

SCRAMBLE’N’GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models

Piotr F. J. Lipiński¹  · Przemysław Szurmak²

Received: 25 January 2017 / Accepted: 26 May 2017 / Published online: 5 June 2017
© The Author(s) 2017. This article is an open access publication

Abstract A common practice in modern QSAR modelling is to derive models by variable selection methods working on large descriptor pools. As pointed out previously, this is intrinsically burdened with the risk of finding random correlations. Therefore it is desirable to perform tests showing the performance of models built on random data. In this contribution, we introduce a simple and freely available software tool SCRAMBLE’N’GAMBLE that is aimed at facilitating data preparation for y -randomization and pseudo-descriptors tests. Then, four close-to-real-world modelling situations are analysed. The tests indicate what the quality of obtained QSAR models is like in comparison to chance models derived from random data. The non-randomness is not the only requirement for a good QSAR model, however, it is a good practice to consider it together with internal statistical parameters and possible physical interpretations of a model.

Keywords y -Randomization · Chance correlations · QSAR · QSAR validation

Electronic supplementary material The online version of this article (doi:10.1007/s11696-017-0215-7) contains supplementary material, which is available to authorized users.

✉ Piotr F. J. Lipiński
plipin@icm.edu.pl

¹ Department of Neuropeptides, Mossakowski Medical Research Centre Polish Academy of Sciences, 02-106 Warsaw, Poland

² ChemPharmSoft, 01-926 Warsaw, Poland

Introduction

Quantitative Structure-Activity Relationship (QSAR) modelling is an important field of research in current medicinal chemistry. QSAR models relate the structure of chemical compounds to their biological activities:

$$\text{activity} = f(\text{structure}).$$

The aim of building such models is to explain and/or to predict the activity of a group of compounds and thus to facilitate and direct search for new active substances.

In QSAR, the structure of a chemical compound is represented mathematically by molecular descriptors. These can be based on physicochemical properties measured experimentally (e.g. partition coefficient LogP), quantities calculated by quantum chemistry methods (e.g. HOMO/LUMO energies) (Karelson et al. 1996) or be derived from other theoretical bases (e.g. chemical graph theory, (Balaban 1985; Helguera et al. 2008) theory of quantitative chirality (Ostrowski et al. 2012; Jamróz et al. 2012 etc.). The number of currently available descriptors is enormous (Dearden 2016). There are several applications designed specifically at their calculation [for example DRAGON by Talete Srl that computes ca. 5000 descriptors (Talete Srl 2010)] and such a functionality is present in probably all drug design and discovery suites like Accelrys Discovery Studio (Accelrys Software Inc. 2009), Schrödinger Suite (2017), molecular operating environment (Chemical Computing Group ULC 2017) to mention only a few.

In a typical situation a researcher has at his or her disposal a scarce number of compounds with determined activity (like 20 to several dozen) and an alluring plenitude of molecular descriptors (hundreds or thousands) to be used for constructing QSAR equations. This makes the danger of overfitting data a very likely one.

The common statistical parameters, like coefficient of determination, standard deviation, significance etc. are not able to discern ‘good’ models from overfitted ones (Rücker et al. 2007). This cannot be also done by any kind of internal validation procedures, like leave-one-out, leave-many-out etc. An ultimate test of validity and utility of a given QSAR model is always the external validation on an independent, large enough, properly designed set of new derivatives (Gramatica 2007). This is, however, rarely possible due to the lack of resources and/or time. In such circumstances, perhaps the only affordable way to see if studied QSAR models work better than the pure chance is to simulate the ‘predictive power’ of the pure chance. Two tests could be of help here: *y*-scrambling and pseudo-descriptors test (Clark et al. 2001; Rücker et al. 2007).

The *y*-scrambling (*y*-randomization, response randomization) is a form of a permutation test, where the values of the response variable (*y*) are randomly ascribed (scrambled) to different compounds, while the descriptors values (*x*’s) are left intact. Scrambled data are then used for training QSAR models. In the pseudo-descriptors test, the descriptors (*x*’s) are replaced by random numbers (pseudo-descriptors) that are also subsequently used to train QSAR equations.

Both tests are run over several to several dozen times, and from each run best coefficient of determination r^2 , leave-one-out cross-validation correlation coefficient q^2 and perhaps other adequate statistical parameters are collected. The mean highest r^2 (mhr^2) and q^2 (mhq^2) along with their standard deviations (SD) are calculated. This allows to assess the ‘predictive power’ of the pure chance, and the truly good models should have their r^2 and q^2 significantly better than this.

Unfortunately, these simple tests are very often not included into QSAR studies. One of the reasons, apart from their time-consuming character, might be in a difficulty in obtaining random data for simulations. Not every researcher is enough computer proficient to generate them on his own, and not everyone has access to good statistical software that could accomplish this without much trouble. The software, in majority, if not in all cases, is also not suited to working with common formats of chemical table files like SDF (Dalby et al. 1992) that are usually accepted by QSAR modelling software. The need for manual operations on numerous, large spreadsheets of numbers and chemical files can be an actual obstacle, and discouraged researchers omit these insightful tests.

In order to facilitate data preparation for the tests, a simple and free software tool SCRAMBLE’N’GAMBLE is proposed. It is a stand-alone Java application with both graphic user interface as well as a command-line manageability. SCRAMBLE’N’GAMBLE reads in comma-separated files (csv) and chemical table files by MDL (sdf)

containing descriptors and activity data. It can perform *y*-scrambling as well as generate pseudo-descriptors given number of times and output the results into a csv file, but also directly into a sdf file immediately usable in most QSAR programs. SCRAMBLE’N’GAMBLE is available free of charge at: <http://www.drugdesign.pl/scramble-n-gamble/>.

In order to demonstrate the importance of simulating random chance performance along with building QSAR models, let us expose the following Cases: I. classical QSAR modelling (descriptors based on 2D structures) of steroids’ affinity for the sex-hormone-binding globulin, II. classical QSAR modelling (descriptors based on 2D and 3D structures) of steroids’ affinity for the corticosteroid-binding globulin, III. Fujita-Ban QSAR modelling of the effective dose of some fentanyl in the mouse hot plate test and IV. a classification model for discerning glucocorticoid receptor binders and non-binders.

Experimental

Molecules, activity data, descriptor calculation and modelling procedure

In all Cases, a general workflow was as follows. First, molecules with activity data for a given molecular target were collected and divided into a training set and a test set. Second, molecular descriptors were calculated. Constant and near-constant descriptors were deleted from the pool, and further reduction was done by checking intercorrelations between descriptors. In pairs where the coefficient of correlation was larger than 0.90, one of the descriptors was randomly excluded. Third, QSAR models were trained. Fourth, random data for *y*-scrambling and pseudo-descriptors test were generated using SCRAMBLE’N’GAMBLE and the tests were performed by training QSAR models in the same way as the ones based on true data were trained. Fifth, the performance of the latter was checked on test sets.

Details of the workflow for singular Cases are given in Table 1.

Evaluation of regression models

For regression models in Cases I and II, standard statistical metrics were applied. These are:

- r^2 coefficient of determination in the training set,
- q^2 cross-validated coefficient of determination in the training set (internal validation, leave-one-out procedure)
- R^2 coefficient of determination in the test set (external validation).

Table 1 Details of the workflow for singular cases

Case	I	II	III	IV
QSAR task	Linear equation of up to 3 variables, descriptors based on 2D structure	Linear equation of up to 3 variables, descriptors based on 3D structure	Linear equation, Fujita-Ban model (Fujita and Ban 1971)	Classification model
Molecules	Various steroids (Fig. 2)	Various steroids (Fig. 2)	Fentanyl derivatives (3-methyl-1,4-disubstituted piperidines) (Table SI-1 in Electronic supporting material)	Various compounds (Tables SI-5 to SI-8)
Dependent variable	Logarithm of binding affinity to sex-hormone-binding globulin (SHBG) (Fig. 2)	Logarithm of binding affinity to corticosteroid-binding globulin (CBG) (Fig. 2)	Effective dose ED ₅₀ in mouse hot plate test (analgesic activity test) (Table SI-1)	Whether a molecule binds or does not bind the glucocorticoid receptor (Tables SI-5 to SI-8)
Training set	21 steroids of the benchmark Cramer data set (Cramer et al. 1988; Coats 1998) (S1-S21 in Fig. 2)	21 steroids of the benchmark Cramer data set (Cramer et al. 1988; Coats 1998), chosen by clustering from 31 molecules: S2, S3, S4, S5, S7, S8, S9, S10, S11, S14, S15, S18, S19, S20, S22, S24, S25, S26, S28, S29, S31 in Fig. 2	36 active derivatives (Lalinde et al. 1990) (Table SI-1)	100 active molecules and 3600 decoys randomly chosen from Directory of useful decoys, enhanced (DUD-E). (Mysinger et al. 2012) (Tables SI-5 and Table SI-7)
Test set	Up to 12 molecules (within the applicability domains of the models found) taken from the extended benchmark steroid data set (Cherkasov et al. 2008) (S32-S44 in Fig. 2)	10 steroids of the benchmark Cramer data set (Cramer et al. 1988; Coats 1998), chosen by clustering from 31 molecules: S1, S6, S12, S13, S16, S17, S21, S23, S27, S30 in Fig. 2	10 inactive derivatives (Lalinde et al. 1990) (Table SI-1)	Other 50 active molecules and 1800 decoys randomly chosen from Directory of Useful Decoys, Enhanced (DUD-E). (Mysinger et al. 2012) (Tables SI-6 and Table SI-8)
Calculation of molecular descriptors	2D descriptors in DRAGON 6 (Talete Srl 2010)	Structure optimization (B3LYP/6-31G*) in Gaussian 09 (Frisch et al. 2009) (full citation in Electronic supporting material); vibrational frequencies to check imaginary frequencies; atomic charges (q-descriptors) calculated using the CHELPG algorithm (Breneman and Wiberg 1990); 2D and 3D molecular descriptors in Accelrys Discovery Studio (Accelrys Software Inc. 2009); Sinister-Rectus Chirality Measures (^{SR} CMs) (Ostrowski et al. 2012; Jamróz et al. 2012; Ostrowski et al. 2013) and Continuous Chirality Measures (CCMs) (Zabrodsky and Avnir 1995), ^{SR} CMs calculated using the CHIMEA software (Jamróz 2010) available at http://www.smmg.pl , while CCMs using a web page from Hebrew University of Jerusalem (Zayit et al. 2011)	Indicator variables	2D descriptors in DRAGON 6 (Talete Srl 2010)
Descriptors calculated	3764	127	9	3764

