Monte-Carlo method based QSAR model to discover phytochemical urease inhibitors using SMILES and GRAPH descriptors.

Kumar Sambhav Chopdar[1], Ganesh Chandra Dash[2], Pranab Kishor Mohapatra[3], Binata Nayak[4] & Mukesh Kumar Raval[3, #, *]

1. Department of Zoology, Rajendra College, Balangir, Odisha 767002, India

2. Department of Chemistry, APS College, Roth, Balangir, Odisha 767061, India

3. Department of Chemistry, C. V. Raman Global University, Bidyanagar, Mahura, Janla, Bhubaneswar, Odisha 752054, India

4. School of Life Sciences, Sambalpur University, Sambalpur, Odisha 768019, India

Present Address: Stone Building, Opposite Mission School, Balangir, Odisha 767001, India

*Author for correspondence

E-mail: mraval@yahoo.com

Phone - +91 94371 10137

Abstract

Urease inhibitors are known to play a vital role in the field of medicine as well as agriculture. Special attention is attributed to the development of novel urease inhibitor with a view to treating *Helicobacter pylori* infection. Amongst a number of urease inhibitors, a large numbers of molecules fail *in vivo* and in clinical trials due to their hydrolytic instability and toxicity profile. The search for potential inhibitors may require screening of large and diverse databases of small molecules and to design novel molecules. We developed a Monte-Carlo method based QSAR model to predict urease inhibiting potency of molecules using SMILES and GRAPH descriptors on an existing diverse database of urease inhibitors. The QSAR model satisfies all the statistical parameters required for acceptance as a good model. The model is applied to identify urease inhibitors among the wide range of compounds in the phytochemical database, NPACT, as a test case. We combine the ligand-based and structure-based drug discovery methods to improve the accuracy of the prediction. The method predicts pIC_{50} and estimates docking score of compounds in the database. The method may be applied to any other database or compounds designed *in silico* to discover novel drugs targeting urease.

Keywords: QSAR . Monte-Carlo method . Urease inhibitors . Phytochemicals . Drug Designing

Introduction

Urease has emerged as a therapeutic target for the design and discovery of antibacterial, antifungal drugs. Urease occurs as a virulent factor in many pathogenic bacteria and fungi [1-5]. Pathogen *Helicobactor pylori (H. Pylori)* is unable to form a colony in the stomach of the host in the absence of urease [6]. Inhibition of urease has also utility in agriculture by preventing soil urease mediated hydrolysis of urea applied as fertilizer [7]. Therefore, search for urease inhibitors has gained momentum due to its application in the field of medicine as well as agriculture.

Several studies reported a number of urease inhibitors of diverse class ranging from urease derivatives to metal complexes [8, 9]. Structure-activity relationship studies have also been reported to optimize the leads of the individual class of compounds [10-13]. However, a large number of molecules with high potency as urease inhibitor fail during *in vivo* and in clinical trials due to their hydrolytic instability and adverse toxicity profile [14]. In order to overcome these hurdles, an intensive large-scale screening of existing as well as designed novel molecules need to be carried out to find ultimately a urease inhibitor of desired pharmacokinetic properties and toxicity free profile. However, the process will be time consuming and high cost incurring if implemented only through experimental methods. Introduction of an *in-silico* screening method prior to experimental validation will drastically reduce the time and the expenditure. There is a need to develop a reliable QSAR model which can predict the potency of a molecule as a urease inhibitor. The model should also provide leads to improve upon the structure of the molecule to enhance its potency and pharmacokinetic properties while diminishing the toxicity.

In the present work, we have developed a Monte-Carlo method based QSAR model using SMILES and GRAPH descriptors taking a database containing a diverse class of urease inhibiting compounds. This ligand-based drug discovery method is combined with structure-

based drug discovery by including docking score with an aim to improve the quality of the method. The method predicts urease inhibition properties of a wide range of compounds included in the phytochemical database NPACT, as a test case. The model may also be applied to other databases and compounds designed *in silico*. The model may be useful in prediction and designing novel urease inhibitors and drastically reduce the number of compounds for synthesis and experimental screening leading to drug discovery.

Materials and Methods

The Database

A database of 436 urease inhibitors was collected from the BindDB database (please refer to the supplementary data file Dataset S1) [15]. The inhibitory activity IC_{50} in nM unit in the database were converted into M and finally to pIC_{50} (-log IC_{50} in M), which were used for QSAR modelling. The structures of the inhibitors were obtained in sdf format from the database. Subsequently, they were converted to SMILES by Open Babel [16]. The dataset was randomly split into four subsets: Training, Invisible Training, Calibration, and Validation. Three such independent splits were generated. They were random, not identical and they had a similar range of pIC_{50}. The splits were used to build up the models. Training, Invisible Training, and Calibration subsets were used for the model building while Validation subset was used for testing the quality of the model. The model was tested on a phytochemical database, NPACT (Naturally occurring Plant-based Anti-cancer Compound-activity-Target database), which is a curated database of 1574 phytochemicals that exhibit anti-cancer activities [17]. Besides anticancer activity, these phytochemicals are also known to have other medicinal applicability.

The SMILES and GRAPH descriptors

Execution of molecular structure by both SMILES and molecular graphs result in a hybrid descriptor that develops a QSAR model with better statistical quality [18]. The hybrid optimal

descriptors used to spring up the model for the urease inhibitors were computed by the following equation

$$^{Hybrid}DCW(T^*, N^*)=^{SMILES}DCW(T^*, N^*)+^{GRAPH}DCW(T^*, N^*) \quad (1)$$

Where DCW is the correlation weight descriptor. *T** is the preferable threshold value for classification of molecular features into active and rare structural attributes (*SAk*). If the frequency of *SAk* in the training set is < *T**, then these attributes were taken to be rare and hence not included in the modelling process. N* is the preferable number of epochs of the Monte-Carlo optimization [18].

The SMILES based optimal descriptors were calculated using the following equation:

$$^{SMILES}DCW(T^*, N^*) = \Sigma CW(Sk) + \Sigma CW(SSk) + \Sigma CW(SSSk) + CW(PAIR) + CW(NOSP) + CW(HALO) + CW(BOND) \quad (2)$$

where CW is the correlation weight for a structural feature extracted from SMILES; Sk, SSk and SSSk are SMILES attributes which contain one-, two-, and three SMILES elements respectively; PAIR reflects the possible combination of atom pairs and/or bonds that are present in the structure together but disconnected from each other; NOSP, a global molecular descriptor related to the presence or absence of nitrogen, oxygen, sulphur and phosphorus atoms; HALO, a global molecular descriptor related to the presence or absence of fluorine, chlorine and bromine atoms; and BOND indicates the presence of double (=), triple (#) or stereo-chemical bonds (@ or @@) in SMILES [17].

