Received: 10 October 2017 | Revised: 28 November 2017 | Accepted: 1 December 2017

DOI: 10.1002/cem.2996

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Modeling of Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities: A nonparametric comparison of models for prediction of drug absorption sites

Darija Obradović[1] | Filip Andrić[2] | Mario Zlatović[2] | Danica Agbaba[1]

[1] Faculty of Pharmacy, University of Belgrade, Vojvode Stepe 450, 11000 Belgrade, Serbia

[2] Faculty of Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia

Correspondence
Danica Agbaba, Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Belgrade.
Email: danica@pharmacy.bg.ac.rs

Funding information
Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant/Award Numbers: 172033 and 172017

Abstract

Aripiprazole and ziprasidone are atypical antipsychotic drugs with the effect on positive and negative symptoms of schizophrenia, mania, and mixed states of bipolar disorder. Hansen's solubility parameters, δ_d, δ_p, and δ_h, which account for dispersive, polarizable, and hydrogen bonding contributions to the overall cohesive energy of a compound, are often used to assess pharmacokinetic properties of drugs. However, no data exist of solubility parameters for the drugs of interest in this study. Therefore, in the present study, partial least square regression (PLS), artificial neural networks (ANNs), regression trees (RT), boosted trees (BT), and random forests (RF) were applied to estimate Hansen's solubility parameters of ziprasidone, aripiprazole, and their impurities/metabolic derivatives, targeting their biopharmaceutical classes and absorption routes. A training set of 47 structurally diverse and pharmacologically active compounds and 290 molecular descriptors and pharmaceutically important properties were used to build the prediction models. The modeling approaches were compared by the sum of ranking differences, using the consensus values as a reference for the unknowns and the experimentally determined values as a gold standard for the calibration set. In both instances, the PLS models, together with ANNs, demonstrated better performance than RT, BT and especially RF. Based on the best scored models, we were able to pinpoint the most probable absorption sites for each drug and the corresponding metabolite, i.e., the upper parts of the gastrointestinal tract, small intestine, or absorption along entire length of gastrointestinal tract.

KEYWORDS

artificial neural network (ANN), drug absorption, Hansen's solubility parameters, partial least square regression (PLS), sum of ranking difference (SRD)

1 | INTRODUCTION

Aripiprazole and ziprasidone are piperazine derivatives, atypical antipsychotic drugs with an effect on positive and negative symptoms of schizophrenia, mania, and mixed state of bipolar disorder.[1] Interesting pharmacological profile of investigated compounds is associated with their binding to serotonin and dopamine receptors.[2-5] Orally administered ziprasidone of 80 to 160 mg/day with the food was a clinically valuable treatment option for stable patients with schizophrenia or schizoaffective disorder, who experienced suboptimal efficacy or poor tolerability with haloperidol, olanzapine, or risperidone.[2] Also, the oral rout of aripiprazole shows effective and safe characteristics with patients who are nonresponsive to standard antidepressants.[1]

In the case of oral rout of administration, the delivery by the bloodstream to the site of action is related to the process of absorption of the drug in the gastrointestinal tract (GIT), and in this case, it may result in lesser availability and bloodstream concentration fluctuations, due to incomplete absorption. When a tablet or capsule is swallowed, it must dissolve before it can be absorbed. Highly water-soluble medications dissolve more readily in the GIT, while the fat-soluble drugs dissolve slower. Tablets that dissolve too early are problematic, because they give bad taste and are difficult to swallow. Special formulations or coatings can be used to delay dissolution, thereby protecting the drug from stomach acid, or allowing gradual release of the drug to intentionally lengthen the absorption process. These are referred to as the delayed or sustained release formulations.[6]

With an aim to predict the in vivo pharmacokinetic properties of drugs, the Food and Drug Administration set up the Biopharmaceutical Classification System (BSC), which is based on measuring permeability and solubility. Biopharmaceutical Classification System classifies drugs into 4 classes and 2 groups[7]. Class 1 is high solubility and high permeability, Class 2 is low solubility and low permeability, Class 3 is high solubility and low permeability, and Class 4 is low solubility and low permeability. Group 1 belongs to Class 1, which includes drugs that rapid absorb along the first part of jejunum. Contrary to Class 1, Group 2 can belong to Classes 2 to 4, characterizing with an incomplete absorption when passing GIT.[7]

