Multi-model *in silico* characterization of 3-benzamidobenzoic acid derivatives as partial agonists of Farnesoid X receptor in the management of NAFLD

Soumya Mitra [a], Amit Kumar Halder [a,c], Nilanjan Ghosh [b], Subhash C. Mandal [b], M. Natália D.S. Cordeiro [c,*]

[a] *Dr. B.C. Roy College of Pharmacy & Allied Health Sciences, Durgapur, 713206, India*
[b] *Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India*
[c] *LAQV@REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007, Porto, Portugal*

ARTICLE INFO

ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) is a pathological condition which is strongly correlated with fat accumulation in the liver that has become a major health hazard globally. So far, limited treatment options are available for the management of NAFLD and partial agonism of Farnesoid X receptor (FXR) has proven to be one of the most promising strategies for treatment of NAFLD. In present work, a range of validated predictive cheminformatics and molecular modeling studies were performed with a series of 3-benzamidobenzoic acid derivatives in order to recognize their structural requirements for possessing higher potency towards FXR. 2D-QSAR models were able to extract the most significant structural attributes determining the higher activity towards the receptor. Ligand-based pharmacophore model was created with a novel and less-explored open access tool named *QPhAR* to acquire information regarding important 3D-pharmacophoric features that lead to higher agonistic potential towards the FXR. The alignment of the dataset compounds based on pharmacophore mapping led to 3D-QSAR models that pointed out the most crucial steric and electrostatic influence. Molecular dynamics (MD) simulation performed with the most potent and the least potent derivatives of the current dataset helped us to understand how to link the structural interpretations obtained from 2D-QSAR, 3D-QSAR and pharmacophore models with the involvement of specific amino acid residues in the FXR protein. The current study revealed that hydrogen bond interactions with carboxylate group of the ligands play an important role in the ligand receptor binding but higher stabilization of different helices close to the binding site of FXR (e.g., H5, H6 and H8) through aromatic scaffolds of the ligands should lead to higher activity for these ligands. The present work affords important guidelines towards designing novel FXR partial agonists for new therapeutic options in the management of NAFLD. Moreover, we relied mainly on open-access tools to develop the *in-silico* models in order to ensure their reproducibility as well as utilization.

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) has become one the most common manifestations of chronic liver diseases in the last decade or so in the global population. The concern regarding the disease has multiplied many folds due to its disruption of different extra-hepatic systems and not only confining itself to metabolic disorders [1,2]. Non-alcoholic fatty liver is a pathological condition which is strongly correlated with fat accumulation in the liver with presence of >5% steatosis of

hepatocytes and has been well established as one of the major reasons for morbidity and mortality in 1/4th of the global adult population [3, 4]. Disease progression of NAFLD has been perceived in two stages: firstly the benign non-alcoholic fatty liver and then its more aggressive form, namely non-alcoholic steatohepatitis (NASH) [5]. NASH is considered to be the progressed stage in fatty liver disease, which is characterized by excessive steatosis, hepatocytic enlargement with rarefied cytoplasm (hepatocytic ballooning) and fibrosis [6,7]. NASH is capable of advancing to stages of hepatic diseases, namely cirrhosis of

liver and hepatocellular carcinoma that has shown high prevalence in population with other complicated metabolic disorders [8]. The underlying etiology behind development of NAFLD is associated with obesity and insulin resistance which make it a close accomplice to type 2 diabetes mellitus (T2DM) among others. The intertwining relation between T2DM and NAFLD is often complex as T2DM can severely accelerate the disease complications like cirrhosis, hepatocellular carcinoma of NASH [9,10].

The underlying pathophysiology of the disease is often associated with uncontrolled lipid supply from gut to liver due to uninhibited dietary intake or dysregulated microbiota of the gut, escalated free fatty acid influx from the white adipose tissue and elevated lipogenesis [11]. Patients with early stages of NAFLD are often recommended lifestyle modifications involved with calorie restriction and increased physical activity. Loss in weight may help individuals with early stage complications, but targeted therapeutic approach may be useful against NAFLD in advanced stages. The major therapeutic approaches include use of peroxisome proliferator-activator receptors (PPARs) agonists, lipid lowering agents, incretin-based therapies, agents acting on inflammation, antioxidants and farnesoid X receptor (FXR) agonists [12].

Farnesoid X receptor (FXR) is a nuclear receptor transcription factor which plays an important role in maintaining bile acid homeostasis in the liver by downregulation of the bile acid synthesis and reduced gluconeogenesis as well as lipogenesis [13]. FXR agonists have become one of the most sought-after therapeutic approaches in management of NAFLD/NASH after its synthetic agonist Obeticholic acid (OCA) had produced promising results in clinical investigations [14]. OCA is the synthetic analogue of the naturally occurring FXR agonist chenodeoxycholic acid (CDCA), and it has been identified as a lead compound due to its effect on improving insulin resistance in T2DM and its anti-fibrotic potential [15,16]. The effects exerted by FXR agonism have been derived from its multifaceted effects in genetic regulation which are predominant in bile acid axis for regulation of metabolic disorders. However, in a recently completed study, the effect of OCA in the treatment of NAFLD through activation of FXR has come under radar due to its adverse effect of causing hepatic injury by cholesterol accumulation and interleukin - 1β pathway [17]. Overactivation of FXR receptors has proven to diminish CYP7A1 gene expression, which is responsible for the synthesis of cholesterol 7α-hydroxylase, an enzyme, effective in activating cholesterol metabolism and synthesis of bile acids [18–20]. In fact, the United States Food and Drug Administration (USFDA) recently restricted the use of OCA in patients suffering from primary biliary cholangitis and advanced cirrhosis due to the fact that its administration may lead to severe hepatotoxicity and liver failure [21]. This has led researchers to look into partial FXR agonists as potential therapeutic agents in the management of NAFLD/NASH.

A recent investigation by Hu et al. on novel 3-benzamidobenzoic acid derivatives as partial FXR agonists has established potential of the synthesized molecules in combatting NAFLD/NASH [22]. The compounds were synthesized on the basis of their interaction between the phenoxyphenyl moiety in the large hydrophobic cavity and potency of the synthesized compounds was expressed in 50% effective concentration (EC_{50}) by the Lanthascreen™ Time-Resolved Fluorescence resonance energy transfer (TR-FRET) assay and indicated activity range between 5.5 and 7.6 μM. The highest active compound was subjected to *in vivo* assay in acetaminophen (APAP) induced hepatotoxic animals and showed satisfactory results. The structure activity relationship (SAR) study conducted by Hu et al. also revealed the effect of biphenyl moiety in binding to the hydrophobic pocket of the receptor [22]. A further *in silico* study needs to be carried out to interpret the effects of different structural features on their activity. In the present work, we employed a range of *in silico* modeling tools to characterize the biological activity variations observed in these recently developed 3-benzamidobenzoic acid derivatives in a bid to facilitate further development of FXR partial agonists in near future. Starting with a 2D-quantitative structure activity relationship (2D-QSAR) modeling analysis, the latter was

followed by pharmacophore mapping, 3D-QSAR modeling, molecular docking and molecular dynamics (MD) simulation analyses.

2. Materials and methods

2.1. Dataset collection and preparation

In a recently published study, Hu et al. reported 35 novel 3-benzamidobenzoic acid derivatives as partial FXR agonists [22]. The activity of the compounds was checked by TR-FRET assay for their potential in the treatment of NAFLD/NASH. The 50% effective concentration (EC_{50}) of the synthesized compounds were collected and converted to pEC_{50} ($= -\log_{10}(EC_{50}/10^6)$) and subsequently, used as response variables for development of the 2D and 3D QSAR models. The structures and biological activity of the 35 compounds are supplied in the supplementary data (Table S1). The SMILES provided by the group were converted to .sdf files by Discovery Studio Visualization tool and then numbered accordingly. Standardization of the structures were performed by Chemaxon Standardizer tool through the following steps: (a) addition of explicit hydrogen atoms, (b) aromatization, (c) clean 2D, (d) clean 3D, (e) neutralize and (f) strip salts. After standardization of the structures, these were submitted to 2D and 3D QSAR modeling analysis efforts [23].

2.2. 2D-QSAR modeling

2.2.1. Descriptor calculation

Descriptor calculation of the thirty-five 3-benzamidobenzoic acid derivatives was carried out by the OCHEM webserver-based AlvaDesc v.2.0.4 (https://www.alvascience.com/alvadesc/) module [24,25]. The Corina tool was employed for geometry optimization of the 3D structures of the compounds [26]. The pEC_{50} values of the compounds were incorporated with the calculated descriptors to obtain the final dataset for 2D-QSAR model development.

2.2.2. Dataset division and model development

The entire dataset was divided into a training set and a test set by the help of Python coded open-access tool SFS-QSAR (https://github.com/ncordeirfcup/SFS-QSAR-tool) and the dataset division *per se* was made with the activity sorting method taking 2 as the starting point [27]. The 2D QSAR modeling was completed in two stages. In the first stage, only descriptors with better interpretability were utilized for the development of the model. The latter included constitutional descriptors, functional group counts, 2D-atom pairs, drug-like indices, ring descriptors, atom-centered fragments, pharmacophore descriptors and molecular properties. In the second stage, all AlvaDesc descriptors were deployed for the model development to check how much improvement in the model quality was achieved as compared to the model generated with interpretable descriptors. The models were developed by applying two feature selection techniques, namely: (a) the sequential forward selection (SFS) and (b) the genetic algorithm (GA). The SFS based feature selection method resources to the open-access tool SFS-QSAR (https://github.com/ncordeirfcup/SFS-QSAR-tool), which utilizes the 'Feature Selector' module of the Mlxtend library (http://rasbt.github.io/mlxtend/) [25]. The variance cut-off was set at 0.0001 to remove constant and near-constant descriptors while the correlation cut-off was kept at 0.99 during data treatment. The determination coefficient (R^2), the negative mean absolute error (*NMAE*), the negative mean Poisson deviance (*NMPD*) and the negative mean gamma deviance (*NMGD*) were the four scoring functions employed for the sequential forward selection (SFS) module during model development with no cross-validation as well as 5-fold cross-validation. Therefore, eight models ($= 4 \times 2$) were produced from SFS-QSAR. On the other hand, the genetic algorithm (GA), which follows a stochastic algorithm unlike SFS, was employed to generate randomized 2D QSAR models with the help of the open access tool GeneticAlgorithm v.4.1_2 (accessed from https://dtclab.webs.com/software-tools) [28]. GA utilizes techniques including

cross-validation and mutation to improve the fitting of the independent variables against the dependent variable. Nevertheless, since the model generation in GA-MLR largely depends on randomness, 25 runs were performed with the training set and the best GA-MLR model was chosen with respect to its internal and external predictivities. For both SFS-MLR and GA-MLR, the maximum number of descriptors was fixed at 5.

2.2.3. Statistical evaluation of the models

Statistical parameters such as the determination coefficient R^2, adjusted R^2 (R^2_{Adj}), F-statistics, mean absolute error (MAE) and internal cross-validation parameter (Q^2_{LOO}) were used for validation of the developed 2D QSAR models [29]. External predictivity of the developed models was evaluated by the external validation parameter R^2_{Pred} [30]. Along with these, various r^2_m metrics including r^2_{mLOO} and Δr^2_{mLOO} were utilized as internal validation parameters while external validation parameters like r^2_{mtest} and Δr^2_{mtest} were used [31]. Inter-collinearity between the descriptors of the best generated models was evaluated from the cross-correlation matrices generated from the descriptors. On the other hand, the multi-collinearity of the final models was estimated by calculating the variation inflation factor (VIF $= 1/1 - R^2_i$, where R^2_i is the determination coefficient (R^2) determined by regressing the ith descriptor on the other descriptors) [32]. Additionally, 1000 runs of the Y-randomization test were carried out to generate models with randomized response variables. The $^cR^2_p$ value of the original model was calculated to confirm that the models were not developed by chance [33].

