Indian Journal of Biochemistry & Biophysics
Vol. 60, November 2023, pp. 844-856
DOI: 10.56042/ijbb.v60i11.6067

Generating a potent inhibitor against MCF7 breast cancer cell through artificial intelligence based virtual screening and molecular docking studies

Latha V[1]*, Gomathi V[2], Rajeshkanna A[3] & Hari Ram S[4]

[1]Department of Chemistry; & [3]Dartmenent of Computer Science, Sri S. Ramasamy Naidu Memorial College, Sattur-626 203, Tamil Nadu, India

[2]Department of CSE; & [4]Department of CSE, National Engineering College, Kovilpatti-628 503, Tamil Nadu, India

Received 06 October 2023; *revised 04 November 2023*

Artificial Intelligence (AI) has been widely adopted by pharmaceutical industry to aid rationally drug design and development by fostering the quick delivery of drug targets with optimized structures in spite of huge chemical space of $>10^{60}$ drug molecules. Tamoxifen, Selective Estrogen Receptor Modulator (SERM), is the drug for breast cancer cell, MCF 7 with many side effects. Tamoxifen may cause side effects like increased bone or tumor pain, pain or reddening around the tumor site, hot flashes, nausea and excessive tiredness *etc*., Therefore, compound which can resist ER's bioactivity is considered as an important target for treating breast cancer. In this study, AI based Virtual Screening (VS) method using an efficient Generative Neural Network (GNN) model has been experimented to generate high inhibitory potential hit drug-like inhibitors. Physicochemical, Pharmacokinetic and toxicity analysis are carried out for conforming the sub-selection of drug-likeness of inhibitors. Additionally, Molecular Docking studies with DNA (355D) and protein (3EU7) are performed for the evaluation of binding affinity, prediction of intermolecular interactions and inhibition constant. The docked results of the inhibitor M22 (methyl 2-[(2-benzoylphenyl) carbamoyl] benzoate) has low free energy of binding (-8.61 Kcal/mol and -8.05 Kcal/mol) and low Inhibition constant, Ki, value (0.486 μM and 1.25 μM) as compared to Tamoxifen (-6.7 Kcal/mol & -5.62 Kcal/mol and 12.2 μM & 75.85 μM). Thus, minimum amount of the M22 inhibitor is enough as compared to Tamoxifen and M22 has 3 benzene rings, extended conjugation, amide linkage and huge number of labile electrons which facilitates as a potent drug. This study provides a greenish path to synthesise a potent inhibitor, M22, for further experimental studies rather than preparing number of inhibitors on the atom economy way.

Keywords: ADMET analysis, Selective Estrogen Receptor, Toxicity analysis

World Health Organization medical officer for cancer control, Dr. Ben Anderson said that "With breast cancer now the most common cancer globally and the most likely reason a woman will die from cancer, countries need to embrace the concept of improving breast cancer outcomes if they are going to address cancer as a health priority". As per the data given by National Cancer Institute, in women about 67% to 80% breast cancers are Estrogen Receptor (ER) positive. In India breast cancer is ranked as number one cancer disease in women. To block the estrogen receptors, Tamoxifen (Nolvadex), is used as the drug[1].

The basic principles of green chemistry like preventing waste, maximizing atom economy, designing less hazardous chemical and saving time drive us to use compound activity prediction based computational models to screen potential active compounds. Recently, enormous open access virtual chemical spaces such as, ChEMBL[2], PubChem[3]. DrugBank[4], Purple Book[5] and ZINC[6] are released to support the search based on the molecular properties and similar bioactivity related compounds.

Virtual screening (VS) helps to computationally identify the bioactive compounds with structure-based or ligand-based drug design approaches[7,8]. VS provides faster elimination of non-lead compounds based on 1D/2D/3D molecular descriptors, SMILES strings, fingerprints or graphical molecular structures, and also the selections of lead compounds are highlighted with physical, chemical and toxicological properties in terms of ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiles[9,10].

Physicochemical properties are calculated by involving the AI-based drug discovery program (DDP)[11] to measure the set of distinct factors to identify the drug-likeness of a chemical entity. Recently, various AI-based strategies for ADMET

*Correspondence:
E-mail: latha@srnmcollege.ac.in
Suppl. Data available on respective page of NOPR

and drug-likeness are being exploited by the researchers[12] as a part of DDP to avoid unnecessary wastage of time, budget, and manpower.

Binding affinity is one of the key factors in AI based VS during drug discovery. The quantitative structure activity relationship (QSAR) is a conventional method[13] to predict the protein-ligand binding affinity, by using various machine learning techniques such as support vector machine (SVM), random forest (RF), Bayesian algorithm[14], artificial neural networks (ANNs) and so on.

Deep learning models are more powerful to learn the hidden characteristics of drug discovery. Convolutional Neural Network (CNN)[15,16] models are adopted for binding affinity predictions, and have also been applied for virtual screening. The variational auto encoders (VAEs) and generative adversarial network (GAN) models are widely used for small molecular generation. The adversarial auto encoder (AAE) helps to train the molecular SMILES representations to generate the drug-like ligands. The objective-reinforced generative adversarial network[17] (ORGAN) merges the adversarial and reinforcement learning methods to generate the desired molecules by using Wasserstein-1 distance to improve the learning stability. OnionNet[18], a modified CNN model for predicting the rotation-free element pair-specific contacts between ligands and protein had been extended further as OnionNet-2 with multiple distance shells, to describe the ligand interactions.

Recently, using Graph Neural Network (GNN)[19] generative models, the molecule is represented as an undirected graph, where each node is a specific atom type and the edge is the bond type. MolGAN, a graph generative GAN model is capable of producing meaningful drug-like molecules via the annotation matrix and adjacency tensor[20]. Using the adjacency tensor and annotation matrix formats, the similar chemical compounds can be categorically grouped. These are generally grouped as single-shot graphs and iterative graphs. Graph learning methods such as MPN[21], Scaffold[22], GraphNVENT[23] have the potential to improve drug discovery efficiency dramatically for their ability to amplify insights available from existing drug-related datasets.

To solve issues of 3D ligand generation, MolAICal[24], a soft tool is used for drug design by employing the sequence based generative model (FDAMol) and GNN generative model (ZINCMol) for producing the ligand set and small molecular fragments. It is capable of generating optimized structures of ligands in the active pocket of receptors.

AI based VS creates smaller ligands, which have anticancer properties as equivalent to the potential compounds in medical practices presently for cancer patients. Throughout the worldwide, Tamoxifen [1-(p-dimethylaminoethoxyphenyl)-1,2-diphenyl-1-butene], a selective estrogen receptor modulator (SERM), has been widely used for the treatment and prevention of recurrence for patients with hormone receptor positive breast cancers[25].

Here, we propose the efficacy of AI based VS technique in predicting efficient inhibitor for ER and molecular docking of the hit compounds with DNA and protein of MCF7 cell. These findings will be useful in synthesizing of novel SERMs in the future.

Materials and Methods

AI based virtual screening for Drug Discovery

Novel synthesizable and drug-like ligands can be generated by MolAICal package, where, ZINCMol model is designed over molecular docking with generative deep learning model trained on ZINC database, which is built using Wasserstein generative adversarial networks (WGANs)[26] and reinforcement learning (RL). Here, the fragment growths of new compounds are carried out with Fibonacci random perturbation search algorithm[27].

