Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(4): 112
https://doi.org/10.31083/j.fbl2704112

Original Research

QSAR based on hybrid optimal descriptors as a tool to predict antibacterial activity against *Staphylococcus aureus*

Karel Nesměrák[1,]*, Andrey Toropov[2], Ilkay Yildiz[3]

[1]Department of Analytical Chemistry, Faculty of Science, Charles University, 128 43 Prague 2, Czech Republic
[2]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, 20156 Milano, Italy
[3]Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Ankara University, Yenimahalle, 06560 Ankara, Turkey
*Correspondence: karel.nesmerak@natur.cuni.cz (Karel Nesměrák)

Academic Editor: Graham Pawelec
Submitted: 8 December 2021 Revised: 19 January 2022 Accepted: 25 January 2022 Published: 1 April 2022

Abstract

Background: *Staphylococcus aureus* bacterial infections are still a serious health care problem. Therefore, the development of new drugs for these infections is a constant requirement. Quantitative structure–activity relationship (QSAR) methods can assist this development. **Methods**: The study included 151 structurally diverse compounds with antibacterial activity against *S. aureus* ATCC 25923 (Endpoint 1) or the drug-resistant clinical isolate of *S. aureus* (Endpoint 2). QSARs based on hybrid optimal descriptors were used. **Results**: The predictive potential of developed models has been checked with three random splits into training, passive training, calibration, and validation sets. The proposed models give satisfactory predictive models for both endpoints examined. **Conclusions**: The results of the study show the possibility of SMILES-based QSAR in the evaluation of the antibacterial activity of structurally diverse compounds for both endpoints. Although the developed models give satisfactory predictive models for both endpoints examined, splitting has an apparent influence on the statistical quality of the models.

Keywords: antibacterial activity; CORAL software; hybrid optimal descriptors; Monte Carlo method; QSAR; SMILES

1. Introduction

Quantitative structure-property/activity relationships (QSPRs/QSARs) are a tool to model different biological activities, such as antimicrobial [1], anti-HIV [2], anti-cancer [3], antimycobacterial [4], enzyme selectivity [5], multi-targets drug discovery [6–8], absorption, distribution, metabolism, excretion and toxicity (ADMET) analysis [9], finally, the influence of QSPR/QSAR to epistemological processes in natural sciences is also a significant object of study [10].

The dramatic increase in numerous multidrug-resistant bacterial infections in recent decades has become a serious health care problem. In particular, multidrug-resistant strains of Gram-positive bacterial pathogens, namely *Staphylococcus aureus*, which dominate world-wide bacterial infection rates, are a problem of very serious significance [11,12]. Although various antimicrobial drugs are used in treatment, a high mortality rate is still a serious problem in *S. aureus* bacteremia, and the development of new drugs or the elaboration of new types of previously known drugs remains a very actual task [13–15].

In our previous work, we have dealt with the synthesis of new antibacterial agents, determined their minimum inhibitory concentrations (MIC) against a number of microorganisms, and evaluated their properties using various QSAR approaches. First, in 2010 a novel series of *N*-(2-hydroxyphenyl)benzamides and *N*-(2-hydroxyphenyl)-2-phenylacetamides was synthesized (Fig. 1A) [16]. The microbiological results indicated that they possess a broad spectrum of activity against various pathogens (MIC values between 1.95 and 500 μg/mL). A follow-up study [17] using classical QSAR and 3D-common-feature pharmacophore hypothesis approaches showed that the insertion of a methylene group between the phenyl and carboxyamido moiety decreases MIC. In contrast, the substituent at position R^1 is important for the increase of activity, and similarly, substituting position R^3 with a group enhancing the electron-donor capability of the phenolic ring system increased the potency of a compound. Finally, it was found that the benzamide derivatives exhibited the greatest activity against drug-resistant bacteria, including *S. aureus*. These findings led to the synthesis of the series of 2-(*p*-substituted benzyl)-5-(2-substituted acetamido)benzoxazoles (Fig. 1B) [18]. The microbiological assay showed that the compounds possessed a large spectrum of MIC between 7.8–250 μg/mL. The 2D-QSAR showed that the width and hydrophobicity of the R^1 substituent are directly proportional to MIC against methicillin-resistant *S. aureus*. No methylene bridge should take place between the benzoxazole moiety and the *p*-substituted phenyl group. These observations led to the design of the next series of 5(or 6)-nitro/amino-2-(substituted phenyl/benzyl)benzoxazoles (Fig. 1C) [19]. Antibacterial evaluation indicated a broad spectrum of activity against the tested microorganisms (including *S. aureus*) with MIC values between 12.5