Table 1 continued

Case	I	II	III	IV
Descriptors included in training the models	89	49	9	1090
Training QSAR model	Genetic function approximation (GFA) algorithm (Rogers and Hopfinger 1994) in Discovery Studio, GFA settings were as default	Genetic function approximation (GFA) algorithm (Rogers and Hopfinger 1994) in Discovery Studio, GFA settings were as default	Linear regression routine in Microsoft Excel 2016	In-house script based on a Python scikit-learn library for machine learning. (http://scikit-learn.org/); the depth of the tree was set to be maximally 3
Number of random data sets generated	300 (y-scrambling, pseudo-descriptors with original distributions, pseudo-descriptors with uniform distributions)	300 (y-scrambling, pseudo-descriptors with original distributions, pseudo-descriptors with uniform distributions)	300 (y-scrambling, pseudo-descriptors with original distributions, pseudo-descriptors with uniform distributions)	25 (pseudo-descriptors with original distributions, pseudo-descriptors with uniform distributions)

The r^2 and q^2 values were compared to mean highest r^2 (mhr²) and q^2 (mhq²) from y-scrambling and pseudo-descriptors tests in order to check whether the models perform better than chance models.

Additionally, ${}^cR_p^2$ and $r_{m(\text{test})}^2$ parameters were applied calculated as proposed by the Roy group (Pratim Roy et al. 2009; Mitra et al. 2010). Both these metrics should be greater than 0.5 for an acceptable model. The fulfilment of this criterion with regard to $r_{m(\text{test})}^2$ parameter ensures that a model predicts the exact values of the response data. High ${}^cR_p^2$ values allow to consider a model to be robust and not just the outcome of a chance correlation.

For the models in Case II, additional parameters were checked. First, the internal cross-validation was performed also in leave-three-out procedure, giving $q_{(L3O)}^2$ —a cross-validated coefficient of determination in the training set (leave-three-out). Furthermore, another type of randomization experiment was performed (Wold et al. 1998). Here y-scrambled data (25 runs) were used to refit the Case II models. The obtained r^2 and q^2 values were then plotted against the correlation coefficients of original y and permuted y data. The resulting intercepts (R_{int}^2 and Q_{int}^2) are expected to be below 0.4 and 0.05 respectively for valid models.

Evaluation of decision trees

For all decision trees (Case IV) the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) was collected. The following metrics were used for assessment of the decision trees: accuracy (ACC), precision (PREC), sensitivity (SENS), specificity (SPEC), fall-out (FALL) and F1-score (F1). They are given by the expressions:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{PREC} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

$$\text{FALL} = 1 - \text{SPEC} \quad (5)$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = 2 \times \frac{\text{PREC} \times \text{SENS}}{\text{PREC} + \text{SENS}} \quad (6)$$

Results and discussion

Software description

SCRAMBLE'N'GAMBLE is a fast and user-friendly software for generation of random data for the purposes of QSAR model validation. The program can read and output both comma-separated files (csv) as well as chemical table files by MDL (sdf) containing molecular descriptors and activity data. Upon selecting which fields should be scrambled or replaced with random data (pseudo-descriptors), the user is able to obtain a required number of randomized data sets in csv or sdf files. The latter are most often accepted by QSAR modelling software. SCRAMBLE'N'GAMBLE may be run in a graphic user interface mode (Fig. 1), but it is also manageable in the command-line mode.

The generation of random (or to be said more precisely: pseudo-random) numbers is achieved using Mersenne Twister 19937 generator (Matsumoto and Nishimura 1998)

The screenshot displays the SCRAMBLE'N'GAMBLE software interface. The top window is titled 'ScrambleNGamble' and shows a 'New Task' configuration for 'pseud-descriptors'. The 'Number of runs' is set to 200. The 'Pseudo-descriptors generation' option is selected. The 'Independent variables' list includes q1 through q7, and the 'Dependent variables' list includes q14 through q1, ccmH, srcm0, srcm2, ALogP, ALogP_MR, Molecular, and PolarSurfaceArea. The 'Save as SDF file' option is selected. The bottom window is a 'CSV File Explorer' showing a table with 14 columns (q1 to q14) and 21 rows of steroid names and their corresponding values.

Name	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14
5-androsten...	-0.072193	-0.078468	0.325587	0.077711	-0.247704	-0.255886	0.129632	0.153224	-0.056169	0.450014	0.07862	-0.160247	0.305252	0.059761
17a-hydrox...	-0.256736	0.08377	0.338727	-0.062207	-0.13321	-0.261427	0.126133	-0.06214	0.145581	0.36252	0.031379	-0.111257	0.119077	0.230711
estrone	-0.059116	-0.321741	0.38463	-0.294651	0.067458	-0.019057	-0.012027	0.191437	0.057434	-0.104938	0.058367	-0.157184	0.280334	-0.067753
estril	-0.057535	-0.312827	0.377769	-0.269406	0.04124	-0.019437	0.030031	0.282214	0.051307	-0.098021	0.043374	-0.031243	0.174082	-0.007451
16a-methyl...	0.050951	-0.232938	0.628226	-0.42222	-0.022454	0.124222	-0.123488	0.180168	-0.053364	0.32083	0.011204	-0.098976	0.27284	0.131567
2a-methyl-9...	-0.024297	0.099945	0.525642	-0.388265	-0.009765	0.106186	-0.087398	0.150177	0.209177	0.246806	0.288316	-0.21047	0.335633	-0.002982
4-androsten...	0.092574	-0.254699	0.62844	-0.400945	-0.035136	0.112155	-0.116549	0.262993	-0.181416	0.321538	0.169947	-0.221295	0.257609	-0.108327
19-nortesto...	0.070448	-0.184587	0.610616	-0.398556	0.029184	0.036401	-0.025403	0.198583	0.081312	0.053466	-0.031226	-0.128859	0.463475	-0.047415
progesterone	0.058656	-0.244806	0.628417	-0.401635	-0.044702	0.115415	-0.11024	0.130337	-0.054753	0.33541	0.050262	-0.105642	0.281478	0.089481
4-pregnene...	-0.038312	-0.19722	0.61744	-0.4042	-0.062135	0.110083	-0.098729	0.218483	-0.2679	0.474518	0.530896	-0.248003	0.326266	0.014362
cortisone	-0.000705	-0.218414	0.624778	-0.412701	-0.032999	0.094062	-0.075647	0.185897	-0.207467	0.40617	0.483575	-0.253964	0.414352	-0.034867
epicorticost...	0.052246	-0.242969	0.637475	-0.439578	0.000946	0.090798	-0.062266	0.139561	-0.142238	0.3324	0.331438	-0.217617	0.298864	0.096362
prednisolone	-0.13744	-0.317496	0.657359	-0.39436	0.016701	0.033576	-0.074294	0.203447	-0.04952	0.397338	0.41255	-0.220462	0.154293	0.092858
11-deoxyco...	0.088251	-0.22776	0.625223	-0.411105	-0.038581	0.142695	-0.123404	0.146198	-0.099849	0.320607	0.100829	-0.189694	0.27137	0.110236
cortisol	0.074165	-0.211267	0.606705	-0.397111	-0.034792	0.084174	-0.105062	0.164859	-0.041245	0.329936	0.325827	-0.250911	0.350643	-0.020782
11-deoxyco...	0.070342	-0.236335	0.630591	-0.425	-0.015537	0.11418	-0.084127	0.189041	-0.110666	0.31596	0.037684	-0.158456	0.348845	0.019194
19-norproge...	0.070746	-0.185387	0.613042	-0.402851	0.047631	0.012399	-0.020196	0.111403	0.112853	0.039605	-0.064252	-0.066798	0.284319	0.02743
androsterone	0.077753	-0.175041	0.351193	-0.166645	0.251739	-0.156032	0.009129	0.114595	0.01216	0.250984	0.099045	-0.205806	0.267961	-0.029282
androstane...	-0.147361	-0.042278	0.403789	-0.269948	0.207702	-0.085588	-0.051168	0.086818	0.021297	0.340522	0.02293	-0.13382	0.253829	0.110891
dehydroepia...	-0.058888	-0.135699	0.363562	-0.108305	-0.103958	-0.297189	0.130944	0.150661	-0.066212	0.317575	0.127958	-0.196056	0.274221	-0.089993
17a-hydrox...	0.088928	-0.224418	0.616399	-0.406646	-0.018805	0.094565	-0.084912	0.111388	-0.018161	0.280484	0.07382	-0.158049	0.125353	0.173015

Fig. 1 General view of SCRAMBLE'N'GAMBLE interface

implemented in UncommonMaths Java library (Dyer 2006). The generator has been shown to generate high quality random numbers and pass many statistical tests for randomness. It is possible to select a distribution from which random numbers will be generated: uniform, normal, binomial, Poisson or exponential. The user may also want to keep original distributions of variables, and in such case the program will perform *x*-scrambling. SCRAMBLE'N'GAMBLE is available free of charge at: <http://www.drugdesign.pl/scramble-n-gamble/>.