The probability of identifying the difference between newly generated compounds f(x') and real data f(x) from distributions is required to be indistinguishable. It is measured in terms of Earth Mover (EM) Distance or Wasserstein-1[28] as,

$$W(P_r, P_\theta) = \sup_{\|f\|_{L\leq 1}} E_{x\sim P_r}[f(x)] - E_{x'\sim P_\theta}[f(x')] \quad \ldots(1)$$

where, the supremum is over all the L-Lipschitz functions, for some L. Also, the loss function at this point is an estimate of the EM distance. During training phase of the ZINCMol model, one could considerably solve the above equation by optimizing the value function of WGAN as,

$$\min_G \max_D E_{x\sim P_r}[D(x)] - E_{z\sim P_{(z)}}[D(G(z))] \quad \ldots(2)$$

where, G is the generative block of ZINCMol that tries to generate the chemical compounds and minimizes the distinguishable probability of fake data. Here, D is the discriminative block of ZINCMol that tries to discriminate the chemical compounds and maximizes the distinguishable probability of real data.

Finally, representative ligands (with formation of rigid fragments and rotational bonds) are chosen based on the Vinardo score[29] and evaluated by Equations (3) to further estimate the affinity between ligands and receptors.

$$S = \sum_i w_g . G(l_i) + w_r . R(l_i) + w_{hy} . Hy(l_i) + w_h . H(l_i) \quad \ldots(3)$$

where, S is the binding score between the ligand and protein. l_i is the distance between two atoms. w_g, w_r, w_{hy} and w_h are the weight parameters. The steric interaction is evaluated by Equations (4) and (5). The hydrophobic and H-bond interactions are assessed using Equations (6) and (7):

$$G(l) = e^{-((l-o)/s)^2} \quad \ldots(4)$$

$$R(l) = \begin{cases} l^2, \text{if } l < \dot{0}\ A \\ 0, \text{if } l \geq \dot{0}\ A \end{cases} \quad \ldots(5)$$

$$Hy(l) = \begin{cases} 1, \text{if } l < p_1 \\ (p_2 - l), \text{if } p_1 \leq l \leq p_2 \\ 0, \text{if } l > p_2 \end{cases} \quad \ldots(6)$$

$$H(l) = \begin{cases} 1, \text{if } l \leq h_1 \\ (p_2 - l), \text{if } h_1 < l < 0 \\ 0, \text{if } l \geq \dot{0}\ A \end{cases} \quad \ldots(7)$$

where, o, s, p_1, p_2 and h_1 are the tuning parameters. The combination of an attractive Gaussian function (Eq. 4) with a repulsive parabolic function (Eq. 5) reproduces the general shape of a typical Lennard-Jones interaction. Non-steric interactions are refined using (Eq. 6), if the two atoms are having hydrophobic bond, and/or by using (Eq.7), if the two atoms are having hydrogen bond.

As depicted in (Fig. 1), initially, using ZINCMol model of MolAICal, ligand samples can be generated which are structurally similar to Tamoxifen drug (Reference Compound). The synthetically generated compounds are filtered out based on their similarity with Tamoxifen drug using their SMILES string by computing the dice-coefficient based similarity score as given below.

$$\text{Dice\_coefficient}(A, B) = \left(c/(1/2(a + b))\right) \quad \ldots(8)$$

where, a counts the number of features present in A; b are of those present in B and c refers to the features shared by both A and B. Among these generated samples, some of the top similar molecules are selected as test compounds for further analysis of their suitability towards drug discovery.

Physicochemical, Pharmacokinetic and Toxicity parameters

Swiss ADME tool has been adopted to predict the ADMET characteristics for the test compounds. Based on Lipinski filter,[30-34] "the rule of 5", the physicochemical properties of the generated ligands

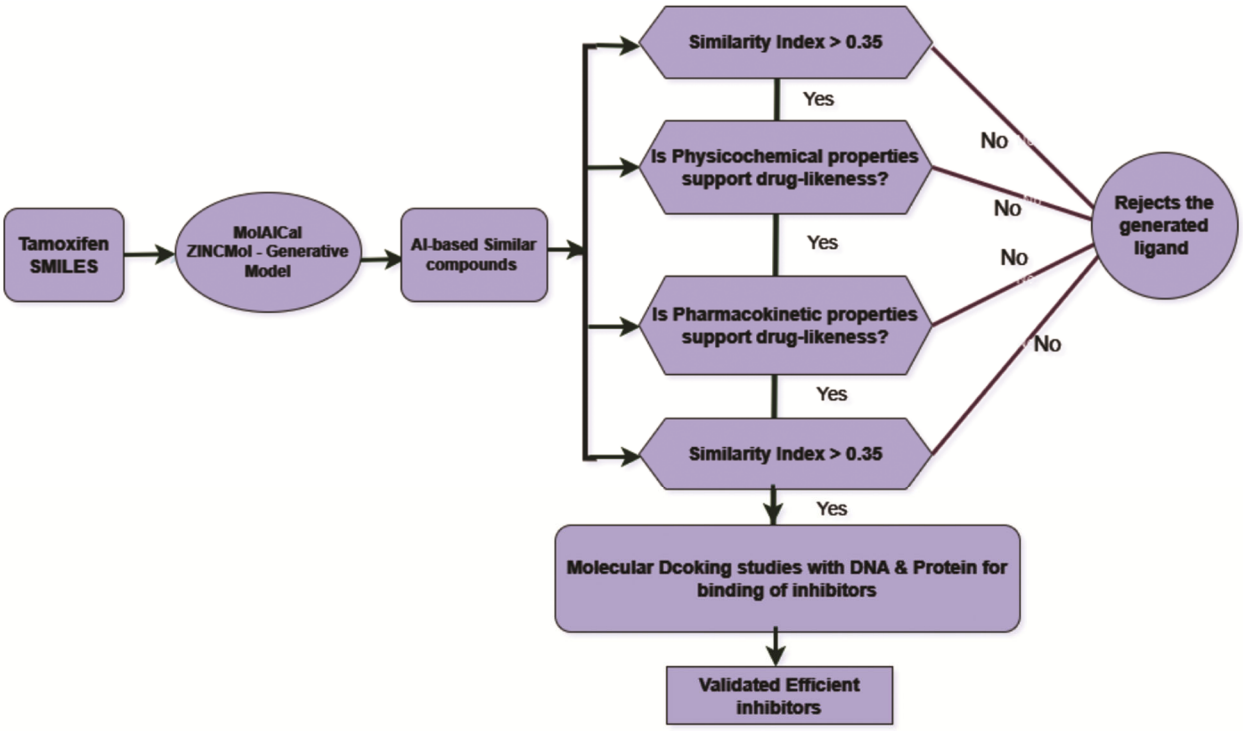

Fig. 1 — Schematic flow of proposed work

help to sub-select only the efficient drug-like compounds.

Pharmacokinetic parameters are important determinants of developmental toxicity[35] measured as absorption, distribution, metabolism, and excretion and/or toxicity (ADMET). For any drug-like compound, during oral administration an adequate absorption from the gut wall into the blood circulatory system is needed.

Then, it will be transported to the liver, undergoes a modification by a panel of hepatic microsomal enzymes; where few molecules may be metabolized and few may be excreted via the bile. If a drug-like compound survives during metabolism, it will enter arterial circulation and is subsequently distributed to the target tissue. Once the drug has triggered the desirable therapeutic response, it should be steadily eliminated from the body; otherwise bioaccumulation may also become yet another concern.