2.2.4. Applicability domain of the models

The Williams plot, a plot between leverage values and standardized coefficients, was utilized for assessing the applicability domain of the models. When the leverage value of any data-point is greater than the *hat* value h^* ($h^* = 3p/n$, where p stands for the number descriptors in the model $+ 1$ and n for the number of data-points in the training set), it is alleged to be a structural outlier. If the standardized residual of any particular data-point is $> \pm 2.5$, it is considered a response outlier [34].

2.3. Ligand-based pharmacophore mapping

2.3.1. Conformation generation technique

For pharmacophore modeling, 100 conformations were generated for each dataset compound using RDKit (https://www.rdkit.org/). More specifically, the EmbedMultipleConfs program of the AllChem module of RDKit was used to generate these conformations by choosing the energy cut-off (conformers with energy difference from the lowest one) as 100 and the RMS cut-off was selected as 0.5 (*i.e.*, only conformers with RMS higher than the threshold were kept).

2.3.2. Model development

In the current work, we explored a recently developed open-access tool named *QPhAR* or Quantitative Pharmacophore to develop pharmacophore models as well as to align the compounds for 3D-QSAR mapping. The methodology of *QPhAR*-based model development has been described in detail in the original article published by Kohlbacher et al. [35]. Briefly in this method, a training library with multiple conformations of each compound is first selected. From this library, a template structure is selected and in the current work the most rigid compound of the training set was selected as a template using *QPhAR*. The template is then used to align the remaining training set compounds by pharmacophore alignment in which the greedy optimization technique is adopted to choose the best fitting conformations. It is to be noted here that, during this alignment, each pharmacophore is associated with the activity value of the parent pharmacophore. This alignment results in a merged pharmacophore like outcome (named as pharmacophore feature container) that is then clustered by the minimum Euclidean distance hierarchical clustering algorithm. For

clustering, the default cut-off value of 1.5 Å was chosen for the radius of pharmacophoric feature.

Two post-processing steps followed after clustering to obtain the most relevant pharmacophoric features. Initially, representative features were chosen for each cluster and secondly, the compounds with non-conclusive activity data were discarded. Subsequently, input vectors were generated for each training sample with the help of the training set and the reference merged pharmacophore. Weights were then added to input vectors where the features having higher number of merged features are weighted heavier than features with fewer data. After generating the input features for each aligned training sample, the random forest (RF) regression technique was employed to produce statistical results for the training set. In this work, we firstly divided the data into 21 training set and 14 test set compounds randomly using the *splitData.py* program of *QPhAR*. Important to mention here is that the test set, which comprises 40% of the total dataset, was not used for generating the pharmacophore model but used solely for external validation of the generated model. The *train.py* program of *QPhAR* was employed for generating the quantitative pharmacophore and after obtaining satisfactory results for the training set, the *predict.py* program of the same tool was used for estimating the predictivity towards the 14 test set compounds [35]. In order to obtain the aligned poses of the dataset compounds with the pharmacophore, we used the *profile3-DActivity.py* program of the *QPhAR-applications tool* (https://github.com/StefanKohlbacher/qphar-applications), an extension of *QPhAR* recently released by Kohlbacher et al. [36]. All conformations of input dataset structures were submitted to this program in .sdf format. The aligned structures were subsequently utilized for 3D-QSAR modeling. The statistical results of the ligand-based pharmacophore models were checked using the statistical metrics R^2, Q^2, root mean square error (RMSE), mean error (ME), standard error (SE) and R^2_{Pred}.

2.4. 3D-QSAR modeling

2.4.1. Alignment methods

In the current work, two methods have been utilized for alignment of the structures, namely: (a) *QPhAR*-based pharmacophore alignment, and (b) constrained docking-based alignment. The pharmacophore-based alignment was described in previous section The molecular docking was performed with the Autodock Vina software for alignment purposes (described later).

2.4.2. Model development

3D-QSAR model development was carried out with the Open3-DQSAR which is an open-source software [37,38]. Details about this methodology for model development can be found in our previous works [39,40]. Briefly, the compounds in the dataset are probed with a carbon and volume-less positively ($+1$) charged probe for calculation of the steric as well as electrostatic fields. A smart region definition (SRD) cut-off level of 2.0 was set and removal of N-level variables were performed for field pre-treatment. The software utilizes SRD for variable grouping by assessing the closeness of the variable in 3D space along with two different variable selection algorithms, i.e.: (a) Fractional Factorial Design-based variable SELection (FFD-SEL), and (b) Uninformative Variable Elimination-based Partial Least Square (UVE-PLS). The quality of the developed 3D-QSAR-based PLS models was assessed by R^2 with the standardized error of calibration (SDEC) values, F-test results, Q^2_{LOO}, leave-two-out Q^2 (Q^2_{LTO}), leave-many-out Q^2 (Q^2_{LMO}, with groups $= 5$, runs $= 20$), as well as by R^2_{Pred} along with the associated standardized error of prediction (SDEP). Progressive scrambling operations were employed to check whether the model was unique and not created by chance. The scrambling operations were deployed for selected models using the following criteria: critical value: 0.80, type: LMO groups $= 5$, runs $= 20$ and scrambling $= 20$. The Q^2_S in this work denotes the output of the scrambling test and depicts 'fitted q^2 values'. Robustness of the generated 3D-QSAR model was validated by its low value

when compared with Q_{LMO}^2. The contour maps were generated with isocontour values at PLS coefficients of $+0.005$ (green) and -0.005 (yellow) for the steric field and $+0.003$ (blue) and -0.003 (red) for the electrostatic field [40].

2.5. Molecular docking

Molecular docking methods were used to serve two purposes. First, docking based alignment was used for 3D-QSAR analyses as described above. For this, each dataset compound was docked at the binding site of FXR protein (pdb: 6HL1) [41] using the Autodock Vina software [42]. Second, docking of the most potent and the least potent compounds with the FXR protein was performed with Autodock 4.2 [43]. For both Autodock 4.2 and Autodock Vina, a grid box was generated with a center having coordinates $X = 12.86$ Å, $Y = -13.62$ Å and $Z = 12.25$ Å with extensions of $40 \times 40 \times 40$ Å^3. The amino acids present in the grid are listed in Table S2. Note that the amino acid residues were renumbered in the current work and these new residue numbers are also provided in that table. The protein pdbqt file was prepared stepwise by removing the water molecules and bound heteroatoms, by adding hydrogen atoms and by adding Gasteiger–Marsili partial atomic charges. Nine and ten docked poses were generated from Autodock Vina and Autodock 4.2, respectively. The default parameters were applied while performing the molecular docking with Autodock 4.2.

2.6. Molecular dynamics simulations

MD simulations of 50 ns with the docked poses as well as the apo-protein were carried out with Amber 12 following the procedure described earlier [40,44]. The protein structures were protonated at pH 7.0 using the PDB2PQR webserver (https://server.poissonboltzmann.org/pdb2pqr) [45]. Trajectory analyses were performed using the PTTRAJ and CPPTRAJ software codes to obtain plots for the root mean square deviation (RMSD) and radius of gyration (RG) of proteins and ligands along with the root mean square fluctuation (RMSF) for the backbone atoms using the QtGrace software (https://sourceforge.net/projects/qtgrace/). Additionally, trajectory analyses were also employed to determine the hydrogen bonding between the ligands and the binding site amino acid residues of the FXR protein. The energy contributions of the binding site amino acid residues were computed using the per residue free energy decomposition method with help of the Amber MM-GBSA module [40,46]. All energy components (van der Waals, electrostatic, polar solvation, and nonpolar solvation contributions) were calculated using 200 snapshots extracted from the last 10 ns MD trajectories.

3. Results and discussions

3.1. 2D-QSAR modeling

Following the strategy outlined before, and keeping in mind the importance of mechanistic interpretation of 2D-QSAR models, we began by seeking the best linear models based on (a) selected interpretable

descriptors and on (b) all descriptors from the AlvaDesc descriptors [40]. These two sets of descriptors were separately employed for setting up the 2D-QSAR models in order to check which set provides a more predictive model and how much difference in the statistical quality may exist among the most developed predictive models. Two different feature selection methods namely SFS and GA were also separately employed to develop the linear 2D-QSAR models after splitting the dataset into a training set and test set. A summary of the statistical results of the obtained models is provided in Table 1, in which the two parameters Q_{LOO}^2 and R_{Pred}^2 were initially considered for assessing their internal and external predictivity, respectively.

As can be seen, GA-MLR provided statistically more predictive models (Model-1 and Model-2) as compared to the models generated with SFS-MLR. Undoubtedly, a statistically more predictive GA-MLR model is retrieved when all descriptors are employed instead of selected interpretable descriptors. With all descriptors, the GA-MLR model (Model-2) secured Q_{LOO}^2 and R_{Pred}^2 values of 0.862 and 0.957, respectively. Nevertheless, the GA-MLR model generated with selected interpretable descriptors (Model-1) also produced a statistically reliable 2D-QSAR model, the overall predictivity of which (i.e., Q_{LOO}^2: 0.831 and R_{Pred}^2: 0.944) is not significantly poorer than Model-1. Moreover, Model-2 can provide greater insight into the structural requirement of these compounds for possessing higher activity that further improves its significance as a 2D-QSAR model. These two GA-MLR models are presented in Table 2 and, as it can be checked from various statistical parameters, both these models provide satisfactory statistical significance.

The observed vs. predicted activity plots of these models are shown in Supplementary Fig. S1. From Table 2, it is evident that both models

Table 2
Detailed description of the GA-MLR models developed with interpretable descriptors (Model-1) and all descriptors (Model-2).

Model	Equation	Statistical results[a]
Interpretable Descriptors (Model-1)	pEC$_{50}$ = +4.963($\pm$0.182) +1.085 ($\pm$0.080) CATS2D_08_NL+0.304 ($\pm$0.063) F08[C–N]–0.033 ($\pm$0.018) H-046 + 0.406($\pm$0.111) Depressant-80 –0.136 ($\pm$0.030) F08[C–O]	$N_{training}$ = 28; R^2 = 0.896; R_{Adj}^2 = 0.872; F (22;5) = 37.844; Q_{LOO}^2 = 0.832; MAE$_{LOO}$ = 0.127; r_{mLOO}^2 = 0.772; Δr_{mLOO}^2 = 0.060; N_{test} = 8; R_{Pred}^2 = 0.945; r_{mtest}^2 = 0.930; Δr_{mtest}^2 = 0.029, $^c R_p^2$ = 0.809
All descriptors (Model-2)	pEC$_{50}$ = +22.741($\pm$1.635)+6.328 ($\pm$0.744) WHALES20_Isol–0.146 ($\pm$0.015) TPSA(NO)+0.065 ($\pm$0.067) CATS2D_08_AA–0.266 ($\pm$0.085) Mor04v–7.233 ($\pm$1.763) E1i	$N_{training}$ = 28; R^2 = 0.911; R_A^2 = 0.891;F (22;5) = 45.204; Q_{LOO}^2 = 0.863;MAE$_{LOO}$ = 0.123; r_{mLOO}^2 = 0.811; Δr_{mLOO}^2 = 0.060; N_{test} = 8; R_{Pred}^2 = 0.958; r_{mtest}^2 = 0.877;Δr_{mtest}^2 = 0.034, $^c R_p^2$ = 0.819

[a] $N_{training}$: Number of data-points present in the training set; N_{test}: Number of data-points present in the test set.

Table 1
Summary of the statistical results obtained from SFS-MLR and GA-MLR models.

