**ABSTRACT**

A comparative study of classification models of brain penetration by different approaches was carried out on a training set of 1000 chemicals and drugs, and an external test set of 100 drugs. Ten approaches were applied in this work: seven medicinal chemistry approaches (including “rule of 5” and multiparameter optimization) and also three SAR techniques: logistic regression (LR), random forest (RF) and support vector machine (SVM). Forty-one different medicinal chemistry descriptors representing diverse physicochemical properties were used in this work. Medicinal chemistry approaches based on the intuitive estimation of preference zones of CNS or non-CNS chemicals, with different rules and scoring functions, yield unbalanced models with poor classification accuracy. RF and SVM methods yielded 82% and 84% classification accuracy respectively for the external test set. LR was also successful in CNS/non-CNS (denoted in this study as CNS+/CNS−) classification and yielded an overall accuracy equivalent to that of SVM and RF. At the same time, LR is especially valuable for medicinal chemists because of its simplicity and the possibility of clear mechanistic interpretation.

**Introduction**

Statistics indicates that the percentage of people over 65 years of age in the population of many industrialized countries is permanently increasing. It has been estimated that the treatment of central nervous system (CNS) diseases affected by aging, such as Alzheimer’s disease, Parkinson’s disease, brain cancer and stroke, will cost trillions of dollars.

The human brain is a uniquely complex organ, which has evolved a sophisticated protection system to prevent injury from external insults and toxins. Designing molecules that can overcome this protection system and achieve optimal concentration at the desired therapeutic target in the brain is a specific and major challenge for medicinal chemists working in CNS drug discovery [1].

The physicochemical properties needed for blood-brain barrier (BBB) penetration have been extensively studied by many researchers. Early work in this area included that of Young et al. (relationships of logBB with logPoctanol–water and H-bond capacity) [2], van der Waterbeemd et al. (modeling of BBB penetration by plotting various combinations of two physicochemical properties, such as molecular size, shape, H-bonding descriptors and polar surface area (PSA), and formulation of various guidelines for use by medicinal chemists, e.g. PSA of \( \leq 90 \text{Å}^2 \) and molecular mass of \(< 450 \text{Da} \) [3–5], Kelder et al. (estimation of upper limit of PSA for CNS drugs) [6], and Abraham et al. (Linear Free Energy Relationships study) [7–9]. A summary of selected literature of this century describing the physicochemical properties required for optimal brain exposure is presented in [1]. Several physicochemical properties have consistently been found to be important for permeability in CNS, including lipophilicity, expressed by logP; logarithm of the octanol/water partition coefficient at physiological pH 7.4 (logD); the number of hydrogen bond donors (HBD) and acceptors (HBA); PSA; ionization state (pK\(_a\)); rotatable bond (RB) count, molecular weight (MW) and others [1–10].

In 1997, Lipinski proposed his “rule of 5” whereby four physicochemical parameters (MW, HBD, HBA and logP) were utilized for CNS drug-likeness prediction [11]. This publication had a significant influence on the development of medicinal chemistry and has been cited hundreds of times. There is also the “rule of fours” (compounds with \( N+O \geq 8, MW > 400 \) and acid pK\(_a\) > 4 are likely to be Pgp substrates, whereas compounds with \( N+O \leq 4, MW < 400 \) and base pK\(_a\) < 8 are likely to be non-substrates) [12].

In the current century, multivariate models that interrogate multiple descriptors simultaneously have been developed. In particular, Wager et al. proposed multiparameter optimization (MPO) specifically for CNS drug discovery [13]. This method utilizes six physicochemical parameters (logP, logD, MW, TPSA, HBD and pK\(_a\) of the most basic center). Here, instead of hard cutoffs an algorithm constructs desirability scores (within limits 0–1) for each of the six properties and calculates an overall desirability score using summation, ranging from 0 (undesirable) to 6 (highly desirable). Multiple physicochemical property graphical analysis using Radar charts was proposed by Ghose et al. [10]. In a recently published perspective [1], it was indicated that “over the recent years MPO methods have been steadily gaining a wider acceptance in drug discovery, with their application to the prospective design of new compounds being particularly appealing. Machine-learning algorithms designed to differentiate CNS from non-CNS penetrating compounds on the basis of their substructure fingerprints have
also been reported in recent years. However, these approaches are based on molecular descriptors and fingerprint patterns that are less intuitive and difficult for medicinal chemists’.

(Q)SAR methods (including machine-learning techniques) with the construction of classification and regression models of BBB penetration were also successfully developed during this century. Particular attention is drawn to the publications in this field of Adenot et al. [14], Abraham et al. [15] and Raevsky et al. [16], which showed that hydrogen-bonding properties of compounds play a crucial role in the modeling of BBB penetration. Our review of (Q)SAR approaches is presented in [17].