The examples and importance of performing random data tests in QSAR validation are provided by considering four close-to-real-world modelling situations.

Case I

Sex-hormone-binding globulin (SHBG) is a transport glycoprotein produced in all vertebrates except for birds.

SHBG binds preferentially sex hormones (androgens and oestrogens) in the bloodstream and in this way it has impact on the concentration of their free, supposedly biologically active, fractions. Its role in various endocrine disorders is well described (Anderson 1974; Cunningham et al. 1983; Key et al. 2002; Hammond 2011; Caldwell and Jirikowski 2014). Environmental toxicology points also to the importance of SHBG in the endocrine disruption in men and animals caused by exogenous substances (Wilson et al. 2007; Saxena et al. 2014; Hong et al. 2015).

In QSAR studies, the Cramer data set of 21 steroids (Fig. 2) binding to SHBG became a benchmark set for validating novel QSAR methodologies or descriptors (Cramer et al. 1988; Coats 1998). Therefore, it is a good point for illustrating the danger of chance correlations.

In our study, we trained QSAR models of up to 3 independent variables, using 89 2D molecular descriptors. The top 10 models are presented in Table 2. Their

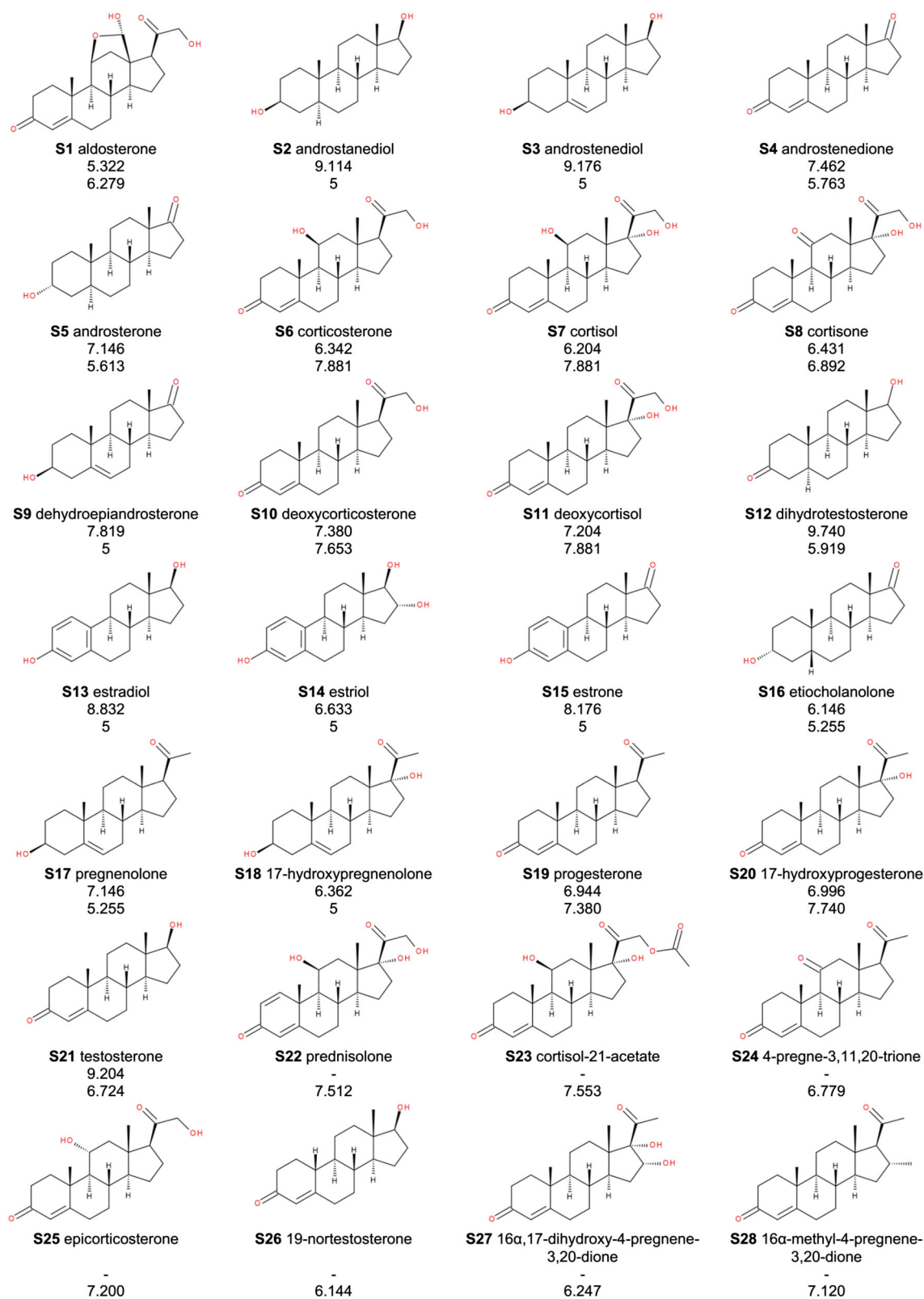


Fig. 2 Steroid molecules used in Cases I and II. The figures under names are binding affinities to corticosteroid-binding globulin (*upper figures*, Case II) and sex-hormone-binding globulin (*lower figures*, Case I)

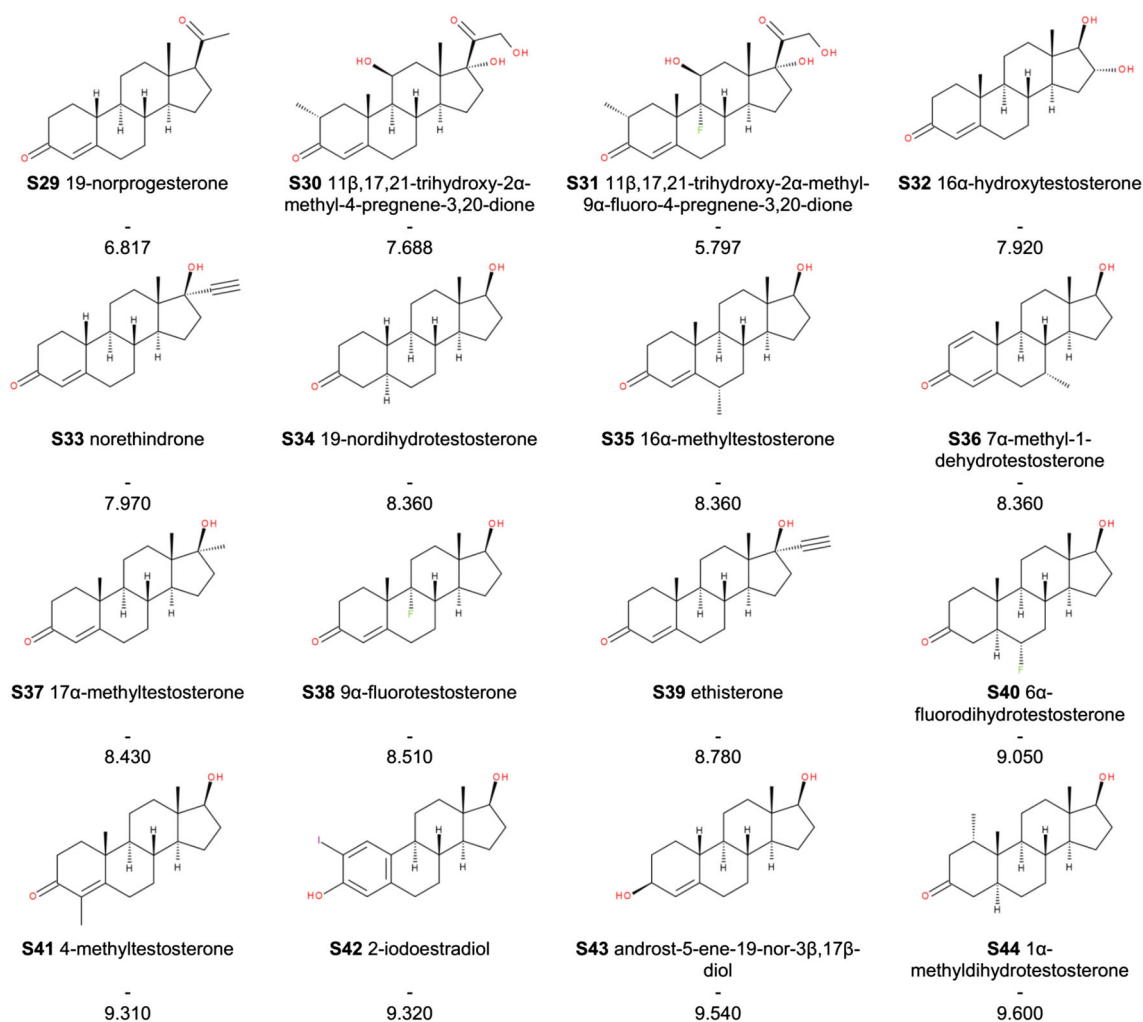


Fig. 2 continued

statistical parameters are not the best ones, but they could be perceived as acceptable by some QSAR modellers ($r^2 = 0.762\text{--}0.811$, $q^2 = 0.613\text{--}0.706$). On the other hand, the equations are physically uninterpretable as almost all descriptors (except for $P_VSA_s_4$ and $NssCH$) cannot be translated (at least without great effort) into the language of atoms, functional groups or other chemical structures. Still, many authors ‘interpret’ similar models just by providing brief descriptions of how the descriptors are calculated and conclude that the equation(s) could serve for screening chemical libraries in search of new active compounds.

