The use of machine learning-based sequential virtual screening in the search for new ligands of 5-HT$_6$ receptor

Michał Sapa[1], **Alicja Gawalska**[1], **Marcin Kołaczkowski**[1], **Adam Bucki**[1]

[1]Jagiellonian University Medical College, Faculty of Pharmacy, Department of Medicinal Chemistry, Kraków, Poland

Farmacja Polska, ISSN 0014-8261 (print); ISSN 2544-8552 (on-line)

Corresponding author
Adam Bucki,
Jagiellonian University Medical College,
Faculty of Pharmacy,
Department of Medicinal Chemistry.
Medyczna 9, 30-688 Kraków, Poland
e-mail: adam.bucki@uj.edu.pl

Sources of financing
No sources of financing were indicated.

Conflict of interest
No sources of financing were indicated.

Received: 2022.12.16
Accepted: 2023.02.20
Published on-line: 2023.04.05

DOI
10.32383/farmpol/161474

ORCID
Michał Sapa - 0000-0001-9251-9261
Alicja Gawalska - 0000-0002-3131-9300
Marcin Kołaczkowski - 0000-0001-8402-1121
Adam Bucki - 0000-0003-0451-9814

The use of machine learning-based sequential virtual screening in the search of new ligands of 5-HT$_6$ receptor

5-HT$_6$ receptor takes part in learning and memory processes. For this reason, the use of ligands of this receptor in the treatment of neurodegenerative diseases such as Alzheimer's disease, depression, or autism is being investigated. The development of machine learning (ML) and access to large compound databases allow for the increasing use of these methods in search of new drugs. The use of ML in pre-clinical tests allows for a reduction in time and costs of drug discovery. In this study, we used a sequential virtual screening approach in search of new structures with potential affinity for the 5-HT$_6$ receptor. Data from the ChEMBL database containing ligand binding affinities, measured as an inhibition constant (Ki), was used as the training dataset. Each step of the screening was based on machine learning models, the task of which was to classify compounds as potentially active and inactive. The first step included a ligand-based drug discovery (LBDD) approach, in which, using Klekota-Roth fingerprints and molecular descriptors of the ligands, a classification model was developed to select a preliminary group of candidates from the Otava chemical compound database. In the second step, a structure-based drug discovery (SBDD) approach was used. For this purpose, compounds were docked to the AlphaFold database-derived model of the 5-HT$_6$ receptor, previously optimized by the Induced-Fit Docking tool and molecular dynamics. Docking poses were scored by a trained Extra Trees classifier. Interactions of a reference ligand with 14 binding site residues were used as features for the trained model.
The use of machine learning as a scoring function allowed for improvement in the virtual screening parameters compared to the Glide GScore scoring function. Based on the obtained model, it was also confirmed that the location of a ligand near Ser5.43 and Phe5.38 residues is important for binding. The procedure allowed to select 20 candidates characterized by novel chemical structure and a relatively low basic pKa compared to known ligands, and thus suspected to have a low affinity for hERG channels and good brain penetration.

Keywords: molecular docking, machine learning, structure-activity relationship, serotonin 6 receptor.

© Farm Pol, 2022, 78(11): 607-614

Introduction

Serotonin 6 receptors belong to the large family of serotonin receptors and like most of them are G protein-coupled receptors (GPCR). They are present mainly in the central nervous system and are assigned to participate in the regulation of learning and memory processes. The presence of the 5-HT_6 receptors on GABAergic and glutamatergic neurons suggests that they may be involved in the regulation of the transmission of these systems [1].

Due to their role in learning and memory processes, ligands of the 5-HT_6 receptor are being investigated for the treatment of neurodegenerative diseases. The main studies are focused on the use of 5-HT_6 receptor antagonists in the treatment of Alzheimer's disease. Although there were 6 candidates in clinical trials for the treatment of Alzheimer's disease, none of them passed phase II [2]. All compounds tested in clinical trials were antagonists, however, both antagonists and agonists have been shown to be pro-cognitive in animal models [3, 4]. Additionally, 5-HT_6 receptor antagonists have been tested for use in the treatment of depression, anxiety, autism, IBS disease, or obesity [5–9].

The increasing amount of available data and rapid development of machine learning (ML) allow for the widespread use of these methods in the search for new drugs. Machine learning can help reduce the costs associated with testing new drugs in the early stages of drug discovery. Thanks to machine learning, for example, QSAR modeling or docking scoring has been improved [10, 11].

In this study, we used machine learning methods in a 2-step sequential virtual screening, in search of new ligands of the 5-HT_6 receptor. In the first step, ligand-based drug discovery (LBDD) was used to reduce the number of compounds in the screened database. In the second step, structure-based drug discovery (SBDD) was used by docking compounds to the 5-HT_6 receptor model. ML classifier, trained by poses of known ligands, served as a scoring function. The sequential approach aims to reduce the time and computational costs of molecular docking in the second step [12]. The use of the machine learning model, in turn, can improve the parameters of virtual screening compared to classic scoring functions [13].

Aim

The aim of the study was to develop a sequential virtual screening procedure based on machine learning models and to use this procedure to search for new potential 5-HT_6 receptor ligands.