The pharmacokinetic parameters considered are, GI absorption, P-gp(P-Glycoprotein) substrate, Caco-2 Permeability, BBB (Blood-Brain-Barrier) permeant, Lipinski filter, Eagen filte, Muegge (Bayer) filter, Veber filter and Ghose filter. In addition, a drug-like compound must not cause any toxic side effects[36]. Toxicity prediction can be estimated by PreADMET online database[37]. This is helpful to have a broad-spectrum of health benefits, including anti-oxidative, anti-inflammatory, anti-mutagenic, and anti-carcinogenic properties.

Molecular Docking

The existing drug discovery approaches normally start with the identification of a target[38], which plays an important role in the protein interaction network of breast cancer disease. Thus an ideal target is essential for designing inhibitory drugs for MCF7 breast cancer cell line[39-41].

Molecular Docking studies with DNA

Deoxyribonucleic acid (DNA) is an important genetic substance in the organism and is the primary intracellular target of anticancer drugs. It not only carries and expresses the hereditary information but also decides the type and function of cells. It plays a decisive role in growth, breeding, heredity, variation and transformation. Calf thymus DNA (CT-DNA) is widely used in studies of DNA binding anticancer agents and DNA binding agents that modulate DNA structure and function. The investigation of drug–DNA interaction is important for understanding the molecular mechanism of drug action and for the design of specific DNA targeted drug.

Molecular dockingis performed for AI based generated drug-like ligands with the duplex sequence DNA d(CGCGAATTCGCG)2 dodecamer (PDBID: 355D)

Drugs can interact with DNA through the following three non-covalent modes: (i) Electrostatic binding (ii) Intercalative binding and (iii) Groove binding. The intercalative binding is stronger than other two binding modes because the surface of intercalative molecule is sandwiched between the aromatic and heterocyclic base pairs of DNA.

Molecular Docking studies with Protein

Protein - ligand docking is carried out using Autodock (v4.2.6) software by following the user guide's instructions provided by the Scripps Research Institute. In the docking experimental setup, AI generated compounds are docked with the target protein (PDB ID: 3EU7) which is encoded by breast cancer cell as, crystal structure of a PALB2 / BRCA2 complex, In all these experiments, the target protein is set as flexibly binding with ligand compounds, to ensure maximum degree of freedom. Here, the ligands could be able to explore various conformations during binding with target protein as an active site. For each ligand-protein docking output, 100 conformations of the ligand are generated. The ligand binding with the lowest free energy is considered for further analysis in every case.

The docked structures are further analyzed through PYMOL-v2.4.2 tool[42], Discovery Studio Visualizer-v21.1.0.20298[43] and LigPlot+ v.2.2.8[44]. Using AutoDockTools-v4.2.6, all the ligands and receptor are prepared for docking simulation and protonated. The solvation and default Kollman charge parameters were designate to the macromolecule atoms. Addition of Gasteiger charges to molecule as a ligand atom. The Lamarckian genetic algorithm (LGA) specifications are 100 runs, elitism of 1, the mutation rate of 0.02, the population size of 150, and a crossover rate of 0.8 band 2 500 000 energy evaluations. The docking outcomes are imaged using Discovery Studio Visualizer. The aim of this study is to find the efficient compound that binds well with the ER.

Results and Discussion

AI based virtual screening results

For generating new ligands, the SMILES of Tamoxifen drug is fed as input to the MolAICal by

selecting ZINCMol model and by varying the number of desired output ligands (1000, 8000, 12000), virtual screening experiments are conducted. Generated compounds are selected based on their structural dice similarity index >0.35 compared with Tamoxifen. While generating 1000 ligands, only 10 compounds have similarity index within the range 0.25-0.29. Similarly, for 8000 as the desired ligands, 21 compounds have similarity index range from 0.25-0.35. Whereas, for 12000 as ligands, 28 compounds (M1-M28) have similarity index above 0.35 dice-coefficient score and their SMILES string are shown in (Suppl. Fig. S1).

Analysis of Physicochemical properties

Using SWISS tool, physicochemical properties are extracted for the 28 compounds (M1-M28) and are described in (Table 1). If the compounds contain more than 3 aromatic rings are associated with poor developability. Compounds with higher rotatable bonds (limited to 10) have lower binding affinity with the target group, as the ligand affinity decreases by 0.5 Kcal for every 2 rotatable bonds. Other restrictions are, molar refractivity (MR) ranges between 40 and 130; and transport property correlated with TPSA (Topological Polar Surface Area) should be less than 140Å^2. Further, XLogP3 value is expected between -0.7 and 4.99. Poor absorption or permeation will be observed, when there are more than 5 H-bond donors, 10 H-bond acceptors, and the molecular weight (MWT) greater than 500 g/mol.

From Table 1, it is observed that, the molecular weights of all these 28 compounds M1-M28 are

Table 1 — Physicochemical parameters of the AI generated compounds (M1-M28)

Compound ID	Formula	MW (g/mol)	No. of Heavy atoms	No. of Aromatic heavy atoms	No. of Rotatable bonds	No. of H-bond acceptors	No. of H-bond donors	MR	TPSA (Å^2)	XLogP3
M1	C15H14FNOS	275.34	19	12	5	2	1	77.14	54.4	3.74
M2	C15H15NO2	241.29	18	12	5	2	1	71.51	38.33	2.18
M3	C17H20N2O2	284.35	21	12	7	2	2	82.56	50.36	2.71
M4	C15H15F2NO	263.28	19	12	5	3	0	71.5	12.47	3.66
M5	C17H16FNO3	301.31	22	12	7	4	1	81.51	55.4	3.34
M6	C13H15NS	217.33	15	11	4	0	1	67.88	40.27	3.91
M7	C15H14FNO	243.28	18	12	4	2	1	70.39	29.1	3.08
M8	C18H24N2O	284.4	21	12	8	1	2	90.45	33.29	4.82
M9	C14H12FNO	229.25	17	12	4	2	1	63.88	29.1	2.45
M10	C16H16O3	256.3	19	12	6	3	0	73.51	35.53	3.61
M11	C17H19NO2	269.34	20	12	7	2	1	81.13	38.33	4.7
M12	C13H10FNO	215.22	16	12	3	2	1	60.61	29.1	2.69
M13	C15H14FNO2	259.28	19	12	5	3	1	69.78	38.33	2.78
M14	C13H21NO3	239.31	17	6	8	4	2	66.42	50.72	0.66
M15	C11H16N2O2	208.26	15	6	7	2	2	57.91	50.36	0.72
M16	C15H14FNO	243.28	18	12	5	3	1	68.69	29.1	2.8
M17	C17H19NO2	269.34	20	12	7	2	1	79.76	38.33	3.73
M18	C16H17NO2	255.31	19	12	6	2	1	75.22	38.33	3.62
M19	C14H14O2	214.26	16	12	4	2	0	63.91	18.46	3.13
M20	C15H13NO3	255.27	19	12	5	3	1	71.94	55.4	3.31
M21	C15H15NO	225.29	17	12	4	1	1	70.43	29.1	2.98
M22	C22H17NO4	359.37	27	18	7	4	1	101.81	72.47	4.09
M23	C12H15FO3	226.24	16	6	7	4	0	57.77	35.53	1.96
M24	C11H14O4	210.23	15	6	7	4	0	54.53	44.76	1.26
M25	C17H19NO2	269.34	20	12	7	2	1	80.03	38.33	3.49
M26	C16H16O2	240.30	18	12	5	2	0	72.42	26.3	3.57
M27	C15H14O2	226.27	17	12	4	2	0	67.62	26.3	3.67
M28	C15H13FO2	244.26	18	12	5	3	0	67.41	26.3	3.31
Tamoxifen	C26H29NO	371.51	28	18	8	2	0	119.72	12.47	7.14

ranging between 208.26 and 359.37 g/mol only. Also, these compounds are satisfying all the other Lipinski's rule of 5 conditions; compounds have maximum of 18 heavy aromatic atoms and 8 rotatable bonds; MR with acceptable range between 54.53 and 101.81; TPSA ranges from 12.47Å^2 to 72.47Å^2; XLogP3 values are from 0.66 to 4.82. Based on the analysis of physicochemical parameters, all the 28 compounds (M1-M28) are validated as drug-like compounds.