Method	Score	CV	All descriptors			Interpretable descriptors		
			Q_{LOO}^2	R_{Pred}^2	Average	Q_{LOO}^2	R_{Pred}^2	Average
GA-MLR	na	na	**0.862**	**0.957**	**0.910**	**0.831**	**0.944**	**0.888**
SFS-MLR	NMAE	0-fold	0.939	0.876	0.908	0.808	0.411	0.609
SFS-MLR	NMAE	5-fold	0.929	0.840	0.884	0.819	0.619	0.719
SFS-MLR	NMGD	0-fold	0.852	0.504	0.678	0.768	0.423	0.595
SFS-MLR	NMGD	5-fold	0.779	0.162	0.470	0.803	0.135	0.469
SFS-MLR	NMPD	0-fold	0.879	0.274	0.577	0.768	0.423	0.595
SFS-MLR	NMPD	5-fold	0.827	0.874	0.850	0.803	0.135	0.469
SFS-MLR	R^2	0-fold	0.878	0.860	0.869	0.812	0.580	0.696
SFS-MLR	R^2	5-fold	0.907	0.688	0.798	0.819	0.619	0.719

were generated with satisfactory statistical predictivity. For example, the r^2_{mOO} and Δr^2_{mOO} values of Model-1 are 0.772 and 0.060, respectively, whereas for Model-2, these values were found to be 0.811 and 0.060, respectively. These values were well above (for r^2_{mOO} cut-off value > 0.5) or below (for Δr^2_{mOO} cut-off value < 0.2) their cut-off values for having satisfactory internal predictivity. The external predictivity of these models were even more satisfactory as r^2_{mtest} of Model-1 and Model-2 were found to be 0.930 and 0.877, respectively, whereas the $\Delta r^2_{\text{mtest}}$ of these models are 0.029 and 0.034, respectively. Next, we checked the correlation matrix of the independent parameters of these two models, which revealed that no high intercorrelation does exist. Simultaneously, multi-collinearity of these models was also checked with the variation inflation factor (VIF) values. Model-1 and Model-2 have VIF values ranging from 1.16 to 1.76 and 1.18−1.50, respectively. These values clearly indicate that the multi-collinearity does not exist in these models since these values lie well below the cut-off value of 5. High $^cR^2_p$ values were also obtained from the *Y*-randomization tests for both models, indicating that these models were not developed by chance.

Furthermore, it is absolutely necessary to look at the applicability domain (AD) of the established 2D-QSAR models. The Williams plots generated for Model-1 and Model-2 are presented in Fig. 1. One training set compound of Model-2 was found to be a structural outlier but it was predicted satisfactorily by the model and therefore it was retained. By contrast, no structural outlier was detected in Model-1 but each of these models contained one response outlier.

The relative significance of the descriptors of Model-1 and Model-2 are shown in Fig. 2. As expected, the models developed with interpretable descriptors (Model-1) provide further insight into the structural requirements of 3-benzamidobenzoic acid derivatives. At the same time, the significance of different descriptors of Model-1 varied consistently whereas for Model-2, two descriptors have high significance while the other three descriptors have almost negligible significance.

Therefore, we first describe and discuss the descriptors of Model-1. The CATS2D_08_NL was found to be the most significant descriptor of Model-1 with positive correlation with the response variable. This descriptor belongs to the category of 2D Chemically Advanced Template Search (CATS) descriptors that are highly useful descriptors to understand the required 2D pharmacophoric features in the compounds and at the same time, these descriptors also specify the required topological distance between the required features [47]. For example, CATS2D_08_NL indicates the presence of negative ionizable and lipophilic features separated at the topological distance of 8. Most of the potent derivatives showed higher values (i.e., 2) for this parameter whereas the lower active compounds had lower values (i.e., 0 or 1), as shown in Fig. 3.

The second most significant descriptor of Model-1 is Depressant-80, which is a drug-like index representing the Ghose-Viswanadhan-Wendoloski antidepressant-like index at 80% [48]. This drug-like index is originated from studies of the physicochemical properties of a specific group of drugs, antidepressants in this case with a qualifying range of 80%. The physicochemical properties that are considered for calculating such descriptors include lipophilicity (ALogP, range: 1.4−4.9), molar refractivity (AMR, range: 62−114), molecular weight (MW, range: 210−380) and number of atoms (nAT, range: 32−56). The positive correlation of this descriptor with the biological activity indicates that higher values of this descriptor favors higher biological activity.

The third most influential descriptor of Model-1 is F08[C–N], which is basically a 2D atom pair descriptor representing the frequency of carbon and nitrogen atoms located at topological distance of 8 [49]. Again, a positive correlation with the descriptor indicates that higher values for this descriptor should lead to higher biological activity. For example, one of the most potent derivatives SN04 (pEC$_{50}$: 7.01) has F08[C–N] value of 2 while one of the least potent derivatives SN31 has for this descriptor a value of 0 (Fig. 4).

The last two descriptors of the models are F08[C–O] and H-046, both having negative correlation with the FXR partial agonistic activity of 3-benzamidobenzoic acid derivatives. Among these, F08[C–O] possesses higher significance and similar to F08[C–N], this descriptor is a 2D atom-pair descriptor but the former stands for carbon and oxygen atoms located at topological distance of 8 [49]. Naturally, higher values for this descriptor were found mostly in the compounds with lower partial agonistic activity. One example is shown in Fig. 4 with the most potent (SN01) and the least potent (SN35) compounds of the dataset. The differences in the values of F08[C–O] stems from the fact that n-butoxy side chain in SN35 gives rise to an additional center for oxygen leading to higher value of F08[C–O] that results in its lower activity. The same argument is valid for one of the most potent derivative SN04 (structure shown in Fig. 4) with high F08[C–O] value of 8 and it explains why compounds with biphenyl scaffolds were found to be more potent as FXR partial agonists when compared to those containing the phenoxyphenyl scaffold.

The atom-centered fragment descriptor H-046, in contrast, represents a specific fragment in which the hydrogen atom is attached to a sp^3 hybridized carbon (C0) whereas no heteroatom (X) is attached to the next carbon. Therefore, this atom centered fragment descriptor indicates the presence of terminal alkyl groups without heteroatom. The presence of such fragment was found to diminish activity. For example, the value of this descriptor was found to be 0 in the most potent compound SN01 whereas the least active compound SN35 has a value of 7, which clearly demonstrates the unfavorable influence of long aliphatic chains for greater biological activity towards FXR.

Model-2, being the most statistically significant 2D-QSAR model, comprises five descriptors among which WHALES20_Isol and E1i possess the maximum significance. Details about these five descriptors are provided in Table 3. Both WHALES20_Isol and E1i are 3D descriptors
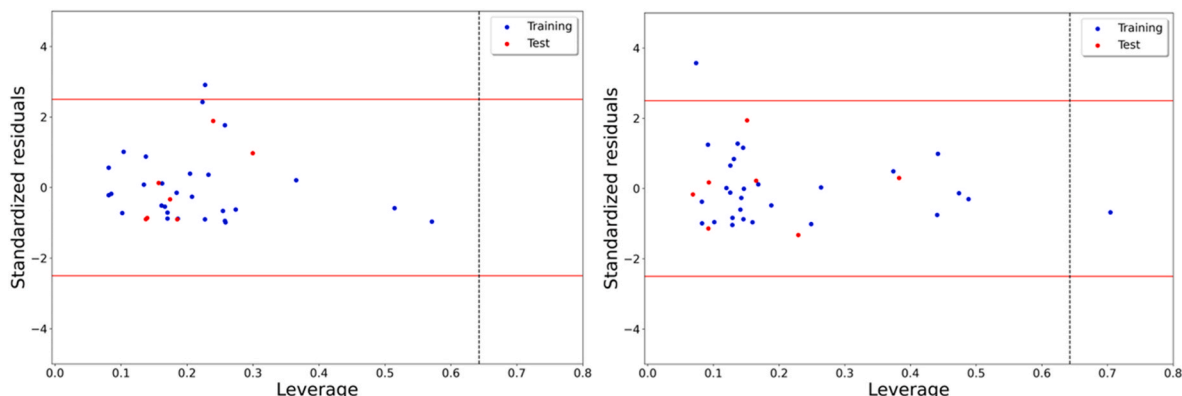

Fig. 1. Williams plots for the 2D-QSAR models, Model-1 (left) and Model-2 (right).
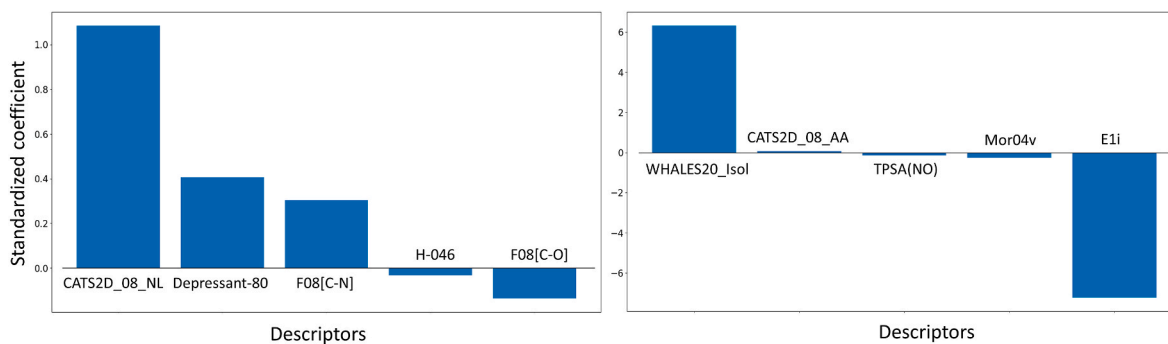
Fig. 2. Relative importance of the descriptors of Model-1 (left) and Model-2 (right).
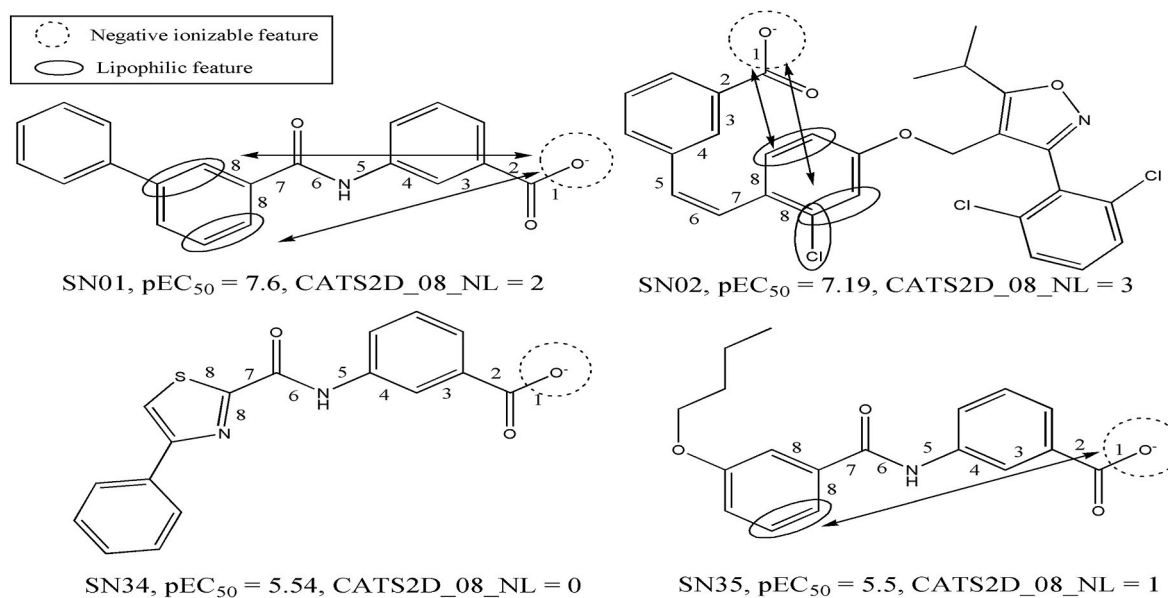

