Journal of Molecular Structure 1130 (2017) 711–718

Contents lists available at ScienceDirect

Journal of Molecular Structure

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

QSAR modeling for anti-human African trypanosomiasis activity of substituted 2-Phenylimidazopyridines

Vijay H. Masand [a, *], Nahed N.E. El-Sayed [b, f], Devidas T. Mahajan [a], Andrew G. Mercader [c], Ahmed M. Alafeefy [d], I.G. Shibi [e]

[a] Department of Chemistry, Vidya Bharati College, Camp, Amravati, Maharashtra, 444 602, India
[b] Department of Chemistry, College of Science, "Girls Section", King Saud University, P.O. Box 2457, Riyadh 11451, Saudi Arabia
[c] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), UNLP, CCT La Plata-CONICET, Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina
[d] Department of Chemistry, Kulliyyah of Science, International Islamic University, P.O. Box 141, 25710 Kuantan, Malaysia
[e] Department of Chemistry, Sree Narayana College, Chempazanthy, Kerala, India
[f] National Organization for Drug Control and Research, Giza 35521, Egypt

ARTICLE INFO

Article history:
Received 20 August 2016
Received in revised form
2 November 2016
Accepted 3 November 2016
Available online 5 November 2016

Keywords:
Anti-HAT activity
QSAR
Substituted 2-Phenylimidazopyridines

ABSTRACT

In the present work, sixty substituted 2-Phenylimidazopyridines previously reported with potent anti-human African trypanosomiasis (HAT) activity were selected to build genetic algorithm (GA) based QSAR models to determine the structural features that have significant correlation with the activity. Multiple QSAR models were built using easily interpretable descriptors that are directly associated with the presence or the absence of a structural scaffold, or a specific atom. All the QSAR models have been thoroughly validated according to the OECD principles. All the QSAR models are statistically very robust (R^2 = 0.80–0.87) with high external predictive ability (CCC_{ex} = 0.81–0.92). The QSAR analysis reveals that the HAT activity has good correlation with the presence of five membered rings in the molecule.

1. Introduction

Sleeping sickness or human African trypanosomiasis (HAT), transmitted by tsetse flies (genus *Glossina*), has a major occurrence in rural populations in sub-Saharan Africa. HAT, considered as a neglected tropical disease, was nearly eradicated in the mid-1960s. The resurgence in the late 1990s, due to poor sanitation and suitable habitats for its vector in the Democratic Republic of the Congo (DRC), Angola, Central African Republic, southern Sudan, and Uganda, received considerable attention of the researchers to develop better diagnosis and treatment for the disease [1–3]. Recent reports indicate that in humans the disease is thought to be mainly caused by *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense*, however the non-human-pathogenic trypanosome species *Trypanosoma brucei brucei*, *Trypanosoma congolense*, and *Trypanosoma evansi* are also responsible in some instances [1]. The disease has two stages; in stage 1 (hemolymphatic) the peripheral infection with non-specific clinical symptoms occur; and in stage 2 the parasite crosses the blood brain barrier (BBB) and intrudes the central nervous system (CNS) [1–3].

Suramin and pentamidine are the recommended drugs for stage 1 infection, whereas for stage 2, the therapeutic options are melarsoprol, eflornithine and the currently used combination therapy NECT (nifurtimox and eflornithine combination therapy). Unfortunately, vaccine cannot be developed due to a high degree of antigenic variation. In addition, the treatment is parasite- and stage-specific, depending on the ability of the compound to cross the BBB. For BBB clearance the drug must be sufficiently lipophilic, which results in poor water solubility, hence, such drugs are mostly toxic and problematical to administer. Consequently, the available drugs for stage 2 of the disease exhibit high toxicity, involve the complexity of administration procedures and progressive loss of efficacy in some geographical regions. Recent efforts identified Fexinidazole, furamidine, DB289 (parafuramidine), CPD-0802 (an aza analogue of parafuramidine) and SCYX-7158 (a boron based compound) as attractive lead/targets in the drug pipeline for

Abbreviations: HAT, Human African Trypanosomiasis; GA, Genetic algorithm; MLR, Multiple linear Regression; QSAR, Quantitative structure-activity analysis; WHO, World health organization; ADMET, Absorption, Distribution, Metabolism, Excretion and Toxicity; OLS, Ordinary Least Square; QSARINS-Chem, QSAR Insubria-Chemistry; OECD, Organization for Economic Co-operation and Development.

* Corresponding author.
E-mail address: vijaymasand@gmail.com (V.H. Masand).

http://dx.doi.org/10.1016/j.molstruc.2016.11.012
0022-2860/

developing better therapeutics for HAT (see Fig. 1). Despite the previous efforts executed high toxicity, poor oral bioavailability and blood–brain barrier penetration are the major obstacles ahead for these clinical candidates. Thence, the search for a drug candidate with adequate activity, ADME and toxicity profile still persists [1–4].

Recently, Tatipaka et al. [5] identified substituted oxazolopyridine **1** (see Fig. 2) as an attractive lead due to good whole-cell activity on *T. brucei*, no cytotoxicity on mammalian cell lines, acceptable exposure in the central nervous system, and satisfactory aqueous solubility. But, its poor metabolic stability in liver microsomes appeared as a severe liability. Later, to design a better analogue of **1** with the desired profile, they synthesized and screened a series of substituted 2-Phenylimidazopyridines. Since, the mechanism of action and the specific target with which these analogues interact is unknown [5]; in such a situation, a good strategy for lead optimization is to employ computer aided drug design (CADD) using the available information. Hence, in the present work, ligand based drug design technique, viz. QSAR (2D- and 3D-) has been executed to determine the structural features having a significant correlation with the HAT activity profile of substituted 2-Phenylimidazopyridines.