When considering the physicochemical characteristics of the drugs, an important impact on the absorption process is exerted by the potential of the H-bond formation.[8,9] When the molecule of a drug passes through the membranes, it has to disrupt hydrogen bonds with its aqueous environment. Building strong H-bonds requires more energy, and consequently, more energy is then needed for their disruption. Thus, high capacity of building the H-bond connections is not a desired property, and it is often associated with poor permeability and poor absorption.[10] By calculating the Hansen's solubility parameters of the investigated compounds based on the H-bond potential (δ_h) and the δ_v parameter (a combined influence of the dispersive [δ_d] and polar forces [δ_p]), it is possible to determine location and duration of the absorption process in GIT.[8,11-13]

According to the BSC criteria, ziprasidone[14] and aripiprazole[15] belong to Class 2[7,16] and Group 2, while for their impurities, there are no data available in the literature (Figure 1). In the finished formulation (tablet) taken by the patients, one can find the impurities of the production and degradation process.

Instability of ziprasidone is almost a consequence of reactivity of the alpha position of the benzoxindol moiety in the molecule of ziprasidone. The methylene moiety next to the lactam carbonyl is susceptible to nucleophiles, and it can easily form an oxidative degradant (Imp Z2; Figure 1). Imp Z2 can be involved in the reaction of aldol condensation with the molecule of ziprasidone, when a new degradant is formed[17] (Imp Z3; Figure 1). A degradant that is formed at ambient temperature under the influence of the daylight in the solid-state ziprasidone is a product of the side reaction of benzisothiazole at the alpha position of the benzoxindol moiety of ziprasidone (Imp Z5; Figure 1). Imp Z6 is the dehydration product of Imp Z3, while Imp Z7 is the result of the opening of the benzisothiazole ring.

Aripiprazole is synthesized by coupling 1-(2,3-dichlorophenyl)-piperazine or its derivative with the compounds that could be regarded as 7-derivatives of 3,4-dihydroquinolin-2(1*H*)-one. The synthetic route and the type of substituent at position 7 of 3,4 dihydroquinolon-2(1*H*)-one allow formation of several structurally related impurities such, as Imp A1, Imp A2, Imp A3, and Imp A4. Depending on the reaction conditions applied in the course of the synthesis, the additional side reactions can occur and the consecutive byproduct impurities can be expected, including Imp A5, Imp A6 and Imp A7, and the degradant Imp A8 (Figure 1).[18]

The preliminary studies have shown that piperazine derivatives that include aripiprazole and ziprasidone are the compounds that can be absorbed from the GIT.[19] However, no data exist suggesting from which part of the GIT the main components and especially their impurities and the degradation products are absorbed. Within the scope of the present study, a series of such impurities have been targeted. For the investigated compounds, there are also no data available of

