Internet Electronic Journal of Molecular Design **2004**, *2*, 000–000
BioChem Press http://www.biochempress.com

General and independent approaches to predict HERG affinity values

Elena Fioravanzo[1,*], Nicola Cazzolla[2], Lucia Durando[2], Cristina Ferrari[1], Massimo Mabilia[1], Rosella Ombrato[2], Marco Daniele Parenti[1]

[1] S-IN Soluzioni Informatiche Via Salvemini 9, 36100, Vicenza, Italy (www.s-in.it)
[2] A.C.R.A.F. S.p.A, P.le Stazione I-00040, S. Palomba, Rome, Italy

Received xxx; Preprint published xxx; Accepted xxx ; Published xxx

Internet Electron. J. Mol. Des. **2004**, *1*, 000–000

Abstract

Motivation. The protein product of the human ether-a-go-go gene (hERG) is a potassium channel that when inhibited may lead to cardiac arrhythmia. At present various in vivo and in vitro models for QT prolongation and subsequent arrhythmia exist but they may not be entirely predictive for humans. Consequently a fast and reliable in silico screen for the prediction of hERG affinity values would increase the screening rate and would also lower the cost compared to experimental assay methods.

Method. In this communication different approaches were employed to predict hERG K+ channel affinities. First of all different QSAR models were developed employing various molecular descriptors. Then independent software were used to predict hERG activity values: Qikprop and PASS. The software QikProp (Schrödinger, L.L.C) allows to predict pharmaceutically relevant properties for organic molecules, starting from their 3D structures and employing calculated physically significant descriptors. In addition to cell permeability, logP, solubility, blood/brain barrier permeability, the program can also predict hERG K+ channel affinity values. As an independent approach, the program PASS - Prediction of Activity Spectra for Substances - (V. Poroikov, D. Filimonov & Associates) that can predict several hundreds biological activity probability values, such as pharmacological effects, mechanisms of action, toxicity and metabolism reactions, was trained to predict the probability of hERG activity.

Conclusions. The availability of different and independent methods and models able to predict hERG activity allow the application of a consensus criterion to be used as a filter in the discovery process.

Availability. DRAGON (Talete, srl,) www.telemacus.it - MODDE (Umetrics AB) www.umetrics.com - PASS www.ibmh.msk.su/PASS/ - QikProp (Schrödinger, L.L.C) www.schrodinger.com - SIMCA-P+ (Umetrics AB) www.umetrics.com - Spartan'02 (Wavefunction, Inc) www.wavefun.com -

Keywords. ADMET, HERG, QSAR, molecular descriptors.

Abbreviations and notations

EVA, Eigen VAlues	PLS, Partial Least Squares
OSC, Orthogonal Signal Correction	QSAR, Quantitative Structure-Activity Relationships
PCA, Principal Component Analysis	VIP, Variable Importance Plot

* Correspondence author; phone: ++39-0444-240341; fax: ++39-0444-533954; E-mail: elena.fioravanzo@s-in.it

1 INTRODUCTION

Drug-induced QT interval prolongation, as measured on the human electrocardiogram, was once considered a trivial physiological finding. Now it is believed that drug-induced QT interval prolongation, that has been identified as a critical side effect for numerous drugs, might result in sudden cardiac death. As a consequence, a number of prescription medications associated with QT prolongation have been removed from the market. The normal quasiperiodic electrical activity of the heart is the result of the flow of ions through channels in the membranes of myocardial cells. Drugs affect ventricular repolarization by interfering with the opening and closing of these channels. The focus of many in vitro studies to date is the membrane-bound inward (rapid activating delayed) rectifier potassium channel (IK_r) also known as the product of the human ether-a-go-go gene (hERG). Drugs or their metabolites may block this channel, thereby prolonging the QT interval and in same cases leading to the potentially life-threatening ventricular arrhythmia that may degenerate into ventricular fibrillation and sudden death. At present blockade of hERG K^+ channel is an unwanted side effect that must be detected as early as possible during drug development [1].

Since various in vivo and in vitro models for QT prolongation and subsequent arrhythmia exist but they may not be entirely predictive for humans, the availability of in silico methods in the early phase of drug development would dramatically increase the screening rate and would also lower the costs compared to experimental assay methods. The possibility of a computational hERG model to be used as a filter in the discovery process would add an extra dimension to lead optimization. Both a quantitative and a qualitative model would theoretically enable virtual selection of candidates with the lowest potential to cause hERG inhibition.

Recent studies on hERG K^+ channels involve pharmacophore mapping and CoMFA study. Both approaches, however, are based on the assumption that different compounds bind to the same binding site of the channel using similar binding modes. On the contrary, it is reasonable to assume that the binding affinity of a given compound may vary as a function of the channel states (activated/inactivated), and that structurally diverse molecules may adopt different binding modes. Such considerations are not compatible with a single pharmacophore model nor with a common alignment criterion.

In this study different and independent computational approaches are used to predict hERG K^+ channel affinities in order to allow a consensus criterion in classify compounds as active or inactive towards hERG K^+ channel.

2 MATERIALS AND METHODS

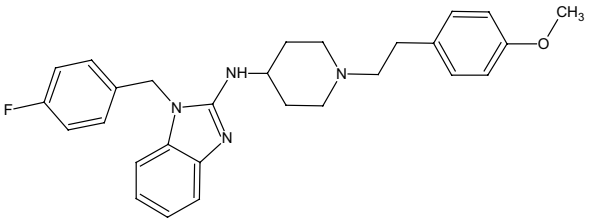
2.1 Chemical Data

Table 1 – pIC_{50} values of the data set.

