Quick and Efficient Quantitative Predictions of Androgen Receptor Binding Affinity for Screening Endocrine Disruptor Chemicals Using 2D-QSAR and Chemical Read-Across

Arkaprava Banerjee[a], Priyanka De[a], Vinay Kumar[a], Supratik Kar[b], Kunal Roy[a,*]

[a]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata700032, India

[b]Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, Mississippi 39217, United States

**For Correspondence*

E-mail: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in;

URL: https://sites.google.com/site/kunalroyindia;

Tel: +91 98315 94140

Abstract

Endocrine Disruptor Chemicals are synthetic or natural molecules in the environment that promote adverse modifications of endogenous hormone regulation in humans and/or in animals. In the present research, we have applied two-dimensional quantitative structure-activity relationship (2D-QSAR) modeling to analyze the structural features of these chemicals responsible for binding to the androgen receptors (logRBA) in rats. We have collected the receptor binding data from the EDKB database (https://www.fda.gov/science-research/endocrine-disruptor-knowledge-base/accessing-edkb-database) and then employed the **DTC-QSAR** tool, available from https://dtclab.webs.com/software-tools, for dataset division, feature selection, and model development. The final partial least squares was evaluated using various stringent validation criteria. From the model, we interpreted that hydrophobicity, steroidal nucleus, bulkiness and a hyrdrogen bond donor at an appropriate position contribute to the receptor binding affinity, while presence of electron rich features like aromaticity and polar groups decrease the receptor binding affinity. Additionally we have also performed chemical Read-Across predictions using **Read-Across-v3.1** available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, and the results for the external validation metrics were found to be better than the QSAR-derived predictions. To explore the essential features responsible for the receptor binding, pharmacophore mapping, molecular docking along with molecular dynamics simulation were also performed, and the results are in accordance with the QSAR findings.

Keywords: Endocrine disruptors; Androgen receptor binding affinity; QSAR; Read-across; docking; Pharmacophore

1. Introduction

It is fascinating that our brain is responsible for almost every physiological function that our body performs. The hypothalamus, also known as our "built-in thermostat" is the control centre for the endocrine system, which comprises various ductless chemical messengers commonly termed as hormones. In nature, there is existence of molecules which can potentially mimic these chemical messengers and bring about "disruption" in the normal physiological functioning of the body. Such compounds are classified as Endocrine Disrupting Chemicals (EDCs) as they mimic the natural hormones, bind to the specific receptors and bring about endocrine disruption in humans and wildlife [1-4]. In 2011, *Schug et al.* [5] reported that EDCs show various neurological, reproductive and cardiovascular adverse effects by interfering with the synthesis, transport, metabolism and release of hormones. However, it has also been observed that EDCs can act on transcriptional coactivators, synthesis and metabolism of steroids, non-steroidal receptors and various other mechanisms that ultimately converge to endocrine and reproductive systems [5]. The complexity in the mechanism of disruption in endocrine functions and activation of signaling pathways probably explains the reason for the lack of experimental toxicity data of EDCs [6]. As compared to estrogenic mode of disruption, little is known about how EDCs adversely affect the androgen receptors and hinders the male reproductive tract health [7]. Among various other targets, chemicals like DDTs, industrial chemical phthalates, organophosphate insecticides like parathion and herbicides of phenylurea derivatives like linuron can potentially bind to the androgen receptor and bring about disruption thus resulting in the toxicity [1].

Development and maintenance of male sexual characteristics is controlled by Androgen Receptors (AR), a class of ligand-activated transcriptional regulatory protein [8]. Most androgenic EDCs perform activation of transcription through receptor mediated mechanism

[9]. Using this information, it is possible to identify the potential EDCs through the competitive binding assay at the AR. **Figure S1 (Supplementary Material SI-1)** represents the potential of EDCs in inhibition of the androgen receptor inside the mammalian cell.

The Organization for Economic Co-operation and Development (OECD) promotes the use of *in silico* approaches wherever applicable. As the resources are limited, it is highly impractical to perform toxicity assessment of all EDCs against all possible end points in the exploration of different disruption mechanisms experimentally [6]. Thus, with the aim of data gap filling, efficient *in silico* approaches with scientifically well defined algorithms are adopted. In recent times, there has been an increase in non-testing methods which comply with the 3Rs (Reduction, Replacement and Refinement in animal experiments) in scientific experimentations [10]. Among various other non-testing methods, Quantitative Structure-Activity Relationship (QSAR) and Chemical Read-Across are two of the most widely used methods for prediction of toxicity associated with chemicals [10-11]. The advantages associated with *in silico* approaches in general are: a) they reduce experimental time, cost and b) they speed up obtaining the desired results. The basic concept behind regression-based QSAR lies in the development of a model consisting of the dependent variable (response) and one or more features in the molecules (independent variables) which contribute to the response values either positively or negatively and is expressed in numerical terms. Read-Across, on the other hand, is performed by extrapolating the outcome of hazard identification from certain source chemicals to one or more target chemicals based on “similarity” between the source compound(s) and the target compound [11] and it does not involve the development of supervised learning models. Both of these approaches are mainly used for two purposes: 1) to predict end point values of a completely new set of chemicals for the purpose of filling data gaps (predictive models) and 2) mechanistic and physicochemical interpretation of the structural features in a molecule which are responsible to elicit the response [12].

In the recent past, efforts have been made to predict the binding affinity of various EDCs to the androgen receptors using computational approach. *Hong et al.* [13] in 2003 studied the binding affinity of natural, synthetic and environmental chemicals to the androgen receptor by Comparative Molecular Field Analysis (CoMFA) (a 3D-QSAR approach), and they inferred that the steric and electronic properties of the training compounds are essential in describing the binding affinity of EDCs to the androgen receptor. In 2002, *Serafimova et al.* [14] studied the active formulation ingredients of pesticides and their ability to bind to the androgen receptor and performed their evaluation using COREPA method. They have utilized stereochemical properties like the inter-atomic distances between the nucleophilic sites and their charges and used them to predict the binding affinity in terms of pK_i. *Piir et al.* [15] in 2020 performed binary and multi-class classifications for antagonists, agonists and binders to the AR by implementing random forest classification models. They stated that the accuracy obtained in their multi-class classification was good considering the large size of the training set that they have utilized.

