European Journal of Medicinal Chemistry 46 (2011) 618–630

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Predicting hERG activities of compounds from their 3D structures: Development and evaluation of a global descriptors based QSAR model

Nandita Sinha, Srikanta Sen*

Molecular Modeling Section, Biolab, Chembiotek, TCG Lifesciences Ltd., Bengal Intelligent Park, Tower-B, 2nd Floor, Block-EP & GP, Sector-V, Salt Lake Electronic Complex, Calcutta-700091, India

ARTICLE INFO

Article history:
Received 26 May 2010
Received in revised form
27 October 2010
Accepted 27 November 2010
Available online 8 December 2010

Keywords:
Neural network
Enrichment factor
Domain of applicability
Classification
Statistical analysis

ABSTRACT

A QSAR based predictive model of hERG activity in terms of 'global descriptors' has been developed and evaluated. The QSAR was developed by training 77 compounds covering a wide range of activities and was validated based on an external 'test set' of 80 compounds using neural network method. Statistical parameters and examination of enrichment factor indicated the effectiveness of the present model. Randomization test demonstrated the robustness of the model and cross-validation test further validated the QSAR. Domain of applicability test indicated to the high degree of reliability of the predicted results. Satisfactory performance in classifying compounds into 'active' and 'inactive' groups was also obtained. The cases where the QSAR failed, the possible sources of errors have been discussed.

© 2010 Elsevier Masson SAS. All rights reserved.

1. Introduction

The potassium channel encoded by the human *ether-a-go-go related gene* (hERG) is well known for its important role to the electrical activity of the heart that coordinates the heart's beating. hERG potassium channel allows for the rapid efflux of potassium ions from the cell, which is required for repolarization of the ventricles in preparation for the next cardiac cycle [1–3]. Blockage of this channel is associated with prolongation of the QT interval known as 'long QT syndrome', potentially leading to ventricular arrhythmia or other adverse cardiovascular effects and causing sudden death [1,4,5]. hERG channels are extraordinarily sensitive to blockage by large number of structurally diverse drugs [4–6].

The blockage of hERG channel by a wide number of diverse compounds has become a major concern in drug design as such blocking agents can cause sudden cardiac death [4–7]. In the recent years, several blockbuster drugs, like *terfenadine*, *droperidol*, *lidoflazine*, *sartindole*, *levomethadyl*, *cisapride* etc., have been withdrawn from the market due to their proven adverse hERG activities [5–13]. The potential of a drug candidate to inhibit the hERG channel is therefore of significant concern to the pharmaceutical industry [14,15]. Pharmaceutical companies have successfully developed and utilized higher throughput techniques for the early detection of compounds that block hERG, and this has helped reducing the number of compounds that fail in the late stages of development. However, this screening-based approach is expensive and consumes significant amount of new chemical entities (NCE). Given the cost, labor associated with *in vitro* assays to assess the hERG liability of NCEs, the development of *In Silico* tools to filter out potential hERG channel inhibitors in early stages of the drug discovery process is of great interest.

Considerable amount of computational work has been done in predicting hERG activities of compounds and classifying compounds into hERG active and hERG-inactive groups *In Silico* to aid the drug-design work. Various techniques have been applied in finding appropriate molecular descriptors for modeling structure–activity relationships, for example, substructure analysis, self-organizing maps, principal component analysis, partial least squares fitting, and neural networks [16–26]. However, in such a developing area, always there is a need of new models for achieving better accuracy and improving insights.

In the present work we have developed a computational model using 'global molecular descriptors' to predict the potential of given compounds as hERG channel inhibitors. The performance of the generated model assessed in terms of statistical quantities and its proven ability of fishing out the hERG active compounds from a mixture of hERG active and inactive compounds indicate the

---

* Corresponding author. Fax: +91 33 23572743.
E-mail addresses: nandita@chembiotek.com (N. Sinha), srikanta@chembiotek.com (S. Sen).

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.11.042

usefulness of the present predictive model. Prediction of $p\text{IC}_{50}$ values of individual compounds as well as classification of predicted 'active' and 'inactive' compounds have been discussed here. In this model the selection of relevant 'global descriptors' was made based on physical reasoning. The training set compounds were selected from a collection of compounds from literature with known experimental activity data. Selection of the training set was done carefully avoiding any activity value bias. Multiple statistical parameters have been used in characterizing the performance of the present model. High degree of enrichment in prediction has been demonstrated explicitly. Finally, it has been suggested that a single QSAR cannot assess hERG activities of all varieties of compounds in general. Several new features make the present work distinct from the other similar *In-Silico* hERG prediction works. These are outlined below.

(i) Only simple 'global descriptors' derivable directly from the 3D structures of the compounds have been used. No quantum chemical descriptors have been used.
(ii) Robustness of the QSAR has been demonstrated explicitly by using the method of randomization of experimental data. To the best of our knowledge this has not been done in earlier QSAR related publications on hERG prediction.
(iii) Enrichment effect in the results obtained by applying the QSAR for hERG activities has been demonstrated for the first time.
(iv) Analysis of cross-validation (leave one out) test and randomization test on the hERG classification data for the training set molecules have been included.
(v) Consideration of 'domain of application' has provided better confidence on the reliability of the predicted activities of test compounds and other unknown compounds.

The present work provides a simple predictive *In-Silico* model that is reasonably accurate in predicting the $p\text{IC}_{50}$ values of individual compounds from its 3D structures and also can effectively discriminate between hERG active and inactive compounds. The present QSAR has been validated on a larger independent test set compounds. The data analysis also provides significant insight on the applicability of such a statistical model as well as identifying the features relevant for hERG activity. Any predictive model based on statistical methods is bound to have some failures which are characteristics of statistical method itself. Likewise, our model also failed in some cases. We have discussed the possible reasons behind the observed failures.

2. Methods

2.1. Basis of selection of descriptors

Over the last decade extensive modeling of the hERG channel has been carried out, including homology modeling and site-directed mutagenesis, to identify key binding sites [19,27–29]. As diverse compounds are known to bind to hERG it is logical that the binding is driven mostly by non-specific interactions which indicates that hydrophobic interaction should play a major role in determining the binding potential of a given compound. In order to get better insight on this issue, 30 most active ($p\text{IC}_{50} \geq 6.0$) and 25 most inactive ($p\text{IC}_{50} \leq 4.0$) compounds were selected from our collected dataset and the average logP and volume were calculated for both the most active and the most inactive sets. The average value of the predicted logP values for the most active molecules was found to be 5.1 ($rmsd = 1.8$) while for the most inactive compounds the respective value was found to be 2.2 ($rmsd = 1.7$). This signifies that most of the hydrophobic or non-polar molecules have the tendency to be hERG active. Similarly the average volume of the most hERG active compounds was 501 Å^3 ($rmsd = 118$ Å^3) and for the most inactive ones it was 347 Å^3 ($rmsd = 131$ Å^3) indicating that the molecules having smaller size and shape will be less probable to be hERG active. Also, the non-specific nature of binding implies that molecular descriptors that describe the molecule in global fashion may be the most important generalized descriptors to address large varieties of compounds. By 'global descriptors' we mean molecular descriptors that characterize a compound as a whole. For example, descriptors like polar surface area or volume characterize a compound in a wholistic fashion and thus are global characteristics of molecules. In that sense, molecular weight, volume, number of H-bond acceptors, number of H-bond donors, number of carbon, polar surface area and non-polar surface area, etc., are expected to be good global descriptors in this case. For the size, shape and flexibility, volume and number of rotatable bond may be considered as suitable descriptors in this respect. Average distance between acceptor–acceptor, acceptor–donor and donor–donor may also be important descriptors to have an idea how the acceptors and donors are distributed in 3D. During generating the 3D models of the compounds, the geometry of the molecules are optimized by molecular mechanics force field of Discovery Studio platform. This geometry-optimized 3D conformation was used for computing the conformation dependent descriptors, as there is no rational available for considering the correct conformation. Moreover, in considering the stereoisomeric compounds, if the stereoisomeric nature is explicitly given in the reference, we have used that. In the other cases we have used any of the two possibilities. We have checked that as we are using descriptors of global nature, the values of the descriptors and the predicted activities depend insignificantly on the actual stereoisomeric form in the present work. Thus, based on the physical reasoning on binding, we have selected 12 descriptors and they are (i) number of acceptor, (ii) number of basic nitrogen, (iii) number of carbon, (iv) acceptor–acceptor average distance, (v) donor–donor average distance (vi) volume,(vii) polar surface area, (viii) non-polar surface area,(ix) number of rotatable bond, (x) donor charge,(xi) acceptor charge and (xii) dipole moment. Even though logP appears to be an important determinant of the hERG activity we have not considered logP explicitly in the descriptor set and its effect has been taken care of through other basic descriptors like polar surface area, non-polar surface area, number of acceptor, number of carbon etc. All the global molecular descriptors were computed from the 3D models by computer (FORTRAN) codes developed in-house. Use of 3D descriptors are advantageous in considering features like polar surface area, non-polar surface area, dipole moment etc., that have important roles in determining the interactions between the compound and the receptor. Such descriptors cannot be obtained from 2D structures. We have used global descriptors that contain both conformation independent and dependent descriptors. The main reason for choosing 'global molecular descriptors' that are not very specific to individual compounds, is to capture diverse compounds in the QSAR.

2.2. Collection and compilation of dataset

All the molecular structures and the experimental activity data were collected from the published literatures. As a result, the experimental conditions and accuracies of the data may vary significantly. However, we have paid equal status to all the data used in the present work. It may further be pointed out that many of the compounds reported in the literature were common or very similar and thus were discarded. In some cases we were not very sure about the structures and those were ignored. Considering all these factors finally, a total of 157 compounds with their known

Table 1
The Smiles strings of the training set compounds along with their experimental $p\mathrm{IC}_{50}$ values, predicted $p\mathrm{IC}_{50}$ values, errors ($p\mathrm{IC}_{50}^{\mathrm{expt}} - p\mathrm{IC}_{50}^{\mathrm{pred}}$) in predictions and the references to literature sources are summarized.