Methods

A dataset of known ligands and their binding affinity (measured as an inhibition constant, K_i) from the ChEMBL database [14, 15] (date of database download: June 14, 2022) was used to train and evaluate ML models (table 1). From a dataset containing 5,205 records, missing values were removed, resulting in 4,269 records. Subsequently, duplicates were removed, resulting in 3,620 records eventually. Ligands were labeled based on binding affinity – compounds with K_i < 100 nM were labeled as "active", and those with K_i > 700 nM as "inactives" (Dataset A). These thresholds ensured the separation of active from inactive compounds, while providing a possibly large number of "inactives", which were sparsely represented (ratio of inactives to actives above 0.25). Additionally, after clustering the active compounds based on fingerprints, a group of 65 structurally diverse ligands was selected and 3099 DUDE decoys were generated based on their structures [16] (Dataset B). This dataset was used to calculate the enrichment parameters of virtual screening. In prospective virtual screening, the Otava In-house Stock dataset (date of database download: July 21, 2022) was used after filtering structures with molecular weight (MW) values ranging from 300 to 500.

The sequential virtual screening consisted of 2 main steps grounded in ligand-based (LBDD) and structure-based drug discovery (SBDD). In the first step, 1D, end 2D descriptors and Klekota-Roth fingerprints were generated using PaDEL-Descriptor software [17, 18]. Low variance features were removed by VarianceThreshold (threshold = 0.2) algorithm the from scikit-learn library. The number of features was reduced from

Table 1. Datasets used in training and evaluations of ML models. For SBDD procedure, the number of poses is given. K_i – inhibition constant.

Dataset	Active molecules	Inactive molecules
Dataset for training and testing ML models (Dataset A)	2168 structures (K_i < 100 nM) 6127 poses	615 structures (K_i > 700 nM) 1482 poses
Dataset for evaluation of virtual screening (Dataset B)	65 structures (K_i < 100 nM) 185 poses	3099 structures (DUDE decoys) 7469 poses

6304 to 616. PyCaret was used for further data preprocessing and to compare the created ML models [19]. Perfect collinear features were removed, data was standardized and the imbalance between active and inactive ligands was removed by SMOTE [20]. PyCaret comparison was performed with 10-fold cross-validation for 14 classifiers (table 2 and table 4). The best model was chosen based on Matthew's correlation coefficient (MCC). The chosen model was additionally tuned using the default grid of hyperparameters. The developed model (LBDD model) was used in the virtual screening of the Otava dataset.

The model of the 5-HT_6 receptor was downloaded from the AlphaFold platform [21, 22]. Then, structure refinement was performed by Protein Preparation Wizard (Schrödinger, LLC) [23]. Using the induced-fit protocol, a selected group of rigid ligands [24, 25] was docked to the refined receptor structure [26]. The complex with **CHEMBL2165519** ligand (a high-affinity, arylsulfonyl ligand of the 5-HT_6 receptor, K_i = 1.0 nM) was selected for further optimization as the best protein-ligand complex, on the grounds of the literature description of binding mode [27–30]. Next, a system of the protein complex, membrane (DPPC 325K, based on 4IAR – crystal structure of 5-HT_{1B} receptor), water (SPC model), and salt (NaCl 0.15 M) was set. A 100 ns molecular dynamics simulation was carried out for 325K temperature, 1.01325 bar pressure, and 200 bar×Å surface tension using Desmond Molecular Dynamics software and OPLS2005 force field [31]. The obtained 500 trajectories were clustered into 12 models, to which the actives from set B were docked by Glide Ligand Docking [32]. The best model was chosen based on the number of successfully docked active compounds (met the criteria of constraint – H-bond with Asp3.32). For the selected model, only one active compound was not docked.

During these studies, a crystal structure of the 5-HT_6 receptor was published (PDB: 7XTB) [33]. This automatically became a reference structure for further modeling studies; however, the alignment of both models showed an RMSD value of 1.945 Å and a favorable alignment score of 0.154. In addition, only 26.2% of actives (from Dataset B) were successfully docked (meeting the constraint criterion – H-bond with Asp3.32), so the crystal structure would require further optimization to challenge the optimized AlphaFold model. Eventually, we decided to continue work on the latter structure.

The second step started with generating 3D structures of compounds by LigPrep (for pH = 7.0 and OPLS2005 force field) [34]. Then, the compounds were docked to the selected model by Glide Ligand Docking [32], setting SP protocol and Asp3.32 as a constraint. For each compound, up to 3 conformations were generated, increasing the possibility of correct binding, regardless of the GScore value. Interaction energy, van der Waals interaction, Coulomb interaction, H-bond score, and minimum distance (calculated by Glide) for 14 residues of the binding site were used as features. Results of Dataset A docking were used to train and test classifiers. Using PyCaret [19], perfect collinear features were removed, data was standardized and the imbalance between active and inactive ligands was removed by SMOTE [20]. Model