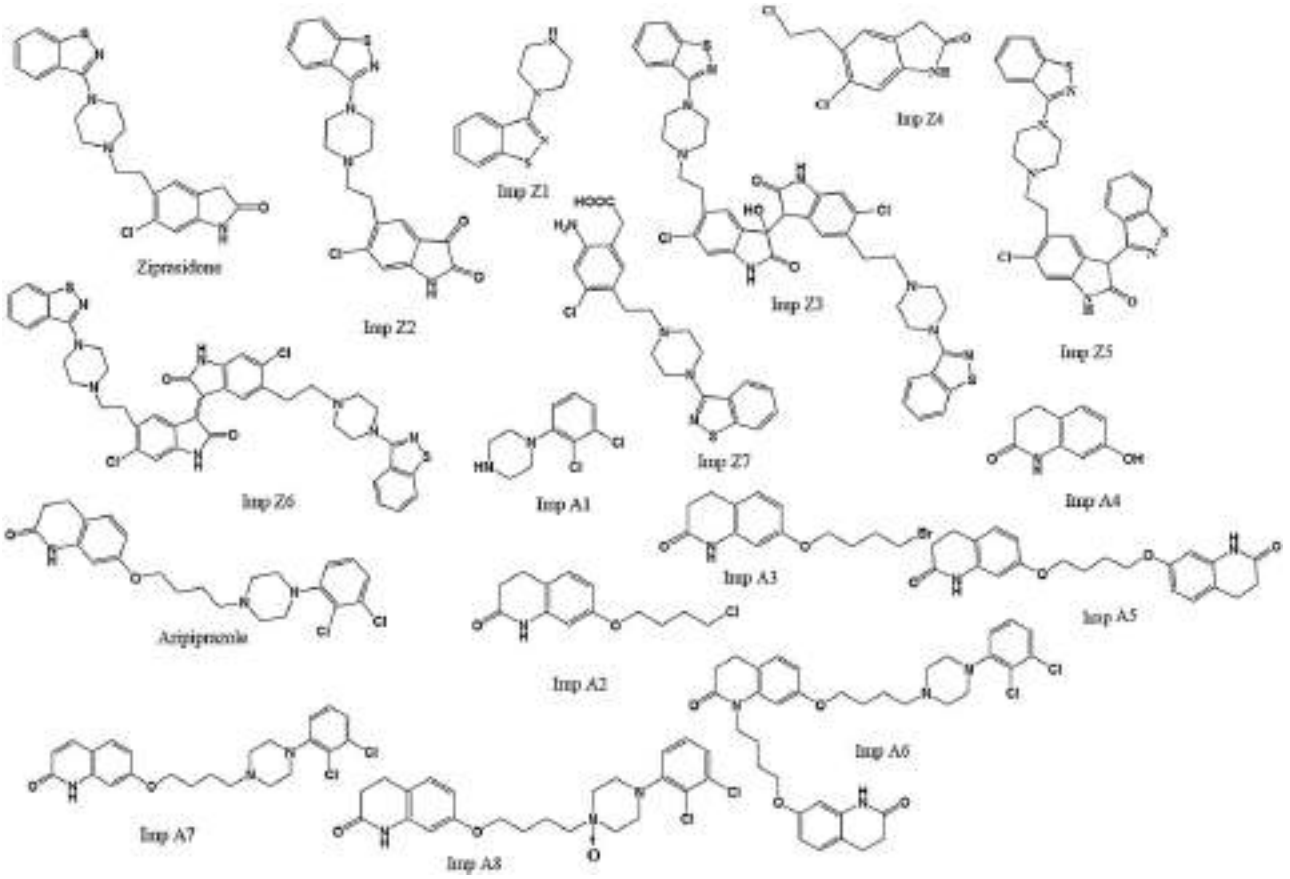

FIGURE 1 Structures of the target molecules: aripiprazole, ziprasidone, and their impurities

the Hansen's solubility parameters. The aim of the present study was to estimate the Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities and to pinpoint the most probable sites of their absorption.

2 | METHODS

2.1 | Selection of the training compounds set, structural optimization, and computation of molecular descriptors

The training set was composed of 47 carefully selected and pharmaceutically important substances (Table 1) of various molecular shapes, sizes, and abilities to establish specific interactions (the hydrogen bond donating, the hydrogen-bond accepting, dipolar, and polarizable interactions). Absorption from different parts of GIT has been entirely covered by the training set. All molecular structures were built by using Maestro (Maestro, version 10.7, Schrödinger, LLC, New York, 2016). Calculations of physically significant molecular descriptors and pharmaceutically relevant properties were performed by using QikProp (QikProp, version 4.9, Schrödinger, LLC, New York, 2016). The single point calculations using the RM1 method from the semiempirical neglect of diatomic differential overlap module of Schrodinger Suite 2016-3 (semiempirical neglect of diatomic differential overlap driver, version 3.5, Schrödinger, LLC, New York, 2016) were performed with semiempirical parameters. In total, 298 2D molecular descriptors were calculated; 198 reflecting molecular topology, size, branching, and shape such as valence connectivity indices (χ 0-5), Gutman, Zagreb, and Wiener topological indices; 27 physicochemical properties, such as human oral absorption, variously estimated octanol-water partition coefficients, solvent accessible surface area, and hydrogen bond donating and accepting properties; and 26 descriptors obtained from molecular-orbital computations, e.g., total energy, ionization potential, core-core repulsion, dipole, and electron affinity. The rest of the descriptors were related to different structural or physicochemical properties. The calculated descriptors are given in the Supporting Information (Data sheets 1). The new machine-learning application (AutoQSAR) for the validation and deployment of the QSAR models built in Schrödinger was tested in 2