No	Smiles string of the compound	Expt $p\mathrm{IC}_{50}$	Pred $p\mathrm{IC}_{50}$	Error	Ref
1	O=C2NC(=O)NC(=O)C2(CC)c1ccccc1	2.52	3.73	−1.21	[21]
2	Clc2ccc1n(cc(c1c2)C(CC)CC)c3ccc(F)cc3	2.83	3.77	−0.94	[22]
3	Clc2ccc1n(cc(c1c2)C(=O)CC)c3ccc(F)cc3	2.70	4.22	−1.52	[22]
4	Clc2ccc1n(cc(CC)c1c2)c3ccc(F)cc3	2.46	4.01	−1.55	[22]
5	Fc5c(N2CC1C(NCCC1)C2)c(OC)c4N(C3CC3)\C=C(/C(=O)c4c5)C(O)=O	3.89	3.96	−0.07	[23]
6	O(C)c1cc(cc(OC)c1OC)Cc2cnc(N)nc2N	3.62	3.73	−0.11	[22]
7	O=C/3\C=C/C2(C)CC(O)C1C(OC(=O)C1C)/C2=C\3C	3.96	3.97	−0.01	[22]
8	O=C(NCCN(CC)CC)c1ccc(N)cc1	3.86	5.00	−1.14	[21]
9	Fc4c(N1CCNCC1)cc3N(C2CC2)\C=C(/C(=O)c3c4)C(O)=O	3.02	4.27	−1.25	[21]
10	C1(CCCN1C)c2cnccc2	3.61	4.16	−0.55	[18]
11	Fc4c(N1CC(NCC1)C)c(OC)c3N(C2CC2)\C=C(/C(=O)c3c4)C(O)=O	3.89	3.84	0.05	[23]
12	c1ccc2c(c1)SC(C(C(N2CCN(C)C)=O)OC(=O)C)c3ccc(OC)cc3	4.76	4.75	0.01	[18]
13	OC(c1ccccc1)(c2ccccc2)C4CCN(CCCC(O)c3ccc(C(C)(C)C(O)=O)cc3)CC4	4.64	5.51	−0.87	[23]
14	Fc4c(N1CC(NCC1)C)cc3N(C2CC2)\C=C(/C(=O)c3c4C)C(O)=O	4.30	4.47	−0.17	[23]
15	c1c(ccc(c1)C(=O)O)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	4.13	5.39	−1.26	[19]
16	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)CCC	4.59	3.66	0.93	[19]
17	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(CC)(CC)O	4.80	5.10	−0.30	[19]
18	O=C(N)C(CCN(C(C)C)C(C)C)(c1ccccc1)c2ncccc2	4.04	5.12	−1.08	[21]
19	S(=O)(=O)(N1CCN(CC1)CC)c2cc(c(OCC)cc2)C=4Nn3c(nc(c3C(=O)N=4)C)CCC	4.89	5.47	−0.58	[21]
20	Clc2ccc(C(CCN(C)C)c1ncccc1)cc2	4.68	5.04	−0.36	[21]
21	O1C(CCC1)C(N2CCN(CC2)c3nc4c(c(n3)N)cc(OC)c(OC)c4)=O	4.75	5.33	−0.58	[21]
22	c1ccccc1C(c2ccc(cc2)Cl)N3CCN(CC3)CCOCC(O)=O	4.50	5.86	−1.36	[20]
23	FC(F)Oc1ccc(cc1OC(C)C)C(\C2=C\NC(\C=C2)C(O)(C)C)C=3\C=N/C(=O)CC=3	4.25	5.23	−0.98	[22]
24	N\1=C(\c3c(Nc2c/1cccc2)sc(c3)C)N4CCNCC4	4.90	5.07	−0.17	[20]
25	N\1=C(\c3c(Nc2c/1cc(cc2)Cl)cccc3)N4CCNCC4	4.50	5.05	−0.55	[20]
26	c1(c(cccc1)[N^{+}]([O^{-}])=O)C\2/C(=C(\N/C(=C/2C(=O)OC)C)C)C(=O)OC	4.30	4.19	0.11	[20]
27	c1(c(OCC)ccc(c1)S(=O)(N2CCN(CC2)C)=O)C=3NC(c4c(N=3)c(nn4C)CCC)=O	5.48	6.05	−0.57	[18]
28	N1(c3c(CCc2c1cccc2)cccc3)CCCNC	5.86	5.91	−0.05	[18]
29	c1(cc(c(cc1)C3(Cn2cncc2)OCC(O3)COc4ccc(cc4)N5CCN(CC5)C(=O)C)Cl)Cl	5.72	5.93	−0.21	[18]
30	O=C2N(CCc1n(c3c(c12)cccc3)C)Cc4ncnc4C	5.50	5.89	−0.39	[18]
31	N1c2c(CCC1=O)cc(cc2)N3CCN(CC3)C(c4cc(c(cc4)OC)OC)=O	5.96	6.71	−0.75	[18]
32	C(C1CCCCC1)(CC2CCCCN2)C3CCCCC3	5.11	4.71	0.40	[18]
33	N2(c1ccccc1Sc3c2cc(Cl)cc3)CCCN(C)C	5.83	6.28	−0.45	[18]
34	C2(c1c(cccc1)CCc3c2cccc3)=[C@H]CCN(C)C	5.00	4.81	0.19	[18]
35	FC(F)(F)c2nc3c(c(C(O)C1NCCCC1)c2)cccc3C(F)(F)F	5.26	4.16	1.10	[23]
36	c1c(cccc1)C(c2ccccc2)CC(C)NC(C)(C)C	5.24	4.12	1.12	[23]
37	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(C)O	5.34	4.73	0.61	[19]
38	Clc2cc1nccc(NC(CCCN(CC)CC)C)c1cc2	5.60	6.09	−0.49	[21]
39	O(C(c1ccccc1)c2ccccc2)CCN(C)C	5.59	5.39	0.20	[21]
40	FC(F)(F)COc2c(C(=O)NCC1CCCCN1)cc(OCC(F)(F)F)cc2	5.41	4.63	0.78	[21]
41	c1c2c(c(cc1Cl)C(O)CN(CCCC)CCCC)c4c(/C2=C\c3ccc(cc3)Cl)cc(cc4)Cl	5.09	5.66	−0.57	[21]
42	Fc4cc1c(C(C(OC(=O)COC)(CC1)CCN(CCCc2nc3c(n2)cccc3)C)C(C)C)cc4	5.84	5.98	−0.14	[21]
43	Clc4cc(N1CCN(CC1)CCCN2\N=C\3/N(C2=O)\C=C/C=C/3)ccc4	5.54	5.48	0.06	[21]
44	O=C(N1CCN(CC1)c2nc3c(c(n2)N)cc(OC)c(OC)c3)c4occc4	5.80	4.85	0.95	[21]
45	c1c(cc(cc1)C/2/C(=C(\N\C(=C\2C(OCC)=O)C)C)C(OC)=O)[N^{+}]([O^{-}])=O	5.00	5.17	−0.17	[20]
46	O=C2N(C)c1ncn(C)c1C(=O)N2C	5.31	4.12	1.19	[22]
47	c1(noc2c1ccc(c2)F)C3CCN(CC3)CC\C4=C(\N=C5/N(C4=O)CCCC5)C	6.83	6.32	0.51	[18]
48	Clc5c(CCN1CCN(CC1)c2nsc3c2cccc3)cc4c(NC(=O)C4)c5	6.90	6.71	0.19	[18]
49	S2c4c(N(CCC1N(CCCC1)C)c3c2cccc3)cc(S(=O)C)cc4	6.50	6.42	0.08	[18]
50	OC(c1ccccc1)(c2ccccc2)C4CCN(CCCC(O)c3ccc(C(C)(C)C)cc3)CC4	6.67	6.28	0.39	[18]
51	C(C#N)(CCCN(CCc1cc(OC)c(cc1)OC)C)(C(C)C)c2ccc(c(c2)OC)OC	6.85	5.91	0.94	[18]
52	Clc2c1cc(C(O)CCN(CCCC)CCCC)c3c(c1cc(Cl)c2)cc(cc3)C(F)(F)F	6.71	5.94	0.77	[18]
53	c1cc(ccc1Cl)c2ccc(o2)\C=N\N3CC(N(C3=O)CCCCN4CCN(CC4)C)=O	6.25	6.32	−0.07	[23]
54	c1c(ccc(c1)CC(=O)O)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	6.24	5.42	0.82	[19]
55	C1C(CCCC1)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	6.86	7.49	−0.63	[19]
56	c1c(ccc(c1)CC(=O)OC)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	6.88	6.92	−0.04	[19]
57	c1(ccccc1)C(c2ccccc2)(O)C3CCN(CC3)CCCC(O)c4ccc(cc4)C(O)(C)C	6.34	6.27	0.07	[19]
58	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C4CCNCC4	6.69	6.05	0.64	[19]
59	Ic3c(OCCN(CC)CC)c(I)cc(C(=O)c1c(oc2c1cccc2)CCCC)c3	6.00	5.78	0.22	[21]
60	S(=O)(=O)(Nc4cc(OC)c(Nc1c3c(nc2c1cccc2)cccc3)cc4)C	6.68	4.92	1.76	[21]
61	Clc2cc1NC(=O)N(c1cc2)C3CCN(CC3)CCCN4c5c(NC4=O)cccc5	6.79	5.62	1.17	[21]
62	C1N(CCC(C1)OC(c2ccccc2)c3ccccc3)CCCC(=O)c4ccc(cc4)C(C)(C)C	6.48	6.82	−0.34	[21]
63	c1c2c(ccc1)NC(N(C2=O)CCN3CCN(CC3)C(=O)c4ccc(cc4)F)=O	6.97	5.99	0.98	[21]
64	Brc1cccc(c1)c2nc(nc2)C3CCC(CC3)C(=O)NC(=O)C	6.30	5.24	1.06	[22]
65	c1cc2c(cc1)n(cc2\C3=C(\C(NC3=O)=O)c4cnc5c4cccc5)CCCN(C)C	6.00	5.93	0.07	[21]
66	N5(C1CCN(CC1)CCCC(c2ccc(cc2)F)c3ccc(cc3)F)c4ccccc4NC5=O	7.26	7.15	0.11	[18]
67	Clc2cc1c(cn(c1cc2)c3ccc(F)cc3)C4CCN(CC4)CCN5CCNC5=O	7.83	7.05	0.78	[18]
68	S2c4c(N(CCC1CCCCN1C)c3c2cccc3)cc(SC)cc4	7.02	6.85	0.17	[21]
69	c1c(cccc1)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	7.06	6.90	0.16	[19]
70	c1c(ccc(c1)C(=O)OC)n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5C(NCC5)=O	7.44	7.05	0.39	[19]
71	c1c(ccc(c1)F)N2CC(c3c2cccc3)C4CCN(CC4)CCN5C(NCC5)=O	7.63	7.52	0.11	[19]
72	Fc5ccc(Cn3c(NC1CCN(CC1)CCc2ccc(OC)cc2)nc4c3cccc4)cc5	7.94	7.57	0.37	[21]
73	c1c(ccc(c1)F)C(c2ccc(cc2)F)CCCN3CCN(CC3)CC(Nc4c(cccc4C)C)=O	7.80	8.06	−0.26	[21]

Table 1 (*continued*)

No	Smiles string of the compound	Expt pIC_{50}	Pred pIC_{50}	Error	Ref
74	n2c1ccccc1c(c2c3ccccc3)C4CCN(CC4)c5ccccc5	7.10	6.75	0.35	[22]
75	O=C(NCc1ccc(CC(O)CO)c(OC)c1)CCCCCCCC	7.00	5.78	1.22	[22]
76	c1c(ccc(c1)F)n2cc(c3c2cccc3)C4CCN(CC4)CCN5C(NCC5)=S	8.21	8.62	−0.41	[19]
77	c1c(ccc(c1)F)n2cc(c3c2cccc3)C4CCN(CC4)CCN5C(NCC5)=O	8.16	7.77	0.39	[19]

experimental activity data have been collected from the literature [16–23,30]. We have used pIC_{50} as our activity parameter which is related to IC_{50} as $pIC_{50} = -\log_{10}IC_{50}$. In selecting the 'training set' compounds, care was taken to select more or less equal number of representatives for each sub-range of pIC_{50} values to avoid any bias resulting from over representation of a specific range. First the lowest and the highest pIC_{50} values of the compounds were identified. Then the entire range was divided into 10 consecutive zones and compounds were randomly selected from the individual zones. We tried to keep the number of representative compounds from the individual zones reasonably similar to avoid any bias towards one or a few particular zones. The residual compounds were considered as the test set compounds. In this way a training set consisting of 77 compounds (Table 1) covering the maximum range (2.46–8.21) of experimental pIC_{50} values were generated. The remaining 80 compounds were considered as the test set compounds (Table 2). The test set compounds covered a range (2.36–8.66) of experimental pIC_{50} values. The Smiles Strings of the training set and test set compounds along with their experimental pIC_{50} values and the literature sources are given in Table 1 and Table 2 respectively. The 2D structures of the compounds have been included as supplementary materials for the ready reference to the readers. The structures and activity data were collected from different published literature and the experimental details were not available for all of them. However, for many of the compounds patch clamp hERG current inhibition assay using mammalian cell lines HEK or CHO were employed to measure the respective experimental pIC_{50} values.

