

A *k*-nearest neighbor classification of hERG K⁺ channel blockers

Swapnil Chavan¹ · Ahmed Abdelaziz² · Jesper G. Wiklander¹ · Ian A. Nicholls^{1,3} 

Received: 7 December 2015 / Accepted: 28 January 2016 / Published online: 10 February 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract A series of 172 molecular structures that block the hERG K⁺ channel were used to develop a classification model where, initially, eight types of PaDEL fingerprints were used for *k*-nearest neighbor model development. A consensus model constructed using Extended-CDK, PubChem and Substructure count fingerprint-based models was found to be a robust predictor of hERG activity. This consensus model demonstrated sensitivity and specificity values of 0.78 and 0.61 for the internal dataset compounds and 0.63 and 0.54 for the external (PubChem) dataset compounds, respectively. This model has identified the highest number of true positives (i.e. 140) from the PubChem dataset so far, as compared to other published models, and can potentially serve as a basis for the prediction of hERG active compounds. Validating this model against FDA-withdrawn substances indicated that it may even be useful for differentiating between mechanisms underlying QT prolongation.

Keywords Classification model · hERG blockers · Ikr · KCNH2 · *k*-nearest neighbor (*k*-NN) · Toxicity

Abbreviations

CDK	Chemistry development kit
CV	Cross validation
hERG	Human ether-a-go-go-related gene
IUPAC	International union of pure and applied chemistry
<i>k</i> -NN	<i>k</i> -nearest neighbor
MACCS	Molecular ACCESS system
NER	Non-error rate
QSAR	Quantitative structure–activity relationship
SMARTS	SMILES arbitrary target specification
SMILES	Simplified molecular-input line-entry system

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9898-z) contains supplementary material, which is available to authorized users.

✉ Swapnil Chavan
swapnil.chavan@lnu.se

✉ Ian A. Nicholls
ian.nicholls@lnu.se

¹ Bioorganic and Biophysical Chemistry Laboratory, Department of Chemistry and Biomedical Sciences, Linnaeus University Centre for Biomaterials Chemistry, Linnaeus University, 391 82 Kalmar, Sweden

² eADMET GmbH, Lichtenbergstraße 8, 85748 Garching, Munich, Germany

³ Department of Chemistry-BMC, Uppsala University, Box 576, 751 23 Uppsala, Sweden

Introduction

The human ether-a-go-go related gene (hERG, KCNH2) encodes for a voltage dependent K⁺ ion channel (Kv11.1). Blocking of this channel has been associated with potential severe heart arrhythmia, and because of this, several drugs have been withdrawn from the market [1–6]. Further, the drug-induced long QT syndrome may cause avoidable sudden cardiac arrest [3, 4]. With the intention of protecting clinical trial participants and patients, the International Conference of Harmonization published a guideline (S7B) recommending that “all new drugs” should be tested pre-clinically for hERG sensitivity and cardiac safety before submitting an application to regulatory reviews [7]. Accordingly, the early assessment of hERG-related cardiotoxicity has become a common practice in drug discovery.

Many in vitro assays exist for the pre-clinical evaluation of hERG-related cardiotoxicity [8], examples include rubidium-flux assays, radioligand binding assays, in vitro electrophysiology measurements, and fluorescence-based assays [9]. In addition, in silico models have been proposed for identifying potential hERG blockers in drug discovery processes [10, 11].

Efforts to use computational methods for the prediction of hERG blocking effects have ranged from the use of simple rules based on structural and functional features, through to more complex quantitative structure–activity relationship (QSAR) models [12–16]. A number of QSAR models have been developed for the hERG toxicity endpoint using different machine learning algorithms, such as multiple linear regressions [17], partial least squares (PLS) [18], *k*-nearest neighbor algorithms (*k*-NN) [19], artificial neural networks [20], support vector machines (SVM) [21], random forest [22] and naive Bayesian classifications [23]. Despite these efforts there is significant scope for development of more powerful and more easily deployed predictive models.