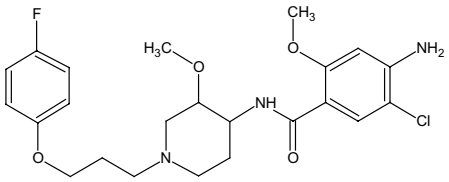
Primary ID	No.	pIC_{50} *	pIC_{50} **	pIC_{50} ***	Primary ID	No.	pIC_{50} *	pIC_{50} **	pIC_{50} ***
astemizole	**1**	9	-	8	imipramine	**36**	5.5	5.5	5.5
cisapride	**2**	8.2	8.2	7.4	granisetron	**37**	5.4	5.4	-
E-4031	**3**	8.1	7.7	7.7	flecainide	**38**	-	-	5.4
ibutilide	**4**	-	-	8	citalopram	**39**	-	-	5.4
dofetilide	**5**	7.9	-	8	norclozapine	**40**	-	-	5.4
sertindole	**6**	7.9	7.8	8	mefloquine	**41**	-	-	5.3
pimozide	**7**	7.7	7.3	7.3	cocaina	**42**	5.2	-	5.1
haloperidol	**8**	7.6	7.6	7.5	dolasetron	**43**	5.2	4.9	4.9
norastemizole	**9**	7.6	-	7.6	perhexiline	**44**	5.1	5.1	5.1
droperidol	**10**	7.5	-	7.5	amitriptyline	**45**	-	-	5
thioridazine	**11**	7.5	-	6.4	nitrendipine	**46**	-	-	5
terfenadine	**12**	6.9	6.7	6.7	amiodarone	**47**	-	-	5
verapamil	**13**	6.8	6.8	6.9	2-Hydroxymethyl olanzapine	**48**	-	4.9	-
ziprasidone	**14**	6.8	6.9	6.9	carvediol	**49**	-	-	4.9
domperidone	**15**	6.8	-	-	desmethyl olanzepine	**50**	-	4.9	
risperidone	**16**	6.8	6.8	6.8	diltiazem	**51**	4.8	4.8	4.8
loratadine	**17**	6.8	-	6.8	chlorpheniramine	**52**	-	-	4.7
clozapine	**18**	6.7	6.5	6.5	fexofenadine	**53**	4.7	-	-
halofantrine	**19**	6.7	6.7	6.7	sparfloxacin	**54**	4.6	-	4.7
olanzapine	**20**	-	6.6	6.7	diphenhydramine	**55**	-	-	4.6
terikalant	**21**	-	-	6.6	cetirizine	**56**	-	-	4.5
mesoridazine	**22**	-	6.5	6.5	N-des methylclozapine	**57**	-	4.5	-
quinidine	**23**	-	-	6.5	A 56268	**58**	-	-	4.5
mizolastine	**24**	6.5	-	6.4	nifedipine	**59**	-	-	4.3
bepridil	**25**	6.3	-	6.3	glibenclamide	**60**	4.1	-	-
azimilide	**26**	6.3	-	5.9	grepafloxacin	**61**	4.1	-	4.3
ondansetron	**27**	-	6.1	6.1	disopyramide	**62**	-	-	4
vesnarinone	**28**	-	6	6	sildenafil	**63**	4	5.5	5.5
9-OH risperidone	**29**	-	5.9	-	epinastine	**64**	-	-	4
desipramine	**30**	-	5.9	5.9	moxifloxacin	**65**	3.9	-	3.9
mibefradil	**31**	5.8	-	5.8	gatifloxacin	**66**	3.9	-	3.9
chlorpromazine	**32**	5.8	5.8	5.8	trimethoprin	**67**	-	-	3.6
fluoxetine	**33**	-	-	5.8	nicotine	**68**	-	3.6	3.6
ketoconazole	**34**	-	-	5.7	levofloxacin	**69**	-	-	3
alosetron	**35**	-	5.5	5.5	ciprofloxacin	**70**	-	-	3

* Experimental data from [2a] ** Experimental data from [2c] *** Experimental data from [2b]