In the past decades, CADD has appeared as a thriving option to conventional 'trial and error' methodology of drug design/discovery to unknot the mysteries of structural patterns that govern the activity, pharmacokinetics, pharmacodynamics and toxicity profiles of a drug candidate. CADD is relatively fast, economical and significantly result oriented successful *in-silico* technique [6–10]. It encompasses a combination of different ideas, algorithms, tools and techniques of various scientific fields like computer, mathematics, statistics, etc. Its major emphasis is on simulation of interactions of different molecules, to determine the reasons behind the specific interactions of different molecules and identification of effective structural features associated with activity/toxicity. QSAR, molecular docking, pharmacophore modeling, etc. are established CADD methods, which when used in harmony provide significant and unrivalled information essential for lead/drug optimization [4,11–15]. These methods have been widely used for identification of the structural patterns that govern the specific activity/toxicity of drug candidates and provide better insight into the mechanism of drug action.

The main objective of the present work is to develop statistically robust and easily interpretable, in terms of structural fragments or specific atom, QSAR models with high external predictive ability.

2. Experimental methodology

2.1. Experimental datasets

In the present work, HAT inhibition activities of sixty substituted 2-Phenylimidazopyridines comprising different heterocyclic scaffolds and diverse substituents at various positions covering a meaningful portion of the chemical space were subjected to QSAR modeling [5]. The reported EC_{50} (μM) values for HAT activity were converted to pEC_{50} ($-\log_{10}EC_{50}$) before QSAR analysis. The EC_{50}, pEC_{50} and the substituents on 2-Phenylimidazopyridine moiety have been listed in Table 1.

2.2. Modeling and molecular descriptors calculation

In the present work, a QSAR analysis following the standard procedure recommended by OECD and different researchers was exercised [16–26]. The chemical structures were drawn using ChemSketch 12 freeware followed by energy minimization using MMFF94 force field in TINKER [4,12,15]. The optimized structures were used as input for the calculation of a good number of 1-3D, electro-topological, fingerprints and other descriptors. Two descriptor calculating softwares were used: PaDEL 2.21 and e-Dragon. Since, all the calculated descriptors (>18,000) do not contain significant information; objective feature selection was employed to reduce the descriptor pool. Nearly constant (>95%), constant, and highly correlated (|R| > 95%) descriptors were eliminated before subjective feature selection (SFS) using QSARINS-Chem 2.2.1 [16,17,20]. This resulted in a reduced cluster of 345 descriptors only. The very next step involved the elimination of highly esoteric descriptors, the descriptors for which an exact explanation is not available or it is difficult to interpret it in terms of structural features [26]. This led to a set of only 253 easily interpretable descriptors. The reduced set still consists a wide range of theoretical molecular descriptors that takes into account different structural features, viz. constitutional (0D-), mono-dimensional (1D-), bi-dimensional (2D-) and three-dimensional (3D-), capturing and magnifying the diverse aspects of the chemical structures.

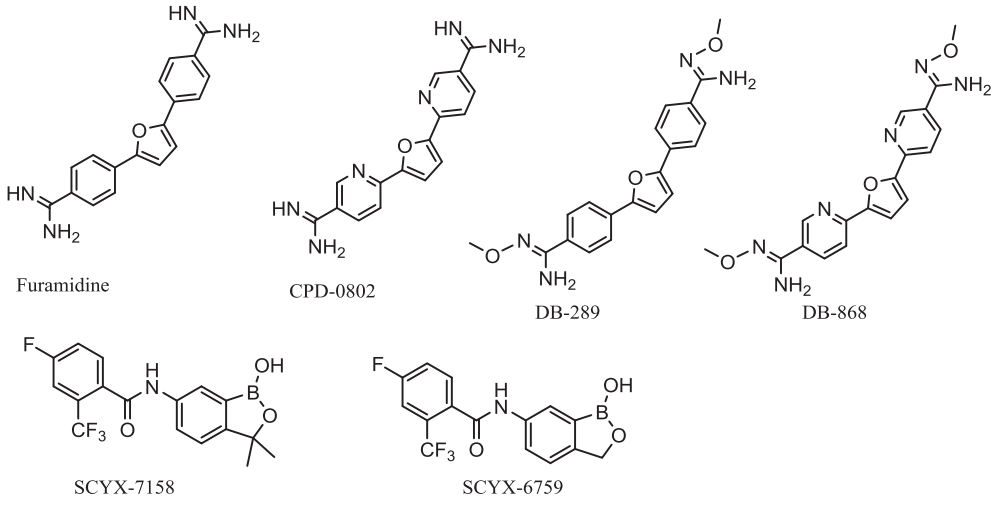
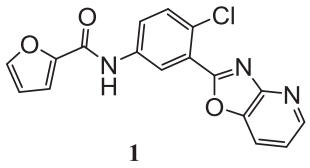


Fig. 1. Chemical structures of clinical drug candidates against HAT.



1

Ease of synthesis,
No stereogenic carbons,
Low MW and Lipinski rule-of-five compliance,
T. brucei in vitro EC_{50} = 220 nM,
Acceptable aqueous solubility,
Not cytotoxic to mammalian cells (L6, CRL-8155 and HepG2 cells),
Poor metabolic stability in liver microsomes

Fig. 2. Chemical structure and profile of substituted oxazolopyridine **1**.

2.3. QSAR model development

2.3.1. QSAR model

The very first principles and applications of QSAR analysis are to gain maximal information of activity related structural features and to predict the desired activity of a molecule before its actual synthesis and bio-screening. Hence, in order to achieve these goals easily interpretable descriptors were considered during model generation and multiple QSAR models were developed using divided and undivided datasets [19,27,28]. The dataset was divided into a training (80%) and a prediction (or test) (20%) sets in random fashion before descriptor selection. Multiple splitting were employed to develop multiple QSAR models [12,15], in this way a molecule in the training set of a splitting may or may not be in the training set of another splitting. Therefore, the multiple QSAR modeling approach ensured that maximum number and information is gained for molecular descriptors that govern the biological profile of the molecules. GA (Genetic Algorithm) module of QSARINS-Chem 2.2.1 was utilized for the selection of optimum number and set of descriptors. For the sake of simplicity and to avoid the problem of over-fitting, the heuristic search of descriptors was limited to four descriptors using the default settings in QSARINS-Chem 2.2.1. Q^2_{loo} was used as a fitness function to avoid the problem of naïve Q^2. The strategy used in QSAR model development has been summarized in Fig. 3 [12,15,24].