The recent development of open source fingerprints, such as PaDEL fingerprints, which are libraries of descriptors [24], allows for ready access to tools for predicting biological endpoints. A recent report on the use of PaDEL fingerprints in conjunction with a *k*-NN strategy aimed at the prediction of chronic toxicity [25] prompted us to apply this approach to hERG-channel blockers, a far more focused system. It was envisaged that publicly available data on a series of hERG-channel blockers could function as a starting point for model construction, and a series of 1953 PubChem compounds could act as basis for validation.

Methodology

Description of dataset

IC₅₀ data for 172 I_{kr} (‘rapid’ delayed rectifier current) channel blockers were retrieved from the webserver OCHEM [26] and Fenichel [27]. These 172 compounds are structurally diverse and belong to different therapeutic classes. The compounds were authenticated with respect to structure and IUPAC name. After authentication, the SMILES notations for all the 172 compounds were verified using ChemSpider [28], SigmaAldrich [29] and PubChem [30]. A PubChem dataset comprised of 1953 entries was chosen for the external validation [31]. Dataset entries that were mixtures or salts were discarded, leading to a final PubChem validation set of 1795 compounds. More details about the training and test set compounds are provided in the Online Resources 1 and 2, respectively.

Descriptor calculation

The descriptor calculation was a primary requirement for the construction of the classification model. Eight types of PaDEL fingerprints were calculated for both the training and test set compounds using PaDEL software [24]. These consisted of the CDK, Extended CDK, CDK Graph, Estate, MACCS, PubChem, Sub-structure and Sub-structure count fingerprints. Each of the eight types of fingerprints was then used, separately, to develop a classification model.

Class assignment

The training set compounds were split into one of the two classes (active and inactive) using an IC₅₀ threshold value of 5 μM. The PubChem dataset derived test set compounds were similarly classified, i.e. as either active or inactive, here using a % inhibition threshold of 20 %. A summary of the numbers of the compounds and their classes is provided in Table 1.

Software and modules

The Matlab module “classification_toolbox” [32] was employed for the development of the *k*-NN classification model. The Matlab module is freely available at [33].

Classification model development

The *k*-nearest neighbor (*k*-NN) classification method employed used cross validation (CV) to identify optimal *k* values [34, 35]. A series of *k* values (from 1 to 10) were assigned to construct the model, and by determining the lowest class error, optimal *k* values were identified.

A five-step cross validation was implemented by first dividing the training set into five equal groups, four of which were used for model construction and the remaining for validation. This procedure was repeated so that each of the five groups was used for validating the models constructed using the remaining four. After cross validation, the models were subjected to external validation using the 1795 PubChem compounds. The performance of each classification model was assessed by means of statistical parameters, such as non-error rate (NER), sensitivity, specificity, precision and error rate [36]. The models were

Table 1 Classification of training and test set compounds

	Class 1 (hERG active)	Class 2 (hERG inactive)	Total
Training	93	79	172
Test	221	1574	1795

then analysed and compared on the basis of these statistical parameters.

Results and discussion

Construction of eight k -NN classification models

The k -nearest neighbor (k -NN) classification method was employed to construct classification models using each of the eight PaDEL fingerprints. Employing the k -NN algorithm requires that the optimal value of k is determined [34]. There are several ways to determine the k value, e.g. through application of a risk function or empirical rules, or

through cross validation. Here, cross validation was used to determine the optimal k value.

A series of eight k -NN classification models was constructed using each of the PaDEL fingerprints, and compared with respect to a series of statistical parameters, Table 2.

