

Title: Synergistic Discovery of Novel Telomerase Inhibitors: From Structure-Based Pharmacophore Screening to Machine Learning Validation.

Introduction

Cancer remains a prominent global cause of mortality, impacting numerous individuals and placing significant strain on healthcare systems. Despite advancements in cancer therapy, there is an ongoing requirement for new and efficient therapies that can alleviate severe side effects caused by conventional treatments (Pucci et al., 2019). Cellular immortality, as envisaged by upregulated telomerase, is an established hallmark of all cancer cells (Hanahan & Weinberg, 2000). The potential of the enzyme as an anticancer drug target is underscored by its significant upregulation in more than 85% of cancer malignancies while it remains undetectable or present in minimal amounts in somatic cells (N. W. Kim et al., 1994; Shay, 1997). Developing drugs that target telomerase remains a hot area as it gives leverage to selectively disrupt the replication of cancer cells and induce their senescence or death meanwhile sparing the healthy cells.

Recent advances in structural biology have shed light on the complex architecture of telomerase, a large ribonucleoprotein complex that plays a critical role in cellular immortality. The catalytic core of telomerase is formed by the telomerase reverse transcriptase (TERT) and the telomerase RNA component (TERC), which together maintain the integrity of chromosome ends by adding repetitive DNA sequences to telomeres (Nguyen et al., 2018). The thumb domain of TERT, in particular, has emerged as a key target for therapeutic intervention due to its role in stabilizing the enzyme's interaction with telomeric DNA (Robart & Collins, 2011).

Recent advances in the structural biology of the enzyme have indicated that telomerase is a large ribonucleoprotein complex, architecturally bilobular in shape. The catalytic core is made of the telomerase reverse transcriptase (TERT) ribonucleoprotein complex and H/ACA lobe-ribonucleoprotein complex is majorly involved in the non-canonical roles of the enzyme (Nguyen et al., 2018). TERT, along with the shelterin complex of proteins, uses TERC as a template to add repeats of TTAGGG at the 3' end of the lagging strand of DNA, thereby circumventing the Hayflick limit and leading to the evolution of most cancer malignancies (Shay & Wright, 2000; Wright & Shay, 1992). Structurally, the catalytic TERT subunit of the enzyme in humans is mainly composed of four subunits, which are ordered from N-terminal to C-terminal as: Telomerase Essential N-terminal domain (TEN), Telomerase RNA Binding Domain (TRBD), Reverse Transcriptase domain (RT), and THUMB domains. The last three subunits form a TERT ring structure where the RNA template TERC integrates (Robart & Collins, 2011) (**Figure 1**).

Targeting either of the two telomerase subunits (TERT or TERC) has been the goal of anticancer medicine (Eckburg et al., 2020; Guterres & Villanueva, 2020; Jäger & Walter, 2016). Many telomerase-based therapies and drugs based on different mechanisms are currently under investigation. The most successful is the 13-mer modified oligonucleotide GRN163L (Imetelstat), which is in Phase I and Phase II clinical trials for multiple cancer subtypes (Eckburg et al., 2020). Other classes of inhibitors in trials include G-quadruplex

stabilizers, heat shock protein90 (HSP90) inhibitors, and tankyrase inhibitors. Additionally, alternative approaches such as telomerase peptide vaccines or adaptive T-cell therapy, have gained attention to target telomerase in cancer cells (Guterres & Villanueva, 2020). It is pertinent to mention here that despite extensive efforts and time invested in the cause, none of the small molecule telomerase inhibitors have been clinically approved yet. A highly selective small molecule telomerase inhibitor, BIBR1532 is still in the preclinical stages due to cytotoxicity issues (Ding et al., 2019). Likewise, other small molecule inhibitors like MST-312, TMPI, and costunolide are also in the preclinical stages (Vishwakarma et al., 2023). Therefore, the field of cancer treatment has opened a great avenue for advancing the preclinical and clinical development of small molecules as telomerase inhibitors. Developing small molecule inhibitors for targeted therapy has emerged as a compelling approach, owing to their myriad advantageous characteristics like low molecular weight, ease of oral administration, low cost, and efficient intracellular penetration to reach their target. These attributes, coupled with their ability to target telomerase directly, underscore their potential as effective anticancer agents.

Despite its discovery in 1984, the development of telomerase inhibitors has been hampered by significant challenges, including the lack of structural information (Greider & Blackburn, 1985; Welfer & Freudenthal, 2023). A breakthrough in the field came when the crystal structure of the *Tribolium castaneum* TERT (tcTERT) bound to a highly specific telomerase inhibitor, BIBR1532 was reported. (PDB ID: 5CQG). This study indicated the specificity of the inhibitor in binding to a conserved hydrophobic FVYL motif in the thumb domain of tcTERT. The FVYL pocket was found to be highly conserved in human TERT, and as a binding site for the P6.1 and CR4/5 loops of the RNA template (TERC) (Bryan et al., 2015).

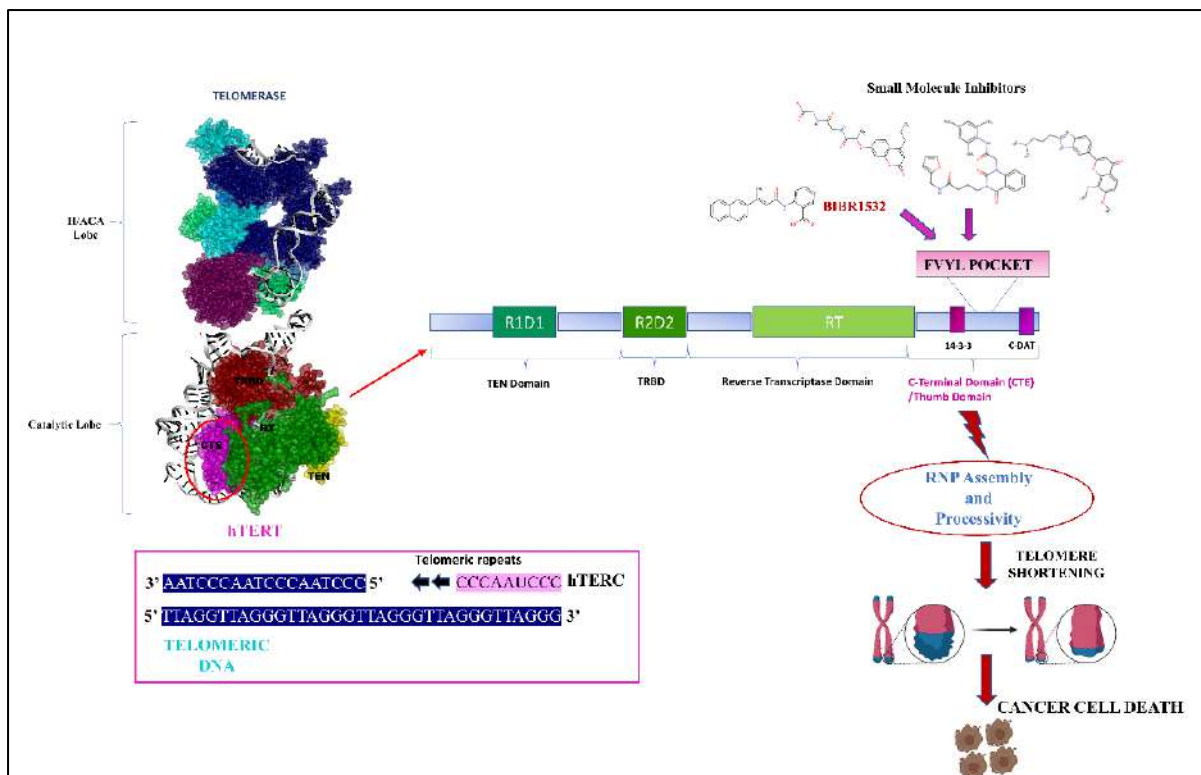


Figure 1. Schematic representation of Telomerase enzyme activity inhibited by small molecule inhibitors binding to the hTERT thumb domain will hamper its assembly and processivity to cause cancer cell death. The four domains of hTERT catalytic subunit from N-terminal to C-terminal end, i.e., TEN, TRBD, RT, and CTE/thumb domain, are illustrated.

Later in 2017, the Crystal structure of the human telomerase thumb domain was reported from 965 to 1122 residues with 2.61 Å resolution (PDB ID:5UGW) (Hoffman et al., 2017). The mutational studies performed in this work provided deep insights into the clinically relevant residues in the thumb domain of the enzyme and their role in a variety of telomeropathies. In addition, it also confirmed that the thumb domain is responsible for the processivity and assembly of the ribonucleoprotein complex (Hoffman et al., 2017). Thus, targeting the thumb domain will hamper the telomerase activity resulting in the cessation of tumour cell replication potential and eventual cell death.

In the present-day scenario, cues from the literature based on sequence similarity and phylogeny suggest that BIBR1532 binds to hTERT in the FVYL pocket in a manner similar to that of tcTERT, disrupting the crucial TERT-TER interaction (Bryan et al., 2015; Hoffman et al., 2017; B. Liu et al., 2022). However, due to the lack of a co-crystallized structure, it is still not confirmed.

In response to these challenges, the present study focuses on developing novel small-molecule inhibitors targeting the thumb domain of human telomerase. Leveraging the conserved FVYL pocket as a binding site, we employed a structure-based drug design strategy to generate pharmacophore models. We first attempted to understand how BIBR1532 binds to the pocket using *in silico* molecular docking followed by molecular dynamics studies. Subsequently, we followed a structure-based drug design strategy to develop pharmacophore models which were screened through the ChemDiv anticancer library, Otava drug-like green collection library, and Binding database. The top hits from each database were filtered based on three-level molecular docking energy scores and drug-likeness parameters. Finally, the hits were validated using molecular dynamics (MD) simulations and binding free energy calculations using MM-PBSA as well as the MM-GBSA method. These models were used to screen compound three libraries, leading to the identification of ten promising lead compounds with the potential to disrupt telomerase activity and inhibit tumor cell replication.

Recognizing the complementary strengths of pharmacophore modelling and machine learning (ML) in drug discovery, we implemented an integrated approach to enhance the robustness of our findings. Initially, pharmacophore-based virtual screening focused on identifying ten lead compounds targeting the thumb domain of telomerase from three diverse databases. These leads were then subjected to further validation using an ML classification model designed specifically for predicting telomerase inhibition across the entire enzyme. By applying the ML model to the top 10 proposed leads from the pharmacophore-based screening, we identified three compounds that demonstrated consistent inhibitory potential across both methodologies.

This two-tiered approach—starting with precise structural targeting and followed by comprehensive validation—not only solidifies the credibility of our results but also underscores the value of integrating multiple computational techniques in drug discovery. The

successful identification of the three novel telomerase inhibitors from large databases containing millions of compounds through this methodology represents a significant advancement in the development of targeted anti-cancer therapies, highlighting the potential of combining pharmacophore modelling with machine learning to optimize drug discovery efforts.

Objectives:

Objective 1: Structure based approach for identification of Novel telomerase inhibitors: Pharmacophore screening, Molecular docking and Molecular dynamics studies

- To investigate the potential of the thumb domain of the telomerase enzyme as a therapeutic target.
- To generate a structure-guided pharmacophore model for the thumb domain of the telomerase enzyme using BIBR1532 as a reference molecule.
- To identify hit compounds from novel compound libraries using a virtual screening approach as potential telomerase inhibitors.
- To eliminate false positives through secondary screening using various drug-likeness parameters, including ADME, Lipinski's Rule of Five, and TOPKAT toxicity prediction.
- To funnel down the lead molecules using molecular docking techniques, evaluating interactions with critical residues in the hTERT thumb domain, and shortlisting lead compounds based on binding energy scores and their biological mode of action.
- To validate the stability of selected lead compounds using molecular dynamics simulations and binding free energy studies with MM-PBSA and MM-GBSA methods.

Objective 2: Development of Machine learning model for identification of small molecules as telomerase inhibitors and Validation of Top Novel leads

- To develop machine learning models for the classification of telomerase inhibitors using molecular fingerprint descriptors.
- To create a high-quality dataset of known telomerase inhibitors with annotated activity values.
- To preprocess the dataset by generating molecular fingerprints that capture the structural characteristics of the compounds.
- To explore and evaluate different machine learning algorithms suitable for classification tasks, such as Random Forest, Support Vector Machines, XG-Boost etc.
- To validate the identified lead compounds using the ML classification model to predict telomerase inhibition across the entire enzyme, ensuring their effectiveness and consistency.
- Evaluation of ML model on test set, 5-fold CV and decoy set based on various parameters
- To advance the development of targeted anti-cancer therapies by integrating pharmacophore modelling with machine learning validation.

Material and Methods

Objective 1: Structure based approach for identification of Novel telomerase inhibitors: Pharmacophore screening, Molecular docking and Molecular dynamics studies

1.1 Protein preparation and binding site identification in hTERT

The crystal structure of *Tribolium castaneum* TERT (tcTERT) co-crystalized with BIBR1532 (PDB ID:5CQG) and human telomerase thumb domain (PDB ID:5UGW) with a resolution of 2.30 Å and 2.31 Å respectively, were retrieved from protein data bank (PDB) (Bryan et al., 2015; Hoffman et al., 2017). Protein structures were first prepared in Biovia Discovery Studio (DS) 2020 using the Prepare protein module to add missing hydrogen atoms and loops. All water molecules were removed from the structures. The thumb domains of tcTERT and hTERT were aligned and superimposed in Biovia DS 2020 following the sequence alignment studies reported based on structural and sequence similarity (Bryan et al., 2015). Based on the literature study, a hydrophobic FVYL pocket (Binding site I) located in the hTERT thumb domain was identified to be exploited in this study.

1.2 Molecular docking of BIBR1532 in FVYL pocket of hTERT (Binding site I)

The hydrophobic FVYL pocket identified in the hTERT thumb domain was defined as the binding site sphere for docking studies. BIBR1532 was extracted from the tcTERT structure, and its energy was minimized using the steepest descent method (2000 steps), followed by the conjugate gradient method (2000 steps) using the energy minimization tool of Biovia DS 2020. The ligand was docked into the FVYL pocket of hTERT (PDB ID: 5UGW) using LibDock, which is a high-throughput hotspot-based docking module in DS, wherein both the receptor and ligands are kept rigid. Alternatively, BIBR1532 was also docked using CDOCKER protocol, a CHARMM-based simulated annealing/molecular dynamics method available in DS 2020 (Rao et al., 2007; Wu et al., 2003). Finally, the docking was also performed using a flexible docking module which allowed both the receptor and ligand to be flexible with the highest level of accuracy (Rarey et al., 1996). The active site residues (Q1024, V1025, W1026, V1087, T1088, Y1089, L1090) were made flexible to dock BIBR1532 into the binding site.

1.3 Identification of new putative binding cavity (Binding site II).

Owing to the undesired results, the methodology was improvised. BIBR1532 was placed in the putative FVYL pocket of hTERT and in-situ ligand energy minimization was performed. The in-situ ligand minimization protocol minimizes the ligand in the presence of a rigid receptor. However, residues with atoms inside the specified binding sphere or flexible residues are allowed to move. A Molecular Dynamics (MD) simulation run of 20 ns was performed in the CHARMM36 forcefield to confirm the

stability of the ligand in the pocket. Subsequently, the simulation was extended up to 100 ns.

The protein topology of the complex was generated using pdb2gmx module in the CHARMM36 force field using the Gromacs 2020.1 package in the Linux environment. The ligand was fully hydrogenated using Avogadro software before generating its topology using an external CGenFF server (Hanwell et al., 2012; Vanommeslaeghe et al., 2010). The protein-ligand complex was then solvated using the TIP3P water model in a cubic box, which defined its periodic boundary conditions. Suitable concentrations of Na⁺ and Cl⁻ ions were added to the solvated system to neutralize the charges, followed by energy minimization using the steepest descent algorithm (50000 steps). The energy-minimized complex was equilibrated under NVT and NPT (constant number of particles, volume, temperature, and pressure) conditions for 100 ps each. The bond lengths were restrained using the LINear Constraint Solver (LINCS) algorithm and long-range electrostatics were calculated using the particle mesh Ewald (PME) method (Essmann et al., 1995). Short-range interactions were computed using a van der Waals cutoff at 1.2 nm (Vařeková et al., 2004).

Initially, a MD run of 20 ns was performed and snapshots were saved for every 10ps frame. Trajectory analysis was performed post MD and graphs were extracted for RMSD, RMSF, Radius of gyration, and Ligand RMSD specifically. The last 5% of the 20 ns MD simulation trajectory was manually analyzed to screen out the best pose with the lowest RMSD and maximum set of required interactions using PyMOL and Biovia DS 2020. The simulation run was later extended up to 100 ns to confirm the stability of BIBR1532 in the new site (Binding site II).

1.4 Revalidation of BIBR1532 stability in the binding site II by redocking

To ensure that BIBR1532 binds well onto the newly observed pocket. The selected protein-ligand complex pose from the MD simulation run was exported into the Biovia DS 2020 and the binding site sphere was defined at the exact pose with coordinates X= 61.148 Y=61.587 Z= 30.441 using the ‘define and edit binding site’ tool in DS. The ligand was prepared and docked onto this binding site using the CDOCKER module. Additionally, the binding free energy of the protein-ligand complex was evaluated using the MM-PBSA (Molecular Mechanics Poisson-Boltzmann surface area) implicit solvation method before and after simulation at site I and site II (Thompson et al., 2008).

1.5 Structure-based pharmacophore generation

A pharmacophore comprises steric and electronic features extracted from a ligand-binding site within a protein complex, serving as a framework for identifying ligands akin to a known selective inhibitor. To explore ligands resembling a selective inhibitor targeting the FVYL pocket within the thumb domain of human telomerase, we employed structure-based pharmacophore generation. While structural investigations

strongly suggest the binding of BIBR1532 to this pocket, the absence of its co-crystallized structure with hTERT led us to adopt pharmacophore modeling in a unique manner (B. Liu et al., 2022).

Our methodology involved two distinct approaches. In the first approach, pharmacophore models were generated from the pose at site I, wherein BIBR1532 was placed in the FVYL pocket and in-situ minimized ligand followed by using the Receptor-Ligand pharmacophore Generation (RLPG) module in Discovery Studio 2020. Excluded volumes were automatically generated to prevent steric clashes.

In the second approach, following a 20 ns MD simulation run on this complex, pharmacophore models were generated from the selected pose at site II, with the least RMSD and maximum interactions in the last 5% of the 20 ns MD simulation trajectory. Finally, the pharmacophore model with the required feature set and highest selectivity score predicted by genetic function approximation (GFA) was selected from the set of pharmacophores generated by both approaches, as described (Arooj et al., 2013). The models were named Pharmacophore A1 and Pharmacophore B1, corresponding to the two different pockets named site I and site II, respectively from which they were generated.

1.6 Validation of pharmacophores

A well-validated pharmacophore virtually screens out molecules with chemical properties equivalent to or better than the control ligand, ensuring high specificity and selectivity to the binding site (X. Liu et al., 2010).

The control ligand BIBR1532 was mapped onto pharmacophore models A1 and B1 using the ligand-pharmacophore mapping protocol of Biovia Discovery Studio 2020 to obtain a standard fit value for the control ligand with the pharmacophores. The fit value of the control BIBR1532 was subsequently used as a cut-off to filter top hits from a virtually screened set of compounds.