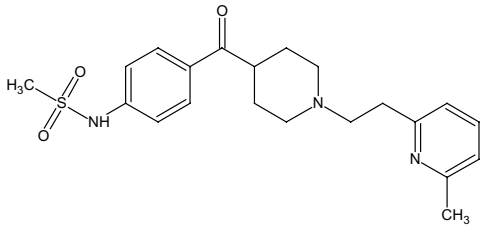
Figure 1 – Structures of the data set



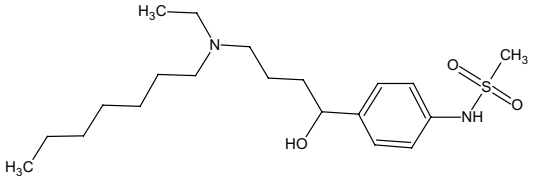
1



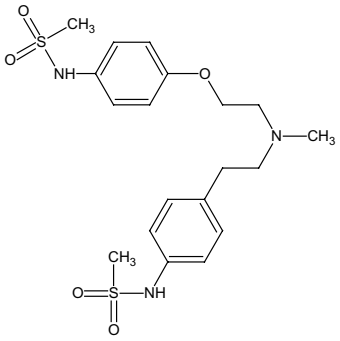
2



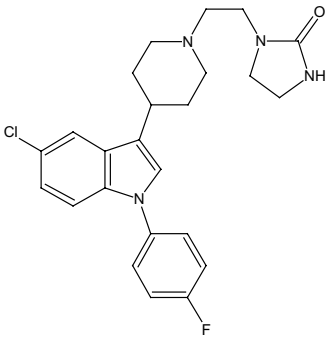
3



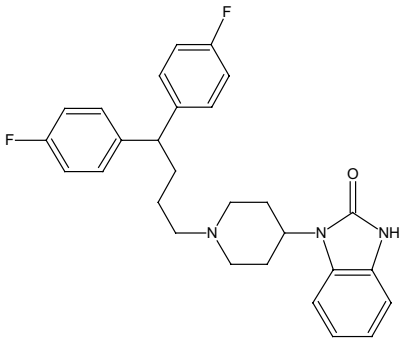
4



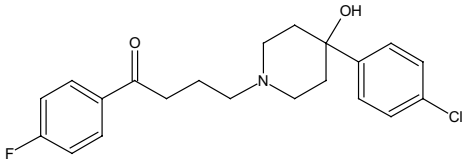
5



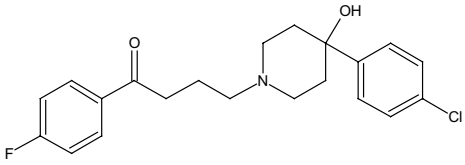
6



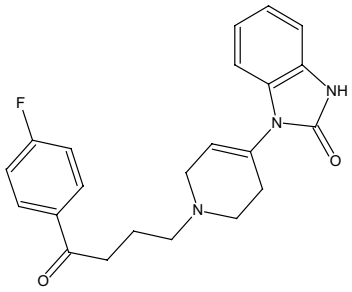
7



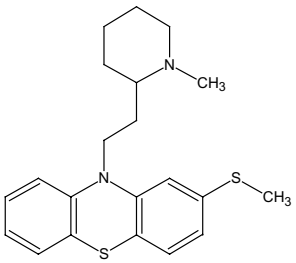
8



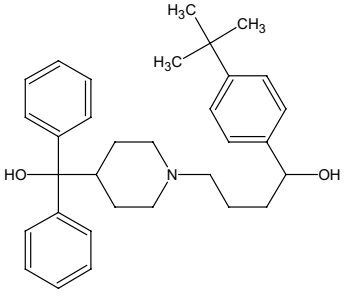
9



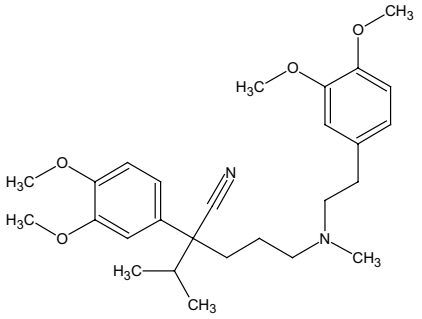
10



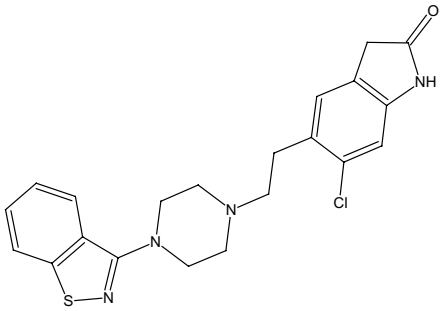
11



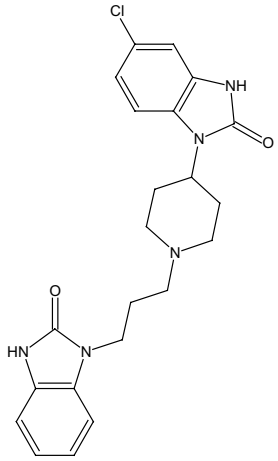
12



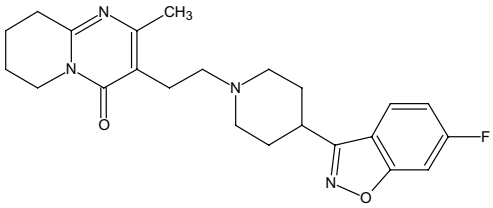
13



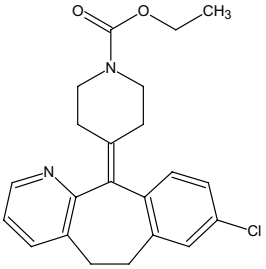
14



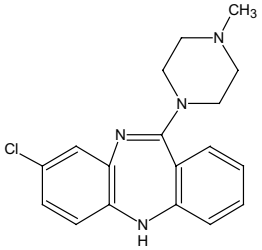
15



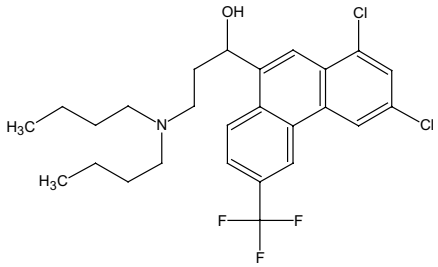
16



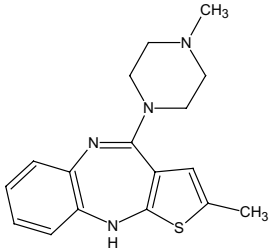
17



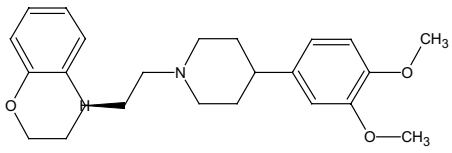
18



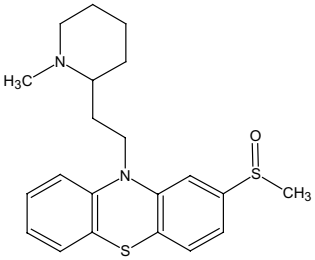
19



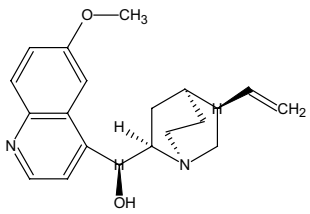
20



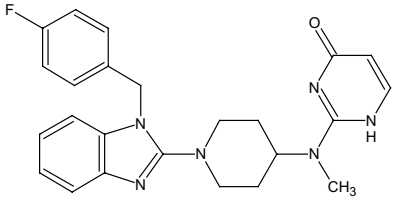
21



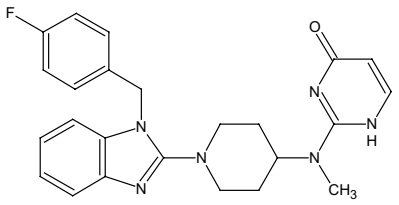
22



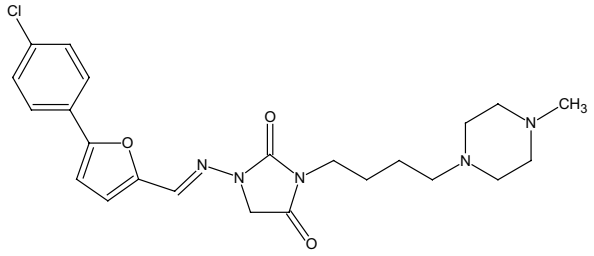
23



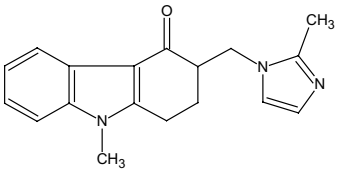
24



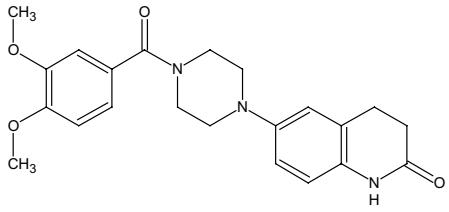
25



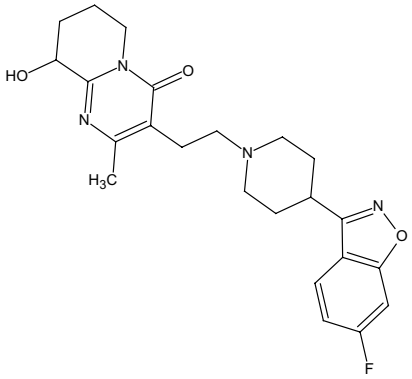