Copyright: © 2022 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

Publisher's Note: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

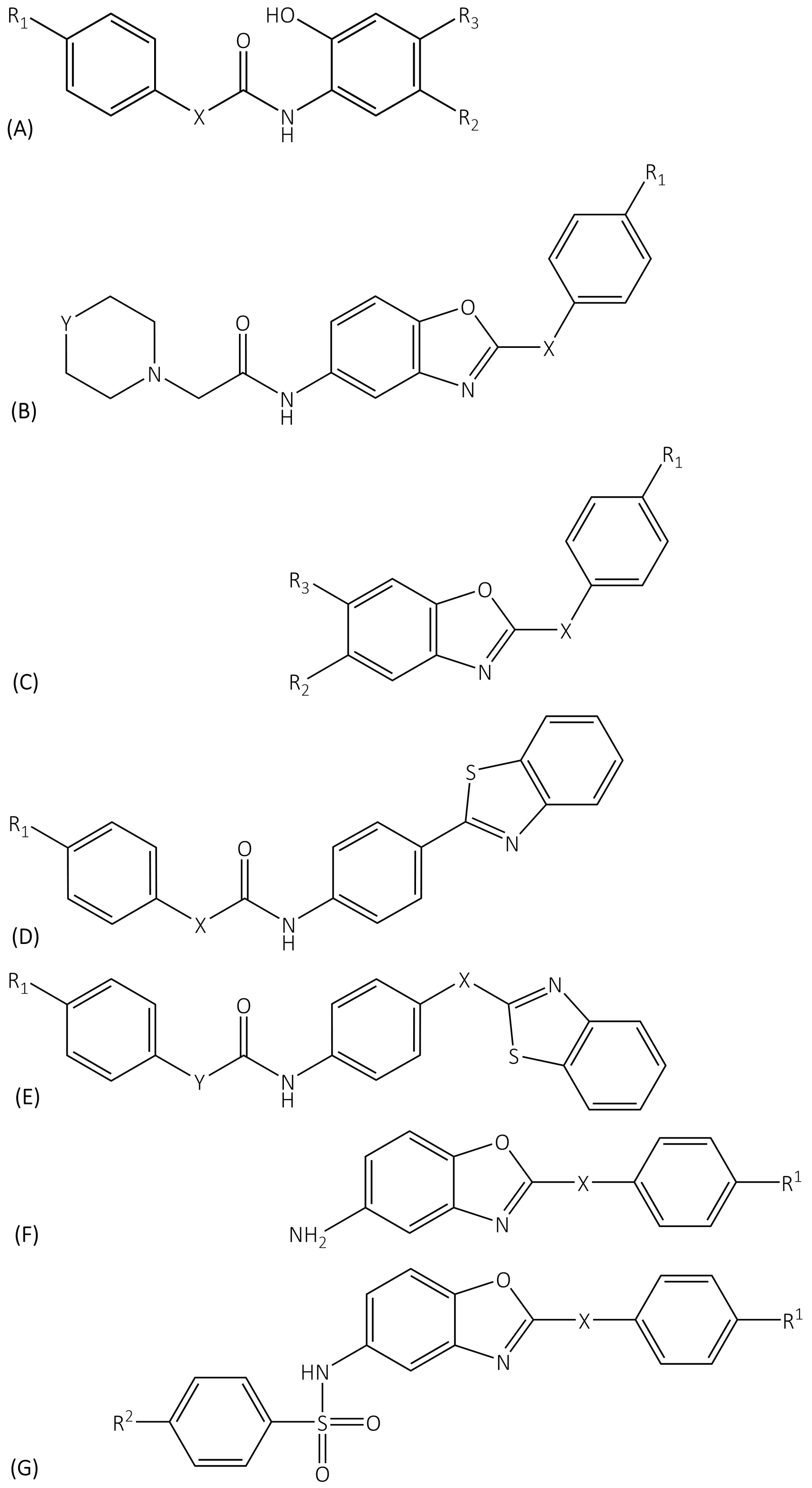

Fig. 1. General chemical structures of the examined class of compounds with activity against *Staphylococcus aureus*. (A) *N*-(2-hydroxyphenyl)benzamides (X = –), respectively *N*-(2-hydroxyphenyl)-2-phenylacetamides (X = $–CH_2–$). (B) 2-(*p*-substituted benzyl)-5-(2-substitutedacetamido)benzoxazoles (X = –, $–CH_2–$; Y = –, –O–, $–CH_2–$, $>N–CH_3$, >N–phenyl, >NH, $>N–CH_3$). (C) 2-(phenyl/benzyl)benzoxazoles (X = –, $–CH_2–$). (D) 2-[4-(4-substitutedbenzamido/phenylacetamido)phenyl]benzothiazoles (X = –, $–CH_2–$). (E) 2-[4-(4-substituted benzamido/phenylacetamido/phenylpropionamido)benzyl/phenyl]benzothiazoles (X = –, $–CH_2–$; Y = –, $–CH_2–$, $–C_2H_4–$). (F) 5-amino-2-(4-substituted phenyl/benzyl)benzoxazoles (X = –, $–CH_2–$). (G) 2-substituted-5-(4-nitro/amino-phenylsulfonamido)benzoxazoles (X = –, $–CH_2–$).

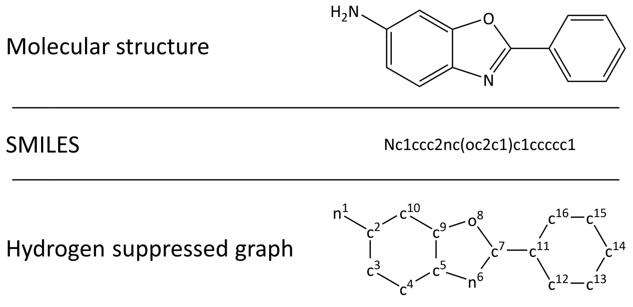

Fig. 2. An example of the molecular structure together with SMILES and hydrogen suppressed graph for compound 76.

and >400 μg/mL. However, the 2D-QSAR analysis using the multivariable regression analysis was performed only for *Bacillus subtilis*. Next, the series of 2-[4-(4-substituted benzamido/phenylacetamido)phenyl] benzothiazoles was prepared and evaluated, the MIC values ranged between 6.25 and 100 μg/mL (Fig. 1D) [20]. The majority of the compounds showed more antibacterial activity against the screened drug-resistant clinical isolate of *S. aureus* compared to the non-resistant of *S. aureus* ATCC 25923. Further development based on the use of structural effects already identified led to the synthesis of a series of 2-[4-(4-substituted benzamido/phenylacetamido/phenylpropionamido)benzyl/phenyl] benzothiazoles (Fig. 1E) [21]. Evaluation of their activity against various bacterial pathogens showed MIC values between 6.25 and 200 μg/mL. Two compounds exhibited great antimicrobial activity against the drug-resistant clinical isolate of *S. aureus*, but the structural effects were not evaluated by QSAR. Finally, two groups of antibacterial compounds were designed and synthesized using two pharmacologically compatible moieties in one molecule by attaching a sulfonamide group to a benzoxazole [22]. The first was derivatives of 5-amino-2-(4-substituted phenyl/benzyl)benzoxazole (Fig. 1F). The derivatives of 2-substituted-5-(4-nitro/amino-phenylsulfonamido)benzoxazole (Fig. 1G) form the second group. Minimal inhibitory concentrations of these derivatives are between 8 and 256 μg/mL. The structural effects on the MIC of these compounds have been evaluated only for their activity against *Mycobacterium tuberculosis* [23,24].

An irreplaceable step in the targeted search for suitable antibacterial compounds is the analysis of the relationship between the structure and the biological effect of a substance. The structural diversity of the compounds mentioned above does not allow the use of classical QSAR procedures to evaluate their antibacterial activity against *S. aureus*. Therefore, in this work, we used hybrid optimal descriptors calculated with the molecular graph, i.e., based on the description of the entire structure of a molecule. A simplified molecular input-line entry system (SMILES) represents an appealing alternative to representing the molecular structure by a graph, and the development of SMILES-based QSAR becomes a promising way of research work in the field of QSAR theory and applications [25,26]. From the medicinal chemistry point of view, only one SMILES-based QSAR model describing the effect of structure on antibacterial activity against *S. aureus* has been published yet. In 2020, Lotfi *et al.* [27] studied the possibility of predicting the MIC of 204 ionic liquids against *S. aureus* and found that developed QSAR models are at a high level.

Consequently, this study aims to combine the results of previous work and evaluate the antibacterial effects of a total of 151 compounds against *S. aureus* using SMILES-based hybrid optimal descriptors.

2. Materials and methods

2.1 Data

The structures of the examined compounds and their MIC against (i) *S. aureus* ATCC 25923, and (ii) drug-resistant clinical isolate of *S. aureus* were taken from previous publications [16,18–22]. The molecular structure of the compounds was transferred to the SMILES notation using ACD/ChemSketch software [28]. Due to the structural diversity of examined compounds, the MIC values were recalculated from μg/mL to mol/L units and expressed as logarithms of reciprocal values. This makes it possible to unify the antibacterial activity of the test substances with respect to the number of molecules (and not the weight). The complete data are represented in Table 1.

2.2 Optimal hybrid descriptor

The molecular structure can be represented by SMILES and/or a molecular graph (hydrogen suppressed graph). Fig. 2 contains an example of the molecular structure together with the SMILES and the hydrogen suppressed graph for compound 78.

The hybrid optimal descriptors [10] are sensitive to both above-mentioned representations of the molecular structure. Hybrid optimal descriptors are calculated by optimization of the so-called correlation weights of the SMILES attributes together with the correlation weights of the graph invariants. The optimal hybrid descriptor *DCW*(*T*,*N*) is applied for a predictive model of endpoint via the equation:

$$Endpoint = C_0 + C_1 \times DCW(T, N) \qquad (1)$$

$$DCW(T, N) = DCW_{SMILES}(T, N) + DCW_{graph}(T, N) \qquad (2)$$

Table 1. Structures of the examined compounds (for general structures A–G see Fig. 1), SMILES notation, and experimental minimal inhibition concentrations (mol/L) against *Staphylococcus aureus* ATCC 25923, and drug-resistant clinical isolate of *Staphylococcus aureus*.