2.3. *Generation of a QSAR based on the selected descriptors*

Feed forward back propagation neural network method with linear activation function has been used in generating a QSAR involving the selected molecular descriptors [31,32]. As in the present work, we have used a single linear neural network (a single neuron only) and a small set of 12 molecular descriptors as inputs, hence the possibility of over fitting is negligible. The initial values of the weights of the descriptors were assigned randomly in the range (0–1) and a small learning rate was used. Then the weights of the input descriptors were optimized in iterative fashion by minimizing the squared error $Err^2 = [\sum_{i=1}^{N}(x_i - y_i)^2]$, where $x_i = pIC_{50}^{\text{pred}}(i)$ represents the predicted pIC_{50} value of the ith compound, $y_i = pIC_{50}^{\text{expt}}(i)$ corresponds to the respective experimental value of pIC_{50} and N is the total number of data points. The mean squared error $mse = Err^2/N$ was monitored and the iteration was continued until the mse reached a predefined sufficiently low value or it reached saturation.

2.4. *External validation of the generated QSAR*

The generated QSAR was validated on the basis of an external test set consisting of 80 compounds with known experimental pIC_{50} values but were not used in the training set. The pIC_{50} values of the compounds were predicted by the present QSAR and were compared with the respective experimental data. The statistical parameters were also computed to characterize the quality of the predictions for the test set.

2.5. *Details of data analysis*

2.5.1. *Characterization of the QSAR in terms of statistical quantities*

Both the training and test results are characterized in terms of statistical quantities. The quality of the QSAR both for the training set and the test set compounds was assessed by the following statistical quantities:

$$\text{Correlation coefficient } (r) = \frac{\langle x_i y_i \rangle - \langle x_i \rangle \langle y_i \rangle}{\sqrt{\left\langle x_i^2 \right\rangle - \langle x_i \rangle^2} \cdot \sqrt{\left\langle y_i^2 \right\rangle - \langle y_i \rangle^2}} \quad (1)$$

$$\text{Mean Absolute Error } (mae) = \frac{1}{N}\sum_{i=1}^{N} |(x_i - y_i)| \quad (2)$$

$$\text{Root Mean Squared Error } (rmse) = \left[\frac{1}{N}\left\{\sum_{i=1}^{N}(x_i - y_i)^2\right\}\right]^{1/2} \quad (3)$$

$$\text{Cross} - \text{validated correlation coefficient } \left(q^2\right) = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{\sum_{i=1}^{N}(x_i - \langle x_{Tr} \rangle)^2} \quad (4)$$

The quantity r, known as linear correlation coefficient, measures the strength and the direction of a linear relationship between experimental and prediction activity of the compounds. The quality of the fit in a linear regression is commonly expressed in terms of r^2 and is known as the coefficient of determination. In general a QSAR with $r^2 > 0.60$ is considered as a good model to predict the activities well within the error range.

2.5.2. *Assessing the accuracy of prediction by cross validation test and randomization test*

In order to demonstrate that the observed correlation of the generated QSAR is not a 'chance-correlation' due to choice of training set compounds, we have performed the standard tests (i) the cross-validation test (leave one out test) and (ii) the randomization test. In the cross-validation test we trained the training set compounds leaving one compound out and examined the quality of the generated QSAR considering the remaining training set compounds and then we used this new QSAR to predict the pIC_{50} value of the compound that was left out. The predicted pIC_{50} value of this compound was then compared to the original predicted value of the respective compound. Small differences between such predicted values indicate that the quality of the generated QSAR is not significantly dependent on the transient compounds and thus cross-validate the QSAR. These procedures were repeated for all the 77 compounds of the training set.

Robustness of the QSAR model was checked by the 'randomization test'. In this technique, new QSAR models were developed by randomly shuffling the experimental pIC_{50} values of the training set compounds keeping the molecular descriptors the same. The new QSAR models were expected to have poorer values of the statistical quantities characterizing the quality of the QSAR. If the opposite happens, an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Table 2
The Smiles strings of the test set compounds along with their experimental pIC_{50}values, predicted pIC_{50}values, errors $(pIC_{50}^{expt} - pIC_{50}^{pred})$in predictions and the references to literature sources are summarized.

No	Smiles string of the compound	Expt pIC_{50}	Pred pIC_{50}	Error	Ref
1	O=C3NC(=O)NC3(c1ccccc1)c2ccccc2	3.62	4.58	−0.96	[21]
2	O=C1NC(=O)C(CC)(C(=O)N1)C(C)CCC	3.88	3.33	0.55	[22]
3	O(C3CC1N2C(CC(C1)C(=O)C2)C3)C(=O)c4c5c(nc4)cccc5	4.92	5.08	−0.17	[18]
4	O(C2C(C(OC1OC(C(O)C(OC)(C1)C)C)C(C(OC(C(O)(C(O)C(C(=O)C(CC2(O)C)C)C)C)CC)=O)C)C)C3OC(CC(N(C)C)C3O)C	4.14	3.54	0.60	[21]
5	C1C2(N(CC1)CCC2)CC(Nc3c(cccc3C)C)=O	4.69	5.18	−0.49	[21]
6	O(C2C(C(OC1OC(C(O)C(OC)(C1)C)C)C(C(OC(C(O)(C(O)C(C(=N\OCOCCOC)/C(CC2(O)C)C)C)C)CC)=O)C)C)C3OC(CC(N(C)C)C3O)C	4.44	3.85	0.59	[21]
7	c1c2c(ccc1)ncc2C(=O)OC5CC3CC4CC(N3CC4O)C5	4.92	3.69	1.23	[22]
8	N2(c1ccccc1CCc3c2cccc3)CCCN(C)C	5.47	5.88	−0.41	[18]
9	S2c4c(C(\N1CCN(CC1)CCOCCO)=N/c3c2cccc3)cccc4	5.24	5.99	−0.75	[23]
10	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(C)C	5.00	3.74	1.26	[19]
11	c12c(ccc(c1)C(C)(C)C)n(cc2)c3ccc(cc3)F	5.00	4.47	0.53	[19]
12	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(CC)O	5.66	4.96	0.70	[19]
13	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(C)=O	5.00	4.08	0.92	[19]
14	Fc3ccc(C1(OCc2c1ccc(c2)C#N)CCCN(C)C)cc3	5.40	5.53	−0.13	[21]
15	Brc1cncc(c1)C(=O)N4CCC([NH^{2+}]Cc2n(cnc2)Cc3ccc(C#N)cc3)C4=O	5.22	6.42	−1.20	[22]
16	S4C(=O)N\N=C(\c2ccc1N(CCCc1c2)\C(=N\CC)c3ccc(OC)c(OC)c3)C4C	5.18	5.81	−0.63	[22]
17	O=C1N\N=C(/CC1)c3ccc(Nc2ccncc2)cc3	5.98	4.65	1.33	[22]
18	n2c1ccccc1c(c2c3ccccc3)C4C[NH^{2+}]CCC4	5.31	6.02	−0.71	[22]
19	O(C)c1ccc(cc1OC)Cc3nccc2cc(OC)c(OC)cc23	5.14	4.49	0.65	[22]
20	C\1(=N\c3c(Nc2c/1cccc2)ccc(c3)Cl)N4CCN(CC4)C	5.40	5.84	−0.44	[20]
21	O=C(OC)\C1=C\CC2[NH^{+}](C)C1CC2	5.05	4.16	0.89	[22]
22	C=2(c1c(cccc1)Nc3c(N=2)cc(cc3)Cl)N4CCN(CC4)C	6.50	6.03	0.47	[18]
23	S/3/C/2=N/c4c(NC(\N1CCN(CC1)C)=C\2\C=C\3C)cccc4	6.64	6.06	0.58	[18]
24	c32C(=O)C(Cn1ccnc1C)CCc2n(c4c3cccc4)C	6.09	5.55	0.54	[18]
25	FC(F)(F)c2ccc(OC(CCNC)c1ccccc1)cc2	6.34	5.68	0.66	[21]
26	c1cc(ccc1F)Cn2c5c(nc2N3CCC(CC3)N(C)C=4NC(\C=C/N=4)=O)cccc5	6.36	7.27	−0.91	[21]
27	O4C(C(=O)N1CCN(CC1)c2nc3c(c(n2)N)cc(OC)c(OC)c3)COc5c4cccc5	6.23	5.47	0.76	[21]
28	Cl\C(=C(/c1ccc(OCCN(CC)CC)cc1)c2ccccc2)c3ccccc3	6.74	5.44	1.30	[21]
29	FC(F)(F)c1cccc(c1)c2nc(nc2)C3CCC(CC3)C(=O)NCCO	6.30	5.70	0.60	[22]
30	c12N(CCOc1cccc2)CCN3CCC(CC3)c4cc(c(cc4)OC)OC	6.60	6.53	0.07	[20]
31	c1c2c(ccc1)OCCC2CCN3CCC(CC3)c4ccc(c(c4)OC)OC	6.70	6.12	0.58	[21]
32	n2c1ccccc1c(c2c3ccccc3)C5CC6N(c4ccccc4)C(C5)CC6	7.75	7.19	0.56	[22]
33	S(=O)(=O)(C)Nc1ccc(cc1)C(C2CCN(CC2)CCc3nc(ccc3)C)=O	7.92	6.58	1.34	[21]
34	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)\C4=C\CN(CC4)CCN5C(NCC5)=O	8.00	7.18	0.82	[19]
35	O=C(NC1CC2N(C(C1)CCC2)C)c3nn(c4c3cccc4)C	5.42	5.39	0.03	[16]
36	C12C(C(CC(CC1)N2C)OC(=O)c3ccccc3)C(=O)O	2.40	3.88	−1.48	[16]
37	c1c2c(c(c(c1)OC)OC)c[n^{+}]3c(c2)c4c(CC3)cc5c(c4)OCO5	4.18	4.37	−0.19	[16]
38	ClC(Cl)(Cl)C(C)(C)O	2.36	2.78	−0.42	[16]
39	c1c(cc(cc1)C(F)(F)F)CCNCC	3.27	4.20	−0.93	[16]
40	c1c2c(ccc1NS(=O)(=O)C)oc(c2C(c3ccc(cc3)OCCCN(CCCC)CCCC)=O)CCCC	6.50	6.76	−0.26	[16]
41	C\1(=C\CC2CCC/1N2C)C(OC)=O	3.77	3.50	0.27	[16]
42	O(c3ccc(/C(=C(/CC)c1ccccc1)c2ccccc2)cc3)CCN(C)C	5.80	5.31	0.49	[16]
43	O(C(C(CN(C)C)C)(Cc1ccccc1)c2ccccc2)C(=O)CC	5.56	4.58	0.98	[16]
44	c4(c2c(CNC(Cc1ccc(cc1)OC)=O)cccc2)c(C(NCCc3cccnc3)=O)cccc4	5.00	6.04	−1.04	[30]
45	C2C1N(CCc4c1c(c3c2ccc(c3O)O)ccc4)C	5.62	4.23	1.39	[30]
46	c14c(ccc(c1)OCCCCN2CCN(CC2)c3c(c(ccc3)Cl)Cl)CCC(N4)=O	6.20	7.63	−1.43	[30]
47	N2(c1c(cccn1)Cl)CCN(CC2)C(Nc3ccc(cc3)C(C)(C)C)=O	6.00	5.53	0.47	[30]
48	COc5ccc4cc3c2cc1OCOc1cc2CC[n^{+}]3cc4c5OC	5.51	4.44	1.07	[30]
49	C12(C(C1)CNC2)c3ccc(cc3)C	4.35	4.06	0.29	[30]
50	C(c1ccc(cc1)Br)(c2ccccn2)CCN(C)C	6.05	4.84	1.21	[30]
51	c1cnc(nc1)N2CCN(CC2)CCCCN3C(CC4(CC3=O)CCCC4)=O	5.40	6.02	−0.62	[30]
52	n23c(c(C(NCC1CCN(CC(C)C)CC1)=O)cc(Cl)c2N)nc(c3)C	5.59	5.30	0.29	[30]
53	[C@@]14(/C(/C=C\[C@@H]2[C@@H]1CC[C@]3([C@H]2CC[C@@]3(CCC(=O)O)O)C=C\C(=O)CC4)C	3.98	3.32	0.66	[30]
54	C(N1CCC(CC1)NC(c2c(cc(c(c2)Cl)N)OC)=O)c3ccccc3	6.22	5.63	0.59	[30]
55	c1(c(nc2c(c1)C(C(CN2C3CC3)C(O)=O)=O)N4C/C(C5(C4)CNC5)=N\OC)F	3.66	4.20	−0.54	[30]
56	c1(ccccc1)CN2CCC(CC2)CC3Cc4c(C3=O)cc(c(c4)OC)OC	5.70	5.86	−0.16	[30]
57	c1ccc\2c(c1)OCc3c(C/2=C\CCN(C)C)cccc3	5.19	5.60	−0.41	[30]
58	P(Sc1ccccc1)(Sc2ccccc2)(OCC)=O	4.30	3.84	0.46	[30]
59	c1(cc(c2c(c1)OC(C(C2)OC(c3cc(c(c(c3)O)O)O)=O)c4cc(c(c(c4)O)O)O)O)O	5.22	4.44	0.78	[30]
60	[C@@H]12[C@H]([C@H](C[C@@H](N1C)CC2)O)C(OC)=O	3.38	3.30	0.08	[30]
61	O([C@@]3(C)[C@H]([C@@H](O[C@@H](O[C@@H]1[C@H](C(O[C@@H]([C@](O)(C)[C@@H]([C@H]([C@H]([C@@H](C[C@](O)(C)[C@@H]([C@H]1C)O[C@@H]2O[C@@H](C[C@@H]([C@H]2O)N(C)C)C)C)N)C)O)CC)=O)C)C3)C)O)C	3.50	3.43	0.07	[30]
62	C2(\c1c(cccc1)OCc3c2ccc(c3)F)=C4/CCN(CC4)CC(C(O)=O)(C)C	5.00	5.09	−0.09	[30]
63	c13c(c(c(n1C(c2ccc(cc2)Cl)=O)C)CC(O)=O)cc(cc3)OC	3.52	3.60	−0.07	[30]
64	C\15(C(N(/C(=N/1)CCCC)Cc2ccc(cc2)c4c(c3nnnn3)cccc4)=O)CCCC5	3.71	5.08	−1.37	[30]
65	C\3(c1c2c(ccc1)non2)\C(=C(/N/C(=C/3C(OC)=O)C)C)C(OC(C)C)=O	4.30	3.52	0.78	[30]
66	[C@H]1([C@H]([C@@H](CC(O[C@@H](C\C=C\C=C\[C@@H]([C@@H](C[C@@H]1CC=O)C)O)C)=O)OC(C)=O)OC)O[C@H]2[C@@H]([C@H]([C@@H]([C@H](O2)C)O[C@H]3C[C@@]([C@H]([C@@H](O3)C)OC(CC(C)C)=O)(C)O)N(C)C)O	3.99	2.85	1.14	[30]
67	c1(ccc(cc1)[N^{+}](=O)[O^{-}])CCCCN(CCCCCCC)CC	8.66	7.42	1.24	[30]