TABLE 1 Compound training set

No.	Compound	δ_d	δ_p	δ_h	Method	Ref.	No.	Compound	δ_d	δ_p	δ_h	Method	Ref.
1	Acetaminophen	21.13	8.62	15.61	CGM	13	25	Indomethacin	23.06	5.98	9.42	CGM	13
2	Acyclovir	21.80	14.64	24.06	CGM	13	26	Ipsapirone	21.00	7.86	11.31	CGM	13
3	Allopurinol	25.63	23.68	25.19	CGM	13	27	Isosorbide-mononitrate	20.87	13.37	17.00	CGM	13
4	Amoxicillin	23.13	7.38	15.65	CGM	13	28	Levodopa	22.34	6.26	21.86	CGM	13
5	Ampicillin	21.98	6.69	11.70	CGM	13	29	Metacycline	25.61	8.84	22.66	CGM	13
6	Betamethasone	20.50	5.63	16.18	CGM	13	30	Methaqualone	21.62	7.34	8.01	CGM	13
7	Cyclosporine	18.83	3.25	14.10	CGM	13	31	Metoprolol	18.31	3.68	11.52	CGM	13
8	Ciprofloxacin	22.66	7.30	10.97	CGM	13	32	Metronidazole	19.86	13.94	16.86	CGM	13
9	Digitoxin	20.73	3.69	17.61	CGM	13	33	Nifedipine	19.61	5.15	8.59	CGM	13
10	Dicoumarol	26.25	5.57	17.83	CGM	13	34	Nitrofurantoin	22.14	15.45	16.64	CGM	13
11	Dilazep	19.79	3.97	9.81	CGM	13	35	Nimodipine	19.00	4.36	8.30	CGM	13
12	Diltiazem	20.44	4.86	8.40	CGM	13	36	Nisoldipine	19.00	4.31	7.87	CGM	13
13	Doxycycline	24.70	8.49	22.20	CGM	13	37	Oxytetracycline	25.85	9.08	24.64	CGM	13
14	Erythromycin	18.09	3.35	15.65	CGM	13	38	Oxprenolol	18.32	3.87	11.34	CGM	13
15	Phenylbutazone	20.91	6.41	9.85	CGM	13	39	Piretanide	22.48	6.75	12.18	CGM	13
16	Phenytoin	22.80	6.69	7.74	CGM	13	40	Pyridoxal-5-phosphate	24.79	19.47	16.97	CGM	13
17	Phenoxymethylpenicillin	21.61	6.54	10.05	CGM	13	41	Propranolol	19.57	3.35	11.04	CGM	13
18	Furosemide	23.83	8.00	13.35	CGM	13	42	Prednisolone	20.63	5.59	16.65	CGM	13
19	Glibenclamide	21.44	5.29	8.87	CGM	13	43	Riboflavin	23.75	9.65	22.25	CGM	13
20	Hydrochlorothiazide	23.86	10.07	13.82	CGM	13	44	Sulfametoxydiazine	21.70	9.50	13.55	CGM	13
21	Quinidine	20.72	5.53	11.97	CGM	13	45	Theophylline	17.80	12.85	12.65	CGM	13
22	Chloramphenicol	23.06	9.50	18.68	CGM	13	46	Tetracycline	24.99	8.53	22.25	CGM	13
23	Ibuprofen	16.60	6.91	9.97	EHSA	21	47	Sulfinpyrazone	23.44	7.30	10.79	CGM	13
24	Ibuprofen lysinate	16.97	22.75	12.83	IGC	21							

The reference values of Hansen's solubility parameters were retrieved from Terada and Marchessault and Hansen.[12,20] Hansen's solubility parameter values for ibuprofen and ibuprofen lysinate were determined by extended Hansen solubility approach (EHSA) and inversion gas chromatography (IGC), respectively. For the rest of compounds, contribution group method (CGM) was used for estimation.

independent runs. The tool has becoming increasingly popular in molecular modeling community, especially among those not specialized in statistical modeling. It should provide reliable predictions by combining robust linear modeling, cross-validation, validation, ranking, and selection of the best models among many linear models obtained by variations of partial least square regression (PLS) and multiple linear regression. Naturally, we were highly motivated to include the AutoQSAR in the present study.