Analysis of Pharmacokinetics properties

Drug-like compounds should be P-gp substrate and should have GI absorption above 70%. The BBB permeability values, >2.0, 0.1-2.0 and <0.1, indicates the highest, moderate, and the lowest absorption levels. The Caco-2 permeability levels, <4nm/s, 4-70nm/s and >70nm/s, indicates the low, middle, and high permeability respectively. There should be zero violations for Lipinski, Ghose, Veber, Egan and Mugge drug-likeliness filters.

From Table 2, the observed pharmacokinetics parameters indicate that, GI absorption level is significantly high for all the compounds (M1-M28). BBB permeability ranges between 0.0399798 and 3.59993. All the compounds satisfy the zero violation for various drug-likeliness filters, except M1. Similarly, all the compounds have high Caco-2 permeability, except M19. These compounds have higher aq. solubility ranging from $3.88e^{-8}$ to $1.40e^{-1}$ mol/L than the reference anticancer drug, Tamoxifen ($1.21e^{-09}$ mol/L). Based on the pharmacokinetic analysis, M1 and M19 are discarded, and the remaining 26 compounds (M2-M18, M20-M28) are analysed further for their toxicity.

Toxicity Analysis

With the support of PreADMET, an *in-silico* toxicity computational site, the toxicity predictions are made for carcino_mice and carcino_rat as reported in (Table 3). Based on toxicity analysis, 11

Table 2 — Pharmacokinetic parameters of the AI generated compounds (M1-M28)

Compound ID	GI absorption	P-gp substrate	Number of Violations					BBB Permeability	Caco-2 Permeability	Silicos IT Solubility
			Lipinski	Ghose	Veber	Egan	Muegge			
M1	High	Yes	1	0	0	0	0	0.276389	36.8285	8.04E-07
M2	High	No	0	0	0	0	0	0.234637	26.8016	3.52E-06
M3	High	Yes	0	0	0	0	0	2.59384	21.9625	1.17E-06
M4	High	Yes	0	0	0	0	0	1.69975	58.29	1.78E-06
M5	High	No	0	0	0	0	0	0.074238	33.9972	8.53E-07
M6	High	No	0	0	0	0	0	1.89188	46.9381	3.64E-06
M7	High	Yes	0	0	0	0	0	3.59993	43.1729	1.02E-06
M8	High	Yes	0	0	0	0	0	9.88777	26.3016	7.13E-08
M9	High	No	0	0	0	0	0	1.56631	40.03	2.48E-06
M10	High	No	0	0	0	0	0	0.0836689	53.5401	2.85E-06
M11	High	No	0	0	0	0	0	2.60305	42.3461	5.53E-07
M12	High	No	0	0	0	0	0	1.60281	37.3922	6.30E-06
M13	High	No	0	0	0	0	0	1.00219	40.1155	4.24E-06
M14	High	No	0	0	0	0	0	0.10549	29.6337	1.57E-04
M15	High	No	0	0	0	0	0	0.343625	21.215	1.59E-04
M16	High	No	0	0	0	0	0	0.966196	48.2273	9.77E-07
M17	High	Yes	0	0	0	0	0	2.49285	22.5773	5.53E-07
M18	High	No	0	0	0	0	0	0.574831	50.4171	1.39E-06
M19	High	No	0	0	0	0	0	1.64586	1.64586	6.40E-06
M20	High	No	0	0	0	0	0	0.0399798	21.1028	1.01E-05
M21	High	No	0	0	0	0	0	3.04296	32.6353	1.93E-06
M22	High	No	0	0	0	0	0	0.126027	21.025	3.88E-08
M23	High	No	0	0	0	0	0	0.0431294	51.1481	7.59E-05
M24	High	No	0	0	0	0	0	0.0411434	51.1962	1.40E-01
M25	High	No	0	0	0	0	0	0.946825	50.8197	1.49E-04
M26	High	No	0	0	0	0	0	0.307259	55.8344	3.75E-04
M27	High	No	0	0	0	0	0	1.80741	56.812	8.95E-04
M28	High	No	0	0	0	0	0	1.61368	54.6475	4.90E-04
Tamoxifen	Low	Yes	1	1	0	1	1	14.1639	49.5448	1.29E-09

Table 3 — Toxicity analysis of AI generated compounds (M1-M28)

Compound ID	Carcino_Mice	Carcino_rat	Compound	Carcino_Mice	Carcino_rat
M2	Negative	Negative	M15	Positive	Negative
M3	Positive	Negative	M16	Negative	Negative
M4	Negative	Negative	M17	Positive	Negative
M5	Positive	Negative	M18	Positive	Negative
M6	Positive	Negative	M20	Negative	Negative
M7	Positive	Negative	M21	Negative	Negative
M8	Positive	Negative	M22	Negative	Negative
M9	Positive	Negative	M23	Negative	Positive
M10	Negative	Negative	M24	Negative	Negative
M11	Positive	Negative	M25	Negative	Negative
M12	Positive	Negative	M26	Negative	Negative
M13	Positive	Positive	M27	Negative	Positive
M14	Negative	Negative	M28	Negative	Positive

compounds (M2, M5, M10, M14, M16, M20-M22, M24-M26) have been reported as negative in carcino_mice and carcino_rat toxicity tests. These 11 validated drug-like compounds (Table 3) alone are considered for further experimental analysis.

Molecular Docking studies with DNA (355D)

The mode of binding of drug-like inhibitors (M2, M5, M10, M14, M16, M20, M21, M22, M24, M25, M26) with DNA can be understand through Molecular Docking studies using the duplex sequence DNA d(CGCGAATTCGCG)2 dodecamer (PDBID: 355D). The docking poses of inhibitor's binding positions at the grooves of DNA have been depicted in (Fig. 2). From the docking pose as depicted in Figure 2, M20, M22, M24. M25 & M26 are bind to the minor grooves of DNA and rest of the compounds are bind in intercalative mode. K_i represents the concentration at which the inhibitor occupies 50% of the receptor sites. The smaller the K_i the greater the binding affinity and the smaller the amount of ligand is needed. Generally compounds have high inhibitory potential when $K_i < 1$ μM, moderate when1 μM < Ki < 10 μM, and weak Ki > 10 μM concentration.

It is observed that the compound M22 has high potential inhibitory activity and M2, M20 and M25 showed moderate inhibitory activity, which is similar to Tamoxifen (12.2 μM with -6.7 Kcal/mol), while M5, M10, M14, M16, M21 and M26 have lower activity. From DNA docking studies (Table 4), out of the 11 inhibitors, it is observed that, minimum energy conformations are obtained during docking of the inhibitors M22 & M25 with -8.61 Kcal/mol, -7.84 Kcal/mol as well has efficient inhibition constants as 0.486 M, 1.79 M, respectively.