No.	Structure	R^1	R^2	R^3	X	Y	SMILES	log 1/*c*	
								S. a.	*S. a.* isol.
1	A	–$C(CH_3)_3$	–H	–NO_2	—	—	Oc2cc(ccc2NC(=O)c1ccc(cc1)C(C)(C)C)[N+]([O-])=O	4.61	5.21
2	A	–H	–H	–NO_2	—	—	Oc2cc(ccc2NC(=O)c1ccccc1)[N+]([O-])=O	3.62	3.32
3	A	–F	–H	–NO_2	—	—	Oc2cc(ccc2NC(=O)c1ccc(F)cc1)[N+]([O-])=O	4.55	4.25
4	A	–Br	–H	–NO_2	—	—	O=C(Nc1ccc(cc1O)[N+]([O-])=O)c2ccc(Br)cc2	4.64	2.83
5	A	–C_2H_5	–H	–NO_2	—	—	Oc2cc(ccc2NC(=O)c1ccc(CC)cc1)[N+]([O-])=O	3.66	3.36
6	A	–H	–NO_2	–H	—	—	Oc2ccc(cc2NC(=O)c1ccccc1)[N+]([O-])=O	3.92	4.22
7	A	–C_2H_5	–NO_2	–H	—	—	Oc2ccc(cc2NC(=O)c1ccc(CC)cc1)[N+]([O-])=O	4.56	4.56
8	A	–F	–NO_2	–H	—	—	Oc2ccc(cc2NC(=O)c1ccc(F)cc1)[N+]([O-])=O	4.55	4.25
9	A	–Br	–H	–NO_2	–CH_2–	—	Brc2ccc(CC(=O)Nc1ccc(cc1O)[N+]([O-])=O)cc2	3.15	3.45
10	A	–Cl	–H	–NO_2	–CH_2–	—	Clc2ccc(CC(=O)Nc1ccc(cc1O)[N+]([O-])=O)cc2	3.39	3.39
11	A	–CH_3	–H	–NO_2	–CH_2–	—	Oc2cc(ccc2NC(=O)Cc1ccc(C)cc1)[N+]([O-])=O	3.06	2.76
12	A	–F	–H	–NO_2	–CH_2–	—	Oc2cc(ccc2NC(=O)Cc1ccc(F)cc1)[N+]([O-])=O	3.37	3.67
13	A	–CH_3	–NO_2	–H	–CH_2–	—	Oc2ccc(cc2NC(=O)Cc1ccc(C)cc1)[N+]([O-])=O	3.06	3.36
14	A	–F	–NO_2	–H	–CH_2–	—	Oc2ccc(cc2NC(=O)Cc1ccc(F)cc1)[N+]([O-])=O	3.97	3.97
15	A	–$C(CH_3)_3$	–H	–NH_2	—	—	Oc2cc(N)ccc2NC(=O)c1ccc(cc1)C(C)(C)C	4.26	4.56
16	A	–H	–H	–NH_2	—	—	Oc2cc(N)ccc2NC(=O)c1ccccc1	3.26	4.17
17	A	–F	–H	–NH_2	—	—	Oc2cc(N)ccc2NC(=O)c1ccc(F)cc1	3.60	4.20
18	A	–Br	–H	–NH_2	—	—	O=C(Nc1ccc(N)cc1O)c2ccc(Br)cc2	3.39	4.60
19	A	–C_2H_5	–H	–NH_2	—	—	Oc2cc(N)ccc2NC(=O)c1ccc(CC)cc1	3.31	4.52
20	A	–H	–NH_2	–H	—	—	Oc2ccc(N)cc2NC(=O)c1ccccc1	3.86	3.86
21	A	–C_2H_5	–NH_2	–H	—	—	Oc2ccc(N)cc2NC(=O)c1ccc(CC)cc1	4.22	4.22
22	A	–F	–NH_2	–H	—	—	Oc2ccc(N)cc2NC(=O)c1ccc(F)cc1	4.20	4.20
23	A	–Br	–H	–NH_2	–CH_2–	—	Brc2ccc(CC(=O)Nc1ccc(N)cc1O)cc2	3.41	4.01
24	A	–Cl	–H	–NH_2	–CH_2–	—	Clc2ccc(CC(=O)Nc1ccc(N)cc1O)cc2	3.65	3.95
25	A	–CH_3	–H	–NH_2	–CH_2–	—	Oc2cc(N)ccc2NC(=O)Cc1ccc(C)cc1	3.01	3.61
26	A	–F	–H	–NH_2	–CH_2–	—	Oc2cc(N)ccc2NC(=O)Cc1ccc(F)cc1	3.32	3.62
27	A	–CH_3	–NH_2	–H	–CH_2–	—	Oc2ccc(N)cc2NC(=O)Cc1ccc(C)cc1	3.91	3.61
28	A	–F	–NH_2	–H	–CH_2–	—	Oc2ccc(N)cc2NC(=O)Cc1ccc(F)cc1	3.62	3.92
29	B	–Cl	–H	–H	–CH_2–	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(Cc3ccc(Cl)cc3)nc2c1	3.79	3.49
30	B	–CH_3	–H	–H	–CH_2–	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(Cc3ccc(C)cc3)nc2c1	3.47	3.47
31	B	–H	–H	–H	–CH_2–	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(Cc3ccccc3)nc2c1	3.45	3.45
32	B	–F	–H	–H	–CH_2–	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(Cc3ccc(F)cc3)nc2c1	3.47	3.77
33	B	–Cl	–H	–H	–CH_2–	–CH_2–	O=C(CN1CCOCC1)Nc1ccc2oc(Cc3ccc(Cl)cc3)nc2c1	4.09	3.79
34	B	–CH_3	–H	–H	–CH_2–	–CH_2–	Cc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCCCC1	4.07	3.46
35	B	–H	–H	–H	–CH_2–	–CH_2–	O=C(CN1CCCCC1)Nc1cc2nc(Cc3ccccc3)oc2cc1	3.45	3.45
36	B	–F	–H	–H	–CH_2–	–CH_2–	Fc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCCCC1	3.47	3.77
37	B	–Br	–H	–H	–CH_2–	–CH_2–	Brc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCCCC1	3.84	3.84
38	B	–Cl	–H	–H	–CH_2–	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(Cc3ccc(Cl)cc3)oc2cc1	3.80	3.50
39	B	–CH_3	–H	–H	–CH_2–	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(Cc3ccc(C)cc3)oc2cc1	3.78	3.48
40	B	–H	–H	–H	–CH_2–	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(Cc3ccccc3)oc2cc1	3.16	3.46
41	B	–F	–H	–H	–CH_2–	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(Cc3ccc(F)cc3)oc2cc1	3.49	3.79
42	B	–Br	–H	–H	–CH_2–	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(Cc3ccc(Br)cc3)oc2cc1	3.55	3.85
43	B	–Cl	–H	–H	–CH_2–	>N–Ph	Clc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCN(CC1)c1ccccc1	3.57	3.57
44	B	–CH_3	–H	–H	–CH_2–	>N–Ph	Cc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCN(CC1)c1ccccc1	3.55	3.55
45	B	–H	–H	–H	–CH_2–	>N–Ph	O=C(CN1CCN(CC1)c1ccccc1)Nc1cc2nc(Cc3ccccc3)oc2cc1	3.23	3.83

Table 1. Continued.