The reference values of Hansen solubility parameters are taken from the literature[12,20] and are included in Table 1, along with the estimation and determination methods. Practically, experimentally obtained data were provided only in the case of ibuprofen and ibuprofen lysinate, using the extended Hansen's solubility approach and the inverse gas chromatography, respectively.[20] For the rest of compounds, the values are obtained by a reliable fragmentation method.[12] The values of the Hansen's solubility parameters are also given in the Supporting Information (Data sheets 1).

2.2 | Building the models

Partial least square regression, artificial neural networks (ANNs), regression trees (RTs), random forests (RFs), and boosted trees (BTs) were used for modeling of the Hansen's solubility parameters. Compounds were randomly divided into the training ($n = 36$) and the test set ($n = 11$). Depending on the regression method, the training set was further split into several cross-validation subsets (in the case of PLS and RT), or the single validation set was randomly selected

($n = 8$, in the case of ANN-s, BT, and RF). The validation/cross-validation subsets were used to find the optimal complexity of the models, while the test set was used for an assessment of their prediction abilities. The root mean squared errors and R^2 values of the calibration (cal), validation/cross-validation (val/cv), and prediction (pred) points were used to assess the models' performances.

Prior to any modeling, categorical and ordinal variables, as well as molecular descriptors with low variability (relative standard deviation <10%), were removed from the dataset. In this way, the starting pool of 296 descriptors was initially reduced to 218 variables. To improve model performances, an additional variable selection was performed, if and when necessary.

Partial least square regression was carried out on the standardized data (mean centered and divided by standard deviation), using the SIMPLS algorithm (PLS Tool Box v. 7.02 for MATLAB R2011). The models of optimal complexity (number of latent variables) were selected, based on the cross-validation experiments with 5 or 4 splits of the training set by the venetian-blind resampling. An abnormal behavior of the cross-validation error versus the number of the PLS components was a strong indicator of the presence of influential points. After removal of the outliers, the models were further improved by the stepwise variable selection. All variables with the variable importance to the projection scores higher than 1 were retained. The procedure was repeated, until no improvement in the model performance was achieved.

The data without any pretreatment were used to build the ANN models. The feed-forward multiple perceptron layer (MLP) networks were employed, with the Broyden-Fletcher-Goldfarb-Shanno learning algorithm. The network training was done by applying the self-training function as a part of Statistica 10.0 software (StatSoft Inc.). The number of perceptrons in the hidden layer was determined by trial and error, although the predefined range was set to $n = 8$ to 25, following the rule $n = N^{1/2} + 10$, where N is the number of the descriptors involved in training. Basically, the networks of optimal architecture (topology) were selected of 100 networks trained per each Hansen parameter. Such a high number of trained networks, with random values of initial parameters, were used to achieve global model optimum and avoid local ones. The following activation functions were varied during network construction: identity (ident), exponential (exp), logistic (log), and hyperbolic tangent (tanh). Each of the developed networks was internally and externally validated by using randomly selected subsets of compounds. Mean squared residual error was used as a performance function. Also, architectures with highest determination coefficient for training and validation were selected as the final ones (the best 5). Practically, internal and external validations were necessary to avoid the problem of overfitting. For that purpose, the complete set of compounds was divided into approximately 60% of training and 20% of each—internal and external validation subsets. In that way, compounds 2, 16, 17, 26, 32, 35, 37, and 42 were randomly selected as internal validation set, while the prediction performance of ANNs was assessed by randomly selected external set of compounds: 5, 8, 13, 14, 19, 22, 23, 31, 36, and 38. The rest of the compounds were used for calibration (training).

The RT models were built by using (a) single trees, here denoted plainly as the RTs, (b) BTs, and (c) RFs, all parts of the tree/partitioning modules in Statistica 10.0 (StatSoft Inc.). In the case of RT, for each solubility, parameters about 12 to 16 trees of different sizes were built. The trees of the maximum size were pruned on variance, with 5 objects in the final nodes used as the stopping criteria. The 10-fold cross-validation was used to select the tree of optimal complexity. The minimum of the cross-validation cost function (a sort of trade-off between the tree size and the prediction error) was used as a criterion.