26



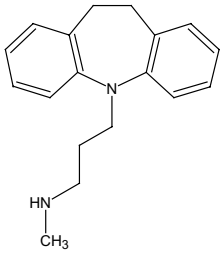
27



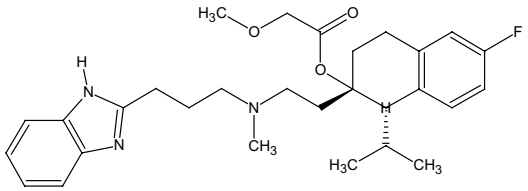
28



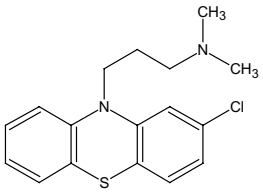
29



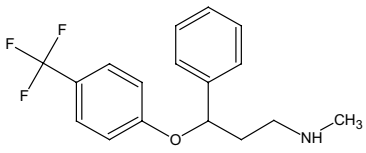
30



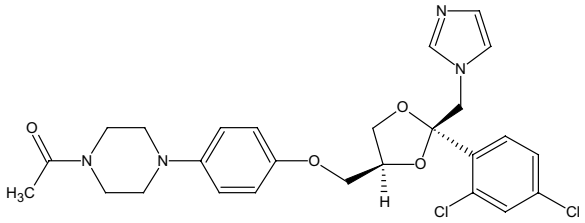
31



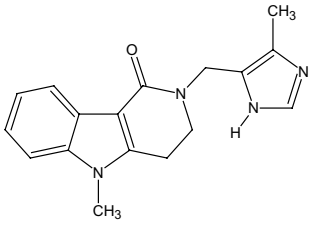
32



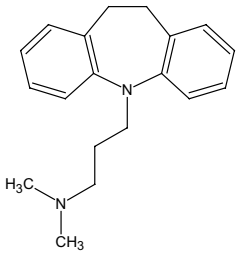
33



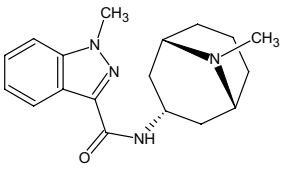
34



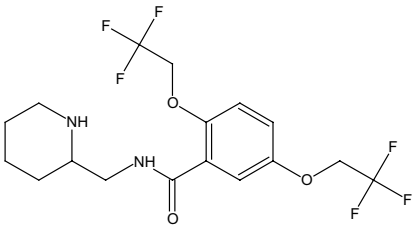
35



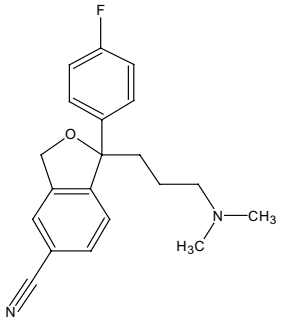
36



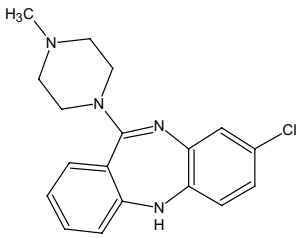
37



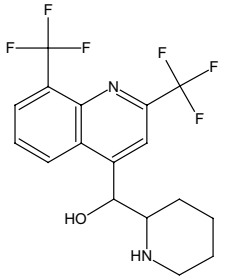
38



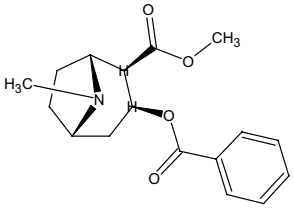
39



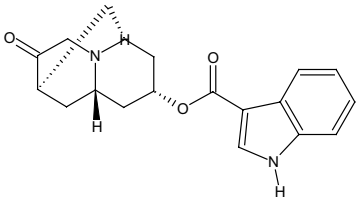
40



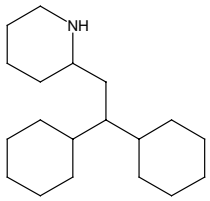
41



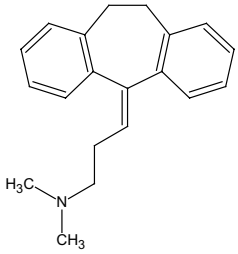
42



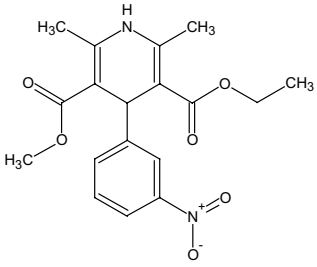
43



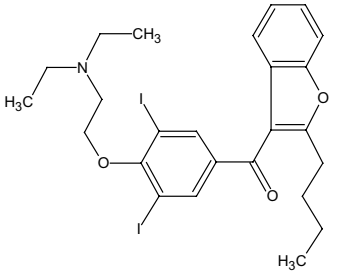
44



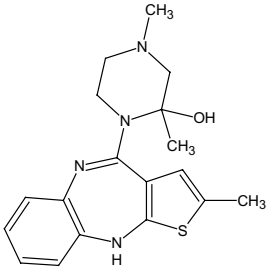
45



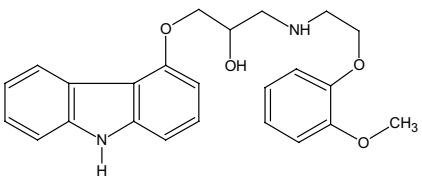
46



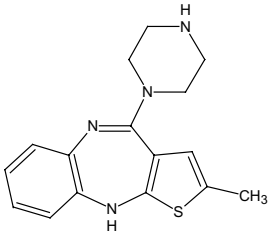
47



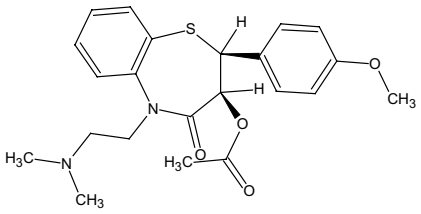
48



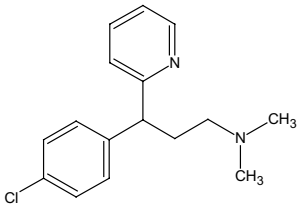
49



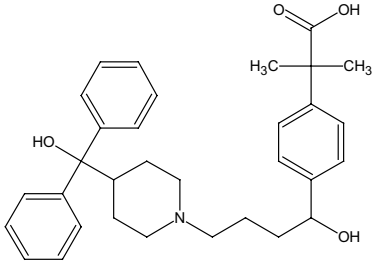
50



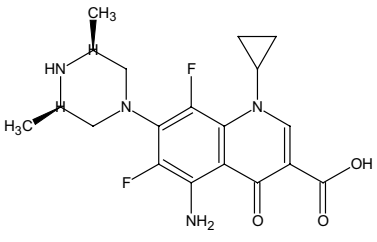
51



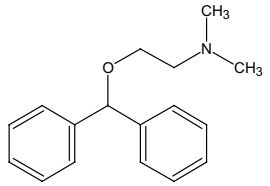
52



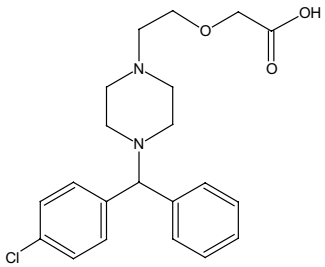
53