The moderately optimistic r^2 and q^2 become not optimistic at all if one looks at the outcomes of the models trained on y -scrambled activity data or those trained on pseudo-descriptors (Table 3). It turns out that none of the obtained ‘real’ models is better than the 99th percentile (+2.3 SD) of the models found in the y -randomization or pseudo-descriptors tests ($mhr^2 + 2.3$ SD of models trained on pseudo-descriptors is as high as 0.825). Further, external

validation on several ligands (6–12, depending on the applicability domain of a given model, (Tables 2 and SI-2) extracted from the extended steroid set (Cherkasov et al. 2008) yields very poor results, with the coefficient of determination in the test set (R^2) not higher than 0.270.

The models in Table 2 are thus: internally quite good but uninterpretable and not better than random models. As such, they could be expected to have poor predictive power, what is then shown in external validation (Tables 2 and SI-2).

Case II

In the second of the studied cases, we used the same Cramer steroid data set (Cramer et al. 1988; Coats 1998), but this time the target property was binding affinity for the corticosteroid-binding globulin (CBG). CBG is another steroid transporting protein, but contrary to SHBG, it binds preferentially corticosteroids and progestogens, while

Table 2 Top 10 QSAR models obtained by the GFA procedure (Case I)

No	$K_{\text{aff}} =$	r^2	q^2	R^2	${}^c R_p^2$	$r_{m(\text{test})}^2$
1	$-69.185 - 1.103 \times X5v + 38.155 \times \text{MATS5m} + 44.524 \times \text{SpMin3_Bh}(m)$	0.811	0.690	0.051, $n = 10$	0.339	0.023
2	$-60.380 + 28.860 \times \text{MATS5m} - 45.301 \times \text{MATS3v} + 34.208 \times \text{SpMin3_Bh}(m)$	0.792	0.706	0.015, $n = 9$	0.312	<0.001
3	$-89.817 + 160.850 \times X2A + 36.696 \times \text{MATS5m} + 30.049 \times \text{SpMin3_Bh}(m)$	0.784	0.701	0.027, $n = 12$	0.300	0.017
4	$22.870 + 127.78 \times \text{VE2_Dt} - 5.549 \times \text{SpDiam_AEA}(dm) - 0.559 \times \text{NsssCH}$	0.782	0.673	0.322, $n = 8$	0.297	0.289
5	$14.787 - 2.262 \times \text{IDDE} + 157.890 \times \text{VE2_Dt} - 3.801 \times \text{SpDiam_AEA}(dm)$	0.777	0.669	0.030, $n = 8$	0.290	0.024
6	$-25.180 + 30.414 \times \text{MATS5m} + 23.871 \times \text{SpMin3_Bh}(m) - 1.935 \times \text{SpDiam_AEA}(dm)$	0.771	0.679	0.146, $n = 9$	0.280	0.097
7	$2.426 + 186.700 \times \text{VE2_Dt} + 0.070 \times \text{P_VSA_s_4} - 4.433 \times \text{SpDiam_AEA}(dm)$	0.766	0.620	0.062, $n = 10$	0.273	0.059
8	$-67.765 + 34.796 \times \text{MATS5m} - 8.129 \times \text{MATS8m} + 40.696 \times \text{SpMin3_Bh}(m)$	0.766	0.615	0.1085, $n = 11$	0.273	0.084
9	$61.058 + 170.730 \times \text{VE2_Dt} - 178.400 \times \text{ChiA_B}(p) - 5.089 \times \text{SpDiam_AEA}(dm)$	0.765	0.613	0.057, $n = 6$	0.271	0.050
10	$35.620 + 34.296 \times \text{MATS5m} - 25.263 \times \text{SpMax4_Bh}(i) + 32.296 \times \text{SpMin3_Bh}(m)$	0.762	0.633	0.016, $n = 10$	0.266	0.003

K_{aff} logarithm of the affinity for SHBG, $\text{ChiA_B}(p)$ average Randic-like index from Burden matrix weighted by polarizability, IDDE mean information content on the distance degree equality, MATS3v Moran autocorrelation of lag 3 weighted by van der Waals volume, MATS5m Moran autocorrelation of lag 5 weighted by mass, MATS8m Moran autocorrelation of lag 8 weighted by mass, NsssCH Number of atoms of type sssCH, P_VSA_s_4 P_VSA-like on I-state, bin 4, $\text{SpDiam_AEA}(dm)$ spectral diameter from augmented edge adjacency matrix weighted by dipole moment, $\text{SpMax4_Bh}(i)$ largest eigenvalue n . 4 of Burden matrix weighted by ionization potential, $\text{SpMin3_Bh}(m)$ smallest eigenvalue n . 3 of Burden matrix weighted by mass, VE2_Dt average coefficient of the last eigenvector from detour matrix, $X2A$ average connectivity index of order 2, $X5v$ valence connectivity index of order 5

r^2 coefficient of determination in the training set

q^2 cross-validated coefficient of determination in the training set (internal validation)

R^2 coefficient of determination in the test set (external validation)

n number of molecules used for the external validation (that is: found in the applicability domain of a given model)

${}^c R_p^2$ a parameter of model non-randomness proposed in (Mitra et al. 2010)

$r_{m(\text{test})}^2$ a parameter describing the prediction of the absolute response rate of the test set, proposed in (Pratim Roy et al. 2009)

Table 3 Predictive power of the chance models (Case I)

	mhr^2	SD	+1 SD	+2.3 SD	+3 SD	mhq^2	SD	+2.3 SD	+3 SD
γ -scrambling	0.544	0.123	0.667	0.827	0.913	0.316	0.183	0.737	0.865
Pseudo-descriptors (original distributions)	0.656	0.066	0.723	0.809	0.855	0.546	0.101	0.778	0.849
Pseudo-descriptors (uniform distributions)	0.669	0.068	0.737	0.825	0.873	0.571	0.091	0.780	0.843

mhr^2 mean highest coefficient of determination from 300 test runs

SD standard deviation

mhq^2 mean highest cross-validated coefficient of determination from 300 test runs

androgens or oestrogens have only moderate affinity for it (Rosner 1990). The protein is implicated in the inflammatory response by modulating the corticosteroid concentration at the site of inflammation (Klieber et al. 2007). On

the other hand, under physiological conditions it buffers blood cortisol levels. In CBG-deficient individuals observed are symptoms of extreme tiredness, hypotension or chronic muscle pain (Marathe and Torpy 2012; Torpy

Table 4 Top 10 QSAR models obtained by the GFA procedure (Case II)

No	$K_{\text{aff}} =$	r^2	$q^2_{(\text{LOO})}$	$q^2_{(\text{L3O})}$	R^2	${}^cR_p^2$	$r^2_{m(\text{test})}$
1	$-9.394 + 5.644 \times q_2 - 0.145 \times \text{ALogP_MR} + 10.252 \times \text{JX}$	0.892	0.823	0.826	0.365,	0.546	0.131
					$n = 7$		
2	$-9.934 - 5.780 \times q_3 + 2.055 \times \text{Shadow_Ylength} - 0.263 \times \text{Shadow_YZ}$	0.885	0.802	0.739	0.325,	0.538	0.178
					$n = 8$		
3	$-4.871 - 6.046 \times q_3 + 5.795 \times \text{JX} - 1.168 \times \text{Shadow_Zlength}$	0.879	0.805	0.786	0.732,	0.531	0.630
					$n = 8$		
4	$8.117 - 6.065 \times q_3 + 1.013 \times \text{CHI_3_C} - 2.035 \times \text{Shadow_Zlength}$	0.877	0.790	0.778	0.635,	0.529	0.366
					$n = 8$		
5	$-1.520 - 6.636 \times q_3 + 0.840 \times \text{Shadow_Ylength} - 1.209 \times \text{Shadow_Zlength}$	0.877	0.775	0.750	0.255,	0.529	0.078
					$n = 7$		
6	$4.437 + 1.986 \times q_2 - 5.312 \times q_3 - 1.112 \times \text{Shadow_Zlength}$	0.871	0.803	0.802	0.654,	0.522	0.484
					$n = 7$		
7	$9.151 - 6.785 \times q_3 + 2.103 \times \text{srm2} - 2.035 \times \text{Shadow_Zlength}$	0.867	0.789	0.736	0.691,	0.518	0.371
					$n = 8$		
8	$3.079 + 2.976 \times q_2 - 4.246 \times q_3 - 0.074 \times \text{ALogP_MR}$	0.867	0.788	0.800	0.619,	0.518	0.425
					$n = 8$		
9	$3.451 - 2.334 \times q_1 - 5.156 \times q_3 - 1.034 \times \text{Shadow_Zlength}$	0.866	0.805	0.102	0.579,	0.516	0.334
					$n = 8$		
10	$-5.392 - 6.237 \times q_3 - 0.076 \times \text{ALogP_MR} + 1.113 \times \text{Shadow_Ylength}$	0.863	0.694	0.600	0.636,	0.513	0.514
					$n = 8$		