Also, docked structures (depicted in Suppl. Fig. S2) are stabilized by electrostatic attractions, extensive Halogen bonds, hydrogen bonds, hydrophobic interaction, Alkyl interactions, and other π-Alkyl, π-donor hydrogen bond, π-sigma, π-sulphur, π-orbitals, and π-π T-shaped interactions with DNA 355D (Table S1). When assuming H-bonds to be perfectly linear, dipolar data indicate time-averaged hydrogen bond lengths of 1.80 +/- 0.03 Å for A-T and 1.86 +/- 0.02 Å for C-G. The inhibitor M22 with very high inhibitory activity is acting both as hydrogen donor and acceptor with DNA but also there is an intra molecular hydrogen bonding within the molecule with low bond lengths in the range of 1.9 Å and 2.2 respectively.

The inhibitor, M2 with moderate activity has both electrostatic interaction (~ 5.4 Å) and hydrogen bond interaction with DNA. There are two types of hydrogen bonds, namely, conventional hydrogen bond (~ 2.2 Å) and carbon- hydrogen bond (3.1 Å). The oxygen atom in the inhibitor M20 is hydrogen bonded with both the strands of DNA (A DNA and B DNA) with the bond length in the range of 1.7 Å to 3 Å. The carbon–hydrogen bond is formed between the carbon atom of M20 and oxygen atom of A strand with the bond length 3.12 Å. There is also Pi-Alkyl interaction with the alkyl group of M20 with B strand of DNA as hydrophobic interaction. M25 acts as hydrogen acceptor in the formation of conventional hydrogen bond formation with both A and B strands of DNA. This bond length ranges from 1.68 Å to 3.06 Å. The C-H bond is also formed between the carbon atoms of M25 with oxygen atom of A-DNA.

From the results of virtual screening, for the 11 drug like inhibitors, the docking nature in terms of electrostatic interactions with nearby amino acid

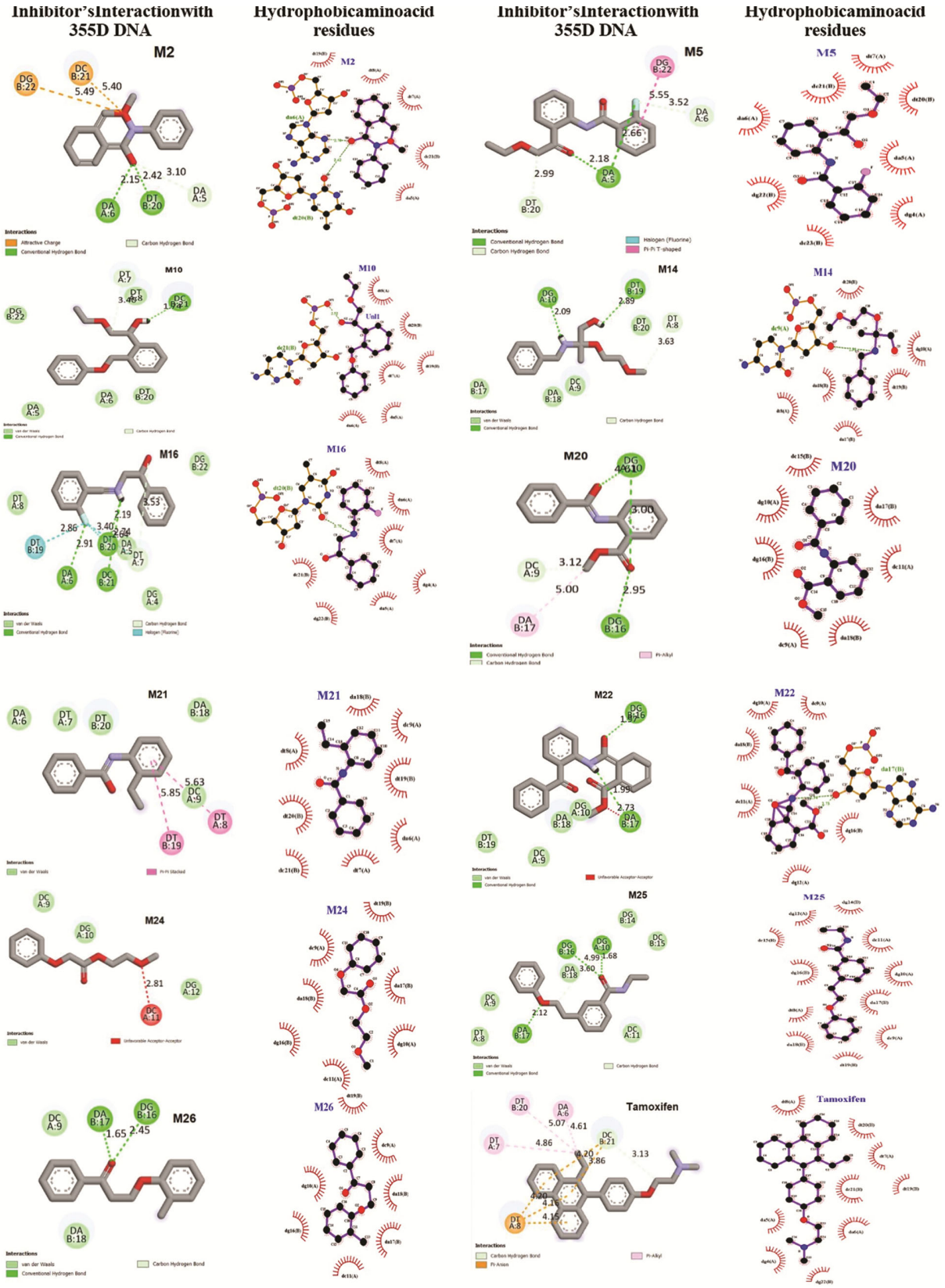

Fig. 2 — Drug-like inhibitor interactions in the binding pocket of 355D DNA receptor (M2, M5, M10, M14, M16, M20-M22, M24-M26)

Table 4 — Calculated binding energy and inhibition constant of the inhibitors with DNA, 355D at the temperature 298.15K

Compound ID	Free Energy Binding (Kcal/mol)	Inhibition Constant Ki (uM) Temp. = 298.15K	Nearby Residues
M2	-7.21	5.22	dcB:21, daA:5, daA:6, dtB:20,dgB:22
M5	-5.52	89.93	daA:6,daA:6, dgB:22,dtB:20
M10	-5.15	168.5	daA:5, daA:6, dtA:7, dtA:8, dt19, dtB:20, dcB:21, dgB:22
M14	-6.03	37.78	daB:17, daB:18, dtA:8, dtB:19, dtB:20, dcA:9, dGS:10
M16	-6.67	12.84	daA:5, daA:6, dtA:7, dtA:8, dtB::19, dtB:20, dcB:21, dgA:4, dgB:22
M20	-6.9	8.81	daB:17, dcA:9, dgA:10, dgB:16
M21	-6.81	10.23	daA:6, daB:18, dtA:7, dtA8, dtB:19, dtB:20, dcA:9
M22	-8.61	0.486	daB:17, daB:18, dcA:9, dtB:19, dgA:10, dgB:16
M24	-6.89	8.93	dcA:9, dcA:11, dgA:10, dgA:12
M25	-7.84	1.79	daB:17, daB:18, dtA:8, dcA:9, dcA:11, dcB:15, dgA:10, dgB:14, dgB:16
M26	-6.73	11.73	daB:17, daB:18, dcA:9, dgB:16
Tamoxifen	-5.62	75.85	daA:6, dtA:7, dtA:8, dtB:20, dcB:21