Table 2. Comparison of 14 classifiers built by PyCaret for the LBDD procedure, sorted by MCC value. AUC – area under the ROC curve, F1 – F1 score, Kappa – Cohen's kappa coefficient, MCC – Matthew's correlation coefficient.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Light Gradient Boosting Machine	0.9162	0.9605	0.9539	0.9401	0.9469	0.7484	0.7495
Extra Trees Classifier	0.9127	0.9575	0.9462	0.9427	0.9443	0.7420	0.7431
Gradient Boosting Classifier	0.9097	0.9511	0.9417	0.9431	0.9423	0.7343	0.7352
Ridge Classifier	0.9037	0.0000	0.9238	0.9521	0.9375	0.7268	0.7292
Random Forest Classifier	0.9052	0.9550	0.9353	0.9432	0.9391	0.7248	0.7261
Logistic Regression	0.8982	0.9378	0.9231	0.9457	0.9341	0.7093	0.7111
SVM – Linear Kernel	0.8962	0.0000	0.9180	0.9480	0.9326	0.7064	0.7094
Linear Discriminant Analysis	0.8941	0.9190	0.9212	0.9424	0.9315	0.6980	0.6998
K Neighbors Classifier	0.8821	0.9114	0.9026	0.9445	0.9230	0.6719	0.6758
Ada Boost Classifier	0.8866	0.9330	0.9238	0.9313	0.9272	0.6700	0.6725
Decision Tree Classifier	0.8645	0.8079	0.9110	0.9159	0.9133	0.6029	0.6037
Naive Bayes	0.8460	0.8525	0.8648	0.9337	0.8978	0.5868	0.5945
Quadratic Discriminant Analysis	0.7908	0.5208	0.9974	0.7904	0.8819	0.0628	0.1547
Dummy Classifier	0.2168	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3. Comparison of 14 classifiers built by PyCaret for the SBDD procedure, sorted by MCC value. AUC – area under the ROC curve, F1 – F1 score, Kappa – Cohen's kappa coefficient, MCC – Matthew's correlation coefficient.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Extra Trees Classifier	0.8977	0.9336	0.9361	0.9368	0.9364	0.6741	0.6745
Random Forest Classifier	0.8877	0.9270	0.9238	0.9361	0.9298	0.6494	0.6508
Light Gradient Boosting Machine	0.8721	0.9176	0.9072	0.9324	0.9194	0.6098	0.6128
K Neighbors Classifier	0.8436	0.8979	0.8489	0.9517	0.8972	0.5741	0.5912
Gradient Boosting Classifier	0.8269	0.8909	0.8375	0.9411	0.8861	0.5299	0.5462
Ridge Classifier	0.7807	0.0000	0.7815	0.9357	0.8514	0.4448	0.4724
Linear Discriminant Analysis	0.7807	0.8370	0.7815	0.9357	0.8514	0.4448	0.4724
Logistic Regression	0.7820	0.8369	0.7843	0.9346	0.8526	0.4454	0.4721
Ada Boost Classifier	0.7957	0.8416	0.8125	0.9248	0.8649	0.4528	0.4694
SVM – Linear Kernel	0.7587	0.0000	0.7636	0.9238	0.8359	0.3951	0.4220
Decision Tree Classifier	0.7995	0.7212	0.8501	0.8956	0.8722	0.4079	0.4111
Naive Bayes	0.4426	0.7955	0.3186	0.9701	0.4723	0.1376	0.2486
Quadratic Discriminant Analysis	0.4631	0.7387	0.3686	0.9364	0.5016	0.1163	0.2070
Dummy Classifier	0.1947	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 4. Evaluation metrics of selected ML classifiers for the LBDD and SBDD procedures. AUC – area under the ROC curve, F1 – F1 score, Kappa – Cohen's kappa coefficient, MCC – Matthew's correlation coefficient.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
LBDD procedure							
Light Gradient Boosting Machine (tuned)	0.9167	0.9617	0.9513	0.9430	0.9470	0.7520	0.7529
Light Gradient Boosting Machine	0.9162	0.9605	0.9539	0.9401	0.9469	0.7484	0.7495
SBDD procedure							
Extra Trees Classifier	0.8977	0.9336	0.9361	0.9368	0.9364	0.6741	0.6745
Extra Trees Classifier (the best poses by GScore)	0.8647	0.9141	0.9222	0.9100	0.9159	0.5687	0.5701

comparison was performed with 10-fold cross-validation for 14 classifiers (table 3 and table 4). Extra Trees Classifier was selected as the best model (SBDD model), further tuning of hyperparameters was unsuccessful. Additionally, the best poses by GScore for each ligand were selected. Then, they were preprocessed and compared likewise but the obtained classifier was worse than the previous one (table 4).

The selected candidates were compared in terms of structural similarity with the compounds from the ChEMBL database, calculating the Tanimoto coefficient for generated fingerprints by the RDKFingerprint algorithm from RDKit Python package [35]. Values of pK_a for the strongest basic atom were calculated by InstantJChem from the ChemAxon package [36].

Results and discussion

The evaluation of the importance of the LBDD model features (table 5) showed that descriptors describing the appearance of nitrogen atoms and van der Waals volumes are crucial for 5-HT_6 ligands. When it comes to fingerprints, the KRFP4853, corresponding to many arylsulfonyl derivates and the KRFP467, corresponding to secondary nitrogen atoms, are the most important in the ligand structure. The LBDD procedure allowed for the selection of 1,599 potential candidates from the Otava database of 195,942 compounds.