1.7 Pharmacophore-based virtual screening

The goal of building pharmacophore models is to search for multiple ligands that can bind to the same receptor with high selectivity and sensitivity. High-throughput virtual screening (HTVS) is crucial to the *in silico* drug discovery process because it enables quick and inexpensive screening of compound libraries that are substantially larger than those that can be screened experimentally, thereby accelerating the drug development process. Structurally diverse and large compound libraries containing novel compounds were chosen for this study. The best pharmacophore models A1 and B1 corresponding to site I and site II, respectively, were virtually screened through an anticancer ChemDiv library containing 61,538 compounds (<https://www.chemdiv.com>) and Otava drug-like green collection library containing 1,69,658 compounds (<https://www.otavachemicals.com>), using the screen library module and FAST method of conformation generation in Biovia DS 2020. Additionally, both the pharmacophores

were also screened through the Binding database containing 990 telomerase binding compounds using a similar protocol (<https://www.bindingdb.org>). BIBR1532 was also included in the dataset for revalidation. The screened hits from each pharmacophore model were filtered based on the fit value cutoffs of the control BIBR1532 i.e. ≥ 3.6 for model A1 and ≥ 2.8 for model B1, respectively.

1.8 Filtering of hit compounds for drug-likeness

ChemDiv anticancer library contains curated compounds with 3400 scaffolds with a diverse chemical space and screened for up to 10426 target-based and phenotypic screens. In contrast, Otava's drug-like green collection contains compounds pre-formatted by Lipinski's rule of five and curated specifically to exclude biologically unstable compounds. The library from the Binding database contained 990 telomerase binding compounds, most of which have been tested using inhibition in *in-vitro* assays but their exact binding mechanisms have not been established yet.

The hits comprised 1459 compounds from the ChemDiv library, 2518 from the Otava library, and 294 from the Binding database after screening by Pharmacophore model A1. Pharmacophore B1-based screening resulted in the identification of 2777, 8259, and 84 hits from these libraries respectively. All the screened hits underwent a rigorous filtering process to prioritize candidates with optimal drug-like properties. This comprehensive assessment utilized state-of-the-art computational tools available in Biovia DS 2020 (Moroy et al., 2012). The evaluation encompassed ADME predictions, which analyzed absorption, distribution, metabolism, and excretion profiles, Veber's rule for assessing molecular size and flexibility to ensure oral bioavailability, and Lipinski's rule of five criteria to confirm adherence to key physicochemical parameters associated with drug-likeness. Additionally, TOPKAT (Toxicity prediction using Komputer-assisted technology) analysis was employed to identify non-mutagenic leads and assess potential adverse effects based on structural alerts and chemical properties. In Computer-aided drug design (CADD), the computational characterization of pharmacokinetic properties and toxicity prediction are of vital importance, as they reduce the likelihood of rejection in clinical development. These properties influence the dosing, benefits, and adverse effects of drugs (Meibohm & Derendorf, 1997).

1.9 Molecular docking: three level screening approach

Molecular docking has become an indispensable tool in drug discovery, facilitating the identification of potential binding poses and estimating the strength of receptor-ligand interactions within the candidate binding sites.

Validation of hit compounds binding at sites I and II corresponding to pharmacophore models A1 and B1 respectively, was a crucial step in our study. Filtered ligands from the ChemDiv anticancer library, Otava drug-like green collection, and Binding database were docked using three types of docking protocols available in DS 2020 (Jha, Singh, et al., 2022). The first binding-site grid was defined at the putative FVYL pocket (site

I), using the Define and Edit tool of Biovia DS 2020. Similarly, a second binding-site grid was defined at site II, corresponding to Pharmacophore B1 in the hTERT thumb domain (PDB ID:5UGW). LibDock, a high-throughput rigid docking tool available in Biovia DS 2020, was used for the first-level screening. This docking protocol enumerates polar and apolar hotspots at protein-ligand interaction sites to predict their binding affinities (Rao et al., 2007). The docked poses were ranked and filtered based on the LibDock score of BIBR1532, which was included as a positive control in the docked datasets.

The top 200 docked hits based on the cut-off LibDock score of BIBR1532 for sites I and II were filtered, prepared, and energy-minimized using the steepest descent method (4000 steps) followed by the conjugate gradient method (4000 steps). The filtered ligands were docked to validate site I and the newly identified site II in the hTERT thumb domain using the CDOCKER module available in Biovia DS 2020. The docked poses were analyzed carefully using the 'analyse ligand pose' tool available in the receptor-ligand interaction module in DS 2020. As discussed previously, CDOCKER is a simulated-annealing-based docking protocol that employs the CHARMM force field to accurately dock ligands flexibly into the rigidly held receptor protein (Gagnon et al., 2016; Wu et al., 2003).

The top 1% ligands, which docked successfully in site II and had CDOCKER energy > -11.62 kcal/mol (control BIBR1532) were finally docked by using the flexible docking protocol available in DS 2020. F1012, V1025, Y1089, L1092, and T1088 residues were made flexible to carry out the third-stage docking to cross-validate the identified binding site. This process was used to funnel down the selective and high-affinity compounds from such large databases with the highest accuracy. The docked poses were ranked and filtered based on the flexible CDOCKER energy of the control BIBR1532 (-20.99 kcal/mol).

1.10 Molecular dynamic simulations and analyses

Molecular dynamics (MD) simulations were performed to ascertain the stability of the chosen candidate molecules in the presence of hTERT receptor interacting for a brief amount of time at the atomic level.

Finally, the top four selected hit molecules each from ChemDiv anticancer library (Compound IDs 26295, 26219, 28088, 45315) and Otava druglike green collection (Compound IDs 77574, 164867, 137187, 83601) and top two hits from Binding database (Compound 834, Compound 637) along with the control BIBR1532, in complex with hTERT thumb domain were subjected to a 100 ns MD run using Gromacs 2020.1 package with CHARMM36 forcefield using the same methodology as discussed in detail in section 2.3 above. (Brooks et al., 2009).

Trajectory analyses were performed after MD simulation in PyMOL software. Graphs for root mean squared deviation (RMSD), root mean square fluctuation (RMSF), Radius of Gyration (RoG) and the number of hydrogen bonds (H-bond) were extracted for all the ligand-protein complexes using Gromacs tools (gmx rms, gmx rmsd, gmx

gyrate, gmx hbond) on the Linux platform. Comparative graphs were plotted using the XMGrace software to understand the stability of the top leads from each database as compared to the control ligand in the hTERT thumb domain (Shaker et al., 2021). The crucial FVYL pocket interactions were evaluated in Biovia DS 2020 after the MD Simulation. RMSD graph indicates fluctuation of the protein backbone atoms in the presence of all the drugs in a binding pocket with time. In contrast, the RMSF graph shows individual residue fluctuations over the simulation period. The RoG graph is a measure of how well the protein structural compactness remains maintained throughout the simulation trajectory. H-bond graph represents the number of hydrogen bonds formed between the protein and ligand throughout the trajectory (Jha, Saluja, et al., 2022; Kant et al., 2022).

1.11 Binding free energy calculations (MM-PBSA and MM-GBSA)

Additionally, the validation of the MD simulation study was performed by enumerating the binding free energies of the selected top ten leads from the three databases and BIBR1532 using the gmx_MMPBSA package version 1.4 (Miller et al., 2012; Valdés-Tresanco et al., 2021). We extracted the last 20 ns of each MD run (80 ns to 100 ns) and removed all periodic boundary conditions as a precondition to calculate the Molecular Mechanics Poisson-Boltzmann surface area (MM-PBSA) scores. Analysis was performed using the gmx_MMPBSA_ana pipeline tool.

The binding free energies of the solvated protein-ligand system can be expressed using the following equation:

$$\Delta G_{\text{binding}} = \Delta G_{\text{complex}} - [\Delta G_{\text{protein}} + \Delta G_{\text{ligand}}]$$

where $\Delta G_{\text{complex}}$ is the receptor-ligand total free energy, and $\Delta G_{\text{protein}}$ and ΔG_{ligand} are the isolated total free energies of the receptor protein and ligand, respectively, in a solvent.

The free energy of the ligand-protein complex can be calculated as follows:

$$\Delta G_{\text{binding}} = \Delta H - T\Delta S$$

Or,

$$\Delta G_{\text{binding}} = \Delta E_{\text{MM}} + \Delta G_{\text{solv}} - T\Delta S$$

where ΔE_{MM} , ΔG_{solv} , and $T\Delta S$ are the changes in gas-phase molecular mechanics, solvation energy, and conformational entropy upon ligand binding, respectively.

ΔE_{MM} helps to calculate the bonded and non-bonded interactions using the following equation:

$$\Delta E_{\text{MM}} = \Delta E_{\text{bond}} + \Delta E_{\text{vdw}} + \Delta E_{\text{elec}}$$

Where ΔE_{bond} corresponds to the internal energy component of the system and ΔE_{vdw} and ΔE_{elec} are the Van der Waals and electrostatic interaction energies, respectively, corresponding to the nonbonding interactions in the MM force field parameters.

The solvation energy has polar and nonpolar components, which can be estimated using the following equation:

$$\Delta G_{\text{solv}} = \Delta G_{\text{polar}} + \Delta G_{\text{Non-polar}}$$

Or

$$\Delta G_{\text{solv}} = \Delta G_{\text{PB/GB}} + \Delta G_{\text{SA}}$$

The polar contribution was estimated using both the Poisson Boltzmann (PB) model and the Generalized Born (GB) model. (Wang et al., 2019). The nonpolar component of the system is usually estimated using the solvent-accessible surface area (SASA). Entropy (ΔS) has a minimal effect on the total energy; hence, it can be neglected.

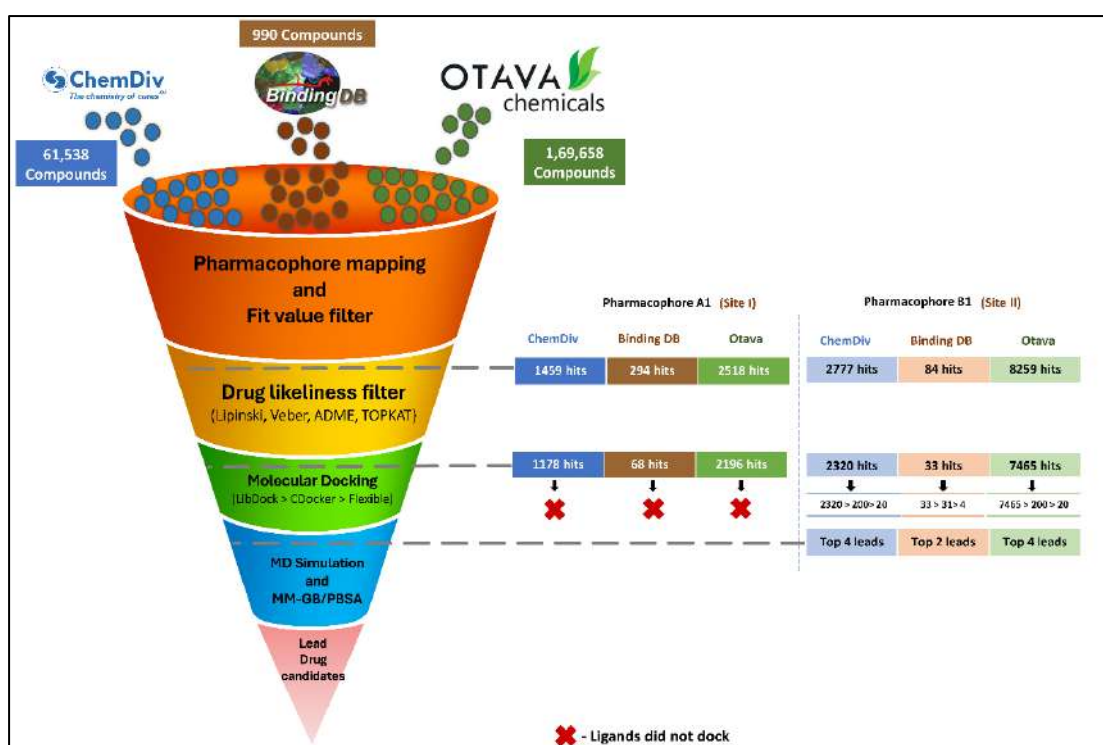


Figure 2. Schematic representation of stepwise Methodology for virtual screening and drug-likeness filtering of ChemDiv anticancer library, Otava druglike green collection, and Binding database at site I and site II.

Material and Methods

Objective 2: Development of Machine learning model for identification of small molecules as telomerase inhibitors and validation of top Novel leads

2.1. Computational Environment Details

Google Colaboratory (Bisong, 2019), a cloud-based Jupyter notebook environment from Google Research was used for all the computational studies and model development. A transparent and repeatable approach was made possible by the easy integration of code, visualizations, and documentation made possible by the interactive nature of Colab notebooks. Most of the necessary libraries used in this study were pre-installed on Google Colab (e.g., Scikit-learn, Pandas, etc.), and the other needed libraries were installed using the 'pip' command. All scripts for model construction were written in Python version 3.10.12 (*Machine Learning Made Easy: A Review of Scikit-Learn Package in Python Programming Language - Jiangang Hao, Tin Kam Ho, 2019, n.d.*).

2.2. Data Preparation

To establish target prediction and virtual screening methods, we framed the original problem as a classification task by labelling compound as 'active' or 'inactive' based on their reported biological activity. The schematic representation of the workflow for development of Machine learning model is depicted in **Figure 3**.

2.2.1. Data retrieval from ChEMBL database

The data on biological activity that has been reported with inhibitory activity (IC₅₀) values expressed as the half-maximal inhibitory concentration for Telomerase marked as 'SINGLE PROTEIN' were retrieved from ChEMBL database (Gaulton et al., 2012). The downloaded raw files in rich CSV format contained a plethora of parameters for target proteins and interacting small molecules, providing a thorough picture of their interactions across several data points.

2.2.2. Data wrangling

The CSV files was subsequently loaded in Google Colab for the data pre-processing step in a similar manner. Rows showing indeterminate outcomes, such as "inconclusive," "undetermined," or with "missing values" in the "chembl id" and "IC50" columns, were excluded from the analysis. During the data cleaning process, median values were computed for entries with multiple IC₅₀ values reported with the same ChEMBL ID to avoid data duplication and subsequent "data leakage" by preventing the inclusion of the same compounds within both the training and testing subsets.

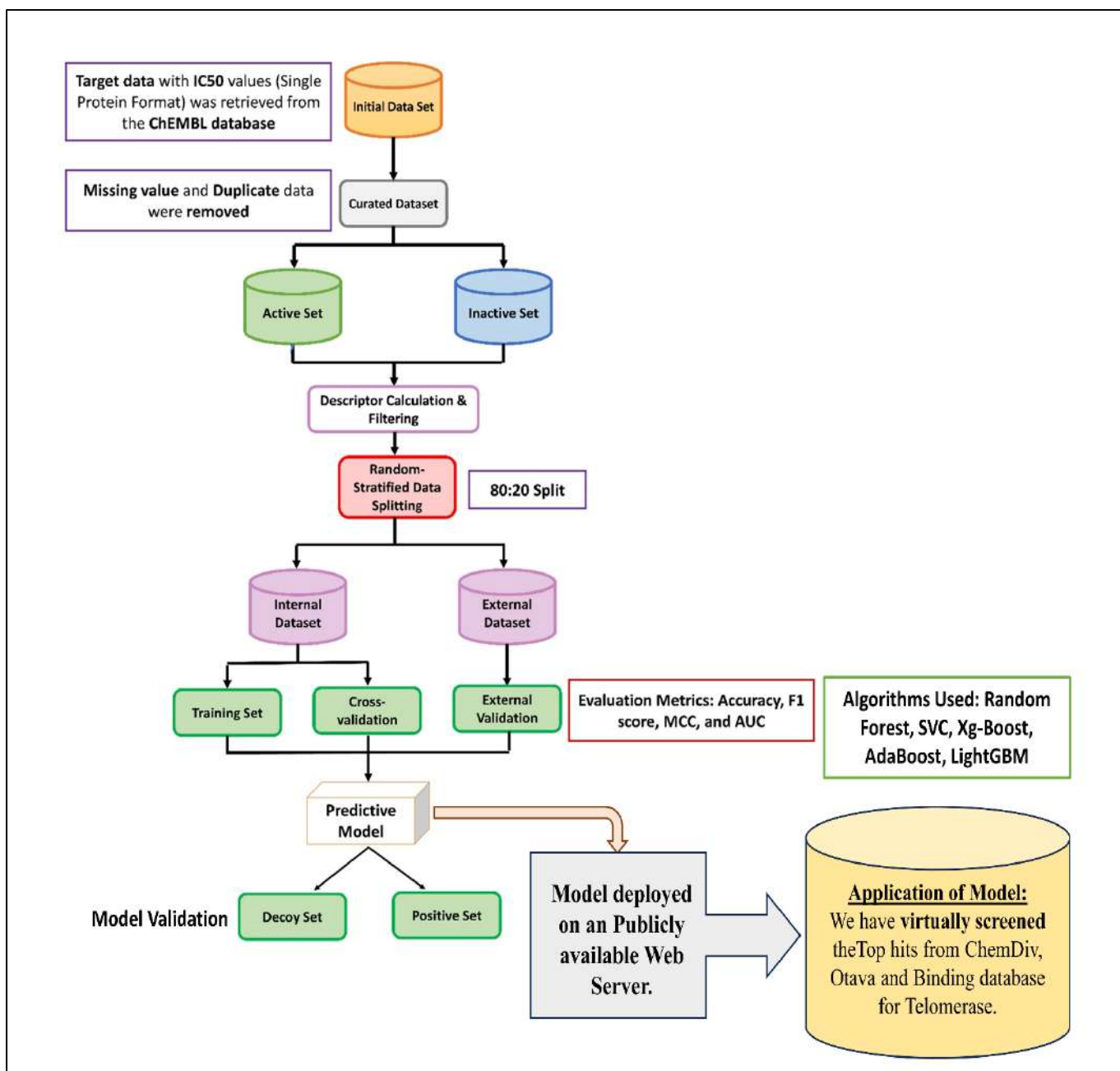


Figure 3. Schematic workflow of Machine Learning model development for Telomerase inhibitor prediction.

We framed the initial problem into a binary classification framework, distinguishing between "active" and "inactive" for the interactions between target and small molecules. To categorize the interactions, we transformed IC_{50} values from the standard value column into pIC_{50} , using the formula $pIC_{50} = -\log_{10}IC_{50}$. Following the transformation, compounds with a pIC_{50} value above 5 (equivalent to an IC_{50} at or below 10 μM), indicative of higher biological activity, were categorized as "active." Conversely, those with a pIC_{50} value of 4.9 or below (equivalent to an IC_{50} at or above 12.3 μM) were classified as "inactive." This distinction was made using the active threshold of $pIC_{50} = 5$ as a benchmark established by several foundational studies in the field. Given the complex nature of telomerase, the indirect measurement techniques used in

assays like Telomerase Repeat Amplification Protocol (TRAP), and the biological efficacy achieved at moderate potencies, a pIC_{50} cutoff (≥ 5) is appropriate (C. Liu et al., 2020; Vishwakarma & Bhatt, 2021). To create a definitive separation between active and inactive classifications, we excluded any data points with pIC_{50} values ranging from 5 and 4.9. This exclusion ensured a distinct threshold between the two categories. The file for each target was saved in “.csv” format.