Table 5 — Calculated binding energy and inhibition constant values of the inhibitors, of the validated test molecules, with the breast cancer protease, 3EU7

Compound ID	Free Energy Binding (Kcal/mol)	Inhibition Constant Ki (uM) Temp. = 298.15K	Nearby Residues
M2	-7.30	4.45	Met875, Pro924, Val925, Val928, Tyr929, Gly1166
M5	-7.28	4.6	Met875, Pro924, Val925, Val928, Tyr929, Gly1166
M10	-5.46	99.85	Val925, Val932, Pro924, Pro926, Met875, Trp877, Leu931, Ile922, Val928, Ser873
M14	-5.58	81.17	Val925, Val932, Pro924, Met875, Val928
M16	-7.1	6.24	Val925, Val928, Pro926, Val932, Met875, Pro924, Asp927, Leu931
M20	-7.54	2.95	Met875,Phe876, Ile888, Ile922, Val925, Pro926, Val928, Val932, Cys933
M21	-7.51	3.1	Ala874, Met875, Ile922, Pro924, Val925, Val928, Val932, Cys933
M22	-8.05	1.25	Ala874, Met875, Pro924, Pro926, Lys1163,Gly1166
M24	-4.93	244.18	Ser873, Val925, Val928, Val932, Met875, Pro924, Ala874
M25	-7.86	1.72	Ala874, Met875, Pro924, Pro926, Val925, Val928, Val932
M26	-7.01	7.21	Phe876, Pro924, Pro926, Val925, Val928, Val932, Asp927, Met875
Tamoxifen	-6.7	12.2	Phe1071, tyr1064, Glu1066, Gln1020, Gly1021, Ser1065, Leu1092, Leu1142, Leu1143, Tyr1108

residue and various types of interactions in terms of bond length are depicted in (Fig. 2). This is further supported by the LigPlot+ v.2.2.8 studies (Fig. 2). Here the attention is not only given to the hydrogen bond interactions but also to the interactions with nearby amino acid residues through hydrophobic interactions (Suppl. Fig. S3).

Molecular Docking Studies with Breast Cancer Protease

Molecular docking of the selected 11 drug-like inhibitors (M2, M5, M10, M14, M16, M20, M21, M22, M24, M25 and M26) with breast cancer protease, 3EU7 is performed to elucidate the molecular mechanism. The studies are compared with that of the drug used for breast cancer, Tamoxifen (75.85 μM with -5.62 Kcal/mol). For each of the inhibitor, their calculated binding energy, inhibition constant values along with nearby residues is tabulated in (Table 5).

By analysing the docking data (Table S2), we can come to the conclusion that, similar to the binding studies of the inhibitors with DNA, the inhibitor M22 has the lowest binding energy -8.05 Kcal/mol with a minimum inhibition constant as 1.25 μM. In addition, the inhibitors (M2, M20, M25, M5, M16, M21, M26) have moderate inhibition value. Based on the inhibition value, the activity of the inhibitors are ranked in the following order as M22 > M25 > M20 > M21 > M2 > M5 > Tamoxifen > M16 > M26. Six of the drug-like inhibitors have more inhibition activity than Tamoxifen.

The greater activity of M22 inhibitor can be explained based on the structural analysis. On analysing the structural features of the inhibitors, the 6 inhibitors (M22, M25, M20, M21, M2 and M5) have amide (-CO-NH-) linkage except M16 and M26. The existence of this amide linkage may be responsible for higher activity than Tamoxifen. The inhibitor M22 alone has three benzene rings and the remaining inhibitors (M25, M20,

M21, M2, M5, M16 and M26) have only two benzene rings. The inhibitor M22 also has conjugation with 12 π bonds (24 π electrons) and 9 lone pair of electrons on 4 oxygen atoms and one nitrogen atom. The higher inhibition activity of M22 may be due to the presence of amide linkage, three benzene rings, extended conjugation and more number of labile electrons.

From Table S2, it is clearly observed that the docked structures are stabilized by hydrogen bonds, hydrophobic interaction, Alkyl interactions, and other π-Alkyl, π-donor hydrogen bond, π-sigma, π-sulphur, π-orbitals, and π-π T-shaped interactions with the protein, 3EU7). The docking results in terms of bond length and bonding nature are depicted (Fig. 3) for