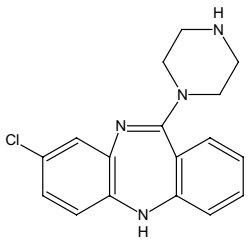
54



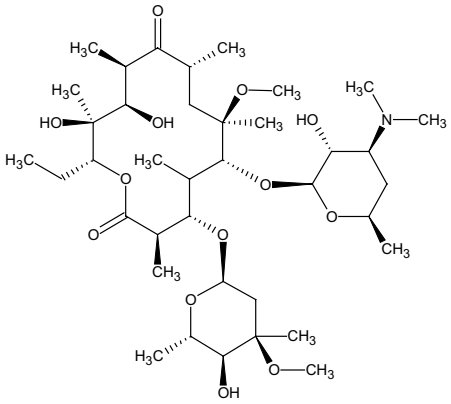
55



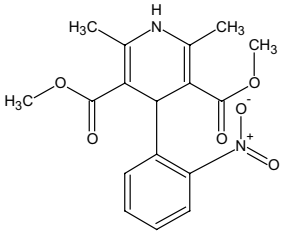
56



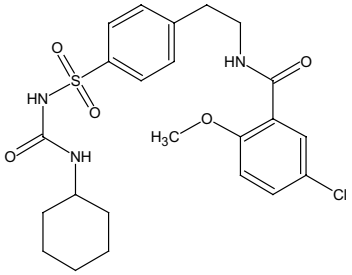
57



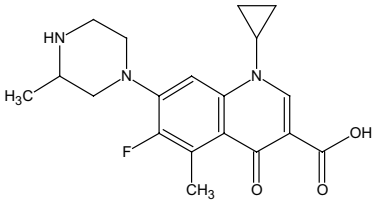
58



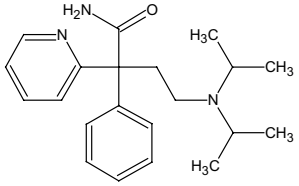
59



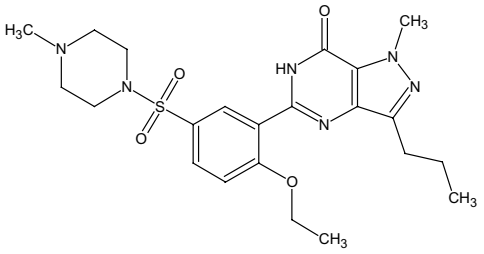
60



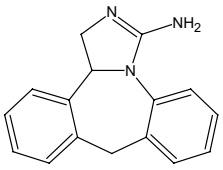
61



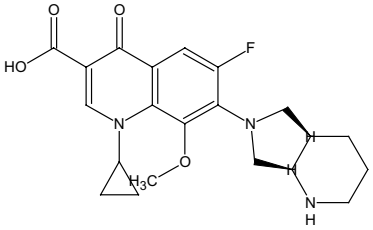
62



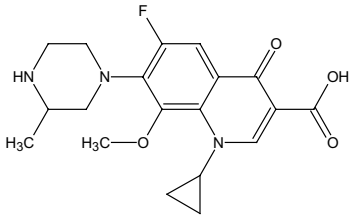
63



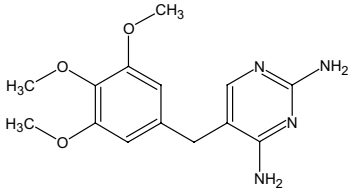
64



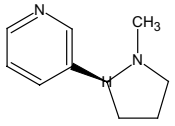
65



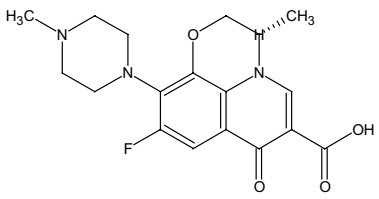
66



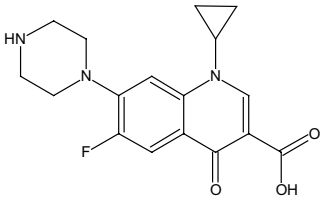
67



68



69



70

The data set is formed by 70 compounds with experimental hERG IC_{50} values retrieved from the literature [2] (Table 1, Figure 1).

2.2 Descriptors

2.2.1 EVA descriptor

The derivation of the EVA descriptor has previously been described elsewhere [3] and only a brief description of the technique will be given here. The descriptor is derived from IR- and Raman-range molecular vibrational frequencies usually calculated through the application of a normal coordinate analysis (NCA) to an energy minimized structure. For a compound with N atoms there are 3N – 6 (or 3N – 5 for a linear structure such as acetylene) normal modes of vibration. Thus, except in the special case where each structure has the same number of atoms, the number of frequencies will be different for each structure; that is, the property is in non-standard form. A technique has thus been developed in order to standardize the property such that each compound is characterized by an equivalent-length descriptor. The frequency set for a given structure is projected onto a linear bounded frequency scale (BFS) covering a range from 1 to 4000 cm^{-1}. A Gaussian kernel of fixed standard deviation (s) is then placed over each and every eigenvalue. The BFS is then sampled at fixed increments of L cm^{-1} and the value of the resulting EVA descriptor at each sample point is the sum of the amplitudes of the overlaid kernels at that point. This procedure is repeated for each dataset compound and then combined to provide a matrix with M rows (compounds) and 4,000/L columns (descriptor variables). Typically, a descriptor set has been derived using a s of 10 cm^{-1} and an L of 5 cm^{-1} giving 800 descriptor variables. For a standard QSAR dataset the number of variables is thus much larger than M and Partial least square to Latent Structure (PLS) is hence used to provide a robust regression analysis.