Table 2 (*continued*)

No	Smiles string of the compound	Expt pIC_{50}	Pred pIC_{50}	Error	Ref
68	[C@@H]1(N(CCCC1)CCCC)C(Nc2c(cccc2C)C)=O	4.68	5.57	−0.89	[30]
69	c1(c(c(c2c(c1)C(/C(=C\N2CC)C(O)=O)=O)F)N3CCNC(C3)C)F	2.62	3.70	−1.08	[30]
70	CN1CCCCC1C(=O)Nc2c(C)cccc2C	3.81	4.96	−1.15	[30]
71	c1ccc(c(c1C)OCC(N)C)C	5.00	4.38	0.62	[30]
72	c1(ccc(cc1)C(NCCN2CCOCC2)=O)Cl	4.10	4.77	−0.67	[30]
73	C(c1c(cc(c(c1)Cl)N)OCC)(=O)NCC2CN(CCO2)Cc3ccc(cc3)F	5.32	5.36	−0.04	[30]
74	C(c1ccccc1)(Oc2ccc(cc2)C(F)(F)F)CCN	5.64	4.82	0.82	[30]
75	C=1(c3c(Nc2c(N=1)cccc2)sc(c3)C)N4CCN(CC4)C	6.69	6.28	0.41	[30]
76	c12c4c(n(c1CCC(C2=O)Cn3c(ncc3)C)C)cccc4	6.09	5.38	0.71	[30]
77	c1cccc(c1)C(c2ccccc2C)OCCN(C)C	6.07	5.15	0.92	[30]
78	c45c(C1CCN(CC1)CCC=2C(N3/C(=N\C=2C)C(CCC3)O)=O)noc4cc(cc5)F	5.89	5.67	0.22	[30]
79	c1cc3c2c(c1)[C@@H]4[C@@H](Cc2cn3)N(C[C@@H](C4)CSC)CCC	6.92	5.92	1.00	[30]
80	O(C(=O)C2(c1ccccc1)CCN(CC2)C)CC	4.13	3.84	0.29	[30]

2.5.3. *Studies on enrichment factor*

Twenty-five known hERG active compounds were mixed with 2475 compounds randomly picked up from the three databases named *Oakwood*, *Maybridge* and *IBScreen* to compile a dataset for performing the enrichment studies. The compounds were ranked according to their predicted pIC_{50} values. The percentage of known active compounds recovered at different top fraction of the ranked list was examined against the fraction of ranked compounds considered. The objective was to examine how successfully the method could fish out the known actives from the mixture. The enrichment factor (EF) is defined as [33]

$$EF = \frac{Hits_{sample}/N_{sample}}{Hits_{total}/N_{total}} \tag{5}$$

In this equation, $Hits_{sample}$ is the number of active compounds recovered in the top N_{sample} compounds in the ranked dataset. $Hits_{total}$ is the total number of active compounds in the dataset and N_{total} represents the total number of compounds in the dataset. Thus, EF can have a maximum value given by $EF_{max} = N_{total}/Hits_{total}$.

2.5.4. *Applicability domain of the generated QSAR*

In the case of a new compound with no experimental data, a predicted value without an idea of reliability of the value is not useful. Thus, in order to use a QSAR model for evaluating new compounds, its domain of application [34,35] needs to be defined and only those predictions that fall within this domain may be regarded as reliable. There are several measures for the domain of applicability of a QSAR [34–37]. Extent of extrapolation is a simple measure to define the applicability domain [34,35]. It is based on the calculation of the leverage h_i for each compound, where the QSAR model is used to predict its activity. The leverage h_i is defined as shown below [37].

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \tag{6}$$

where the vector x_i represents the ith query compound in the descriptor space and x_i^T is the transpose of its representative matrix. X is the $n \times k$ training data matrix containing the k descriptor values in the descriptor space for each of the n training compounds. The leverage of a compound is a measure related to the statistical error of prediction at that compound and can be viewed as a measure of extrapolation. A leverage value $h_i > 3k/n$ [36,37] is considered large. This criterion means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

2.5.5. *Assessing the accuracy of prediction for classification based on the present QSAR*

In order to assess the performance of the present QSAR in classifying compounds into hERG ‘active’ and ‘inactive’ groups, the compounds of the ‘training set’ and the ‘test set’ were classified based on their experimental activity data and the predicted activity values using a predefined cutoff value for pIC_{50}. The quality of classification performance was assessed in terms of a set of standard statistical parameters [38,39].

A brief description of the statistical parameters follows: (i) *Sensitivity* is the percentage of correctly predicted actives out of the total number of experimentally found actives and thus represents how precisely the QSAR is able to predict actives; (ii) *Specificity* is the percentage of correctly predicted non-actives out of the total number of experimentally found inactive and hence corresponds to the precision of prediction of inactive molecules; (iii) *Accuracy* is the fraction of compounds predicted correctly to either true actives or true inactives and hence represents the overall correctly predicted fraction of the result; (iv) *PPV* (positive prediction value, also referred to as precision) is the probability of correct positive prediction giving the percentage of correct active predictions out of the total active predictions; (v) *NPV* (negative prediction value) is the probability of correct negative prediction and gives the percentage of true presentation of inactive prediction out of total inactive prediction; and finally (vi) *Matthew's correlation coefficient* (*MCC*) is a measure of the quality of binary classifications and can be used even if the classes are of very different sizes. *MCC* is limited in the range (+1.0 to −1.0). A value of +1 represents a perfect prediction, 0 corresponds to a random prediction and −1 represents an inverse prediction. The parameters may be calculated by the following equations [38].

$$Sensitivity = \frac{n(TP)}{n(TP) + n(FN)} \tag{7}$$

$$Specificity = \frac{n(TN)}{n(TN) + n(FP)} \tag{8}$$

$$Accuracy = \frac{n(TP) + n(TN)}{n(TP) + n(FP) + n(TN) + n(FN)} \tag{9}$$

$$PPV = \frac{n(TP)}{n(TP) + n(FP)} \tag{10}$$

$$NPV = \frac{n(TN)}{n(TN) + n(FN)} \tag{11}$$

$$MCC = \frac{n(TP)n(TN) - n(FP)n(FN)}{\sqrt{[n(TP) + n(FP)][n(TP) + n(FN)][n(TN) + n(FP)][n(TN) + n(FN)]}} \tag{12}$$

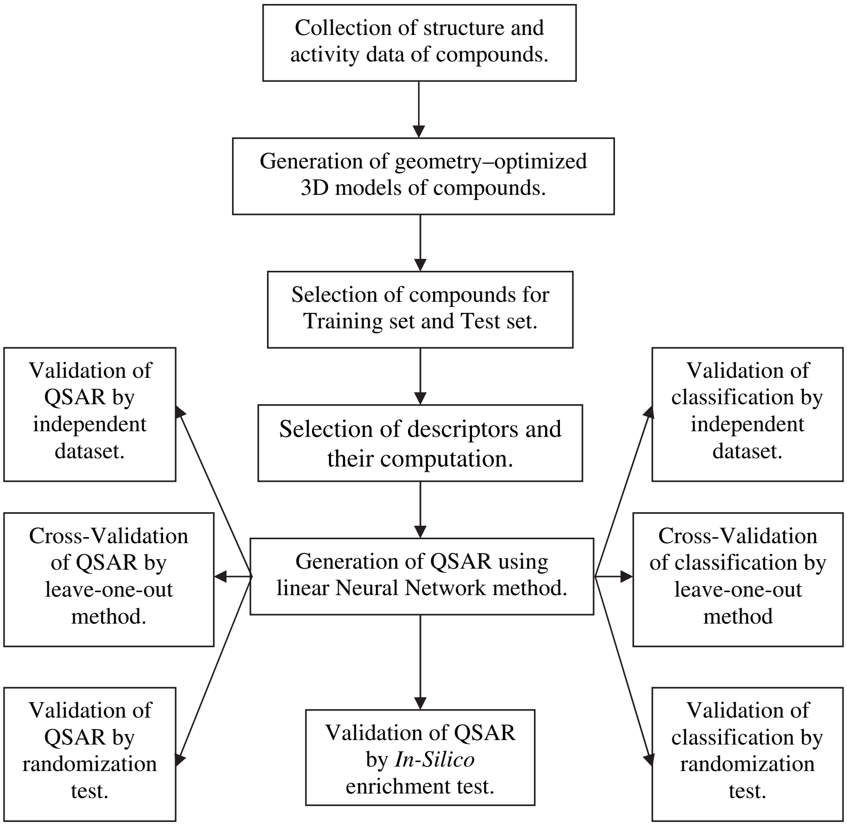

Fig. 1. Flowchart summarizing the overview of the computational work flow.

where, *n* (*TP*) and *n* (*FP*) represent the numbers of true positives and false positives while *n* (*TN*) and *n* (*FN*) represent the numbers of true negatives and false negatives respectively.