CDK fingerprints are one-dimensional 1024 bit long arrays that are arranged based upon the occurrence of particular structural elements. The Extended CDK fingerprints are extended versions of CDK fingerprints that include ring features. Graph fingerprints are specialized versions of the CDK fingerprints that exclude bond orders. Estate fingerprints represent the influence of substituent electronic effects in a given compound. PubChem

Table 2 Summary of statistical parameters for the k -NN classification models

Entry	Fingerprints	NER	k	Sensitivity		Specificity	
				Class 1	Class 2	Class 1	Class 2
1	<i>CDK</i>						
	Fitting	0.68	1	0.72	0.65	0.65	0.72
	CV	0.66	1	0.72	0.61	0.61	0.72
2	<i>External</i>	0.54	1	0.52	0.57	0.57	0.52
	<i>Estate</i>						
	Fitting	0.68	1	0.73	0.62	0.62	0.73
3	CV	0.66	1	0.72	0.61	0.61	0.72
	External	0.53	1	0.49	0.57	0.57	0.49
	<i>Extended CDK</i>						
4	Fitting	0.67	1	0.70	0.63	0.63	0.70
	CV	0.65	1	0.70	0.61	0.61	0.70
	External	0.56	1	0.56	0.57	0.57	0.56
5	<i>CDK graph</i>						
	Fitting	0.64	1	0.69	0.59	0.59	0.69
	CV	0.64	1	0.70	0.58	0.58	0.70
6	External	0.55	1	0.52	0.57	0.57	0.52
	<i>MACCS</i>						
	Fitting	0.68	6	0.76	0.59	0.59	0.76
7	CV	0.67	6	0.76	0.57	0.57	0.76
	External	0.55	6	0.54	0.55	0.55	0.54
	<i>PubChem</i>						
8	Fitting	0.60	3	0.69	0.52	0.52	0.69
	CV	0.60	3	0.71	0.49	0.49	0.71
	External	0.57	3	0.62	0.52	0.52	0.62
9	<i>Sub-structure</i>						
	Fitting	0.68	1	0.70	0.67	0.67	0.70
	CV	0.67	1	0.69	0.66	0.66	0.69
10	External	0.57	1	0.54	0.59	0.59	0.54
	<i>Sub-structure count</i>						
	Fitting	0.67	1	0.74	0.61	0.61	0.74
11	CV	0.68	1	0.72	0.65	0.65	0.72
	External	0.58	1	0.61	0.56	0.56	0.61

Table 3 Statistical parameters for the consensus models

Model ^a	Dataset	TP ^b	FP ^c	TN ^d	FN ^e	TP + TN	Total ^f	Q ^g	Sens. ^h	Spec. ⁱ	Prec. ^j	G-mean ^k
1	Training	72	25	54	21	126	172	0.73	0.77	0.68	0.74	0.73
	Validation	130	654	920	91	1050	1795	0.58	0.59	0.58	0.17	0.59
2	Training	73	31	48	20	121	172	0.70	0.78	0.61	0.70	0.69
	Validation	140	723	851	81	991	1795	0.55	0.63	0.54	0.16	0.59
3	Training	71	31	48	22	119	172	0.69	0.76	0.61	0.70	0.68
	Validation	135	707	867	86	1002	1795	0.56	0.61	0.55	0.16	0.58
4	Training	74	32	47	19	121	172	0.70	0.80	0.59	0.70	0.69
	Validation	128	718	856	93	984	1795	0.55	0.58	0.54	0.15	0.56
5	Training	73	29	50	20	123	172	0.72	0.78	0.63	0.72	0.70
	Validation	132	685	889	89	1021	1795	0.57	0.60	0.56	0.16	0.58
6	Training	73	28	51	20	124	172	0.72	0.78	0.65	0.72	0.71
	Validation	131	675	899	90	1030	1795	0.57	0.59	0.57	0.16	0.58