Calculation of graph-based optimal descriptors were carried out as per the following equation

$$^{GRAPH}DCW(T^*, N^*)= \Sigma CW(Ak)+ \Sigma CW(0Eck)+ \Sigma CW(1Eck) +\Sigma CW(2Eck) \quad (3)$$

Where Ak is an element i.e. carbon, nitrogen, oxygen, etc. for hydrogen suppressed graph; and 0ECk, 1Eck, and 2Eck are the Morgan's extended connectivity of each vertex [20].

The QSAR models were built using CORALSEA 17 ([21], *CORAL* software available at http://www.insilico.eu/coral). The process involved three steps; (i) search for preferable threshold value and number of epoch with various values of the threshold (T = 1-3), the number of epochs (N = 1–30) and number of probes (P =1- 3) (ii) selection of preferable T* and N* corresponding to the maximum correlation coefficient of calibration set and (iii) calculation of the correlation weight descriptor (DCW) to build up the model with T = T* and N = N*. Finally, predictability of the model with the validation set was estimated which contains compounds that are not included in the process of the building up of the model. The DCW was used to calculate the pIC_{50} value as follows

$$pIC_{50} = C0 + C1 \times DCW(T^*, N^*) \quad (4)$$

Statistical Parameters

Coefficient of determination, r^2 and the coefficient for external cross-validation, Q^2_{ext}, are two usual parameters to estimate the efficiency and stability of a QSAR model. They are defined as

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_{train})^2} \quad (5)$$

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} (Y(exp) - Y(pred))^2}{\sum_{i=1}^{test} (Y(exp) - \bar{Y}(train))^2} \quad (6)$$

where, n is the number of compounds in the dataset (training, invisible training, calibration or test), Y_i are the measured values; $\bar{Y}$ is the averaged value for an overall dataset; $\bar{Y}_{train}$ or $\bar{Y}$(train) is the averaged value of the dependent variable for the training set. Y(exp) is the experimental value of the dependent variable; Y(pred) is the predicted value of the dependent variable.

A model is acceptable when r^2 and Q^2_{ext} are greater than 0.5. However, a rigorous approach to test the quality of a model is made by determining a robust parameter, Δr_m^2.

$$\Delta r_m^2 = |r_m^2 - r^{*}_m{}^2| \quad (7)$$

$$r_{mav}^2 = 0.5\,(r_m^2 + r^{*}_m{}^2) \quad (8)$$

$$r_m^2 = r^2 + (1 - |\sqrt{(r^2 - r_0^2)}|) \quad (9)$$

$$r^{*}_m{}^2 = r^{*2} + (1 - |\sqrt{(r^{*2} - r^{*}_0{}^2)}|) \quad (10)$$

where, r_0^2 is coefficient of determination without intercept; r^{*2}, $r^{*}_0{}^2$, coefficient of determination with and without intercept with inter changing the axes of experimental and predicted values. r_m^2, $r^{*}_m{}^2$, should be greater than 0.5 and Δr_m^2 should be less than 0.2 for an acceptable model [20-24].

The robustness of a QSAR model is further validated using the randomization technique.

$$^{C}r^2_p = r(r^2 - r_r^2)^{1/2} \quad (11)$$

Where r_r^2 is the Y-randomized coefficient of determination. $^{C}r^2_p$ should be greater than 0.5 for a model to be acceptable [20].

Other conditions for acceptability of a QSAR model include slopes of regression lines through origin *k* and *kk*(randomized) should be 0.85 <*k* < 1.15 and 0.85 <*kk*< 1.15; $(r^2 - r_0^2)/r^2$and $(r2 - rr_0^2)/r^2$ should be <0.1.

Index of ideality of correlation, IIC, is proposed to estimate the reliability of prediction [25]. Higher is the reliability of prediction, closer the IIC to 1.0. Mean absolute error (MAE) is utilized in estimation of IIC.

$$MAE = (1/n)\Sigma|Y_{exp} - Y_{pred}| \quad (12)$$

$$\Delta_k = \text{Experimental}_k - \text{Predicted}_k \quad (13)$$

$$^{-}MAE_{calibration} = \frac{1}{^{-}N}\sum_{k=1}^{^{-}N}|\Delta_k| \quad \Delta_k<0,\ ^{-}N \text{ is the number of } \Delta_k<0 \quad (14)$$

$$^{+}MAE_{calibration} = \frac{1}{^{+}N}\sum_{k=1}^{^{+}N}|\Delta_k| \quad \Delta_k\geq 0,\ ^{+}N \text{ is the number of } \Delta_k\geq 0 \quad (15)$$

$$IIC = r_{calibration} \times \frac{\min(^{-}MAE_{calibration}, ^{+}MAE_{calibration})}{\max(^{-}MAE_{calibration}, ^{+}MAE_{calibration})} \quad (16)$$

Applicability Domain

Reliable prediction with the QSAR model is expected only when a compound lies in the applicability domain (AD) of the predictor. The AD of a compound is defined in the present work on the basis of DefectSMILES as follows

$$\text{DefectSMILES} = \sum_{\text{ActiveSA}_k} \text{SA}_k\text{defect} \tag{17}$$

A compound is in the AD if

$$\text{DefectSMILES} < 2 \times \overline{\text{DefectSMILES}} \tag{18}$$

Where, $\overline{\text{DefectSMILES}}$ is average DefectSMILES

Docking

Structure of *H. pylori* urease (1e9y B) was taken as the target. 1253 phytochemicals (MW ≤ 600) from NPACT database were selected as the ligand set. Docking scores were obtained by applying AutoDock Vina algorithm [26] using YASARA molecular modeling suite version 18.4.24 [27]. The ligands and the active site residues Ala169, Thr170, His221, Asp223, Trp224, His248, Met317, Cys321, His322, Arg338, Asp362, and Met366 were allowed to be flexible during docking.

Molecular Dynamics Simulation

Molecular dynamics simulation process was applied for *in silico* validation of the stability of the protein ligand complex. The protein-ligand complex was put in the aqueous medium (TIP3P water model, density 0.997 $g.l^{-1}$, NaCl 0.9% as counter ions) in a cubic simulation box setup with at least 5 Å around the complex molecule under periodic boundary conditions. The system was energy minimized by steepest gradient approach (100 cycles) using AMBER14 force field. The molecular complexes were simulated for 20 ns (production period) with frame capture at every 25 ps step to analyze the trajectory by various evaluation parameters. YASARA suite is used for molecular dynamics study [27].