K_{aff} negative logarithm of the affinity for CBG, ALogP_MR the Ghose and Crippen estimate of molar refractivity, CHI_3_C Kier and Hall molecular connectivity index, cluster subgraph of order 3, JX Balaban Index JX , q_1 CHELPG atomic charge at the C1 atom of the steroid skeleton (IUPAC numbering), q_2 CHELPG atomic charge at the C2 atom of the steroid skeleton (IUPAC numbering), q_3 CHELPG atomic charge at the C3 atom of the steroid skeleton (IUPAC numbering), Shadow_Ylength length of molecule in the y dimension, Shadow_YZ area of the molecular shadow in the yz plane, Shadow_Zlength length of molecule in the z dimension, srm2 Sinister-Rectus Chirality Measure weighted by mass

r^2 coefficient of determination in the training set

$q^2_{(\text{LOO})}$ cross-validated coefficient of determination in the training set (internal validation, leave-one-out procedure)

$q^2_{(\text{L3O})}$ cross-validated coefficient of determination in the training set (internal validation, leave-three-out procedure)

R^2 coefficient of determination in the test set (external validation)

n number of molecules used for the external validation (that is: found in the applicability domain of a given model)

${}^cR_p^2$ a parameter of model non-randomness proposed in (Mitra et al. 2010)

$r^2_{m(\text{test})}$ a parameter describing the prediction of the absolute response rate of the test set, proposed in (Pratim Roy et al. 2009)

Table 5 Predictive power of the chance models (Case II)

	mhr^2	SD	+1 SD	+2.3 SD	+3 SD	mhq^2	SD	+2.3 SD	+3 SD
y-scrambling	0.475	0.105	0.580	0.716	0.789	0.341	0.148	0.681	0.785
Pseudo-descriptors (original distributions)	0.570	0.090	0.660	0.776	0.839	0.456	0.118	0.728	0.811
Pseudo-descriptors (uniform distributions)	0.558	0.098	0.656	0.784	0.853	0.431	0.128	0.725	0.815

mhr^2 mean highest coefficient of determination from 300 test runs

SD standard deviation

mhq^2 mean highest cross-validated coefficient of determination from 300 test runs

et al. 2013). Some research has been also made on the role of CBG in glucose metabolism (Fernández-Real et al. 1999), obesity (Ousova et al. 2004) or sperm motility

(Teves et al. 2010). Recently, an interesting proposition was put forward to use engineered CBGs as drug delivery agents (Chan et al. 2014).

Fig. 3 Progesterone in the binding site of CBG (PDB accession code: 4BB2)

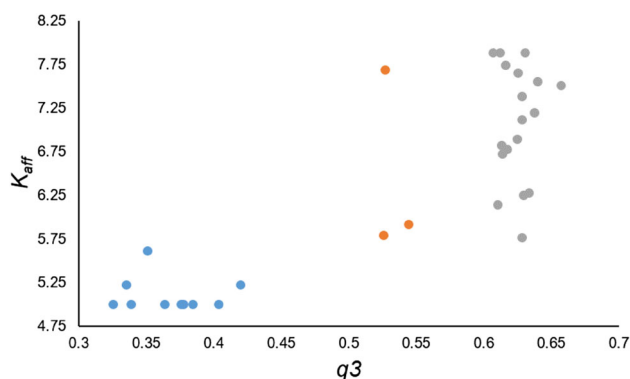
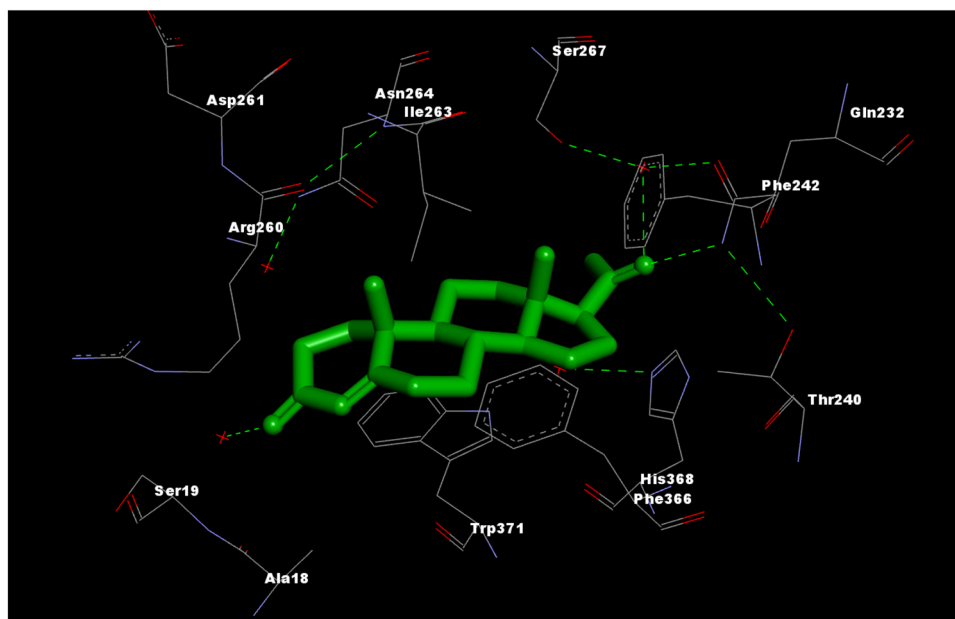


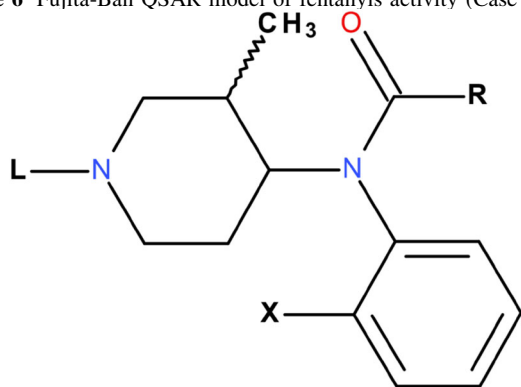
Fig. 4 Plot of $q3$ and K_{aff}

In the study, we divided the CBG set into training and test subsets (in proportion 21:10). The GFA procedure was used to find equations of up to 3 variables, using 49 descriptors derived from 2D and 3D molecular structures. The top 10 models are presented in Table 4.

The presented models have good statistical parameters ($r^2 = 0.863\text{--}0.892$, $q^2_{(LOO)} = 0.694\text{--}0.823$, $q^2_{(L30)} = 0.600\text{--}0.826$). A look at the performance of the chance models allows to conclude that in this modelling situation (21 data points and 49 molecular descriptors) the probability of chance correlations is lower than in the Case I (Table 5). All obtained QSAR models are significantly better than y -scrambled or pseudo-descriptor models. Their additional advantage is clear physical meaning of the variables used (except for two topological descriptors). External validation on several ligands (7–8, depending on the applicability domain of a given model, Tables 4 and SI-3) yields both poor and good results. Three models have

external R^2 much lower than 0.5, but on the other hand in the case of the best two (model 3 and 7) the value is 0.732 and 0.691, which is a decent outcome. Model 3 fulfils also the widely accepted criteria for QSAR model predictive power (Golbraikh and Tropsha 2002): $q^2 > 0.5$, $R^2 > 0.6$, $\frac{R^2 - R_0^2}{R^2} < 0.1$ and $0.85 \leq k \leq 1.15$, where R_0^2 denotes external coefficient of determination forced through the origin, and k is a slope of the regression line through the origin. Here, the value of $r^2_{m(\text{test})}$ parameter is 0.630 and this further supports the predictive power of the model with regard to exact affinity values of the test compounds. Note also that the model has good R^2_{int} and Q^2_{int} metrics (their values provided in Table SI-4 in Electronic Supporting Material).