2.3. Exploratory Data Analysis (EDA)

Chemical Space Visualization by Exploratory Data Analysis (EDA)

To analyze the distribution within the chemical space between active and inactive compounds of each target, six commonly utilized molecular descriptors in ADME/T were calculated using RDKit library (Bento et al., 2020). These descriptors included Molecular Weight (MW), octanol/water partition coefficient (ALogP), number of Hydrogen Bond Donors (NumHDonors), number of Hydrogen Bond Acceptors (NumHAcceptors), count of rotatable bonds, and the total polar surface area (TPSA). Various data visualizations, such as count plots, scatter plots, and box plots, were created to elucidate bioactivity class distribution and relationships among molecular descriptors.

2.4. Calculation of Molecular Fingerprints

The input file (containing only the two columns “smiles” and chembl_id”) was transformed from “.csv” to “.smi” for molecular fingerprint or descriptor calculation. SMILES (Simplified Molecular Input Line Entry System) represents the string of characters translated from 3’D structure of each molecule. We generated twelve types of molecular fingerprints using “python wrapper of PaDEL-Descriptor” padelpy version 0.1.14 (Yap, 2011). These included the Klekota-Roth (Klekota & Roth, 2008), CDK (fingerprinter, extended, and graph) (Steinbeck et al., 2003), PubChem (Helal et al., 2016), 2D AtomPairs (Carhart et al., 1985), Substructure (Capecchi et al., 2020), MACCS (Durant et al., 2002), and E-state (Hall & Kier, 1995).

The following arguments were chosen as “True” during the Padelpy fingerprint calculation: “fingerprints”, “removesalt”, “detectaromaticity”, “standardizenitro”, and “standardizetautomers”, keeping other settings as default.

2. 5. Machine learning (ML) model development and validation

We employed six statistical machine learning algorithms: Random Forests (RF) (Breiman, 2001), Support Vector Machines (SVM) (Heikamp & Bajorath, 2014), Extreme Gradient Boosting (XGBoost) (Babajide Mustapha & Saeed, 2016), Adaptive Boosting (Adaboost), Light Gradient Boosting Machine (LightGBM) (Ferreira & Figueiredo, 2012), and Classification and Regression Tree (CART) (Breiman et al., 2017), to efficiently and accurately classify compounds as active or inactive based on probability. To develop robust classification model in a streamlined manner, addressing class imbalances, a pipeline that included sequential transformations, sampling, and a final estimator was constructed. Intermediate processes, such as transformers employ fit, transform, and sample approaches, with sampling only used during the fitting process. The final estimator, however, only used the fit approach.

The developed ML pipeline for the Telomerase consisted of following sequential steps:

2.5.1. Feature Selection: The first step in feature selection was to use a Variance Threshold filter that removes features with low variance (Cutoff 0.1), which could lower the noise and help in better performance without any effect on the accuracy of the model. Another method used was the Recursive Feature Elimination (RFE) algorithm implemented along with ML models (classifiers) to eliminate those features not needed for model building; this feature elimination is done iteratively, therefore preserving the model's accuracy despite reducing the number of features (Li et al., 2017).

2.5.2. Class Imbalance Correction: The code included the option for imbalanced class correction using SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002). This technique is particularly relevant if the dataset contains a significant imbalance between active and inactive compounds.

2.5.3. Hyperparameter Tuning: A Randomized SearchCV algorithm was implemented to optimize the hyperparameters of each classifier. This approach involves evaluating various combinations of hyperparameter values (e.g., number of estimators, maximum depth) and selecting the configuration that yields the best performance on a validation set.

2.5.4. Model Training: For the classification objectives, a suite of algorithms from the Python scikit-learn library was utilized, encompassing Random Forests (RF), Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (Adaboost), Light Gradient Boosting Machine (LightGBM), and Classification and Regression Tree (CART). These algorithms were selected for their proven effectiveness in similar tasks. The dataset underwent a stratified division, allocating 80% for model training and 20% for evaluation, ensuring proportional representation of both active and inactive compounds. The integrity and consistency of the predictive models were rigorously examined via five-fold cross-validation protocol.

2.6. Model evaluation

To ensure the robustness and generalizability of the developed model, a rigorous external validation was carried out on a set of molecules that were completely new to the model creation process and evaluated on various parameters.

2.6.1. Model validation on test set and 5 fold CV

During the development stage, we used 5-fold cross-validation (5-CV) to evaluate the models' performance on training set. This meticulous technique resulted in a more robust evaluation than a single training-testing split. A wide range of statistical indicators were then used to objectively evaluate the performance of all six ML models on the training as well as the test dataset for each target protein.

These evaluation metrics included the overall accuracy, which reflects the proportion of correct predictions among the total number of cases; sensitivity, indicating the model's ability to correctly identify true positives; specificity, measuring the true negative rate; the F1 score,

which is the harmonic mean of precision and sensitivity; Matthew's correlation coefficient (MCC), offering a balanced measure that accounts for true and false positives and negatives; and the area under the receiver operating characteristic curve (AUC-ROC), which provides an aggregated measure of performance across all classification thresholds. These metrics collectively elucidate the nuanced aspects of the model's predictive power and its effectiveness in discriminating active from inactive compounds (Naidu et al., 2023).

2.6.2. Model validation using enrichment factor analysis

The screening potential of the finalized machine learning model for Telomerase was evaluated using inactive-enriched test data. For this, 15 active compounds were selected from the test dataset, and 300 property-matched decoys were generated from the DUD-E database (Mysinger et al., 2012), maintaining an active-to-decoy ratio of 1:20.

The effectiveness of the models was assessed by calculating the enrichment factor (EF) at various cutoffs—1%, 10%, 25%, 50%, and 100%—with particular attention to EF1%, which measures the model's ability to identify active compounds within the top 1% of ranked predictions. This metric is crucial for evaluating early enrichment in prospective virtual screening. EF_x% (where, x= 1, 10, 25, 50, 100) is determined using the formula: $EF_x\% = (A_x\%/A) \times 100$, where $A_x\%$ represents the number of active compounds in the top x% of ranked occurrences, and A is the total number of active compounds in the dataset.

2.7. Identification of Common Lead Compounds via Structure-Based and Machine Learning Screening

Virtual screening is a pivotal computational technique in drug discovery, enabling the identification of potential drug candidates from extensive compound libraries. To cross-validate the results obtained from our initial structure-based approach, specifically the pharmacophore-based virtual screening, we focused on the top 10 hits (Lead 1 to Lead 10) identified from the ChemDiv, Otava, and Binding databases as potential telomerase inhibitors targeting the thumb domain of hTERT. These selected compounds were further evaluated using our developed machine learning model designed to classify small molecules as telomerase inhibitors.

This dual approach was implemented to ensure the robustness of the selected lead compounds, thereby providing cross-validated candidates that can be confidently proposed as potential telomerase inhibitors for further in-vitro and ex-vivo assays. Moreover, it provides a broader context, ensuring that the compounds not only fit well in the thumb domain but are also likely to inhibit the enzyme's overall activity.

The raw files (in SDF format) of the top 10 compounds, labelled Lead 1 to Lead 10, sourced from ChemDiv, Otava, and Binding databases, were converted to SMILES format (.smi). Corresponding Klekota-Roth fingerprint descriptors were then calculated using PaDELpy.

Compounds were labelled as active or inactive based on their alignment with the machine learning model, using a probability threshold of 0.5. In general, compounds with a predicted probability score above 0.5 were considered promising hits, provided they also exhibited drug-like properties, which are critical for early-stage drug discovery

Figure 4. A) Identification of FVYL binding pocket in hTERT by structural superimposition with tcTERT having co-crystallized BIBR1532 ligand. B) Sequence

alignment of tcTERT and hTERT thumb domains. FVYL residues are highlighted in red colour.

3.2 In-situ ligand minimization of BIBR1532 in FVYL pocket (Binding site I) and MD simulation

In an improvised methodology, BIBR1532 was placed in the putative FVYL pocket (site I) of the human telomerase thumb domain identified by superimposing the structure with tcTERT, and in-situ energy minimization of the ligand. Reported thumb domain residues, such as V1016, L1017, F1012, V1025, Y1089, L1092, T1088, and F1032, showed interactions with BIBR1532. Moreover, the in-situ energy drastically reduced from +4.17374e+09 kcal/mol to -552.737 kcal/mol, indicating that the ligand was stable inside the pocket.

To further confirm the stability of BIBR1532 in this pocket, the complex was extracted in this pose and an initial 20 ns molecular dynamics (MD) simulation run was performed. We observed that the ligand shifted to an adjacent pocket immediately after the NVT equilibration step and remained stable throughout the simulation trajectory in the newly occupied pocket now onward designated as binding site II in the whole manuscript (**Figure 3**). The simulation run was further extended up to 100 ns to confirm the stability of the complex. RMSD, RMSF, Radius of gyration graphs, and specifically the ligand RMSD graph showed that the protein-ligand complex remained stable throughout the trajectory even up to 100 ns (**Figure 5**).

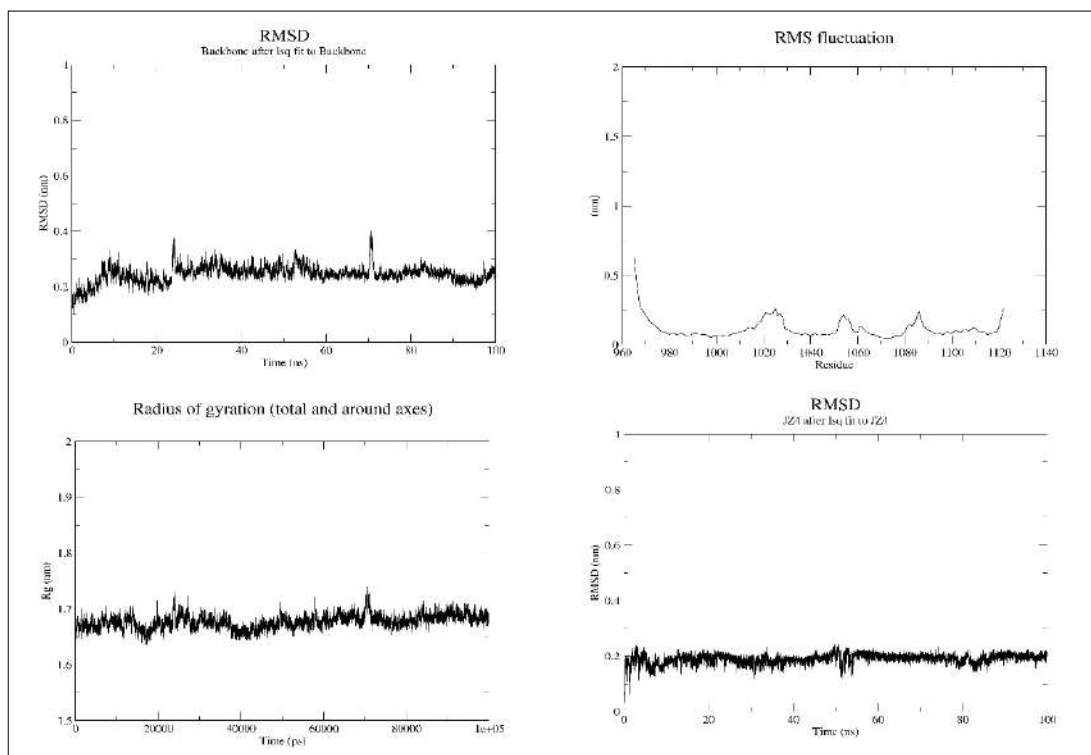


Figure 5. Plots to investigate binding stability of BIBR1532 in the new putative FVYL pocket (site II) after 100ns MD simulation A) represents RMSD (protein-ligand) B) RMSF C) Radius of gyration D) Ligand RMSD

Poses before and after the MD simulation were superimposed, and it was observed that the new pocket (site II) was located away from the putative FVYL pocket (site I), although the ligand still interacted with most of the important residues and was bound in the same orientation as before (**Figure 3**). The binding site RMSD measured 2.67 Å, falling within the expected range. This value also suggested a slight expansion of the binding site to accommodate the ligand, which remained stable within this new pocket throughout the trajectory. Notably, a conformational change in the hTERT binding site became evident after the MD simulation, which had potentially caused the ligand to shift from site I to site II while still interacting with FVYL residues (**Figure 3B**). The putative FVYL pocket and the newly occupied pocket were referred to as site I and site II, respectively.

3.3 Revalidation of BIBR1532 stability in the binding site II by redocking

The validation of BIBR1532 binding affinity in site II was critical because this new binding pocket away from the putative FVYL pocket (site I) was discovered as a result of an MD simulation study. In the next set of experiments, BIBR1532 was docked successfully directly into the site-II pocket, revealing the same set of interactions with FVYL and other essential residues as observed after the MD simulation. The observed CDOCKER energy ranged from -11.62 kcal/mol to -7.62 kcal/mol corresponding to ten poses of the BIBR1532 obtained after the docking. The negative energy obtained

confirmed that the ligand was stable in the new binding pocket. Moreover, the energy bracket helped us decide on a cut-off value at an upper limit (-11.62 kcal/mol) to filter the molecular docking results after the screening in later stages. The binding free energies of the BIBR1532-hTERT complex calculated using the MM-PBSA method at site II after simulation and redocking at site II were -26.805 kcal/mol and -26.455 kcal/mol respectively. The energy for both the poses was comparable and significantly more negative than that of the docked complex at site I before simulation with an energy of -10 kcal/mol. This revalidated that the shift of BIBR1532 from its presumed FVYL pocket (site I) to the new pocket (site II) was not by chance and that it is most likely its actual binding site. Also, the protein underwent a conformation change to accommodate the ligand at an adjacent wider site (**Figure 6**).

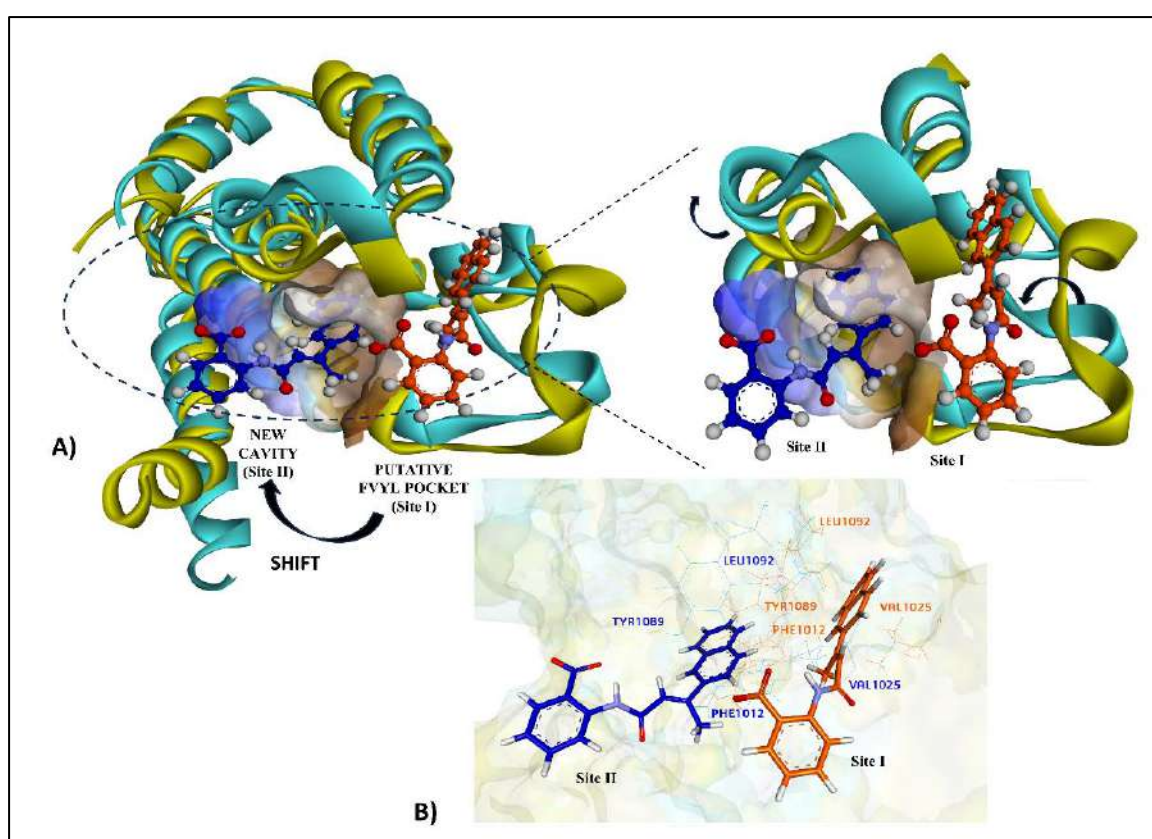


Figure 6. A) hTERT thumb domains superimposed before (yellow) and after (cyan) the MD simulation run. A Shift in the binding cavity of BIBR1532 from a putative FVYL pocket (orange) to a new cavity (blue) was observed. The zoomed view shows the change in the conformation of the binding cavity in hTERT after MD simulation leading to the shift of ligand from site I (putative FVYL pocket) to site II (new pocket) B) Relative shift in the positions of FVYL residues in the binding cavity.

3.4 Structure-based pharmacophore generation

Structure-based pharmacophore models were generated using a two-way approach in the Biovia DS 2020. We obtained six pharmacophore models from the first approach, where BIBR1532 was placed in the putative FVYL pocket (site I), and *in situ* energy

minimization was carried out. Similarly, we obtained ten pharmacophore models from the second approach, wherein the pharmacophore was generated from the selected pose after a 20 ns MD simulation run (site II) (**Figure 7** and **Table 1**).

In approach one, the pharmacophore with the highest selectivity score of 7.1958 and feature set AAHHH (two hydrogen bond acceptors and three hydrophobic groups) was selected (Pharmacophore A1 in **Table 1A**). In the second approach, the top two pharmacophore models had a similar selectivity score of 9.1336, which was the highest. However, the critical distinguishing factor was their feature set (Pharmacophore B1 in **Table 1B**). We purposely chose a pharmacophore model with the feature set AHHNR (one hydrogen bond acceptor, two hydrophobic groups, one negative ionizable group, and one ring aromatic feature) in contrast to AHHHN (**Figure 7**). This choice was made with respect to the previously reported mode of interaction of BIBR1532 in the 3D space of the binding site within the protein. The aromatic ring group of BIBR1532 is an essential contact point between the ligand and protein, making it an essential feature for choosing the correct pharmacophore model (Bryan et al., 2015).

The selected pharmacophore models A1 and B1, corresponding to the two possible binding sites of BIBR1532, site I and site II respectively, with the highest selectivity scores and suitable feature sets, were subsequently used as queries for virtual screening. A good pharmacophore model is sensitive enough to identify potential novel ligands that can bind to a pocket and are selective enough to avoid false positive candidates. This is possible only by controlling the number of features that must be possessed. A pharmacophore model with 4-6 features is usually considered good, as in our case (Kant et al., 2022).