No.	Structure	R^1	R^2	R^3	X	Y	SMILES	log 1/*c*	
								S. a.	*S. a.* isol.
46	B	–F	–H	–H	–CH_2–	>N–Ph	Fc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCN(CC1)c1ccccc1	3.55	3.85
47	B	–Br	–H	–H	–CH_2–	>N–Ph	Brc1ccc(cc1)Cc1nc2cc(ccc2o1)NC(=O)CN1CCN(CC1)c1ccccc1	3.31	3.91
48	B	–H	–H	–H	—	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(nc2c1)c1ccccc1	3.43	3.73
49	B	–F	–H	–H	—	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(nc2c1)c1ccc(F)cc1	3.45	3.75
50	B	–C_2H_5	–H	–H	—	–O–	O=C(CN1CCOCC1)Nc1ccc2oc(nc2c1)c1ccc(CC)cc1	3.77	3.77
51	B	–H	–H	–H	—	>NH	O=C(CN1CCNCC1)Nc1cc2nc(oc2cc1)c1ccccc1	3.73	3.73
52	B	–F	–H	–H	—	>NH	Fc1ccc(cc1)c1nc2cc(ccc2o1)NC(=O)CN1CCNCC1	3.75	3.75
53	B	–C_2H_5	–H	–H	—	>NH	CCc1ccc(cc1)c1nc2cc(ccc2o1)NC(=O)CN1CCNCC1	4.37	4.37
54	B	–$C(CH_3)_3$	–H	–H	—	>NH	CC(C)(C)c1ccc(cc1)c1nc2cc(ccc2o1)NC(=O)CN1CCNCC1	4.40	4.70
55	B	–H	–H	–H	—	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(oc2cc1)c1ccccc1	3.75	4.05
56	B	–F	–H	–H	—	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(oc2cc1)c1ccc(F)cc1	3.47	3.77
57	B	–C_2H_5	–H	–H	—	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(oc2cc1)c1ccc(CC)cc1	4.08	4.08
58	B	–$C(CH_3)_3$	–H	–H	—	>N–CH_3	CN1CCN(CC1)CC(=O)Nc1cc2nc(oc2cc1)c1ccc(cc1)C(C)(C)C	4.42	4.72
59	C	–$C(CH_3)_3$	–H	–NO_2	—	—	CC(C)(C)c1ccc(cc1)c1nc2cc(ccc2o1)[N+]([O-])=O	3.47	3.47
60	C	–H	–H	–NO_2	—	—	[O-][N+](=O)c1cc2nc(oc2cc1)c1ccccc1	3.08	3.38
61	C	–F	–H	–NO_2	—	—	[O-][N+](=O)c1cc2nc(oc2cc1)c1ccc(F)cc1	3.41	3.41
62	C	–Br	–H	–NO_2	—	—	[O-][N+](=O)c1cc2nc(oc2cc1)c1ccc(Br)cc1	3.50	2.90
63	C	–C_2H_5	–H	–NO_2	—	—	[O-][N+](=O)c1cc2nc(oc2cc1)c1ccc(CC)cc1	3.43	3.43
64	C	–H	–NO_2	–H	—	—	[O-][N+](=O)c1ccc2nc(oc2c1)c1ccccc1	3.08	2.78
65	C	–C_2H_5	–NO_2	–H	—	—	[O-][N+](=O)c1ccc2nc(oc2c1)c1ccc(CC)cc1	3.43	3.43
66	C	–F	–NO_2	–H	—	—	[O-][N+](=O)c1ccc2nc(oc2c1)c1ccc(F)cc1	3.41	3.41
67	C	–Br	–H	–NO_2	–CH_2–	—	[O-][N+](=O)c1cc2nc(Cc3ccc(Br)cc3)oc2cc1	3.52	3.52
68	C	–Cl	–H	–NO_2	–CH_2–	—	[O-][N+](=O)c1cc2nc(Cc3ccc(Cl)cc3)oc2cc1	3.16	3.46
69	C	–F	–H	–NO_2	–CH_2–	—	[O-][N+](=O)c1cc2nc(Cc3ccc(F)cc3)oc2cc1	3.43	3.43
70	C	–F	–NO_2	–H	–CH_2–	—	[O-][N+](=O)c1ccc2nc(Cc3ccc(F)cc3)oc2c1	3.43	3.43
71	C	–CH_3	–NO_2	–H	–CH_2–	—	[O-][N+](=O)c1ccc2nc(Cc3ccc(C)cc3)oc2c1	3.73	3.73
72	C	–$C(CH_3)_3$	–H	–NH_2	—	—	CC(C)(C)c1ccc(cc1)c1nc2cc(N)ccc2o1	3.43	3.43
73	C	–F	–H	–NH_2	—	—	Fc1ccc(cc1)c1nc2cc(N)ccc2o1	3.36	3.36
74	C	–Br	–H	–NH_2	—	—	Brc1ccc(cc1)c1nc2cc(N)ccc2o1	3.46	3.46
75	C	–C_2H_5	–H	–NH_2	—	—	CCc1ccc(cc1)c1nc2cc(N)ccc2o1	3.38	3.38
76	C	–H	–NH_2	–H	—	—	Nc1ccc2nc(oc2c1)c1ccccc1	3.32	3.32
77	C	–C_2H_5	–NH_2	–H	—	—	CCc1ccc(cc1)c1nc2ccc(N)cc2o1	3.98	3.38
78	C	–F	–NH_2	–H	—	—	Fc1ccc(cc1)c1nc2ccc(N)cc2o1	3.66	3.36
79	C	–Br	–H	–NH_2	–CH_2–	—	Brc1ccc(cc1)Cc1nc2cc(N)ccc2o1	3.48	3.48
80	C	–Cl	–H	–NH_2	–CH_2–	—	Clc1ccc(cc1)Cc1nc2cc(N)ccc2o1	3.41	3.41
81	C	–F	–H	–NH_2	–CH_2–	—	Fc1ccc(cc1)Cc1nc2cc(N)ccc2o1	3.38	3.38
82	C	–CH_3	–NH_2	–H	–CH_2–	—	Cc1ccc(cc1)Cc1nc2ccc(N)cc2o1	3.38	3.38
83	C	–F	–NH_2	–H	–CH_2–	—	Fc1ccc(cc1)Cc1nc2ccc(N)cc2o1	3.08	3.38
84	D	–F	–H	–H	—	—	Fc1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.84	4.14
85	D	–Cl	–H	–H	—	—	Clc1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.86	4.47
86	D	–Br	–H	–H	—	—	Brc1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.91	4.21
87	D	–C_2H_5	–H	–H	—	—	CCc1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.86	4.76
88	D	–$C(CH_3)_3$	–H	–H	—	—	CC(C)(C)c1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.89	4.49
89	D	–NO_2	–H	–H	—	—	[O-][N+](=O)c1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.88	4.48
90	D	–F	–H	–H	–CH_2–	—	Fc1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.56	3.86
91	D	–Cl	–H	–H	–CH_2–	—	Clc1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.88	3.88

Table 1. Continued.

No.	Structure	R^1	R^2	R^3	X	Y	SMILES	log 1/*c*	
								S. a.	*S. a.* isol.
92	D	–Br	–H	–H	$–CH_2–$	—	Brc1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.93	4.23
93	D	$–CH_3$	–H	–H	$–CH_2–$	—	Cc1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.86	4.16
94	D	$–OCH_3$	–H	–H	$–CH_2–$	—	COc1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.87	4.48
95	E	–H	–H	–H	—	—	O=C(Nc1ccc(cc1)c1nc2ccccc2s1)c1ccccc1	3.82	4.12
96	E	$–OCH(CH_3)C_2H_5$	–H	–H	—	—	CC(CC)Oc1ccc(cc1)C(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.91	4.21
97	E	–H	–H	–H	—	$–CH_2–$	O=C(Cc1ccccc1)Nc1ccc(cc1)c1nc2ccccc2s1	3.84	4.14
98	E	$–NO_2$	–H	–H	—	$–CH_2–$	[O-][N+](=O)c1ccc(cc1)CC(=O)Nc1ccc(cc1)c1nc2ccccc2s1	3.89	4.49
99	E	–H	–H	–H	—	$–C_2H_4–$	O=C(CCc1ccccc1)Nc1ccc(cc1)c1nc2ccccc2s1	3.86	4.46
100	E	–F	–H	–H	$–CH_2–$	—	Fc1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.86	4.16
101	E	–Cl	–H	–H	$–CH_2–$	—	Clc1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	4.18	4.18
102	E	–Br	–H	–H	$–CH_2–$	—	Brc1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	4.83	4.53
103	E	$–NO_2$	–H	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.59	4.19
104	E	$–C_2H_5$	–H	–H	$–CH_2–$	—	CCc1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.87	4.17
105	E	$–C(CH_3)_3$	–H	–H	$–CH_2–$	—	CC(C)(C)c1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.90	4.20
106	E	$–OCH(CH_3)C_2H_5$	–H	–H	$–CH_2–$	—	CC(CC)Oc1ccc(cc1)C(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.92	4.22
107	E	–H	–H	–H	$–CH_2–$	$–CH_2–$	O=C(Cc1ccccc1)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.86	4.16
108	E	–F	–H	–H	$–CH_2–$	$–CH_2–$	Fc1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.58	4.18
109	E	–Cl	–H	–H	$–CH_2–$	$–CH_2–$	Clc1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.90	4.20
110	E	–Br	–H	–H	$–CH_2–$	$–CH_2–$	Brc1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.94	4.24
111	E	$–NO_2$	–H	–H	$–CH_2–$	$–CH_2–$	[O-][N+](=O)c1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.91	4.21
112	E	$–CH_3$	–H	–H	$–CH_2–$	$–CH_2–$	Cc1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	4.17	4.47
113	E	$–OCH_3$	–H	–H	$–CH_2–$	$–CH_2–$	COc1ccc(cc1)CC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.89	4.19
114	E	–H	–H	–H	$–CH_2–$	$–C_2H_4–$	O=C(CCc1ccccc1)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.57	4.17
115	E	$–OCH_3$	–H	–H	$–CH_2–$	$–C_2H_4–$	COc1ccc(cc1)CCC(=O)Nc1ccc(cc1)Cc1nc2ccccc2s1	3.91	4.21
116	F	–H	–H	–H	—	—	Nc1cc2nc(oc2cc1)c3ccccc3	3.52	3.52
117	F	–Cl	–H	–H	—	—	Clc1ccc(cc1)c2nc3cc(N)ccc3o2	3.58	3.58
118	F	–F	–H	–H	—	—	Fc1ccc(cc1)c2nc3cc(N)ccc3o2	3.55	3.55
119	F	–Br	–H	–H	—	—	Brc1ccc(cc1)c2nc3cc(N)ccc3o2	3.65	3.65
120	F	$–C_2H_5$	–H	–H	—	—	CCc1ccc(cc1)c2nc3cc(N)ccc3o2	3.57	3.57
121	F	$–CH_3$	–H	–H	—	—	Cc1ccc(cc1)c2nc3cc(N)ccc3o2	3.54	3.54
122	F	$–OCH_3$	–H	–H	—	—	COc1ccc(cc1)c2nc3cc(N)ccc3o2	3.57	3.57
123	F	–H	–H	–H	$–CH_2–$	—	Nc2cc3nc(Cc1ccccc1)oc3cc2	3.54	3.54
124	F	–Cl	–H	–H	$–CH_2–$	—	Clc1ccc(cc1)Cc2nc3cc(N)ccc3o2	3.61	3.61
125	F	–F	–H	–H	$–CH_2–$	—	Fc1ccc(cc1)Cc2nc3cc(N)ccc3o2	3.58	3.58
126	F	–Br	–H	–H	$–CH_2–$	—	Brc1ccc(cc1)Cc2nc3cc(N)ccc3o2	3.68	3.68
127	F	$–CH_3$	–H	–H	$–CH_2–$	—	Cc1ccc(cc1)Cc2nc3cc(N)ccc3o2	3.57	3.57
128	G	–H	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccccc4	3.15	3.46
129	G	–Cl	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(Cl)cc4	3.19	3.49
130	G	–F	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(F)cc4	3.18	3.48
131	G	–Br	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(Br)cc4	3.24	3.54
132	G	$–C_2H_5$	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(CC)cc4	3.19	3.49
133	G	$–CH_3$	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(C)cc4	3.47	3.47
134	G	$–OCH_3$	$–NH_2$	–H	—	—	Nc1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(OC)cc4	3.19	3.49
135	G	–H	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccccc4	3.49	3.49
136	G	–Cl	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(Cl)cc4	3.53	3.53
137	G	–F	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(F)cc4	3.51	3.51