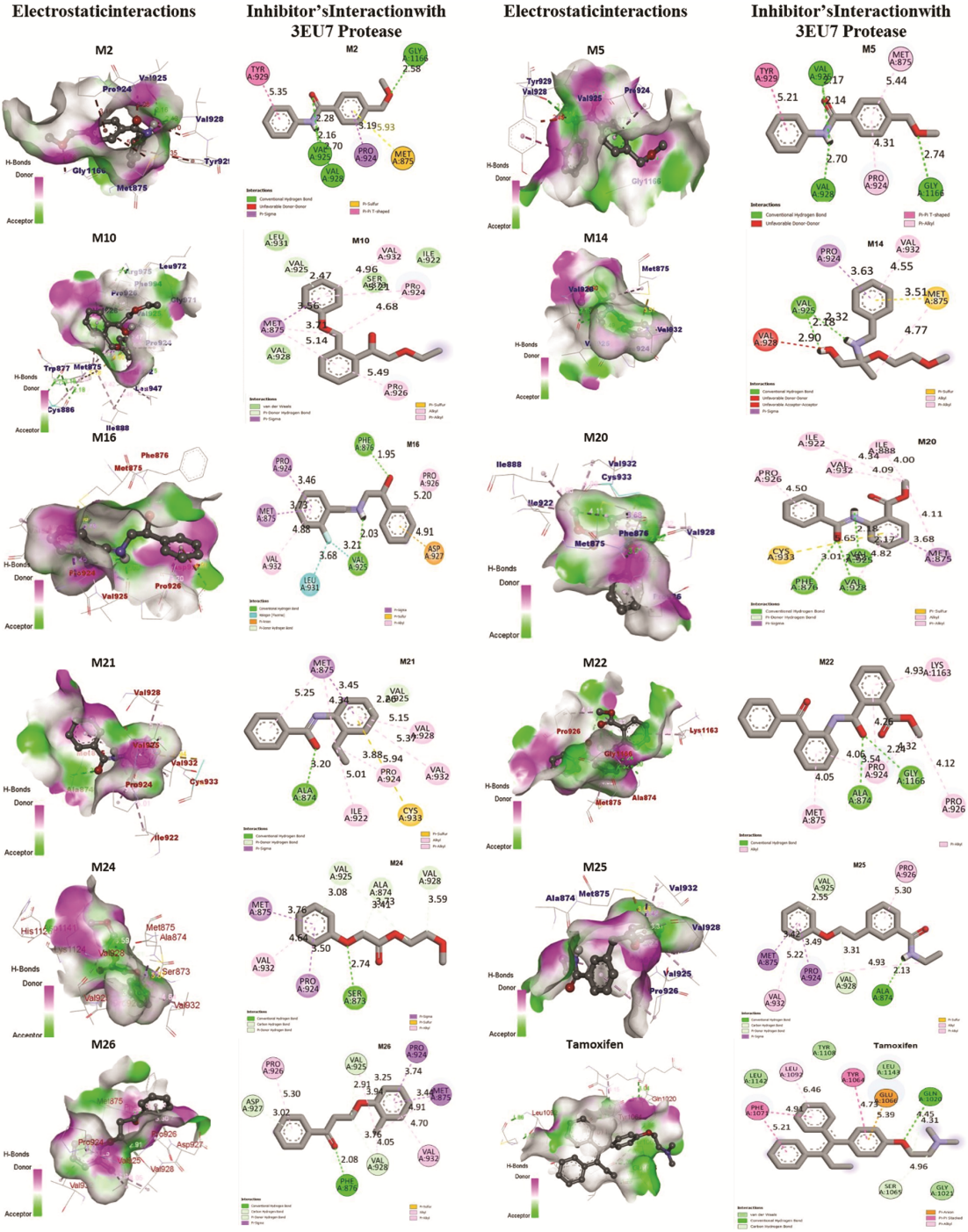

Fig. 3 — Electrostatic interactions and Ligand interactions (M2, M5, M10, M14, M16, M20, M21, M22, M24, M25, M26) in the binding pocket of the 3EU7 protein receptor

each of the 11 compounds. The hydrophobic interactions of the inhibitors with 3EU7 is also analysed with Ligplot program (Fig. 4). Here the attention is not only given to the hydrogen bond interactions but also to the interactions with nearby amino acid residues through hydrophobic interactions (Suppl. Fig. S4). This study helps to analyze their binding significance and interaction with protein 3EU7.

This work clearly illustrates the presence of O, N, F bonds. Halogen bond has gained widespread interest in the past years for hit-to-lead-to-candidate optimization aiming at improving drug–target binding affinity. In general, heavy organohalogens are able to form halogen bonds while organofluorines are not. Tamoxifen has only three π–π–alkyl interaction with Leu 1092(6.46 Å), Leu 1092 (6.46 Å) and Glu 1066 (4.45 Å). As depicted in (Fig. 3), the six inhibitor compounds (M2, M5, M20-M22, M25) have interactions with many amino acids in hydrophobic environment and non - bonded interactions; they show potent anti cancer activity than the drug, Tamoxifen.

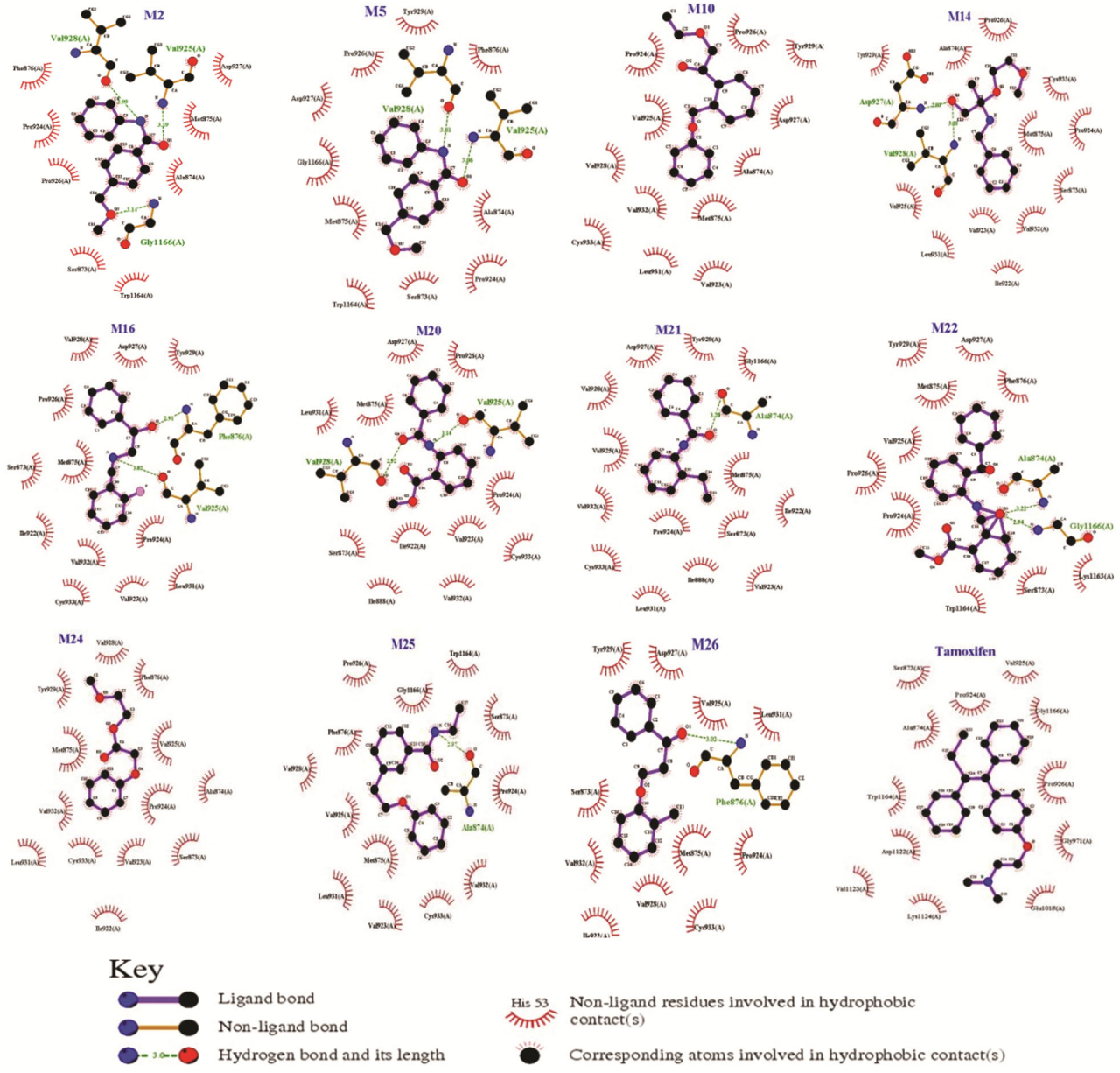

Fig. 4 — The hydrophobic amino acid residues which are in close contact were indicated in two dimension diagrams for ligands. Hydrogen bonds were shown as green dotted lines

Conclusion

In this work, MolAICal (ZINCMol) software is used to generate drug like inhibitors of the drug Tamoxifen for breast cancer (MCF 7 cell). The generated 28 inhibitors are filtered out with more than 0.35 similarity index and based on their physicochemical, pharmacological characteristics and toxicity analysis. The inhibition activity of these validated inhibitors with DNA, 355D and protein, 3EU7 is carried out through molecular docking studies and the results are compared with drug, Tamoxifen. It is detected that the inhibitor, M22 (methyl 2-[(2-benzoylphenyl) carbamoyl] benzoate), has the highest inhibition potential with 3 benzene rings, extended conjugation, amide linkage and huge number of labile electrons.

Hence, this work has explored the scientific approach for identifying a drug's structure efficiently and proved the same with the obtained results and gives the scope to synthesise that inhibitor experimentally for carrying out the *in-vitro* and *in-vivo* studies against breast cancer cell, MCF7.

Acknowledgement

Authors express their sincere thanks for the Managements of Sri S. Ramasamy Naidu Memorial College, Sattur and IP Research Centre, Department of CSE, National Engineering College, Kovilpatti for providing the necessary research infrastructures to carry out this work.