Fig. 3. Importance of the CATS2D_08_NL descriptor with respect to selected dataset compounds.
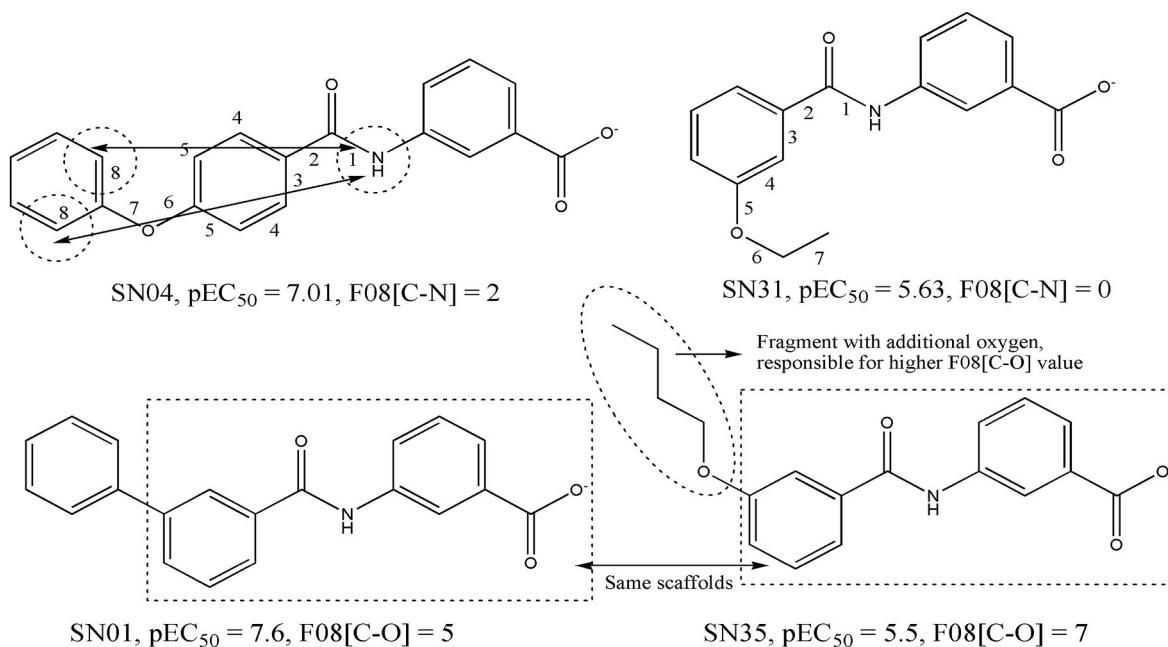

Fig. 4. Importance of the F08[C–N] and F08[C–O] descriptors with respect to selected dataset compounds.

Table 3
Description of the independent variables appearing in Model-2.

Descriptor	Category	Description
E1i	WHIM (3D)	1st component accessibility directional WHIM index/weighted by ionization potentials
WHALES20_Isol	WHALES (3D)	WHALES Isolation degree (Isol) (percentile 20)
Mor04v	MoRSE (3D)	Signal 04/weighted by van der Waals volumes
CATS2D_08_AA	CATS2D	CATS2D Acceptor-Acceptor at lag 08
TPSA(NO)	Molecular properties	Topological polar surface area with N, O polar contributions

and thus their values are highly dependent on the specific 3D conformation of the compounds and clearly related to their shape. As *per* the relative significance, the Weighted Holistic Invariant Molecular (WHIM) descriptor E1i [50] has the maximum significance and is negatively correlated to the response variable. This descriptor stands for relevant 3D information regarding size, shape, symmetry, and atom distribution of molecular structures. At the same time, E1i implies that a high ionization potential of the compound is detrimental to its partial agonistic activity towards FXR.

The WHALES20_Isol, on the other hand, is positively correlated with the biological activity, that is, higher numerical values for this descriptor were found in the more potent derivatives. This descriptor belongs to the category of Weighted Holistic Atom Localization and Entity Shape (WHALES) descriptors, which are comparatively novel descriptors developed to transform structural and pharmacophore information of known biologically active phytochemicals to synthetically accessible iso-functional compounds with the help of similarity-driven approaches [51]. The 3D Morse descriptor Mor04v is originated from the equations used in electron diffraction studies. Mor04v is weighed by the van der Waals volume and at the same time, this descriptor is negatively correlated with the biological activity. As such, that means that along with ionization potentials, higher van der Waals volumes can be detrimental to the biological activity. The last two descriptors of the models are CATS2D_08_AA and TPSA(NO), which are more interpretable than the other three descriptors of the model even though with relatively less significance. Interestingly however is that the presence of CATS2D_08_AA again means that a topological distance of 8 is highly crucial in these compounds and particularly the existence of two hydrogen bond acceptor features located at such topological distance. The final descriptor of the model is TPSA(NO), which reveals the topological polar surface area with polar contributions for nitrogen and oxygen atoms. A negative correlation is observed between TPSA(NO) values and the pEC_{50} values, indicating that lower values for this

descriptor are likely to provide more potency in the compounds as partial agonist of FXR. Two descriptors related to polarity (i.e., E1i and TPSA(NO)) indicate that higher van der Waals interactions and lower polar interactions determine the biological activity of these compounds.

3.2. Ligand-based pharmacophore mapping

The quantitative ligand-based pharmacophore was generated with the most rigid structure automatically selected by the *QPhAR* program, i. e.: SN17. The template pharmacophore aligned with the compound SN17 is shown in Fig. 5 along with the pharmacophore container and final quantitative pharmacophore (i.e., hpmodel), which contained ten pharmacophore features. The pharmacophore features of the hpmodel consist of four aromatic (R), three hydrophobic (H), one donor (D), one acceptor (A) and one negative ionizable (N) characteristics [35,36]. It must be noted that, even though ten features were obtained in the model, it does not mean that a compound needs to bind to all these features of the hpmodel. Rather, these features have their own weightages, and alignment of pharmacophore of a specific compound with different features of the hpmodel gives rise to its predicted activity that is required to be compared with its experimental activity.

The statistics of the quantitative pharmacophore mapping is summarized in Table 4. The model developed with 21 training set compounds afforded a highly predictive model with R^2 of 0.933 and Q^2 of 0.901. However, the true predictivity of the model was tested with the test set containing 14 compounds. Initially, an R^2_{Pred} value of 0.437 was obtained which was below the cut-off value of 0.50. However, it was noticed, that only one compound (SN22) was predicted poorly by the model and removal of it resulted in an improved R^2_{Pred} value of 0.719. At the beginning, we thought that the poor prediction (observed pEC_{50} = 5.77, predicted pEC_{50} = 6.75) was due to the conformers of this

Table 4
Statistical results for the *QPhAR* based quantitative pharmacophore model.

Parameter	Training	Test
N	21	14
R^2	0.933	–
Q^2	0.901	–
RMSE	0.193	–
ME	0.130	–
SE	0.143	–
R^2_{Pred}	–	0.437
R^2_{Pred}[a]	–	0.719

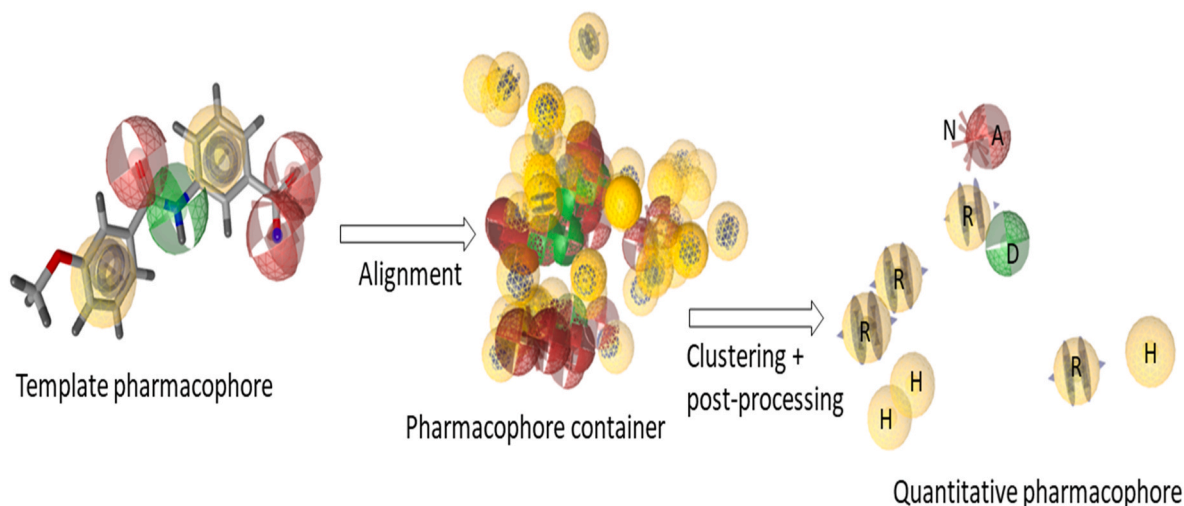[a] Obtained after removal of SN22 from the test set.

Fig. 5. Flow of quantitative pharmacophore (or hpmodel) generation in *QPhAR* with the training set from the present data.

compound. Therefore, we generated additional 50 conformations for this compound with the genetic algorithm technique of the Openbabel tool [52]. However, the predictivity failed to improve.

In spite of having poor prediction for SN22, it may not be denied that the pharmacophore mapping predicted other compounds of the dataset considerably well, and it leads us to the understanding of the structural requirements of the compounds for higher partial agonistic activity towards the FXR protein. In Fig. 6, the pharmacophore-aligned poses of three dataset compounds with varying biological activities are shown. The most potent compound SN01 (pEC$_{50}$ = 7.600, predicted pEC$_{50}$ = 7.096) was mapped with five pharmacophore features (2R+1D+1A+1 N) whereas the least potent compound SN35 (observed pEC$_{50}$ = 5.500, predicted pEC$_{50}$ = 5.579) mapped with four features (2H+1A+1 N). It is therefore clearly understood that the weightages of the aromatic features (R) are higher as compared to the hydrophobic (H) features. Another compound, SN30 with pEC$_{50}$ of 6.690 was mapped with four features (2R+1A+1 N). Noticeably, the terminal benzene of biphenyl scaffold was mapped in SN01 whereas the phenyl group attached to carboxamide was mapped in SN35. It suggests the greater influence of the biphenyl scaffold towards higher biological activity as compared to the phenoxyphenyl scaffold. Additionally, the donor feature (D) mapped with the amino group of SN01 was also found to have high significance as far as biological activity is concerned. The presence of acceptor (A) and negative ionizable (N) features in all these compounds clearly point out the importance of the carboxylic acid group, which is of high significance due to the fact that interactions of the carboxylic acid group with the arginine residue of FXR are well-established for determining the affinity of the compounds towards the FXR protein.

3.3. 3D-QSAR modeling

Considering that the alignment technique plays a crucial role in the 3D-QSAR modeling, two different types of alignment techniques were used to perform 3D-QSAR analyses with the Open3DQSAR tool [38]. First, the obvious choice is the alignment obtained from the above-mentioned pharmacophore modeling. Second, we performed molecular docking analyses with each structure of the dataset at the binding site of FXR enzymes (PDB:6HL1) using the AutoDock Vina

software [42,43]. However, the best poses (with maximum binding energy) failed to provide any statistically reliable 3D-QSAR model (results not shown). Therefore, we picked those poses where the carboxylate group of the compounds interacted with the Arg331 residue of the protein, taking into account the crucial role of this interaction in determining the activity [22,41]. More specifically, we resorted to the constrained docking technique for obtaining the docking-based alignment for the 3D-QSAR modeling. On the other hand, it is important to understand that the *QPhAR* based alignment that we used in this work is basically a template-based alignment that was obtained by the greedy optimization method [35,36]. The results of the 3D-QSAR modeling analysis are shown in Table 5.