3D-QSAR methods involve computational complexity of conformational analysis and alignment and inherit the property of being non-reproducible in nature. The novelty of the current work is predicting the binding affinity of endocrine disruptors to the androgen receptors in a quantitative and reproducible manner. The data was obtained from Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-research/bioinformatics-tools/endocrine-disruptor-knowledge-base) thus avoiding personal, systemic or instrumental error in data collection. It was then divided into a modeling set and a validation set based on the availability of experimental response values in terms of log RBA, where RBA stands for Receptor Binding Affinity. A regression-based 2D-QSAR model was generated using the modeling set, and subsequently similarity-based chemical Read-Across was also performed. The reliability of both of the approaches was evaluated using various strict validation metrics. The physicochemical interpretation of different possible mechanisms influencing the binding of EDCs to the androgen receptor were also discussed and reported which can ultimately help a chemist recognize the features in a molecule that has potential to cause androgen receptor toxicity. In support of this theory, pharmacophore mapping was also performed to serve the purpose of screening of the features in a molecule which contribute to AR binding affinity. Analysis of the binding of the ligand to the various amino acid residues in the receptor was also done with the help of molecular docking and the stability of such binding was evaluated using molecular dynamics (MD) simulation at 100 ns.

2. Materials and methods

2.1 Collection of Androgen Receptor Binding Affinity data of EDCs and curation of their structures

The androgen Receptor Binding Affinity (RBA) data of various EDCs were collected from the Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-research/bioinformatics-tools/endocrine-disruptor-knowledge-base) obeying the strict OECD guidelines. The chemical structures downloaded from PubChem database (https://pubchem.ncbi.nlm.nih.gov/) in .sdf format were represented in Marvin Sketch (https://chemaxon.com/products/marvin) software. Chemical curation of our compounds was performed by the application of a KNIME workflow (https://sites.google.com/site/dtclabdc/) taking the single .sdf file as input. Further details are available in **Supplementary Material SI-1.**

2.2 Calculation of molecular descriptors and data pre-treatment

Descriptors are certain properties in a molecule encoded in numerical terms which can be handled statistically. Two molecules are said to be "identical" or 100% similar if they have identical set of descriptor values. The descriptors for our curated compounds were calculated using alvaDesc v2.0.6 [16]. To enhance simplicity in the interpretation of the developed model, we have used only selected classes of descriptors **(Supplementary Material SI-1).** The inter-correlated descriptors having correlation values >0.95 and variance cut-off 0.00001 were removed using the Java-based tool Data Pretreatment GUI 1.2 available from https://dtclab.webs.com/software-tools.

2.3 Dataset division and model development

Dataset division into training and test sets during a QSAR model development ensures the models' predictive ability. In the present study, the available data set was segregated into two classes: 1) the modeling set which comprises the compounds having reported response values in terms of log RBA and 2) the validation set consisting of compounds for which the response values were not reported. We have eliminated six compounds from our modeling set due to their aberrant nature of activity. The reduced modeling set was taken as an input for the java-based software tool DTC-QSAR v1.0.5 (https://dtclab.webs.com/software-tools), where we performed division into training and test sets in 70:30 ratios based on Euclidean Distance method [17], and feature selection was done by employing Genetic Algorithm technique [18]. The descriptors obtained from the set of GA-MLR models were then pooled and the best descriptor combinations from all possible models were obtained by using Best Subset Selection (BSS) v2.1 available from https://dtclab.webs.com/software-tools. The Best Subset Selection tool generates models based on all possible combination of descriptors, and one can select the best models based on validation metrics like r^2, Q^2_{LOO}, $MAE_{95\%}$, Q^2_{F1} and Q^2_{F2}. To nullify the inter-correlation among descriptors, the final Partial Least Squares (PLS) regression model was obtained with the best descriptor combination taking three latent variables, and various internationally accepted validation metrics were calculated [19-20] (**Supplementary Material SI-1**).

2.4 DModX Applicability Domain Plots

The Applicability Domain can be termed as a theoretical region in chemical space which surrounds both the descriptors and response [21]. The distance to model in X-space (DModX) approach was implemented to check the applicability domain of the model.

2.5 Similarity based Read-Across prediction

What differentiates Read-Across approach from classical QSAR is that Read-Across is entirely a similarity-based approach which does not involve the development of a statitstical model. QSAR models become statistically unreliable when there are limited number of data points [11] and contrastingly, read-across approach not being a hardcore statistical approach tends to yield better results even for small datasets and thus can be aimed for data gap filling. In the present work, after performing feature selection, we have divided the training set compounds into sub-training and sub-test sets based on Euclidean distance-based division. These sets were further used for hyperparameter optimisation in the Read-Across-v3.1 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) tool. The optimised hyperparameters were then used for the original training and test set files as input.

2.6 3D-Pharmacophore mapping

In this investigation, 3D-Pharmacophore mapping was implemented to explore the potential features that are crucial for the interaction at the active site of the androgen receptor. The receptor binding affinity (RBA) expressed as logRBA was used as the dependent variable to develop the pharmacophore models. Molecules prepared for the 2D-QSAR model development were used for this study. The dataset was rationally divided into training (30 compounds for hypothesis development) and test (115 compounds for validation) sets based on the logRBA values spanning four orders of magnitude. 3D-Pharmacophore modeling was performed using HypoGen algorithm as embedded in Biovia Discovery Studio Client 4.1 client [22] following the protocol as discussed by *Kumar et al.* [23]. Details of the protocol performed for 3D-Pharmacophore modeling is provided in **Supplementary Material SI-1**. Validation of the obtained models was executed using different parameters such as cost analysis, the Fischer randomization test (F-test), and test set prediction to evaluate the robustness and predictive ability of models as discussed by *Kumar et al.* 2020 [23].

2.7 Molecular docking study

Molecular docking study was performed to predict the potential of complex formation and explore the binding mode of the compounds showing the highest and lowest binding affinity to the androgen receptor. The crystal structure of the protein was extracted from the protein databank by the PDB ID: 3G0W [24] (available from https://www.rcsb.org/structure/3G0W). A rigid docking approach was applied using the CDOCKER with a grid-based protocol [25] for the aim of the receptor-ligand interaction, as prompted in Biovia Discovery Studio Client 4.1 client [22] following the protocol as discussed by *Kumar et al.* [23]. Details of the protocol performed for molecular docking is provided in **Supplementary Material SI-1**. After molecular docking, the docked inclusion complexes with the best ranked CDOCKER interaction energy and bond formation between compounds and active amino acid residues were chosen for the detailed interpretation and correlation. We have also validated the docking protocol by redocking the bound ligand at the protein's active site (**Figure S2**) (**Supplementary material SI-1)** and calculating the RMSD (**Figure S3**) **(Supplementary**

material SI-1) with the bound ligand and the redocked ligand. The ligplot shows the number of interactions and active amino acids responsible for the important interaction in the crystal structure of androgen receptor and with their bound ligand.