^a Model 1 = substructure (SS) + substructure count (SSC) + extended CDK (ECDK), 2 = PubChem (PC) + SSC + ECDK, 3 = PC + SSC + SS, 4 = PC + SSC + MACCS, 5 = PC + SSC + ECDK + SC + MACCS, 6 = PC + SSC + ECDK + SS + MACCS + CDK + CDK Graph, ^b true positives, ^c false positives, ^d true negatives, ^e false negatives, ^f TP + TN + FP + FN, ^g overall accuracy of prediction, ^h sensitivity, ⁱ specificity, ^j precision, ^k $\sqrt{\text{Sensitivity} \times \text{Specificity}}$

fingerprints are binary substructure fingerprints of length 881. MACCS fingerprints consist of 166 keys that are based on SMARTS patterns [37, 38]. The Sub-structure fingerprints represent 307 SMARTS patterns for different functional groups, whereas the count of these SMARTS patterns is referred to as the Sub-structure count fingerprint [37].

The sensitivity expresses the prediction accuracy of hERG-active compounds, whereas specificity reflects the prediction accuracy for hERG-inactive compounds. The models performed similarly in terms of the statistical parameters examined. Thus, to further improve the predictive power of these models we developed a series of consensus models. Several methods have been reported for consensus model development [39]. For classification models, the majority principle [40] is commonly employed and we have used this strategy to develop consensus models based upon three, five and seven different fingerprint-based models. As it is more important to identify hERG-active compounds than hERG-inactive compounds, the eight models (from Table 2) were examined with respect to their sensitivity in the external prediction. The Estate-fingerprint-based model exhibited relatively poor sensitivity (0.49) and was discarded from the consensus model building procedure to provide an odd number (seven) of fingerprints. Six consensus models were built using different combinations of the seven remaining fingerprint-based models, Table 3.

Although consensus model 1 shows better overall accuracy of prediction (Q), consensus model 2 shows higher sensitivity for test set prediction, and was thus chosen for further studies.

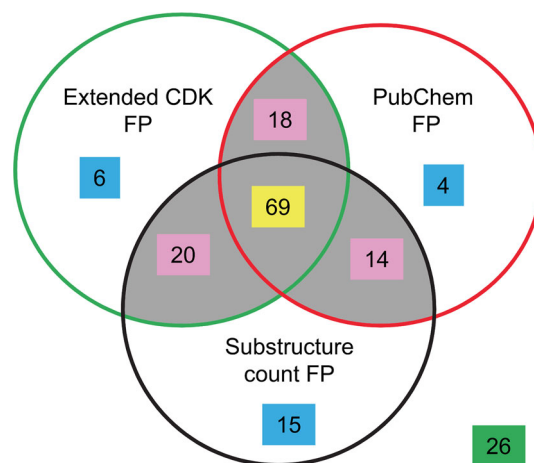


Fig. 1 Venn diagram representing the number of training set compounds correctly predicted by all three models (yellow), by any two models (magenta), by only one model (blue) and by none of the models (green). The shaded area represents compounds correctly predicted by the consensus model

Individual contribution of each model

With consensus model 2 in hand, we then examined how individual training set compounds were handled by the consensus model as well as the individual models, i.e. Extended CDK, PubChem and Substructure count fingerprint based, Fig. 1.

The consensus model correctly predicted 121 of the 172 training set compounds. 69 of these 121 compounds were predicted correctly by all three individual models, while the remaining 52 compounds were correctly predicted by any two of the three models. Conversely, the consensus