The most predictive 3D-QSAR model was obtained from the alignment derived from *QPhAR* by applying the FFD-SEL based feature selection technique. This model yielded values for Q^2_{LOO} and R^2_{Pred} of 0.657 and 0.790, respectively. However, the constrained-docking based alignment also let to a satisfactory 3D-QSAR model using the UVE-PLS feature selection technique, as can be observed from its obtained values for Q^2_{LOO} and R^2_{Pred} (0.635 and 0.716). These results reflect that the constrained docking-based predictions were successful in extracting meaningful information about the structural requirements of these

Table 5
Statistical results of the 3D-QSAR modeling analyses performed with Open3DQSAR.

Parameter	QPhAR		Constrained Docking	
	FFD-SEL	UVE-PLS	FFD-SEL	UVE-PLS
PC [a]	4	5	2	4
$N_{training}$	28	28	28	28
F-test	86.25	120.231	14.056	70.291
R^2/SDEC	0.937/0.135	0.965/101	0.351/0.435	0.924/0.149
Q^2_{LOO}/SDEP	0.657/0.317	0.453/400	0.191/0.486	0.635/0.326
Q^2_{LTO}/SDEP	0.640/0.324	0.440/0.404	nd	0.619/0.333
Q^2_{LMO}/SDEP	0.590/0.339	0.389/0.421	nd	0.527/0.369
N_{test}	7	7	7	7
R^2_{Pred}/SDEP	0.790/0.289	0.807/0.277	nd	0.716/0.336
Q^2_S	0.494	na	nd	0.418

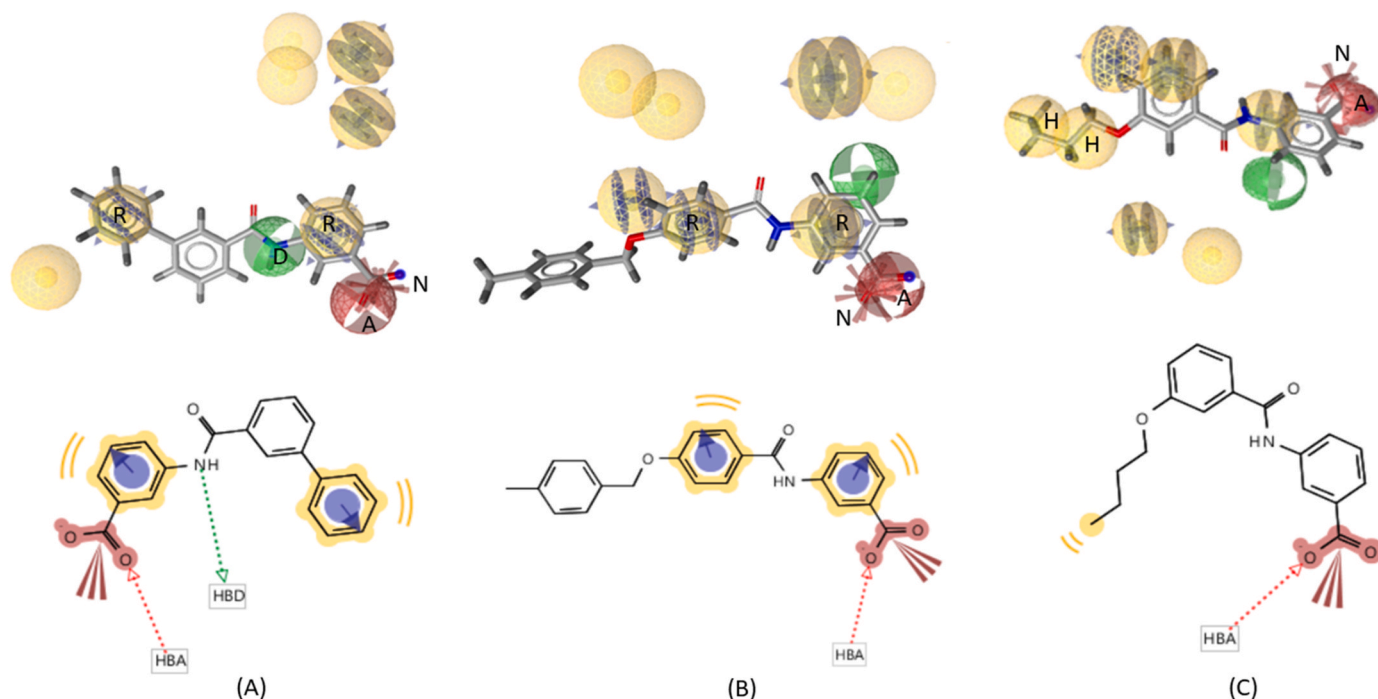[a] PC: Number of principal components.

Fig. 6. Pharmacophore alignment of (A) SN01, (B) SN30 and (C) SN35 with hpmodel.

compounds for higher biological activity. Nevertheless, on the basis of higher statistical predictivity, the *QPhAR* alignment-based model is described and discussed here. The contour maps obtained from this 3D-QSAR model are shown in Fig. 7 with respect to the most potent (SN01) and the least potent (SN35) compounds of the current dataset.

First of all, the significance of steric features (82.67%) is considerably higher than that of the electrostatic features (17.33%) in this 3D-QSAR model. The biphenyl moiety of the SN01 was found to be remarkably close to the steric favorable fields. On the other hand, the lack of any bulky residue near the steric favorable contour map for SN35 indicates that bulky aromatic residues are essential for pertaining to higher activity. Further analysis revealed that the phenoxyphenyl scaffolds of some potent compounds are also inserted into the steric favorable field highlighting the importance of the presence of bulky aromatic residues attached to benzene carboxamide residue. The latter is better reflected from the fact that most of the less potent compounds fail to insert into these same steric fields. Naturally, the question remains whether the biphenyl or phenoxyphenyl moieties fulfil steric requirements only or an electrostatic influence is also involved in establishing their potency. Even though electrostatic interactions were less effective in the 3D-QSAR model, the biphenyl residue of SN01 was found to be close to both electropositive and electronegative fields. Therefore, apart from steric interactions − i.e.: related more to the van der Waals interactions with binding site amino acids, these aromatic rings may also be involved in interactions in which the electron distribution of aromatic rings may play an important role in the binding.

In addition, we noticed that the terminal benzene of phenoxyphenyl residue was close to the electronegative field and more importantly the electron rich oxygen was inserted into the electronegative field, signifying that its replacement with non-polar –CH$_2$ may indeed deteriorate the activity. Furthermore, it can be also observed from Fig. 7 that the aliphatic side chain of SN35 is inserted into the polar field indicating unfavorable interactions. Significantly, no contour map was found near the carboxylate moiety of these compounds in spite of the fact that this moiety is likely to be extremely important as suggested by the 2D-QSAR and pharmacophore modeling. Yet, this can be explained by the fact that all compounds in the dataset contained this moiety and therefore no significant field change was detected by 3D-QSAR and the same argument holds true for the benzamide residue as well. Nevertheless, apart

from having statistically reliable results, the 3D-QSAR model offers some essential information regarding the structural requirements of these compounds for higher activity.

3.4. Molecular dynamics simulations

Finally, in order to understand the dynamic behavior of receptor-bound complexes of 3-benzamidobenzoic acid derivatives, we carried out 50 ns MD simulations with the complexes of two derivatives, i.e. the most potent compound of the dataset (SN01) and the least potent compound of the dataset (SN35), as well as with the apoprotein. Along with these, for comparative purposes, we also performed MD simulations of the chenodeoxycholic acid (CDCA) − i.e., a physiological FXR agonist. Note that, even though CDCA is a considered as a full agonist, the reported EC$_{50}$ values reported for this compound have a range of 3–30 μM. Therefore, exact comparison with the biological activity of this compound with our dataset compounds is not justified. However, inclusion of this compound should help us to differentiate key structural changes associated with the partial and full agonist behaviors. Both dataset compounds were docked with the X-ray crystal structure of FXR protein (PDB: 6HL1) using the Autodock software [42]. Initially, an analysis of the trajectory was carried out to understand the stability of the complexes. Fig. 8 depicts some important outcomes of these MD simulations. As can be seen, the RMSD plots of protein complexes and ligands clearly show that increased fluctuations are observed for the least potent analogue SN35 as compared to the most potent analogue SN01. More importantly, the SN01-FXR complex was more stable than the apoprotein and CDCA-FXR complex. Indeed, the average RMSDs obtained from SN1−FXR, SN35−FXR, apoprotein−FXR and CDCA−FXR were found to be 1.37 Å, 1.89 Å, 1.67 Å and 1.69 Å, respectively. The ligand RMSD plots show that SN01 was stabilized after a 5–6 ns run whereas higher fluctuation was noted in the case of SN35 throughout the simulation even though the ligand RMSD remained less than 2 Å for both these ligands.

Understandably, both complexes were found to be stable during the simulations simply because of the fact that both these were active against FXR protein. Here, our main focus should be on the differences in their potencies. The CDCA, on the other hand, depicted the maximum stability throughout the MD simulation. The RMSF plot provides some
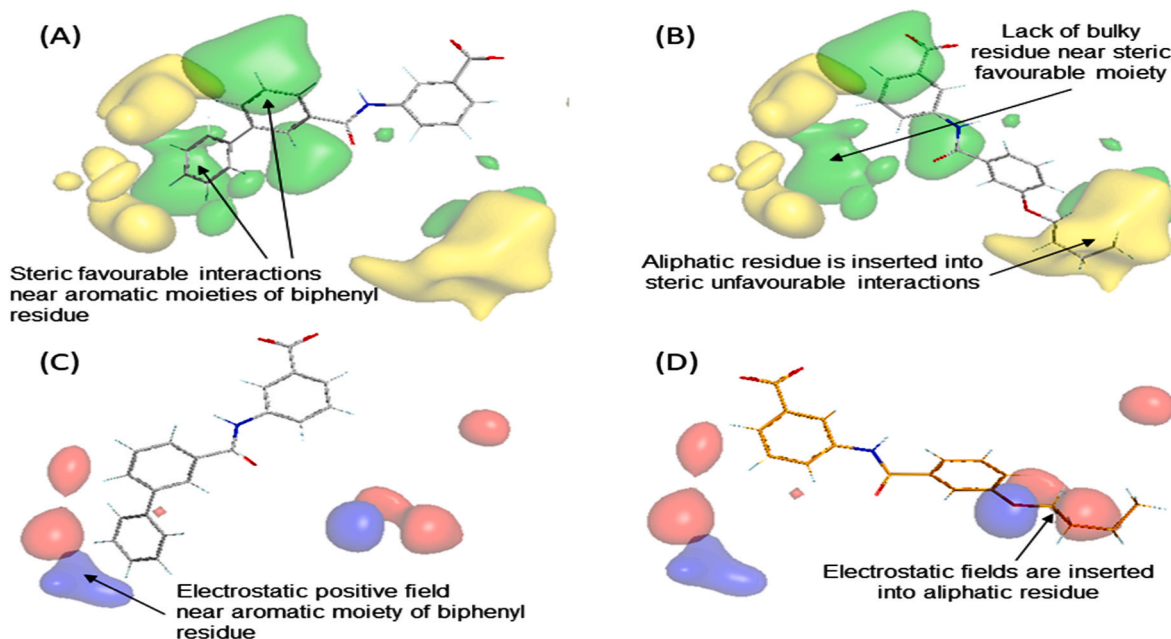