Table 1. Continued.

No.	Structure	R1	R2	R3	X	Y	SMILES	log 1/*c*	
								S. a.	*S. a.* isol.
138	G	–Br	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(Br)cc4	3.57	3.57
139	G	$–C_2H_5$	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(CC)cc4	3.22	3.52
140	G	$–CH_3$	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(C)cc4	3.50	3.50
141	G	$–OCH_3$	$–NO_2$	–H	—	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc2cc3nc(oc3cc2)c4ccc(OC)cc4	3.22	3.52
142	G	–H	$–NH_2$	–H	$–CH_2–$	—	Nc1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccccc2)oc4cc3	3.17	3.47
143	G	–Cl	$–NH_2$	–H	$–CH_2–$	—	Nc1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(Cl)cc2)oc4cc3	3.21	3.51
144	G	–F	$–NH_2$	–H	$–CH_2–$	—	Nc1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(F)cc2)oc4cc3	3.19	3.49
145	G	–Br	$–NH_2$	–H	$–CH_2–$	—	Nc1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(Br)cc2)oc4cc3	3.25	3.55
146	G	$–CH_3$	$–NH_2$	–H	$–CH_2–$	—	Nc1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(C)cc2)oc4cc3	3.79	3.49
147	G	–H	$–NO_2$	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccccc2)oc4cc3	3.20	3.50
148	G	–Cl	$–NO_2$	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(Cl)cc2)oc4cc3	3.24	3.54
149	G	–F	$–NO_2$	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(F)cc2)oc4cc3	3.22	3.52
150	G	–Br	$–NO_2$	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(Br)cc2)oc4cc3	3.28	3.58
151	G	$–CH_3$	$–NO_2$	–H	$–CH_2–$	—	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)Nc3cc4nc(Cc2ccc(C)cc2)oc4cc3	3.22	3.52

Table 2. The adjacency matrix for the hydrogen suppressed graph, which is represented in Fig. 2.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16				
		N	c	c	c	c	n	c	o	c	c	c	c	c	c	c	c	*EC*0	*EC*1	*EC*2	*EC*3
1	N	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	5	14
2	c	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	3	5	14	27
3	c	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	5	10	26
4	c	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2	5	12	28
5	c	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	3	7	18	45
6	n	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	2	6	14	37
7	c	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	3	7	19	45
8	o	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2	6	14	38
9	c	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	3	7	19	44
10	c	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	6	12	33
11	c	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	3	7	17	41
12	c	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2	5	11	26
13	c	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	4	9	19
14	c	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2	4	8	18
15	c	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	4	9	19
16	c	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	5	11	26

Table 3. The statistical characteristics of the models for Endpoint 1 were calculated with Eqs. 13–15 for (A) active training set, (P) passive training set, (C) calibration set, and (V) validation set (*n* is the number of conpounds in the corresponding set).

Eqn.		*n*	R^2	*CCC*	*IIC*	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\overline{R^2_m}$	*RMSE*	*MAE*	*F*
13	A	37	0.6297	0.7728	0.4830	0.5739					0.233	0.193	60
	P	40	0.6298	0.6815	0.4533	0.5866					0.344	0.299	65
	C	37	0.7749	0.8776	0.8799	0.7430	0.8241	0.7577	0.9168	0.6823	0.129	0.105	121
	V	37	0.7201								0.129	0.108	
14	A	37	0.4100	0.5816	0.6066	0.3376					0.309	0.272	24
	P	38	0.5909	0.6108	0.6224	0.5364					0.343	0.274	52
	C	38	0.6626	0.8041	0.8132	0.6280	0.7066	0.6374	0.8833	0.5365	0.152	0.116	71
	V	38	0.7565								0.128	0.105	
15	A	37	0.5779	0.7325	0.6461	0.5261					0.256	0.213	48
	P	39	0.7778	0.7614	0.6454	0.7445					0.268	0.221	130
	C	38	0.7413	0.8446	0.8609	0.7129	0.6594	0.6509	0.8458	0.6361	0.168	0.137	103
	V	37	0.5693								0.260	0.203	

Table 4. Criteria of predictive potential of a model.

Criterion of the predictive potential	Ref.
$R=\frac{n\sum xy-\sum x\sum y}{\sqrt{(n\sum x^2-(\sum x)^2(n\sum y^2-(\sum y)^2}}$	[32]
$Q^2=1-\frac{\sum(y_k-\acute{y}_k)^2}{\sum(y_k-\bar{y}_k)^2}$	[33]
$Q_{F1}^2=1-\frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i-y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}}(y_i-\bar{y}_{TR})^2\right]/N_{EXT}}$	[34]
$Q_{F2}^2=1-\frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i-y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}}(y_i-\bar{y}_{EXT})^2\right]/N_{EXT}}$	[34]
$Q_{F3}^2=1-\frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i-y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{TR}}(y_i-\bar{y}_{TR})^2\right]/N_{TR}}$	[34]
$\overline{R_m^2}=\frac{R_m^2(x,y)+R_m^2(y,x)}{2}$	[35]
$\Delta R_m^2=\left\lvert R_m^2(x,y)-R_m^2(y,x)\right\rvert$	
$CCC=\frac{2\sum(\mathrm{x}-\bar{\mathrm{x}})(\mathrm{y}-\bar{\mathrm{y}})}{\sum(\mathrm{x}-\bar{\mathrm{x}})^2+\sum(\mathrm{y}-\bar{\mathrm{y}})^2+\mathrm{n}(\bar{\mathrm{x}}-\bar{\mathrm{y}})^2}$	[36]
$IIC_{CLB}=r_{CLB}\frac{\min(^{-}MAE_{CLB},^{+}MAE_{CLB})}{\max(^{-}MAE_{CLB},^{+}MAE_{CLB})}$	[37]
$^{-}MAE_{CLB}=\frac{1}{-N}\sum_{k=1}^{-N}\lvert\Delta_k\rvert,\Delta_k0;^{-}N\ is\ the\ number\ of\ \Delta_k<0$	
$^{+}MAE_{CLB}=\frac{1}{+N}\sum_{k=1}^{+}\lvert\Delta_k\rvert,\Delta_k0;^{+}N\ is\ the\ number\ of\ \Delta_k\geq 0$	
$\Delta_k=observed_k-calculated_k$	

Table 5. The statistical characteristics of the models for Endpoint 2 were calculated with Eqs. 16–18 for (A) active training set, (P) passive training set, (C) calibration set, and (V) validation set (*n* is the number of conpounds in the corresponding set).