2.8 Molecular dynamics (MD) simulation and MM/GBSA-Binding energy calculation

Further to study the stability of ligand-receptor complex at biological conditions, molecular dynamics simulation at 100ns was performed [26-29], and receptor binding affinity using MM/GBSA [30] method was calculated.

The whole workflow of multiple cheminformatic applications applied to the ARB data set has been depicted pictorially in **Figure 1**.

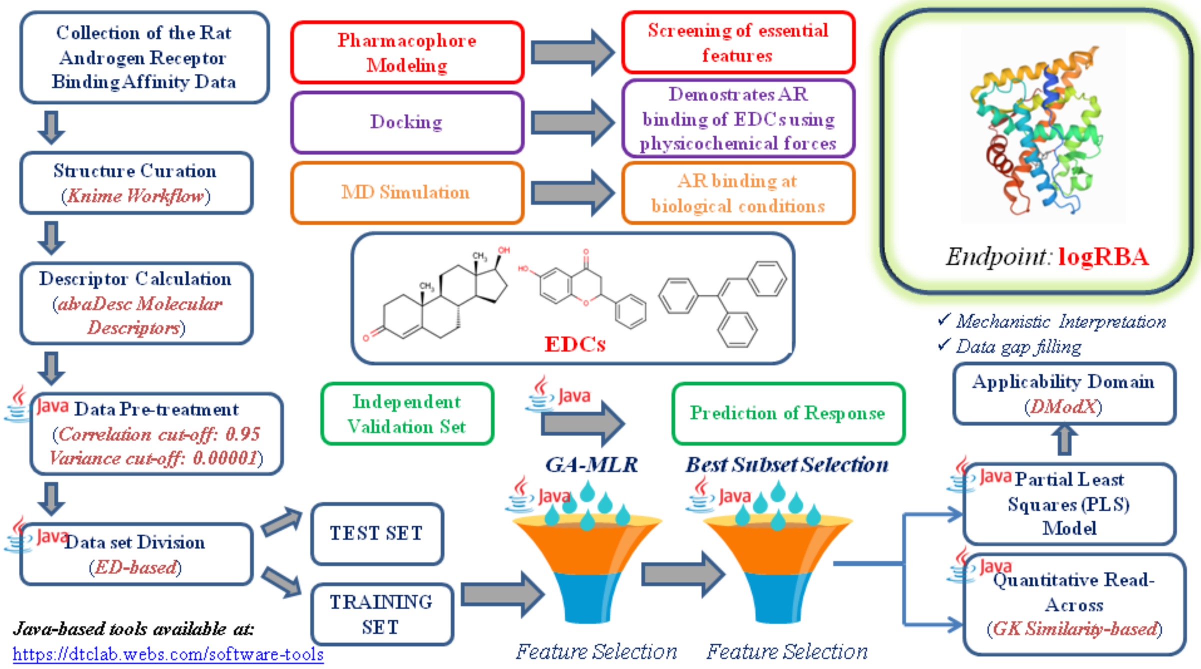
Collection of the Rat Androgen Receptor Binding Affinity Data
Structure Curation (Knime Workflow)
Descriptor Calculation (alvaDesc Molecular Descriptors)
Java Data Pre-treatment (Correlation cut-off: 0.95 Variance cut-off: 0.00001)
Java Data set Division (ED-based)
Pharmacophore Modeling
Docking
MD Simulation
Screening of essential features
Demostrates AR binding of EDCs using physicochemical forces
AR binding at biological conditions
EDCs
Endpoint: logRBA
✓ Mechanistic Interpretation
✓ Data gap filling
Applicability Domain (DModX)
Independent Validation Set
Java
Prediction of Response
GA-MLR
Best Subset Selection
Java
Java
TEST SET
TRAINING SET
Feature Selection
Feature Selection
Java Partial Least Squares (PLS) Model
Java Quantitative Read-Across (GK Similarity-based)
Java-based tools available at:
https://dtclab.webs.com/software-tools

Figure 1. Schematic representation of the workflow of cheminformatic applications used in this study.

3. Results & Discussion

3.1 2D-QSAR analysis

The modeling data set has been provided in an Excel sheet in the **Supplementary Material SI-2**. The training set comprises 103 EDCs that were used for model development while the test set comprises 44 EDCs that were used for prediction and external validation. The final PLS equation with three Latent Variables is shown in Eq. (1). The descriptors have been mentioned in the descending order of importance as per the Variable Importance Plot (**Figure 2**).

$$LogRBA = -3.23 + 0.49 \times SsssCH - 0.41 \times MaxaaCH + 0.23 \times nCconj + 0.35 \times LogP99 - 0.17 \times F10[C-O] + 0.06 \times minsOH + 0.06 \times N\% + 0.67 \times F08[O-F] \tag{1}$$

$$R^2_{(TRAIN)} = 0.74, Q^2_{(LOO)} = 0.68, Q^2_{F1} = 0.58, Q^2_{F2} = 0.58$$

$$Scaled\ average\ r^2_m(Train) = 0.57, Scaled\ average\ r^2_m(Test) = 0.50$$

$$Scaled\ delta\ r^2_m(Train) = 0.18, Scaled\ delta\ r^2_m(Test) = 0.07$$

$$MAE_{(TRAIN)} = 0.46, MAE_{(TEST)} = 0.54, n_{(Training)} = 103, n_{(Test)} = 44$$

The statistical quality and internal and external validation metric values of the QSAR model are satisfactory considering the diversity and heterogeneity of the data set. The descriptors selected in the QSAR model are detailed below **(Figure S4 in Supplementary Materials SI-1).** The different plots [10] related to the PLS model are provided in **Figures S5-S9 in Supplementary Materials SI-1**.