In the case of the RF models, the number of trees in ensembles was in between 70 and 100. The stopping criteria for the tree selections were the maximum size of 100, the minimum number of cases in the child and terminal nodes $n = 5$, and the maximum number of levels $l = 4$. The stopping criteria for boosting trees were a bit stricter, with the maximum number of nodes $n = 3$, to produce small, weak learners able to perform boosting through the modeling of the residual error of a preceding tree.

2.3 | Sum of ranking differences

Sum of ranking difference (SRD) has recently emerged as a robust nonparametric method for a comparison of the methods, models,[21-23] and fusion of multiple criteria.[24-27] The method is entirely general. Sum of ranking difference requires the data to be arranged in a matrix in such a way that the methods or models to be compared are placed in columns, while the objects (in this case, the compounds) are arranged in the rows. Then, a reference column is added to the matrix. It can be the row-wise maxima, minima, the average, or the golden reference standard vector. Here, 2 kinds of the SRD analysis were performed, one using the row-wise average of the experimentally determined Hansen's

solubility parameters (in the case of the standard set of compounds), and the second the row-wise average (in the case of the compounds with unknown experimental values). Then, for each column, the values are ranked in an ascending order by using the arithmetic mean for the tied values, and the ranks are subtracted from the reference. The rank differences are summed up resulting in SRD score associated with each column. In the case when the results of multiple SRDs are to be compared, the SRD scores of each analysis are rescaled to the maximum value (SRDnorm), according to Equation 1.

$$\mathrm{SRDnorm} = 100 \times \mathrm{SRD}/\mathrm{SRDmax} \tag{1}$$

The smaller the SRD scores, the closer the model to the reference. The results of the SRD analysis can be validated in 2 ways, and in this study, only the randomization test was used. The purpose of the test is to distinguish statistically significant SRD values from the random ones. Random distribution of the SRD scores is produced, and if the SRD score falls within the 90% confidence limits of the bell-shaped random distribution curve, then the model is not able to rank the objects (compounds) better than by chance, implying that the results of such model are statistically insignificant (random), compared with the reference (cf. to the Figure 3A-C lower subplots).

The SRD toolkit was provided in a form of the Microsoft Excel visual basic macros freely available from http://aki.ttk.mta.hu/srd/.

2.4 | Classification of the absorption sites of target drugs based on Hansen's solubility approach

Classification of the absorption sites of the drugs based on the Hansen's solubility approach is related to the cohesive energy density (CED) method.[28] The term "solubility parameter" was first used by Hildebrand and Scott.[29] Hildebrand described the total solubility parameter (δ_t) of the nonpolar substances as the square root of CED:

$$\delta_t = \mathrm{CED}^{1/2} = (\Delta E/V_m)^{1/2} \tag{2}$$

where ΔE represents the liquid cohesion energy, which is divided by the molar volume, V_m ($cm^3\ mol^{-1}$)[9].

Hansen extended application of this concept to the polar systems, where the total cohesion energy, E, is the sum of individual energies that make it up:

$$E = E_d + E_p + E_h \tag{3}$$

Dividing E by the molar volume of a compound of interest (V_m) results in the square of the total solubility parameter (δ_t^2), which is the sum of the squares of the Hansen's components, δ_d, δ_p, and δ_h:

$$E/V_m = E_d/V_m + E_p/V_m + E_h/V_m \tag{4}$$

$$\delta_t{}^2 = \delta_d{}^2 + \delta_p{}^2 + \delta_h{}^2 \tag{5}$$

where δ_d, δ_p, and δ_h describe the dispersive, the dipole, and the hydrogen bonding interactions, respectively.[29]

Bagley performed a projection of the 3-dimensional solubility parameter space onto the 2-dimensional plot characterized by the association interaction of the dispersion and the polarization forces, and he introduced the volume-dependent solubility parameter, δ_v, defined by the following equation[8]:

$$\delta_v = \left(\delta_d{}^2 + \delta_p{}^2\right)^{1/2} \tag{6}$$

Transfer of the calculated 3-dimensional solubility parameters into the Bagley diagram classifies gastrointestinal absorption sites according to the following criteria: Group 1, drugs that are only absorbed from the upper parts of the GIT, with the shortest absorption time ($\delta_h > 17$ $[J/cm^3]^{1/2}$); Group 2, drugs that are preferably absorbed from the upper parts of the small intestine but to a lower extent from the other sites also; Group 3, drugs with the longest absorption time that are absorbed from along the whole GIT, or even better from cecum or colon than from small intestine ($\delta_v = 20 \pm 2.5$ $[J/cm^3]^{1/2}$, $\delta_h = 11 \pm 3$ $[J/cm^3]^{1/2}$).[12,30]