For development of QSAR models 6–10, HeuristicLab 3 was employed using different operand using the default settings.

2.4. Model validation

All QSAR models need to be appropriately validated to ascertain its predictive ability and utility. The statistical qualities and validity of the QSAR models were established by means of: (a) internal validation or cross-validation (CV) by leave-one-out (LOO) and leave-many-out (LMO) procedure; (b) using the prediction set; (c) data randomization i.e. Y-scrambling and (d) examining if the following conditions are satisfied [12,15]: $R^2_{tr} \geq 0.6$, $Q^2_{loo} \geq 0.5$, $Q^2_{LMO} \geq 0.6$, $R^2 > Q^2$, $R^2_{ex} \geq 0.6$, $RMSE_{tr} < RMSE_{cv}$, $\Delta K \geq 0.05$, $CCC \geq 0.80$, $Q^2 - F^n \geq 0.60$, $r^2_m \geq 0.6$, $(1 - r^2/r^2_0) < 0.1$, $0.9 \leq k \leq 1.1$ or $(1 - r^2/r'^2_0) < 0.1$, $0.9 \leq k' \leq 1.1$, $|r^2_0 - r'^2_0| < 0.3$ with *RMSE* and *MAE* close to zero. The threshold values of these parameters confirm the robustness and good external predictive ability of a GA-MLR model. Thus, all the models having low internal and external predictive ability were subsequently rejected.

3. Results and discussion

3.1. QSAR models

The GA analysis resulted in the generation of a good number of MLR models with nearly similar statistical performance but encompassing different descriptors. In such a situation, the usual practice followed by a QSAR modeller is to select only one MLR model on the basis of its statistical performance. However, this 'first among equals' approach is with following drawbacks [12,15] (1) a QSAR model consisting of only esoteric descriptors, suitable and realistic description in terms of structural features is highly problematic and challenging, (2) The single QSAR model may not be based on (i) appropriate composition of training and test sets, (ii) sufficient chemical and biologic space i.e. appropriate applicability domain, (3) the single QSAR model might have high predictive on a particular prediction set, but poor predictivity on another prediction set. To overcome these drawbacks of 'first among equals' approach, building and reporting multiple models or consensus modeling are two easy, practicable and efficient solutions. Recently, it has been established that developing multiple QSAR models based on divided and undivided dataset enhance the efficacy of QSAR in determining the dominant and concealed structural features that have significant correlation with the activity [12,15]. Therefore, in the present study, multiple QSAR models have been built following the OECD principles for acceptable QSAR models. This approach led to generation of ten QSAR models possessing excellent statistical performance. The GA-MLR QSAR models along with different statistical parameters, tabulated in Table 2, are as following:

3.2. Model-1 (Undivided dataset)

$$pEC_{50} = 6.486\ (\pm 0.378) - 0.214\ (\pm 0.067)\ *\ \text{C-024} + 0.756\ (\pm 0.273)\ *\ \text{nR05} - 1.087\ (\pm 0.318)\ *\text{F09[N–F]} + 0.897\ ((\pm 0.169)\ *\text{F10[C–F]}.$$

3.3. Model-2 (Undivided dataset)

$$pEC_{50} = 4.218\ (\pm 0.514) + 0.449\ (\pm 0.141)\ *\text{F02[N–N]} + 0.842\ (\pm 0.265)\ *\ \text{nR05} - 0.570\ (\pm 0.305)\ *\text{F09[N–F]} + 0.453\ (\pm 0.147)\ *\text{F10[C–F]}.$$

3.4. Model-3 (Divided dataset)

$$pEC_{50} = 5.148\ (\pm 0.537) + 0.518\ (\pm 0.399)\ *\text{B10[C–F]} + 0.624\ (\pm 0.292)\ *\ \text{nR05} - 0.558\ (\pm 0.476)\ *\text{B09[N–F]} + 1.178\ (\pm 0.404)\ *\text{B07[N–F]}.$$

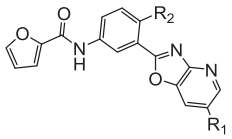
3.5. Model-4 (Divided dataset)

$$pEC_{50} = 5.729\ (\pm 0.258) - 1.328\ (\pm 0.724)\ *\ \text{nArNR2} + 1.450\ (\pm 0.337)\ *\ \text{nPyrrolidines} - 0.507\ (\pm 0.254)\ *\text{F10[N–F]} + 0.308\ (\pm 0.100)\ *\text{F09[C–F]}.$$

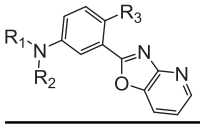
3.6. Model-5 (Divided dataset)

$$pEC_{50} = 5.789\ (\pm 0.258) - 1.375\ (\pm 0.743)\ *\ \text{nArNR2} + 1.322\ (\pm 0.331)\ *\ \text{nPyrrolidines} - 0.015\ (\pm 0.008)\ *\text{G(N..F)} + 0.381\ (\pm 0.121)\ *\text{F09[C–F]}.$$

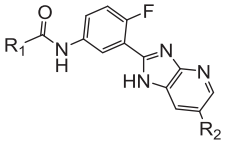
Table 1
Different substituted 2-Phenylimidazopyridines along with reported IC_{50} and pIC_{50}.