The main aim of this work is a comparison of different computer approaches of CNS+/CNS−/C0 classification using the same datasets. Seven intuitive approaches of medicinal chemists (including the “rule of 5”, MPO), chemometric methods (including logistic regression (LR) and also machine-learning techniques of random forest (RF) and support vector machine (SVM)) were applied in this work to classify CNS and non-CNS chemicals (drugs). The quality of all SAR models was evaluated by the use of strict statistical measures of the performance of a binary classification (CNS+/CNS−):

true positive (TP), false negative (FN), true negative (TN), false positive (FP), sensitivity (SE) = TP/(TP + FN), specificity (SP) = TN/(TN + FP), accuracy (ACC) = (TP + TN)/(TP + FN + TN + FP).

Datasets
The datasets used in this research were based on published data collected by us, and contain information about 3000 chemicals. A training set of 1000 chemicals contained 500 CNS+ and 500 CNS− chemicals. Pruning was carried out by a dissimilarity procedure. In this case, all CNS+ and CNS− chemicals (drugs) were separately arranged in order of their Tanimoto dissimilarity indices, and the first 500 CNS+ and CNS− chemicals were selected for our datasets using our program manager CheD [18] and our MolDivs (MOlecular DIVersity & Similarity) program [19]. An external test set was formed from an excellent dataset of supporting information of the article by Ghose et al. [10]. Fifty CNS+ and 50 CNS− drugs which were not included in the above-mentioned training set were selected by the procedures indicated above.

Figure 1 shows the distribution of chemical functional groups in the studied chemicals. It should be noted that the distributions are approximately equal for the training, test and external sets. A comparison of CNS+ and CNS− sets also indicated a higher percentage of hydroxyl, carboxyl and secondary amine chemicals in the CNS− sets compared with the CNS+ sets.

Descriptors
Forty-one molecular physicochemical descriptors which are widely used in medicinal chemistry permeability research were applied in this work. They are: descriptors of the “rule of 5” [11]: MW, number of HBD atoms, number of HBA atoms, Moriguchi octanol-water partition coefficient (MlogP); MPO descriptors [13]: topological PSA using contributions of nitrogen and oxygen atoms (TPSA(N,O)), octanol-water partition coefficient at pH = 7.4 (logOD), acidity constant (pKₐ); descriptors of Radar chart [10]: Ghose–Crippen octanol-water
Table 1. Statistical protocol of CNS+/CNS− classification by some intuitive approaches of medicinal chemists for 1000 chemicals and drugs of the training set (500 CNS+/500 CNS−) and 100 drugs (50 CNS+/50 CNS−) of the external test set.

| Approach | Cutoffs | Sets | TP | FN | TN | FP | SE | SP | ACC |
|----------|---------|-----|----|----|----|----|----|----|-----|
| Model 1, “Rule of 5” [11] | MW < 500, MlogP < 4.15, HBD ≤ 5, HBA ≤ 10 | Training | 397 | 103 | 198 | 302 | 0.794 | 0.396 | 0.595 |
| | | External | 47 | 3 | 11 | 39 | 0.940 | 0.220 | 0.580 |
| Model 2 [5] | MW < 450, PSA < 90Å², logD 1–4 | Training | 219 | 281 | 416 | 84 | 0.438 | 0.832 | 0.635 |
| | | External | 30 | 29 | 43 | 7 | 0.600 | 0.860 | 0.730 |
| Model 3, [30] | N + O ≤ 5 | Training | 496 | 4 | 59 | 441 | 0.992 | 0.118 | 0.555 |
| | | External | 50 | 0 | 5 | 45 | 1.000 | 0.100 | 0.555 |
| Model 4, [30] | ClogP -(N + O) > 0 | Training | 345 | 155 | 304 | 196 | 0.690 | 0.608 | 0.649 |
| | | External | 36 | 14 | 34 | 16 | 0.720 | 0.680 | 0.700 |
| Model 5 [31] | clogP < 4, TPSA 40–80Å² | Training | 178 | 322 | 394 | 106 | 0.356 | 0.644 | 0.572 |
| | | External | 23 | 27 | 37 | 13 | 0.460 | 0.740 | 0.600 |
| Model 6 [32] | PSA < 90Å², HBD < 3, logP 2–5, logD 2–5, MW < 500 | Training | 69 | 431 | 432 | 48 | 0.138 | 0.864 | 0.501 |
| | | External | 15 | 35 | 46 | 4 | 0.300 | 0.820 | 0.610 |
| Model 7 [13] | MPO score ≥ 4.0/<4.0 | Training | 342 | 158 | 256 | 244 | 0.684 | 0.512 | 0.598 |
| | | External | 34 | 16 | 18 | 32 | 0.680 | 0.360 | 0.520 |