3.1.1 Descriptors contributing to Hydrophobicity

In the final PLS model, we have obtained a total set of 6 descriptors contributing positively to the response out of which some are responsible for directly influencing the hydrophobic properties of the molecules (e.g., LOGP99, SsssCH, nCconj, F08[O-F]) while some induce hydrophobicity indirectly (e.g., minsOH). The **SsssCH** descriptor stands for sum of E-states of sssCH (tertiary carbon atoms) [31]. In this data set, the compounds containing a sterioidal (cyclopentanoperhydrophenathrene) nucleus shows higher values for this descriptor. This suggests that for a higher receptor binding affinity, presence of the steroidal nucleus is preferred. The present dataset includes 5α-Androstan-17β-ol (**23**) which has a higher SsssCH descriptor value and shows a higher receptor binding affinity, as compared to 4-Hydroxybiphenyl (**148**) which is devoid of tertiary carbon atoms (**Figure 2**). The **nCconj** descriptor signifies the number of non-aromatic conjugated carbons (sp^2), and it positively correlates with the response values as in the case of Trenbolone (**157**), which has a higher number of non-aromatic conjugated carbon atoms (sp^2) thus resulting in enhanced receptor binding affinity while in case of Aldrin (No. **176**), where the sp2 carbons are not in conjugation, exhibit a much lower receptor binding affinity. In the steroidal structures of the data set, the descriptor nCconj actually signifies the importance of the conjugated enone moiety in ring A (like **67**), as the keto group at 3 position serves as an hydrogen bond acceptor (see molecular docking in a later section). The descriptor **LOGP99** stands for Wildmann-Crippen octanol-water partition coefficient, and it positively contributes to the response values, as an increase in the o/w partition coefficient value increases the lipid solubility. For instance, Dihydrotestosterone benzoate (**134**) has a high LOGP99 value, and thus has a higher receptor binding affinity compared to Diethyl phthalate (**34**) which has a lower partition coefficient value. The descriptor **minsOH** stands for minimum E-state of the sOH hydroxyl group [31]. This can be attributed to the inherent property of the hydroxyl groups to be able to form hydrogen bond interactions with the receptor residues in an appropriate location [32] and thus contributes to the enhancement in the receptor binding affinity of the molecule. A higher minsOH value also signifies that there is a large hydrophobic moiety attached to the hydroxyl group, thus bulkiness of the structure also contributes to the overall hydrophobic property. The presence of OH group at a desired location as well as its attachment to a bulky moiety in Norgestrel (**67**) is the reason for its high receptor binding affinity whereas the molecule Aldrin (**176**) lacks the hydroxyl group and results in lower receptor binding affinity. **N%** denotes the percentage of nitrogen present in the molecular structure, and it shows a positive contribution to the response. In a previous work, *Zhou et al.* stated that nitrogen in the form of primary amino group can be accommodated in the same location as the hydroxyl group (probably due to the bio-isosteric nature of O and NH) and thus can actively participate in hydrogen bonding with the receptor residues like Asn705 [33] resulting in enhanced receptor binding affinity, as also demonstrated in our model. Due to the presence of Nitrogen in Carbaryl (**72**), it exhibits slightly higher receptor binding affinity than Bis(n-octyl) phthalate (**114**) which is devoid of nitrogen atoms. The descriptor **F08[O-F]** stands for frequency of O and F atoms at the topological distance of 8. The presence of F atoms can induce polarity, but the presence of O at the topological distance of 8 suggests that the compounds are bulky in nature, thus overshadowing the polar effects with the hydrophobic properties contributed due to bulkiness of the structure. Presence of a lipophilic -CF3 group in Hydroxyflutamide (**187**) ensures higher receptor binding affinity while 17α-Estradiol (**7**) is devoid of CF_3 atoms and does not tend to bind well to the receptor.

3.1.2 Descriptors contributing to Polarity and Electron Richness

Out of the total 8 descriptors obtained in our model, two of them correlate negatively to the response values and induce polarity and electron richness to the molecules. One of the descriptors is **MaxaaCH,** which stands for maximum E-state of aaCH (aromatic CH groups) [31]. This is probably due to the fact that aromatic compounds are comparatively more polar than their alicyclic counterparts. This can be observed in 3-methyl-estriol (**102**) with an aromatic ring showing reduced receptor binding affinity as compared to 3β-Androstanediol (**183**) which is devoid of any aromatic ring and thus exhibiting higher receptor binding affinity. The other descriptor is **F10[C-O]** which stands for frequency of C and O at the topological distance 10. This descriptor depicts the presence of polar functionalities like hydroxyl, ether or ester groups. It is to be noted that the hydroxyl group as minsOH contributes positively to the receptor binding affinity due to its ability to form a hydrogen bond at a desired location while attached to a bulky scaffold. Therefore, it can be concluded that F10[C-O] descriptor actually acts to compensate that effect with the polar effects of OH and this can be confirmed with the near-equal and opposite values of the standardized coefficients of both these descriptors in our PLS model. Our dataset contains Dexamethasone (**75**) which shows lower receptor binding affinity than Triphenylethylene (**4)** as the latter lacks polar functionalities like hydroxy, ether or ester groups. The hydrogen bond donor group should be present at a specific position like 17 position of the steroidal nucleus as in 5α-Androstan-17β-ol (**23**) to participate in the hydrogen bonding interaction with the receptor functionalities (see Molecular Docking in a later section). Presence of polar functionality at any other locations decrease the RBA.