S.No.	R_1	R_2	EC_{50} (μm)	pEC_{50}(M)
1.	H	Cl	0.22	6.658
2.	H	F	0.12	6.921
3.	H	Me	0.43	6.367
4.	H	CN	0.4	6.398
5.	Br	F	0.23	6.638
6.	CN	F	0.38	6.42
7.	phenyl	F	0.04	7.398
8.	4-fluorophenyl	F	0.05	7.301
9.	3-chlorophenyl	F	0.05	7.301
10.	4-MeO-phenyl	F	0.17	6.77
11.	4-phenylphenyl	F	0.42	6.377

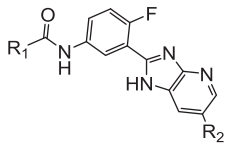


	R_1	R_2	R_3	EC_{50} (μM)	pEC_{50}(M)
12.	5-methylfuran-2-carbonyl	H	F	0.2	6.699
13.	3-methylfuran-2-carbonyl	H	F	0.1	7
14.	3-furanoyl	H	F	0.15	6.824
15.	benzoyl	H	Cl	7.1	5.149
16.	oxazole-5-carbonyl-	H	F	1.9	5.721
17.	2-thiophenoyl	H	Cl	1.5	5.824
18.	3-pyridinecarbonyl-	H	F	7	5.155
19.	pyrazine-2-carbonyl-	H	F	0.9	6.046
20.	N-methylpyrrole-2-carbonyl-	H	F	1.1	5.959
21.	methylsulfonyl	H	Cl	6.1	5.215
22.	2-furancarbothioyl-	H	F	0.41	6.387
23.	2-furanoyl	2-acetyl	F	0.5	6.301
24.	2-furanoyl	2-furanoyl	F	0.12	6.921
25.	benzyl	benzyl	F	1.1	5.959
26.	methylcarbamoyl	H	F	3.8	5.42
27.	isopropylcarbamoyl-	H	Cl	1	6
28.	phenylcarbamoyl-	H	Cl	12	4.921
29.	dimethylcarbamoyl-	H	Cl	0.4	6.398
30.	1-pyrrolidinoyl-	H	Cl	0.09	7.046
31.	1-piperidinoyl-	H	Cl	1.9	5.721



	R_1	R_2	EC_{50} (μM)	pEC_{50}(M)
32.	2-furanyl	H	0.2	6.699
33.	2-furanyl	Cl	0.07	7.155
34.	2-furanyl	F	0.2	6.699
35.	2-furanyl	5-Cl	10	5
36.	2-furanyl	7-Cl	0.12	6.921
37.	N-pyrrolidinyl	Cl	0.05	7.301
38.	N-pyrrolidinyl	phenyl	0.002	8.699
39.	N-pyrrolidinyl	3-methoxyphenyl	0.01	8
40.	N-pyrrolidinyl	2-methoxyphenyl	0.005	8.301
41.	N-pyrrolidinyl	3-Cl-phenyl	0.002	8.699
42.	N-pyrrolidinyl	2-chlorophenyl	0.005	8.301
43.	N-pyrrolidinyl	3-acetylphenyl	0.003	8.523
44.	N-pyrrolidinyl	3-Me-phenyl	0.002	8.699
45.	N-pyrrolidinyl	3-trifluoromethoxyphenyl	0.03	7.523
46.	N-pyrrolidinyl	3-methyl-4-fluorophenyl	0.02	7.699
47.	N-pyrrolidinyl	3-NH_2-phenyl	0.01	8
48.	N-pyrrolidinyl	3-furanyl	0.004	8.398
49.	N-pyrrolidinyl	3-thiophenyl	0.002	8.699
50.	N-pyrrolidinyl	2-thiophenyl	0.004	8.398
51.	N-pyrrolidinyl	3-pyridyl	0.005	8.301

Table 1 (*continued*)



	R_1	R_2	EC_{50} (μM)	pEC_{50}(M)
52.	N-pyrrolidinyl	5-(2-chloropyridyl)	0.01	8
53.	N-pyrrolidinyl	4-(2-chloropyridyl)	0.004	8.398
54.	N-pyrrolidinyl	5-(3-methylpyridyl)	0.01	8
55.	N-pyrrolidinyl	5-(2-methoxypyridyl)	0.005	8.301
56.	N-pyrrolidinyl	5-(3-pyrrolidino)	0.02	7.699
57.	N-pyrrolidinyl	5-(3-chloropyrimidinyl)	0.005	8.301
58.	N-pyrrolidinyl	5-pyrimidinyl	0.03	7.523
59.	N-pyrrolidinyl	5-(2-methoxypyrimidinyl)	0.06	7.222
60.	N-pyrrolidinyl	5-(2-chloropyrimidinyl)	0.04	7.398

3.7. Model-6 (Undivided dataset)

pEC_{50} = 8.455 (±0.292)−0.013 (±0.006) *G(N..F)−1.410 (±0.392) * invsqr-nR05−1.440 (±0.273) *B03[N−O] + 0.318 (±0.134) *F10 [C−F].

3.8. Model-7 (Undivided dataset)

pEC_{50} = 9.122 (±0.407)−0.013 (±0.006) *G(N..F)−1.217 (±0.341) * invsqr-nR05−0.846 (±0.159) * expB03[N−O] + 0.317 (±0.135) *F10 [C−F].

3.9. Model-8 (Divided dataset)

pEC_{50} = 8.162 (±1.455)−0.022 (±0.021) * Ss−2.053 (±0.442) * invcube-nR05 + 0.948 (±0.263) * C-041 + 0.901 (±0.325) *B10 [C−F].

3.10. Model-9 (Divided dataset)

pEC_{50} = 8.540 (±0.458)−0.015 (±0.011) *G(N..F)−1.603 (±0.450) * invsqr-nR05−1.416 (±0.355) *B03[N−O] + 0.340 (±0.160) *F10 [C−F].