Table 1. Selectivity score and characteristic features of pharmacophores generated **A)** by first approach (A1 to A6) **B)** by second approach (B1 to B10)

Pharmacophore	Number of Features	Feature set	Selectivity score
Pharmacophore A1	5	AAHHH	7.1958
Pharmacophore A2	4	AHHH	5.6810
Pharmacophore A3	4	AHHH	5.6810
Pharmacophore A4	4	AAHH	5.6810
Pharmacophore A5	4	AAHH	5.6810
Pharmacophore A6	4	AAHH	5.6810

A)

Pharmacophore	Number of Features	Feature set	Selectivity score
Pharmacophore B1	5	AHHNR	9.1336
Pharmacophore B2	5	AHHHN	9.1336
Pharmacophore B3	4	HHNR	7.9596

Pharmacophore B4	4	HHHN	7.9596
Pharmacophore B5	4	AHNR	7.9596
Pharmacophore B6	4	AHHN	7.9596
Pharmacophore B7	4	AHNR	7.9596
Pharmacophore B8	4	AHHN	7.9596
Pharmacophore B9	4	AHHN	7.9596
Pharmacophore B10	4	AHHR	5.7496

B)

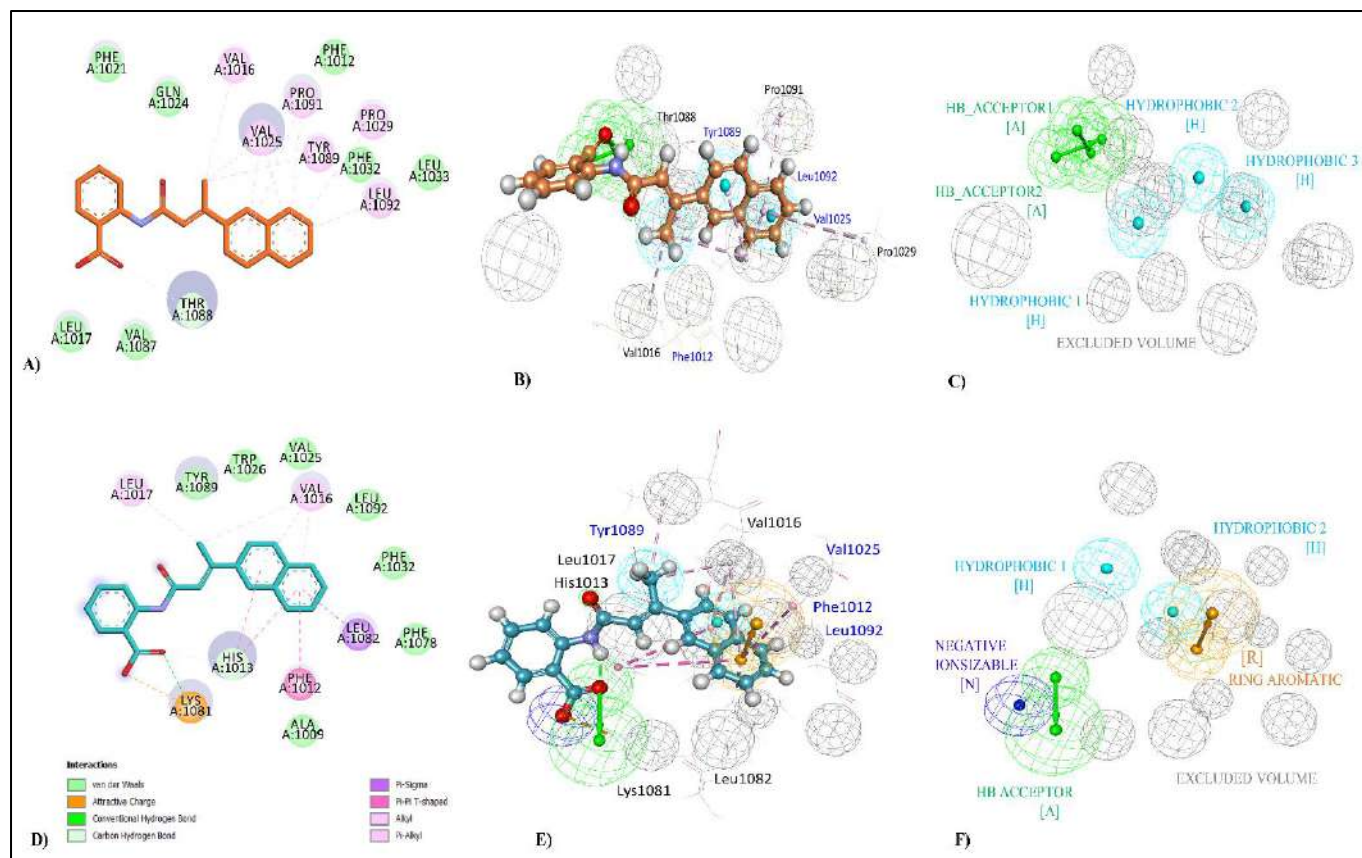


Figure 7. Structure-based pharmacophore models selected from approach 1 (A-C) and approach 2 (D-F). A) 2D interaction diagram of BIBR1532 in-situ energy minimized in FVYL pocket (site I) B) Pharmacophoric features in the binding pocket C) Best pharmacophore model A1 with a feature set AAHHH D) 2D interaction diagram of BIBR1532 in the new pocket (site II) after 20 ns MD simulation E) Pharmacophoric features in the new binding pocket F) Best pharmacophore model B1 with a feature set AHHNR.

3.5 Validation of Pharmacophores

Validation of pharmacophore is an important step to determine the fit value by which filtering of top hits should be performed after virtual screening. This fit value is indicative of how well the features in the pharmacophore map match the chemical features present in BIBR1532 while keeping track of the excluded volume. When

BIBR1532 was mapped onto the two pharmacophore models, we obtained a fit value of 3.65 and 2.85 for the models A1 and B1 respectively. This fit value was subsequently used as a cut-off to filter top hits from a virtually screened set of compounds by both the pharmacophore models, which should be ideally better than the control inhibitor (Jha, Saluja, et al., 2022).

3.6 Pharmacophore-based virtual screening

Structurally novel and potential lead molecule identification from diverse chemical databases is possible through high-throughput virtual screening (HTVS) of the generated pharmacophore models. The best pharmacophore models A1 and B1 from each approach were screened against a total of 61538 compounds in the ChemDiv anticancer library (<https://www.chemdiv.com>), 169658 Otava drug-like green collection (<https://www.otavachemicals.com>) and 990 telomerase binding compounds in the Binding database (<https://www.bindingdb.org>).

As a result of the key interactions at sites I and II, a total of 52413 and 54765 molecules from the ChemDiv anticancer library mapped to pharmacophore models A1 and B1, respectively. They were filtered further based on the fit value cut off ≥ 3.6 and ≥ 2.8 , respectively (obtained by mapping BIBR1532), resulting in 1459 and 2777 molecules in each case. Similarly, 122570 and 132727 molecules from Otava drug-like green collection, mapped to the pharmacophore models A1 and B1, resulting in 2518 and 8259 molecules, respectively. From the Binding database, 619 and 637 molecules mapped to pharmacophore models A1 and B1, respectively that resulted in the identification of 294 and 84 compounds from the Binding database. (**Figure 2 flowchart**).

3.7 Filtering of hit compounds for drug-likeness

Pharmacokinetics (PK) study involves the study of the absorption, distribution, metabolism, and excretion (ADME) of drugs in the body. Determining the PK properties is a crucial step in the drug development process to maximize drug efficacy and safety, optimize dosage regimens, and reduce the likelihood of failure in subsequent stages of the process. Drug-likeness filters such as Lipinski's rule of five, Veber's rule, and ADME were applied to the screened hits from all three selected databases.

1288 and 2457 molecules from the ChemDiv anticancer library passed Lipinski's rule, Veber rule, and ADME filter corresponding to pharmacophore models A1 and B1 respectively. Similarly, 2428 and 8034 molecules from Otava drug-like green collection and 71 and 39 molecules from the Binding database corresponding to pharmacophore models A1 and B1 respectively passed all three filters. The ADME tool helps to eliminate compounds with unfavourable features in terms of human intestinal absorption, blood-brain barrier penetration, hepatotoxicity, aqueous solubility, CYP2D6 binding, plasma protein binding, and undesired functional groups. Finally, after TOPKAT filtering, 1178 ChemDiv molecules, 2196 Otava molecules, and 68

Binding database molecules corresponding to pharmacophore model A1 and 2320 7465, and 33 molecules from each database corresponding to pharmacophore model B1 were left. TOPKAT® (TOXicity Prediction by Komputer Assisted Technology) tool in Discovery Studio helps to assess ames mutagenicity of a compound from its 2D structure alone (Kant et al., 2022) (**Figure 2**).

3.8 Molecular docking

Molecular docking is a very important part of the structure-based drug discovery pipeline. It helps to determine the binding affinity of a particular ligand to a receptor molecule based on a ligand placement algorithm and scoring each pose in which the ligand docks successfully to a protein. 1178 hits from ChemDiv, 2196 hits from Otava, and 68 hits from Binding database when docked onto the site-I, corresponding to the Pharmacophore A1 showed unfavourable non-bonded interactions with the most critical residues in the FVYL pocket (site I), preventing any molecule to properly dock there either by LibDock or by CDOCKER protocol (**Figure 8**).

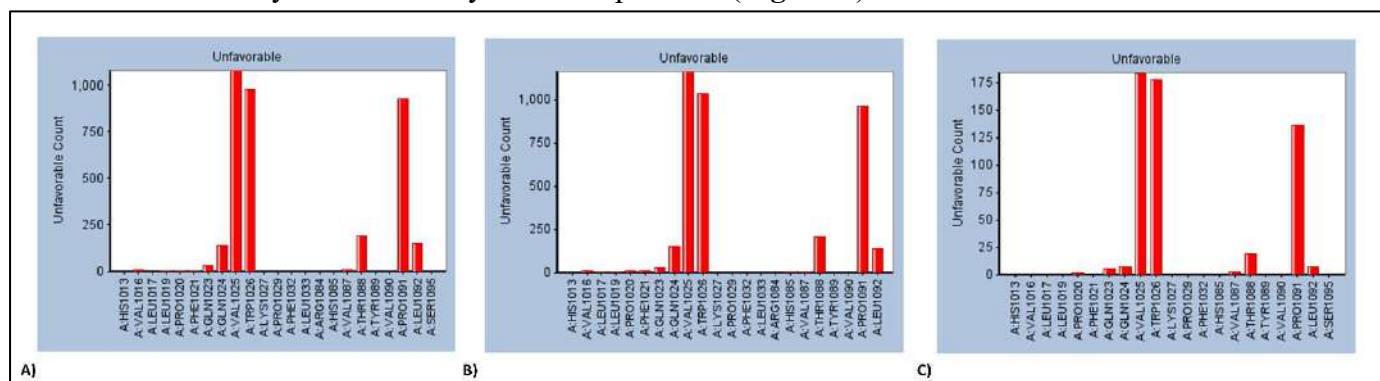


Figure 8. Residue interaction histogram of all the poses analyzed after docking compounds at site I in hTERT thumb domain showing unfavourable bumps with crucial residues in the binding cavity (red). A) ChemDiv B) Otava and C) Binding database

Critical examination revealed that residues VAL1025, TRP1026, PRO1029, THR1088, TYR1089, and PRO1091 of hTERT protrude inside the putative FVYL pocket (site I), decreasing the effective volume of the cavity and making it impossible for any molecule from all the three databases to dock there. These residues also showed steric clashes with the placed BIBR1532, making it impossible to dock inside the pocket whereas the corresponding residues in tcTERT kept the pocket open and accessible (**Figure 9**).

In contrast to this, 2320 compounds from ChemDiv, 7465 compounds from Otava, and 33 compounds from Binding database could dock into site II corresponding to the Pharmacophore B1 using LibDock protocol (first level screening). This pocket though away from the putative FVYL pocket showed almost all essential FVYL interactions in addition to other interactions. The docked hits from each database were filtered based on a better LibDock score as compared to the positive control BIBR1532 (LibDock score = 79.69). Subsequently, the top 200 hits from large ChemDiv and Otava libraries

and 31 hits from the Binding database were further prepared, minimized, and docked using simulated annealing-based CDocker protocol (second level screening). 144 hits from ChemDiv, 166 hits from Otava, and 30 hits from the Binding database docked and were filtered based on CDOCKER energy more than that of control BIBR1532 (≥ -11.62 kcal/mol). The top 1% hits from ChemDiv and Otava databases and 5 hits left from the Binding database when docked using a more accurate flexible docking protocol (third level screening), produced 450 poses, 399 poses, and 60 poses respectively.

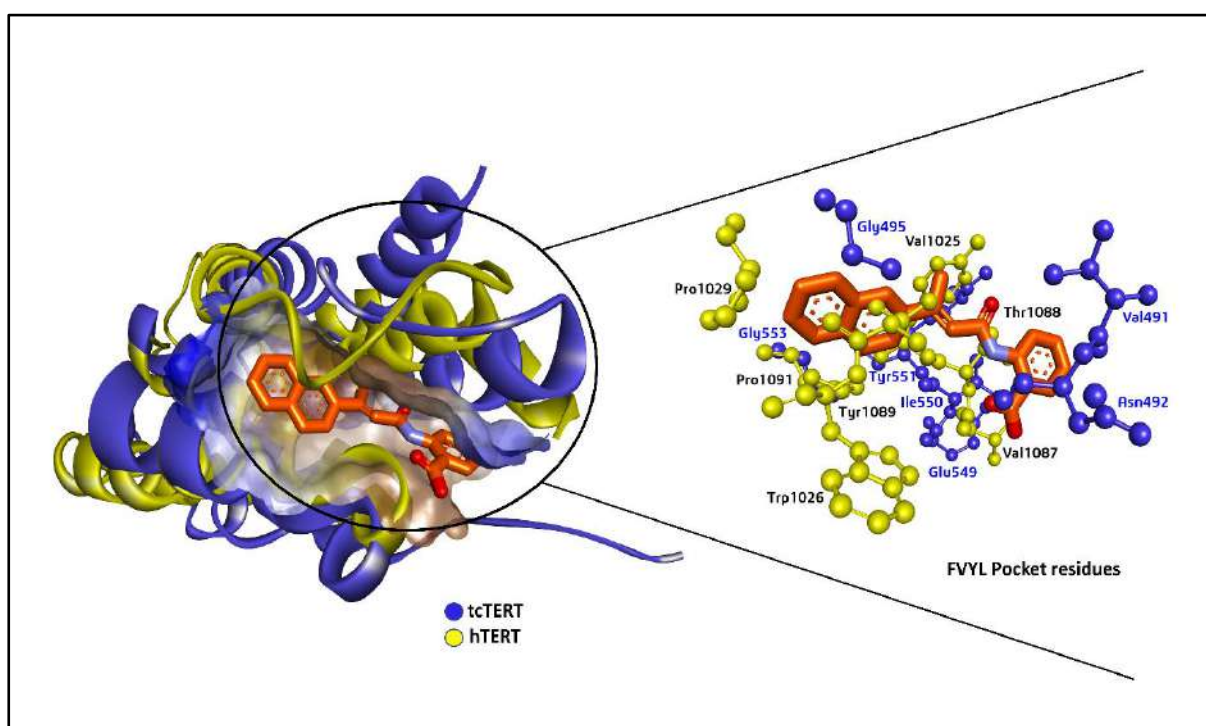


Figure 9. Structural superimposition of tcTERT and hTERT with BIBR1532 (orange) in FVYL pocket. Labelled residues of hTERT (yellow) protrude inside the FVYL pocket decreasing the effective pocket volume whereas tcTERT residues (blue) keep the pocket in open conformation.

F1012, V1025, Y1089, L1092, and T1088 were made flexible to provide conformational flexibility to the protein and if the ligand maintains its interactions with these crucial residues. The CDOCKER energy scores tremendously improved after flexible docking for most of the hits from the three databases respectively. Finally, the top ten compounds comprising four hits each from ChemDiv anticancer library and Otava drug-like green collection, two hits from the Binding database with better flexible CDOCKER energy as compared to BIBR1532 (-20.99 kcal/mol), showing interactions with FVYL and other crucial residues identified from the literature were chosen as leads for further molecular dynamics studies and binding free energy calculations. The chemical structures, known biological activity, and other parameters of these compounds are illustrated in **Table 2**.

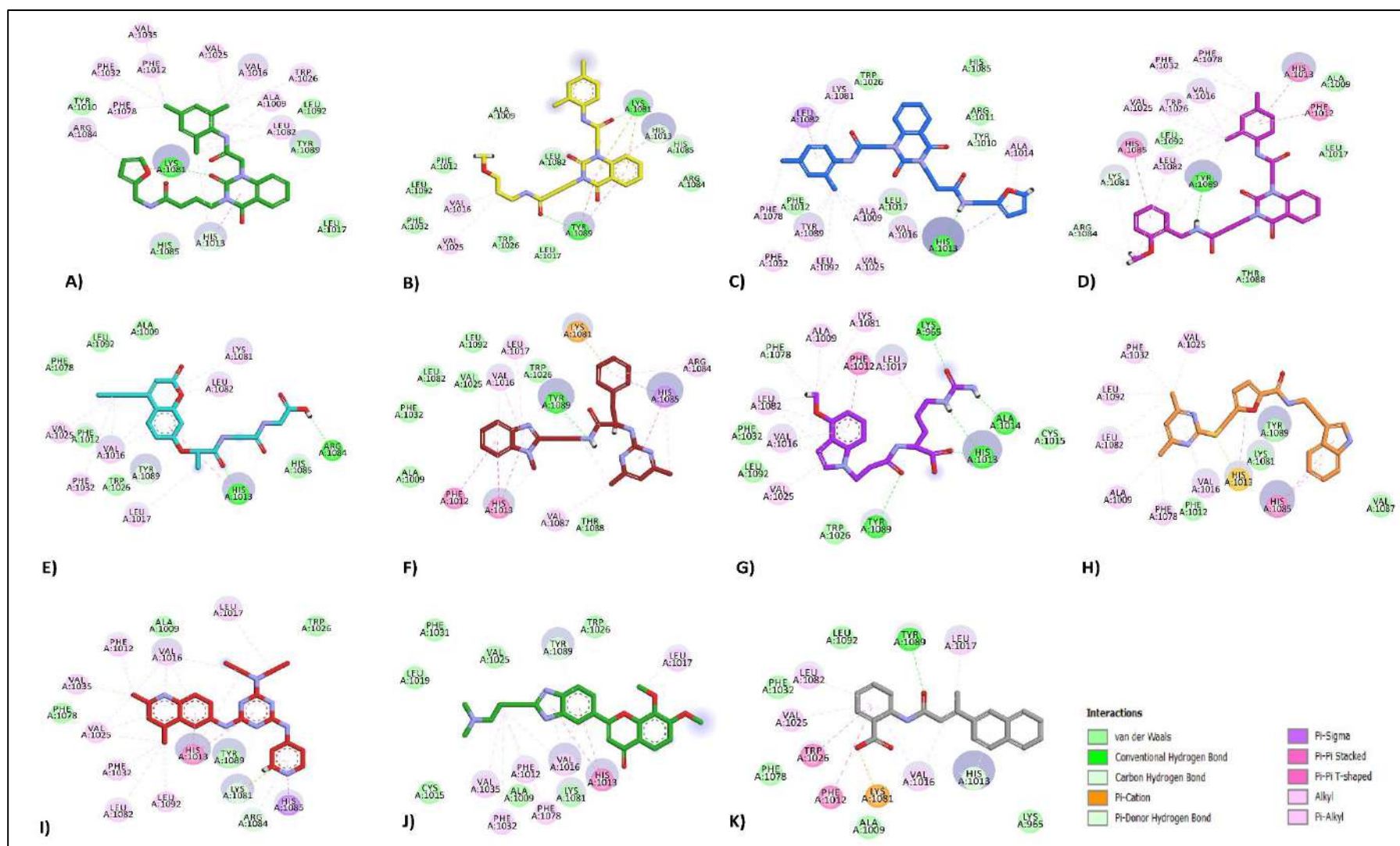


Figure 10. 2D interaction diagrams of top leads from **A-D)** Lead 1, 2, 3,4 from ChemDiv anticancer library **E-H)** Lead 5,6,7,8 from Otava drug-like green collection **I-J)** Lead 9,10 from Binding database **K)** BIBR1532 (control)

Table 2. Top ten shortlisted lead compounds from ChemDiv anticancer library, Otava drug-like lead collection and Binding database with their fit value, docking scores, and crucial interacting residues.