Eqn.		n	R^2	CCC	IIC	Q^2	Q_{F1}^2	Q_{F2}^2	Q_{F3}^2	$\overline{R_m^2}$	$RMSE$	MAE	F
16	A	37	0.7624	0.8652	0.8272	0.7382					0.183	0.133	112
	P	40	0.7357	0.6840	0.5881	0.6931					0.358	0.259	106
	C	37	0.8231	0.8987	0.9068	0.8018	0.8503	0.8056	0.9132	0.7455	0.142	0.098	163
	V	37	0.7486								0.209	0.152	
17	A	37	0.7645	0.8665	0.6662	0.7356					0.179	0.121	114
	P	38	0.7458	0.7604	0.8128	0.7100					0.320	0.230	106
	C	38	0.8079	0.8838	0.8985	0.7828	0.8035	0.8026	0.8488	0.6650	0.184	0.133	151
	V	38	0.8197								0.156	0.119	
18	A	37	0.6952	0.8202	0.7899	0.6409					0.238	0.167	80
	P	39	0.8051	0.8665	0.8934	0.7820					0.234	0.180	153
	C	38	0.7656	0.8742	0.8747	0.7360	0.7529	0.7513	0.8097	0.6704	0.204	0.156	118
	V	37	0.6694								0.269	0.210	

where

$$DCW_{SMILES}(T,N)=\sum_{k=1}^{NA}CW\left(S_k\right)+\sum_{k=1}^{NA-1}CW\left(SS_k\right)+\sum_{k=1}^{NA-2}CW\left(SSS_k\right) \tag{3}$$

$$DCW_{\text{Graph}}\left(T,N\right)=\sum_{k=1}^{NG}CW\left(EC1_k\right)+\sum_{k=1}^{NG}CW\left(EC2_k\right)+\sum_{k=1}^{NG}CW\left(EC3_k\right) \tag{4}$$

If SMILES = ABCD, the *S*, *SS*, and *SSS* can be represented as

S = (A, B, C, D)
SS = (AB, BC, CD)
SSS = (ABC, BCD)

The *EC*1, *EC*2, and *EC*3 are Morgan extended connectivity of first, second, and third order, respectively. The graph invariants are calculated with the adjacency matrix (Table 2).

The *T* is an integer to separate SMILES attributes into rare and non-rare. The non-rare SMILES are applied to build up the model. The rare SMILES are not applied to build up the model.

The *N* is the number of epochs of the optimization of the correlation weights.

Table 6. Promoters of increase/decrease for Endpoint 1 (split 1). N_a, N_p, and N_c are frequencies of a molecular feature in active training, passive training, and calibration sets, respectively. The equivalent promoters are indicated in bold.

SMILES attributes and graph invariants	CWs Run 1	CWs Run 2	CWs Run 3	N_a	N_p	N_c
Increase						
(...........	0.09942	0.42712	0.15207	37	40	37
1...........	0.62056	0.48386	0.71737	37	40	37
EC2-C...12..	0.35486	1.17923	1.59423	37	40	37
c...(.......	**1.05508**	**0.06376**	**1.10144**	**37**	**40**	**37**
c...c...c...	0.23446	1.22105	1.23382	37	40	37
c...c...2...	0.80818	0.18092	0.41245	36	40	33
c...c...(...	0.60409	0.73628	0.07209	35	35	35
C...........	0.23494	1.13657	1.14319	34	34	30
O...........	0.98334	0.98437	1.03588	32	38	30
EC2-C...16..	1.02590	0.72814	0.12875	30	31	28
EC2-C...18..	0.70135	0.63760	0.07998	30	32	35
1...(.......	1.05272	1.21128	1.32806	29	35	33
EC1-O...3...	**0.53006**	**0.43278**	**1.12533**	**29**	**35**	**25**
EC3-C...25..	1.33575	2.18057	0.78050	28	29	31
c...1...(...	1.13685	1.34514	1.20583	27	31	31
Decrease						
EC1-C...5...	**−0.73018**	**−0.59835**	**−0.67517**	**37**	**40**	**37**
c...1.......	−0.37863	−0.26498	−0.10006	37	40	37
c...1...c...	−0.03245	−0.27869	−0.18373	37	40	37
N...........	−0.28596	−0.07395	−0.08210	35	37	33
=...........	−0.60774	−0.93351	−0.77261	32	38	29
O...=.......	−0.08485	−1.20808	−0.74428	32	38	29
N...(.......	−0.15528	−0.95613	−1.93177	30	31	31
=...(.......	−0.32191	−0.64861	−1.14785	28	31	26
EC2-C...13..	−0.09063	−0.08354	−0.10325	28	32	30
O...=...(...	**−0.42882**	**−0.44401**	**−0.24622**	**28**	**31**	**26**
=...O...(...	−0.00103	−0.10204	−0.86064	27	31	26
O...(.......	**−0.47555**	**−0.04726**	**−0.99254**	**27**	**32**	**26**
EC2-C...17..	−0.46045	−1.04920	−0.66526	25	32	29
EC2-O...5...	−0.30910	−0.60111	−1.35192	22	30	22
EC1-C...4...	−0.18550	−0.99501	−0.68634	21	26	21

The S_k is a SMILES atom, i.e., one symbol of the SMILES line (e.g., '=', 'O') or a group of symbols that cannot be examined separately (e.g., 'Cu', '%11').

The $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ are the correlation weights of the above SMILES attributes.

2.3 The Monte Carlo optimization

Eqn. 2 needs the numerical data on the above correlation weights. The Monte Carlo optimization is a tool to calculate those correlation weights. Here, two target functions for the Monte Carlo optimization are examined:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (5)$$

$$TF \ = TF_0 + IIC_C \times 0.5 \quad (6)$$

The r_{AT} and r_{PT} are the correlation coefficient between the observed and predicted endpoints for the active training set and the passive training set, respectively.

The IIC_C is the index of ideality of correlation [29,30], and it is calculated with data on the calibration set as the follows:

$$IIC_C = r_C \frac{min\ ({}^{-}MAE_C, {}^{+}MAE_C)}{max\ ({}^{-}MAE_C, {}^{+}MAE_C)} \quad (7)$$

$$\min(x, y) = \begin{cases} x,\ if\ x < y \\ y, otherwise \end{cases} \quad (8)$$

$$\max(x, y) = \begin{cases} x,\ if\ x > y \\ y, otherwise \end{cases} \quad (9)$$

Table 7. Promoters of increase/decrease for Endpoint 2 (split 1). N_a, N_p, and N_c are frequencies of a molecular feature in active training, passive training, and calibration sets, respectively. The equivalent promoters are indicated in bold.