2.2.1 DRAGON descriptors

DRAGON descriptors are more than 1600 molecular descriptors listed in Table 2 divided into 20 logical blocks [4].

Table 2 – Molecular descriptors calculated by DRAGON

Molecular Descriptors	No.	Molecular Descriptors	No.
Constitutional descriptors	48	Randic molecular profiles	41
Topological descriptors	119	Geometrical descriptors	74
Walk and path counts	47	RDF descriptors	150
Connectivity indices	33	3D-MoRSE descriptors	160
Information indices	47	WHIM descriptors	99
2D autocorrelations	96	GETAWAY descriptors	197
Edge adjacency indices	107	Functional group counts	121
Burden eigenvalues	64	Atom-centred fragments	120
Topological charge indices	21	Charge descriptors	14
Eigenvalue-based indices	44	Molecular properties	28

2.3 Statistical analysis

PLS modeling has been used to investigate likely correlations between EVA and experimental pIC_{50} values and, respectively, descriptors generated by DRAGON and experimental pIC_{50} values. The optimal number of components in each PLS model was determined by SIMCA-P+ default cross-validation procedure. Different approaches were explored in order to obtain the best models in terms of stability and predictivity: (a) variables selection carried out with 2 different protocol: (a1) on the basis of VIP parameter and coefficient values and (a2) employing a genetic algorithm implemented in GAVS (Computer Chemistry Lab., Bracco Imaging SpA); (b) SIMCA-P+ Orthogonal Signal Correction (OSC) algorithm, used to remove from X data matrices information that is orthogonal to Y.

All PLS models here reported were generated considering just the experimental values found in ref. 2(b). Initial models were generated using all 62 compounds - strong outliers were detected and then excluded employing PCA on each X data matrix. The best models were further validated considering half of the compounds as training set and the rest as external test set. Training and test sets were generated by means of Onion/D-Optimal Design.

2.2.1 Software

EVA. Energy minimization and normal coordinate analysis needed to derive EVA descriptor were carried out by means of Spartan'02 (Wavefunction, Inc.) employing Merck Force Field. Calculation of EVA descriptor from vibrational frequencies was carried out using the proprietary program EVA-02 (S-IN).

DRAGON is a software package for the calculation of molecular descriptors developed by Milano Chemometrics and QSAR Research Group. It allows calculation of more than 1600

molecular descriptors for thousands of molecules (Talete, srl).

PASS (Prediction of Activity Spectra for Substances) (V. Poroikov, D. Filimonov & Associates) [5] predicts the probability for any given compound to be active (P_a) or inactive (P_i) for each one of over 1000 biological activities, including pharmacological effects, mechanisms of action, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity. P_a and P_i values vary from 0 to 1, and their sum may be different than 1. PASS predictions are based on the analysis of structure-activity relationships for a training set including a great number of non-congeneric compounds with different biological activities, using the descriptor Multilevel Neighborhoods of Atoms (MNA). PASS training set consists of over 46,000 biologically active compounds: 16,000 are already launched drugs and 30,000 drug-candidates under clinical or advanced preclinical testing.

QikProp (Schrödinger, L.L.C.) [6] has been developed by Prof. Bill Jorgensen at Yale University to rapidly predict ADMET properties of drug candidates. QikProp results have been fitted to datasets of drug-like molecules, based on 2-D and 3-D descriptors reflecting Monte Carlo simulation studies as well as experiment. QikProp predictions are calculation-based, as opposed to fragment based. Fragment-based methods can be problematic when they do not recognize parts of a structure or encounter unfamiliar fragment interactions, whereas QikProp will calculate properties based on the whole molecule. The advantage of this approach is that QikProp can be applied to new and unknown scaffolds.

Statistical analysis. PLS modeling and PCA were carried out with the software SIMCA-P+. MODDE was employed for the Onion/D-Optimal Design.

3 RESULTS AND DISCUSSION

3.1 QSAR models

In order to have homogeneous biological data, QSAR analysis was conducted just on the 62 compounds for which is available the experimental data found in ref. 2(b) (pIC_{50}*** values in Table 1). An initial PCA carried out on either EVA and DRAGON X matrix detected 3 outliers: compounds 44, 58 and 68 whose structure is quite different from all other compounds. These 3 structures were not considered in further analysis. The remaining 59 compounds were divided into a training set and a test set 1 respectively formed by 29 and 30 compounds selected by means of a Onion/D-Optimal Design. Models were generated considering just the training set and their real predictive power was tested with the external test set. Results are reported in Table 3. SDEP values are calculated on the 30 compounds of the test set 1.

These models were also employed to classify molecules as active or inactive, considering 5.0 as threshold value of pIC_{50}, predicted or experimental. A new test set was considered, test set 2 that

contains 38 compounds, formed by 30 structures of test set 1 and 8 structures whose biological data are from ref. 2a or 2c. According to experimental data, training set is formed by 21 molecules classified as active, and 8 molecules classified as inactive; test set is formed by 26 molecules classified as active, and 12 molecules classified as inactive. Fraction of compounds well classified according to predicted value of pIC_{50} are reported in Table 4. Consensus criterion assigns the activity class according to at least 3 of the 5 prediction available.