Support vector machine (SVM)

The SVM method can be applied to create classification and regression models. The basis of the method consists of a transition from the initial space of variables to a space of higher dimension with the help of a kernel function and a search for a separating hyperplane with the maximum margin in this space. For constructing the SVM model, the LibSVM package [25] was used. Both training and test data were scaled in standard deviation units (autoscaling procedure) [26]. During construction of a hyperplane in classification, C-SVC minimization was used. The kernel type was radial basis function.

To construct SVM models, three parameters were used: penalty parameter C (default 1), loss function parameter ε (default 0.1) and kernel parameter γ (default 1/M, where M is the number of features). To obtain optimized values of γ and C parameters, a graphical grid search strategy was used based on three-fold cross-validation (CV). Various pairs of C, γ values were tried and the one with the best CV ACC was selected. The interval for C was 2−5 ... 215 and for γ was 2−15 ... 22.

To reduce linear dependence, the selection of the descriptors was performed by two algorithms.

Algorithm 1
(1) Calculate variation coefficient (VC) and correlation coefficients (CC) for each descriptor with others; (2) choose descriptor with maximum VC; (3) set the CC threshold; (4) estimate nearest neighbors for this descriptor; (5) nearest neighbors with CC higher than the threshold are eliminated.

Algorithm 2
(1) Calculate variation coefficient (VC), correlation coefficients (CC) for each descriptor with others, and correlation coefficient with the activity (CCA); (2) choose descriptor with maximum CCA; (3) set the CC threshold; (4) estimate nearest neighbors for this descriptor; (5) nearest neighbors with CC higher than the threshold are eliminated.

Selections of the descriptors were made for the CC threshold in the range [0.50–1.00] in increments of 0.05.

Results and discussion

Empirical intuitive binary CNS+/CNS− classification

There are many publications in which binary classification of permeability is characterized by medicinal chemists by means of the following: improved chance of CNS penetration if MW < 450 and PSA < 90Å² [5]; upper PSA limit for most of the CNS drugs is <60–70Å² [6]; proposed cutoffs to avoid P-gp efflux liability: MW < 400 [12]; attributes of a successful CNS drug candidate:
MW < 450; cLogP < 5; HBD < 3; HBA < 7; RB < 8; H-bonds < 8; pKa 7.5–10.5; PSA < 60–70 Å² [27]; compounds with TPSA < 60 Å² and pKₐ < 8 are less likely to be P-gp substrates [28]. Such rules are guidelines for structural properties of drug-like compounds. The fastest method for evaluating the drug-like properties of a compound is to apply ‘rules’. Rules are a set of guidelines for the structural properties of compounds that have a higher probability of being well absorbed after oral dosing. The values for the properties associated with rules are quickly counted from an examination of the structure or calculated using software that is widely available. These guidelines are not absolute, nor are they intended to form strict cutoff values for which property values are drug-like and which are not drug-like. Nevertheless, they can be quite effective and efficient.

Such declarations do not ensure complete correct information about the ACC of classification of whole studied datasets which have to contain exact data on the ratio of true and false objects among active and inactive classes. Our attempt to put the above-mentioned declarations on a strict statistical basis is presented in Table 1. Published cutoff descriptor values of seven known intuitive approaches were used in this table to estimate its CNS+/CNS− discriminating power for the same datasets: training (1000 chemicals) and external (100 chemicals).

The “rule of 5” or what has become known as the “Lipinski rules” is the most known and popular. “These rules are a set of property values that were derived from classifying the key physico-chemical properties of drug-like compounds. The rules were used at Pfizer for a few years prior to their publication and since then have become widely used. The impact of these rules in the field has been very high. This acceptance can be attributed to many factors: the rules are easy, fast and have no cost to use; the ‘5’s’ mnemonic makes the rules easy to remember; the rules are intuitively evident to medicinal chemists; the rules are a widely used standard benchmark; the rules are based on solid research, documentation and rationale; the rules work effectively” [29].

As is obvious from Table 1, the binary CNS+/CNS− classification model based on the “rule of 5” showed that 397 from 500 CNS+ chemicals of training were placed below the indicated cutoff values and so were classified correctly. However, in this region also, 302 CNS− chemicals were shown. Clearly, this is because of rather high cutoff values. As a result, although the SE in this region also, 302 CNS− indicated cutoff values and so were classified correctly. However, from 500 CNS− chemicals very poorly. Model 7, which was indicated in [1] as a perspective for drug design, gave total prediction ACC close to 50%, which could be obtained simply by random allocation to the two classes.