As to the SBDD model, the most important features (table 6) revolved around Phe5.38 and Ser5.43 residues located in the orthosteric site of the 5-HT_6 receptor. This confirmed that interactions with these residues, along with aromatic interactions with Phe6.51/6.52 and ionic bonds with Asp3.32 are crucial for ligand binding [27-30].

The use of multiple docking poses allowed for improveing parameters of the machine learning model (table 4); however, it must be taken into account that potentially inappropriate conformations of active compounds may be a bias for the ML model. A possible way to deal with that issue appears to be the use of a pose filter (accessible

Table 5. Feature importance for the LBDD model (Light Gradient Boosting Machine).

Feature	Value	Description / SMARTS
VE3_Dzm	39	Logarithmic coefficient sum of the last eigenvector from Barysz matrix / weighted by mass
VR2_Dt	32	Normalized Randic-like eigenvector-based index from detour matrix
MDEC-33	28	Molecular distance edge between all tertiary carbons
minsssN	24	Minimum atom-type E-State: >N-
SpMAD_Dzs	23	Spectral mean absolute deviation from Barysz matrix / weighted by I-state
AATSC2v	22	Average centered Broto-Moreau autocorrelation - lag 2 / weighted by van der Waals volumes
VE3_Dzv	22	Logarithmic coefficient sum of the last eigenvector from Barysz matrix / weighted by van der Waals volumes
maxaaN	21	Maximum atom-type E-State::N:
SssNH	20	Sum of atom-type E-State: -NH-
KRFP4853	20	S(c1ccccc1)
KRFP467	20	[!#1][CH2][NH][!#1]

in the Schrödinger package), which selects poses upon knowledge-based binding criteria. Nevertheless, the predicted poses might still be burdened with error and since the rate of success in docking novel compounds may be only estimated based on the ability to reproduce the experimental binding pose [37], the method of fast and reliable selecting appropriate docking poses should be considered in the future.

Using Dataset B, the evaluation of virtual screening was carried out. Docking results were sorted by GScore values and the SBDD model scores, and then virtual screening parameters were calculated for both rankings. ML-based scoring improved the ability to identify active compounds compared to the GScore scoring function (figure 1 and table 7). A particular improvement was made in early recognition, where the enrichment factor for 1% of the dataset (EF1%) increased more than 7-fold.

The LBDD procedure reduced the number of compounds from 195,942 to 1,599 in the screened database (compounds with a score greater than 0.9). After docking the selected candidates and the SBDD model prediction, a group of 20 compounds was selected according to the SBDD model score (equal to or greater than 0.85) (figure 2 and table 8). In addition, several of the selected compounds also had a high GScore value (for comparison, for the best 1% Dataset B, the GScore values range was -9.15 to -10.27).

Among the compounds selected from the Otava database, the largest group was arylsulfonyloxazole derivatives. Compared to the compounds from the ChEMBL database, only CHEMBL1714441 (K_i = 61 nM) [38] showed a high structural similarity to the selected candidates (Tanimoto coefficient > 0.8). Two of the highest-rated compounds (14 and 15) are non-bases and therefore they did not form a key ionic bond with Asp3.32. However, provided that *in vitro* tests would verify their activity, there's a chance to find valuable novel non-basic chemotypes of the 5-HT_6 receptor ligands.

Table 6. Feature importance for the SBDD model (Extra Trees Classifier).

Per-residue interaction	Residue	Value
resA193_vdw	Ser5.43	0.034755
resA188_vdw	Phe5.38	0.033314
resA193_dist	Ser5.43	0.032987
resA106_Eint	Asp.3.32	0.028502
resA188_Eint	Phe5.38	0.027725
resA192_Eint	Ala5.42	0.027599
resA193_Eint	Ser5.43	0.027246
resA188_dist	Phe5.38	0.023902
resA106_coul	Asp.3.32	0.023262
resA192_coul	Ala5.42	0.022040

The poses obtained in the docking process were aligned and the binding mode mostly

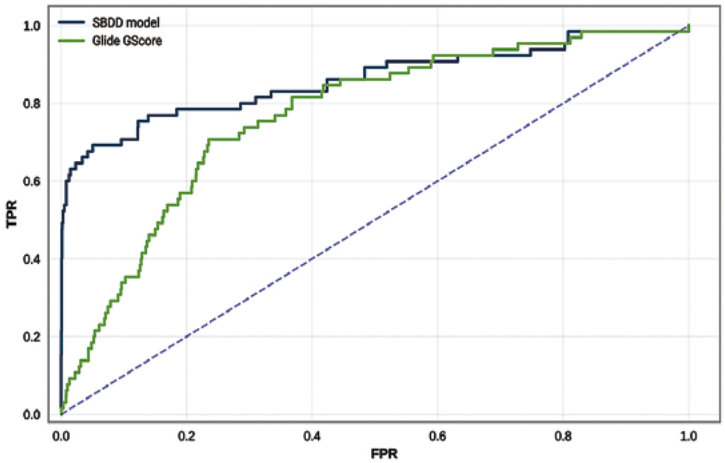

Figure 1. ROC plot for virtual screening results of Dataset B, ranked by the SBDD model scores (blue line) and Glide GScore (green line). TPR – true positive ratio, FPR – false positive ratio.