Fig. 7. (A) Contour maps obtained from the best 3D-QSAR model (Green: Steric favorable; Blue: Electropositive favorable; Red: Electronegative favorable). (B) SN35 with steric features (C) SN01 with electrostatic features and (C) SN35 with electrostatic features.
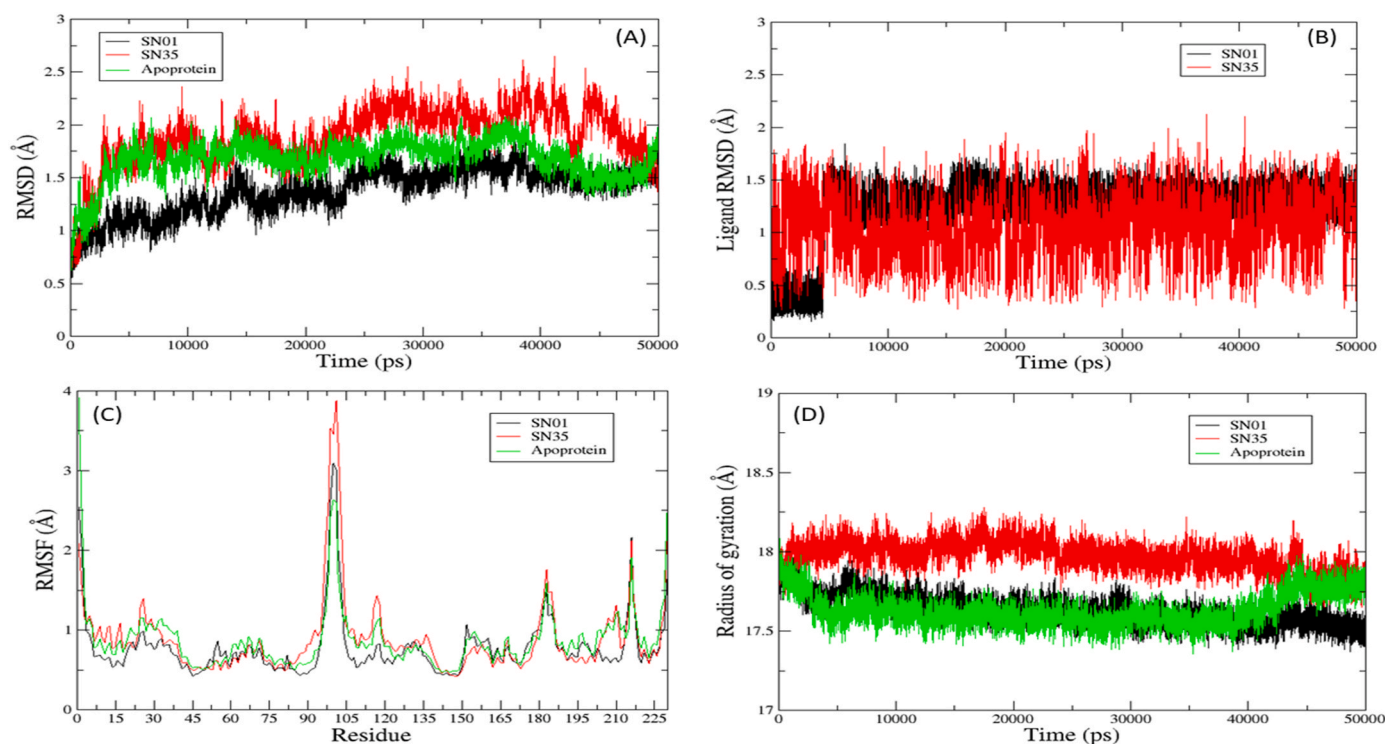
Fig. 8. Plots obtained from an analysis of the trajectories for SN01, SN35 and apoprotein – (A) Protein/complex RMSD, (B) ligand RMSD, (C) RMSF and (D) Radius of gyration.

interesting hints. First, it is clearly observed that the fluctuations of amino acid residue remain reduced in the case of SN01 as compared to SN35 as well as to apoprotein but more importantly, the fluctuation between residues 84−92, 104−114, 117−132 decreased significantly for SN01. These residues belong to H5, H6 and H8 helices, respectively. Recently, Gohda and co-workers pointed out that the larger the fluctuations are in the H5, H6 or H8 helices, the less agonistic activity a compound should possess [53,54]. Moreover, as compared to both the apoprotein and the SN35−complex, the SN01-complex displayed significant stability in the residues 205−212, which are part of H11 (residues 186−214) that is involved directly in the formation of heterodimer with retinoid X receptor (RXR) in order in order to regulate transcription [54]. A sharp difference was observed in the fluctuations of CDCA-FXR complex with other complexes in the H6 domain and this observation is consistent with a previous work where the fluctuations of another full agonist GW4046-FXR complex in this region have been found to be significantly less as compared to a complex formed with a partial agonist

[53]. In addition, as compared to S1− and S35−complexes, low fluctuations were noted for the CDCA-FXR complex in the residues 214−217, which belong to the AF-2 loop that connects H11 and H12 (see Fig. S2). Previously, Merk et al. [41] pointed out that partial agonists (i.e., S1 and S35) destabilize this AF-2 loop and this information complies with our MD simulation results. The radius of gyration plot, on the other hand, points to the fact that more compactness is observed for the FXR−SN01 complex as compared to the FXR-SN35 complex. Clearly, the current MD simulations successfully project the differences in the activities of SN01 and SN35.

Following on, we carried out a hydrogen bond analysis with the produced MD trajectories. For this analysis, we selected those amino acids that lied within 8 Å radius of the bound/docked ligands. Fig. 9 shows the plots of the number of hydrogen bonds recorded during 50 ns runs for the SN01 and SN35 complexes. First, it can be observed that hydrogen bonds are consistently formed among the ligands and surrounding amino acids in both complexes but the total number of
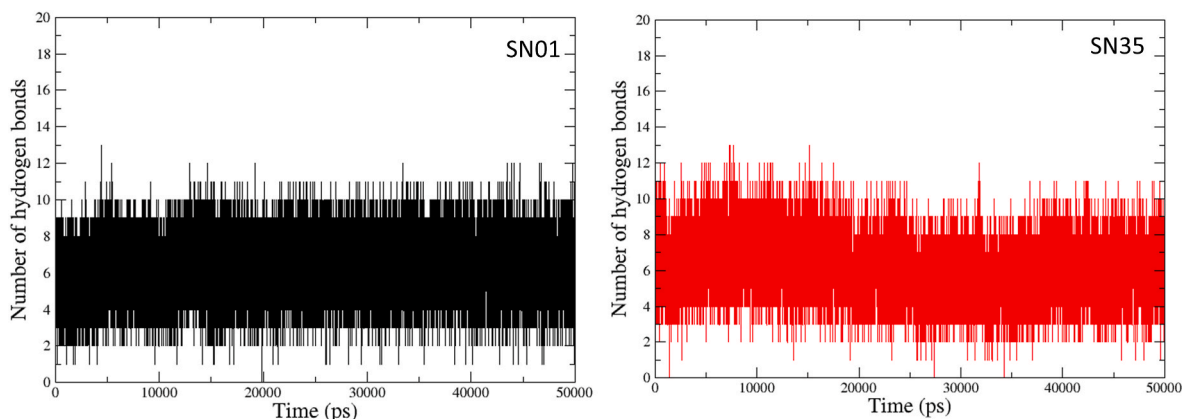

Fig. 9. Number of hydrogen bonds formed among the binding site amino acids and ligands during 50 ns MD simulations.

hydrogen bonds was slightly higher in SN01 in comparison with SN35, especially in the last 20 ns simulation time. The same plot generated with CDCA (see Fig. S3) shows that the number of hydrogen bond formation for SN01 is even higher than that pertaining to CDCA.

Further investigations revealed that the carboxylate groups of both these ligands formed stable hydrogen bond interactions with Arg89 (Arg331 in PDB: 6HL1 since the amino acids were renumbered for MD simulation). More specifically, a total of 60,208 and 53,952 frames were counted where hydrogen bond interactions occurred between any one oxygen atom of the carboxylate group with the side chain of Arg89. Nevertheless, in the case of SN01, around 55% frames were found where hydrogen bond interactions took place between its carboxylate group and the Met23 residue of FXR protein. For SN35, only 38% frames depicted such interactions. It is also important to mention that the carboxamide (−CONH−) group of SN01 interacted with Ser90 by hydrogen bond interaction whereas no such interaction was recorded between the SN35 and Ser90.

The final receptor-bound poses of SN01 and SN35 with the FXR protein were extracted after 50 ns of MD simulations to check the key interactions that are depicted in Fig. 9.

It is clearly visible that the interactions with the biphenyl scaffold of SN01 contributed largely to high binding affinity towards the receptor. The terminal benzene moiety of this scaffold interacted with residues such as Met123, Trp212 and Ile110 with π-Sulfur, π-π and π-alkyl interactions, clearly suggesting that the aromaticity of this ring plays an important role in ligand-receptor interactions. Noticeably, the alkoxy chain of SN35 fails to replicate these interactions. The interactions with Ile110 and Met123 may be of high significance since these residues are in fact part of H6 and H8 helices of FXR protein, respectively. Moreover, the π-π interaction of SN01 with Trp212 should not be ignored especially because the SN01−complex differed from the SN35−complex as the former showed extremely low fluctuations in the residues 205−212. As explained by Merk et al. [41], this tryptophan residue plays a crucial role in the FXR modulation. In fact, endogenous FXR agonist CDCA failed to activate W212A and W212Y mutated FXR. Overall, the interactions obtained from the biphenyl scaffold of SN01 were significantly higher as compared to the interactions obtained from the butoxyphenyl scaffold of SN35. This directly connects with the interpretation derived from the 3D-QSAR modeling and pharmacophore mapping. Note that, complying with the interactions displayed in Fig. 10, the 3D-QSAR model pointed out that not only steric factors but some electrostatic interactions may also contribute to the higher activity of SN01. The interaction between

the Met86 residue and SN01 is also worth mentioning here since such interaction was missing for SN35. It is reasonable to assume that this interaction with SN01 should have contributed to the stabilization of the H5 helix, as shown in the RMSF plot (Fig. 8C).

Since substantial hints already show that stabilization of H5, H6 and H8 helices may play a significant role in determining the activity of these 3-benzamidobenzoic acid derivatives, we finally performed per-residue decomposition analyses with selected residues of these helices that remained close to the bound ligands. The results of such analyses are depicted in Fig. 11.

The residues of H5 and H6 helices that showed maximum differences in the interactions between SN01 and SN35 are Met86, Leu106 and Ile110. For all these amino acid residues, the van der Waals interaction energies were much greater in SN01 as compared to SN35. Additionally, Leu106 and Ile110 were also found to have significant electrostatic interactions with SN01 indicating that π-alkyl interactions may contribute to interactions between these ligands and the FXR protein. Even though a significant change of interactions was not noticed for the residues of H5 helix, SN01 showed increased van der Waals interactions with residues such as Ile120, Met123 and Tyr127. It may therefore be summarized that greater overall stabilization imparted to H5, H6 and H8 helices contributes largely towards the higher agonistic potential of SN01 as compared to SN35. This inference was however obtained not only from the MD simulations but also from the 2D-QSAR, 3D-QSAR and pharmacophore mapping analyses. Similarly, the per residue analyses performed with the CDCA−FXR complex is presented in Fig. S4 and it can be observed that, save for a few residues, both CDCA and SN01 depicted similar binding with the residues of the H5, H6 and H8 helices.