Experimental structures of the corticosteroid-binding globulin co-crystallized with cortisol or progesterone (Fig. 3. PDB accession codes: 2V95, 4BB2) allow to interpret the models in structural terms (Klieber et al. 2007; Gardill et al. 2012). The interaction of corticosteroids or progesterone with CBG depends mainly on hydrogen bonds formed by polar functions at C and D steroidal rings (IUPAC steroid nomenclature). Although in our models, no charge descriptors for C- and D-rings atoms are present, this is accounted for by shape descriptors like Shadow_Zlength or srcm2. The presence or absence of pharmacophoric polar elements (C17 chain with a keto group, C11 hydroxyl group etc.) affects the size of the molecule or non-superposability on its mirror image and thus these important features are indirectly included into equations. On the other hand, $q3$ descriptor depicts electrostatics of the A ring. If we plot $q3$ and K_{aff} , there appear three clusters (Fig. 4). The lowest $q3$ values characterize molecules with a hydroxyl group attached to C3 atom. The

Table 6 Fujita-Ban QSAR model of fentanyls activity (Case III)

	Equation terms	Coefficient	Standard error
3-CH ₃	Intercept (parent)	2.604	0.330
	Cis/trans	-0.001	0.225
L	Phenylethyl	-0.281	0.237
	Tetrazolyethyl	-1.657	0.413
	Thienylethyl	0.000	0.000
R	CH ₂ OCH ₃	0.000	0.000
	CH(CH ₃)OCH ₃	-1.116	0.262
	Furoyl	-0.722	0.247
X	F	0.163	0.262
	Cl	-0.918	0.348

middle three are those with C3-keto group but with the charge modified due to a C2-substituent or saturation of the C4–C5 double bond (dihydrotestosterone). The third cluster contains molecules with C3-keto group. There exists some rough correlation between q_3 and K_{aff} ($r^2 = 0.690$) showing that the C3-keto group (with its geometry and electrostatics) is preferred over C3-hydroxyl, perhaps due

to a formation of more favourable hydrogen bonds network with water and surrounding amino acids of the binding site. The clustering achieved by q_3 is refined by the shape descriptors (bearing also indirectly information on the most important pharmacophoric elements) or the topological JX descriptor (the role of which is not easily interpretable on its own) and thus good QSAR models are obtained.

Concluding, the models obtained in Case II are not only internally good, but also significantly better than chance correlations in this modelling situation. Further, they are well-interpretable. As such, they may be expected to possess some predictive power, what is shown by external validation.

Case III

Case III represents a different modelling situation than the previous two, since it was attempted to build Fujita-Ban models (Fujita and Ban 1971). This type of QSAR analysis uses variables that are discrete indicators (taking 0 or 1 values) of presence or absence of particular structural elements in a molecule. Fujita-Ban models have a clear physical sense, but on the other hand they contain multiple parameters. The ratio of the number of equation variables to the number of data points is usually larger than in 'typical' QSARs with variables of a continuous character.

In this Case, we considered a group of 36 active (training set) and 10 inactive (test set) fentanyl derivatives (3-methyl-1,4-disubstituted piperidines) (Lalinde et al. 1990) (Table SI-1). Fentanyls or more basically 4-anilidopiperidines are one of the most important groups of analgesics. Since the discovery of fentanyl in the late 1950s (Janssen et al. 1963), numerous derivatives with varying

Fig. 5 Plot of predicted and experimental activities for the QSAR model in Case III

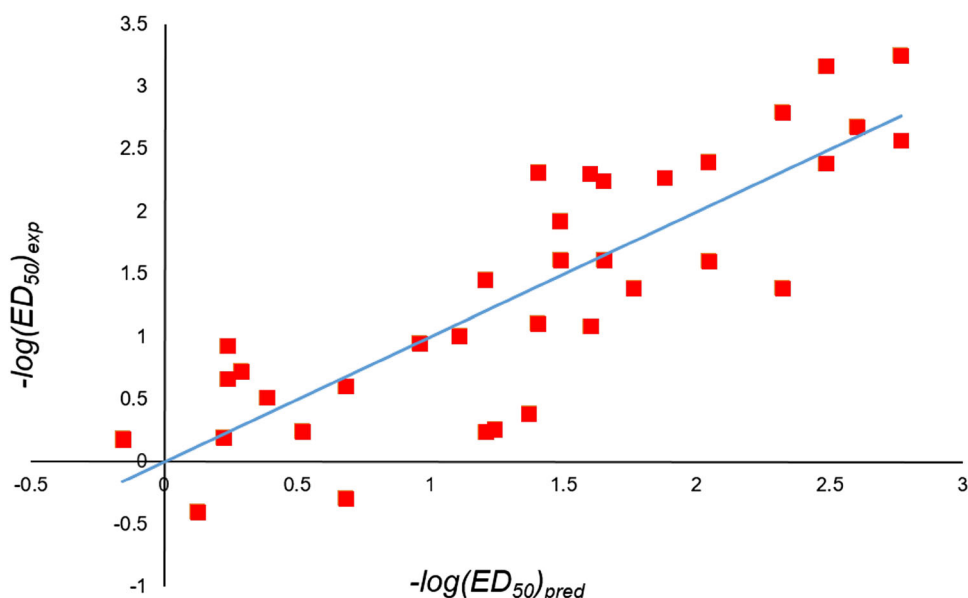


Table 7 Predictive power of the chance models (Case III)

	mhr^2	SD	+1 SD	+2.3 SD	+3 SD	mhq^2	SD	+2.3 SD	+3 SD
Pseudo-descriptors test (original distributions)	0.293	0.108	0.401	0.541	0.617	–	–	–	–
Pseudo-descriptors test (uniform distributions)	0.285	0.112	0.397	0.542	0.621	–	–	–	–
γ -scrambling	0.235	0.092	0.327	0.446	0.511	–	–	–	–

Most q^2 were negative

mhr^2 mean highest coefficient of determination from 300 test runs

SD standard deviation

mhq^2 mean highest cross-validated coefficient of determination from 300 test runs

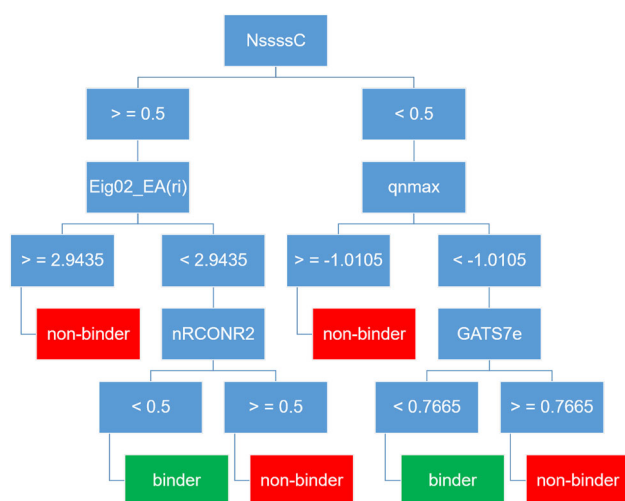


Fig. 6 Scheme of the classification tree (Case IV). $Eig02_EA(ri)$ eigenvalue $n. 2$ from edge adjacency matrix weighted by resonance integral, $GATS7e$ Geary autocorrelation of lag 7 weighted by Sanderson electronegativity, $nRCNR2$ number of tertiary amides (aliphatic), $NssssC$ number of atoms of type $ssssC$ ($>C<$), where $<$ - or $>$ are two single bonds, $qmax$ maximum negative charge

activity have been synthesized and described (Vardanyan and Hruby 2014). Four of them are present in medicinal practice and these are fentanyl, alfentanil, sufentanil and remifentanil. They are used for pain management in terminally ill cancer patients and anaesthesia. Fentanyls act at the μ -opioid receptor (MOR), belonging to the family A of G-protein coupled receptors (GPCR). Unfortunately, this class of analgesics is not free of typical unwanted side effects of opioids (Chaney 1995) nor of their potential for abuse (Skulska et al. 2005; Algren et al. 2013; Mountney et al. 2016).

The dependent variable for QSAR model building was the effective dose ED_{50} in mouse hot plate test (analgesic activity test). Multiple Linear Regression correlated indicator variables with the activity to give an equation the terms of which are presented in Table 6. The plot of

experimental vs predicted activities is given in Fig. 5. The equation has a moderate r^2 of 0.718 and large errors of terms coefficients, rendering a few of the terms insignificant. On the other hand the predictive power of chance models in this particular modelling situation is rather low, and even such moderately good QSAR model is better than the best predictions trained on random data (Table 7). Large errors may be attributed to inaccuracies of the experimental data (in vivo testing), but still the model is able to predict inactivity of six of 10 compounds not used in model training. In the case of the remaining four, it predicts low or very low activity (Table SI-1).

As to the model interpretability, it must be said that statistical insignificance of the terms causes any interpretations to be only rough in their nature, even though all terms are physically well-defined. Nevertheless, the coefficients of L-descriptors (Table 6) seem to fit the Structure-Activity Relationship knowledge on fentanyl derivatives, with the following order of L-substitution preference: thienylethyl (as in sufentanil) $>$ phenylethyl (as in fentanyl) $>$ tetrazolyethyl (as in alfentanil) (Volpe et al. 2011). Regarding the R-part of the molecules, it is clearly visible that R-methoxymethyl is more favourable for analgesic activity than its branched ($R-CH(CH_3)OCH_3$) or rigidified (R-furoyl) counterparts. The freedom of rotation and lack of steric hindrance may allow more facile formation of hydrogen bonds. Unfortunately, the role of 3-Me stereochemistry is not well rendered in the model by the statistically insignificant coefficient. In general, however, it is well-known that 3-*cis* substituents are more active (Vuckovic et al. 2009). No clear conclusions may be drawn about X substituents, again due to the insignificance of the coefficients.