Fig. 1 summarizes the overview of the computational workflow as a flowchart.

3. Results

3.1. Results of training set compounds

The training of the 77 pre-selected compounds along with their experimental pIC_{50} values was performed using our in-house developed codes based on neural network as described in the method section. The statistical parameters quantifying the quality of the training are summarized in Table 3. The predicted values of pIC_{50} of the individual compounds of the training set and the associated errors are given in Table 1. Analysis of the distribution of the errors associated with the predictions indicates that 52% of compounds have error ≤ 0.5 log unit while only 16% compounds have $1.5 \geq$ error > 1.0 log unit. Rest of the molecules (29%) excepting a few is within the error range 0.5–1.0 (log unit). Fig. 2 represents the scatter plot of the experimental and predicted values pIC_{50} for visual comparison. Dashed line corresponds to the perfect match between the experimental and predicted pIC_{50} values. The statistical parameters as well as Fig. 2 indicate that the quality of the training is quite good and hence the resulting QSAR can be used for predicting the hERG activity (pIC_{50}) of a given compound from its 3D structure.

Table 3
Summary of the statistical parameters mean absolute error (*mae*), mean squared error (*mse*), root mean squared error (*rmse*) and coefficient of determination (r^2) for the 'training set', 'test set' and 'failed set'.

Set	No. of compounds	QSAR used	mae	mse	rmse	r^2
Training	77	Training of 77	0.57	0.52	0.72	0.73
Test	80	Training of 77	0.68	0.61	0.78	0.64
Failed	32	Training of 77	2.13	4.71	2.17	0.01
Failed	32	Training of 32	1.08	1.61	1.27	0.77

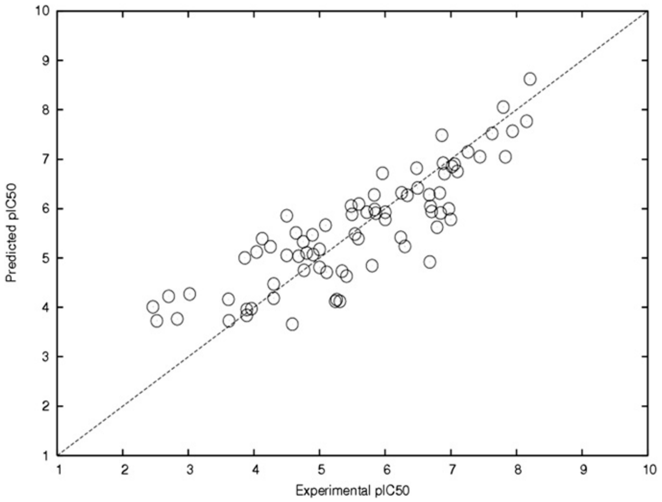

Fig. 2. The scatter plot of the experimental pIC_{50} *vs.* the predicted pIC_{50} values of the individual compounds of the 'training set' for comparison. Dashed line corresponds to the ideal match between the experimental and predicted pIC_{50} values. The best fitted line (not shown) for the data points, follows the equation $pIC_{50}^{pred} = 0.70pIC_{50}^{expt} + 1.66$ where pIC_{50}^{expt} represents the experimental pIC_{50} value and pIC_{50}^{pred} represents the predicted pIC_{50} value.

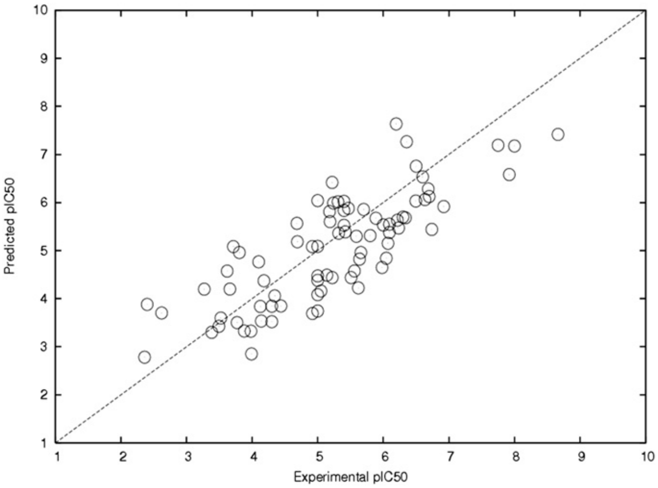

Fig. 3. The scatter plot of the experimental $p\mathrm{IC}_{50}$ *vs.* the predicted $p\mathrm{IC}_{50}$ values of the individual compounds of the ‘test set’ for comparison. Dashed line corresponds to the ideal match between the experimental and predicted $p\mathrm{IC}_{50}$ values. The best fitted line (not shown) for the data points, follows the equation $p\mathrm{IC}_{50}^{\mathrm{pred}} = 0.72p\mathrm{IC}_{50}^{\mathrm{expt}} + 1.27$.

3.2. *Results of external test set compounds*

The external validation ‘test set’ consists of 80 compounds with known experimental $p\mathrm{IC}_{50}$ values in the range (2.36–8.66) and these compounds were not considered for training. The comparison of the experimental and predicted $p\mathrm{IC}_{50}$ values for the test set compounds is summarized in Table 2. Analysis of the errors associated with the predicted $p\mathrm{IC}_{50}$ values indicates that 34% of compounds have error ≤ 0.5 log unit while only 22% compounds have $1.5 \geq$ error >1.0 log unit. Rest of the molecules (44%), is placed within the error range 0.5–1.0 (log unit). The quality of the prediction results are characterized by the statistical parameters r^2, *mae*, *mse* and *rmse* as given in Table 3. The values of these statistical parameters indicate reasonably good quality considering the fact that the prediction is based on a statistical model. Fig. 3 represents the scatter plot of the experimental and predicted $p\mathrm{IC}_{50}$ values of the compounds of the test set for visual comparison. Here also the dashed line corresponds to the perfect match between the experimental and the predicted $p\mathrm{IC}_{50}$ values.

3.3. *Cases where the present QSAR failed*

Table 4 summarizes the predictions on additional 32 test compounds with known experimental hERG activity data collected from literature and were not included in the training set. These compounds are treated here separately as high errors (>1.6 log unit) are associated with their predicted $p\mathrm{IC}_{50}$ values. There may be several reasons behind the observed large errors between the predicted and the respective experimental activity values. First we checked whether these compounds are outside the ‘application domain’ of the current QSAR. The results show that only two of them are outside the ‘application domain’ indicating that the reason of the failure in these cases is something different. The next most probable reason might be the fact that such experimentally found active molecules may have binding modes quite different from that

Table 4
The Smiles strings of additional 32 test compounds along with their experimental $p\mathrm{IC}_{50}$values, predicted $p\mathrm{IC}_{50}$values, errors $(p\mathrm{IC}_{50}^{\mathrm{expt}} - p\mathrm{IC}_{50}^{\mathrm{pred}})$in predictions and the references to literature sources are summarized.

No	Smiles string of the compound	Expt $p\mathrm{IC}_{50}$	Pred $p\mathrm{IC}_{50}$	Error	Ref
1	Clc2ccc1n(cc(CCC)c1c2)c3ccc(F)cc3	1.59	4.21	−2.62	[22]
2	Fc3c(N1CCN(CC1)C)c2OCC(N\4c2c(c3)C(=O)/C(=C/4)C(O)=O)C	2.85	5.13	−2.28	[21]
3	Clc2ccc1n(cc(c1c2)C(O)C)c3ccc(F)cc3	2.34	4.63	−2.29	[22]
4	Clc2ccc1n(cc(c1c2)C(O)CC)c3ccc(F)cc3	2.66	4.86	−2.20	[22]
5	O(C2C(C(OC1OC(C(O)C(OC)(C1)C)C)C(C(OC(C(O)(C(O)C(C(=O) C(CC2(OC)C)C)C)C)CC)=O)C)C)C3OC(CC(N(C)C)C3O)C	4.48	2.68	1.80	[21]
6	Clc3cc(C(=O)NCCc2ccc(S(=O)(=O)NC(=O)NC1CCCCC1)cc2)c(OC)cc3	4.13	5.95	−1.82	[21]
7	c12c(cccc1)c3c(n2)C(N4C(C3)C(N(CC4=O)C)=O)c5ccc6c(c5)OCO6	4.00	5.72	−1.72	[21]
8	c12c3c(nc1cccc2)cccc3OCC(O)CNCCOc4ccccc4OC	4.90	6.97	−2.07	[20]
9	O=C(NCCCN(C)c2nc(C)c1cc(OC)c(OC)cc1n2)C3OCCC3	4.08	5.69	−1.61	[22]
10	C2(C(OC(=O)c1ccccc1)CC3N(C2CC3)C)C(=O)OC	5.14	2.92	2.22	[18]
11	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)C(CC)CC	5.83	3.70	2.13	[19]
12	C2(\c1ncccc1CCc3c2ccc(c3)Cl)=C4\CCN(CC4)C(=O)OCC	6.76	4.71	2.05	[18]
13	OC1(CCN(CC1)CCCC(c2ccc(cc2)F)=O)c3ccc(cc3)Cl	7.57	5.28	2.29	[18]
14	c1c(ccc(c1)F)n2cc(c3c2ccc(c3)Cl)CN(C)C	7.96	5.18	2.78	[19]
15	O(CC(N1CCCC1)CN(Cc2ccccc2)c3ccccc3)CC(C)C	7.64	5.88	1.76	[21]
16	S(=O)(=O)(Nc1ccc(C(O)CCCN(CCCCCCC)CC)cc1)C	7.82	5.92	1.90	[21]
17	c1c(ccc(c1)F)C(=O)CCCN2CCC(CC2)(O)c3ccc(cc3)Cl	7.60	5.79	1.81	[20]
18	C14(CCN(CC1)C2CCc3c(C2)ccc(c3)C#N)CC(c5c(O4)ccc(c5)NS(=O)(=O)C)O	7.68	5.21	2.47	[21]
19	c3(C(=O)NC1C(CN(CC1)CCCOc2ccc(cc2)F)OC)cc(Cl)c(cc3OC)N	8.17	5.14	3.03	[18]
20	Fc4ccc(C(=O)CCCN3CCC(\N1c2c(NC1=O)cccc2)=C/C3)cc4	7.49	5.73	1.76	[17]
21	c1cnccc1N	2.36	4.30	−1.94	[16]
22	c1cc(ccc1)C3(c2ccccc2)CCN(CC3)C(C)(C)C	6.07	4.07	2.00	[16]
23	c1cc(ccc1Cl)CCCC[N$^+$](CCCCCCC)(CC)CC	7.90	5.57	2.33	[16]
24	n1cn(cc1)C(c2ccccc2)(c3ccccc3)c4c(cccc4)Cl	5.52	3.61	1.91	[16]
25	C1(CC2CCC(C1C(OCC)=O)N2C)OC(=O)c3ccccc3	5.92	3.88	2.04	[16]
26	C/1(C(C\C(=C\1CC)C)C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)NC3CCC(CC3)C)=O	4.13	6.71	−2.58	[16]
27	N7(CC1CC1)[C@H]6[C@]45[C@]3(c2c(ccc(c2O[C@H]3[C@@]([C@@H] ([C@@](O)(C)C(C)(C)C)C4)(OC)CC5)O)C6)CC7	5.10	3.40	1.70	[30]
28	CC(C)(O)C(Cl)(Cl)Cl	2.36	5.72	−3.36	[30]
29	C2(Cc1ccccc1)CCN(CC2)C(C(c3ccc(cc3)O)O)C	7.00	5.31	1.69	[30]
30	c1(NC(CN(CC)CC)=O)c(cccc1C)C	3.58	5.35	−1.77	[30]
31	N4([C@H](C(N[C@@H](Cc1ccccc1)C[C@@H]([C@H](Cc2ccccc2) NC(COc3c(cccc3C)C)=O)O)=O)C(C)C)C(NCCC4)=O	5.07	7.64	−2.57	[30]
32	c23[C@H](NC1CC1)[C@H](C(Oc2cc(c(c3)NC(Cc4ccc(cc4)OC)=O) [N$^+$](=O)[O$^-$])(C)C)O	4.36	6.01	−1.65	[30]