Other obvious disadvantages characterize all intuitive approaches of CNS+/CNS− classification: arbitrary choice of cutoff descriptor values, poor discriminating power, absence of analysis of the balance between correctly recognized CNS+ and CNS− chemicals and an ignorance of any validation procedures. Those intuitive models are often named as “rules”. That presumes their universal character although often they are constructed on the basis of small datasets [13]. Hence, the application of such “rules” to all chemical space, containing at least 10¹⁸⁰ chemicals [33], seems too far from reality.

**SAR binary CNS+/CNS− classification by RF and SVM**

At present, a strict chemometric binary classification can be achieved by a whole set of classical and modern methods including discriminant analysis, Naïve Bayes, LR, classification and regression tree, k-nearest neighbor, RF, SVM, Gaussian process, stochastic gradient boosting and others. Yee et al. [34] have reviewed this area, and an example of successful application of the above methods for SAR binary classification has been reported by Zang et al. [35]. Table 2 contains results of CNS+/CNS− classification of studied train and external sets by RF and SVM.

Both methods gave good results of CNS+/CNS− classification. The best RF model was constructed using 17 descriptors connected with volume-related, electrostatic terms and hydrogen bonding. CNS+/CNS− classification of the external test set has excellent statistical criteria. Classification by SVM for the external test set using 26 descriptors was even a little better. Results of CV for both models are also quite satisfactory.

**Logistic regression as an alternative to intuitive CNS+/CNS− classification approaches**

The above unsatisfactory situation connected with the predictive power of “intuitive” approaches of medicinal chemists for CNS+/CNS− chemical classification leads to the problem of selecting an appropriate SAR method(s) from the above-indicated wide set of modern powerful chemometric techniques that can be used in medicinal chemistry. The availability in this work of a suitably large descriptor pool suggests the use of a machine-learning technique such as RF and SVM for CNS+/CNS− classification. However, as pointed out by Rankovic [1], “these approaches are less intuitive and difficult for medicinal chemists to translate into optimization hypotheses”. In other words, medicinal chemists prefer to carry out simple and understandable calculations and have models with clear mechanistic interpretation. That is why we selected in this work an LR method for detailed SAR analysis of predicted
CNS+/CNS− classification power of single descriptors and of multiple descriptors. LR is very similar to linear regression, and in our view, it satisfies the above requirements of medicinal chemists. This method is used to model the probability of the occurrence of some event as a linear function of a set of predictors.

The following equation calculates the probability $Y$:

$$Y = 1/(1 + e^{-z})$$

where $z = a_0 + a_1x_1 + a_2x_2 + \ldots + a_n x_n$ and $a_0$ is an intercept, $x_1, x_2 \ldots x_n$ are descriptors with their corresponding regression coefficients. Given an unknown compound, LR calculates the probability that the compound belongs to a certain target class. For example, in predicting whether an unknown compound is CNS+, or CNS−, LR tries to estimate the probability of the compound being a CNS+ chemical. If the calculated $Y > 0.5$, then it is more likely to be CNS+. It is important for medicinal chemists that, in a similar way to multiple linear regression, the regression coefficients in LR can describe the influence of molecular descriptors on the outcome of the prediction. When the coefficient has a large value, it shows that the molecular descriptor strongly affects the probability of the outcome, whereas a zero value coefficient shows that the molecular descriptor has no influence on the outcome probability. Likewise, the sign of each coefficient affects the probability as well; that is, a positive coefficient increases the probability of an outcome, while a negative coefficient will result in the opposite. As indicated in [34], LR generally cannot handle high dimensional data well, especially if the dataset is large (e.g. more than 60,000 compounds) with large class imbalance (e.g. active:non-active = 1:68), which can be common in a training set intended for virtual screening. Such LR models commonly require increased computational time and do not perform well. The size of our datasets and balance ratio of CNS+/CNS− chemicals permit the use of LR, so we have carried out a detailed analysis of binary classification by this method. A successful recent application of LR in SAR for binary classification of active/inactive chemicals is described in [36–38].

First of all, in this work, we used LR for CNS+/CNS− classification by single descriptors. It should be noted that in the case of application of a small number of descriptors, the LR method is even preferable to machine-learning techniques, from a methodology point of view. The supplementary material of this article includes data about the CNS+/CNS− classification ability of 41 single descriptors used in this work (Table S1).

For the 1000-compound training set, total classification ACC for the whole descriptor set is in the range 0.529–0.682 and for the 100-compound external test set the range is 0.490–0.760.