Table 4 Comparison of the *k*-NN classification model with other models

Model Method	Our study <i>k</i> -NN	Su et al. [42] SVM	Wang et al. [43] Naive Bayesian classifier	Su et al. [18] PLS transformed into binary QSAR	Li et al. [44] SVM
Descriptors	2D PaDEL fingerprints	2D and 3D MOE, 4D fingerprints from MD simulation	Physico-chemical property based and geometry based descriptors, and fingerprints	2D and 3D MOE descriptors and 4D fingerprints	GRIND descriptors derived from docking
<i>Training set</i>					
Cut-off (μM)	5	–	10	40	40
Total	172	546	719	250	495
True positives	73	188	247	–	83
True negatives	48	242	315	–	283
Sensitivity	0.78	0.90	0.89	–	0.55
Specificity	0.61	0.72	0.72	–	0.83
Q	0.70	0.79	0.78	–	0.74
F-measure ^a	0.74	0.76	0.76	–	0.56
G-mean	0.69	0.80	0.80	–	0.67
<i>Test set</i>					
Cut-off (%) ^b	20	20	20	20	20
Total	1795	1668	1953	1668	1877
True positives	140	67	135	121	107
True negatives	851	1298	1247	963	1271
Sensitivity	0.63	0.41	0.54	0.74	0.57
Specificity	0.54	0.86	0.73	0.64	0.75
Q	0.55	0.82	0.71	0.65	0.73
F-measure	0.26	0.31	0.32	0.29	0.30
G-mean	0.59	0.60	0.63	0.69	0.66

^a $2[(\text{precision} \times \text{sensitivity}) / (\text{precision} + \text{sensitivity})]$, ^b % hERG blockage

model incorrectly predicted 51 training set compounds. Of these 51, 25 compounds were predicted correctly by any one of the three models, whereas the remaining 26 compounds were incorrectly predicted by all three models.

In the case of the Extended fingerprint based model, 113 of 172 compounds were correctly predicted, 65 of which were hERG actives. The PubChem fingerprint based model predicted 105 compounds correctly from the training set. Among the 105 correctly predicted compounds, 66 were from class 1 and 39 from class 2. The Substructure count fingerprint based model predicted 118 training set compounds correctly. These 118 compounds were comprised of 67 compounds from class 1 and 51 compounds from class 2.

Compounds for which activities were not correctly predicted by our models are of interest as awareness of factors contributing to the incorrect prediction of compounds can help in the refinement of models. In this case, the IC₅₀ value-based endpoints are derived from a range of studies so impact of inter-laboratory variation in the reported IC₅₀ data on model performance cannot be excluded.

Comparison of our model with other models

External validation provides an assessment of the QSAR model's performance, and to compare models it is necessary that the external validations are performed on the same dataset. The PubChem dataset is comprised of 221 hERG-actives and 1574 hERG-inactives. Sensitivity and specificity are generally used to assess classification performance in imbalanced binary class studies [41]. G-mean, which is a geometric mean of sensitivity and specificity, was also used to measure the performance of the classification method in predicting actives and inactives. In studies aimed at the effective detection of only one class, as in our case where the prediction of hERG-actives is a priority, sensitivity and F-measures are often adopted [41]. Accordingly, we have compared our model with previously published models that were externally validated with the PubChem dataset [18, 42–44], with respect to sensitivity, specificity, G-mean and F-measure, Table 4.

As presented in Table 4, three of the four previously described models demonstrate lower overall sensitivities than our model, though it should be pointed out that IC₅₀

thresholds used in the various studies varied between 5 and 40 μM . From a drug development perspective, it may be argued that it is of more interest to identify the potent hERG blockers (class 1) than hERG inactive compounds (class 2). Comparison on this point reveals that our model demonstrates better performance in predicting the hERG active compounds (True positives = 140, Sensitivity = 0.63) than the other models except that of Su et al. [18] in their model presented 2010. There, 163 hERG actives from the PubChem dataset were used for the external validation, whereas in our study a somewhat more comprehensive external validation was performed using 221 hERG actives.

From a practical perspective, ease of use is an issue of importance and an advantage of our model is that PaDEL fingerprints are fast and easy to calculate and do not involve complicated descriptor selection procedures. This is in contrast with all the other models presented in Table 4 that all employed 3D and 4D descriptors that require geometry optimization, a task necessitating significant computational resources. In addition, the application of different descriptor selection procedures makes these tasks more cumbersome. Therefore, in comparison to the other models, our model has the advantage of being fast, simple and relatively efficient in predicting hERG toxic compounds.