Table 7. Virtual screening parameters for results of Dataset B docking, ranked by the SBDD model scores and Glide GScore. BEDROC – Boltzman enhanced discrimination of the ROC curve, ROC – area under the ROC curve, EF – enrichment factor.

Parameter	Glide GScore	SBDD model (Extra Trees Classifier)
$BEDROC_{alpha=20}$	0.208	0.693
ROC	0.74	0.85
EF1% (NEF1%)	6.08 (12.5%)	44.11 (90.6%)
EF5% (NEF5%)	3.39 (16.9%)	13.25 (66.2%)
EF10% (NEF10%)	3.23 (32.3%)	6.93 (69.2%)

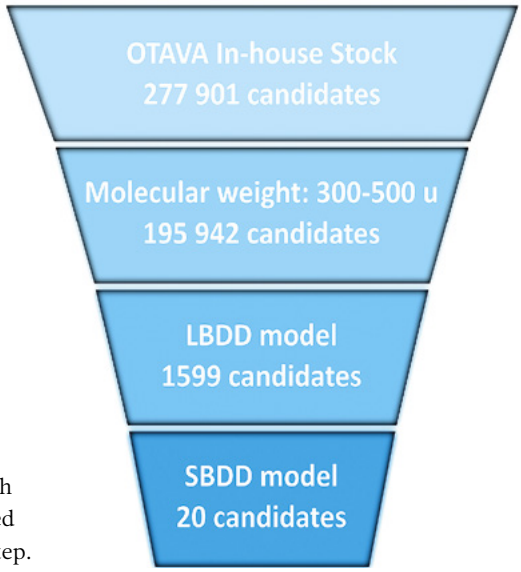

Figure 2. Scheme of sequential virtual screening with the number of selected compounds at each step.

corresponded to that described in the literature [27–30]. The key ionic bond with Asp3.32, the π–π interaction with the Phe6.51/Phe6.52 side chains, the hydrogen bond acceptor (HBA) moiety near Asn6.55 and Ser5.43, and an aromatic fragment near Phe5.38 were present. An example of the binding method of one of the selected compounds is presented in figure 3.

Ma et al. in their publication stated that compounds with a basic pK_a below 5.2 are probably weak inhibitors of hERG channels [39]. Among the basic candidates, only compound **13** met this criterion. Of the 5-HT_6 receptor ligands with an affinity lower than 100 nM, only about 14% have basic pK_a lower than 7.0. For comparison, 11 out of 20 candidates had basic pK_a lower than 7.0 (figure 4 and table 8) and these were compounds that had a basic nitrogen atom in the imidazole or

Table 8. Results for selected 20 compounds.

	Glide GScore	LBDD model	SBDD model	Smiles	Strongest basic pK_a
1	–8.74	0.95	0.96	O=S(=O)(c1ccc2c(c1)OCCO2)c1nc(-c2ccco2)oc1NCCCn1ccnc1	6.53
2	–9.47	0.90	0.93	Cc1ccc(S(=O)(=O)c2nc(-c3ccccc3)oc2NCCN2CCOCC2)cc1	5.23
3	–7.20	0.93	0.93	CN1CCN(c2ccc(NS(=O)(=O)c3ccc4c(c3)CCCC4)cc2)CC1	7.82
4	–9.60	0.96	0.92	Cc1ccc(S(=O)(=O)c2nc(-c3ccccc3Cl)oc2NCCN2CCOCC2)cc1	5.23
5	–9.53	0.96	0.92	O=S(=O)(c1ccccc1)c1nc(-c2ccccc2Cl)oc1NCCCn1ccnc1	6.53
6	–8.20	0.92	0.92	CC(CNC(=O)c1ccc(S(=O)(=O)c2ccccc2)o1)N1CCCCC1	7.90
7	–8.12	0.91	0.92	Cc1ccc(S(=O)(=O)c2nc(-c3ccccc3)oc2NCCCN2CCOCC2)cc1	6.42
8	–8.95	0.94	0.91	Cc1ccc(S(=O)(=O)c2nc(-c3ccccc3Cl)oc2NCCCN2CCOCC2)cc1	6.41
9	–5.44	0.91	0.89	Cc1nc(-c2cccs2)c(CCC(=O)Nc2ccc(N3CCN(C)CC3)cc2)s1	7.86
10	–9.29	0.92	0.88	O=S(=O)(c1ccccc1)c1nc(-c2ccccc2Cl)oc1NCCCN1CCOCC1	5.23
11	–8.57	0.95	0.88	Cc1ccccc1-c1nc(S(=O)(=O)c2ccc(Cl)cc2)c(NCCCN2CCOCC2)o1	6.42
12	–8.33	0.98	0.87	CCN(CC)CCn1c(N)c(S(=O)(=O)c2ccccc2)c2nc(C#N)c(C#N)nc21	8.51
13	–7.62	0.98	0.86	Cc1ccc(S(=O)(=O)c2nc(C)sc2NCCN2CCOCC2)cc1	5.03
14	–5.61	0.94	0.86	CC(C)CCNC(=O)c1cc(S(=O)(=O)C2CCS(=O)(=O)C2)ccc1F	–1.63
15	–5.35	0.95	0.86	COC(CNC(=O)c1cc(S(=O)(=O)C2CCS(=O)(=O)C2)ccc1F)OC	–1.94
16	–5.19	0.92	0.86	Cc1nc(-c2cccs2)c(CC(=O)Nc2ccc(N3CCN(C)CC3)cc2)s1	7.86
17	–8.91	0.98	0.85	O=S(=O)(c1ccccc1)c1nc(-c2cccs2)oc1NCCN1CCOCC1	5.22
18	–7.14	0.92	0.85	O=C(C1=C(O)C(=O)N(CCCN2CCOCC2)C1c1ccco1)c1cc2ccccc2o1	6.69
19	–6.80	0.93	0.85	CN(C)CCCNc1nn2c(=O)c3c(-c4ccc(Cl)cc4)csc3nc2c2ccccc12	8.80
20	–6.74	1.00	0.85	CN1CCC(N(C)c2oc(-c3ccccc3Cl)nc2S(=O)(=O)c2ccccc2)CC1	7.97