The input files for the 2D-QSAR and 3D-QSAR modeling, output files of the pharmacophore mapping and PDB files used for the docked poses, which were also used as input structures in the MD simulations, are provided in our Github repository https://github.com/ncordeirfcup/FXR_agonists.

4. Conclusions

The present work reports a range of cheminformatics and molecular modeling studies with a series of 3-benzamidobenzoic acid derivatives in order to understand their structural requirements for higher potency towards FXR protein as partial agonists. Each study revealed important information pertaining to the structural requirements of these derivatives for higher activity towards the FXR protein. The 2D-QSAR
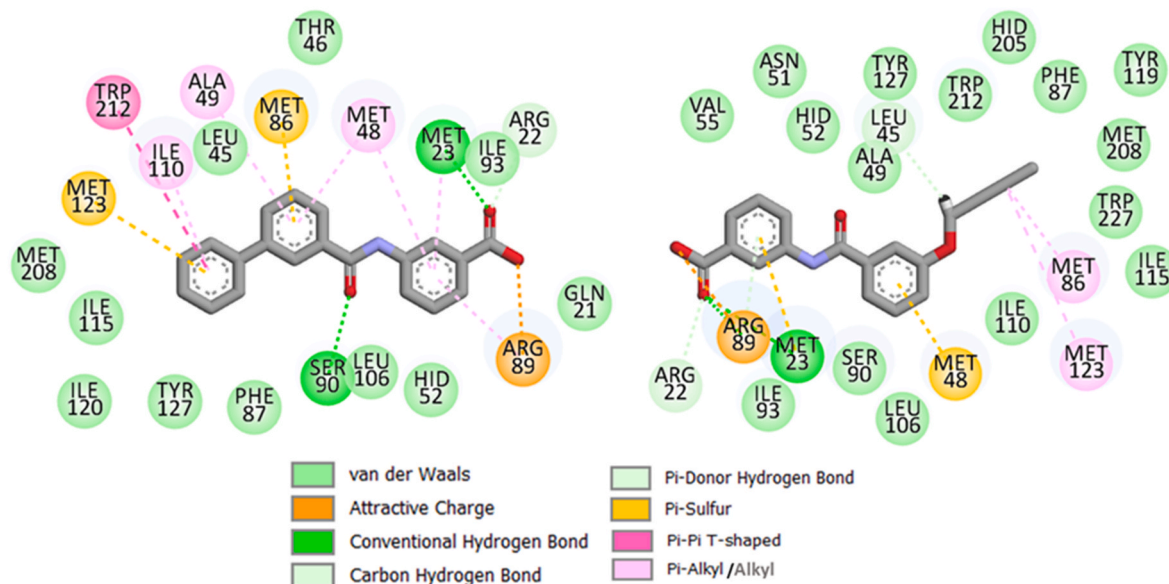

Fig. 10. The receptor-ligand interactions for the SN01−complex (left) and the SN35−complex (right) extracted from the last frame of the MD simulations.
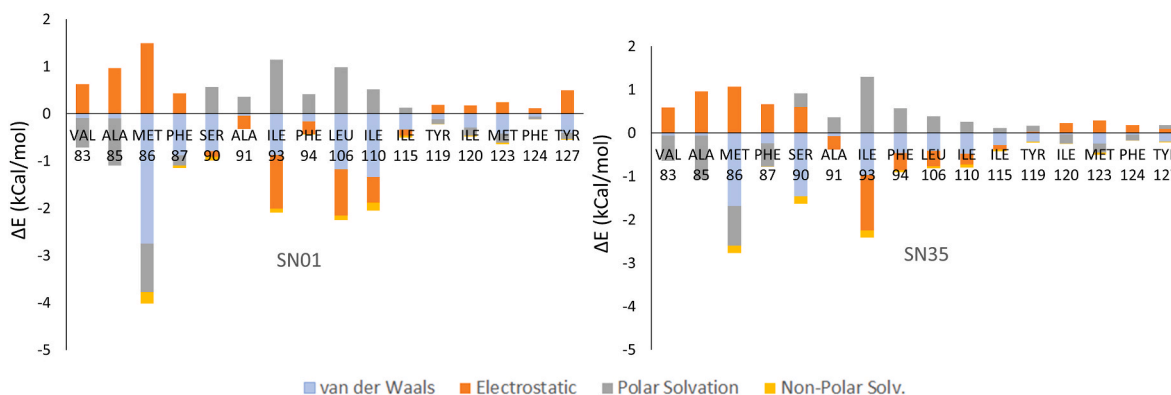
Fig. 11. Per residue decomposition analyses performed with selected amino acids of the H5, H6 and H8 helices of FXR protein.

models highlighted specific 2D-pharmacophoric and topological requirements (cf. Figs. 3–4) as well as charge distributions, surface areas and van der Waals volumes that lead to the biological activity of the compounds. The pharmacophore mapping was performed with a novel as well as less explored ligand-based pharmacophore mapping tool named *QPhAR*. This open-access software tool was found to be extremely useful towards extracting key 3D-pharmacophoric features (cf. Fig. 6), that in turn were found also to be highly compliant with the results obtained from the other *in silico* models developed in this work. 3D-QSAR models also provided us satisfactory predictivity and at the same time, helped us to deduce the key pharmacophoric features, steric and electrostatic factors responsible for determining the higher activity of the 3-benzamidobenzoic acid derivatives (cf. Fig. 7). Even though comparatively faster ligand-based techniques provided insights into the structural requirements, it was the structure-based strategy (i.e.: the MD simulations) that guided us to link the structural interpretations obtained from 2D-QSAR, 3D-QSAR and pharmacophore models with the involvement of specific amino acid residues at the binding site of the FXR protein. The MD simulations revealed that the hydrogen bond interactions with the carboxylate group of the ligands play an important role in ligand receptor binding but it is the higher stabilization of different helices of FXR (e.g., H5, H6 and H8) though aromatic scaffolds of the ligands that may lead to higher activity for these ligands. This work thus comprises important guidelines for designing novel FXR partial agonists as new medicines in the treatment of NAFLD. What is more, it mainly relied on open-access tools in order to develop the *in silico* models that thus ensures their reproducibility as well as utilization.

Previously, a number of studies have been reported with different series of compounds (benzimidazole derivatives, anthranilic acid derivatives, etc.) with agonistic properties for FXR [55–57]. Similar to the current investigation, these studies highlighted the importance of hydrophobic, aromatic and hydrogen bond acceptor features for determining the agonistic activity towards the FXR protein. In addition, the importance of the hydrogen bond interactions between the ligand and the Arg89 residue has been highlighted in all these studies. Furthermore, interactions with amino acid residues like Leu45, Ala49, Arg89, Met123, Tyr127, Trp212 and Met208 were found to be crucial in establishing the biological activity. Chen et al. presented a multi-model *in silico* analysis with anthranilic acid derivatives as partial FXR agonists that depicted the importance of solvent accessible surface area and charges for controlling their activity [56]. Evidently, the later results are well in line with the results of our present investigation. Furthermore, some recent studies have focused on the mechanisms behind full agonism, partial agonism and antagonism properties of various compounds towards the FXR protein. These recent investigations revealed that even though the binding sites of the natural agonist CDCA, full agonists, partial agonists and antagonists do not vary much in FXR, the conformational changes (especially in the H11 helix and AF-2 loop) [41] imparted by these bindings are responsible for different biological outcomes. The latter

therefore leaves us with a future opportunity to look into the characterization of FXR antagonists with similar *in silico* methods described in this work starting from a dataset of FXR antagonists. Detailed MD simulation analyses may be also performed to understand the conformational changes imparted by various FXR ligands (i.e., natural substrate CDCA, full agonists, partial agonists, and antagonists) to facilitate novel lead discovery targeting the FXR protein.

Funding information

CRediT authorship contribution statement

Soumya Mitra, Amit Kumar Halder, Nilanjan Ghosh, Subhash C. Mandal, and M. Natália D. S. Cordeiro: Conceptualization, Methodology. Soumya Mitra, Amit Kumar Halder, and M. Natália D. S. Cordeiro: Software. Amit Kumar Halder, Subhash C. Mandal, and M. Natália D. S. Cordeiro: Writing – review & editing.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.106789.

References

[1] M.J. Armstrong, L.A. Adams, A. Canbay, W.-K. Syn, Extrahepatic complications of nonalcoholic fatty liver disease, Hepatology 59 (2014) 1174–1197, https://doi.org/10.1002/hep.26717.

[2] H.C. Masuoka, N. Chalasani, Nonalcoholic fatty liver disease: an emerging threat to obese and diabetic individuals, Ann. N. Y. Acad. Sci. 1281 (2013) 106–122, https://doi.org/10.1111/nyas.12016.

[3] C.D. Byrne, G. Targher, NAFLD: a multisystem disease, J. Hepatol. 62 (2015) S47–S64, https://doi.org/10.1016/j.jhep.2014.12.012.

[4] Z. Younossi, Q.M. Anstee, M. Marietti, T. Hardy, L. Henry, M. Eslam, J. George, E. Bugianesi, Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention, Nat. Rev. Gastroenterol. Hepatol. 15 (2017) 11–20, https://doi.org/10.1038/nrgastro.2017.109.

[5] E. Cobbina, F. Akhlaghi, Non-alcoholic fatty liver disease (NAFLD) – pathogenesis, classification, and effect on drug metabolizing enzymes and transporters, Drug Metab. Rev. 49 (2017) 197–211, https://doi.org/10.1080/03602532.2017.1293683.

[6] D.E. Kleiner, H.R. Makhlouf, Histology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis in adults and children, Clin. Liver Dis. 20 (2016) 293–312, https://doi.org/10.1016/j.cld.2015.10.011.

[7] S. Caldwell, Y. Ikura, D. Dias, K. Isomoto, A. Yabu, C. Moskaluk, P. Pramoonjago, W. Simmons, H. Scruggs, N. Rosenbaum, T. Wilkinson, P. Toms, C.K. Argo, A.M. S. Al-Osaimi, J.A. Redick, Hepatocellular ballooning in NASH, J. Hepatol. 53 (2010) 719–723, https://doi.org/10.1016/j.jhep.2010.04.031.

[8] Q.M. Anstee, G. Targher, C.P. Day, Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis, Nat. Rev. Gastroenterol. Hepatol. 10 (2013) 330–344, https://doi.org/10.1038/s41575-019-0169-z.

[9] G. Targher, L. Bertolini, R. Padovani, S. Rodella, R. Tessari, L. Zenari, C. Day, G. Arcaro, Prevalence of nonalcoholic fatty liver disease and its association with cardiovascular disease among type 2 diabetic patients, Diabetes Care 30 (2007) 1212–1218, https://doi.org/10.2337/dc06-2247.

[10] R.M. Williamson, J.F. Price, S. Glancy, E. Perry, L.D. Nee, P.C. Hayes, B.M. Frier, L. A.F. Van Look, G.I. Johnston, R.M. Reynolds, M.W.J. Strachan, Prevalence of and risk factors for hepatic steatosis and nonalcoholic fatty liver disease in people with type 2 diabetes: the edinburgh type 2 diabetes study, Diabetes Care 34 (2011) 1139–1144, https://doi.org/10.2337/dc10-2229.

[11] M. Asrih, F.R. Jornayvaz, Diets and nonalcoholic fatty liver disease: the good and the bad, Clin. Nutr. 33 (2014) 186–190, https://doi.org/10.1016/j.clnu.2013.11.003.

[12] A.M. Oseini, A.J. Sanyal, Therapies in non-alcoholic steatohepatitis (NASH), Liver Int. 37 (2017) 97–103, https://doi.org/10.1111/liv.13302.

[13] G. Porez, J. Prawitt, B. Gross, B. Staels, Bile acid receptors as targets for the treatment of dyslipidemia and cardiovascular disease, J. Lipid Res. 53 (2012) 1723–1737, https://doi.org/10.1194/jlr.r024794.