The best classification results were obtained with descriptors TPSA (ACC = 0.760 for external test set), PSAa (0.750), PSAd (0.740), HBD (0.740), $\sum_{\text{C}_2}$ (0.730) and $\sum_{\text{Cad}}$ (0.720). All these descriptors directly or indirectly describe hydrogen bonding ability. For example, $\sum_{\text{Cad}}$ is a thermodynamic descriptor and reflects the total H-bond ability of a molecule [21]. (T)PSA is defined simply as the part of a molecular surface that is polar and can participate in hydrogen bonding. PSA descriptors for the prediction of drug permeability are very popular for several reasons. First, PSA is very easy to interpret, with the notion of “molecular polar surface” and its influence on interactions with a molecule’s environment similar to a medicinal chemist’s own intuition (and is probably also a good approximation of physical reality). Second, PSA is easy to calculate. For TPSA, the calculation is particularly easy and fast, requiring only the identification of polar fragments and then a table lookup to find respective fragment contributions [39]. The high correlation of 3D PSA and TPSA for 34,810 drug-like molecules ($r^2 = 0.982$) helps to confirm the scientific meaning of this descriptor [40].

We expected good CNS+/CNS− classification ability of the above-mentioned descriptors because hydrogen bonding is considered now as one of the most important (if not the most important) intermolecular interactions in general and in drug transport in living organisms in particular.

We have also found that descriptors connected with lipophilicity (MlogP, AlogP, logPP) have the poor CNS+/CNS− discriminating ability, as first shown by Ghose et al. [10]. Single descriptors connected with volume-related terms (MW, $\alpha$) and electrostatic interactions ($\text{maxQ}^+$, $\text{maxQ}^-$) gave accuracies of CNS+/CNS− classification in the range 0.550–0.650. The results of CNS+/CNS− classification by LR using two popular single descriptors, TPSA and MlogP, are given in Table 3.

It is clear from Table 3 that the model using TPSA alone with LR has better statistical criteria compared with models which were obtained by intuitive approaches. The situation with the clogP descriptors was found to be much more complex. A valuable help in the interpretation of those results was obtained from a plot of logP distribution (Figure 2).

Clearly, this graph of MlogP distribution shows that there is very strong overlapping of CNS+ and CNS− clusters, although the cluster of CNS+ objects with a maximum at ~2.5 logP units is situated to the right of the cluster of CNS− chemicals with a maximum at ~1.5. The mean value of those maxima is very close to the threshold estimation of 1.71 by LR procedures (Table 3).

| Method       | Threshold     | Set                  | TP  | FN  | TN  | FP  | SE  | SP  | ACC   |
|--------------|---------------|----------------------|-----|-----|-----|-----|-----|-----|-------|
| Model 10. LR | TPSA < 75.7Å² | Training             | 374 | 126 | 307 | 193 | 0.748 | 0.614 | 0.681 |
|              |               | CV (three-fold)      | 373 | 127 | 305 | 195 | 0.746 | 0.610 | 0.678 |
|              |               | External             | 43  | 7   | 33  | 17  | 0.860 | 0.660 | 0.760 |
| Model 11. LR | MlogP > 1.71  | Training             | 309 | 191 | 294 | 206 | 0.618 | 0.588 | 0.603 |
|              |               | CV (three-fold)      | 308 | 192 | 291 | 209 | 0.616 | 0.582 | 0.599 |
|              |               | External             | 32  | 18  | 27  | 23  | 0.640 | 0.540 | 0.590 |

Further improvement of CNS+/CNS− classification by means of LR is possible by simultaneous application of a number of descriptors. However, it is necessary to be very careful because of possible high pair-wise collinearity of descriptors. For example, the best single descriptors TPSA and PSAa are highly correlated, $r = 0.872$, and similarly for HBD and $\sum_{\text{C}_2}$ with $r = 0.937$. For our dataset containing 1000 chemicals and drugs, using a maximum permissible level of mutual descriptor correlation of $r \leq 0.50$ meant that only 13 from 41 descriptors were kept.

As to the mechanistic interpretation of the results, it is necessary to emphasize that chemical transport in CNS is a very complex process including solvation, absorption, metabolism, interaction with P-glycoprotein and other phenomena. Hence, it is highly unlikely that any single descriptor can ensure satisfactory CNS+/CNS− classification. Roughly, three main components may contribute to chemical transport in the CNS: transmembrane diffusion, efflux effect connected pre-eminently with P-glycoprotein interactions, and influx effects. The majority of CNS drugs are small molecules that cross the BBB via the transcellular passive diffusion route [1]. Kerns and Di [29] make the following comment about...
transmembrane diffusion: “Molecules with a larger molecular size (i.e. higher MW) do not pass through the tightly packed region as readily as smaller molecules. Molecules with higher lipophilicity typically are more permeable than less lipophilic molecules through the highly non-polar central core of the lipid bilayer membrane. Molecules then move through the side chains and polar head groups of the other leaflet of the bilayer and are rehydrated by water molecules and form hydrogen bonds again”.