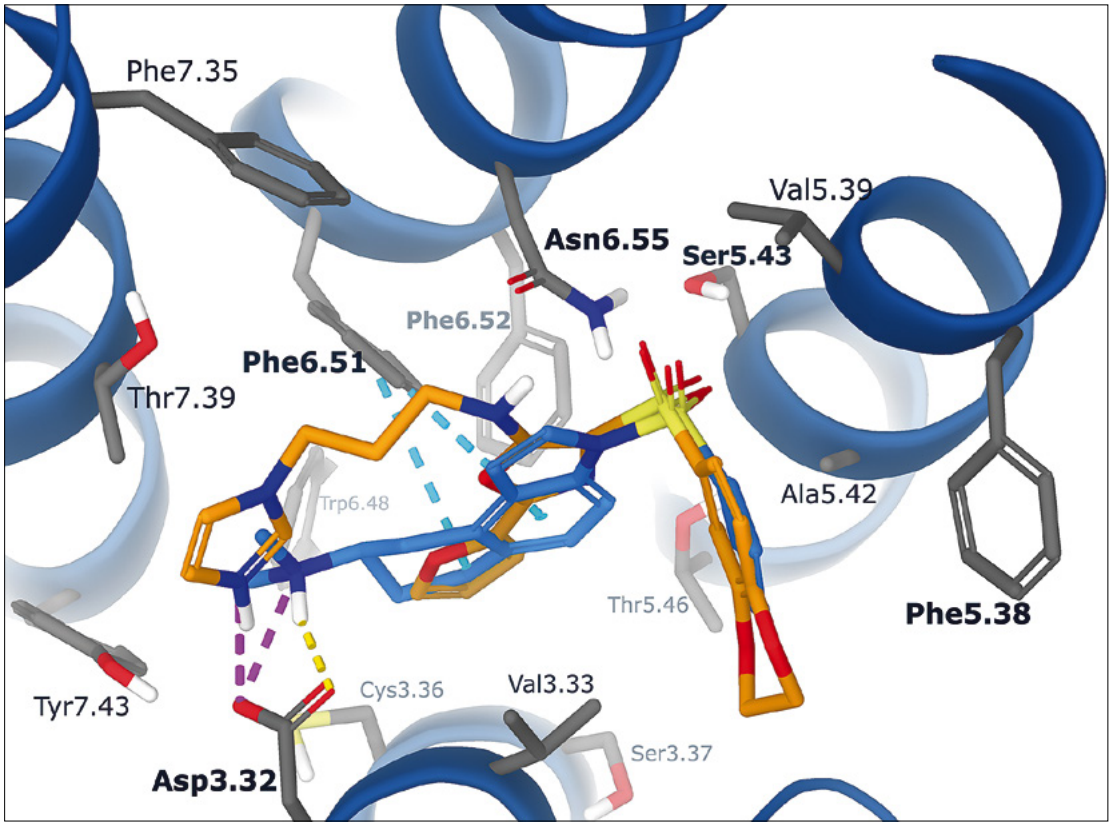

Figure 3. Comparison of the binding mode of compound 1 (orange) and CHEMBL2165519 (blue). Both compounds formed ionic bonds with Asp3.32. 2-(furan-2-yl)-1,3-oxazole and 3H,6H,7H,8H,9H-cyclohexa[e]indole moieties are aligned and formed π-π stacking with Phe6.51. Additionally, arylsulfonyl moieties of both compounds are placed between Asn6.55/Ser5.43 and Phe5.38. Top view, without ECL2. Side chains of selected residues are displayed.

morpholine ring. High basic pK_a can also cause poor brain penetration connected with low permeability and high P-gp (P-glycoprotein) interaction. Therefore, compounds with $pK_a < 8.0$ are thought to avoid the P-gp-mediated efflux [40]. In our study, 18 out of 20 compounds met this criterion (figure 4 and table 8).

Conclusions

The developed procedure allowed us to obtain new chemotypes of potential 5-HT_6 receptor ligands, differing from the existing structures. Most of the selected compounds were arylsulfonyloxazole derivatives, whereas, in the ChEMBL database of known ligands, only one compound had a similar structure. The compounds also had relatively low pK_a, therefore they may have greater brain penetration and a lower risk to bind to the hERG channel. Further studies are needed to verify the potential of these derivatives as 5-HT_6 receptor ligands.