[14] S.L. Friedman, B.A. Neuschwander-Tetri, M. Rinella, A.J. Sanyal, Mechanisms of NAFLD development and therapeutic strategies, Nat. Med. 24 (2018) 908–922, https://doi.org/10.1038/s41591-018-0104-9.

[15] K. Panzitt, M. Wagner, FXR in liver physiology: multiple faces to regulate liver metabolism, Biochim. Biophys. Acta, Mol. Basis Dis. 1867 (2021), 166133, https://doi.org/10.1016/j.bbadis.2021.166133.

[16] B.A. Neuschwander-Tetri, R. Loomba, A.J. Sanyal, J.E. Lavine, M.L. Van Natta, M. F. Abdelmalek, N. Chalasani, S. Dasarathy, A.M. Diehl, B. Hameed, K.V. Kowdley, A. McCullough, N. Terrault, J.M. Clark, J. Tonascia, E.M. Brunt, D.E. Kleiner, E. Doo, Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial, Lancet 385 (2015) 956–965, https://doi.org/10.1016/s0140-6736(14)61933-4.

[17] L. Chuangzhen, Y. Bingqing, C. Lixin, Z. Zhaohui, Y. Weixiang, Z. Hui, B. Wenke, Y. Yuping, N. Biao, Obeticholic acid induces hepatoxicity via FXR in the NAFLD mice, Front. Pharmacol. 13 (2022), 880508, https://doi.org/10.3389/fphar.2022.880508.

[18] C.R. Pullinger, C. Eng, G. Salen, S. Shefer, A.K. Batta, S.K. Erickson, A. Verhagen, C. R. Rivera, S.J. Mulvihill, M.J. Malloy, J.P. Kane, Human cholesterol 7α-hydroxylase (CYP7A1) deficiency has a hypercholesterolemic phenotype, J. Clin. Invest. 110 (2002) 109–117, https://doi.org/10.1172/jci15387.

[19] S. Mudaliar, R.R. Henry, A.J. Sanyal, L. Morrow, H.U. Marschall, M. Kipnes, L. Adorini, C.I. Sciacca, P. Clopton, E. Castelloe, P. Dillon, M. Pruzanski, D. Shapiro, Efficacy and safety of the farnesoid X receptor agonist obeticholic acid in patients with type 2 diabetes and nonalcoholic fatty liver disease, Gastroenterology 145 (2013) 574–582, https://doi.org/10.1053/j.gastro.2013.05.042.

[20] G. Xu, H. Li, L. Pan, Q. Shang, A. Honda, M. Ananthanarayanan, S.K. Erickson, B. L. Shneider, S. Shefer, J. Bollineni, B.M. Forman, Y. Matsuzaki, F.J. Suchy, G. S. Tint, G. Salen, FXR mediated down-regulation of CYP7A1 dominates LXRα in long-term cholesterol-fed NZW rabbits, J. Lipid Res. 44 (2003) 1956–1962, https://doi.org/10.1194/jlr.m300182-jlr200.

[21] C. Lin, B. Yu, L. Chen, Z. Zhang, W. Ye, H. Zhong, W. Bai, Y. Yang, B. Nie, Obeticholic acid induces hepatoxicity via FXR in the NAFLD mice, Front. Pharmacol. 13 (2022), https://doi.org/10.3389/fphar.2022.880508.

[22] L. Hu, Q. Ren, L. Deng, Z. Zhou, Z. Cai, B. Wang, Z. Li, Design, synthesis, and biological studies of novel 3-benzamidobenzoic acid derivatives as farnesoid X receptor partial agonist, Eur. J. Med. Chem. 211 (2021) 113106, https://doi.org/10.1016/j.ejmech.2020.113106.

[23] ChemAxon Standardizer, Version 15.9.14.0 Software, ChemAxon, Budapest, Hungary, 2010.

[24] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I.V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, J. Comput. Aided Mol. Des. 25 (2011) 533–554, https://doi.org/10.1007/s10822-011-9440-2.

[25] A. Mauri, alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints, Ecotoxicological QSARs, 2020, pp. 801–820, https://doi.org/10.1007/978-1-0716-0150-1_32.

[26] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of automatic three-dimensional model builders using 639 X-ray structures, J. Chem. Inf. Comput. Sci. 34 (2002) 1000–1008, https://doi.org/10.1021/ci00020a039.

[27] A.K. Halder, A.H.S. Delgado, M.N.D.S. Cordeiro, First multi-target QSAR model for predicting the cytotoxicity of acrylic acid-based monomers, Dent. Mater. 38 (2022) 333–346, https://doi.org/10.1016/j.dental.2021.12.014.

[28] P. Ambure, R.B. Aher, A. Gajewicz, T. Puzyn, K. Roy, NanoBRIDGES‟ software: open access tools to perform QSAR and nano-QSAR modeling, Chemometr. Intell. Lab. Syst. 147 (2015) 1–13, https://doi.org/10.1016/j.chemolab.2015.07.007.

[29] I.V. Tetko, V.Y. Tanchuk, A.E.P. Villa, Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices, J. Chem. Inf. Comput. Sci. 41 (2001) 1407–1421, https://doi.org/10.1021/ci010368v.

[30] A. Golbraikh, A. Tropsha, Beware of q2, J. Mol. Graph. Model. 20 (2002) 269–276, https://doi.org/10.1016/s1093-3263(01)00123-1.

[31] P.P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, Molecules 14 (2009) 1660–1701, https://doi.org/10.3390/molecules14051660.

[32] W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. Peter He, J.W. Lillard Jr., A study of effects of MultiCollinearity in the multivariable analysis, Int. J. Appl. Sci. Technol. 4 (2014) 9–19.

[33] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometr. Intell. Lab. Syst. 109 (2011) 146–161, https://doi.org/10.1016/j.chemolab.2011.08.007.

[34] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701, https://doi.org/10.1002/qsar.200610151.

[35] S.M. Kohlbacher, T. Langer, T. Seidel, QPHAR: quantitative pharmacophore activity relationship: method and validation, J. Cheminf. 13 (2021) 57, https://doi.org/10.1186/s13321-021-00537-9.

[36] S. Kohlbacher, M. Schmid, T. Seidel, T. Langer, Applications of the novel quantitative pharmacophore activity relationship method QPhAR in virtual screening and lead-optimisation, Pharmaceuticals 15 (2022) 1122, https://doi.org/10.3390/ph15091122.

[37] P. Tosco, T. Balle, F. Shiri, Open3DALIGN: an open-source software aimed at unsupervised ligand alignment, J. Comput. Aided Mol. Des. 25 (2011) 777–783, https://doi.org/10.1007/s10822-011-9462-9.

[38] P. Tosco, T. Balle, Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields, J. Mol. Model. 17 (2010) 201–208, https://doi.org/10.1007/s00894-010-0684-x.

[39] A.K. Halder, S.A. Amin, T. Jha, S. Gayen, Insight into the structural requirements of pyrimidine-based phosphodiesterase 10A (PDE10A) inhibitors by multiple validated 3D QSAR approaches, SAR QSAR Environ. Res. 28 (2017) 253–273, https://doi.org/10.1080/1062936x.2017.1302991.

[40] A. Ghosh, P. Panda, A.K. Halder, M.N.D.S. Cordeiro, In silico characterization of aryl benzoyl hydrazide derivatives as potential inhibitors of RdRp enzyme of H5N1 influenza virus, Front. Pharmacol. 13 (2022), 1004255, https://doi.org/10.3389/fphar.2022.1004255.

[41] D. Merk, S. Sreeramulu, D. Kudlinzki, K. Saxena, V. Linhard, S.L. Gande, F. Hiller, C. Lamers, E. Nilsson, A. Aagaard, L. Wissler, N. Dekker, K. Bamberg, M. Schubert-Zsilavecz, H. Schwalbe, Molecular tuning of farnesoid X receptor partial agonism, Nat. Commun. 10 (2019) 2915, https://doi.org/10.1038/s41467-019-10853-2.

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

[43] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A. J. Olson, AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility, J. Comput. Chem. 30 (2009) 2785–2791, https://doi.org/10.1002/jcc.21256.

[44] A.K. Halder, B. Honarparvar, Molecular alteration in drug susceptibility against subtype B and C-SA HIV-1 proteases: MD study, Struct. Chem. 30 (2019) 1715–1727, https://doi.org/10.1007/s11224-019-01305-0.

[45] T.J. Dolinsky, P. Czodrowski, H. Li, J.E. Nielsen, J.H. Jensen, G. Klebe, N.A. Baker, PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations, Nucleic Acids Res. 35 (Web Server) (2007) W522–W525, https://doi.org/10.1093/nar/gkm276.

[46] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, D.A. Case, Continuum solvent studies of the stability of DNA, RNA, and Phosphoramidate−DNA helices, J. Am. Chem. Soc. 120 (1998) 9401–9409, https://doi.org/10.1021/ja981844+.

[47] M. Reutlinger, C.P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, G. Schneider, Chemically advanced template Search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules, Mol. Inform. 32 (2013) 133–138, https://doi.org/10.1002/minf.201200141.

[48] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, 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 (1998) 55–68, https://doi.org/10.1021/cc9800071.

[49] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, John Wiley & Sons, Ltd, WILEY-VCH, Germany, 2000, 2000.

[50] R. Todeschini, P. Gramatica, R. Provenzani, E. Marengo, Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons, Chemometr. Intell. Lab. Syst. 27 (1995) 221–229, https://doi.org/10.1016/0169-7439(95)80026-6.

[51] F. Grisoni, D. Merk, V. Consonni, J.A. Hiss, S.G. Tagliabue, R. Todeschini, G. Schneider, Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity, Commun. Chem. 1 (2018) 44, https://doi.org/10.3929/ethz-b-000290645.

[52] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33, https://doi.org/10.1186/1758-2946-3-33.

[53] K. Gohda, Y. Iguchi, A. Masuda, K. Fujimori, Y. Yamashita, N. Teno, Design and identification of a new farnesoid X receptor (FXR) partial agonist by computational structure–activity relationship analysis: ligand-induced H8 helix fluctuation in the

ligand-binding domain of FXR may lead to partial agonism, Bioorg. Med. Chem. Lett. 41 (2021), 128026, https://doi.org/10.1016/j.bmcl.2021.128026.

[54] L. Yang, D. Broderick, Y. Jiang, V. Hsu, C.S. Maier, Conformational dynamics of human FXR-LBD ligand interactions studied by hydrogen/deuterium exchange mass spectrometry: insights into the antagonism of the hypolipidemic agent Z-guggulsterone, Biochim. Biophys. Acta 1844 (2014) 1684–1693, https://doi.org/10.1016/j.bbapap.2014.06.007.

[55] T. Sindhu, P. Srinivasan, Pharmacophore modeling, 3D-QSAR and molecular docking studies of benzimidazole derivatives as potential FXR agonists, J. Recept.

Signal Transduction 34 (2014) 241–253, https://doi.org/10.3109/10799893.2014.885048.

[56] M. Chen, X. Yang, X. Lai, J. Kang, H. Gan, Y. Gao, Structural investigation for optimization of anthranilic acid derivatives as partial fxr agonists by in silico approaches, Int. J. Mol. Sci. 17 (2016) 536, https://doi.org/10.3390/ijms17040536.

[57] S. Zhao, W. Peng, X. Li, L. Wang, W. Yin, Y.-D. Wang, R. Hou, W.-D. Chen, Pharmacophore modeling and virtual screening studies for discovery of novel farnesoid X receptor (FXR) agonists, RSC Adv. 11 (2021) 2158–2166, https://doi.org/10.1039/D0RA09320C.