Table 3 – QSAR models results

description	X selection	X	PCs	Obj.	R^2	Q^2	SDEC	SDEP
DRAGON	VIP, coef	320	2	29	0.843	0.736	0.591	**0.932**
	OSC	1421	1	29	0.995	0.979	0.092	**0.980**
	GAVS	82	3	29	0.939	0.854	0.318	**1.206**
EVA	VIP, coef	112	3	29	0.915	0.649	0.374	**1.023**
	OSC	615	2	29	0.998	0.718	0.060	**0.991**

Table 4 – Fraction of compounds well classified

Description	X selection	Training set		Test set 2	
		% active well predicted	% inactive well predicted	% active well predicted	% inactive well predicted
DRAGON	VIP, coef	86%	75%	88%	50%
	OSC	90%	88%	77%	58%
	GAVS	95%	100%	73%	58%
EVA	VIP, coef	86%	88%	73%	67%
	OSC	90%	100%	77%	67%
Consensus 1*		**86%**	**100%**	**88%**	**67%**
Consensus 2**				**96%**	**67%**

* consensus according to QSAR models (5 predicted values)

** consensus according to QSAR models, PASS and QikProp predictions (7 predicted values)

3.2 PASS and QikProp predictions

The program PASS was trained to predict the probability of hERG activity, using a set of molecules with pIC_{50} values grater than or equal to 5.0. Preliminary studies choosing a PASS probability value (p_a) of 0.3 show that 75% of active molecules and, respectively, 73% of inactive molecules are predicted correctly, leading to 12 false negative and 6 false positive.

Results of QikProp predictions: considering 5 as the pIC_{50} threshold value between active and inactive compounds, 2 molecules are predicted as false negatives and 12 as false positives, corresponding to 96% and 45% of active and inactive molecules well predicted.

Employing test set 2 it is possible to calculate the fraction of molecules well predicted as active or inactive according to a new consensus criterion (Table 4): assign the activity class according to at least 4 of the 7 prediction available.

4 CONCLUSIONS

Employing different and independent approaches it is possible to obtain a consensus activity prediction able to well classify 96% of active molecules (1 false negative among 26 molecules) and 67% of inactive molecules (4 false positives among 12 molecules) of a test set formed by 38 compounds extracted from the literature.

Acknowledgment

The authors would like to thank Dr. Alessandro Maiocchi for providing the software GAVS.

Supplementary Material

An SDF file containing the structures of the 70 compounds considered was deposited.

5 REFERENCES

[1] (a) J.S. Mitcheson, J. Chen, M. Lin, C. Culberson, M.C. Sanguinetti, A Structural Basis for Drug-Induced Long QT Syndrome, *PNAS*, **2000**, 97, 12329. (b) D.A. Doyle, J. Morais Cabral, R.A. Pfuetzner, A. Kuo, J.M. Gulbis, S.L. Cohen, B.T. Chait, R. MacKinnon, The Structure of the Potassium Channel: Molecular Basis of K^+ Conduction and Selectivity, *Science*, **1998**, 280, 69.

[2] (a) A. Cavalli, E. Poluzzi, F. De Ponti, M. Recanatini, Towards a Pharmacophore for Drugs Inducing the Long QT Syntome: Insights from a CoMFA Study of HERG K^+ Channel Blockers, *J. Med. Chem.* **2002**, 45, 3844. (b) G.M. Keserü, Prediction of hERG Potassium Channel Affinity by Traditional and Hologram QSAR Methods, *Bioorg. Med. Chem. Lett*. **2003**, 13, 2773. (c) S. Ekins, W.J. Crumb, R. Dustan Sarazan, J.H. Winkel, S.A. Wrighton, Three-Dimensional Quantitative Structure-Activity Relationship fot Inhibition of Human Ether-a-Go-Go-Related Gene Potassium Channel, J. *Pharma. Exp. Therapeutics (JPET)* **2002**, 301, 427.

[3] A.M. Ferguson, T. Heritage, P. Jonathan, S.E. Pack, L. Phillips, J. Rogan, P.J. Snaith, Evaluation Of A Novel Infrared Range Vibration-Based Descriptor (EVA) For QSAR Studies. 1.General Application, *J. Comput. –Aided Mol. Design*, **1997,** 11, 143.

[4] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry*, WILEY-VCH, Weinheim, Germany, 2000.

[5] (a) V.V. Poroikov, D. Filimonov (2001) *Computer-aided prediction of biological activity spectra. Application for finding and optimization of new leads. Rational Approaches to Drug Design*, Eds. H.-D. Holtje, W.Sippl, Prous Science, Barcelona, p.403-407 (b) V.V. Poroikov, D.A. Filimonov, Yu.V. Borodina, A.A. Lagunin, A. Kos, *Robustness of Biological Activity Spectra Predictnig by Computer Program PASS for Noncongeneric Sets of Chemical Compounds*, *J. Med. Chem*., **2001**, 44, 2432-2437.

[6] E.M. Duffy, W.L. Jorgensen, Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water, *J. Am. Chem. Soc*. **2000**, 122, 2878-2888.