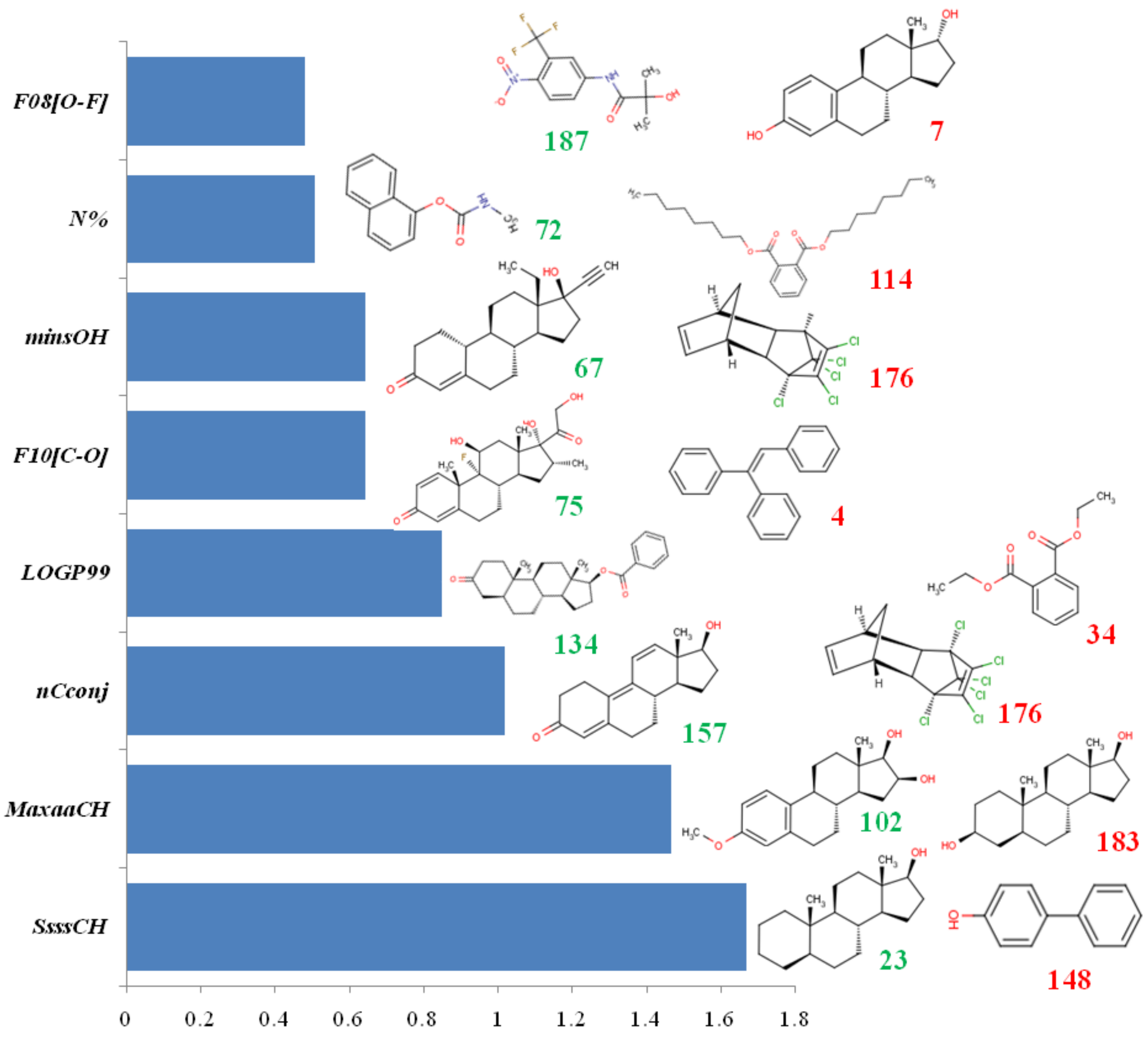

Figure 2. Variable Importance Plot. Structures of representative compounds having higher and lower values of individual descriptors are also shown

3.1.3 Predictions for the validation set

Prediction of the receptor binding affinities for the compounds which constitute the validation set was performed using a java-based software tool Prediction Reliability Indicator (PLS Version) [34] available from https://dtclab.webs.com/software-tools. The results obtained depicts that out of the 55 compounds, 12 were outside the applicability domain with Bad/Unreliable prediction quality and among the remaining 43 compounds, two of them have moderate prediction quality and the others have good prediction quality. The results of this prediction is provided in an Excel sheet in the **Supplementary Material SI-2**.

3.2 Chemical Read-Across results

After QSAR model development, the same training and test set compounds were taken as inputs for quantitative Read-Across-based predictions using the same input features as descriptors, while implementing three different similarity functions: the Euclidean Distance-based, the Gaussian Kernel Similarity-based and the Laplacean Kernel Similarity-based predictions, and after optimization of the hyper-parameters, it was found that the external validation results obtained from quantitative Read-Across algorithm using Gaussian Kernel Similarity-based functions were better compared to the results obtained using QSAR and also the other two read-across approaches **(Figure S10) (Supplementary Material SI-1)**. The higher values of Q^2_{F1} (0.64), Q^2_{F2} (0.64) and lower $MAE_{(TEST)}$(0.47) in Read-Across suggest that predictive ability of the Read-Across algorithm was even better for predictions as compared to the classical QSAR approach. It appears that the local similarity-based approach gives better predictions over model-derived predictions obtained from the whole training data set. The results of this prediction is provided in an Excel sheet in the **Supplementary Material SI-2**.

3.3 Comparison of present 2D-QSAR and Read-Across with previous models

We have developed here an easily reproducible and transferable 2D-QSAR model using simple interpretable descriptors. *Hong et al.* [13] employed Comparative Molecular Field Analysis (CoMFA) (a 3D-QSAR approach) by taking similar number of data points and the corresponding quality and validation metrics were $r^2 = 0.902$ and $q^2 = 0.571$ which suggests that their model is less robust due to a high difference between r^2 and q^2 values. Also it is important to note that CoMFA methodology requires conformation analysis and alignment of the molecules making the results less reproducible. *Piir et al.* [15] applied binary and multi-class classification techniques generating only qualitative results whereas our model generates quantitative predictions. Thus, it can be concluded that our model is robust, predictive (due to acceptable values of the external validation metrics) and reproducible. **Table 1** depicts how our QSAR model and Read-Across based predictions supersedes the previous results in the quantitative prediction quality.

Table 1: Comparison with the previous studies

Authors	**Method**	$n_{(Train)}$	$n_{(Test)}$	**End Point**	R^2	Q^2	Q^2_{F1}	Q^2_{F2}	**Inference**
***Hong et al.* [13]**	3D-QSAR (CoMFA) (Regression)	146	8	logRBA	0.90	0.57	-	-	Less robust, non-reproducible
***Piir et al.* [15]**	Classification-based QSAR	1688	5273	AR Activity	-	-	-	-	Graded predictions only
Our work	2D-QSAR (Regression)	103	44	logRBA	0.74	0.68	0.58	0.58	Robust, Predictive, Reproducible
Our work	Quantitative Read-Across	103	44	logRBA	-	-	0.64	0.64	Predictive, Reproducible