ADMET study

FAF-Drug4, a free web service (http://fafdrugs4.mti.univ-paris-diderot.fr/), was used to predict the physicochemical and biological properties of the ligand molecules in the database and as a filter to accept the suitable molecules for further drug development studies [28]. FAF-Drugs4 is an ADME-Tox (adsorption, distribution, metabolism, excretion and toxicity) prediction tool.FAF-*Drugs*4 employs pre-defined filters, but users can also customize their own filtering parameters by using the Filter-Editor service. The filters applied in the present study were Drug-like soft, PAINS (Pan-Assay Interference Compounds: A, B and C) [29], and LillyMedChem rules (relaxed). FAF-QED module was run to obtain the quantitative estimates of drug-likeness, which ranges from 0.0 (most unfavourable) to 1.0 (most favourable) [28].

Flow Chart of the Work

The flow chart adopted in the present work is graphically represented in Fig. 1. The experimental validation has not been done in the present work.

Molecular Visualization

Molecular structures are visualized using Biovia Discovery Studio Visualizer 16.1.0 tools. (*https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/visualization-download.php)*

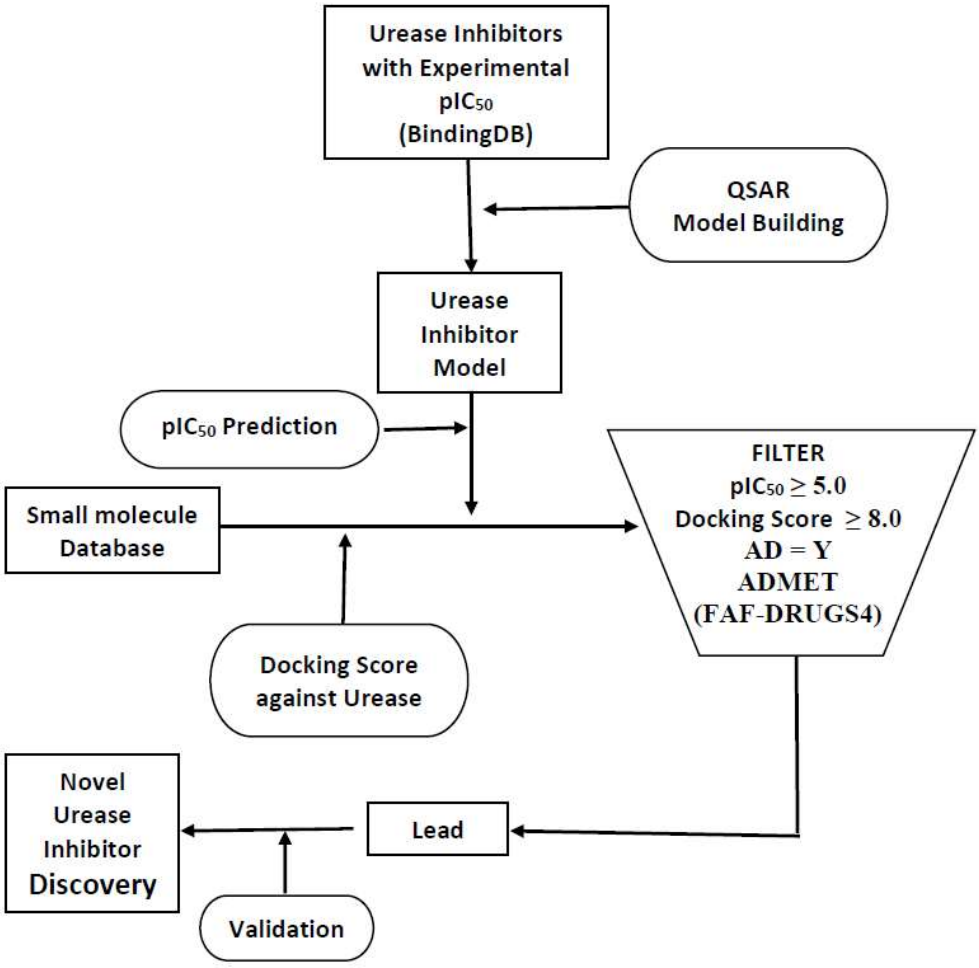

Fig. 1 Process flow for novel urease inhibitor discovery

Results and Discussion

There are 436 compounds in the urease inhibitor database, which were distributed randomly into four categories (training, invisible training, calibration and validation) in three splits. The identities of compounds in four categories distributed in three splits are presented in Table 1. There were no duplicate compounds in the database.

Table 1. Percentage identity of Training (Train), Invisible training (Inv), Calibration (Cal), and Validation (Val) sets in different splits.

	Category	Split 1	Split 2	Split 3
Split 1	Train	100	17.8	13.6
	InvTrain	100	13.6	18.8
	Calib	100	19.6	21.5
	Valid	100	16	21
Split 2	Train	17.8	100	23
	InvTrain	13.6	100	14.4
	Calib	19.6	100	21.5
	Valid	16	100	19.8
Split 3	Train	13.6	23	100
	InvTrain	18.8	14.4	100
	Calib	21.5	21.5	100
	Valid	21	19.8	100

Identity (%) = ($N_{i,j}$/(0.5(N_i+N_j))x100
$N_{i,j}$ is the number of substances which are distributed into the same set for both i-th and j-th split
N_i is the number of substances which are distributed into the set for i-th split;
N_j is the number of substances which are distributed into the set for j-th split.

The statistical parameters obtained by QSAR study employing a Monte-Carlo method using SMILES and GRAPH descriptors applying CORALSEA 17 software are presented in Table 2. The parameters suggested that all the three splits yielded acceptable QSAR models for prediction of pIC_{50} of compounds as urease inhibitors. However, QSAR model from split 1, which had lowest Δr_m^2value (0.05) among the three splits was used for further prediction. End point was estimated as follows.

Endpoint = 3.3576082 (± 0.0065335)+ 0.0277408 (± 0.0001133) * DCW(1,25) (19)

The average of ^{Defect}SMILES = 1.51321

Table 2. Statistical parameters of predictive models with three randomly split datasets.