SMILES attributes and graph invariants	CWs Run 1	CWs Run 2	CWs Run 3	N_a	N_p	N_c
Increase						
2...........	0.22016	0.98780	0.08809	37	40	37
EC1-C...6...	0.24278	0.11086	0.18377	37	40	37
c...(.......	**1.03101**	**1.42056**	**0.08329**	**37**	**40**	**37**
c...2.......	1.22829	0.13557	0.33326	37	40	37
c...c.......	0.11189	0.16482	1.20026	37	40	37
c...2...c...	0.61136	1.12282	0.07782	36	40	33
N...........	1.04136	0.35614	0.23026	35	37	33
C...........	0.26780	0.04934	0.37202	34	34	30
EC2-C...11..	0.18299	0.99930	0.06830	34	36	32
C...(.......	1.01889	0.98424	1.05146	33	34	29
EC3-C...27..	0.47895	0.00866	1.40866	33	31	28
O...........	0.15846	0.24855	0.23234	32	38	30
EC2-C...16..	0.26331	1.04331	0.12793	30	31	28
N...(.......	0.35805	0.33794	0.79600	30	31	31
EC1-O...3...	**1.04302**	**0.41286**	**0.39454**	**29**	**35**	**25**
Decrease						
EC1-C...5...	**−0.56374**	**−0.35414**	**−0.06509**	**37**	**40**	**37**
c...c...1...	−0.15919	−0.05077	−0.11664	37	40	37
EC2-C...10..	−0.00196	−0.30027	−0.47412	32	32	30
O...=.......	**−0.18356**	**−0.20414**	**−0.51466**	**32**	**38**	**29**
EC2-C...18..	−0.31891	−0.26150	−0.04142	30	32	35
EC2-C...13..	−0.29896	−0.04766	−0.26914	28	32	30
O...=...(...	**−0.35727**	**−0.28264**	**−0.38639**	**28**	**31**	**26**
EC2-O...5...	−0.47640	−0.42986	−0.34099	22	30	22
c...C...(...	−0.46936	−0.18807	−0.15492	20	16	17
EC3-C...38..	−0.61296	−0.50030	−1.18697	18	19	19
N...c...1...	−1.29351	−0.16679	−0.53877	18	21	17
EC3-C...30..	−0.20519	−0.32076	−1.23300	17	18	18
c...n...1...	−0.58666	−0.30928	−0.76703	14	12	13
EC3-N...37..	−0.17496	−0.38205	−0.48940	13	13	20
EC1-N...5...	−1.39223	−0.41065	−0.50609	10	19	11

$$^{-}MAE_C = \frac{1}{^{-}N}\sum|\Delta_k|\,,\ ^{-}N\ is\ the\ number\ of\ \Delta_k < 0 \tag{10}$$

$$^{+}MAE_C = \frac{1}{^{+}N}\sum|\Delta_k|\,,\ ^{+}N\ is\ the\ number\ of\ \Delta_k \geq 0 \tag{11}$$

$$\Delta_k = observed_k\text{–}calculated_k \tag{12}$$

The observed and calculated are the corresponding values of the endpoint.

The Monte Carlo optimization that used the IIC_C is described in the literature [29,30].

3. Results and discussion

QSARs based on hybrid optimal descriptors were performed for 151 examined compounds (Table 1). Two endpoints were studied: (i) the first was the MIC against *S. aureus* ATCC 25923, and (ii) the second was the MIC against the drug-resistant clinical isolate of *S. aureus*.

The examined compounds were randomly split into an active training set (≈25%), passive training set (≈25%), calibration set (≈25%), and validation set (≈25%). Each of the above sets has a defined task. The active training set is used to build the model: molecular features extracted from quasi-SMILES of the active training set are involved in the process of Monte Carlo optimization aimed to provide correlation weights for the above features, which give maximal correlation coefficient between descriptors (the sum of the correlation weights) and endpoint on the active training set. The task of the passive training set is to check whether the model obtained for the active training set is satisfactory for quasi-SMILES that were not involved in the active training set. The calibration set should detect the start of the overtraining (overfitting). At the beginning of the optimization, the correlation coefficients between experimental values of the endpoint and the descriptor contemporaneously increase for all sets, but the correlation coefficient for the calibration set reaches a maximum (this is the start of the overtraining), and further optimization leads to decrease of the correlation coefficient for the calibration set. Optimization should be stopped when overtraining starts. After stopping the Monte Carlo optimization procedure, the validation set is used to assess the predictive potential of the obtained model.

3.1 MIC against S. aureus ATCC 25923 (i.e., Endpoint 1)

In order to check up the reproducibility of the CORAL [31] models, one should test several splits into the training sub-system (i.e., active training, passive training, and calibration sets) and validation sub-system. The described scheme for three random splits gives the following models:

(a) Split 1:

$$\log 1/c(S.a.) = 2.363(\pm 0.034) + 0.03939(\pm 0.00110) \times DCW(1,15) \quad (13)$$

(b) Split 2:

$$\log 1/c(S.a.) = 2.786(\pm 0.029) + 0.02335(\pm 0.00086) \times DCW(1,15) \quad (14)$$

(c) Split 3:

$$\log 1/c(S.a.) = 2.755(\pm 0.026) + 0.06817(\pm 0.00178) \times DCW(1,15) \quad (15)$$

Table 3 contains the statistical quality of these models. Table 4 (Ref. [32–37]) contains the statistical criteria of the predictive potential of a model.

3.2 MIC against drug-resistant clinical isolate of S. aureus (i.e., Endpoint 2)

For Endpoint 2, the described scheme for three random splits gives the following models:

(a) Split 1:

$$\log 1/c(S.a.,\ \text{isol.}) = 2.340(\pm 0.019) + 0.02649(\pm 0.00037) \times DCW(1,15) \quad (16)$$

(b) Split 2:

$$\log 1/c(S.a.,\ \text{isol.}) = 2.695(\pm 0.017) + 0.03511(\pm 0.00060) \times DCW(1,15) \quad (17)$$

(c) Split 3:

$$\log 1/c(S.a., \text{isol.}) = 1.575(\pm 0.056) + 0.08254(\pm 0.00227) \times DCW(1,15) \quad (18)$$

Table 5 contains the statistical quality of these models.

3.3 Mechanistic interpretation

An example of the technical details for Split 1, i.e., the calculated values for Endpoint 1 (Eqn. 13) and Endpoint 2 (Eqn. 16), and the corresponding correlation weights for the SMILES attributes and graph invariants, is presented in **Supplementary Material**.

Having numerical data on the correlation weights obtained in several runs of the described Monte Carlo method optimization, one can find molecular features extracted from SMILES or hydrogen suppressed graphs which have solely positive correlation weights. These should be interpreted as promoters of increase for the corresponding endpoint. If a molecular feature has a stable negative correlation weight in several runs of the optimization, it should be interpreted as a promoter of decrease for an endpoint. Table 6 contains a collection of the above promoters for Endpoint 1 and Table 7 contains similar data for Endpoint 2, respectively. One can see (Tables 6,7), that Endpoint 1 and Endpoint 2 have five equivalent promoters (indicated by bold). In other words, these endpoints are far from to be identical ones.

3.4 The statistical quality of the models

The statistical quality of the models for Endpoint 1 and Endpoint 2 is quite good (Tables 3,5). Reproducibility of the results for both endpoints is observed. However, the predictive potential observed for three random splits is not identical. For both endpoints, the best predictive potential is observed in the case of split 2. The statistical quality of the models for Endpoint 2 is slightly better than that of the models for Endpoint 1. The models suggested here are traditional, that is, multi-targets approach [6–8], and ADMET [9] are not used here. However, in principle, the approach can be available for the corresponding analyses in the future.

4. Conclusions

The application of hybrid optimal descriptors has been proposed and tested to develop a predictive model for 151 structurally diverse compounds with antibacterial activity against *S. aureus* ATCC 25923 (Endpoint 1) or the drug-resistant clinical isolate of *S. aureus* (Endpoint 2) has been proposed and tested. The predictive potential of these models has been checked with three random splits into the training, passive training, calibration, and validation sets. The proposed models give satisfactory predictive models for both endpoints examined, but it has been found that splitting has an apparent influence on the statistical quality of these models, and the best predictive potential is observed in the case of split 2 for both endpoints. The statistical quality of the models is slightly better for the Endpoint 2 models. The results of the study show the possibility of SMILES-based QSAR in the evaluation of the antibacterial activity of structurally diverse compounds.

Author contributions

KN and AT designed the study and participated in writing the manuscript. AT performed the study, software, and calculation. IY provided data and participated in writing the manuscript. KN handled the funding acquisition. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research was funded by Charles University, the project Cooperatio.

Conflict of interest

The authors declare no conflict of interest.

Supplementary material

Supplemental data contains the technical details for Split 1: the calculated values for Endpoint 1 (Eqn. 13) and Endpoint 2 (Eqn. 16), and the corresponding correlation weights for the SMILES attributes and graph invariants. Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2704112.

References

[1] González-Díaz H, Prado-Prado FJ, Santana L, Uriarte E. Unify QSAR approach to antimicrobials. Part 1: Predicting antifungal activity against different species. Bioorganic & Medicinal Chemistry. 2006; 14: 5973–5980.

[2] Liu Q, Zhou H, Liu L, Chen X, Zhu R, Cao Z. Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: an in-silico study. BMC Bioinformatics. 2011; 12: 294.