3.4 3D Pharmacophore modeling analysis

In this analysis, we have developed ten different 3D- pharmacophore hypotheses from a training set of 30 compounds. The robustness of the generated models in terms of fitness, stability, classical fitness metrics, and predictability was examined using stringent validation metrics. In terms of internal validation, all the developed models were showing excellent results, thus for the selection of the best hypothesis, we have checked the performance on the test set. External validation of the developed models was implemented by mapping the test set compounds with the same settings applied for the pharmacophore generation by the FAST method. After analysis (**Table S1**) **(Supplementary Material SI-1),** Hypo-8 was found to be the best one among the ten hypotheses with one Hydrogen bond acceptor (HBA), two Hydrophobic (HYD), and one Hydrogen bond donor (HBD) features (**Figure S11**) **(Supplementary Material SI-1).** In terms of internal validation, the best pharmacophore model (Hypo 8) was obtained (**Table S1**) (**Supplementary Material SI-1)** in the cost analysis with a higher correlation coefficient (R: 0. 757), total cost (329.866), maximum fit (10.809), configuration cost (12.097) and higher cost difference (287.88). These values stated that the selected model was appropriate in terms of internal quality metrics. After mapping, we found that 27 compounds from the data set were correctly mapped and predicted, whereas 88 compounds were not mapped due to the absence of features found in the select pharmacophore model. Out of these 88 compounds, 79 compounds have the ARB affinity lower than the training set mean suggesting that these are low affinity compounds due absence of the required pharmacophoric features (and hence not mapped). The observed and predicted values of the training and test set molecules obtained from the analysis using Hypo-8 are given in **Sheets 2 and 3 (Supplementary Material SI-3).** We have developed a Java-based software tool **Klassification1.0** for calculating the classification metrics and the tool is now made available online at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The test set statistics are based only on the mapped compounds. The Fisher validation test confirms the non-randomness of the selected pharmacophore (Hypo-8) model. The total correlation and cost values obtained from the original and randomized models of the hypothesis for the Fisher validation test are stated in **Sheets 4 and 5** in the **Supplementary Material SI-3**. Additionally, the validated pharmacophore model was used to estimate the affinity of the external dataset of 55 compounds, with no quantitative observed response values in the source file. After prediction, we have found that only 13 compounds were correctly mapped and predicted, whereas 42 compounds were not mapped due to the absence of features found in the select pharmacophore model, out of the listed 55 compounds (see **Sheet 6** in **Supplementary Material SI-3**). We have also predicted the 6 compounds omitted from the original dataset because of their outlier behavior in the initial modeling (2D-QSAR) exercises. After prediction, we have found that only 4 compounds were properly mapped and estimated and whereas 2 compounds were not mapped due to the absence of features found in the select pharmacophore model (see **Sheet 7** in **Supplementary Material SI-3**).

3.5 Molecular docking analysis

3.5.1 Molecular docking analysis of the compounds with the highest and lowest binding affinities from the dataset

We have implemented the molecular docking using the three compounds with the highest ARB (compound **157, 193,** and **207**) and three compounds with the lowest ARB (compound **34, 87,** and **114**) from the whole dataset, to explore the potential interactions at the active pocket of androgen receptor. The detailed information of docking interactions, CDOCKER interaction energy, and their correlation with the features derived from the developed best 2D-QSAR model are illustrated in **Table S2** in **Supplementary Material SI-1.**

3.5.1.1 Molecular docking analysis of the compounds with the highest binding affinity from the dataset

One of the highest ARB compounds from the dataset is compound **157,** which interacted with the active site pocket of the receptor (**Figure 3**) *via* hydrogen bonding with the amino acid residues ASN A: 705, ARG A: 752 in the distance of 1.95, 2.79 and 2.20 Å respectively, π-alkyl hydrophobic bond with amino acid residue PHE A: 764 in the distance of 5.34 Å, and alkyl hydrophobic bonding with the amino acid residues LEU A: 704, MET A: 742, MET A: 745, LEU A: 873 in the distance of 4.81, 4.73, 5.21, 5.36, 4.40, 5.11, 5.10 Å respectively.

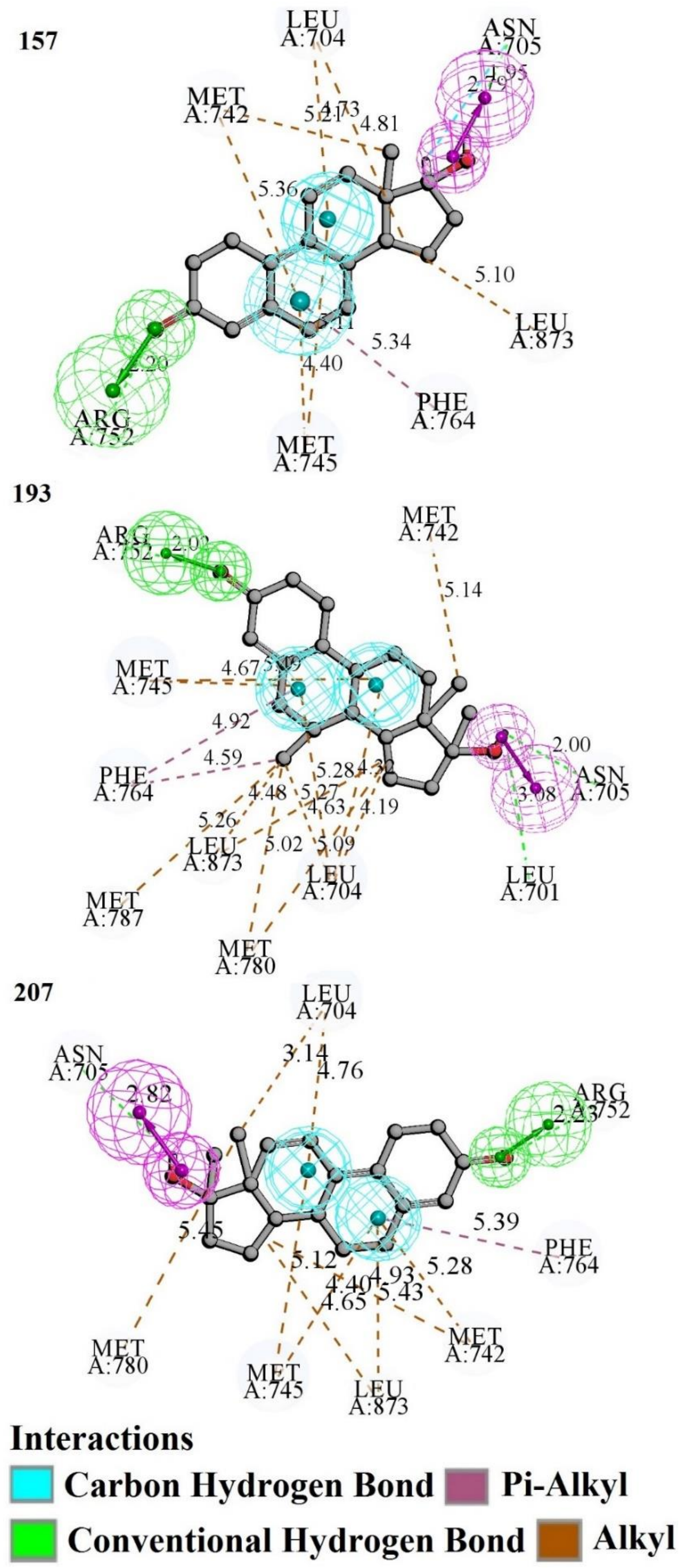

Figure 3. Molecular docking interactions and correlation with pharmacophore model of the compound with the highest binding affinity (Compound **157, 193, 207**) from the dataset.