Statistical parameters	Split 1	Split 2	Split 3
n (Train, Inv, Cal, Val)	118, 117, 102, 99	117, 118, 100, 101	118, 113, 106, 99
r^2(Train, Inv, Cal)	0.838, 0.838, 0.759	0.811, 0.811, 0.814	0.821, 0.821, 0.737
Q^2(Train, Inv, Cal, Val)	0.832, 0.832, 0.750, 0.667	0.804, 0.804, 0.804, 0.745	0.815, 0.814, 0.726, 0.763
s (Train, Inv, Cal, Val)	0.363, 0.402, 0.409, 0.452	0.380, 0.385, 0.371, 0.442	0.374, 0.360, 0.408, 0.426
MAE (Train, Inv, Cal, Val)	0.259, 0.300, 0.309, 0.320	0.281, 0.280, 0.281, 0.340	0.281, 0.258, 0.318, 0.329
F (Train, Inv, Cal, Val)	600, 594, 316, 211	493, 497, 429, 306	534, 509, 292, 375
IIC (Train, Inv, Cal, Val)	0.772, 0.832, 0.871, 0.715	0.827, 0.628, 0.902, 0.836	0.690, 0.889, 0.859, 0.723
$^{C}r^2_p$ (Train, Inv, Cal)	0.836, 0.835, 0.755	0.806, 0.809, 0.813	0.820, 0.816, 0.732
Parameters of Validation			
r^2	0.685	0.755	0.772
r^2_0	0.667	0.714	0.694
rr^2_0	0.642	0.752	0.772
$(r^2 - r^2_0)/ r^2$	0.026	0.055	0.101
$(r^2 - rr^2_0)/ r^2$	0.063	0.005	0.001
k	0.975	0.993	0.982
kk	1.017	0.999	1.011
r^2_m	0.594	0.601	0.556
r^{*2}	0.685	0.755	0.772
r^{*2}_0	0.642	0.752	0.772
rr^{*2}_0	0.667	0.714	0.694
$(r^{*2} - r^{*2}_0)/ r^{*2}$	0.063	0.005	0.001
$(r^{*2} - rr^{*2}_0)/ r^{*2}$	0.026	0.055	0.101
k*	1.017	0.999	1.011
kk*	0.975	0.993	0.982
r^{*2}_m	0.543	0.710	0.758
r^2_{mav}	0.568	0.656	0.657
Δr^2_m	0.051	0.109	0.202

The total numbers of structural attributes (SA) were 879 out of which 790 were active. Top SAk associated with increase and decrease of Endpoint is listed in Table 3.

Table 3. List of top 25 structural attributes associated with increasing and decreasing Endpoints

Increasing Endpoint SA			Decreasing Endpoint SA		
SAk	CW(SAk)	ID	SAk	CW(SAk)	ID
c...(...S...	26.003	695	EC2-Cl..7...	-9.630	389
NOSP00000000	25.249	549	$10000100100	-9.245	22
$00000001000	23.066	5	N...C...1...	-7.434	502
C...(...-...	20.629	211	C...1...=...	-7.185	227
o...(...c...	18.753	833	c...C...N...	-6.630	747
EC2-O...4...	18.498	419	C...N...3...	-6.623	257
=...C...1...	17.940	197	S...N...1...	-6.441	578
c...C...=...	17.377	745	C...\...C...	-6.380	271
C...(...3...	13.379	215	C.../...1...	-6.314	221
F...(...F...	12.809	283	C...[...(...	-6.128	264
++++F---Br==	12.190	69	1...c...1...	-6.005	123
HALO10100000	11.750	471	c...3...F...	-5.311	737
NOSP01100000	11.496	552	4...n...(...	-5.126	172
$00011001010	11.125	18	s...(.....	-5.004	848
C...4.....	11.063	238	C...1.../...	-4.999	225
EC2-N...24..	11.060	405	EC2-s...13..	-4.940	444
EC2-o...16..	10.877	441	F...3.....	-4.939	287
$00011000000	10.876	13	s...(...c...	-4.620	849
C...O...C...	10.753	262	s...1...(...	-4.561	851
C...O...2...	10.253	261	EC1-N...2...	-4.309	327
c...C...(...	10.251	743	c...[...H...	-4.255	766
EC1-O...2...	10.246	336	n...[...n...	-4.249	817
[...c...2...	9.938	655	C...=...1...	-4.186	240
4...c...-...	9.874	170	s...c...c...	-4.183	867
n...[...1...	9.248	816	1...O...(...	-4.000	119

Phytochemicals are known to have diverse medicinal properties [30, 31] and also considered to be safe drugs with low adverse effect [32]. There are several databases, which could be tested to predict urease inhibition property. However, in the present study we select the NPACT database, which contains phytochemicals with known anti-cancer properties. Besides the anti-cancer properties, the phytochemicals in NPACT also have other known (may be unknown) medicinal values. The recent trend is to systematically search for a phytochemical as an alternative remedy to many diseases. Phytochemical urease inhibitors were searched for the treatment of pathogenic infections where urease was considered as a therapeutic target. Also,

a phytochemical urease inhibitor can mitigate the aerial loss of nitrogen from fertilizer urea in an eco-friendly manner [7].

QSAR model from split 1was applied to predict pIC_{50} of 1253 phytochemicals from the NPACT database. The experimental IC_{50} values of EGCG (2.2 μM or pIC_{50}= 5.6575), myricitin (98.7 μM or pIC_{50} = 4.0057), and baicalin (2740 μM or pIC_{50} = 2.5622) had been reported earlier by others and compiled in a review [8]. Ahmed et al [33] reported curcumin as urease inhibitor (IC_{50}19.74μM or pIC_{50} = 4.7046). The values predicted by the present QSAR model were 5.4034, 4.2971, 2.8252, and 5.1964 for EGCG, myricitrin, baicalin, and curcumin respectively. The pIC_{50} values predicated are very close to the experimentally reported values indicating the good predictability of the present model.

Docking scores of phytochemicals in the database were estimated parallelly by AutoDock Vina algorithm [26] using YASARA suite [27]. A filter of "pIC_{50}> 5.0 and docking score >8.0" was applied to obtain a set of compounds, which fall in the applicability domain (AD). These compounds were predicted as strong urease inhibitors. Further, ADMET filter was applied using FAF-Drugs4. Drug-likeness were estimated applying FAF-QED module [28]. Finally, a set of leads that pass through all the filters are presented in Table 4. The details of the parameters predicted by FAF-Drugs4 and FAF-QED are presented in supplementary Tables S1 and S2.

The docking best poses of the lead molecules in interaction with the active site residues are depicted in supplementary Fig. S1. Ni ions (3001, 3002), Gly279, Cys321, His322, Arg338, and Ala365 are common interacting moieties in various docking poses. Cys321, His322 sit on the tip of the flexible flap of the cavity.