To further assess the potential of our consensus model, we turned our attention to the series of 47 substances withdrawn from use on account of QT-prolongation, which can be hERG-derived, as present in the WITHDRAWN database [45] (database last updated December 2015). Our training set had included 32 of these 47 drugs (shown in bold in Online Resource 1) of which our model had correctly predicted the IC_{50} -based classes of 22. We interrogated the remaining 15 withdrawn substances (see Online Resource 3) using our model, which correctly predicted the IC_{50} -based classes of 11 (73 %, see Online Resource 4). It is important to note that our model is solely based upon in vitro data (hERG IC_{50}), while the basis for withdrawal, QT prolongation, is in vivo data-derived. The interpretation of the QT prolongation endpoint is itself a major challenge as mechanisms other than hERG activity can also underlie QT prolongation [4, 46, 47]. This is reflected in the fact that substances were correctly classified as class 1 or class 2, five and six substances respectively, based on their hERG IC_{50} . This observation suggests that the model may even be useful for differentiating between mechanisms underlying QT prolongation.

A general reflection upon examining the hERG active compounds predicted by our model was the prevalence of aromatic and basic functionalities in these compounds (for example, see Online Resource 2). These features have previously been identified as essential components in a pharmacophore for central nervous system activity [48, 49] and we believe should be considered in future model

development. Moreover, this may be considered indicative of a common evolutionary origin for the hERG voltage dependent K^+ ion channel and CNS receptors [50, 51].

Conclusion

In conclusion, PaDEL fingerprint-based *k*-NN classification models presented here show potential as tools for the prediction of the hERG toxicity endpoint, an important issue in modern drug development. In particular, the consensus model developed using the Extended CDK, PubChem and Sub-structure count fingerprint-based models performed comparably with models employing more complicated descriptors in the validation with external datasets. Moreover, the model presented here, in terms of the prediction of hERG toxicity, compares most favorably with these previously published models. Moreover, validating this model against FDA-withdrawn substances indicates that the model may be useful for differentiating between hERG-derived QT prolongation and other QT prolongation mechanisms. Accordingly, we believe that this model may provide a basis for improved drug design.

Acknowledgments We acknowledge financial support from the EU FP-7 Environmental Chemoinformatics (ECO) project (Grant Number-238701) and Linnaeus University, Sweden, and express our sincere thanks to Dr. Igor Tetko for valuable advice, comments and guidance during this work. The authors also thank Dr. Yuri Sushko, Dr. Robert Körner and Dr. Sergii Novotarskyi from eADMET, Germany, for their assistance with data collection and technical support. Finally, the authors sincerely thank Prof. Roberto Todeschini (Chemometrics and QSAR research group, University of Milan, Italy) for sharing the classification_toolbox Matlab routines for the *k*-NN model development.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Warmke JW, Ganetzky B (1994) A family of potassium channel genes related to eag in *Drosophila* and mammals. *Proc Natl Acad Sci* 91(8):3438–3442
2. Choe H, Nah KH, Lee SN, Lee HS, Jo SH, Leem CH, Jang YJ (2006) A novel hypothesis for the binding mode of HERG channel blockers. *Biochem Biophys Res Commun* 344(1):72–78