3.11. Model-10 (Divided dataset)

pEC_{50} = 7.675 (±0.331) + 0.388 (±0.308) * Cl-089−1.485 (±0.412) * invcube-nR05−1.118 (±0.316) *B03[N−O] + 0.672 (±0.308) *B10[C−F].

The statistical symbols have their usual meaning [12,16,17,20,23] and are available in the supplementary material, also. From statistical analysis (Table 2), it is clear that all the developed QSAR models are statistically robust and possess good external predictive ability, especially, the models 6−10. The models 6−10 outperform their counterparts apropos of fitting, internal validation and external predictivity criteria. As expected, establishing the models 6−10 helped in the identification of less dominant descriptors like Ss (sum of Kier-Hall electrotopological indices), G(N..F) (sum of geometrical distances between N and F), C-041 (an atom-centered fragment representing X−C(=X)−X), and Cl-089 (an atom-centered fragment representing Cl attached to sp^2 hybridized C) with correlation with the activity, which were not discovered while building the models 1−5. For all the developed models, the value of $R^2_{adj.}$ is quite close to R^2_{tr} suggesting that the number of descriptors in the models are not too high, thereby, indicating that the models are free from over-fitting [26]. This is further supported by the low value of LOF (Lack of fit) for all the models. The low Kxx

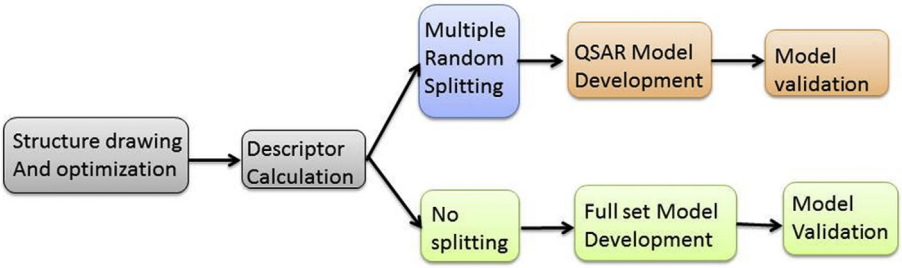


Fig. 3. Strategy used for developing the QSAR models.

Table 2
Statistical performance of different QSAR models.

S.No.	Statistical Parameter	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6	Model-7	Model-8	Model-9	Model-10
1.	N_{tr}	60	60	48	48	48	60	60	48	48	48
2.	N_{ex}	–	–	12	12	12	–	–	12	12	12
3.	Number of Descriptors	4	4	4	4	4	4	4	4	4	4
Fitting Criteria											
4.	R^2_{tr}	0.807	0.806	0.822	0.823	0.812	0.863	0.862	0.839	0.876	0.874
5.	$R^2_{adj.}$	0.793	0.792	0.806	0.806	0.794	0.853	0.852	0.824	0.865	0.863
6.	R^2_{tr} - $R^2_{adj.}$	0.014	0.014	0.016	0.017	0.018	0.010	0.010	0.015	0.012	0.012
7.	LOF	0.288	0.289	0.292	0.301	0.312	0.205	0.206	0.246	0.214	0.217
8.	Kxx	0.255	0.300	0.398	0.365	0.406	0.441	0.440	0.306	0.478	0.374
9.	ΔK	0.119	0.139	0.119	0.064	0.058	0.064	0.063	0.131	0.080	0.104
10.	$RMSE_{tr}$	0.465	0.466	0.450	0.457	0.466	0.392	0.394	0.413	0.385	0.389
11.	MAE_{tr}	0.354	0.358	0.349	0.362	0.383	0.312	0.313	0.320	0.307	0.314
12.	RSS_{tr}	12.981	13.028	9.733	10.034	10.414	9.232	9.302	8.190	7.119	7.248
13.	CCC_{tr}	0.893	0.893	0.903	0.903	0.896	0.926	0.926	0.913	0.934	0.933
14.	s	0.486	0.487	0.476	0.483	0.492	0.410	0.411	0.436	0.407	0.411
15.	F	57.446	57.191	49.749	49.841	46.345	86.363	85.610	56.177	76.220	74.679
Internal Validation Criteria											
16.	R^2_{cv} (Q^2loo)	0.776	0.770	0.784	0.777	0.753	0.840	0.839	0.800	0.847	0.844
17.	R^2-R^2_{cv}	0.031	0.036	0.038	0.046	0.059	0.022	0.023	0.039	0.029	0.030
18.	$RMSE_{cv}$	0.501	0.508	0.496	0.513	0.533	0.423	0.425	0.461	0.428	0.432
19.	MAE_{cv}	0.382	0.394	0.389	0.411	0.442	0.340	0.341	0.358	0.344	0.351
20.	$PRESS_{cv}$	15.041	15.472	11.824	12.629	13.650	10.740	10.837	10.190	8.809	8.974
21.	CCC_{cv}	0.877	0.874	0.882	0.881	0.867	0.914	0.914	0.892	0.919	0.917
22.	Q^2_{LMO}	–	–	0.774	0.792	0.771	0.830	0.827	0.787	0.842	0.841
23.	R^2_{Yscr}	–	–	0.081	0.086	0.087	0.069	0.069	0.086	0.087	0.082
24.	Q^2_{Yscr}	–	–	−0.151	−0.178	−0.231	−0.137	−0.137	−0.144	−0.144	−0.144
External Validation Criteria											
25.	θ*	–	–	−10.983	3.116	2.325	–	–	−2.183	−2.328	1.261
26.	$RMSE_{ex}$	–	–	0.570	0.479	0.417	–	–	0.512	0.441	0.483
27.	MAE_{ex}	–	–	0.421	0.351	0.310	–	–	0.442	0.350	0.402
28.	$PRESS_{ext}$	–	–	3.894	2.756	2.085	–	–	3.148	2.330	2.801
29.	R^2_{ex}	–	–	0.673	0.838	0.924	–	–	0.819	0.779	0.787
30.	Q^2-F^1	–	–	0.690	0.748	0.830	–	–	0.807	0.759	0.711
31.	Q^2-F^2	–	–	0.673	0.713	0.800	–	–	0.804	0.752	0.702
32.	Q^2-F^3	–	–	0.716	0.805	0.849	–	–	0.753	0.838	0.806
33.	CCC_{ex}	–	–	0.805	0.886	0.916	–	–	0.905	0.882	0.875
34.	r^2m aver.	–	–	0.541	0.654	0.776	–	–	0.741	0.687	0.657
35.	r^2m delta	–	–	0.257	0.180	0.094	–	–	0.040	0.069	0.197
36.	R^2-ExPy	–	–	0.785	0.780	0.756	–	–	0.801	0.848	0.845
37.	R'^2_o	–	–	0.740	0.747	0.712	–	–	0.765	0.829	0.823
38.	k'	–	–	0.996	0.997	0.993	–	–	0.996	0.998	0.997
39.	1-(R^2/R'^2_o)	–	–	0.057	0.042	0.059	–	–	0.044	0.022	0.025
40.	r'^2m	–	–	0.413	0.639	0.597	–	–	0.650	0.731	0.722
41.	R^2_o	–	–	0.784	0.777	0.753	–	–	0.800	0.847	0.844
42.	k	–	–	0.999	0.998	1.002	–	–	0.999	0.999	0.999
43.	1-(R^2-ExPy/R^2_o)	–	–	0.001	0.004	0.004	–	–	0.001	0.001	0.001
44.	r^2m	–	–	0.767	0.737	0.716	–	–	0.780	0.829	0.829