The next highest ARB compound in this series from the dataset is compound **193**, which interacted with the active site pocket of the receptor (**Figure 3**) *via* hydrogen bonding with the amino acid residues ARG A: 752, ASN A: 705, LEU A: 701 in the distance of 2.02, 2, 3.08 Å respectively, π-alkyl hydrophobic bond with amino acid residue PHE A: 764 in the distance of 4.92, 4.59 Å, and alkyl hydrophobic bonding with the amino acid residues MET A: 742, MET A: 745, MET A: 787, MET A: 780, LEU A: 873, LEU A: 704 in the distance of 5.14, 4.67, 5.49, 5.26, 5.02, 5.09, 4.48, 5.27, 4.63, 5.28, 4.32, 4.19 Å respectively.

The third highest ARB compound from the dataset is **207,** which interacted with the active site pocket of the receptor (**Figure 3**) *via* hydrogen bonding with the amino acid residues ASN A: 705, ARG A: 752 in the distance of 2.82, 2.23 Å respectively, π-alkyl hydrophobic bond with amino acid residue PHE A: 764 in the distance of 5.39 Å, and alkyl hydrophobic bonding with the amino acid residues LEU A: 704, MET A: 780, MET A: 745, LEU A: 873, MET A: 742 in the distance of 3.14, 4.76, 5.45, 5.12, 4.40, 4.65, 4.93, 5.43, 5.28 Å respectively.

The molecular docking analysis of the compounds with the lowest binding affinity from the data set is given in **Figures S12-S14 in Supplementary Material SI-1**. The results of molecular dynamic simulation are also given in **Supplementary Material SI-1**.

3.6 Correlation of the 3D-pharmacophore model with the molecular docking analysis, 2D QSAR, and Read-across models

We have mapped the highest and least ARB compounds from the data set using the selected pharmacophore model (Hypo 8) and superimposed the mapped highest ARB compounds in the pharmacophore with its docking interaction showing important amino acids (**Figure 3**). From **Figures S15 and S16 (Supplementary Material SI-1)** we can see that the highest ARB compounds of the dataset set **157** (logRBA: 2.05) and **193** (logRBA: 2.27) mapped entirely on Hypo-8 with all of the three features appearing in the model. From **Figures S15, S16, and 3** we can see that B and C rings of the steroid nucleus lie in the hydrophobic region and interact with hydrophobic amino acids (MET A: 745, PHE A: 764, MET A: 742, LEU A: 704) via alkyl and π-alkyl bonding (hydrophobic bond), ketone group is in the hydrogen bond acceptor region, interacting with the ARG A: 752 amino acid by hydrogen bond and hydroxy group lies in the hydrogen bond donor region, interacting with ASN A: 705, LEU A: 701 amino acids via hydrogen bond. These features are well corroborated with the SsssCH, nCconj, LOGP99, and minsOH descriptors of the 2D-QSAR models and Read-across hypotheses. On the other hand, the least ARB compounds of the dataset set do not map entirely due to the lack of hydrogen bond donor feature in the case of compound **34** (logRBA: -3.44) (**Figure S17** in **Supplementary Materials SI-1**) and hydrogen bond acceptor in case of compound **92** (logRBA: -3.15) (**Figure S18** in **Supplementary Materials SI-1**). Thus, we can conclude from the above discussion that the absence of any of these three features in compounds reduces the receptor binding affinity against androgen receptor.

.

4. Overview and Conclusion

This study reports a highly robust, reproducible, easily interpretable and sufficiently predictive regression-based 2D-QSAR model which is developed in accordance to the OECD guidelines. This model predicts that various structural features like o/w partition coefficient, bulkiness of the structure, presence of a steroid (cyclopentanoperhydrophenanthrene) nucleus, number of non-aromatic conjugated carbon (sp^2) and hydrogen bonding to the specific receptor residues contribute positively to the receptor binding affinity leading to the toxicity while features like aromaticity in a molecule and presence of polar functionalities like hydroxyl, ether or ester groups at additional locations in the structures lower receptor binding affinity. The similarity-based Quantitative Read-Across approach was also implemented according to the Gaussian-kernel similarity function using an java-based software tool, and it was found that the predictive ability of the Read-Across approach supersedes that of the QSAR approach as the external validation metrics were slightly better in the Read-Across based predictions. The response values of our validation set were calculated using the Prediction Reliability Indicator tool (https://dtclab.webs.com/software-tools) thus making a successful attempt to data gap filling. Pharmacophore mapping was done to screen the essential features, and it was found that a hydrogen bond acceptor, two hydrophobic and one hydrogen bond donor features are essential for receptor binding affinity. This information was supported by performing molecular docking analysis and it was found that the molecules having highest receptor binding affinity possess all the three different features that our pharmacophore hypothesis suggested. Furthermore, the docking results explained the possible amino acid residues present at the surface of the androgen receptor interacting with the compounds resulting in greater receptor binding affinity of the ligand. Additionally, to demonstrate the receptor binding at the biological conditions, Molecular Dynamics Simulation was performed. We believe that our developed QSAR model and read-across approach will be useful in the screening of compounds with lower androgen receptor binding affinity and will possibly tend to reduce environmental hazards.

Author contributions

AB: computation, validation, software tool development, initial draft, PD: computation, validation and editing, VK: computation, validation and initial draft, SK: computation, initial draft, editing, KR: conceptualization, supervision and editing

Declaration of Competing Interest

The authors declare that there are no competing interests.

Acknowledgements

AB conveys his sincere gratitude to Jadavpur University, Kolkata for a scholarship. PD and VK thanks the Indian Council of Medical Research, New Delhi for Senior Research Fellowships. KR thanks SERB, New Delhi for financial support under the MATRICS scheme (MTR/2019/000008).