- Raschi E, Ceccarini L, De Ponti F, Recanatini M (2009) hERG-related drug toxicity and models for predicting hERG liability and QT prolongation. *Expert Opin Drug Metab Toxicol* 5(9):1005–1021
- Redfern W, Carlsson L, Davis A, Lynch W, MacKenzie I, Palethorpe S, Siegl P, Strang I, Sullivan A, Wallis R (2003) Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development. *Cardiovasc Res* 58(1):32–45
- De Ponti F, Poluzzi E, Montanaro N (2000) QT-interval prolongation by non-cardiac drugs: lessons to be learned from recent experience. *Eur J Clin Pharmacol* 56(1):1–18
- Meyer T, Boven KH, Günther E, Fejtł M (2004) Micro-electrode arrays in cardiac safety pharmacology. *Drug Saf* 27(11):763–772
- Darpo B, Nebout T, Sager PT (2006) Clinical evaluation of QT/QTc prolongation and proarrhythmic potential for nonantiarrhythmic drugs: the international conference on harmonization of technical requirements for registration of pharmaceuticals for human use E14 guideline. *J Clin Pharmacol* 46(5):498–507
- Mitcheson JS (2008) hERG potassium channels and the structural basis of drug-induced arrhythmias. *Chem Res Toxicol* 21(5):1005–1010
- Polak S, Wiśniowska B, Brandys J (2009) Collation, assessment and analysis of literature in vitro data on hERG receptor blocking potency for subsequent modeling of drugs' cardiotoxic properties. *J Appl Toxicol* 29(3):183–206
- Cavalli A, Poluzzi E, De Ponti F, Recanatini M (2002) Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K⁺ channel blockers. *J Med Chem* 45(18):3844–3853
- Wang S, Li Y, Xu L, Li D, Hou T (2013) Recent developments in computational prediction of HERG blockage. *Curr Top Med Chem* 13(11):1317–1326
- Perry M, Stansfeld PJ, Leaney J, Wood C, de Groot MJ, Leishman D, Sutcliffe MJ, Mitcheson JS (2006) Drug binding interactions in the inner cavity of HERG channels: molecular insights from structure–activity relationships of clofilium and ibutilide analogs. *Mol Pharmacol* 69(2):509–519
- Sánchez-Chapula JA, Ferrer T, Navarro-Polanco RA, Sanguinetti MC (2003) Voltage-dependent profile of human ether-a-go-go-related gene channel block is influenced by a single residue in the S6 transmembrane domain. *Mol Pharmacol* 63(5):1051–1058
- Milnes JT, Crociani O, Arcangeli A, Hancox JC, Witchel HJ (2003) Blockade of HERG potassium currents by fluvoxamine: incomplete attenuation by S6 mutations at F656 or Y652. *Br J Pharmacol* 139(5):887–898
- Kamiya K, Niwa R, Mitcheson JS, Sanguinetti MC (2006) Molecular determinants of HERG channel block. *Mol Pharmacol* 69(5):1709–1716
- Aronov AM (2005) Predictive in silico modeling for hERG channel blockers. *Drug Discov Today* 10(2):149–155
- Pourbasheer E, Beheshti A, Khajehsharifi H, Ganjali MR, Norouzi P (2013) QSAR study on hERG inhibitory effect of kappa opioid receptor antagonists by linear and non-linear methods. *Med Chem Res* 22(9):4047–4058
- Su BH, Shen MY, Esposito EX, Hopfinger AJ, Tseng YJ (2010) In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J Chem Inf Model* 50(7):1304–1318
- Gunturi SB, Archana K, Khandelwal A, Narayanan R (2008) Prediction of hERG potassium channel blockage using kNN-QSAR and local lazy regression methods. *QSAR Comb Sci* 27(11–12):1305–1317
- Thai KM, Ecker GF (2009) Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Mol Divers* 13(3):321–336
- Yap C, Cai C, Xue Y, Chen Y (2004) Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci* 79(1):170–177
- Wiśniowska B, Mendyk A, Polak M, Szłęk J, Polak S (2010) Randomforest based assessment of the hERG channel inhibition potential for the early drug cardiotoxicity testing. *BAMS* 6:131–136
- Sun H (2006) An accurate and interpretable Bayesian classification model for prediction of hERG liability. *Chem Med Chem* 1(3):315–322
- Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32(7):1466–1474
- Chavan S, Friedman R, Nicholls IA (2015) Acute toxicity-supported chronic toxicity prediction: a k-nearest neighbor coupled read-across strategy. *Int J Mol Sci* 16(5):11659–11677
- Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25(6):533–554
- Fenichel dataset. <http://www.fenichel.net/pages/Professional/subpages/QT/Tables/pbydrug.htm>. Accessed 11 Sept 2015
- Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87(11):1123–1124
- Lenga RE, Votoupal KL (1993) The Sigma-Aldrich library of regulatory and safety data, vol 3. Aldrich Chemical Company, Wisconsin
- Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem* 4:217–241
- PubChem Bioassay: hERG channel activity. <https://pubchem.ncbi.nlm.nih.gov/assay/assaydata.html?aid=376>. Accessed 11 Sept 2015
- Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal. Methods* 5(16):3790–3798
- Classification Toolbox. <http://micchem.disat.unimib.it/chm/download/classificationinfo.htm>. Accessed 11 Sept 2015
- Kowalski B, Bender C (1972) k-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal Chem* 44(8):1405–1411
- Chavan S, Nicholls IA, Karlsson BC, Rosengren AM, Ballabio D, Consonni V, Todeschini R (2014) Towards global QSAR model building for acute toxicity: Munro database case study. *Int J Mol Sci* 15(10):18162–18174
- Ballabio D, Todeschini R (2009) In: Sun D-W (ed) *Infrared spectroscopy for food quality analysis and control*. Elsevier, Amsterdam, p 2009
- Chem Des. Molecular fingerprints library. <http://www.scbdd.com/chemdes/list-fingerprints/>. Accessed 11 Sept 2015
- Daylight Chemical Information Systems theory manual. <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html>. Accessed 11 Sept 2015
- Mansouri K, Ringsted T, Ballabio D, Todeschini R, Consonni V (2013) Quantitative structure–activity relationship models for ready biodegradability of chemicals. *J Chem Inf Model* 53(4):867–878
- Pavan M, Worth A, Netzeva T (2015) Preliminary analysis of an aquatic toxicity dataset and assessment of QSAR models for narcosis. https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/information-sources/qsar-document-area/Report_QSAR_