covered in the training set and thus the descriptors should have different weights to take care of these compounds. Since a QSAR is a representative of a specific or similar binding mode of compounds, a different QSAR might be more successful for these compounds. In other words, a more successful and different QSAR may provide some confidence in favor of this possible reasoning. In order to check the above possibility, we trained these 32 compounds for generating a new QSAR based on the same set of descriptors. The statistical parameters as given in Table 3 of the training process and the associated errors in individual cases indicate that the new QSAR describes these compounds in a much better way supporting our anticipation explained above. It may be noted that the *mae* value obtained even in this separate training of the 32 compounds showed relatively large value (1.08). This may be due to the possibility that these 32 compounds may represent more than one class of compounds and thus, the new QSAR is also not capable of taking care of all of them. This is a common problem associated with predictive tools based on statistical methods. Our finding thus suggests that as hERG blockade may occur through different types of binding, a single model may not assess hERG activity for all kinds of compounds.

The other possible causes of failure include the inadequacy of the descriptor set used for the proper description of these compounds and the large errors associated with the experimental activity data in such cases.

3.4. Studies on accuracy of prediction by cross-validation test and randomization test

In the cross-validation test the training set compounds were trained leaving one compound out and the quality of the generated QSAR was evaluated. Then this new QSAR was used to predict the $p\mathrm{IC}_{50}$ value of the compound that was left out. These procedures were repeated for all the 77 compounds of the training set and every time the result was close to the original one and the predicted $p\mathrm{IC}_{50}$ values were within the error limit. At the end of the cross-validation test 77 different QSAR were generated with r^2 in the range of 0.72–0.75 where as the r^2 value of our original QSAR was 0.73. The ranges of the *mse* and *mae* were (0.48–0.52) and (0.55–0.57) respectively while the *mse* and *mae* of our original QSAR was 0.52 and 0.57 respectively. The predicted $p\mathrm{IC}_{50}$ value of the left out compound by the new QSAR and the original QSAR is almost the same for the individual compounds with a few exceptions and the differences covered a range from 0.00 to 0.63. Moreover, the q^2 values were found in range from 0.58 to 0.64. All these observations indicated that the observed correlation of the original QSAR is not a chance-correlation. The results are summarized in Table 5 and the detailed results of 77 different QSAR generated by training of the training set compounds after leaving one compound out are given in the supplementary materials.

In the randomization test, we developed new QSAR models after randomly shuffling the experimental $p\mathrm{IC}_{50}$ values of the training set compounds keeping the molecular descriptors the same. The new QSAR models generated based on the randomized experimental data, were expected to show poorer quality as compared to the original QSAR. If the opposite happens, an acceptable QSAR model cannot be obtained for the specific descriptor set and data. We generated twelve such randomized training set using different random number seeds and found that in all the cases, the statistical parameters were very poor, for example, r^2, *mse* and *mae* values were in the ranges (0.03–0.14), (1.70–2.31) and (1.02–1.23) respectively. In addition, poor values (−7.48–0.01) were found for the corresponding parameter q^2 also. This clearly showed the robustness of the QSAR and the major dependence of $p\mathrm{IC}_{50}$ values on the descriptors used in the present case. The results are summarized in Table 6.

Table 5
Summary of the statistical parameters obtained in the cross-validation test.

r^2 range	q^2 range	*rmse* range	Min error (cross)[a]	Max error (cross)[a]	Min error[b]	Max error[b]
0.72–0.75	0.58–0.64	0.69–0.72	0.01	1.89	0.00	0.63

[a] Max and min refer to the absolute error ($= p\mathrm{IC}_{50}^{\mathrm{expt}} - p\mathrm{IC}_{50}^{\mathrm{cross}}$), where $p\mathrm{IC}_{50}^{\mathrm{expt}}$ and $p\mathrm{IC}_{50}^{\mathrm{cross}}$ represent the experimental $p\mathrm{IC}_{50}$ value of the left out candidate and the respective value predicted by the cross-validation QSAR respectively.

[b] Max and min refer to the absolute error ($= p\mathrm{IC}_{50}^{\mathrm{orig}} - p\mathrm{IC}_{50}^{\mathrm{cross}}$), where $p\mathrm{IC}_{50}^{\mathrm{orig}}$ and $p\mathrm{IC}_{50}^{\mathrm{cross}}$ represent the $p\mathrm{IC}_{50}$ value of the left out candidate predicted by the original QSAR and the cross-validation QSAR respectively.

Table 6
Summary of the statistical parameters obtained in the randomization test.

Trial number	r^2	q^2	*mae*	*rmse*
1	0.12	−3.68	1.02	1.31
2	0.03	0.01	1.23	1.52
3	0.08	−5.92	1.23	1.47
4	0.12	−3.54	1.14	1.43
5	0.09	−4.16	1.14	1.46
6	0.08	−5.16	1.17	1.46
7	0.05	−7.00	1.16	1.49
8	0.12	−3.54	1.13	1.42
9	0.11	−3.71	1.17	1.45
10	0.14	−3.36	1.13	1.42
11	0.05	−7.48	1.16	1.47
12	0.10	−3.80	1.19	1.46

3.5. Studies on enrichment factor

Twenty-five known active compounds were mixed with 2475 compounds randomly picked from the *Oakwood* database to compile a dataset for performing the enrichment studies. The objective was to examine how successfully the present model could

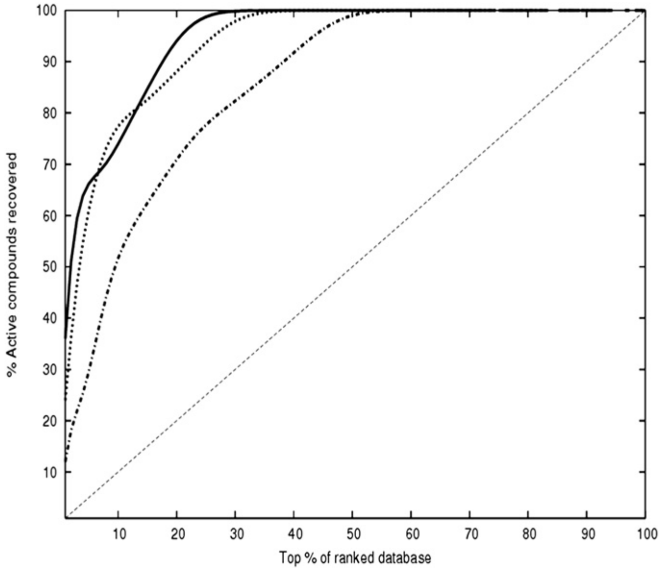

Fig. 4. Plot of % of known active compounds recovered within a given % of the top ranked compiled mixed compounds set against the given top % of the ranked mixed compounds set showing the resulting ‘Enrichment effect’. The mixed compounds set was ranked in the descending order of predicted $p\mathrm{IC}_{50}$ values. The solid line, dotted line and the dashed-dotted line represent the recovery of active compounds from *Oakwood*, *Maybridge*, and *IBScreen* database respectively. The dashed straight line represents the recovery of active compounds in random fashion.

Table 7
The value of the observed EF over the top 1%, 5% and 10% of the ranked three databases are summarized. N_{active} represents the number of known active compounds used in this exercise and N_{random} corresponds to the number of compounds randomly selected from each database to compile the composite compounds sets.

Name of the database	N_{active}	N_{random}	EF over top 1%	EF over top 5%	EF over top 10%	100% recovered in top
Oakwood	25	2475	36.0	13.6	7.2	22.0%
Maybridge	25	2475	24.0	12.8	8.0	29.0%
IBScreen	25	2475	12.0	4.8	5.6	46.0%

fish out the known actives from the mixture. The solid line in Fig. 4 clearly represents that a high degree of enrichment was achieved for compounds from *Oakwood* database. It is seen that within the top 15% of the ranked database, 84% of the known actives were recovered and within the top 22% of the ranked database all the known actives were recovered. In order to check the generality of the observations, we repeated the same process for two other databases *Maybridge* and *IBScreen* of commercially available compounds. In Fig. 4, the dotted line represents the case for compounds from *Maybridge* database and the dash-dotted one corresponds to the same for the compounds from *IBScreen* database. The dotted line shows that within the top 15% of the ranked database 80% of the known actives were recovered and within the top 29% of the ranked database all the known actives were recovered. For compounds from *IBScreen* database, 64% of the known actives were recovered within the top 15% of the ranked database, and within top 46% of ranked database 100% known actives were completely recovered. Compared to the recovery of active compounds in random fashion (the dashed line) the actual recovery curves represent considerable enrichment effect.

These results demonstrate the effectiveness of the present QSAR in identifying the active compounds from a large compound set. The enrichment factors within top 1%, 5% and 10% of each of the three databases are given in Table 7. Fig. 5 demonstrates the variation in the value of the enrichment factor (EF) with the increase in the size of the top % of the ranked databases. However, it may be pointed out that all the 2475 compounds selected randomly from the respective database were considered as hERG inactives even though we do not have any experimental data on their hERG activities. Some of these compounds may be hERG actives in reality and in that case the actual value of EF will be much higher compared to what presented here. Thus, the EF values reported here are the lowest values of it.