R^2 – correlation coefficient, Q^2 – leave-one-out 'cross validated R^2', R^2_{adj} - adjusted R^2, SEE – standard error of estimates, RMSE - root mean squared error, MAE - mean absolute error, CCC - concordance correlation coefficient, for the training (tr), and test (ex) sets; LOF - lack of fit, F – Fischer's value; R^2_{LMO} and Q^2_{LMO} – leave many-out correlation coefficient and cross-validation coefficients; R^2_{Ycr} and Q^2_{Ycr} – Y–- scramble correlation and cross-validation coefficients.

(representing inter-correlation among descriptors) value in all the models indicates that low correlation among the descriptors used in a model [16,17,20]. The condition $RMSE_{tr} < RMSE_{cv}$ is satisfied by all the developed models. The values of cross validation parameters Q^2, Q^2_{LMO} and CCC_{CV} for the developed models are high, thereby, indicating the statistical robustness of the models. The low value of R^2_{Yscr} and Q^2_{Yscr} for all the models indicates that the models have not been developed by chance. For the models 3–5 and 8–10, the

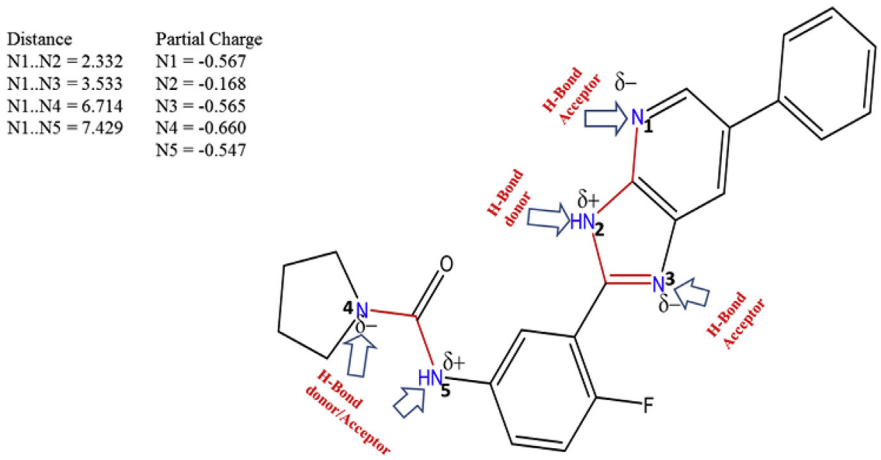


Fig. 4. Exemplification of F02[N–N] descriptor and H-bond donor/acceptor pattern associated with F02[N–N] using molecule **38** as a representative.

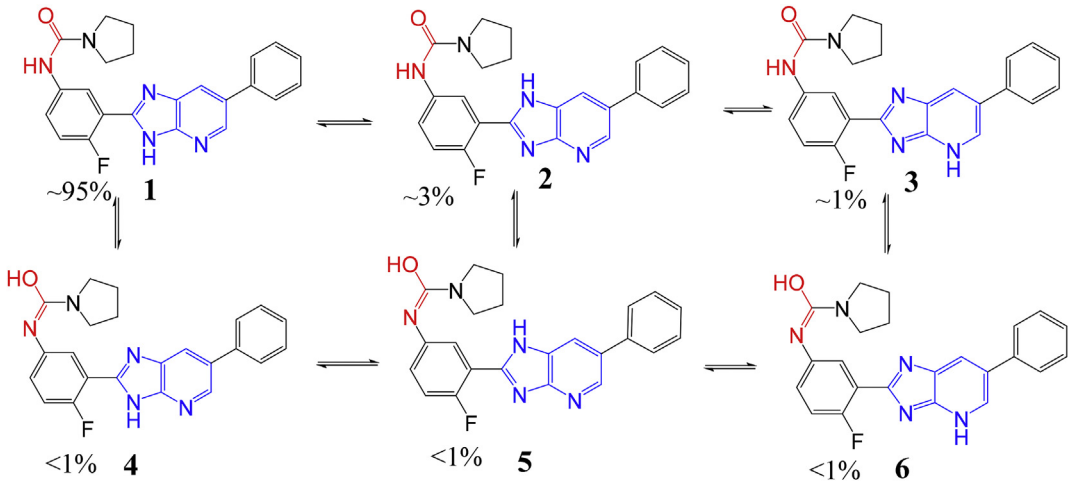


Fig. 5. Different possible tautomeric forms of **38**.