<i>Name</i>	<i>Compound IDs And Database</i>	<i>Fit value</i>	<i>First level screening (LibDock score)</i>	<i>Second Level screening (-CDOCKER energy)</i>	<i>Third level screening (Flexible -CDOCKER energy)</i>	<i>Class</i>	<i>Interacting residues</i>	<i>Type of interaction</i>
Lead 1	26295 (ChemDiv library)	3.00398	127.551	42.3433	51.9919	Anticancer	A1009, <u>F1012</u> , H1013, V1016, <u>V1025</u> , W1026, F1032, V1035, F1078, K1081, L1082, R1084	Hydrophobic
							K1081	H-bond
							TYR1010, LEU1017, H1085, <u>Y1089</u> , <u>L1092</u>	Van der Waals
Lead 2	26219 (ChemDiv library)	3.02468	123.978	40.8855	45.5744	Anticancer	H1013, V1016, <u>V1025</u> , Y1089, K1081	Hydrophobic
							A1009, K1081, <u>Y1089</u>	H-bond
							<u>F1012</u> , L1017, W1026, F1032, L1082, R1084, H1085, <u>L1092</u>	Van der Waals
Lead 3	28088 (ChemDiv library)	2.90623	124.713	41.9005	45.3444	Anticancer	A1009, A1014, V1016, <u>V1025</u> , F1032, F1078, K1081, L1082, <u>Y1089</u> , <u>L1092</u>	Hydrophobic
							H1013, Y1010	H-bond

Lead 4	45315 (ChemDiv library)	2.85486	125.236	38.4081	44.058	Anticancer	<u>F1012</u> , L1017, R1011, W1026, H1085	Van der Waals
							<u>F1012</u> , H1013, V1016, <u>V1025</u> , W1026, F1032, F1078, L1082, H1085	Hydrophobic
							K1081, R1084, <u>Y1089</u>	H-bond
							A1009, L1017, T1088, <u>L1092</u>	Van der Waals
Lead 5	77574 (Otava Chemicals)	3.20966	126.866	33.3204	51.53	Drug like leads	H1013, V1016, L1017, <u>V1025</u> , F1032, K1081, L1082	Hydrophobic
							H1013, R1084, <u>Y1089</u>	H-bond
							A1009, <u>F1012</u> , W1026, F1078, H1085, <u>L1092</u>	Van der Waals
Lead 6	164867 (Otava Chemicals)	2.94493	128.87	37.0166	45.031	Drug like leads	<u>F1012</u> , H1013, V1016, L1017, K1081, R1084, H1085, V1087	Hydrophobic
							<u>Y1089</u>	H-bond
							A1009, <u>V1025</u> , W1026, F1032, L1082, T1088, <u>L1092</u>	Van der Waals
Lead 7	137187 (Otava Chemicals)	3.336	124.409	40.6183	45.767	Drug like leads	<u>A1009</u> , F1012, V1016, L1017, <u>V1025</u> , K1081, L1082	Hydrophobic
							K965, H1013, A1014, F1078, <u>Y1089</u>	H-bond
							W1026, F1032, <u>L1092</u>	Van der Waals

Lead 8	83601 (Otava Chemicals)	3.00106	125.174	39.0021	45.0864	Drug like leads	A1009, H1013, V1016, <u>V1025</u> , F1032, F1078, L1082, H1085, <u>L1092</u> , <u>F1012</u> , L1081, <u>Y1089</u> , V1087	Hydrophobic Van der Waals
Lead 9	Compound 834 (Binding database)	3.15788	93.5414	31.5322	34.2981	Telomerase inhibitor (IC ₅₀ = 3.38 μM)	<u>F1012</u> , H1013, V1016, L1017, <u>V1025</u> , F1032, V1035, L1082, R1084, H1085, <u>L1092</u> , K1081, R1084, A1009, W1026, F1078, <u>Y1089</u>	Hydrophobic H-bond Van der Waals
Lead 10	Compound 637 (Binding database)	3.10202	97.628	14.8486	24.5634	Telomerase inhibitor (IC ₅₀ = 0.47 μM)	<u>F1012</u> , H1013, V1016, L1017, V1035, F1032, F1078, <u>Y1089</u> A1009, C1015, L1019, <u>V1025</u> , W1026, F1031, K1081	Hydrophobic H-bond Van der Waals
Positive Control	BIBR1532	2.85945	79.6947	1.6247	20.9923	Well-established telomerase inhibitor (IC ₅₀ = 100 nM)	<u>F1012</u> , V1016, L1017, <u>V1025</u> , W1026, K1081, L1082 H1013, <u>Y1089</u> K965, A1009, F1032, F1078, <u>L1092</u>	Hydrophobic H-bond Van der Waals

#Source: ChemDiv anticancer library (<https://www.chemdiv.com/>), Otava drug-like green collection (<https://otavachemicals.com/>), Binding database (<https://www.bindingdb.org/>)

3.9 Molecular dynamics simulations

Molecular dynamics was performed for 100 ns using an explicit solvent model to get further insights into the stability and dynamic behavior of the lead compounds inside the newly discovered site II in the thumb domain of hTERT. It helped us to gain a comprehensive understanding of the behavior of the lead compounds from each database in comparison to the control ligand BIBR1532 based on metrics like Root mean squared deviation (RMSD), Root mean squared fluctuation (RMSF), Radius of gyration (RoG) and hydrogen bonds formed between protein and ligand throughout the trajectory (**Figure 11**).

RMSD plot for backbone atoms of the protein in the presence of each compound helped us to assess their structural and conformational stability in the binding pocket during simulations. A lower RMSD value indicates that the system has achieved a state of equilibrium, maintaining the stability and structural integrity of the protein-ligand complex. In contrast, a higher RMSD value suggests a significant deviation of the simulated structure from the reference structure. All the lead compounds from the ChemDiv anticancer database (Lead 1, Lead 2, Lead 3, and Lead 4) showed RMSD less than 0.4 nm which converged steadily to as low as 0.2 nm for Lead 1 and Lead 2 (**Figure 11A**). Lead 5 and Lead 8 from the Otava drug-like green collection library showed a similar trend to the control BIBR1532 with the least RMSD converging near 0.2 nm. Lead 7 displayed an RMSD that was steady up to 80 ns at 0.2 nm, but it converged at 0.4 nm and remained within the acceptable range. In contrast, lead 6 showed the most fluctuating RMSD, reaching up to 0.9 nm and converging near 0.5 nm indicating some conformational change in the protein (**Figure 11B**). Lastly, lead 9 from the Binding database showed stable RMSD similar to control BIBR1532 converging near 0.2 nm, and Lead 10 showed a stable RMSD up to 30 ns, however, it converged near 0.4 nm still lying within the acceptable range (**Figure 11C**).

RMSF plots helped us to gain further insights into the flexibility of each amino acid residue of the thumb domain of hTERT in the presence of each lead compound in the binding pocket. Higher RMSF values in the binding site residues suggest that the complex undergoes conformational changes or fluctuations associated with ligand binding. The RMSF values for all the lead compounds from the ChemDiv database complexed with the protein fell within an acceptable range. However, leads 3 and 4 displayed fluctuations of up to 0.5 nm near residues L1056, G1057, A1058, and K1059, which were not part of the binding cavity (**Figure 11D**). Similarly, RMSF plots for all the lead compounds from the Otava library showed an average RMSF of 0.2 nm, a trend similar to that of the control BIBR1532-hTERT complex. Although, lead 6 showed a sharp fluctuation up to 1 nm in the region of S1055, L1056, G1057, A1058, K1059, G1060, and A1061 residues which were however, not a part of the binding site including the FVYL residues (**Figure 11E**). Lead 9 and Lead 10 from the Binding database showed a stable RMSF plot comparable with the control ligand, with an average RMSF of 0.3 nm lying in the acceptable range (**Figure 11F**).

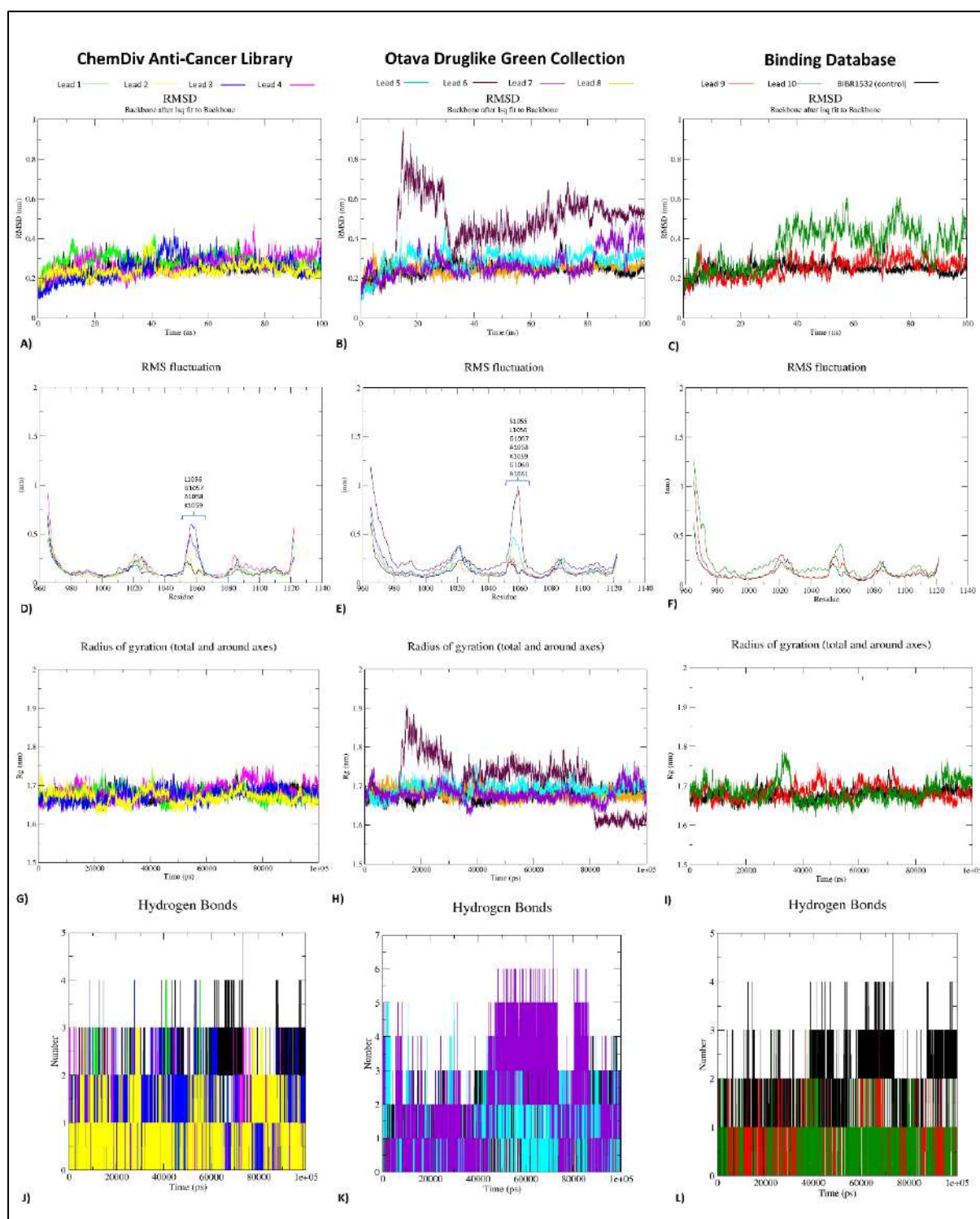


Figure 11. 100 ns MD Simulation graphs of the top lead compounds selected from ChemDiv database, Otava drug-like green collection and Binding database representing A-C) RMSD D-F) RMSF G-I) Radius of gyration J-L) H-bond analysis

Next, we analyzed the radius of gyration (RoG) plots extracted from MD trajectories to study the protein compactness and folding dynamics in the presence of lead compounds. The RoG values of all complexes from the ChemDiv database were observed to be an average of 1.7 nm, which was low and within the acceptable range. All complexes

showed stable protein-folding dynamics, with their compactness maintained throughout the trajectory (**Figure 11G**). Similarly, Lead 5, Lead 7, and Lead 8 from the Otava database showed a stable average RoG of 1.7 nm similar to that of the control BIBR1532. In contrast, Lead 6 exhibited a comparatively high fluctuating RoG which peaked up to 1.9 nm initially and gradually lowered down to 1.7 nm up to 80 ns. The fluctuation steeply decreased and converged near 1.6 nm at the end of the run. The dynamic fluctuations throughout the trajectory of lead 6 suggest a lack of conformational stability and possible unfolding of the protein due to ligand binding (**Figure 11H**). Finally, Leads 9 and 10 from the Binding database showed a similar trend as control BIBR1532, with an average RoG of 1.7 nm within the acceptable range (**Figure 11I**).

Finally, we analyzed the number and duration of hydrogen bonds formed by all the lead compounds from the three databases and BIBR1532 in the binding pocket of the protein throughout the 100 ns simulation run. Hydrogen bonds are a stronger and more stable type of interaction that determines the stability and binding affinity of a ligand inside the binding pocket of a protein. All ChemDiv leads formed a minimum of 3 hydrogen bonds throughout the trajectory. Notably, lead 1 formed up to four hydrogen bonds, similar to the control BIBR1532 (**Figure 11J**). In contrast, Lead 5 and Lead 7 from the Otava database formed up to five and seven hydrogen bonds, respectively, being more than the control. This indicated that they have a stronger binding affinity for the protein than lead 6 or 8 from the same database (**Figure 11K**). Finally, in the Binding database, lead 9 formed at least three hydrogen bonds for the majority of the trajectory in comparison to lead 10 and the control ligand, which formed up to four hydrogen bonds (**Figure 11L**). However, further validation studies were performed using binding free energy calculations of the protein-ligand complexes.

3.10 Binding free energy calculations

The binding free energies of the top ten shortlisted lead compounds from the ChemDiv, Otava, and Binding databases and BIBR1532 were evaluated for the last 20 ns trajectory to gain a better understanding of their stability and binding affinity in the thumb domain binding site. Both MM-PBSA and MM-GBSA binding free energy methods were employed to validate the consistency and reliability of the calculations. PB models are usually more accurate but computationally expensive, and GB models are faster but less accurate; however, both rely on different solvation models and hence can help cross-validate the results. All ten lead complexes and the positive control BIBR1532 exhibited negative effective binding free energies, albeit with varying degrees of affinity. We identified the top five lead compounds through comparative cross-validation analysis, prioritizing those with the most negative binding free energy estimates in both categories, indicating more robust binding interactions. In comparison to the control, which has MM-GBSA and MM-PBSA estimates of -32.64 ± 3.35 kcal mol⁻¹ and -52.56 ± 4.32 kcal mol⁻¹, respectively, the top five lead compounds exhibited promising binding characteristics (**Figure 12**). Lead 1 from ChemDiv database exhibited the most negative MM-PBSA and a comparable MM-GBSA estimate to the

control BIBR1532, indicating the strongest affinity with the protein in comparison to other leads. On the other hand, Lead 2 exhibited a more negative ($-34.5 \text{ kcal mol}^{-1}$) MM-GBSA score than the control and a moderately high MM-PBSA score ($-23.5 \text{ kcal mol}^{-1}$) comparable to Lead 10 which is a known telomerase inhibitor, suggesting a robust binding with the hTERT. Lead 3 from ChemDiv database and Lead 5, and 6 from Otava database also exhibited comparable MM-GBSA estimates to the control BIBR1532 in addition to comparable MM-PBSA to Lead 10, suggesting strong binding interactions. Thus, the top five novel lead compounds from the ChemDiv anticancer library and Otava drug-like green collection, shortlisted based on their energetics, are Lead 1, Lead 2, Lead 3, Lead 5, and Lead 7 (**Table 3**). These compounds show promise as potential candidates with a high binding affinity for hTERT, as indicated by their comparatively higher estimated binding free energy scores compared to the control. Additionally, the energy estimates are comparable to that of Lead 10 (ChEMBL ID: CHEMBL376279), which is a known telomerase inhibitor.

Furthermore, Lead 10 from the Binding database exhibited reasonably high binding free energy in both methods, suggesting a strong affinity to the thumb domain of the telomerase enzyme complex. In comparison, Lead 9 showed a lower binding affinity with the thumb domain having the least MM-GBSA as well as MM-PBSA scores. A detailed analysis of the energetic contribution of each type of polar and nonpolar interaction to the effective binding free energy exhibited by each lead compound is given in **Tables S4** and **S5**.