Conflict of interest

All authors declare no conflict of interest.

References

1 Kaan D, Assessment of cranberry bush on MCF-7 Human breast cancer cells. *Indian J Biochem Biophys*, 59 (2022) 985.

2 Mak KK & Pichika MR, Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today*, 4 (2019) 773.

3 Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B & Overington JP, ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 40 (2012) D1100.

4 Kim S, Thiessen P.A, Bolton E.E, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker B.A, Wang J, Yu B, Zhang J & Bryant SH, PubChem substance and compound databases. *Nucleic Acids Res*, 44 (2016) D1202.

5 Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C & Wilson M, DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, 46 (2018) D1074.

6 Purple Book: Database of Licensed Biological Products, U.S. Food and Drug Administration.

7 Irwin JJ & Shoichet BK, ZINC: a free database of commercially-available compounds for virtual screening. *J Chem Inf Model*, 45 (2005) 177.

8 Guedes IA, Pereira FSS, Dardenne LE, Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects & Challenges. *Front Pharmacol,* 9 (2018) 1089.

9 Daina A, Michielin O & Zoete V, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep,* 7 (2017) 42717.

10 Taskinen J & Yliruusi J, Prediction of physicochemical properties based on neural network modeling. *Adv Drug Deliv Rev*, 55 (2003) 1163.

11 Chandrabose S, Ishwar C & Sanjeev Kumar S, Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. *Mol Divers*, 26 (2022) 1893.

12 Geppert H, Horváth T, Gärtner T, Wrobel S & Bajorath J, Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J Chem Inf Model*. 48 (2008) 742.

13 Kang NS, Ahn JH, Kim SS, Chae CH & Yoo SE, Docking-based 3D-QSAR study for selectivity of DPP4, DPP8, and DPP9 inhibitors. *Bioorg Med Chem Lett,* 17 (2007) 3716.

14 Abdo A, Chen B, Mueller C, Salim N & Willett P, Ligand-based virtual screening using Bayesian networks. *J Chem Inf Model*, 50 (2010) 1012.

15 Zhang T, Leng J & Liu Y, Deep learning for drug-drug interaction extraction from the literature: a review. *Brief Bioinform*, 25 (2020):1609.

16 Xue D, Gong Y, Yang Z, Chuai G & Qu S, Advances and challenges in deep generative models for de novo molecule generation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9 (2018) e1395.

17 Alves LA, da Silva Ferreira NC, Maricato V, Alberto AVP, Dias EA & Coelho NJA, Graph Neural Networks as a Potential Tool in Improving Virtual Screening Programs. *Front Chem*, 9 (2021).

18 Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Cunha Faris LC & Aspuru-Guzik A, Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv*, 1705.10843

19 Wang Z, Zheng L, Liu Y, Qu Y, Li YQ, Zhao M, Mu Y & Li W, OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Front Chem,* 9 (2021).

20 Peter C. St. John, Caleb Phillips, Travis W. Kemper, A. Nolan Wilson, Yanfei Guan, Michael F. Crowley, Mark R. Nimlos & Ross E. Larsen, Message-passing neural networks for high-throughput polymer screening. *J Chem Phys,* 150 (2019) 34111.

21 Lim J, Hwang SY, Kim S, Moon S & Kim WY, Scaffold-based molecular design using graph generative model. *Chem. Sci,* 11 (2020) 1153.

22 Shen WX, Zeng X, Zhu F & Wang YL, Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat Mach Intell,* 3 (2021) 334.

23 Mercado R, Rastemo T, Lindelöf E, Klambauer G, Engkvist O, Chen H & Jannik B, Graph networks for molecular design, *Mach Learn: Sci Technol,* 2 (2021) 2632.

24 Qifeng B, Shuoyan T, Tingyang X, Huanxiang L, Junzhou H. & Xiaojun Y, MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Briefings in Bioinformatics*, 22 (2021) 1.

25 Kiyotani K, Mushiroda T, Nakamura Y & Zembutsu H, Pharmacogenomics of tamoxifen: Roles of drug metabolizing enzymes and transporters. *Drug Metab Pharmacokinet,* 27 (2012) 122.

26 Arjovsky M, Chintala S & Bottou L, Wasserstein Generative Adversarial Networks. *In Procs. of the 34th International Conference on Machine Learning*, PMLR 70 (2017).

27 Wang X & Ge F Quasi-sine Fibonacci M set with perturbation. *Nonlinear Dyn,* 69 (2012) 1765.

28 Leobacher G & Steinicke A, Exception Sets of Intrinsic and Piecewise Lipschitz Functions. *J Geom Anal,* 34 (2022) 118.

29 Quiroga R, Villarreal MA & Vinardo: a scoring function based on Autodock Vina improves scoring, docking, and virtual screening. *PLoS One*, 11 (2016) e0155183.

30 Muegge I, Heald SL & Brittelli D, Simple selection criteria for drug-like chemical matter. *J Med Chem,* 44 (2001) 1841.

31 Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW & Kopple KD, Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem,* 45 (2002) 2615

32 Ghose AK, Viswanadhan, VN & Wendoloski JJ, A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J Comb Chem*, 1(1999) 55.

33 Toppo AL, Yadav M, Dhagat S, Ayothiraman S & Jujjavarapu SE, Molecular docking and ADMET analysis of synthetic statins for HMG-CoA reductase inhibition activity. *Indian J Biochem Biophys*, 58 (2021) 127.

34 Egan WJ, Merz KM & Baldwin JJ, Prediction of Drug Absorption Using Multivariate Statistics. *J Med Chem,* 43 (2000) 3867.

35 Priyadharshini S, Akey Krishna Saroop, Jubie S, Jawahar N & Divecha V, Molecular Docking and cytotoxicity interactions of naringenin and its nano-structured lipid carriers in Era positive breast cancer. *Indian J Biochem Biophys*, 60 (2023) 141.

36 Kaan D, Assessment of cranberry bush on MCF-7 Human breast cancer cells. *Indian J Biochem Biophys*, 59 (2022) 985.

37 Hermansyah O, Bustamam A & Yanuar A, Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. *Comput Biol Chem,* 95 (2021) 107597.

38 Lipinski C, Lombardo F, Dominy BW & Feeney PJ, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv Drug Deliv Rev*, 46 (2001), 3.

39 Bennion BJ, Be NA, McNerney MW, Lao V, Carlson EM, Valdez CA, Malfatti MA, Enright HA, Nguyen TH, Lightstone FC & Carpenter TS, Predicting a Drug's Membrane Permeability: A Computational Model Validated with in Vitro Permeability Assay Data. *J Phys Chem B,* 121 (2017) 5228,

40 Asif M, Alvi SS, Azaz T, Khan AR, Tiwari B, Hafeez BB & Nasibullah M, Novel Functionalized Spiro [Indoline-3,5′-pyrroline]-2,2′dione Derivatives: Synthesis, Characterization, Drug-Likeness, ADME, and Anticancer Potential. *Int J Mol Sci,* 24 (2023) 7336.

41 Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS & Olson AJ, Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat Protoc,* 11 (2023) 905.

42 PyMOL: A user-sponsored molecular visualization system on an open-source foundation, maintained and distributed by Schrödinger.

43 BIOVIA: Discovery Studio Visualizer: a free, feature-rich molecular modeling application for viewing, sharing and analyzing protein and small molecule data.

44 Laskowski RA, Swindells MB, LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model*, 51 (2011) 2778.