The model presented in Case III is most probably not a random one, but still it is rather inaccurate. As mentioned, large coefficient errors are attributable to the inaccuracies of in vivo data. Thus, even though the model is not random and partially interpretable, it may be of only partial utility.

Table 8 Statistical parameters of the decision tree (Case IV) and comparison with different random models and no-model predictions

	TP ^a	TN ^b	FP ^c	FN ^d	ACC ^e	PREC ^f	SENS ^g	SPEC ^h	FALL ⁱ	F1 ^j
Training set										
QSAR model trained on real data										
	90	3371	226	10	0.94	0.28	0.90	0.94	0.06	0.43
No-model: all binders										
	100	0	3600	0	0.03	0.03	1.00	0.00	1.00	0.06
No-model: all non-binders										
	0	3600	0	100	0.97	NaN ^k	0.00	1.00	0.00	NaN
No-model: coin toss										
	50	1800	1800	50	0.50	0.03	0.50	0.50	0.50	0.06
QSAR models trained on pseudo-descriptors (original distributions)										
Mean	80.00	1785.40	1811.60	19.00	0.50	0.04	0.81	0.50	0.50	0.08
SD ^l	9.20	402.30	402.30	9.20	0.11	0.01	0.09	0.11	0.11	0.02
+1 SD	89.20	2187.80	2213.90	28.20	0.61	0.05	0.90	0.61	0.62	0.10
+2.3 SD	101.20	2710.80	2736.90	40.20	0.75	0.06	1.02	0.75	0.76	0.13
QSAR models trained on pseudo-descriptors (uniform distribution)										
Mean	82.30	1666.20	1930.80	16.70	0.47	0.04	0.83	0.46	0.54	0.08
SD	13.50	461.70	461.70	13.50	0.12	0.01	0.14	0.13	0.13	0.02
+1 SD	95.80	2127.90	2392.60	30.30	0.60	0.05	0.97	0.59	0.67	0.10
+2.3 SD	113.40	2728.10	2992.80	47.90	0.75	0.06	1.15	0.76	0.83	0.13
Test set										
QSAR model trained on real data										
Mean	45	1674	126	5	0.93	0.26	0.90	0.93	0.07	0.40

^a True positives^b True negatives^c False positives^d False negatives^e Accuracy^f Precision^g Sensitivity^h Specificityⁱ Fall-out^j F1-score^k Not a number^l Standard deviation

Case IV

In the last Case, the objective was to create a classification model able to discern glucocorticoid receptor (GR) binders and non-binders. GR is a nuclear receptor-binding corticosteroid and acts as a transcription factor to up- or downregulate the expression of certain genes (Luisi et al. 1991; Yudit and Cidowski 2002). It is involved in maintaining homeostasis by affecting inflammatory responses, cellular proliferation and differentiation in target tissues (Funder 1997). GR ligands include classical steroidal glucocorticoids which are used for tackling diseases involving inflammation (van der Velden 1998; Barnes 1998), for immunosuppression (Coutinho and Chapman 2011) or for

cancer treatment (Coleman 1992; Vaidya et al. 2010). Current medicinal chemistry focuses on development of selective glucocorticoid receptor modulators (based on scaffolds different from the steroidal), which would be void of typical side effects of steroidal glucocorticoids (De Bosscher 2010).

For the modelling purposes, we decided to mimic a most common real-world situation (as for example in virtual screening experiments), where the number of receptor binders is much smaller than that of non-binders. Therefore, we decided to keep the original proportion of actives vs decoys occurring in the DUD-E data set (Mysinger et al. 2012) that is 1:36. The machine learning algorithm obtained the classification model presented in Fig. 6. It is a

simple decision tree with a maximal node depth being three. The model has good statistical parameters of internal predictions (Table 8). Models trained on random data have significantly lower accuracies, precisions and specificities and significantly higher fall-out rates, but on the other hand they are comparably sensitive. F1-score, a measure considering both precision and sensitivity, is however much better for the model trained on true data. The quality of the decision tree may also be assessed by comparison to no-model predictions: ‘all binders’, ‘all non-binders’ or ‘coin-toss’. The analysis of their parameters (Table 8) gives optimistic results, with precision and F1-score again much better in the case of the model trained on true data.

Regarding the interpretability of the model, it must be concluded that even though some of the descriptors used in the model are physically well understandable, the tree does not allow to provide explicit statements about what structural features are important for GR binding. The model is thence uninterpretable. Still, when applied for classification of the test set containing 1850 molecules (50 binders and 1800 decoys), it performs correctly for about 93% of cases. The precision (0.26) and F1-score (0.40) are here similar to the ones for the internal predictivity.

Thus, the classification tree in the Case IV is both internally good as well as better than random. The model is not easily interpretable, however physical interpretability is not what is usually expected of classification models. The most important here is good predictivity, what is shown in external validation.

Conclusions

Since the danger of overfitting QSAR models, when working on large descriptor pools, is very high, it is desirable to perform tests showing the performance of models built on random data. In this study we introduce a simple software tool SCRAMBLE’N’GAMBLE that is aimed at facilitating data preparation for y-scrambling and pseudo-descriptors tests. As shown in the Cases studied in the paper, these tests may be applied to all sorts of QSAR techniques, including both classical linear equations, Fujita-Ban models or classification trees. Their results indicate what the quality of a studied model is like in comparison to chance models obtained from random data. While the non-randomness is not the ultimate hallmark of QSAR models’ possible utility, it is a good practice to consider it along with internal statistical parameters and interpretability of the model. On the other hand, if a model performs no better than chance, it is very probable that it will not be of any use in predicting activities of novel compounds. SCRAMBLE’N’GAMBLE (available for free at: <http://www.drugdesign.pl/scramble-n-gamble/>) is hoped

to help QSAR researchers to perform y-scrambling and pseudo-descriptors testing.

Acknowledgements The study was supported by National Science Centre in Poland. The software writing and the research in Cases I, II and IV were funded by Grant 2012/05/N/NZ7/01952, while the research on Case III was funded by Grant 2013/11/B/ST4/00785. Computational Grant G63-10 from the Interdisciplinary Centre for Mathematical and Computer Modelling (ICM) at the University of Warsaw is gratefully acknowledged. PFJL thanks the National Medicines Institute in Warsaw for the internship, during which part of the research was performed. We are indebted to Prof. Jan Cz. Dobrowolski for his comments and suggestions.

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

References

- Accelrys Software Inc. (2009) Discovery studio modeling environment, Release 2.5, San Diego
- Algren DA, Monteilh CP, Punja M et al (2013) Fentanyl-associated fatalities among illicit drug users in Wayne County, Michigan (July 2005–May 2006). *J Med Toxicol* 9:106–115. doi:10.1007/s13181-012-0285-4
- Anderson DC (1974) Sex-hormone-binding globulin. *Clin Endocrinol (Oxf)* 3:69–96. doi:10.1111/j.1365-2265.1974.tb03298.x
- Balaban AT (1985) Applications of graph theory in chemistry. *J Chem Inf Model* 25:334–343. doi:10.1021/ci00047a033
- Barnes PJ (1998) Anti-inflammatory actions of glucocorticoids: molecular mechanisms. *Clin Sci (Lond)* 94:557–572
- Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J Comput Chem* 11:361–373. doi:10.1002/jcc.540110311
- Caldwell JD, Jirikowski GF (2014) Sex hormone binding globulin and corticosteroid binding globulin as major effectors of steroid action. *Steroids* 81:13–16. doi:10.1016/j.steroids.2013.11.010
- Chan WL, Zhou A, Read RJ (2014) Towards engineering hormone-binding globulins as drug delivery agents. *PLoS One* 9:e113402. doi:10.1371/journal.pone.0113402
- Chaney MA (1995) Side effects of intrathecal and epidural opioids. *Can J Anaesth* 42:891–903. doi:10.1007/BF03011037
- Chemical Computing Group ULC (2017) Molecular operating environment (MOE), 2013.08. Montreal, QC, Canada
- Cherkasov A, Ban F, Santos-Filho O et al (2008) An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *J Med Chem* 51:2047–2056. doi:10.1021/jm7011485
- Clark RD, Sprous DG, Leonard JM (2001) Validating models based on large data sets. In: Hoeltje HD, Sippl W (eds) *Rational approaches to drug design*. Prous Science, Barcelona, pp 475–485
- Coats EA (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. In: Kubinyi H, Folkers G, Martin YC (eds) *Three-dimensional quantitative structure activity relationships*. Kluwer Academic Publishers, Dordrecht, pp 199–213