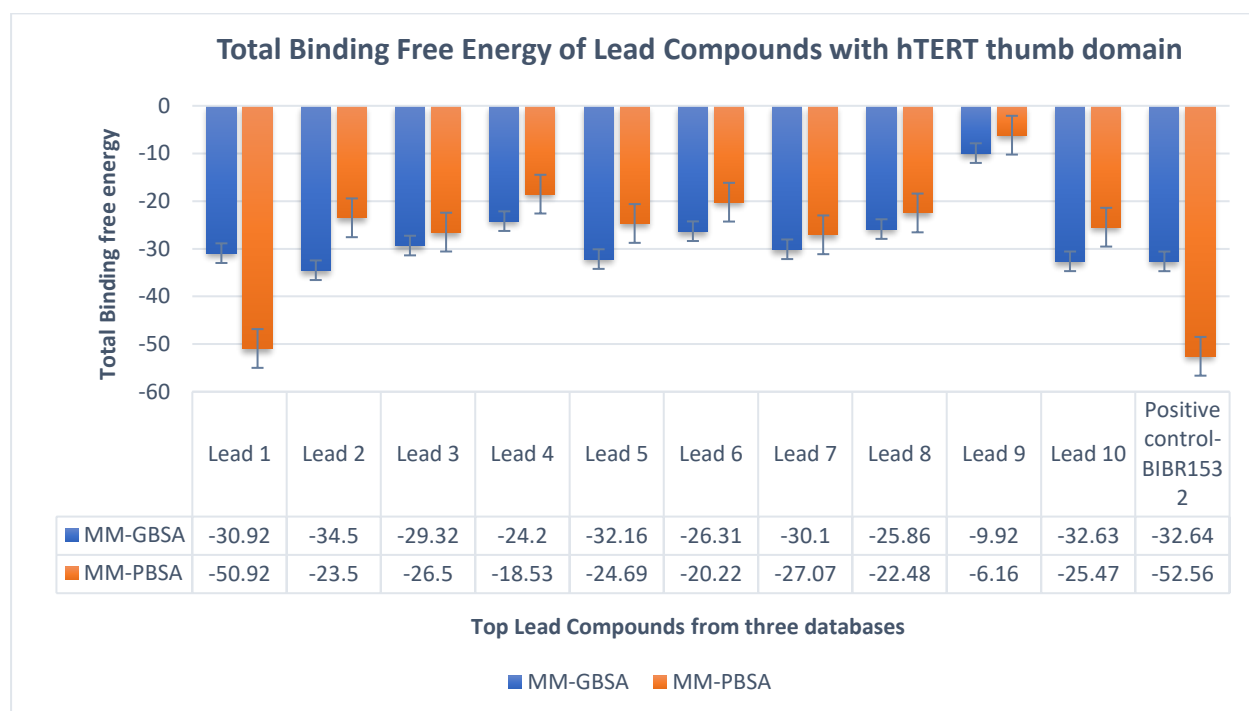
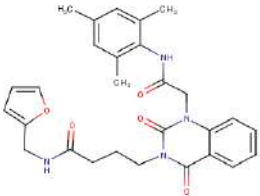
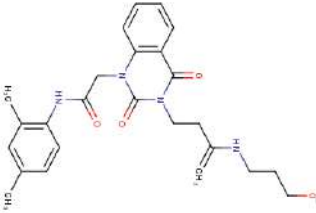
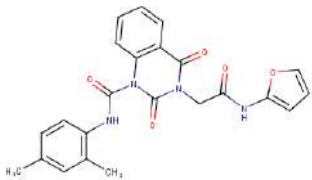
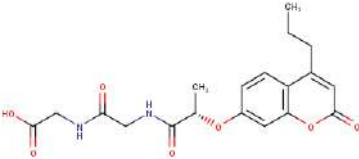
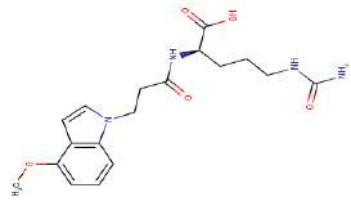
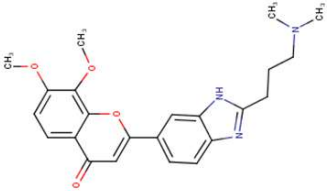


Figure 12. Binding free energy calculations using MM-GBSA (Blue) and MM-PBSA (orange).

Table 3. Top shortlisted leads from ChemDiv, Otava, and Binding databases.

<i>Name</i>	<i>Compound ID and Database</i>	<i>2D structure</i>	<i>Passed ADME and Veber rule filter</i>	<i>Toxicity profile</i>	<i>Known Biological profile</i>
Lead 1	26295 (ChemDiv library)		YES	Non-mutagen	Anticancer
Lead 2	26219 (ChemDiv library)		YES	Non-mutagen	Anticancer Has been checked for 519 targets until now but is active for only Cyt P450-family 2, subfamily C and member 19 (CYP2C19) and -member 9 (CYP2C9) (S. Kim et al., 2022)
Lead 3	28088 (ChemDiv library)		YES	Non-mutagen	Anticancer
Lead 5	77574 (Otava Chemicals)		YES	Non-mutagen	Drug like lead
Lead 7	137187 (Otava Chemicals)		YES	Non-mutagen	Drug like lead
Lead 10	Compound 637 (Binding database) CHEMBL376279		YES	Non-mutagen	Telomerase inhibitor with IC ₅₀ = 0.47 μM.

Results

Objective 2: Machine learning Model development for Telomerase inhibitor prediction and Validation of top leads from ChemDiv, Otava and Binding database.

4.1 Dataset Analysis

The bioactivity data for Telomerase retrieved from the ChEMBL database consisted of 388 compounds data. Subsequently, after removing duplicate entries, indeterminate outcomes, and missing values, the unique data points for each target were kept. The dataset analyses based on pIC_{50} for all the targets showed a bell-shaped distribution as depicted in **Figure 13A**. The remaining unique molecules for each target were divided into active, and inactive based on their pIC_{50} cut-off values for binary classification. as shown in **Table 4** and **Figure 13B&C**. To clearly distinguish between classes and eliminate noise, molecules with activities near the fuzzy borders were omitted during the creation of the dataset for model development.

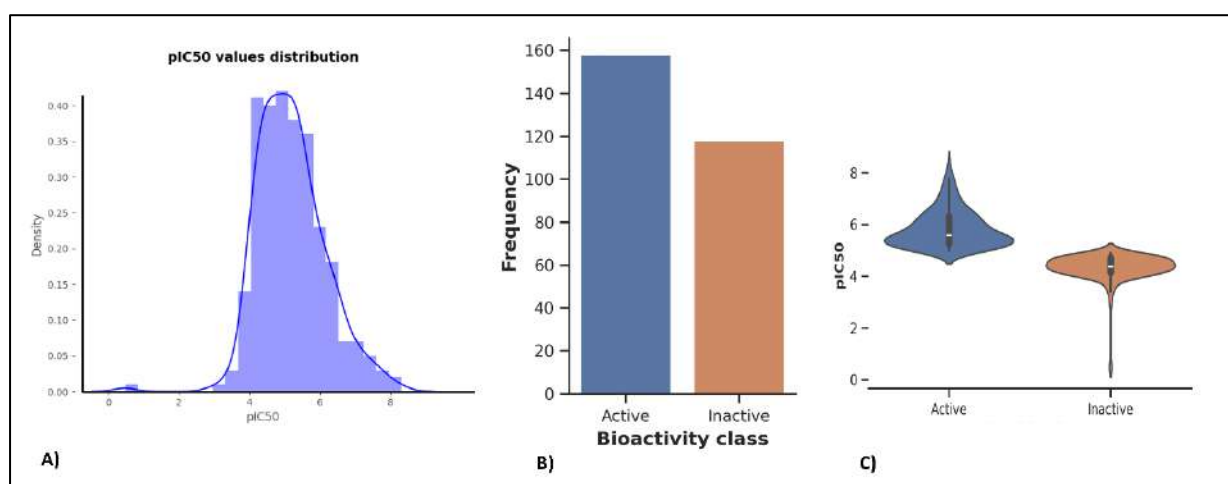


Figure 13. A) Bell shaped distribution of ChEMBL dataset for Telomerase inhibitors based on their pIC_{50} values. B) Frequency of active and inactive telomerase inhibitors after application of pIC_{50} cutoff. C) Distribution of active and inactive dataset

Table 4. Detail of the dataset used in this study

Target	Raw dataset	Unique dataset	Dataset after the removal of borderline compounds	Dataset for model development	
				Active ($\text{pIC}_{50} \geq 5$)	Inactive ($\text{pIC}_{50} \leq 4.9$)
Telomerase	388	281	276	158	118

4.2. Exploratory Data Analysis (EDA)

The initial stage of this study involved conducting exploratory data analysis (EDA) to gain insights into the physicochemical properties of the pre-processed dataset for Telomerase.

The chemical space distribution of the active and inactive datasets was analyzed using the cheminformatics library RDKit for the widely used molecular descriptors such molecular weight (MW), octanol/water partition coefficients (logP), no. of hydrogen acceptors (HBA), no. of hydrogen donors (HBD), no. of heavy atoms, total polar surface area (TPSA), number of rotational bonds, and aromaticity (Zuccotto, 2003) (**Figure 14**).

Molecular weight (MW) is an important feature that affects a compound's solubility, cell permeability, and overall drug-like properties. Octanol/water partition coefficient, often known as log P, is a measure of a molecule's lipophilicity, which influences its ability to pass biological membranes and distribute throughout the body (Al-Lazikani et al., 2007). The number of hydrogen bond acceptors and donors determines a substance's ability to generate favourable interactions with target proteins, regulating binding affinity and selectivity (Chen et al., 2016; Kenny, 2022). The number of heavy atoms and total polar surface area provides information on the size and polarity of the molecule, which might influence absorption, distribution, metabolism, and excretion (ADME) qualities. The molecule's flexibility is reflected by the number of rotational bonds which is an important factor in molecular recognition and binding. Subsequently, aromaticity is a structural characteristic that can affect a compound's pharmacokinetic and pharmacodynamic effects (Bissantz et al., 2010; Chen et al., 2016; Meyer et al., 2003).

We conducted comparisons between active and inactive compounds using Mann-Whitney U test ($p < 0.05$) based on these properties. Our analysis revealed that for Telomerase, all properties except LogP and the number of Heavy atoms differed significantly between actives and inactive. This suggests that the active and inactive samples are from different populations and can be easily distinguished based on a majority of molecular descriptors and certain physicochemical properties correlate with bioactivity. A violin plot for each of these eight properties between active and inactive datasets for all the targets is shown in **Figure 14A**.

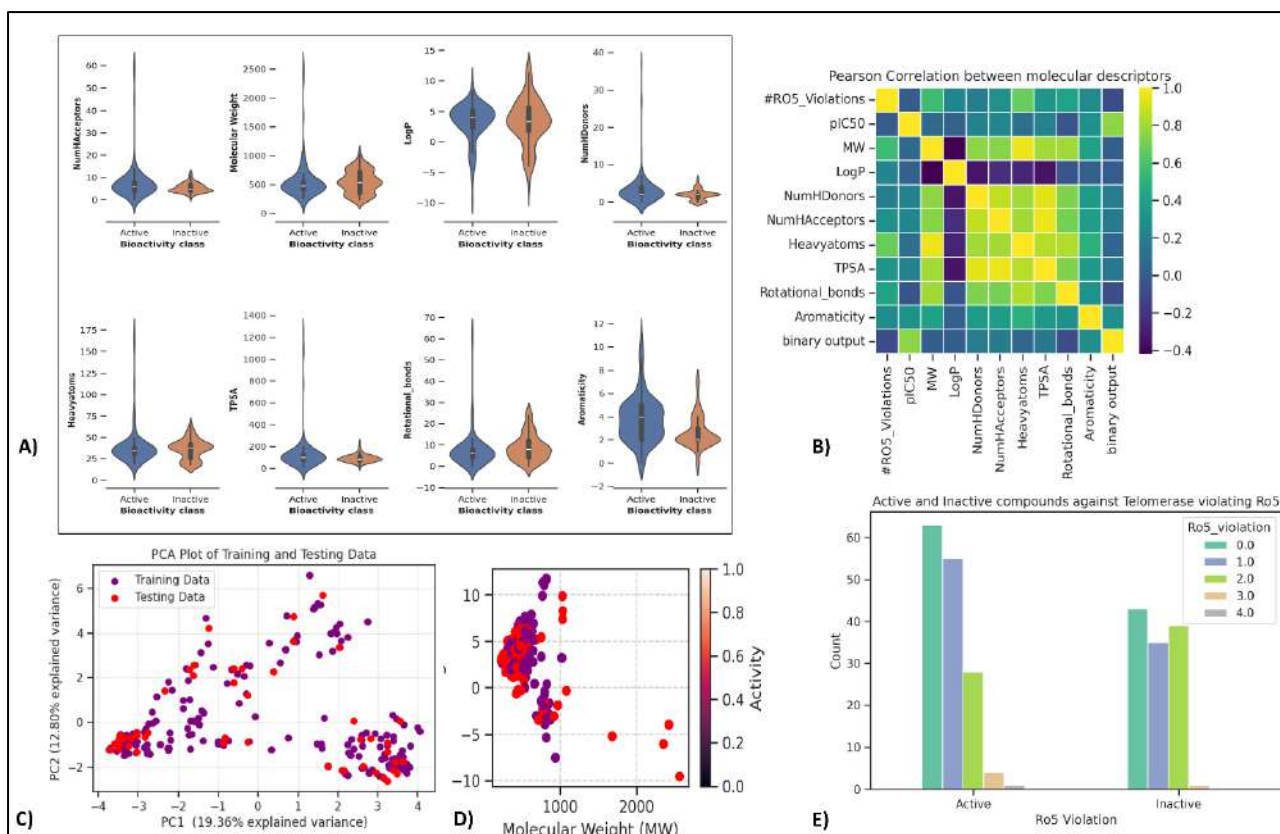


Figure 14. Exploratory analysis of active and inactive datasets for Telomerase **A)** Violin plots for eight different properties **B)** Heat plot indicating correlation between different properties **C)** PCA plot of training and test set distribution **D)** Molecular weight distribution across active and inactive dataset **E)** Ro5 Violation plot.

4.3 Descriptor calculation and feature selection

In the development of our machine learning model for predicting telomerase inhibitors, we first utilized PaDEL-Descriptor software to calculate molecular fingerprint descriptors for each compound in our dataset. This process generated a comprehensive set of descriptors, encompassing twelve fingerprint types across nine distinct classes. To ensure the robustness and reliability of our model, we conducted an initial filtering step, where descriptors with minimal variability—specifically, those with a standard deviation below 0.1—were identified and removed from the analysis. This threshold was chosen to eliminate constant or nearly constant variables, which could potentially introduce noise into the model.

Subsequently, we employed Recursive Feature Elimination (RFE), a feature selection technique that iteratively removes the least significant descriptors, to further refine our descriptor set. By systematically reducing the descriptor space, RFE allowed us to isolate a subset of highly informative and varied molecular fingerprints. This refined descriptor set served as the foundation for building a robust and predictive machine learning model, ultimately enhancing the accuracy of our telomerase inhibitor predictions.

4.4 ML model training and evaluation

In this study, we developed various machine learning models for binary classification by combining six algorithms—random forest (RF), support vector machines for classification (SVC), extreme gradient boosting (XGBoost), LightGBM, Classification, And Regression Trees (CART)- with twelve types of fingerprints i.e Klekota-Roth, Klekota-Roth count, CDK, CDK extended, CDK graph only, PubChem, 2D atom pairs, 2D atom pairs count, Substructure, Substructure count, MACCS, E-state (Table 4.3) available in the PaDEL module. The general workflow of this study is illustrated in **Figure 3**. The dataset was used to train different classification models, with hyperparameter tuning performed via Bayesian optimization and five-fold cross-validation. Detailed information on the hyperparameters and the corresponding search space can be found in **Table 5**.

Table 5. List of hyper parameters, their values and objectives tuned for optimization across all the six algorithms

Model	ML Algorithm	Tuned Hyperparameters
Classification	Random Forest (RF)	<pre> pipeline = ('variance threshold', VarianceThreshold(threshold=0.1)), ('smote', SMOTE(random_state=123)), #('pca', RFE(estimator=RandomForestClassifier(n_estimators=200, random_state=123), n_features_to_select=200)), ('clf',RandomizedSearchCV(estimator=RandomForestClassifier(), n_jobs=-1, random_state=123, scoring='accuracy',cv=5, param_distributions={ 'n_estimators': range(200, 600, 50), 'max_depth': range(2, 22, 1), 'min_samples_split': range(2, 22, 1), 'min_samples_leaf': range(2, 22, 1), 'max_features': ['sqrt', 'log2'] </pre>
	Support vector classifier (SVC)	<pre> pipeline = ('variance_threshold', VarianceThreshold(threshold=0.1)), ('smote', SMOTE(random_state=123)), ('rfe',RFE(estimator=RandomForestClassifier(n_estimators=20, random_state=123), n_features_to_select=20)), ('clf',RandomizedSearchCV(estimator=SVC(random_state=123), param_distributions={'C': np.linspace(0.01, 1, 10), 'kernel': ['linear', 'rbf'], 'gamma': [0.1, 20, 10] },scoring='accuracy',cv=5, n_jobs=-1)) </pre>
	Extreme Gradient Boost XGBoost	<pre> pipeline = ('variance_threshold', VarianceThreshold(threshold=0.1)), ('smote', SMOTE(random_state=123)), ('rfe',RFE(estimator=RandomForestClassifier(n_estimators=40, random_state=123), n_features_to_select=30)), ('clf',RandomizedSearchCV(estimator=xgb(objective='binary:logistic', eval_metric='auc', seed=123), n_jobs=- 1,random_state=123,scoring='accuracy',cv=5, </pre>

Classification		<pre> param_distributions = 'n_estimators': range(100, 500, 50), 'learning_rate': [0.05, 0.1, 0.2], 'max_depth': range(2, 6, 1), 'min_child_weight': range(3, 10, 1), 'gamma': [0.1, 0.5, 1], 'subsample': [0.5, 0.7, 1], 'colsample_bytree': [0.5, 0.7, 1], 'reg_alpha': [0.001, 0.01, 0.1, 1], 'reg_lambda': [1, 10, 100] </pre>
	Adaptive Boosting ADABOOST	<pre> pipeline = ('variance_threshold', VarianceThreshold(threshold=0.1)), ('smote', SMOTE(random_state=123)), ('rfe', RFE(estimator=RandomForestClassifier(n_estimators=50, random_state=123), n_features_to_select=50)), ('clf', RandomizedSearchCV(estimator=AdaBoostClassifier(random_state=123), n_jobs=-1, random_state=123, scoring='accuracy', cv=5, param_distributions= 'n_estimators': range(100, 600, 50), 'learning_rate': [0.1, 0.5, 1] </pre>
	Light Gradient Boosting Machine (LightGBM)	<pre> pipeline = ('variance_threshold', VarianceThreshold(threshold=0.1)), ('smote', SMOTE(random_state=123)), ('rfe', RFE(estimator=RandomForestClassifier(n_estimators=50, random_state=123), n_features_to_select=50)), ('clf', RandomizedSearchCV(estimator=lgb(random_state=123), n_jobs=-1, random_state=123, scoring='accuracy', cv=5, param_distributions = 'boosting_type': ['gbdt'], 'objective': ['binary'], 'metric': ['binary_error', 'binary_logloss'], 'num_leaves': [25, 50, 75], 'learning_rate': [0.01, 0.05, 0.1], 'feature_fraction': [0.8, 0.9, 1.0], 'bagging_fraction': [0.7, 0.8, 0.9], 'bagging_freq': [3, 5, 7], 'verbose': [0] Pipeline = ('variance_threshold', VarianceThreshold(threshold=0.1)), #try different values ('smote', SMOTE(random_state=123)), ('rfe', RFE(estimator=DecisionTreeClassifier(random_state=123), n_features_to_select=100)), ('clf', RandomizedSearchCV(estimator=DecisionTreeClassifier(random_state=123), n_jobs=-1, random_state=123, scoring='accuracy', cv=5, param_distributions= 'max_depth': range(1, 20), 'min_samples_split': range(2, 20), 'min_samples_leaf': range(1, 20) </pre>
Classification	Classification and regression tree (CART)	

A total of 72 ML models were created using 12 different types of features and 6 different algorithms. **Table 5** present the evaluation metrics for the RF, SVM, and XGBoost models, AdaBoost, LightGBM and CART utilizing 12 different types of fingerprints. These results are based on runs conducted on the training set (220 compounds), test set (55 compounds), and 5-fold cross-validation. The tables offer a comprehensive overview of each model's performance across various evaluation metrics. The best performing algorithms, SVC for Telomerase with f evaluation are highlighted in green. **Figure 15** provides a comprehensive evaluation of the machine learning model for Telomerase.