In [5], an equation of transmembrane diffusion on the basis of MW, HBA and HBD descriptors was proposed. So, we first of all used those three descriptors for CNS\(^+\)/CNS\(^-\)/C\(-\) classification of our dataset. Equation parameters of discriminating function (Table S2) of model 12 based on those three descriptors demonstrate the predominant contribution of HBA, followed by MW. The HBD contribution is no better than its standard deviation. Classification results for model 12 are included in Table 4, showing that three out of four chemicals may be correctly predicted using this model which describes transmembrane diffusion.

Further improvement of CNS\(^+\)/CNS\(^-\)/C\(-\) classification is possible by application of descriptors connected with other components of blood–brain and brain–blood chemical transport. Table 4 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (parameters of LR discriminating function (Table S2) of model 12 contains results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by the two best models obtained by LR for the sets considered. Model 13 is based on four descriptors, including the best single descriptor TPSA, which has in this model maximal contribution to classification with an autoscale coefficient value \(k = -1.008\) (Table S2)) of model 12 based on those three descriptors demonstrate the predominant contribution of HBA, followed by MW. The HBD contribution is no better than its standard deviation. Classification results for model 12 are included in Table 4, showing that three out of four chemicals may be correctly predicted using this model which describes transmembrane diffusion.

In order to analyze the changes of classification results obtained when the contributions of transports are included or excluded, the “rule of four” of Didziapetris et al. [12] was used to eliminate a number of P-glycoprotein substrates from the training set of model 13. As a result, descriptor coefficient values of discriminating functions were essentially changed: contributions of TPSA and \(\text{maxQ}^*/\text{C}\(-\) were remarkably decreased, the contribution of \(\text{maxCa}^*\text{maxCd}\) essentially increased, whilst the contribution of \(\text{maxCa}\) was unaltered. Total ACC was slightly better as a result of better recognition of CNS\(^+\).

Examination of Tables 2 and 4 shows that the overall ACC is approximately equal for the three SAR methods used (LR, RF and SVM). At the same time, LR is a simple method with a strict chemometric platform and allows the investigation of contributions of each descriptor in CNS\(^+\)/CNS\(^-\)/C\(-\) classification. Hence, we propose the LR method in medicinal chemistry as a reasonable alternative to the abovementioned intuitive approaches which apply arbitrary cutoffs.

**Figure 2. Distribution of MlogP values in the CNS\(^+\) and CNS\(^-\) drugs.**

**Table 4. Results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by few descriptors by means of LR.**

| Model | Descriptors | Set | TP | FN | TN | FP | SE | SP | ACC |
|-------|-------------|-----|----|----|----|----|----|----|-----|
| 12    | MW          | Training | 377 | 123 | 289 | 211 | 0.754 | 0.578 | 0.666 |
|       | HBA         | CV(three-fold, it = 10) | 376 | 124 | 289 | 211 | 0.752 | 0.578 | 0.665 |
|       | HBD         | External | 42  | 8   | 36  | 14  | 0.840 | 0.720 | 0.780 |
| 13    | maxQ\(^+\)  | Training | 370 | 130 | 339 | 161 | 0.740 | 0.678 | 0.709 |
|       | maxCa\(^+\) | CV(three-fold, it = 10) | 369 | 131 | 333 | 167 | 0.738 | 0.666 | 0.702 |
|       | maxCa\(^+\)maxCd | External | 42  | 8   | 39  | 11  | 0.840 | 0.780 | 0.810 |
|       | TPSA(N,O)   | Training, 721 | 343 | 71  | 170 | 137 | 0.829 | 0.554 | 0.712 |
|       |             | CV(three-fold, it = 10) | 340 | 74  | 169 | 138 | 0.821 | 0.550 | 0.706 |
|       |             | External | 43  | 7   | 39  | 11  | 0.860 | 0.780 | 0.820 |
| 14    | maxQ\(^+\) , maxCa\(^+\) | Training | 382 | 118 | 340 | 160 | 0.764 | 0.680 | 0.722 |
|       | maxCa\(^+\)maxCd | CV(three-fold, it = 10) | 384 | 116 | 336 | 164 | 0.768 | 0.672 | 0.720 |
|       | MW, HBD     | External | 43  | 7   | 38  | 12  | 0.860 | 0.760 | 0.810 |
|       | NCC, logD   |         |     |     |     |     |     |     |     |

Figure 2. Distribution of MlogP values in the CNS\(^+\) and CNS\(^-\) drugs.