- Coleman RE (1992) Glucocorticoids in cancer therapy. *Biotherapy* 4:37–44
- Coutinho AE, Chapman KE (2011) The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights. *Mol Cell Endocrinol* 335:2–13. doi:10.1016/j.mce.2010.04.005
- Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110:5959–5967. doi:10.1021/ja00226a005
- Cunningham SK, Loughlin T, Culliton M, McKenna TJ (1983) Plasma sex hormone-binding globulin and androgen levels in the management of hirsute patients. *Eur J Endocrinol* 104:365–371. doi:10.1530/acta.0.1040365
- Dalby A, Nourse JG, Hounshell WD et al (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Model* 32:244–255. doi:10.1021/ci00007a012
- De Bosscher K (2010) Selective glucocorticoid receptor modulators. *J Steroid Biochem Mol Biol* 120:96–104. doi:10.1016/j.jsbmb.2010.02.027
- Dearden JC (2016) The history and development of quantitative structure-activity relationships (QSARs). *Int J Quant Struct Relationsh* 1:1–44
- Dyer DW (2006) *Uncommons Maths*
- Fernández-Real JM, Grasa M, Casamitjana R et al (1999) Plasma total and glycosylated corticosteroid-binding globulin levels are associated with insulin secretion. *J Clin Endocrinol Metab* 84:3192–3196. doi:10.1210/jcem.84.9.5946
- Frisch MJ, Trucks GW, Schlegel HB, et al (2009) Gaussian 09, Revision D. 01
- Fujita T, Ban T (1971) Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J Med Chem* 14:148–152. doi:10.1021/jm00284a016
- Funder JW (1997) Glucocorticoid and mineralocorticoid receptors: biology and clinical relevance. *Annu Rev Med* 48:231–240. doi:10.1146/annurev.med.48.1.231
- Gardill BR, Vogl MR, Lin H-Y et al (2012) Corticosteroid-binding globulin: structure-function implications from species differences. *PLoS One* 7:e52759. doi:10.1371/journal.pone.0052759
- Golbraikh A, Tropsha A (2002) Beware of q 2! *J Mol Graph Model* 20:269–276. doi:10.1016/S1093-3263(01)00123-1
- Gramatica P (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 26:694–701. doi:10.1002/qsar.200610151
- Hammond GL (2011) Diverse roles for sex hormone-binding globulin in reproduction. *Biol Reprod* 85:431–441. doi:10.1095/biolreprod.111.092593
- Helguera AM, Combes RD, González MP, Cordeiro MNDS (2008) Applications of 2D descriptors in drug design: a DRAGON tale. *Curr Top Med Chem* 8:1628–1655
- Hong H, Branham WS, Ng HW et al (2015) Human sex hormone-binding globulin binding affinities of 125 structurally diverse chemicals and comparison with their binding to androgen receptor, estrogen receptor, and α -fetoprotein. *Toxicol Sci* 143:333–348. doi:10.1093/toxsci/ktu231
- Jamróz MH (2010) CHIMEA-program calculating discrete chirality measures of molecules
- Jamróz MH, Rode JE, Ostrowski S et al (2012) Chirality measures of α -amino acids. *J Chem Inf Model* 52:1462–1479. doi:10.1021/ci300057h
- Janssen PAJ, Niemegeers CJ, Dony JGH (1963) The inhibitory effect of fentanyl and other morphine-like analgesics on the warm water induced tail withdrawal reflex in rats. *Arzneimittelforschung* 13:502–507
- Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 96:1027–1044. doi:10.1021/cr950202r
- Key T, Appleby P, Barnes I, Reeves G (2002) Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *Cancer Spectr Knowl Environ* 94:606–616. doi:10.1093/jnci/94.8.606
- Klieber MA, Underhill C, Hammond GL, Muller YA (2007) Corticosteroid-binding globulin, a structural basis for steroid transport and proteinase-triggered release. *J Biol Chem* 282:29594–29603. doi:10.1074/jbc.M705014200
- Lalinde N, Moliterni J, Wright D et al (1990) Synthesis and pharmacological evaluation of a series of new 1,4-disubstituted 3-methyl-piperidine analgesics. *J Med Chem* 33:2876–2882. doi:10.1021/jm00172a032
- Luisi BF, Xu WX, Otwinowski Z et al (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* 352:497–505. doi:10.1038/352497a0
- Marathe CS, Torpy DJ (2012) Corticosteroid-binding globulin gene mutations and chronic fatigue/pain syndromes: an overview of current evidence. In: Snell CR (ed) *An international perspective on the future of research in chronic fatigue syndrome*. InTech, Rijeka
- Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comput Simul* 8:3–30. doi:10.1145/272991.272995
- Mitra I, Saha A, Roy K (2010) Exploring quantitative structure-activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol Simul* 36:1067–1079. doi:10.1080/08927022.2010.503326
- Mountney J, Griffiths P, Sedefov R et al (2016) The drug situation in Europe: an overview of data available on illicit drugs and new psychoactive substances from European monitoring in 2015. *Addiction* 111:34–48. doi:10.1111/add.13056
- Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 55:6582–6594. doi:10.1021/jm300687e
- Ostrowski S, Jamróz MH, Rode JE, Dobrowolski JC (2012) On stability, chirality measures, and theoretical VCD spectra of the chiral C₅₈X₂ fullerenes (X = N, B). *J Phys Chem A* 116:631–643. doi:10.1021/jp208687c
- Ostrowski S, Jamróz MH, Dobrowolski JC (2013) A study on the stability, chirality, and theoretical spectra of the heterofullerenes C₆₉X (X = N, P, As, B, Si, Ge). *Tetrahedron Asymmetry* 24:1097–1109. doi:10.1016/j.tetasy.2013.07.022
- Ousova O, Guyonnet-Duperat V, Iannuccelli N et al (2004) Corticosteroid binding globulin: a new target for cortisol-driven obesity. *Mol Endocrinol* 18:1687–1696. doi:10.1210/me.2004-0005
- Pratim Roy P, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. *Molecules* 14:1660–1701. doi:10.3390/molecules14051660
- Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J Chem Inf Model* 34:854–866. doi:10.1021/ci00020a020
- Rosner W (1990) The functions of corticosteroid-binding globulin and sex hormone-binding globulin: recent advances. *Endocr Rev* 11:80–91. doi:10.1210/edrv-11-1-80
- Rücker C, Rücker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357. doi:10.1021/ci700157b
- Saxena AK, Devillers J, Pery ARR et al (2014) Modelling the binding affinity of steroids to zebrafish sex hormone-binding globulin.

- SAR QSAR Environ Res 25:407–421. doi:[10.1080/1062936X.2014.909197](https://doi.org/10.1080/1062936X.2014.909197)
- Skulska A, Kała M, Parczewski A (2005) Fentanyl and its analogues in clinical and forensic toxicology. *Przegląd Lek* 62:581–584
- Taleta Srl (2010) DRAGON for windows (software for molecular descriptor calculations)
- Teves ME, Guidobaldi HA, Uñates DR et al (2010) Progesterone sperm chemoattraction may be modulated by its corticosteroid-binding globulin carrier protein. *Fertil Steril* 93:2450–2452. doi:[10.1016/j.fertnstert.2009.09.012](https://doi.org/10.1016/j.fertnstert.2009.09.012)
- Torpy DJ, Bachmann AW, Grice JE et al (2013) Familial corticosteroid-binding globulin deficiency due to a novel null mutation: association with fatigue and relative hypotension. *J Clin Endocrinol, Metab*
- Vaidya JS, Baldassarre G, Thorat MA, Massarut S (2010) Role of glucocorticoids in breast cancer. *Curr Pharm Des* 16:3593–3600
- van der Velden VH (1998) Glucocorticoids: mechanisms of action and anti-inflammatory potential in asthma. *Mediat Inflamm* 7:229–237. doi:[10.1080/09629359890910](https://doi.org/10.1080/09629359890910)
- Vardanyan RS, Hrubby VJ (2014) Fentanyl-related compounds and derivatives: current status and future prospects for pharmaceutical applications. *Future Med Chem* 6:385–412. doi:[10.4155/fmc.13.215](https://doi.org/10.4155/fmc.13.215)
- Volpe DA, Tobin GAMM, Mellon RD et al (2011) Uniform assessment and ranking of opioid μ receptor binding constants for selected opioid drugs. *Regul Toxicol Pharmacol* 59:385–390. doi:[10.1016/j.yrtph.2010.12.007](https://doi.org/10.1016/j.yrtph.2010.12.007)
- Vucković S, Prostran M, Ivanović M et al (2009) Fentanyl analogs: structure-activity-relationship study. *Curr Med Chem* 16:2468–2474
- Wilson VS, Cardon MC, Gray LE, Hartig PC (2007) Competitive binding comparison of endocrine-disrupting compounds to recombinant androgen receptor from fathead minnow, rainbow trout, and human. *Environ Toxicol Chem* 26:1793–1802. doi:[10.1897/06-593R.1](https://doi.org/10.1897/06-593R.1)
- Wold S, Sjoestroem M, Eriksson L (1998) Partial least squares projections to latent structures (PLS) in chemistry. In: von Schleyer RP (ed) *Encyclopedia of computational chemistry*. Wiley, New York, pp 2006–2021
<http://scikit-learn.org/Scikit-learn>. <http://scikit-learn.org/>. Accessed 24 Jan 2017
- Yudt MR, Cidlowski JA (2002) The glucocorticoid receptor: coding a diversity of proteins and responses through a single gene. *Mol Endocrinol* 16:1719–1726. doi:[10.1210/me.2002-0106](https://doi.org/10.1210/me.2002-0106)
- Zabrodsky H, Avnir D (1995) Continuous symmetry measures. 4. Chirality. *J Am Chem Soc* 117:462–473. doi:[10.1021/ja00106a053](https://doi.org/10.1021/ja00106a053)
- Zayit A, Pinsky M, Elgavi H et al (2011) A web site for calculating the degree of chirality. *Chirality* 23:17–23. doi:[10.1002/chir.20807](https://doi.org/10.1002/chir.20807)
- (2017) *Small-Molecule Drug Discovery Suite 2017-1*, Schrödinger, LLC, New York