Table 5. Results of training, test set evaluation metrics along with cross validation accuracy across all six algorithms using which ML model has been developed

Classification Algorithm	CV_Acc	Train Acc	Train F1-Score	Train Precision	Train Recall	Train MCC	Train Spec	Test Acc	Test F1-Score	Test Precision	Test Recall	Test MCC	Test Spec
RF	0.7294	0.8364	0.8356	0.8361	0.8364	0.6640	0.7766	0.8545	0.8537	0.8551	0.8545	0.7035	0.7917
SVC	0.8451	0.8909	0.8910	0.8913	0.8909	0.7778	0.8830	0.8727	0.8724	0.8726	0.8727	0.7406	0.8333
XGB	0.7614	0.8409	0.8417	0.8466	0.8409	0.6832	0.8723	0.8545	0.8537	0.8551	0.8545	0.7035	0.7917
ADB	0.7496	0.7909	0.7920	0.8015	0.7909	0.5881	0.8404	0.7636	0.7644	0.7779	0.7636	0.5394	0.8333
LGB	0.7774	0.8455	0.8456	0.8460	0.8455	0.6852	0.8298	0.8000	0.8004	0.8013	0.8000	0.5957	0.5957
CART	0.7502	0.8364	0.8371	0.8413	0.8364	0.6729	0.8617	0.7636	0.7630	0.7629	0.7636	0.5176	0.7083

The performance of our machine learning models for Telomerase showcases its potential for effective virtual screening in drug discovery. The SVC model for Telomerase displayed strong performance with a test accuracy of 87.27%, an F1-Score of 87.26%, and an MCC of 0.7406, ensuring a reliable balance in prediction accuracy and specificity. A comparative AUC-ROC plot tested across various algorithms against the test dataset is shown in **Figure 15C**.

The SVC-KleotaRoth model on the Telomerase dataset, despite its smaller size, maintained high true positive (20) and true negative (28) rates with AUC values of 0.93 (training) and 0.89 (testing). The model was further validated using five-fold cross-validation, showing performance closely aligned with the training and test set accuracies, as detailed in **Table 5** highlighted in green. Collectively, these results highlight the model's consistent ability to accurately classify active and inactive compounds, demonstrating strong discriminative power and reliability across different targets.

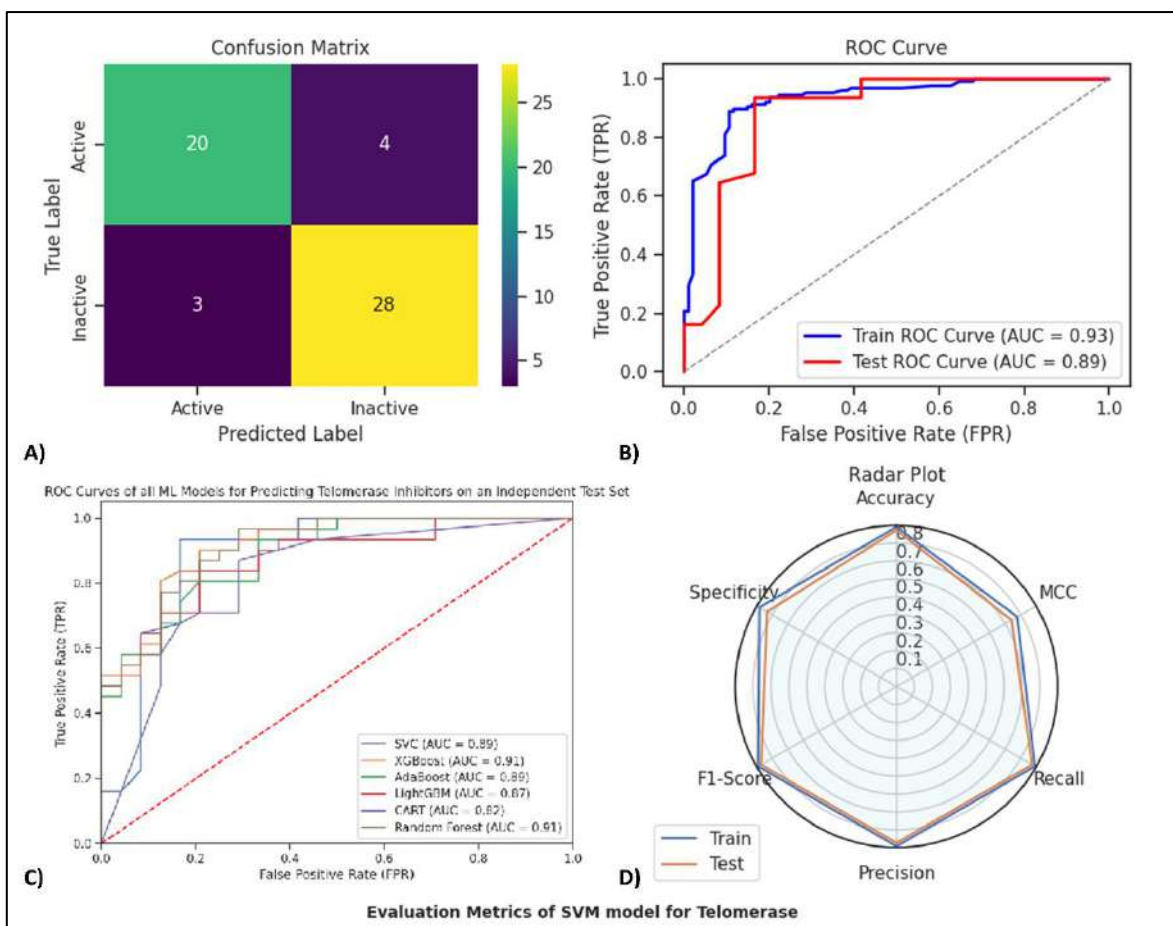


Figure 15. A) Evaluation metrics Telomerase using SVC-KlekotaRoth model. B) Confusion matrix. C) ROC curve for train and test set, D) Combined AUC-ROC for test set performance across all algorithms (Bottom Left) and Radar Plot.

4.5 Interpretability of the model based on SHAP values

The SHAP (SHapley Additive exPlanations) framework, as implemented by Lundberg and Lee, was utilized to identify the most impactful features for each target-specific model. Positive and negative SHAP values indicate features that significantly contribute to the classification of active and inactive compounds, respectively. This analysis helps us understand the key substructures driving the predictions **Figure 16A**.

The top features for Telomerase were KRFP3143, KRFP504, and KRFP4291. These results are based on the mean SHAP value and SHAP value plots, offer valuable insights into the potential substructures influencing model predictions (**Figure 16B**).

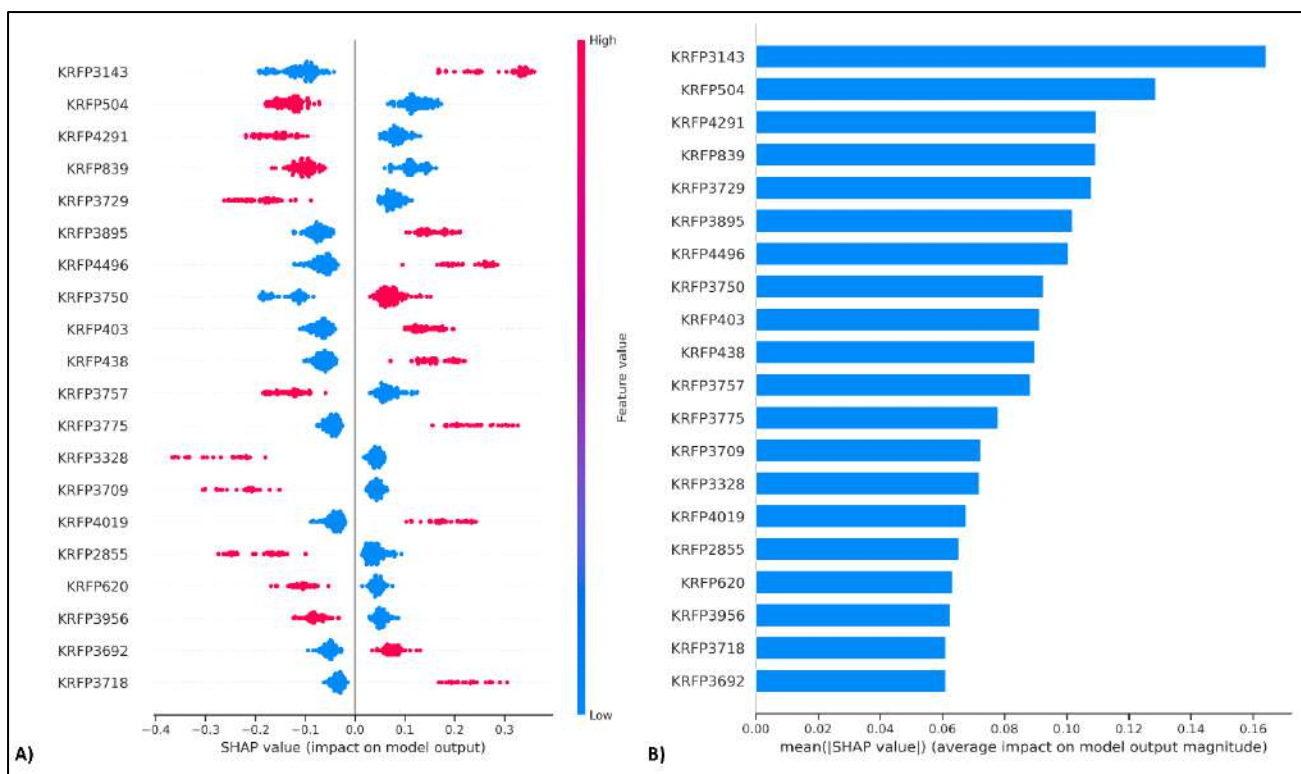


Figure 16. A) SHAP summary plot for the Telomerase's SVC Model **B)** Bar plot of top features from the best SVC model built using KlekotaRoth fingerprint

4.6 Probability Score Distribution for Active and Inactive Classes for the test dataset

The probability score distribution for active and inactive classes in the test datasets was analyzed using predictions from the best-performing model (**Figure 17A**). The boxplot illustrating the distribution for the test dataset for SVC for Telomerase. The results show that most active compounds have high probability scores, clustering near 1.0, with a significantly higher median probability score compared to inactive compounds across all targets. Inactive compounds typically have low probability scores, with a median well below the 0.5 threshold. The dashed red line at a probability score of 0.5 marks the decision threshold for distinguishing active from inactive compounds. Most active compounds score above this threshold, while most inactive compounds fall below it, demonstrating the model's effectiveness in accurately separating the two classes for each target.

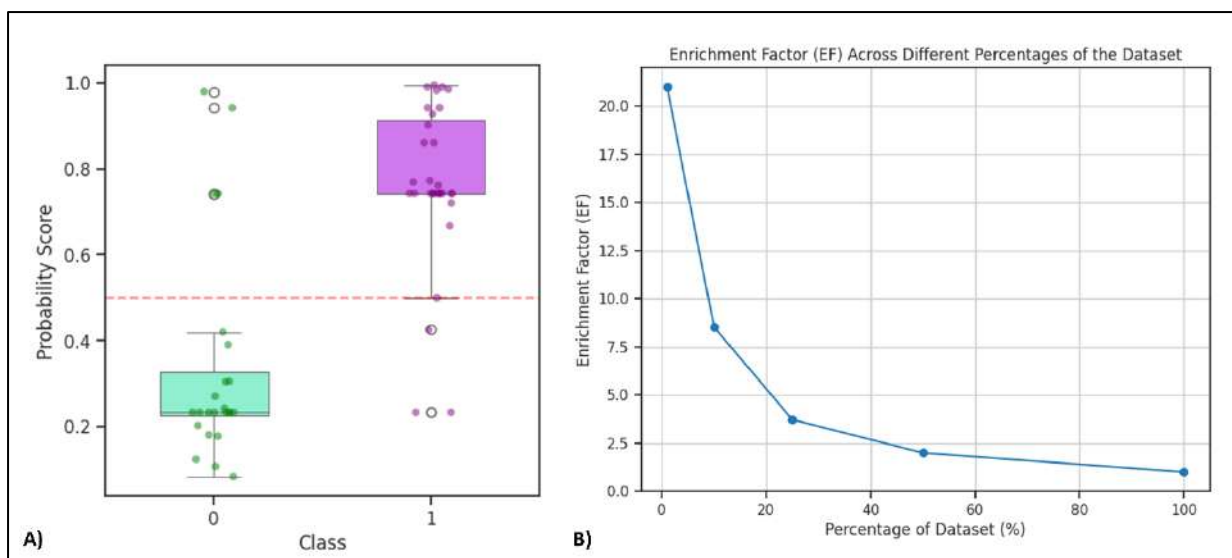


Figure 17. A) Boxplot for probability distribution of active and inactive test dataset.

B) Enrichment factor at various dataset percentages

4.7 Decoy Set Validation and *enrichment factor analysis*

The DUDE dataset [43], widely recognized for benchmarking virtual screening (VS) protocols, was utilized to assess the performance of our target-specific models. We selected the top 15 active test compounds against 300 decoys for telomerase, maintaining an active-to-decoy ratio of 1:20.

To rigorously evaluate the model's performance, we calculated the Enrichment Factor (EF) at various dataset percentages: 1%, 10%, 25%, 50%, and 100%. EF is a key metric in virtual screening (VS) assessments, quantifying the ratio between the hit rate of the top X% of ranked compounds and the hit rate of random selection. As illustrated in **Figure 17B**, our models achieved an EF of 21 at 1%, indicating that all true actives were identified within the top 1% of ranked test compounds, demonstrating a 21-fold enrichment compared to random selection. As the percentage of the dataset increases, the number of true positives identified decreases, and the EF value gradually approaches 1. This trend, shown in Figure reflects that while the models are highly effective at enriching the top ranks with genuine actives initially, this enrichment effect diminishes as the proportion of the dataset considered increases.

4.8 Identification of Common Lead Compounds via Structure-Based and Machine Learning Screening

The top-performing Klekota-Roth-SVC model for telomerase was employed to virtually screen the top 10 lead in Table 2 identified in Objective 1 through a structure-based drug design approach, which included pharmacophore-based screening and molecular dynamics validation studies using a probability cutoff set above 0.6 here. **Table 6.** summarizes the 6 out of 10 screened out Novel compounds from ChemDiv anti-cancer library, Otava drug-like lead collection and Binding database, as potential Telomerase inhibitors by Machine learning model.

Table 6: Screened out candidates as Telomerase inhibitors from ChemDiv, Otava and Binding database by ML model.

S.No	Names and Database	M scaffold	Probabilities	Predictions
1	Ligand-77574 (Otava drug like)	<chem>O=c1ccc2ccccc2o1</chem>	0.785435	Active
2	Compound637 (Binding database)	<chem>O=c1cc(-c2ccc3nc[nH]c3c2)oc2ccccc12</chem>	0.909248	Active
3	137187 (ChemDiv anticancer)	<chem>c1ccc2[nH]ccc2c1</chem>	0.542191	Active
4	Ligand-164867 (Otava drug like)	<chem>O=C(NCCc1nc2ccccc2[nH]1)[C@@H](Cc1ccccc1)Nc1ncccn1</chem>	0.828805	Active
5	Compound834 (Binding database)	<chem>c1cnc2ccc(Nc3ncnc(Nc4ccncc4)n3)cc2c1</chem>	0.954648	Active
6	Ligand-26295 (ChemDiv anticancer)	<chem>O=C(CCCn1c(=O)c2ccccc2n(CC(=O)Nc2ccccc2)c1=O)NCc1ccco1</chem>	0.716987	Active

It was observed that ML model screened 3 common hits successfully corroborating the findings from Objective 1 and Objective 2. Compounds 26295 and 138187 from ChemDiv anticancer database and Compound 77574 from Otava druglike lead collection were identified by both the structure-based approach and the machine learning model are proposed as strong candidates for further in-vitro and experimental studies to evaluate their potential as telomerase inhibitors (**Table 6**-highlighted in green).

The consistent identification of these compounds across both methodologies underscores their robust potential as telomerase inhibitors. Moreover, our machine learning model successfully identified compounds 834 and 637 from the binding database as active telomerase inhibitors with a high probability of 0.95 and 0.90 respectively, further validating the robustness of our model and confirming the accuracy of our findings. This convergence of results not only enhances the credibility of our findings but also demonstrates the effective synergy between computational and predictive modelling approaches in identifying viable drug candidates.

Discussion:

The discovery of telomerase, an enzyme essential for maintaining telomere length, has opened up new avenues in drug development due to its significant roles in aging, cancer, and regenerative medicine (Fragkiadaki et al., 2022). Telomerase is a pivotal factor in combating telomere shortening, a hallmark of nearly 90% of cancer malignancies, while its notable absence in somatic cells underscores its potential as a promising therapeutic target (Vishwakarma et al., 2023). Despite this high therapeutic potential, drug development targeting telomerase has been hindered by the structural ambiguity of the enzyme and the lack of approved small-molecule inhibitors. This study addresses these challenges through a synergistic approach combining structure-based drug discovery and machine learning models.

Utilizing the first reported structure of the hTERT thumb domain (PDB ID: 5UGW), our study focused on identifying novel lead compounds through a structure-based drug discovery approach. Inspired by the binding characteristics of BIBR1532, a highly selective telomerase inhibitor, we targeted the FVYL pocket in the thumb domain. This pocket, crucial for ribonucleoprotein (RNP) assembly and enzymatic activity, has been identified as a promising target for drug discovery (Bryan et al., 2015; Hoffman et al., 2017). Our research represents the first comparative analysis of the FVYL pocket between the human and truncated catalytic telomerase (tcTERT) thumb domains, revealing structural differences that impact drug binding.

Our molecular dynamics simulations uncovered a novel binding site (site II) distinct from the previously identified site I. This new site demonstrates significant structural divergence from site I, with a large backbone RMSD of 6.41 Å for the thumb domain and 2.5 Å for the FVYL pocket. The observed conformational changes suggest that BIBR1532 may shift from site I to site II, a finding supported by binding free energy calculations using MM-PBSA, which indicated a more favorable binding energy at site II. This novel site II, while maintaining critical FVYL interactions, is confirmed as the primary binding site for BIBR1532. However, the conclusive validation awaits a co-crystallized structure of BIBR1532 with hTERT.

To validate these findings, we generated pharmacophore models based on binding sites I and II. Screening through curated small molecule libraries from ChemDiv, Otava, and the Binding database provided a comprehensive assessment of potential inhibitors. Notably, molecules from Pharmacophore B1 docked effectively at site II, while those from Pharmacophore A1 failed to dock at site I. This discrepancy underscores the importance of the newly identified site II in our drug discovery efforts.

From our structure-based drug discovery approach employing pharmacophore screening and three levels of molecular docking studies, we identified 10 leads from the three databases, which were validated using molecular dynamic simulation and binding free energy calculations using MM-PBSA and MM-GBSA methods. Five of these were proposed as novel telomerase inhibitor candidates, alongside compound 637 (Lead 10, ChEMBL ID: ChEMBL376279) from the Binding database, which was established as binding to the thumb domain of telomerase.

Our machine learning (ML) model, trained on Klekota-Roth fingerprints and various algorithms, cross-validated the hits from pharmacophore-based screening. The ML model accurately identified three out of the five common validated hits, highlighting its robustness. Additionally, the ML model successfully screened compounds from the Binding database, reinforcing the reliability of our approach and the potential of compounds 834 and 637 as telomerase inhibitors. This convergence of results not only enhances the credibility of our findings but also demonstrates the effective synergy between computational and predictive modelling approaches in identifying viable drug candidates.