The use of machine learning allowed for improving the parameters of virtual screening, compared to the standard scoring function. Additionally, on the basis of the ML model, it was possible to determine the importance of the interactions necessary for active compound binding. The SBDD model confirmed that Phe5.38 and Ser5.43 residues are the most important in recognizing active molecules.

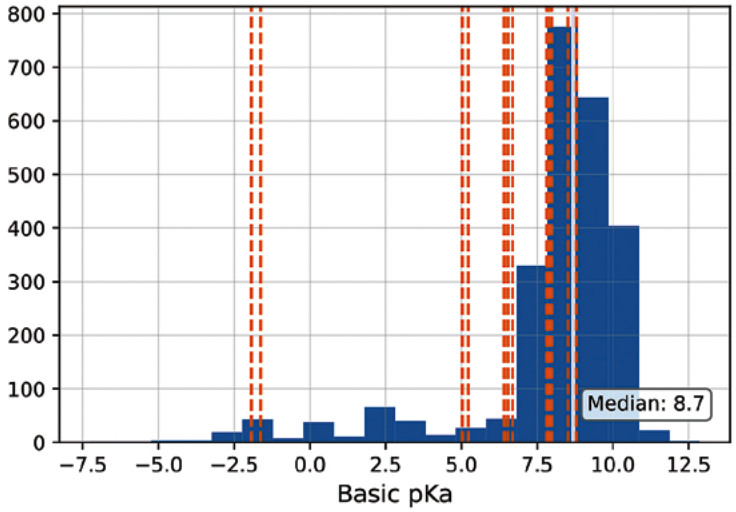

Figure 4. Histogram of basic pKa values for the 5-HT_6 receptor ligands with an affinity lower than 100 nM. Additionally, the basic pK_a values for candidates were marked (orange dashed lines).