References

1) Fang, H., Tong, W., Branham, W.S, Moland, C.L., Dial, S.L., Hong, H., Xie, Q., Perkins, R., Owens, W., Sheehan, D.M., 2003. Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. Chem. Res. Toxicol. 16, 1338-1358

2) Falco, M.D., Forte, M., Laforgia, V., 2015. Estrogenic and anti-androgenic endocrine disrupting chemicals and their impact on the male reproductive system; Front. Environ. Sci. 3, 1-12

3) Tan, H., Wang, X., Hong, H., Benfenati, E., Giesy, J.P., Gini, G.C., Kusko, R., Zhang, X., Yu, H., Shi, W., 2020. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. Environ. Sci. Tech. 54, 11424-11433

4) Kucheryavenko, O., Vogl, S., Marx-Stoelting, P., Endocrine disruptor effects on estrogen, androgen and thyroid pathways: recent advances on screening and assessment. In: Mantovani, A., Fucic, A. (Eds.), Challenges in Endocrine Disruptor Toxicology and Risk Assessment, Royal Society of Chemistry, London, 2021, 1-24

5) Schug, T.T., Janesick, A., Blumberg, B., Heindel, J.J., 2011. Endocrine disrupting chemicals and disease susceptibility. J. Ster. Biochem. Mol. Bio.. 127, 204-215

6) Khan, K., Roy, K., 2019. Ecotoxicological QSAR modeling of endocrine disruptor chemicals. J. Hazard Mater. 369, 707-718

7) Luccio-Camelo, D.C., Prins, G.S., 2011. Disruption of androgen receptor signaling in males by environmental chemicals. J. Ster. Biochem. Mol. Bio. 127, 74-82

8) Zhao, C.Y., Zhang, R.S., Zhang, H.X., Xue, C.X., Liu, H.X., Liu, M.C., Hu, Z.D., Fan, B.T., 2005. QSAR study of natural, synthetic and environmental endocrine disrupting compounds for binding to the androgen receptor; SAR QSAR Environ. Res. 16 (4), 349-367

9) Davey, R.A., Grossmann, M., 2016. Androgen receptor structure, function and biology: from bench to bedside. Clin. Biochem. Rev. 37(1), 3-15

10) Seth, A., Roy, K., 2020. QSAR modelling of algal low level toxicity values of different phenol and aniline derivatives using 2D descriptors. Aquat. Toxicol. 228, 1-11

11) Chatterjee, M., Banerjee, A., De, P., Gajewicz, A., Roy, K., 2022. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ. Sci.: Nano 9, 189-203

12) Ambure, P., Gajewicz, A., Cordeiro, M.N.D.S., 2019. Roy, K., New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler. integration of data curation, exhaustive double cross-validation and a set of optimal model selection techniques. J. Chem. Inf. Model. 59, 4070-4076

13) Hong, H., Fang, H., Xie, Q., Perkins, R., Sheehan, D.M., Tong, W., 2003. Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. SAR QSAR Environ. Res. 14(5-6), 373-388

14) Serafimova, R., Walker, J., Mekenyan, O., 2002. Androgen receptor binding affinity of pesticide "active" formulation ingredients. QSAR evaluation by COREPA method. SAR QSAR in Environ. Res 13(1), 127-134

15) Piir, G., Sild, S., Maran, U., 2021. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. Chemosphere 262, 128313

16) Mauri, A., alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In: Roy K. (Ed.), Ecotoxicological QSARs. Methods in Pharmacology and Toxicology, Humana, New York, NY, 2020, pp. 801-820.

17) Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H., Tropsha, A., 2012. Does rational selection of training and test sets improve the outcome of qsar modeling?. J. Chem. Inf. Model. 52, 2570-2578

18) Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. J. Chemom. 14, 643-655

19) Roy, K., Mitra, I., 2011. On various metrics used for validation of predictive qsar models with applications in virtual screening and focused library design. Comb. Chem. High Throughput Screen. 14, 450-474

20) Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom. Int. Lab. Sys. 152, 18-33

21) Roy, K., Kar, S., Das, R..N., Understanding The Basics Of QSAR For Applications In Pharmaceutical Sciences And Risk Assessment, Elsevier Inc, NY, 2015

22) Discovery Studio Predictive Science Application | Dassault Systèmes BIOVIA. https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/

23) Kumar, V., Ojha, P.K., Saha, A.,Roy, K.,. 2020. Exploring 2D-QSAR for prediction of beta-secretase 1 (BACE1) inhibitory activity against Alzheimer's disease, SAR QSAR Environ. Res. 31(2), 87-133

24) Nirschl, A. A., Zou, Y., Krystek Jr, S. R., Sutton, J. C., Simpkins, L. M., Lupisella, J. A., & Hamann, L. G., 2009. N-Aryl-oxazolidin-2-imine muscle selective androgen receptor modulators enhance potency through pharmacophore reorientation. J. Med. Chem. 52(9), 2794-2798.

25) Momany F.A., Rone R., 1992. Validation of the general purpose QUANTA® 3.2/CHARMm® force field. J. Comput. Chem. 13(7), 888-900.

26) Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., Lindahl, E., 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 1, 19-25.

27) Zoete, V., Cuendet, V., Grosdidier, A., Michielin, O., 2011. SwissParam: a fast force field generation tool for small organic molecules. J. Comput. Chem. 32, 11, 2359-2368.

28) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. M., Klein, M. L., 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79, 2, 926-935.

29) Chatterjee, S., Maity, A., Chowdhury, S., Islam, M. A., Muttinini, R. K., Sen, D., 2021. In silico analysis and identification of promising hits against 2019 novel coronavirus 3C-like main protease enzyme. J. Biomol. Struct. Dyn. 39(14), 5290-5303.

30) Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A., Moreno, E., 2021. gmx_MMPBSA: a new tool to perform end-state free energy calculations with GROMACS. J. Chem. Theory Comput. 17, 10, 6281-6291.

31) Butina, D., 2004. Performance of Kier-Hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules. Molecules 9, 1004-1009

32) Wahl, J., Smieško, M.; 2018. Endocrine disruption at the androgen receptor: employing molecular dynamics and docking for improved virtual screening and toxicity prediction. Int. J. Mol. Sci. 19, 1784

33) Zhou, W., Duan, M., Fu, W., Pang, J., Tang, Q., Sun, H., Xu, L., Chang, S., Li, D., Hou, T., 2018. Discovery of novel androgen receptor ligands by structure-based virtual screening and bioassays. Genom. Proteom. Bioinform. 16, 416-427

34) Roy, K., Ambure, P., Kar, S., 2018. How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? ACS Omega 3; 11392-11406