- [model_for_narcosis.pdf](#). Joint research center, European Commission, Ispra, Italy, 2005. Accessed 5 Nov 2015
41. Tang Y, Zhang YQ, Chawla NV, Krasser S (2009) SVMs modeling for highly imbalanced classification. *IEEE Trans Syst Man Cybern* 39(1):281–288
 42. Su BH, Tu YS, Esposito EX, Tseng YJ (2012) Predictive toxicology modeling: protocols for exploring hERG classification and *Tetrahymena pyriformis* end point predictions. *J Chem Inf Model* 52(6):1660–1673
 43. Wang S, Li Y, Wang J, Chen L, Zhang L, Yu H, Hou T (2012) ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol Pharm* 9(4):996–1010
 44. Li Q, Jørgensen FS, Oprea T, Brunak S, Taboureau O (2008) hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol Pharm* 5(1):117–127
 45. WITHDRAWN: A resource for withdrawn and discontinued drugs. <http://cheminfo.charite.de/withdrawn/>. Accessed 26 Jan 2016
 46. Gupta A, Lawrence AT, Krishnan K, Kavinsky CJ, Trohman RG (2007) Current concepts in the mechanisms and management of drug-induced QT prolongation and torsades de pointes. *Am Heart J* 153(6):891–899
 47. Yap YG, Camm AJ (2003) Drug induced QT prolongation and torsades de pointes. *Heart* 89(11):1363–1372
 48. Lloyd EJ, Andrews PR (1986) A common structural model for central nervous system drugs and their receptors. *J Med Chem* 29(4):453–462
 49. Andrews P, Lloyd E (1983) A common structural basis for CNS drug action. *J Pharm Pharmacol* 35(8):516–518
 50. Moran Y, Barzilai MG, Liebeskind BJ, Zakon HH (2015) Evolution of voltage-gated ion channels at the emergence of Metazoa. *J Exp Biol* 218(4):515–525
 51. Ranganathan R (1994) Evolutionary origins of ion channels. *Proc Natl Acad Sci* 91(9):3484–3486