high value of various statistical parameters like R^2_{ex}, Q^2-F^1, Q^2-F^2, Q^2-F^3, CCC_{ex}, etc. indicate the models possess high external predictivity. This is further supported by the low value of $RMSE_{ex}$, MAE_{ex}, and $PRESS_{ext}$. In short, the developed models satisfy the recommended interrelations and threshold values for various statistical parameters suggested by different researchers. Interestingly, for model 5 a very rare observation is identified, the value of Q^2-F^1 (= 0.830) is greater than R^2 (= 0.812), thus, the model is able to predict new data better than fitting available ones, indicating that the molecules for which this model fits better are present in the prediction set [15,29,30]. This is again supported by the higher value of R^2_{ex} than R^2.

A commonly encountered problem associated with QSAR analysis is the interpretation or correlation of descriptors with specific structural fragment or atom. Therefore, in the present work, easily interpretable descriptors that are directly associated with the presence or the absence of a structural scaffold, or a specific atom were considered. This could be highly beneficial to synthetic chemists for future synthetic strategies.

In models 1–3, the 2D-constitutional type of descriptor 'nR05', representing the presence of a five membered ring, is present with a positive coefficient. The descriptors 'invsqr-nR05' and 'invcube-nR05' represent the inverse of square or cube of 'nR05', respectively, and have negative coefficients in models 6–10. Therefore, it appears that the presence of five membered rings in the molecules increases the activity. This observation is vindicated by comparing the activity of **1** with **15**, and **2** with **18**, as representatives. Therefore, in future modifications five membered ring must be retained to increase the activity. The 2D-descriptors, F09[N–F] and B09[N–F], which correspond to the frequency and presence/absence of N and F at a topological distance of 9, respectively, have a negative correlation with the activity. Interestingly, in models 1–3, the positive correlation of F10[C–F] and B10[C–F], which correspond to frequency and presence/absence of C and F at a topological distance of 10, respectively, with HAT activity indicates that the presence of F favors the activity. This is again supported by the positive correlation of B07[N–F], which represents the presence/absence of N and F at topological distance of 7, in model 3.

The descriptor C-024, an atom centered descriptor corresponds to R–CH–R fragment, has a negative coefficient in model-1, hence, its value must be kept as low as possible. The descriptor 'nPyrrolidines' indicates the presence of number of pyrrolidine rings in the molecule, its presence is beneficial for activity, as evident from its positive coefficient in model 4 and 5. The descriptor 'nArNR2' which stands for the presence of number of tertiary amines (aromatic) in the molecule has a negative association with the activity, hence such groups must be avoided in future modifications. F02[N–N] is a finger print descriptor that represents the frequency of

the presence of two nitrogen atoms at a topological distance of 2, its positive coefficient in the model 2 points out its positive influence on the activity. This descriptor has been depicted by red bonds in Fig. 4 using the molecule **38** as a representative. This descriptor is associated with an interesting pharmacophoric pattern of H-bond donor and acceptor nitrogen atoms present in a specific arrangement, in which a H-bond donor (like N2) is at a topological distance of 2 from at least one H-bond acceptor N (like N1 or N3).

The importance of this structural pattern is further supported by the fact that the presence of 5-chloropyridin-2-yl, as in *N*-(3-((5-chloropyridin-2-yl)carbamoyl)-4-fluorophenyl)furan-2-carboxamide, has negative influence on the activity. The compound *N*-(3-((5-chloropyridin-2-yl)carbamoyl)-4-fluorophenyl)furan-2-carboxamide ($EC_{50} = 7\ \mu M$) was not incorporated during the QSAR and pharmacophore model building, as the aim of the present work was to analyze the bicyclic derivatives only.

This interesting pattern of H-bond donor/acceptor nitrogen atoms also helps in attaining various tautomeric forms, thereby, providing additional flexibility to the molecules to acquire bioactive tautomeric form(s) while interacting with the target receptor. It has been established that a less stable tautomeric form of a molecule could be the true bioactive tautomeric form [12]. Thus, tautomeric transformations could be a possible reason behind higher activity of the molecules bearing imidazopyridine ring. Such a useful flexibility is diminished when an oxazolopyridine ring is present instead of an imidazopyridine ring. The various tautomeric forms along with their calculated distribution (using the software MarvinSketch 5.0) for the molecule **38**, as a representative, have been depicted in Fig. 5.

The descriptor B03[N–O], which represents the presence/absence of N and O at a topological distance of 3, or its exponential expB03[N–O] have negative coefficients in the model 6,7, 9, and 10. Therefore, the presence/absence of N and O at a topological distance of 3 has negative influence on the activity. Hence, such a combination of N and O must be avoided for better activity profile. This is supported by the observation that in the present series the molecules bearing oxazolopyridine ring, in general, are less active than those having an imidazopyridine ring.

4. Conclusions

The QSAR analysis revealed important information and observations about the structural features that steer the HAT activity of substituted 2-Phenylimidazopyridines. The developed models are statistically robust with high external predictive ability. The development of multiple QSAR models helped in identifying less dominant, but very useful descriptors, which have significant correlation with the activity. The QSAR models pointed out that the presence of five membered rings, especially the pyrrolidine ring, is beneficial for the HAT activity of the present series molecules.

Acknowledgements

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding this Research (Group No. RG-1435-083). One of the authors VHM is thankful to Dr. Paola Gramatica for providing these software QSARINS-Chem 2.2.1.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.molstruc.2016.11.012.