The leads were lignans and flavonoids. Flavonoids are reported to have anti- *H. pylori* and urease inhibitory activities [8, 34]. Flavonoids are food supplements without significant

side effects and resistance. Hence may be suitable as a drug against urease as the target. Inophyllum E, curcumin, and (2S)-2'-methoxykurarinone have QED values > 0.5 in the range of 0.0 to 1.0 (Table 4). Therefore, the focus may be put on these three leads for further study.

Table 4. List of lead compounds obtained after filtration.

Compound ID	Name	Class	pIC50 calc	AD	Dock Score	ADMET FAF-Drugs4	QED FAF-QED
21635715	Agastenol	Lignan	5.6008	YES	-8.694	Accepted	0.391
637406	Agastinol	Lignan	5.4721	YES	-8.974	Accepted	0.366
455251	Inophyllum E	Flavonoid	5.3672	YES	-8.995	Accepted	**0.625**
NPACT01531	Vibsanin C	Terpenoid	5.3471	YES	-9.127	Accepted	0.406
969516	Curcumin	Flavonoid	5.1964	YES	-8.203	Accepted	**0.619**
5281810	Tectoridin	Flavonoid	5.1313	YES	-8.548	Accepted	0.245
11982641	(2S)-2'-methoxykurarinone	Flavonoid	5.0960	YES	-8.128	Accepted	**0.507**

PubChem/NPACT Identities of compounds are listed. AD, applicability domain, QED, quantitative estimation of drug-likeness.

Leads were optimized by *in silico* study of derivatives designed by taking clue from the Structural attributes (SAk) increasing Endpoint (Table 3). c….(…S), sp2 carbon branching extension containing sulphur, is the top most attribute contributing to the increase of Endpoint. Besides presence of fluorine in the structure is another high contributing attribute. Including these two attributes in structures of curcumin, inophyllum E, and methoxykurarinone 18 derivatives are designed. The results of pIC50 prediction, FAF Drugs4 filtration and determination of QED are presented in Table 5. The derivative LD10 is the same as curcumin derivative reported by Ahmed et al [33]. It was reported to have IC_{50} value 2.44 μM (pIC_{50} = 5.6126) while the prediction in the present work is somewhat higher, pIC_{50} = 6.3993. Some derivatives with a predicted pIC50 value higher than 6.0 is enlisted (Table 5 and Fig.2). These may serve to improve the design for high potential urease inhibitors from phytochemicals.

However, experimental validation of the proposed leads will have the final say in further drug development.

Table 5. Smiles and properties of sulfur and fluoride derivatives of inophyllum E and curcumin|

Comp ID	SMILES	pIC50	QEDuw	FAFDrugs4
LD1	O1[C@@H](C=c2[nH]c(=S)[nH]c3c2c1c(C[C@@H](CC=C(C)C)C(=C)C)c(O)c3)c1c(OC)cc(O)cc1	5.394	0.348	Accepted
LD2	O1[C@@H](CC(=O)c2c1c(C[C@@H](CC=C(C)C)C(=C)C)c(O)cc2SC)c1c(OC)cc(O)cc1	5.4814	0.392	Rejected
LD3	O1[C@@H](CC(=O)c2c1c(C[C@@H](CC=C(C)C)C(=C)C)c(O)cc2OC)c1c(OC)cc(S)cc1	6.2655	0.404	Rejected
LD4	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1S)(C)C)c1c2oc(=O)cc1c1ccc(cc1)F)C)C	5.8981	0.495	Accepted
LD5	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1S)(C)C)c1c2oc(=O)cc1c1ccccc1)C)C	5.78	0.512	Accepted
LD6	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1)(C)C)c1c2oc(=O)cc1c1ccc(cc1)S)C)C	6.3262	0.518	Accepted
LD7	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1F)(C)C)c1c2oc(=O)cc1c1ccc(cc1)S)C)C	6.4541	0.511	Accepted
LD8	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1F)(C)C)c1c2oc(=O)cc1c1ccccc1)C)C	5.5122	0.617	Accepted
LD9	O1[C@@H]([C@@H](C(=O)c2c1c1c(OC(C=C1)(C)C)c1c2oc(=O)cc1c1ccc(cc1)F)C)C	5.5024	0.605	Accepted
LD10	O(c1cc(ccc1O)/C=C/C1=NC(=S)N=C(C1)/C=C/c1cc(OC)c(O)cc1)C	6.3993	0.633	Accepted
LD11	S(c1cc(cc(c1O)F)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(O)cc1)C	6.211	0.498	Accepted
LD12	O(c1cc(cc(c1O)S)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(O)cc1)C	6.0465	0.431	Accepted
LD13	O(c1cc(cc(c1S)F)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(O)cc1)C	6.3311	0.514	Accepted
LD14	O(c1cc(c(cc1O)S)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(O)cc1)C	6.138	0.431	Accepted
LD15	O(c1cc(ccc1O)/C=C/C(=O)[C@H](C(=O)/C=C/c1cc(OC)c(O)cc1)S)C	5.5634	0.408	Accepted
LD16	O(c1cc(cc(c1S)F)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(S)c(c1)F)C	7.0951	0.477	Accepted
LD17	O(c1cc(ccc1S)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(S)cc1)C	6.5344	0.508	Accepted
LD18	O(c1cc(cc(c1O)F)/C=C/C(=O)CC(=O)/C=C/c1cc(OC)c(O)c(c1)F)C	5.3967	0.592	Accepted