The consistent identification of compounds 26295 and 138187 from the ChemDiv anticancer database, and compound 77574 from the Otava drug-like lead collection, across both methodologies underscores their potential as strong candidates for further in vitro and experimental studies. This study's innovative approach in uncovering a novel binding site in hTERT represents a significant advancement in telomerase inhibitor research. By integrating molecular dynamics, pharmacophore modelling, and machine learning, we have enhanced the accuracy of inhibitor identification and provided a new perspective on telomerase drug targeting.

The novel compounds identified, particularly those from ChemDiv, Otava, and the Binding database, offer promising new candidates for telomerase inhibition. These compounds, validated through rigorous computational methods and in silico models, provide a crucial step forward in overcoming the historical challenges of telomerase drug development. The ML model developed in this study will soon be deployed on a publicly available web server for the entire research community to screen small molecule inhibitors as potential telomerase inhibitors.

Despite the advancements achieved, it is important to acknowledge the limitations of computational models in predicting biological outcomes. Variability in physicochemical properties, metabolism, and pharmacokinetics can affect the accuracy of in silico predictions. Therefore, comprehensive in vitro and in vivo validations are essential to confirm the efficacy, selectivity, and safety of these novel inhibitors.

In conclusion, this study demonstrates the power of a synergistic approach combining structure-based drug discovery with machine learning to identify novel telomerase inhibitors. The identification of a new binding site and the validation of lead compounds through rigorous computational methods mark a significant milestone in telomerase inhibitor research. These findings not only address a critical gap in cancer therapeutics but also provide a foundation for future drug development efforts. The proposed leads hold potential for integration with conventional therapies, offering a comprehensive approach to enhance treatment efficacy against cancer. As we advance, experimental validation and further refinement of these inhibitors will be crucial for translating these computational predictions into effective therapeutic interventions, ultimately benefiting the field of oncology and patient care.

Impact of the research in the advancement of knowledge or benefit to mankind

Telomerase, an enzyme essential for maintaining telomere length, plays a pivotal role in the unchecked growth of cancer cells, positioning it as a prime target for anti-cancer therapies. Despite its therapeutic potential, the development of telomerase-targeted drugs has been significantly hampered by the enzyme's structural complexity and the lack of approved small-molecule inhibitors. Our research addresses these challenges through a groundbreaking synergistic approach that combines structure-based drug discovery with advanced machine learning (ML) techniques, leading to the identification of novel lead compounds with the potential to revolutionize cancer treatment.

Innovative Discovery of Novel Leads: By utilizing the first reported structure of the human telomerase reverse transcriptase (hTERT) thumb domain, our study has identified and validated crucial new binding sites, particularly the FVYL pocket. Inspired by the known inhibitor BIBR1532, we conducted molecular dynamics simulations and binding free energy calculations, revealing a previously unidentified binding site (site II) with superior binding potential. This discovery represents a significant advancement in telomerase-targeted drug development, addressing a critical gap in current cancer therapy options.

Synergistic Approach with Machine Learning: The integration of machine learning into our drug discovery pipeline has markedly enhanced both the accuracy and efficiency of lead identification. Our ML model, trained on Klekota-Roth fingerprints, effectively cross-validated hits from pharmacophore-based screening, successfully identifying three out of five validated leads. This dual-validation approach highlights the robustness of our findings and demonstrates the transformative potential of combining computational methods to expedite drug discovery processes.

Global Scientific Contribution: A major impact of our study is the forthcoming public availability of our machine learning model for predicting telomerase inhibitors. By making this model accessible to the global research community, we are facilitating the screening of potential telomerase inhibitors by scientists worldwide, thereby accelerating collective efforts to develop effective cancer therapies. This contribution not only advances scientific knowledge but also promotes international collaboration, magnifying the impact of our work.

Immediate and Long-Term Benefits: The lead compounds identified through our rigorous computational methods are currently undergoing validation in our lab using various in-vitro and ex-vivo assays. Successfully developing these compounds could lead to new, targeted cancer therapies with greater efficacy and fewer side effects compared to traditional treatments. Additionally, our approach sets a new standard for integrating structure-based drug discovery with machine learning, offering a model that can be applied to other challenging drug targets in the future.

Future Directions: Moving forward, our research will focus on further validating these promising leads and refining our computational models. The recent advancements in the structural understanding of telomerase, including the full-length Cryo-EM structure, provide new opportunities to enhance the precision of our drug discovery efforts. The leads identified in this study hold promise not only for advancing cancer therapy but also for contributing to the development of treatments for other diseases involving telomerase dysregulation.

- I would also like to emphasize that this research work has also been **recognized and awarded** internationally by the **European Molecular Biology Laboratory (EMBO)** in 2022 and **American Association of Cancer Research (AACR)** and **European Association of Cancer Research (EACR)** in 2024 to be presented at their international conferences held in Portugal, San Diego, California and Rotterdam, Netherlands respectively.
- This research work in part **has already been accepted for publication in the Journal of Biomolecular Structure and Dynamics** in 2024. Another segment of the work has also been submitted for publication.

Literature References:

- Al-Lazikani, B., Gaulton, A., Paolini, G., Lanfear, J., Overington, J., & Hopkins, A. (2007). The Molecular Basis of Predicting Druggability. In *Bioinformatics-From Genomes to Therapies* (pp. 1315–1334). <https://doi.org/10.1002/9783527619368.ch36>
- Arooj, M., Sakkiyah, S., Kim, S., Arulalapperumal, V., & Lee, K. W. (2013). A combination of receptor-based pharmacophore modeling & QM techniques for identification of human chymase inhibitors. *PLoS One*, 8(4), e63030. <https://doi.org/10.1371/journal.pone.0063030>
- Babajide Mustapha, I., & Saeed, F. (2016). Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules*, 21(8), Article 8. <https://doi.org/10.3390/molecules21080983>
- Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L. J., De Veij, M., & Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12(1), 51. <https://doi.org/10.1186/s13321-020-00456-1>
- Bisong, E. (2019). Google Colaboratory. In E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7
- Bissantz, C., Kuhn, B., & Stahl, M. (2010). A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry*, 53(14), 5061–5084. <https://doi.org/10.1021/jm100112j>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315139470>
- Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., ... Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), 1545–1614. <https://doi.org/10.1002/jcc.21287>
- Bryan, C., Rice, C., Hoffman, H., Harkisheimer, M., Sweeny, M., & Skordalakes, E. (2015). Structural Basis of Telomerase Inhibition by the Highly Specific BIBR1532. *Structure (London, England : 1993)*, 23(10), 1934–1942. <https://doi.org/10.1016/j.str.2015.08.006>
- Capecchi, A., Probst, D., & Reymond, J.-L. (2020). One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1), 43. <https://doi.org/10.1186/s13321-020-00445-4>
- Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2), 64–73. <https://doi.org/10.1021/ci00046a002>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, D., Oezguen, N., Urvil, P., Ferguson, C., Dann, S. M., & Savidge, T. C. (2016). Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Science Advances*, 2(3), e1501240. <https://doi.org/10.1126/sciadv.1501240>
- Ding, X., Cheng, J., Pang, Q., Wei, X., Zhang, X., Wang, P., Yuan, Z., & Qian, D. (2019). BIBR1532, a Selective Telomerase Inhibitor, Enhances Radiosensitivity of Non-Small Cell Lung Cancer Through Increasing Telomere Dysfunction and ATM/CHK1 Inhibition. *International Journal of*

*Radiation Oncology*Biology*Physics*, 105(4), 861–874.

<https://doi.org/10.1016/j.ijrobp.2019.08.009>

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6), 1273–1280.

<https://doi.org/10.1021/ci010132r>

Eckburg, A., Dein, J., Berei, J., Schrank, Z., & Puri, N. (2020). Oligonucleotides and microRNAs Targeting Telomerase Subunits in Cancer Therapy. *Cancers*, 12(9), Article 9.

<https://doi.org/10.3390/cancers12092337>

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19), 8577–8593.

<https://doi.org/10.1063/1.470117>

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Boosting Algorithms: A Review of Methods, Theory, and Applications. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp. 35–85). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

Fragkiadaki, P., Renieri, E., Kalliantasi, K., Kouvidi, E., Apalaki, E., Vakonaki, E., Mamoulakis, C., Spandidos, D. A., & Tsatsakis, A. (2022). Telomerase inhibitors and activators in aging and cancer: A systematic review. *Molecular Medicine Reports*, 25(5), 158.

<https://doi.org/10.3892/mmr.2022.12674>

Gagnon, J. K., Law, S. M., & Brooks, C. L. (2016). Flexible CDOCKER: Development and application of a pseudo-explicit structure-based docking method within CHARMM. *Journal of Computational Chemistry*, 37(8), 753–762. <https://doi.org/10.1002/jcc.24259>

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue), D1100–D1107.

<https://doi.org/10.1093/nar/gkr777>

- Greider, C. W., & Blackburn, E. H. (1985). Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell*, 43(2), 405–413. [https://doi.org/10.1016/0092-8674\(85\)90170-9](https://doi.org/10.1016/0092-8674(85)90170-9)
- Guterres, A. N., & Villanueva, J. (2020). TARGETING TELOMERASE FOR CANCER THERAPY. *Oncogene*, 39(36), 5811–5824. <https://doi.org/10.1038/s41388-020-01405-w>
- Hall, L. H., & Kier, L. B. (1995). Electrotological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences*, 35(6), 1039–1045. <https://doi.org/10.1021/ci00028a014>
- Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1), 57–70. [https://doi.org/10.1016/s0092-8674\(00\)81683-9](https://doi.org/10.1016/s0092-8674(00)81683-9)
- Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., & Hutchison, G. R. (2012). Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1), 17. <https://doi.org/10.1186/1758-2946-4-17>
- Heikamp, K., & Bajorath, J. (2014). Support vector machines for drug discovery. *Expert Opinion on Drug Discovery*, 9(1), 93–104. <https://doi.org/10.1517/17460441.2014.866943>
- Helal, K. Y., Maciejewski, M., Gregori-Puigjané, E., Glick, M., & Wassermann, A. M. (2016). Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *Journal of Chemical Information and Modeling*, 56(2), 390–398. <https://doi.org/10.1021/acs.jcim.5b00498>
- Hoffman, H., Rice, C., & Skordalakes, E. (2017). Structural Analysis Reveals the Deleterious Effects of Telomerase Mutations in Bone Marrow Failure Syndromes. *The Journal of Biological Chemistry*, 292(11), 4593–4601. <https://doi.org/10.1074/jbc.M116.771204>
- Jäger, K., & Walter, M. (2016). Therapeutic Targeting of Telomerase. *Genes*, 7(7), 39. <https://doi.org/10.3390/genes7070039>
- Jha, P., Saluja, D., & Chopra, M. (2022). Structure-guided pharmacophore based virtual screening, docking, and molecular dynamics to discover repurposed drugs as novel inhibitors against

- endoribonuclease Nsp15 of SARS-CoV-2. *Journal of Biomolecular Structure & Dynamics*, 1–11. <https://doi.org/10.1080/07391102.2022.2079561>
- Jha, P., Singh, P., Arora, S., Sultan, A., Nayek, A., Ponnusamy, K., Syed, M. A., Dohare, R., & Chopra, M. (2022). Integrative multiomics and in silico analysis revealed the role of ARHGEF1 and its screened antagonist in mild and severe COVID-19 patients. *Journal of Cellular Biochemistry*, 123(3), 673–690. <https://doi.org/10.1002/jcb.30213>
- Kant, R., Jha, P., Saluja, D., & Chopra, M. (2022). Identification of novel inhibitors of *Neisseria gonorrhoeae* Muri using homology modeling, structure-based pharmacophore, molecular docking, and molecular dynamics simulation-based approach. *Journal of Biomolecular Structure & Dynamics*, 1–14. <https://doi.org/10.1080/07391102.2022.2121943>
- Kenny, P. W. (2022). Hydrogen-Bond Donors in Drug Design. *Journal of Medicinal Chemistry*, 65(21), 14261–14275. <https://doi.org/10.1021/acs.jmedchem.2c01147>
- Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L., Coviello, G. M., Wright, W. E., Weinrich, S. L., & Shay, J. W. (1994). Specific association of human telomerase activity with immortal cells and cancer. *Science (New York, N.Y.)*, 266(5193), 2011–2015. <https://doi.org/10.1126/science.7605428>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2022). PubChem 2023 update. *Nucleic Acids Research*, 51(D1), D1373–D1380. <https://doi.org/10.1093/nar/gkac956>
- Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics (Oxford, England)*, 24(21), 2518–2525. <https://doi.org/10.1093/bioinformatics/btn479>
- Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Comput. Surv.*, 50(6), 94:1–94:45. <https://doi.org/10.1145/3136625>

- Liu, B., He, Y., Wang, Y., Song, H., Zhou, Z. H., & Feigon, J. (2022). Structure of active human telomerase with telomere shelterin protein TPP1. *Nature*, 604(7906), 578–583.
<https://doi.org/10.1038/s41586-022-04582-8>
- Liu, C., Zhou, H., Sheng, X. B., Liu, X. H., & Chen, F. H. (2020). Design, synthesis and SARs of novel telomerase inhibitors based on BIBR1532. *Bioorganic Chemistry*, 102, 104077.
<https://doi.org/10.1016/j.bioorg.2020.104077>
- Liu, X., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., Zheng, S., Li, Z., Li, H., & Jiang, H. (2010). PharmMapper server: A web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Research*, 38(suppl_2), W609–W614.
<https://doi.org/10.1093/nar/gkq300>
- Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language—Jiangang Hao, Tin Kam Ho, 2019. (n.d.). Retrieved August 29, 2024, from*
<https://journals.sagepub.com/doi/abs/10.3102/1076998619832248>
- Meibohm, B., & Derendorf, H. (1997). Basic concepts of pharmacokinetic/pharmacodynamic (PK/PD) modelling. *International Journal of Clinical Pharmacology and Therapeutics*, 35(10), 401–413.
- Meyer, E. A., Castellano, R. K., & Diederich, F. (2003). Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angewandte Chemie International Edition*, 42(11), 1210–1250.
<https://doi.org/10.1002/anie.200390319>
- Miller, B. R. I., McGee, T. D. Jr., Swails, J. M., Homeyer, N., Gohlke, H., & Roitberg, A. E. (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, 8(9), 3314–3321. <https://doi.org/10.1021/ct300418h>
- Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O., & Miteva, M. A. (2012). Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discovery Today*, 17(1), 44–55.
<https://doi.org/10.1016/j.drudis.2011.10.023>

- Mysinger, M. M., Carchia, M., Irwin, John. J., & Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14), 6582–6594. <https://doi.org/10.1021/jm300687e>
- Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In R. Silhavy & P. Silhavy (Eds.), *Artificial Intelligence Application in Networks and Systems* (pp. 15–25). Springer International Publishing. https://doi.org/10.1007/978-3-031-35314-7_2
- Nguyen, T. H. D., Tam, J., Wu, R. A., Greber, B. J., Toso, D., Nogales, E., & Collins, K. (2018). Cryo-EM structure of substrate-bound human telomerase holoenzyme. *Nature*, 557(7704), 190–195. <https://doi.org/10.1038/s41586-018-0062-x>
- Pucci, C., Martinelli, C., & Ciofani, G. (2019). Innovative approaches for cancer treatment: Current perspectives and new challenges. *Ecancermedicalscience*, 13, 961. <https://doi.org/10.3332/ecancer.2019.961>
- Rao, S. N., Head, M. S., Kulkarni, A., & LaLonde, J. M. (2007). Validation Studies of the Site-Directed Docking Program LibDock. *Journal of Chemical Information and Modeling*, 47(6), 2159–2171. <https://doi.org/10.1021/ci6004299>
- Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3), 470–489. <https://doi.org/10.1006/jmbi.1996.0477>
- Robart, A. R., & Collins, K. (2011). Human telomerase domain interactions capture DNA for TEN domain-dependent processive elongation. *Molecular Cell*, 42(3), 308–318. <https://doi.org/10.1016/j.molcel.2011.03.012>
- Shaker, B., Ahmad, S., Lee, J., Jung, C., & Na, D. (2021). In silico methods and tools for drug discovery. *Computers in Biology and Medicine*, 137, 104851. <https://doi.org/10.1016/j.combiomed.2021.104851>

- Shay, J. W. (1997). Telomerase in human development and cancer. *Journal of Cellular Physiology*, 173(2), 266–270. [https://doi.org/10.1002/\(SICI\)1097-4652\(199711\)173:2<266::AID-JCP33>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1097-4652(199711)173:2<266::AID-JCP33>3.0.CO;2-B)
- Shay, J. W., & Wright, W. E. (2000). Hayflick, his limit, and cellular ageing. *Nature Reviews. Molecular Cell Biology*, 1(1), 72–76. <https://doi.org/10.1038/35036093>
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493–500. <https://doi.org/10.1021/ci025584y>
- Thompson, D. C., Humblet, C., & Joseph-McCarthy, D. (2008). Investigation of MM-PBSA rescoring of docking poses. *Journal of Chemical Information and Modeling*, 48(5), 1081–1091. <https://doi.org/10.1021/ci700470c>
- Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A., & Moreno, E. (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation*, 17(10), 6281–6291. <https://doi.org/10.1021/acs.jctc.1c00645>
- Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & MacKerell, A. D. (2010). CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4), 671–690. <https://doi.org/10.1002/jcc.21367>
- Vařeková, R. S., Koča, J., & Zhang, C.-G. (2004). Complexity and Convergence of Electrostatic and van der Waals Energies within PME and Cutoff Methods. *International Journal of Molecular Sciences*, 5(4), Article 4. <https://doi.org/10.3390/i5040154>
- Vishwakarma, K., & Bhatt, H. (2021). Molecular modelling of quinoline derivatives as telomerase inhibitors through 3D-QSAR, molecular dynamics simulation, and molecular docking

techniques. *Journal of Molecular Modeling*, 27(2), 30. <https://doi.org/10.1007/s00894-020-04648-2>

Vishwakarma, K., Dey, R., & Bhatt, H. (2023). Telomerase: A prominent oncological target for development of chemotherapeutic agents. *European Journal of Medicinal Chemistry*, 249, 115121. <https://doi.org/10.1016/j.ejmech.2023.115121>

Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., & Hou, T. (2019). End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16), 9478–9508. <https://doi.org/10.1021/acs.chemrev.9b00055>

Welfer, G. A., & Freudenthal, B. D. (2023). Recent advancements in the structural biology of human telomerase and their implications for improved design of cancer therapeutics. *NAR Cancer*, 5(1), zcad010. <https://doi.org/10.1093/narcan/zcad010>

Wright, W. E., & Shay, J. W. (1992). The two-stage mechanism controlling cellular senescence and immortalization. *Experimental Gerontology*, 27(4), 383–389. [https://doi.org/10.1016/0531-5565\(92\)90069-c](https://doi.org/10.1016/0531-5565(92)90069-c)

Wu, G., Robertson, D. H., Brooks III, C. L., & Vieth, M. (2003). Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMM-based MD docking algorithm. *Journal of Computational Chemistry*, 24(13), 1549–1562. <https://doi.org/10.1002/jcc.10306>

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474. <https://doi.org/10.1002/jcc.21707>

Zuccotto, F. (2003). Pharmacophore Features Distributions in Different Classes of Compounds. *Journal of Chemical Information and Computer Sciences*, 43(5), 1542–1552. <https://doi.org/10.1021/ci034068k>

