradze, A. Liwo, H.A. Scheraga, Relation between free energy landscapes of proteins and dynamics, Journal of chemical theory and computation, 6 (2010) 583-595.
23. M.R. Jensen, M. Zweckstetter, J.-r. Huang, M. Blackledge, Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy, Chemical reviews, 114 (2014) 6632-6660.
24. S.-H. Chong, S. Ham, Folding free energy landscape of ordered and intrinsically disordered proteins, Scientific reports, 9 (2019) 1-9.
25. C.-H. Ngan, D.R. Hall, B. Zerbe, L.E. Grove, D. Kozakov, S. Vajda, FTSite: high accuracy detection of ligand binding sites on unbound protein structures, Bioinformatics, 28 (2012) 286-287.

26. R. Brenke, D. Kozakov, G.-Y. Chuang, D. Beglov, D. Hall, M.R. Landon, C. Mattos, S. Vajda, Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques, Bioinformatics, 25 (2009) 621-627.
27. D. Kozakov, D.R. Hall, G.-Y. Chuang, R. Cencic, R. Brenke, L.E. Grove, D. Beglov, J. Pelletier, A. Whitty, S. Vajda, Structural conservation of druggable hot spots in protein–protein interfaces, Proceedings of the National Academy of Sciences, 108 (2011) 13528-13533.
28. B. Das, A.T. Baidya, A.T. Mathew, A.K. Yadav, R. Kumar, Structural modification aimed for improving solubility of lead compounds in early phase drug discovery, Bioorganic & Medicinal Chemistry, (2022) 116614.
29. M. Kuroda, R. Watanabe, T. Esaki, H. Kawashima, R. Ohashi, T. Sato, T. Honma, H. Komura, K. Mizuguchi, Utilizing public and private sector data to build better machine learning models for the prediction of pharmacokinetic parameters, Drug Discovery Today, (2022) 103339.
30. T. Esaki, R. Ohashi, R. Watanabe, Y. Natsume-Kitatani, H. Kawashima, C. Nagao, K. Mizuguchi, Computational model to predict the fraction of unbound drug in the brain, Journal of Chemical Information and Modeling, 59 (2019) 3251-3261.
31. T. Esaki, R. Ohashi, R. Watanabe, Y. Natsume-Kitatani, H. Kawashima, C. Nagao, H. Komura, K. Mizuguchi, Constructing an in silico three-class predictor of human intestinal absorption with Caco-2 permeability and dried-DMSO solubility, Journal of pharmaceutical sciences, 108 (2019) 3630-3639.
32. R. Watanabe, T. Esaki, R. Ohashi, M. Kuroda, H. Kawashima, H. Komura, Y. Natsume-Kitatani, K. Mizuguchi, Development of an In Silico Prediction Model for P-glycoprotein Efflux Potential in Brain Capillary Endothelial Cells toward the Prediction of Brain Penetration, Journal of Medicinal Chemistry, 64 (2021) 2725-2738.
33. D. Probst, J.-L. Reymond, A probabilistic molecular fingerprint for big data settings, Journal of cheminformatics, 10 (2018) 1-12.
34. H.O. Villar, R. Mandayan, M.R. Hansen, Molecular Diversity Assessment using Chemotypes, Current Computer-Aided Drug Design, 18 (2022) 1-8.
35. J. Williams, V. Siramshetty, D.-T. Nguyen, E.C. Padilha, M. Kabir, K.-R. Yu, A.Q. Wang, T. Zhao, M. Itkin, P. Shinn, Using in vitro ADME data for lead compound selection: An emphasis on PAMPA pH 5 permeability and oral bioavailability, Bioorganic & medicinal chemistry, (2022) 116588.