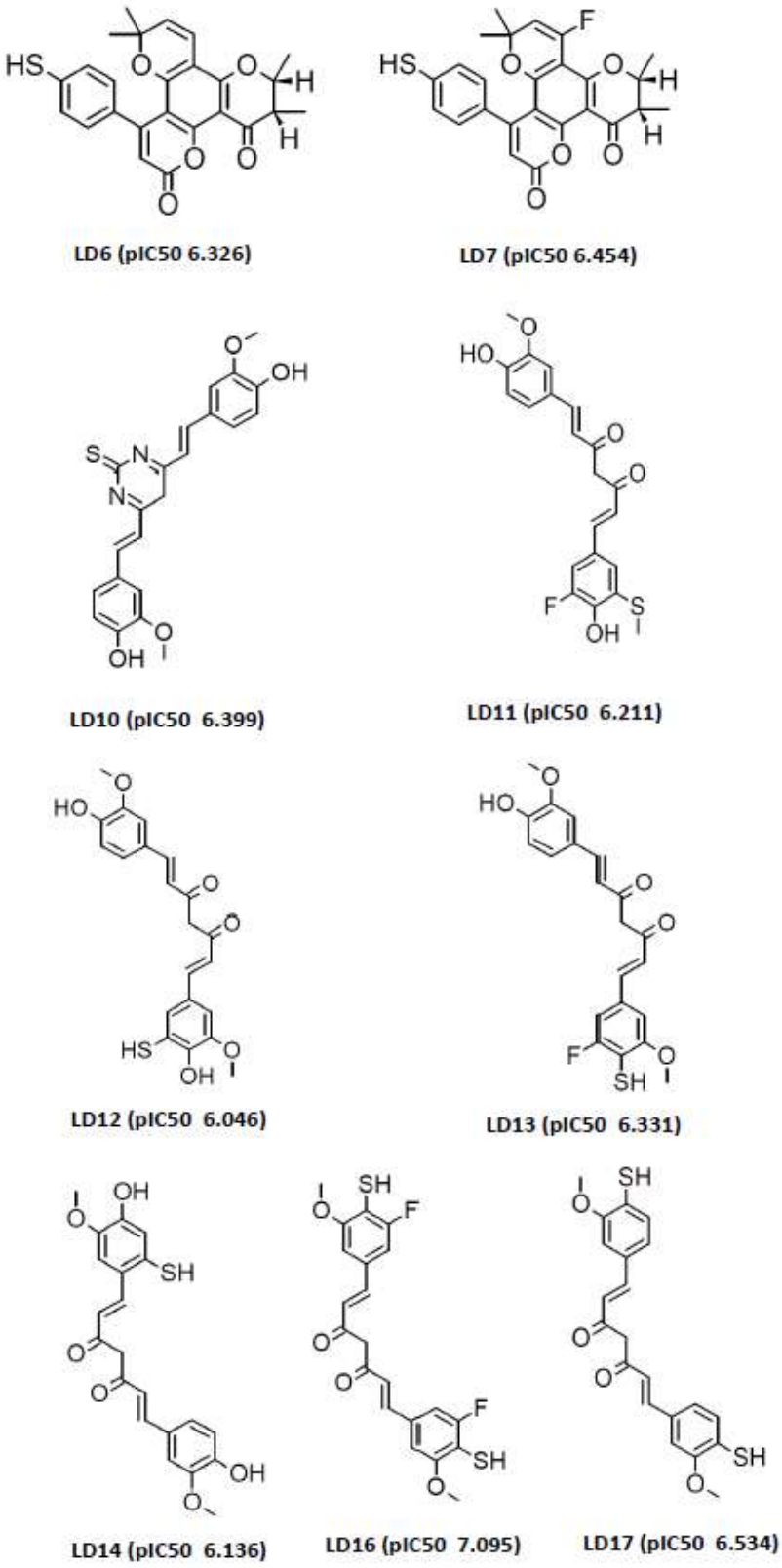

Fig. 2 Structures of sulfur and fluoride derivatives of inophyllum E and curcumin with pIC_{50} value greater than 6.0 (Please refer to table 5 for details of smiles and other properties)

Protein-Ligand Complex

Analysis of simulation data of protein-ligand complexes shows that all the complexes have RMSD values about 3.0 Å or less (Fig. 3). The RMSF values are most fluctuating (>5.0) in the region 462-470 in case of curcumin, 464-468 in case of LD16, Ser466 in case of inophyllum E and (2S)-2'-methoxykurarinone (Fig. 4).

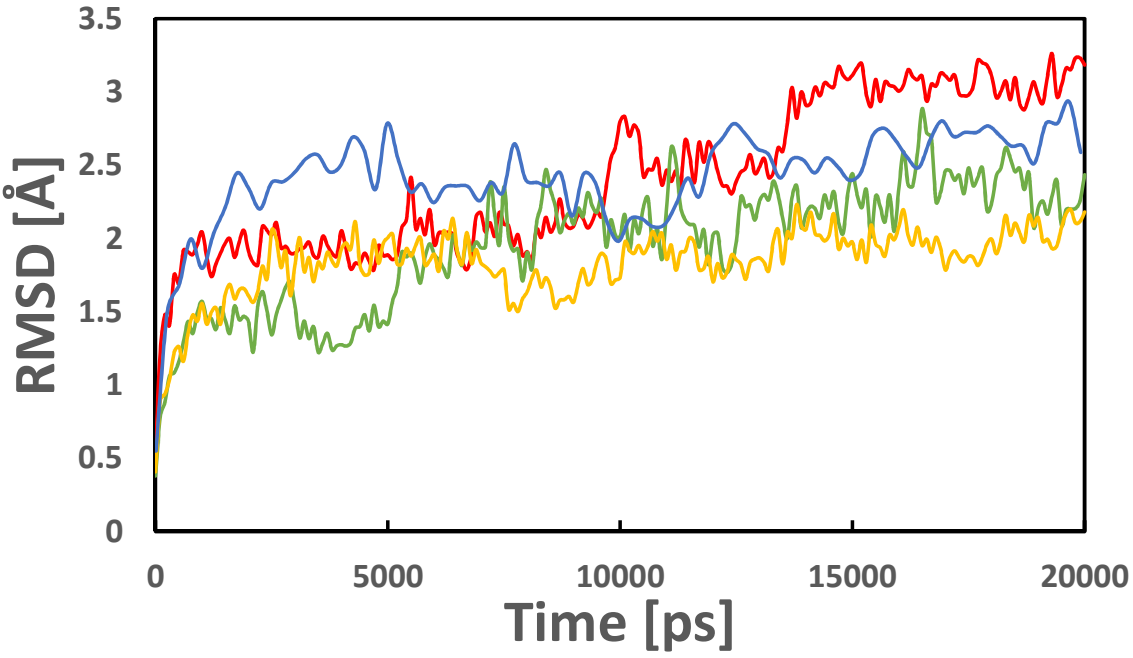

Fig. 3. *Simulation trajectories of H. pylori urease LD16 (blue), curcumin (red), inophyllum E (green) and (2S)-2'-methoxykurarinone (yellow) bound complexes for 20 ns in aqueous medium.*