References

1. Ferrero H, Solas M, Francis PT, Ramirez MJ. Serotonin 5-HT_6 Receptor Antagonists in Alzheimer's Disease: Therapeutic Rationale and Current Development Status. *CNS Drugs*. 2017; 31(1): 19–32. doi: 10.1007/s40263-016-0399-3.
2. Kucwaj-Brysz K, Baltrukevich H, Czarnota K, Handzlik J. Chemical update on the potential for serotonin 5-HT_6 and 5-HT_7 receptor agents in the treatment of Alzheimer's disease. *Bioorg Med Chem Lett*. 2021; 49. doi: 10.1016/j.bmcl.2021.128275.
3. Bokare AM, Bhonde M, Goel R, Nayak Y. 5-HT_6 receptor agonist and antagonist modulates ICV-STZ-induced memory impairment in rats. *Psychopharmacology (Berl)*. 2018; 235(5): 1557–1570. doi: 10.1007/s00213-018-4866-z.
4. Pereira M, Martynhak BJ, Andreatini R, Svenningsson P. 5-HT_6 receptor agonism facilitates emotional learning. *Front Pharmacol*. 2015; 6: 200. doi: 10.3389/fphar.2015.00200.
5. Doucet E, Grychowska K, Zajdel P, Bockaert J, Marin P, Bécamel C. Blockade of serotonin 5-HT_6 receptor constitutive activity alleviates cognitive deficits in a preclinical model of neurofibromatosis type 1. *Int J Mol Sci*. 2021; 22(18). doi: 10.3390/ijms221810178.
6. Geng F, Tian J, Wu JL, et al. Dorsomedial prefrontal cortex 5-HT_6 receptors regulate anxiety-like behavior. *Cogn Affect Behav Neurosci*. 2018; 18(1): 58–67. doi:10.3758/s13415-017-0552-6.
7. Amodeo DA, Oliver B, Pahua A, et al. Serotonin 6 receptor blockade reduces repetitive behavior in the BTBR mouse model of autism spectrum disorder. *Pharmacol Biochem Behav*. 2021; 200. doi: 10.1016/j.pbb.2020.173076.
8. Hagsäter SM, Lisinski A, Eriksson E. 5-HT_6 receptor antagonism reduces defecation in rat: A potential treatment strategy for irritable bowel syndrome with diarrhea. *Eur J Pharmacol*. 2019; 864. doi: 10.1016/j.ejphar.2019.172718.
9. Dudek M, Marcinkowska M, Bucki A, Olczyk A, Kołaczkowski M. Idalopirdine – a small molecule antagonist of 5-HT_6 with therapeutic potential against obesity. *Metab Brain Dis*. 2015; 30(6): 1487–1494. doi:10.1007/s11011-015-9736-3.
10. Zhu H. Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu Rev Pharmacol Toxicol*. 2020; 60(1): 573–589. doi: 10.1146/annurev-pharmtox-010919-023324.
11. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem Rev*. 2019; 119(18): 10520–10594. doi: 10.1021/acs.chemrev.8b00728.
12. Vázquez J, López M, Gibert E, Herrero E, Luque FJ. Merging Ligand--Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. *Molecules* 2020; 25(20): 4723. doi: 10.3390/molecules25204723.
13. Ji B, He X, Zhai J, Zhang Y, Man VH, Wang J. Machine learning on ligand–residue interaction profiles to significantly improve binding affinity prediction. *Brief Bioinform*. 2021; 22(5). doi: 10.1093/bib/bbab054.
14. Davies M, Nowotka M, Papadatos G, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res*. 2015; 43(W1): W612–W620. doi:10.1093/nar/gkv352.
15. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*. 2019; 47(D1): D930–D940. doi: 10.1093/nar/gky1075.
16. Mysinger MM, Carchia M, Irwin JohnJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J Med Chem*. 2012; 55(14): 6582–6594. doi10.1021/jm300687e.
17. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011; 32(7): 1466–1474. doi: 10.1002/jcc.21707.
18. Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics* 2008; 24(21): 2518–2525. doi: 10.1093/bioinformatics/btn479.
19. Moez Ali. PyCaret: An open source, low-code machine learning library in Python. Published online April 2020. Accessed August 18, 2022. https://pycaret.org/.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002; 16: 321–357. doi: 10.1613/jair.953.
21. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021; 596(7873): 583–589. doi: 10.1038/s41586-021-03819-2.
22. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022; 50(D1): D439–D444. doi: 10.1093/nar/gkab1061.
23. Schrödinger Release 2021-1: Protein Preparation Wizard. Published online 2021.
24. Nirogi RV, Konda JB, Kambhampati R, et al. N,N-Dimethyl-[9-(arylsulfonyl)-2,3,4,9-tetrahydro-1H-carbazol-3-yl]amines as novel, potent and selective 5-HT_6 receptor antagonists. *Bioorg. Med. Chem. Lett*. 2012; 22(22): 6980–6985. doi: 10.1016/j.bmcl.2012.06.002.
25. Liu KG, Robichaud AJ. 5-HT_6 antagonists as potential treatment for cognitive dysfunction. *Drug Development Researc*. 2009; 70(2): 145–168. doi: 10.1002/ddr.20293.
26. Schrödinger Release 2021-1: Induced Fit Docking. Published online 2021.
27. López-Rodríguez ML, Benhamú B, de la Fuente T, Sanz A, Pardo L, Campillo M. A Three-Dimensional Pharmacophore Model for 5-Hydroxytryptamine 6 (5-HT_6) Receptor Antagonists. *J Med Chem*. 2005; 48(13): 4216–4219. doi: 10.1021/jm050247c.
28. González-Vera JA, Medina RA, Martín-Fontecha M, et al. A new serotonin 5-HT_6 receptor antagonist with procognitive activity – Importance of a halogen bond interaction to stabilize the binding. *Sci Rep*. 2017; 7: 41293. doi:10.1038/srep41293.
29. Grychowska K, Kurczab R, Śliwa P, et al. Pyrroloquinoline scaffold-based 5-HT_6R ligands: Synthesis, quantum chemical and molecular dynamic studies, and influence of nitrogen atom position in the scaffold on affinity. *Bioorg Med Chem*. 2018; 26(12): 3588–3595. doi: 10.1016/j.bmc.2018.05.033.
30. Wichur T, Godyń J, Góral I, et al. Development and crystallography-aided SAR studies of multifunctional BuChE inhibitors and 5-HT_6R antagonists with β-amyloid anti-aggregation properties. *Eur J Med Chem*. 2021; 225: 113792. doi: 10.1016/j.ejmech.2021.113792.
31. Schrödinger Release 2021-1: Desmond Molecular Dynamics System. Published online 2021.
32. Schrödinger Release 2021-1: Glide. Published online 2021.
33. Huang S, Xu P, Shen D, et al. GPCRs steer Gi and Gs selectivity via TM5-TM6 switches as revealed by structures of serotonin receptors. *Molecular Cell*. 2022; 82(14): 2681–2695.e6. doi: 10.1016/j.molcel.2022.05.031.
34. Schrödinger Release 2021-1: LigPrep. Published online 2021.
35. RDKit: Open-source cheminformatics 2020.9.1. Published online 2020. https://www.rdkit.org.
36. ChemAxon: Instant JChem 22.6.0. Published online 2022.
37. Ballante F, Marshall GR. An Automated Strategy for Binding--Pose Selection and Docking Assessment in Structure-Based Drug Design. *J. Chem. Inf. Model*. 2016; 56(1), 54–72. doi: 10.1021/acs.jcim.5b00603.
38. Staroń J, Kurczab R, Warszycki D, et al. Virtual screening-driven discovery of dual 5-HT6/5-HT2A receptor ligands with pro--cognitive properties. *Eur J Med Chem*. 2020; 185: 111857. doi: 10.1016/j.ejmech.2019.111857.
39. Ma S, McGann M, Enyedy IJ. The influence of calculated physicochemical properties of compounds on their ADMET profiles. *Bioorg Med Chem Lett*. 2021; 36: 127825. doi: 10.1016/j.bmcl.2021.127825.
40. Rankovic Z. CNS Drug Design: Balancing Physicochemical Properties for Optimal Brain Exposure. *J Med Chem*. 2015; 58(6): 2584–2608. doi: 10.1021/jm501535r.