Table 4. Results of CNS\(^+\)/CNS\(^-\)/C\(-\) classification by few descriptors by means of LR.
Four CNS+/CNS− classification models (models 8, 9, 13 and 14) were developed in this work, with overall accuracies of prediction of 0.81–0.84. Each method is independent and based on its own set of descriptors and calculation method. Analysis of those data showed that 41 from 50 CNS+ drugs were correctly recognized by all four models, plus additionally one drug was correctly recognized by three models. For CNS− drugs, the numbers are 30 and 6, respectively. So, 78% of drugs from the external test set were reliably classified by joint consideration of the four models. This allowed the construction of consensus models combining different combinations of three models (Table S4). The consensus model based on models (9), (13) and (14) can be considered as the optimal consensus model not only because of its high ACC (81%) but also because it has the best balance of correctly predicted CNS+ and CNS− drugs. The last column in Table S4 contains results of the program for BBB+/BBB− classification by Bayesian approach [41] with open access on Internet [42] for studied in this work external set. Obtained in this case overall ACC equals only 0.675.

Conclusion
Medicinal chemistry rules for the design of new effective drugs, although probably useful for property optimization in sets of related compounds, are intuitive (i.e. subjective) extreme models because the cutoff takes only one part of the data into account and does not consider other parts. Because no single physicochemical parameter can be used to predict properties related to the very complicated phenomenon of brain penetration, more complex multivariate models that interrogate multiple descriptors simultaneously are required. Such well-known methods of medicinal chemists as “rule of 5” and MPO did not demonstrate in this work any essential advantages compared with one descriptor dependence. However, LR and modern machine-learning techniques of RF and SVM enabled stable predictable SAR models to be obtained with strict validation procedures. Because LR is simple to use and allows of mechanistic interpretation with a description of the influence of molecular descriptors on the outcome of the prediction, this method may be very useful for medicinal chemists.

Disclosure statement
The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Funding information
This work was partly supported by Russian Scientific Foundation (project no.: 14-23-00160).