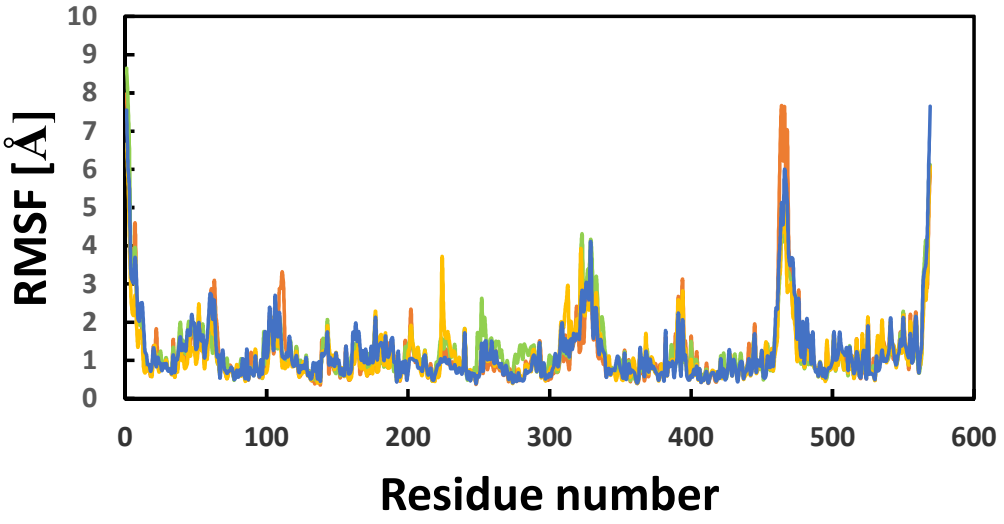

Fig. 4. *Root mean square fluctuation during simulation of H. pylori urease LD16 (blue), curcumin (red), inophyllum E (green) and (2S)-2'-methoxykurarinone (yellow) bound complexes for 20 ns in aqueous medium.*

Ligand Analysis

Two important trajectories of ligands are analyzed: Ligand movement and Ligand conformation. Ligand movement measures displacement of ligand during simulation and ligand conformation measures change in conformation of the ligand during simulation. Both the parameters are expressed in RSMD in Å. The ligand conformational RSMD values were within 2.0 Å for all the ligands including derivatives (not shown here). The ligand movement of inophyllum E was significantly higher (between 6-10 Å) than the RMSD value of protein-complex trajectory (< 3.0 Å), after 10 ns of simulation (Fig. 5). However, the analysis of the structure by alignments of structures before and after MD (20 ns) reveals that the ligand has moved and reoriented inside the cavity and has not gone out (Fig 2S). TM-align is used for the structural alignment [35]

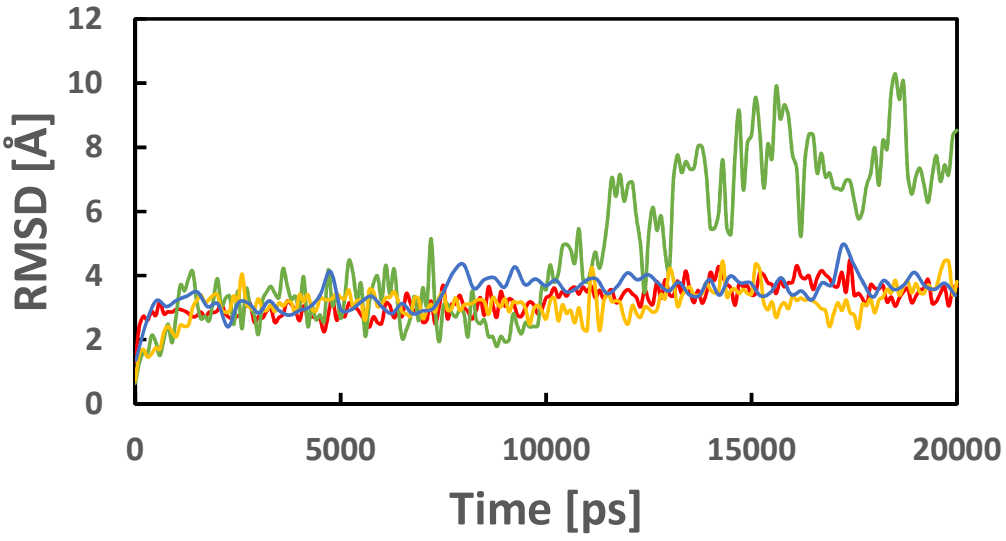

Fig. 5. *Root mean square deviation of movement of ligands during simulation of H. pylori urease-ligand complexes LD16 (blue), curcumin (red), inophyllum E (green) and (2S)-2'-methoxykurarinone (yellow) bound complexes for 20 ns in aqueous medium.*

Conclusion

An acceptable QSAR model has been developed to predict the urease inhibitory potentials of small molecules in terms of their pIC_{50} values. The model may be useful in identifying the potential urease inhibitors in a database of newly designed compounds before going for actual isolation, purification, or chemical synthesis and experimental validation. The reduction of the size of the huge database by computational method to a small set of compounds for experimental validation may drastically reduce the time and cost of the novel drug discovery. Besides, the model may be helpful in optimizing the lead molecule by predicting the increase or decrease of potency due to the modification of the structure of the lead. The work is in progress in our laboratory to apply the QSAR model to the screened hit molecules from ZINC database and designed catechol based molecules for lead identification and optimization [36,

37]. However, the final conclusion awaits experimental validation of the drug candidates selected by *in silico* methods.

References

1. Jones, B. D., Lockatell, C. V., Johnson, D. E., Warren, J. W. & Mobley, H. L. Construction of a urease-negative mutant of Proteus mirabilis: analysis of virulence in a mouse model of ascending urinary tract infection. *Infect. Immun.* **58,** 1120-1123 (1990).

2. Eaton, K. A., Brooks, C. L., Morgan, D. R. & Krakowka, S. Essential role of urease in pathogenesis of gastritis induced by Helicobacter pylori in gnotobioticpiglets. *Infect. Immun.* **59,** 2470-2475 (1991).

3. Cox, G. M., Mukherjee, J., Cole, G. T., Casadevall, A. & Perfect, J. R. Urease as a virulence factor in experimental cryptococcosis. *Infect. Immun.* **68,** 443 – 448 (2000).

4. Loes, A. N., Ruyle, L., Arvizu, M., Gresko, K. E., Wilson, A. L. & Deutch, C. E. Inhibition of urease activity in the urinary tract pathogen Staphylococcus saprophyticus. *Lett. Appl. Microbiol.* **58,** 31-41 (2013).

5. Rutherford, J. C. The emerging role of urease as a general microbial virulence factor. *PLoS Pathog.* **10,** e1004062 (2014).

6. Tsuda, M., Karita, M., Morshed, M. G., Okita, K. & Nakazawa, T. A. A urease-negative mutant of Helicobacter pylori constructed by allelic exchange mutagenesis lacks the ability to colonize the nude mouse stomach. *Infect. Immun.* **62,** 3586-3589 (1994).