[3] Rosenbaum L, Dörr A, Bauer MR, Boeckler FM, Zell A. Inferring multi-target QSAR models with taxonomy-based multi-task learning. Journal of Cheminformatics. 2013; 5: 33.

[4] Speck-Planche A, Cordeiro M. Simultaneous Modeling of Antimycobacterial Activities and ADMET Profiles: a Chemoinformatic Approach to Medicinal Chemistry. Current Topics in Medicinal Chemistry. 2013; 13: 1656–1665.

[5] Nicolotti O, Giangreco I, Miscioscia TF, Carotti A. Improving Quantitative Structure-Activity Relationships through Multiobjective Optimization. Journal of Chemical Information and Modeling. 2009; 49: 2290–2302.

[6] Nicolaou CA, Brown N. Multi-objective optimization methods in drug design. Drug Discovery Today. Technologies. 2013; 10: e427–e435.

[7] Sánchez-Rodríguez A, Pérez-Castillo Y, Schürer SC, Nicolotti O, Mangiatordi GF, Borges F, *et al*. From flamingo dance to (desirable) drug discovery: a nature-inspired approach. Drug Discovery Today. 2017; 22: 1489–1502.

[8] Cummins DJ, Bell MA. Integrating everything: the Molecule Selection Toolkit, a System for Compound Prioritization in Drug Discovery. Journal of Medicinal Chemistry. 2016; 59: 6999–7010.

[9] Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H. Chemical predictive modelling to improve compound quality. Nature Reviews. Drug Discovery. 2013; 12: 948–962.

[10] Toropov AA, Toropova AP. QSPR/QSAR: State-of-art, weirdness, the future. Molecules. 2020; 25: 1292.

[11] Jernigan JA, Hatfield KM, Wolford H, Nelson RE, Olubajo B, Reddy SC, *et al*. Multidrug-Resistant Bacterial Infections in U.S. Hospitalized Patients, 2012–2017. New England Journal of Medicine. 2020; 382: 1309–1319.

[12] Dayan GH, Mohamed N, Scully IL, Cooper D, Begier E, Eiden J, *et al*. *Staphylococcus aureus*: the current state of disease, pathophysiology and strategies for prevention. Expert Review of Vaccines. 2016; 15: 1373–1392.

[13] Horino T, Hori S. Metastatic infection during *Staphylococcus aureus* bacteremia. Journal of Infection and Chemotherapy. 2020; 26: 162–169.

[14] Mohammed YHE, Manukumar HM, Rakesh KP, Karthik CS, Mallu P, Qin H. Vision for medicine: *Staphylococcus aureus* biofilm war and unlocking key's for anti-biofilm drug development. Microbial Pathogenesis. 2018; 123: 339–347.

[15] Das S, Dasgupta A, Chopra S. Drug repurposing: a new front in the war against *Staphylococcus aureus*. Future Microbiology. 2016; 11: 1091–1099.

[16] Ertan T, Yildiz I, Ozkan S, Temiz-Arpaci O, Kaynak F, Yalcin I, *et al*. Synthesis and biological evaluation of new *N*-(2-hydroxy-4(or 5)-nitro/aminophenyl)benzamides and phenylacetamides as antimicrobial agents. Bioorganic & Medicinal Chemistry. 2007; 15: 2032–2044.

[17] Yildiz I, Ertan T, Bolelli K, Temiz-Arpaci O, Yalcin I, Aki E. QSAR and pharmacophore analysis on amides against drug-resistant *S. aureus*. SAR and QSAR in Environmental Research. 2008; 19: 101–113.

[18] Arisoy M, Temiz-Arpaci O, Yildiz I, Kaynak-Onurdag F, Aki E, Yalcin I, *et al*. Synthesis, antimicrobial activity and QSAR studies of 2,5-disubstituted benzoxazoles. SAR and QSAR in Environmental Research. 2008; 19: 589–612.

[19] Ertan T, Yildiz I, Tekiner-Gulbas B, Bolelli K, Temiz-Arpaci O, Ozkan S, *et al*. Synthesis, biological evaluation and 2D-QSAR analysis of benzoxazoles as antimicrobial agents. European Journal of Medicinal Chemistry. 2009; 44: 501–510.

[20] Bolelli K, Yalcin I, Ertan-Bolelli T, Özgen S, Kaynak-Onurdag F, Yildiz I, *et al*. Synthesis of novel 2-[4-(4-substitutedbenzamido/phenylacetamido)phenyl]benzothiazoles as antimicrobial agents. Medicinal Chemistry Research. 2012; 21: 3818–3825.

[21] Yilmaz S, Yalcin I, Kaynak-Onurdag F, Yildiz I, Aki E. Synthesis and *In vitro* Antimicrobial Activity of Novel 2-(4-(Substituted-carboxamido)benzyl / phenyl)benzothiazoles. Croatica Chemica Acta. 2013; 86: 223–231.

[22] Ertan-Bolelli T, Yildiz I, Ozgen-Ozgacar S. Synthesis, molecular docking and antimicrobial evaluation of novel benzoxazole derivatives. Medicinal Chemistry Research. 2016; 25: 553–567.

[23] Acar C, Yalçın G, Ertan-Bolelli T, Kaynak Onurdağ F, Ökten S, Şener F, *et al*. Synthesis and molecular docking studies of some

novel antimicrobial benzamides. Bioorganic Chemistry. 2020; 94: 103368.

[24] Nesměrák K, Toropov AA, Toropova AP, Ertan-Bolelli T, Yildiz I. QSAR of antimycobacterial activity of benzoxazoles by optimal SMILES-based descriptors. Medicinal Chemistry Research. 2017; 26: 3203–3208.

[25] Toropov AA, Toropova AP, Benfenati E, Nicolotti O, Carotti A, Nesmerak K, *et al*. QSPR/QSAR analyses by means of the CORAL software: results, challenges, perspectives. In Roy K (ed.) Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment (pp. 560−585). 1st edn. Hersey, PA: Medical Information Science Reference. 2015.

[26] Toropova AP, Toropov AA. CORAL: Monte Carlo Method to Predict Endpoints for Medical Chemistry. Mini Reviews in Medicinal Chemistry. 2018; 18: 382–391.

[27] Lotfi S, Ahmadi S, Zohrabi P. QSAR modeling of toxicities of ionic liquids toward *Staphylococcus aureus* using SMILES and graph invariants. Structural Chemistry. 2020; 31: 2257–2270.

[28] ACD/ChemSketch. 2021. Available at: www.acdlabs.com (Accessed: 8 December 2021).

[29] Toropova AP, Toropov AA. The index of ideality of correlation: a criterion of predictability of QSAR models for skin permeability? The Science of the Total Environment. 2017; 586: 466–472.

[30] Toropov AA, Carbó-Dorca R, Toropova AP. Index of Ideality of Correlation: new possibilities to validate QSAR: a case study. Structural Chemistry. 2018; 29: 33–38.

[31] CORAL. 2020. Available at: http://www.insilico.eu/coral (Accessed: 8 December 2021).

[32] Hemmateenejad B, Javidnia K, Miri R, Elyasi M. Quantitative structure–retention relationship study of analgesic drugs by application of combined data splitting-feature selection strategy and genetic algorithm-partial least square. Journal of the Iranian Chemical Society. 2012; 9: 53–60.

[33] Shayanfar A, Shayanfar S. Is regression through origin useful in external validation of QSAR models? European Journal of Pharmaceutical Sciences. 2014; 59: 31–35.

[34] Chirico N, Gramatica P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Zsing the Concordance Correlation Coefficient. Journal of Chemical Information and Modeling. 2011; 51: 2320–2335.

[35] Roy K, Kar S. The ${r_m}^2$ metrics and regression through origin approach: reliable and useful validation tools for predictive QSAR models (Commentary on ‘Is regression through origin useful in external validation of QSAR models?’). European Journal of Pharmaceutical Sciences. 2014; 62: 111–114.

[36] Lin LI. Assay Validation Using the Concordance Correlation Coefficient. Biometrics. 1992; 48: 599–604.

[37] Toropov AA, Toropova AP. The index of ideality of correlation: a criterion of predictive potential of QSPR/QSAR models? Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2017; 819: 31–37.