References

[1] R. Brun, J. Blum, F. Chappuis, C. Burri, Human African trypanosomiasis, Lancet 375 (2010) 148–159.
[2] R. Brun, R. Don, R.T. Jacobs, M.Z. Wang, M.P. Barrett, Development of novel drugs for human African trypanosomiasis, Future Microbiol. 6 (2011) 677–691.
[3] P.G. Kennedy, Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness), Lancet Neurol. 12 (2) (2013) 186–194.
[4] S.N. Bukhari, X. Zhang, I. Jantan, H.L. Zhu, M.W. Amjad, V.H. Masand, Synthesis, molecular modeling, and biological evaluation of novel 1, 3–Diphenyl-2-propen-1-one based pyrazolines as anti-inflammatory agents, Chem. Biol. Drug Des. 85 (2015) 729–742.
[5] H.B. Tatipaka, J.R. Gillespie, A.K. Chatterjee, N.R. Norcross, M.A. Hulverson, R.M. Ranade, P. Nagendar, S.A. Creason, J. McQueen, N.A. Duster, A. Nagle, F. Supek, V. Molteni, T. Wenzler, R. Brun, R. Glynne, F.S. Buckner, M.H. Gelb, Substituted 2-phenylimidazopyridines: a new class of drug leads for human African trypanosomiasis, J. Med. Chem. 57 (2014) 828–835.
[6] D.T. Mahajan, V.H. Masand, K.N. Patil, T. Ben Hadda, R.D. Jawarkar, S.D. Thakur, V. Rastija, CoMSIA and POM analyses of anti-malarial activity of synthetic prodiginines, Bioorg. Med. Chem. Lett. 22 (2012) 4827–4835.
[7] V.H. Masand, D.T. Mahajan, T. Ben Hadda, R.D. Jawarkar, A.M. Alafeefy, V. Rastija, M.A. Ali, Does tautomerism influence the outcome of QSAR modeling? Med. Chem. Res. 23 (2014) 1742–1757.
[8] V.H. Masand, D.T. Mahajan, K.N. Patil, T.B. Hadda, M.H. Youssoufi, R.D. Jawarkar, I.G. Shibi, Optimization of antimalarial activity of synthetic prodiginines: QSAR, GUSAR, and CoMFA analyses, Chem. Biol. Drug Des. 81 (2013) 527–536.
[9] V. Rastija, S. Nikolic, V.H. Masand, Quantitative relationships between structure and lipophilicity of naturally occurring polyphenols, Acta Chim. Slov. 60 (2013) 781–789.
[10] Y.K. Yoon, M.A. Ali, A.C. Wei, T.S. Choon, K.Y. Khaw, V. Murugaiyah, H. Osman, V.H. Masand, Synthesis, characterization, and molecular docking analysis of novel benzimidazole derivatives as cholinesterase inhibitors, Bioorg. Chem. 49 (2013) 33–39.
[11] S.N. Bukhari, I. Jantan, V.H. Masand, D.T. Mahajan, M. Sher, M. Naeem-ul-Hassan, M.W. Amjad, Synthesis of alpha, beta-unsaturated carbonyl based compounds as acetylcholinesterase and butyrylcholinesterase inhibitors: characterization, molecular modeling, QSAR studies and effect against amyloid beta-induced cytotoxicity, Eur. J. Med. Chem. 83 (2014) 355–365.
[12] V.H. Masand, D.T. Mahajan, P. Gramatica, J. Barlow, Tautomerism and multiple modelling enhance the efficacy of QSAR: antimalarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl, Med. Chem. Res. 23 (2014) 4825–4835.
[13] V.H. Masand, A.A. Toropov, A.P. Toropova, D.T. Mahajan, QSAR models for antimalarial activity of 4-aminoquinolines, Curr. Comput. Aided Drug Des. 10 (2014) 75–82.
[14] V. Rastija, V.H. Masand, QSAR of antitrypanosomal activities of polyphenols and their analogues using multiple linear regression and artificial neural networks, Comb. Chem. High. Throughput Screen 17 (2014) 709–717.
[15] V.H. Masand, D.T. Mahajan, G.M. Nazeruddin, T. Ben Hadda, V. Rastija, A.M. Alfeefy, Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model, Med. Chem. Res. 24 (2015) 1241–1264.
[16] P. Gramatica, S. Cassani, N. Chirico, QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. 35 (2014) 1036–1044.
[17] P. Gramatica, S. Cassani, N. Chirico, QSARINS-Chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. 35 (2014) 1036–1044.
[18] P. Gramatica, External evaluation of QSAR models, in addition to cross-validation verification of predictive capability on totally new chemicals, Mol. Inf. 33 (2014) 311–314.
[19] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, Baskin II, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57 (2014) 4977–5010.
[20] P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: a new software for the development, analysis, and validation of QSAR MLR models, J. Comput. Chem. 34 (2013) 2121–2132.
[21] P. Gramatica, On the development and validation of QSAR models, Methods Mol. Biol. 930 (2013) 499–526.
[22] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, P. Gramatica, Daphnia and fish toxicity of (benzo)triazoles: validated QSAR models, and interspecies quantitative activity-activity modelling, J. Hazard Mater. 258–259 (2013) 50–60.
[23] P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap, E. Papa, QSAR modeling is not push a button and find a correlation: a case study of toxicity of (Benzo-)triazoles on algae, Mol. Inf. (2012) 817–835.
[24] D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation, J. Chem. Inf. Comput. Sci. 43 (2003) 579–586.
[25] J. Huang, X. Fan, Why QSAR fails: an empirical evaluation using conventional computational approach, Mol. Pharm. 8 (2011) 600–608.

[26] J.C. Dearden, M.T. Cronin, K.L. Kaiser, How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR), SAR QSAR Environ. Res. 20 (2009) 241–266.
[27] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (2012) 2570–2578.
[28] A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, Data set modelability by QSAR, J. Chem. Inf. Model. 54 (2014) 1–4.
[29] N. Chirico, P. Gramatica, Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 51 (2011) 2320–2335.
[30] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, J. Chem. Inf. Model. 52 (2012) 2044–2058.