7. Modolo, L. V., da-Silva, C. J., Brandão, D. S. & Chaves, I. S. A mini review on what we have learned about urease inhibitors of agricultural interest since mid-2000s. *J. Adv. Res.* **13,** 29-37 (2018).

8. Kafarski, P. & Talma, M. Recent advances in design of new urease inhibitors: A review. *J. Adv. Res.* **13,** 101-112 (2018).

9. Habala, L., Devinsky, F. & Egger, A. E. Metal complexes as urease inhibitors. *J. Coord. Chem.* **71,** 907-940 (2018).

10. Xiao, Z. P., Wang, X. D., Peng, Z. Y., Huang, S., Yang, P., Li, Q. S., Zhou, L. H., Hu, X. J., Wu, L. J., Zhou, Y & Zhu, H. L. Molecular docking, kinetics study, and structure-activity relationship analysis of quercetin and its analogous as Helicobacter pylori urease inhibitors. *J. Agri. Food Chem.* **60,** 10572-10577 (2012).

11. Ul-Haq, Z., Ashraf, S., Al-Majid, A. M. & Barakat, A. 3D-QSAR studies on barbituric acid derivatives as urease inhibitors and the effect of charges on the quality of a model. *Int. J. Mol. Sci.* **17,** E657 (2016).

12. Arora, R., Issar, U. & Kakkar, R. Identification of novel urease inhibitors: pharmacophore modeling, virtual screening and molecular docking studies. *J. Biomol. Struct. Dyn.* **37,** 4312-4326 (2019).

13. Yang, X., Koohi-Moghadam, M., Wang, R., Chang, Y-Y., Woo, P. C. Y., Wang, J., Li, H. & Sun, H. Metallochaperone UreG serves as a new target for design of urease inhibitor: A novel strategy for development of antimicrobials. *PLOS Biol.* **16,** e2003887 (2018).

14. Hassan, S. T. S. & Sudomova, M. The development of urease inhibitors: what opportunities exist for better treatment of Helicobacter pylori infection in children? *Children (Basel)* **4,** 2 (2017).

15. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L. & Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045-D1063 (2016).

16. Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **3,** 33 (2011).

17. Mangal, M., Sagar, P., Singh, H., Raghava, G. P. S. & Agarwal, S. M. NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Res.* **41,** D1124–D1129 (2016).

18. Toropova, A. P., Toropov, A. A., Veselinovic, J. B., Miljkovic, F. N. & Veselinovic, A. M. QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method. *Eur. J. Med. Chem.* **77,** 298–305 (2014).

19. Todeschini, R. & Consonni, V. (2000) Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany

20. Toropov, A. A., Toropova, A. P., Raska Jr, I. QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES. *Eur. J. Med. Chem.* **43,** 714–740 (2008).

21. Toropov, A. A., Toropova A. P., Lombardo, A., Roncaglioni, A., Benfenati, E. & Gini, G. CORAL: Building up the model for bio-concentration factor and defining it's applicability domain. Eur. J. Med. Chem. **46,** 1400–1403 (2011).

22. Roy, P. P., Paul, S., Mitra, I. & Roy, K. On two novel parameters for validation of predictive QSAR models. *Molecules* **14,** 1660-1701 (2009).

23. Golbraikh, A. & Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **20,** 269-276 (2002).

24. Roy, P. & Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* **27,** 302–313 (2008).

25. Toropova, A. P. & Toropov, A. A. The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability. *Sci. Total Environ.* **586,** 466-472 (2017).

26. Trott, O. & Olson, A. J. AutoDockVina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** 455–461 (2010).

27. Krieger, E. & Vriend, G. YASARA view – molecular graphics for all devices – from smartphones to workstations. *Bioinformatics* **30,** 2981-2982 (2014).

28. Lagorce, D., Bouslama, L., Becot, J., Miteva, M. A. & Villoutreix, B. O. FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics* **33,** 3658–3660 (2017).

29. Baell, J. B. & Nissink, J. W. M. Seven year itch: Pan-assay interference compounds (PAINS) in 2017-utility and limitations. *ACS Chem. Biol.* **13,** 36-44 (2018).

30. Karimi, A., Majlesi, M. & Rafieian-Kopaei, M. Herbal versus synthetic drugs; beliefs and facts. *J. Nephropharmacol.* **4,** 27-30 (2015).

31. NagoorMeeran, M. F., Goyal, S. N., Suchal, K., Sharma, C., Patil, C. R. & Ojha, S. K. Pharmacological properties, molecular mechanisms, and pharmaceutical development of asiatic acid: A pentacyclic triterpenoid of therapeutic promise. *Front. Pharmacol.* **9,** 892 (2018).

32. Barbieri, R., Coppo, E., Marchese, A., Daglia, M., Sobarzo-Sánchez, E., Nabavi, S. F. & Nabavi, S. M. Phytochemicals for human disease: An update on plant-derived compounds antibacterial activity. *Microbiol. Res.* **196,** 44-68 (2017).

33. Ahmed, M., Qadir, M. A., Hameed, A., Arshad, M. N, Asiri, A. M. & Muddassar, M. Azomethines, isoxazole, N-substituted pyrazoles and pyrimidine containing curcumin derivatives: Urease inhibition and molecular modeling studies. *Biochem. Biophys. Res. Commun.* **490,** 434-440 (2017).

34. Hassan, S. T. S. & Zemlicka, M. Plant-Derived urease inhibitors as alternative chemotherapeutic agents. *Arch. Pharm. Chem. Life Sci.* **349,** 1–16 (2016).

35. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on TM-score, Nucleic Acids Research, **33,** 2302-2309 (2005).

36. Chopdar, K. S., Dash, G. C., Mohapatra, P. K., Nayak, B. & Raval M. K. *In silico* design of covalently bonding catechol based urease inhibitors as potential candidates for treatment of Helicobacter pylori infection. Int J Pharma Res Health Sci **7**, 3111-3116 (2019).

37. Chopdar, K. S., Dash, G. C., Mohapatra, P. K., Nayak, B. & Raval M. K. *In silico* screening of ZINC database for discovery of novel urease inhibitors as a remedy to gastro-duodenal ulcer caused by *Helicobacter pylori.* Int J Pharm Sci Drug Res **12**, 46-52 (2020).

Author contributions

All authors conceived the work. KSC, PKM and MKR performed the experiments; PKM, KSC, GCD analyzed the data with considerable input from BN and MKR. KSC, PKM and MKR wrote the manuscript with considerable input from all the authors. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.