References
1. Rankovic Z. CNS drug design: balancing physicochemical properties for optimal brain exposure. J Med Chem 2015;58:2584–608.
2. Young RC, Mitchell RC, Brown TH, et al. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. J Med Chem 1988;31:656–71.
3. van de Waterbeemd HD, Kansy M. Hydrogen-bonding capacity and brain penetration. Chimia 1992;46:299–303.
4. van de Waterbeemd HD, Camenisch G, Folgers G, Raevsky OA. Estimation of Caco-2 cell permeability using calculated molecular descriptors. Quant Struct-Act Rel 1996;15:480–90.
5. van de Waterbeemd HD, Camenisch G, Folgers G, Chretien JR, et al. Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. J Drug Target 1998;6:151–65.
6. Kelder J, Grootenhuis PDJ, Bayada DM, et al. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharm Res 1999;16:1514–19.
7. Gratton JA, Abraham MH, Bradbury MW, Chadha HS. Molecular factors influencing drug transfer across the blood-brain barrier. J Pharm Pharmacol 1997;49:1211–16.
8. Abraham MH, Platts JA. Physicochemical factors that influence brain uptake. In: Begley DJ, Bradbury MW, Kreuter J, eds. The blood-brain barrier and drug delivery to the CNS. New York/Basel: Marcel Dekker; 2000:9–32.
9. Platts JA, Abraham MH, Zhao YH, et al. Correlation and prediction of a large blood-brain distribution data set – an LFER study. Eur J Med Chem 2001;36:719–20.
10. Ghose AK, Herbertz T, Hudkins RL, et al. Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery. ACS Chem Neurosci 2012;3:50–68.
11. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 1997;23:3–26.
12. Didziapetris R, Japertas P, Avdeef A, Petrauskas A. Classification analysis of P-glycoprotein substrate specificity. J Drug Target 2003;11:391–406.
13. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond rules: the development of a central nervous system multi-parameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem Neurosci 2010;1:435–9.
14. Adenot M, Lahana R. Blood-brain barrier permeation models: discriminating between potential CNS and Non-CNS drugs including P-glycoprotein substrates. J Chem Inf Comput Sci 2004;44:239–48.
15. Zhao YH, Abraham MH, Ibrahim A, et al. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. J Chem Inf Model 2007;47:170–5.
16. Raevsky OA, Solodova SL, Raevskaya OE, et al. The computer classification models on the relationship between chemical structures of compounds and drugs with their blood brain barrier penetration ability. Biochem (Mosc) Suppl Ser B Biomed Chem 2012;6:31–8.
17. Raevsky OA, Solodova SL, Lagunin AA, Poroikov VV. Computer modeling of blood-brain barrier permeability for physiologically active compounds. Biochem (Mosc) Suppl Ser B Biomed Chem 2013;7:95–107.
18. Trepalin SV, Yarkov AV. CheD: chemical database compilation tool, internet server, and client for SQL servers. J Chem Inf Comput Sci 2001;41:100–7.
19. Gerasimenko VA, Trepalin SV, Raevsky OA. MOLDIVS – a new program for molecular similarity and diversity calculations. In: Gudertofte K, Jorgensen FS, eds. Molecular modeling and prediction of bioactivity. New York: Kluwer Academic/Plenum Publishers; 2000:423–4.
20. Raevsky OA, Grigor’ev YV, Trepalin SV. HYBOT program, registration by Russian State Patent Agency No. 990090 of 26.02.99.
21. Raevsky OA, Skvortsov VS. Quantifying hydrogen bonding in QSAR and molecular modeling. SAR QSAR Environ Res 2005;12:287–300.
22. Raevsky OA. H-bonding parameterization in quantitative structure – activity relationships and drug design. In: Mannhold R, ed. Molecular drug properties, measurement and prediction. Wiley-VCH Verlag GmbH, 2008:128–54.
23. Hansch C, Kurup A. QSAR of chemical polarizability and nerve toxicity. 2. J Chem Inf Comput Sci 2003;43:1647–51.
24. Available from: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_examples/prog.f [last accessed 15 Jun 2004].
25. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm [last accessed 14 Dec 2015].
26. Bro R, Smilde AK. Centering and scaling in component analysis. J Chemom 2003;17:16–33.
27. Pajouhesh H, Lenz GR. Medicinal chemical properties of successful central nervous system drugs. NeuroRx 2005;2:541–53.
28. Desai PV, Sawada GA, Watson IA, Raub TJ. Integration of in silico and in vitro tools for scaffold optimization during drug discovery: predicting P-glycoprotein efflux. Mol Pharm 2013;10:1249–61.
29. Kerns EH, Di L. Drug-like properties concepts, structure design and methods: from ADME to toxicity optimization. Amsterdam: Elsevier; 2008.
30. Norinder U, Haeberlein M. Computational approaches to the prediction of the blood-brain distribution. Adv Drug Deliv Rev 2002;54:291–313.
31. Raub, TJ, Lutzke, BS, Andrus, PK, et al. Early preclinical evaluation of brain exposure in support of hit identification and lead optimization. In: Borchardt RT, Kerns EH, Hageman MJ, Thakker DR, Stevens JL, eds. Optimizing the ‘drug-like’ properties of leads in drug discovery, 1st ed. New York: Springer; 2006:355–410.
32. Hitchcock SA, Pennington LD. Structure-brain exposure relationships. J Med Chem 2006;49:7559–83.
33. Warr WA. Some trends in chem(o)informatics. In: Bajorath J, ed. Methods in molecular biology. New York: Springer; 2011:1–37.
34. Yee LC, Wei YC. Current modeling methods used in QSAR/ QSPR. In: Dehmer M, Varmuza K, Bonchev D, eds. Statistical modelling of molecular descriptors in QSAR/QSPR. Weinheim: Wiley-VCH; 2012:1–31.
35. Zang Q, Rotroff DM, Judson RS. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. J Chem Inf Model 2013;53:3244–61.
36. Singh N, Chaudhury S, Liu R, et al. QSAR classification model for antibacterial compounds and its use in virtual screening. J Chem Inf Model 2012;52:2559–69.
37. Riniker S, Wang Y, Jenkins JL, Landrum GA. Using information from historical high-throughput screens to predict active compounds. J Chem Inf Model 2014;54:1880–91.
38. Iwata H, Sawada R, Mizutani S, Yamanishi Y. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. J Chem Inf Model 2015;55:446–59.
39. Ertl P. Polar surface area. In: Mannhold R, ed. Molecular drug properties, measurement and prediction. Weinheim: Wiley-VCH; 2008:111–26.
40. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 2000;43:3714–17.
41. Martins IF, Teixeira AL, Pinheiro L, Falcao AO. A Bayesian approach to in silico blood-brain barrier penetration modelling. J Chem Inf Model 2012;52:1686–97.
42. B3Info web site. Available from: http://b3info.lasige.di.fc.ul.pt [last accessed 2012].