36. J.M. Gally, S. Bourg, Q.T. Do, S. Aci-Sèche, P. Bonnet, VSPrep: a general KNIME workflow for the preparation of molecules for virtual screening, Molecular informatics, 36 (2017) 1700023.
37. J.-P. Ebejer, G.M. Morris, C.M. Deane, Freely available conformer generation methods: how good are they?, Journal of chemical information and modeling, 52 (2012) 1146-1158.
38. T. Arakhamia, C.E. Lee, Y. Carlomagno, M. Kumar, D.M. Duong, H. Wesseling, S.R. Kundinger, K. Wang, D. Williams, M. DeTure, Erratum: Posttranslational Modifications Mediate the Structural Diversity of Tauopathy Strains (Cell (2020) 180 (4)(633–644. e12),(S0092867420301082),(10.1016/j. cell. 2020.01. 027)), Cell, 184 (2021) 6207-6210.
39. S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, Journal of computational chemistry, 29 (2008) 1859-1865.
40. J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B.L. De Groot, H. Grubmüller, A.D. MacKerell, CHARMM36m: an improved force field for folded and intrinsically disordered proteins, Nature methods, 14 (2017) 71-73.
41. J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B.L. De Groot, H. Grubmüller, A.D. MacKerell Jr, CHARMM36m: an improved force field for folded and intrinsically disordered proteins, Nature methods, 14 (2017) 71-73.
42. E.E. Ong, J.-L. Liow, The temperature-dependent structure, hydrogen bonding and other related dynamic properties of the standard TIP3P and CHARMM-modified TIP3P water models, Fluid Phase Equilibria, 481 (2019) 55-65.
43. B. Hess, H. Bekker, H.J. Berendsen, J.G. Fraaije, LINCS: a linear constraint solver for molecular simulations, Journal of computational chemistry, 18 (1997) 1463-1472.
44. D. Sindhikara, Y. Meng, A.E. Roitberg, Exchange frequency in replica exchange molecular dynamics, The Journal of chemical physics, 128 (2008) 01B609.
45. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chemical physics letters, 314 (1999) 141-151.
46. B. Devi, S.S. Vasishta, B. Das, A.T. Baidya, R.S. Rampa, M.K. Mahapatra, R. Kumar, Integrated use of ligand and structure-based virtual screening, molecular dynamics, free energy calculation and ADME prediction for the identification of potential PTP1B inhibitors, Molecular Diversity, (2023) 1-21.

47. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, GROMACS: fast, flexible, and free, Journal of computational chemistry, 26 (2005) 1701-1718.
48. B. Keller, X. Daura, W.F. Van Gunsteren, Comparing geometric and kinetic cluster algorithms for molecular simulation data, The Journal of chemical physics, 132 (2010) 02B610.
49. O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, Journal of computational chemistry, 31 (2010) 455-461.
50. M.M. Jaghoori, B. Bleijlevens, S.D. Olabarriaga, 1001 Ways to run AutoDock Vina for virtual screening, Journal of computer-aided molecular design, 30 (2016) 237-249.
51. B. Das, A.T. Baidya, B. Devi, T. Rom, A.K. Paul, B. Thakur, T. Darreh-Shori, R. Kumar, Synthesis, single crystal X-ray, DFT, spectroscopic, molecular docking studies and in vitro biological evaluation of compound N-benzyl-4-(4-chlorophenyl)-2-oxobutanamide, Journal of Molecular Structure, 1276 (2023) 134782.
52. N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, Journal of cheminformatics, 3 (2011) 1-14.
53. D.S. Goodsell, G.M. Morris, A.J. Olson, Automated docking of flexible ligands: applications of AutoDock, Journal of molecular recognition, 9 (1996) 1-5.
54. S. Sharma, A. Sharma, U. Gupta, Molecular Docking studies on the Anti-fungal activity of Allium sativum (Garlic) against Mucormycosis (black fungus) by BIOVIA discovery studio visualizer 21.1. 0.0, Annals of Antivirals and Antiretrovirals, 5 (2021) 028-032.
55. B.L. Jejurikar, S.H. Rohane, Drug designing in discovery studio, Asian J. Res. Chem, 14 (2021) 135-138.
56. V. Mulpuru, N. Mishra, In silico prediction of fraction unbound in human plasma from chemical fingerprint using automated machine learning, ACS omega, 6 (2021) 6791-6797.
57. A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Scientific reports, 7 (2017) 1-13.
58. B. Bakchi, A.D. Krishna, E. Sreecharan, V.B.J. Ganesh, M. Niharika, S. Maharshi, S.B. Puttagunta, D.K. Sigalapalli, R.R. Bhandare, A.B. Shaik, An overview on applications of SwissADME web tool in the design and development of anticancer, antitubercular

and antimicrobial agents: A medicinal chemist's perspective, Journal of Molecular Structure, (2022) 132712.

59. J. Lee, X. Cheng, J.M. Swails, M.S. Yeom, P.K. Eastman, J.A. Lemkul, S. Wei, J. Buckner, J.C. Jeong, Y. Qi, CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field, Journal of chemical theory and computation, 12 (2016) 405-413.
60. M.S. Valdés-Tresanco, M.E. Valdés-Tresanco, P.A. Valiente, E. Moreno, gmx_MMPBSA: a new tool to perform end-state free energy calculations with GROMACS, Journal of chemical theory and computation, 17 (2021) 6281-6291.
61. A.T. Baidya, B. Das, B. Devi, B. Långström, H. Ågren, T. Darreh-Shori, R. Kumar, Mechanistic Insight into the Inhibition of Choline Acetyltransferase by Proton Pump Inhibitors, ACS Chemical Neuroscience, (2023).
62. R. Salomon-Ferrer, D.A. Case, R.C. Walker, An overview of the Amber biomolecular simulation package, Wiley Interdisciplinary Reviews: Computational Molecular Science, 3 (2013) 198-210.