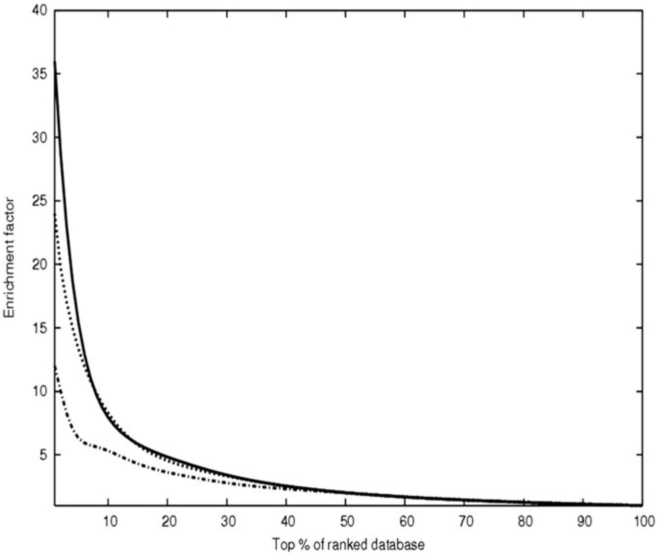

Fig. 5. Plot of computed values of the enrichment factor against the top % of the ranked mixed compounds set used for computing its values. The mixed compounds set was ranked in the descending order of predicted $p\text{IC}_{50}$ values. The solid line, dotted line and the dashed-dotted line represent the *Oakwood*, *Maybridge*, and *IBScreen* database respectively.

3.6. Examination of applicability domain

In the present case the 'training descriptor matrix' X was of order 77 × 12 and thus the threshold leverage value $3k/n$ was 0.4675. The h_i values for the individual test compounds were computed as described in the methods section using in-house developed computer codes and the results have been summarized in Table 8. The results indicate that out of the 80 test compounds (not present in the training set) only 9 have values greater than the threshold leverage value $3k/n$. Thus, 89% of the test compounds are within the applicability domain indicating that their predicted activity values are reliable. Further analysis of the data indicates that for 28 compounds $h_i < 0.1$, for 57 compounds $h_i < 0.2$. So, 71% of compounds have $h_i < 0.2$ indicating significantly low degree of extrapolation. All these together clearly demonstrate the good quality of the present QSAR even in predicting the $p\text{IC}_{50}$ values of new compounds in a reliable fashion.

3.7. Classification of hERG active and inactive compounds by the generated QSAR

The performance of the present QSAR in classifying compounds into hERG 'active' and 'inactive' groups was assessed by classifying the compounds of the 'training set' and the 'test set' based on their experimental activity data and the predicted activity values using a cutoff value for $p\text{IC}_{50}$. As in literature we found three cutoff values used for such classification, we have used all three and have compared the results [13,24,26]. In one case compounds with $p\text{IC}_{50} \leq 5.0$ were considered as hERG inactive ones [13] and compounds with $p\text{IC}_{50} > 5.0$ were taken to be hERG active. In the other cases, classification of active and inactive compounds was

Table 8
The leverage (h) values of the test set compounds are summarized. The threshold leverage value is given by $h^* = 3k/n$, where k is the number of descriptors and n is the number of compounds in training set. Thus here $h^* = 0.4675$. The compounds which are outside of the 'application domain' are highlighted in bold.

Mol No	h	Mol no	h	Mol no	h	Mol no	h
1	0.1884	21	0.1085	41	0.0544	**61**	**2.7442**
2	0.1218	22	0.0650	42	0.1224	62	0.1362
3	0.0911	23	0.1302	43	0.1107	63	0.1627
4	**2.3254**	24	0.0996	44	0.1796	**64**	**2.8187**
5	0.0399	25	0.0867	45	0.1888	**65**	**0.5547**
6	**2.5092**	26	0.1520	46	0.1713	**66**	**1.9188**
7	0.1051	27	0.2806	47	0.0681	**67**	**1.8958**
8	0.0813	28	0.0909	48	0.3097	68	0.0698
9	0.1149	29	0.1975	49	0.1059	69	0.0861
10	0.0996	30	0.1561	50	0.0716	70	0.0473
11	0.1291	31	0.1796	51	0.0875	71	0.0877
12	0.0592	32	0.2082	52	0.1438	72	0.0907
13	0.0827	33	0.1124	53	0.2629	73	0.1126
14	0.1534	34	0.0854	54	0.0836	74	0.0716
15	0.3831	35	0.1888	55	0.2781	75	0.1091
16	0.2624	36	0.0872	56	0.1490	76	0.0640
17	0.3774	37	0.3162	57	0.1907	77	0.0995
18	0.2960	38	0.4554	**58**	**0.8334**	78	0.3110
19	0.2027	39	0.1644	**59**	**2.4584**	79	0.1158
20	0.0545	40	0.2485	60	0.0517	80	0.0473

Table 9
Summary of the statistical parameters characterizing the quality of the classification of compounds of the ‘training set’, ‘test set’ and ‘failed set’ using the present QSAR.

Cutoff pIC_{50}	Data set	*N*	*Sensitivity*	*Specificity*	*Accuracy*	*PPV*	*NPV*	*MCC*
4.4	Training set	77	0.93	0.69	0.88	0.92	0.73	0.64
	Test set	80	0.88	0.82	0.86	0.93	0.72	0.67
	Failed set	32	0.65	0.17	0.47	0.57	0.22	−0.20
5.0	Training set	77	0.84	0.54	0.73	0.76	0.65	0.39
	Test set	80	0.79	0.82	0.80	0.86	0.73	0.60
	Failed set	32	0.61	0.36	0.50	0.55	0.42	−0.03
6.0	Training set	77	0.72	0.92	0.84	0.84	0.85	0.66
	Test set	80	0.55	0.93	0.83	0.75	0.84	0.53
	Failed set	32	0.00	0.80	0.50	0.00	0.57	−0.29

done in the same way but using cutoff values 4.4 and 6.0 respectively [26]. For the three cutoff values the statistical parameters (Eq. (7)–Eq. (12)) were calculated for the training set of 77 molecules, the test set of 80 molecules, and the failed set of 32 compounds. For the cutoff value 5.0, the *accuracy* of the training set indicates that 73% of the total training set molecules were correctly classified while, for the test set 80% of the compounds were correctly classified. Thus *accuracy* for the training set and the test set are similar. The *sensitivity* for the training set (0.84) was also quite similar to that of the test set (0.79). Some other parameters like *PPV* and *NPV* were also similar for the training set and the test set. However, significant differences were observed in the parameters *specificity* and *MCC* and both were found to be better for the test set. It is interesting to note that even for the failed compounds set, the values of most of the statistical parameters for classification are not very poor. The reason seems to be the fact that classification is done on the basis of low resolution criterion (a single cutoff value). For the cutoff value 4.4 all the statistical parameters like *sensitivity*, *accuracy* etc, except *specificity*, were found to be similar for both the training set and the test set. Here the accuracy for the training set and the test set are 0.88 and 0.86 respectively. It is further interesting to note that all the parameters for the training set compounds were improved significantly for the smaller cutoff value 4.4 with respect to the corresponding parameters obtained with the cutoff value 5.0 but this picture is not true for the test set compounds. For the cutoff value 6.0, the *accuracy* of the training set indicates that 84% of the total training set molecules were correctly classified while, for the test set 83% of the compounds were correctly classified. All the statistical parameters for classification like *accuracy*, *specificity* etc. except *sensitivity* for the training set and the test set are also found to be similar for the cutoff value 6.0. The detail results are summarized in Table 9.

One more interesting feature we observed in the classification results. The *sensitivity* and *specificity* values obtained for the training set compounds with the cutoff value 4.4 are found to be 0.93 and 0.69 respectively. This means that with this cutoff the active compounds are predicted with much higher accuracy compared to that for the inactive ones. On the other hand, calculation with cutoff value 6.0 produced the *sensitivity* and *specificity* values as 0.72 and 0.92 respectively, meaning that in this case the inactive compounds are predicted with higher accuracy compared to the actives. The same trend was found for the test set compounds also. Hence, it appears that the choice of cutoff value in classification depends on the purpose. For example, if identification of hERG inactive compounds is required then the cutoff value of 6.0 seems to be a better choice.

We have further performed the cross-validation test (leave one out) for the classification of the training set compounds using the cutoff values 5.0, 4.4 and 6.0 for pIC_{50}. For all the cutoff values it is found that the ranges of all the individual statistical parameters are very similar to the respective values for the original QSAR. The results are summarized in Table 10.

In randomization test with cutoff value 5.0 we found that the ranges of all the statistical parameters excepting *MCC* are similar to the respective values obtained with the original QSAR. The range of *MCC* was (−0.04–0.46) with most of the values in the lower range while for the original training set, the value of *MCC* was 0.39. On the other hand, with the cutoff value 4.4, it is found that the *specificity* (0.00–0.25) and the *MCC* (−0.06–0.39, with most of the values in the lower range) were significantly poorer compared to the original QSAR. For the cutoff value 6.0 it was seen that all the individual statistical parameters except *specificity* became very poor compared to the respective values obtained with the original QSAR. The detailed observations are summarized in Table 10.

It is interesting to point out that as classification is done based on cutoff values of pIC_{50}, it is a low resolution description and hence many of the statistical parameters characterizing the classification are not significantly sensitive to the randomization of the pIC_{50} values as found in the present work.

3.8. Comparison of the present model with other hERG predictive models

Quite a number of publications have reported the results of successful *In-Silico* models for predicting hERG activities. Early works employed pharmacophore based approaches. Ekins et al. developed a pharmacophore based model that resulted in r^2 value of 0.90 for the training set compounds [18]. In another paper [23] a regression analysis using HQSAR method produced $r^2 = 0.52$ and $q^2 = 0.35$. Discrimination models have also been constructed for filtering out potential hERG blockers that has reported a capability of 83% correct prediction of hERG blocker and 87% correct prediction of non-blocker compounds [40]. Aronov et al. used 2-dimensional topological similarity filter and achieved 82% *accuracy* in classification of hERG blockers [24]. Roche et al. published a model that produced a classification *accuracy* of 93% and 71% for hERG non-blockers and blockers respectively [41]. Comparison with these results it appears that regarding the performance of classification or discrimination of hERG blockers non-blockers, the present model if not substantially better, is at least at per to the best models. However, in a recent paper significantly higher accuracy (95%) has been reported but in that case a much larger number of molecular descriptors have been used [26].

Table 10
Summary of the statistical parameters obtained in the cross-validation test and randomization test on classification data.

	Cutoff pIC_{50}	*Sensitivity* range	*Specificity* range	*Accuracy* range	*PPV* range	*NPV* range	*MCC* range
Cross-validation	4.4	0.93–0.95	0.67–0.75	0.87–0.90	0.90–0.93	0.71–0.79	0.59–0.68
	5.0	0.83–0.85	0.52–0.63	0.72–0.76	0.76–0.80	0.64–0.71	0.38–0.48
	6.0	0.69–0.82	0.87–0.94	0.83–0.88	0.79–0.88	0.84–0.89	0.63–0.74
Randomization	4.4	0.95–1.00	0.00–0.25	0.77–0.83	0.78–0.83	0.00–1.00	−0.06–0.39
	5.0	0.79–0.91	0.17–0.52	0.56–0.75	0.61–0.75	0.33–0.75	−0.04–0.46
	6.0	0.07–0.41	0.77–0.94	0.56–0.74	0.33–0.80	0.62–0.73	−0.02–0.43

4. Discussions

In principle, all such *In-Silico* predictive models present theoretical representations of reality, based on approximate molecular descriptors, which are never expected to give a completely accurate picture of the real system studied. Another problem may stem from the fact that the experimental data for the training set and the test set compounds are obtained from multiple literatures that generally differ considerably in method, accuracy and other detailed features. This also influences the predictive quality of the generated model. Thus, all these factors together dictate that it is quite unlikely that the predictive power and the accuracy of the resulting QSAR will be very high in general. Moreover, the results based on statistical models are valid in a collective fashion only and generally are not accurate at the individual molecular level. Also the training set dependence of the resulting QSAR leaves the possibility of restricted application and may not be applicable to all varieties of compounds. However, still a single QSAR-based prediction model may provide acceptable results within a limit and there are compounds that are predicted reasonably by one method and not the other.

Considering all the above mentioned intrinsic shortcomings of such statistical methods, the present model, based on simple molecular descriptors derived from 3D structures, seems to be reasonably good and also helps in improving insights. We have further checked that consideration of 3D global descriptors produce more accurate results compared to that based on only 2D features. Availability of much larger number of compounds with more accurate experimental data should further improve the quality of the model generated in the present fashion.

It should be pointed out here that the QSAR was obtained by optimizing the error between the predicted and experimental activities of a set of compounds by adjusting the weights of the different descriptors in the QSAR. Since the optimization was done by considering the QSAR as a whole, and that not all the descriptors used are independent of each other, it is not expected that the contributions of the individual descriptors to the activity will be possible to partition in a meaningful way. Thus, we have not tried to interpret the model at the level of individual descriptors.

5. Conclusions

In the present work an effective QSAR in terms of only simple descriptors derivable directly from the 3D structures of the compounds has been developed and analyzed. The resulting QSAR predicts the hERG activities reasonably well for wide varieties of compounds from its 3D structures only.

It has also been demonstrated that a rather small set of simple 'global molecular descriptors' can predict the hERG activities in a reasonably accurate fashion. The examination of the domain of application has indicated that 89% of the test set compounds has their predicted values in the reliability zone indicating that the prediction of the present QSAR is mostly reliable.

Enrichment effect in the results obtained by applying the QSAR has been demonstrated explicitly. This has not been presented before in QSAR related publications. Cross-validation (leave-one-out test) and the results of external test set compounds have shown the reliability of the QSAR. The application of the method of randomization of experimental data has explicitly demonstrated the robustness of the present QSAR.

Our findings suggest that hERG blockade may occur through different types of binding modes and hence a single QSAR cannot predict the activities of all varieties of compounds in general. The performance of the present QSAR model for classification of compounds appeared to be quite good as judged by the relevant statistical parameters. Finally, our analysis suggests that if identification of hERG non-blocker is required then a classification cutoff value 6.0 for pIC_{50} seems to produce the highest accuracy for the present QSAR. Similarly, for identifying the hERG blockers a cutoff value 4.4 appears to be a better choice.

Acknowledgements

The authors would like to thank Ms. Sohini Basu for discussions and critically reading the manuscript.

Appendix. Supporting information

Supporting information associated with this article can be found, in the online version, at doi:10.1016/j.ejmech.2010.11.042.

References

[1] M.E. Curran, I. Splawski, K.W. Timothy, G.M. Vincent, E.D. Green, M.T. Keating, A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome, Cell 80 (1995) 795–803.

[2] E. Marban, Cardiac channelopathies, Nature 415 (2002) 213–218.

[3] J. Tamargo, R. Caballero, R. Gomez, C. Valenzuela, E. Delpon, Pharmacology of cardiac potassium channels, Cardiovasc. Res. 62 (2004) 9–33.

[4] M.L. De Bruin, M. Pettersson, R.H.B. Meyboom, A.W. Hoes, H.G.M. Leufkens, Anti-HERG activity and the risk of drug-induced arrhythmias and sudden death, Eur. Heart J. 26 (2005) 590–597.

[5] S. Rajamani, L.L. Eckhardt, C.R. Valdivia, C.A. Klemens, B.M. Gillman, C.L. Anderson, K.M. Holzem, B.P. Delisle, B.D. Anson, J.C. Makielski, C.T. January, Drug-induced long QT syndrome: hERG K^+ channel block and disruption of protein trafficking by fluoxetine and norfluoxetine, Br. J. Pharmacol. 149 (2006) 481–489.

[6] J.T. Barbey, R. Lazzara, D.P. Zipes, Spontaneous adverse event reports of serious ventricular arrhythmias, QT prolongation, syncope, and sudden death in patients treated with cisapride, J. Cardiovasc. Pharmacol. Ther. 7 (2002) 65–76.

[7] W.A. Ray, K.T. Murray, S. Meredith, S.S. Narasimhulu, K. Hall, C.M. Stein, Oral erythromycin and the risk of sudden death from cardiac causes, N. Engl. J. Med. 351 (2004) 1089–1096.

[8] P.K. Honig, D.C. Wortham, K. Zamani, D.P. Conner, J.C. Mullin, L.R. Cantilena, Terfenadine–ketoconazole interaction pharmacokinetic and electrocardiographic consequences, JAMA 269 (1993) 1513–1518.

[9] K.S. Lee, E.W. Lee, Ionic mechanism of ibutilide in human atrium: evidence for a drug-induced Na^+ current through a nifedipine inhibited inward channel, J. Pharmacol. Exp. Ther. 286 (1998) 9–22.

[10] I. Shamovsky, S. Connolly, L. David, S. Ivanova, B. Norden, B. Springthorpe, K. Urbahns, Overcoming undesirable hERG potency of chemokine receptor antagonists using baseline lipophilicity relationships, J. Med. Chem. 51 (2008) 1162–1178.

[11] C. Jamieson, E.M. Moir, Z. Rankovic, G. Wishart, Medicinal chemistry of hERG optimizations: highlights and hang-ups, J. Med. Chem. 49 (2006) 5029–5046.

[12] A.M. Aronov, Predictive in-silico modeling for hERG channel blockers, Drug Discov. Today 10 (2005) 149–155.

[13] K.M. Thai, G.F. Ecker, A binary QSAR model for hERG potassium channel blocker, Bioorg. Med. Chem. 16 (2008) 4107–4119.

[14] M. J. Killen, Drug-Induced arrythmias and sudden cardiac death: implications for the pharmaceutical industry, Drug Discov.Today. 14 (2009) 589-597.

[15] B. Fermini, A.A. Fossa, The impact of drug-induced QT interval prolongation on drug discovery and development, Nat. Rev. Drug Discov. 2 (2003) 439–447.

[16] A. Coi, I. Massarelli, L. Murgia, M. Saraceno, V. Calderone, A.M. Bianucci, Prediction of hERG potassium channel affinity by the CODESSA approach, Bioorg. Med. Chem. 14 (2006) 3153–3159.

[17] M. Seierstad, D.K. Agrafiotis, QSAR model of hERG binding using a large, diverse, and internally consistent training set, Chem. Biol. Drug Des. 67 (2006) 284–296.

[18] S. Ekins, W.J. Crumb, R.D. Sarazan, J.H. Wikel, S.A. Wrighton, Three–dimensional quantitative structure–activity relationship for inhibition of human ether-a-go-go-related gene potassium channel, J. Pharmacol. Exp. Ther. 301 (2002) 427–434.

[19] R.A. Pearlstein, R.J. Vaz, J. Kang, X.L. Chen, M. Preobrazhenskaya, A.E. Shchekotikhin, A.M. Korolev, L.N. Lysenkova, O.V. Miroshnikova, J. Hendrixa, D. Rampea, Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches, Bioorg. Med. Chem. Lett. 13 (2003) 1829–1835.

[20] E. Fioravanzo, N. Cazzolla, L. Durando, C. Ferrari, M. Mabilia, R. Ombrato, M.D. Parenti, General and independent approaches to predict HERG affinity values, Internet Electron. J. Mol. Des. 2 (2004).

[21] M. Song, M. Clark, Development and evaluation of an in silico model for hERG binding, J. Chem. Inf. Model. 46 (2006) 392–400.

[22] Q. Li, F.S. Jørgensen, T. Oprea, S. Brunak, O. Taboureau, hERG classification model based on a combination of support vector machine method and grind descriptors, Mol. Pharm. 5 (2008) 117–127.

[23] G. Cianchetta, Y. Li, J. Kang, D. Rampe, A. Fravolini, G. Crucianic, R.J. Vaza, Predictive models for hERG potassium channel blockers, Bioorg. Med. Chem. Lett. 15 (2005) 3637–3642.

[24] A.M. Aronov, B.B. Goldman, A model for identifying HERG K^+ channel blockers, Bioorg. Med. Chem. 12 (2004) 2307–2315.

[25] R. Pearlstein, R. Vaz, D. Rampe, Understanding the structure–activity relationship of the human ether-a-go-go-related gene cardiac K^+ channel. A model for bad behavior, J. Med. Chem. 46 (2003) 2017–2022.

[26] M. Tobita, T. Nishikawa, R. Nagashima, A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors, Bioorg. Med. Chem. Lett. 15 (2005) 2886–2890.

[27] J.S. Mitcheson, J. Chen, M. Lin, C. Culberson, M.C. Sanguinetti, A structural basis for drug-induced long QT syndrome, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 12329–12333.

[28] R. Rajamani, B.A. Tounge, J. Li, C.H. Reynolds, A two-state homology model of the hERG K^+ channel: application to ligand binding, Bioorg. Med. Chem. Lett. 15 (2005) 1737–1741.

[29] R. Farid, T. Day, R.A. Friesner, R.A. Pearlstein, New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies, Bioorg. Med. Chem. 14 (2006) 3160–3173.

[30] hERG inhibition compound set. http://www.lifescienceinformatics.uni-bonn.de/0-Seiten/downloads/down.html.

[31] L. Fausett, Fundamentals of Neural Networks. Prentice Hall, New York, 1994.

[32] Introduction to neural networks. http://www.idsia.ch/NNcourse/intro.html.

[33] Z. Zhou, A.K. Felts, R.A. Friesner, R.M. Levy, Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets, J. Chem. Inf. Model. 47 (2007) 1599–1608.

[34] R.W. Stanforth, E. Kolossov, B. Mirkin, A measure of domain of applicability for QSAR modelling based on intelligent K-means clustering, QSAR Comb. Sci. 26 (2007) 837–844.

[35] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 48 (2008) 1733–1746.

[36] A. Afantitis, G. Melagraki, H. Sarimveis, O.I. Markopoulou, G. Kollias, A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-*N*-aryl-[1, 4] diazepane ureas, Eur. J. Med. Chem. 44 (2009) 877–884.

[37] R. Darnag, E.L.M. Mazouz, A. Schmitzer, D. Villemin, A. Jarid, D. Cherqaoui, Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives, Eur. J. Med. Chem. 45 (2010) 1590–1597.

[38] S. Saha, G.P.S. Raghava, AlgPred: prediction of allergenic proteins and mapping of IgE epitopes, Nucleic Acids Res. 34 (2006) W202–W209.

[39] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[40] G.M. Keseru, Prediction of hERG potassium channel affinity by traditional and hologram QSAR methods, Bioorg. Med. Chem. Lett. 13 (2003) 2773–2775.

[41] O. Roche, G.J.Z. Trube, P. Pflimlin, A. Alanine, G.A. Schneider, Virtual screening method for prediction of the HERG potassium channel liability of compound libraries, ChemBioChem. 3 (2002) 455–459.