3 | RESULTS AND DISCUSSION

3.1 | Predictive performance of models and nonparametric model selection

Generally, the PLS and ANN models showed somewhat better performance compared with RTs, BTs, or RFs, with R^2pred values ranging between 0.75 and 0.95, versus R^2pred for BTs 0.40 to 0.50 (Table 2). In the case of δ_d, the BT model was obtained with extremely poor predictability R^2pred = 0.01, but with good calibration performance. The prediction parameters remain in agreement with the ANN models already published in the literature.[31,32] Járvás et al[32] obtained the ANN models based on 14σ COSMO moments with the mean absolute errors of 1.09, 1.70, and 1.96 for δ_d, δ_p, and

TABLE 2 Statistical performance parameters of the obtained models

Models	Statistics, Model Complexity, and Other Details	Hansen's Solubility Parameter		
		δ_d	δ_p	δ_h
PLS	RMSE (cal; CV; pred) R^2 (cal; CV; pred) Complexity Excluded compounds	0.30; 0.83; 0.91 0.978; 0.833; 0.896 n(LV) = 5 7, 10, 40, 45	0.84; 1.58; 1.08 0.931; 0.756; 0.771 n(LV) = 3 23, 24, 40	1.71, 2.23, 2.05 0.892, 0.810; 0.752; n(LV) = 2 8, 10, 22, 45
ANN1	RMSE (cal; val; pred) R^2 (cal; val; pred) Complexity and structure Excluded compounds	0.11; 0.46; 1.19 0.977; 0.880; 0.802; (210-16-1), log-log 10, 25, 45	<0.01; 0.42; 0.62 1.000; 0.970; 0.878 (211-14-1), log-ident 23, 24, 48	0.37; 2.27; 2.85 0.984; 0.956; 0.831 (211-18-1), tanh-log 10
ANN2	RMSE (cal; val; pred) R^2 (cal; val; pred) Complexity and structure Excluded compounds	0.02; 0.36; 0.42 0.996; 0.914; 0.940 (210-14-1), log-log 10, 25, 45	<0.01; 0.50; 0.41 1.000; 0.976; 0.907 (211-14-1), exp-ident 23, 24, 48	0.27; 2.36; 3.96 0.988; 0.956; 0.764 (211-20-1), log-log 10
ANN3	RMSE (cal; val; pred) R^2 (cal; val; pred) Complexity and structure Excluded compounds	0.11; 0.52; 0.72 0.983; 0.877; 0.890 (210-14-1), exp-log 10, 25, 45	<0.01; 0.33; 0.51 1.000; 0.965; 0.928 (211-14-1), exp-ident 23, 24, 48	0.37; 2.35; 2.50 0.984; 0.960; 0.856 (211-15-1), tanh-tanh 10
ANN4	RMSE (cal; val; pred) R^2 (cal; val; pred) Complexity and structure Excluded compounds	0.04; 0.39; 0.88 0.991; 0.896; 869; (210-14-1), log-log 10, 25, 45	<0.01; 0.35; 0.58 1.000; 0.976; 0.890 (211-14-1), exp-tanh 23, 24, 48	0.21; 2.54; 3.47 0.991; 0.963; 0.789 (211-23-1), log-log 10
ANN5	RMSE (cal; val; pred) R^2 (cal; val; pred) Complexity and structure Excluded compounds	0.15; 0.57; 1.09 0.968; 0.865; 0.824 (210-14-1), log-log 10, 25, 45	0.03; 0.71; 0.58 0.999; 0.965; 0.856 (211-14-1), exp-tanh 23, 24, 48	0.01; 1.96; 1.89 1.000; 0.967; 0.891 (211-14-1), tanh-log 10
RT	RMSE (cal) R^2 (cal) Complexity Excluded compounds	1.22 0.747 TS = 3; cost cut-off = 5.23 16, 32, 36	2.83 0.640 TS = 2; cost cut-off = 12.51 1	3.27 0.510 TS = 2; cost cut-off = 18.48 2, 19, 20, 37, 44
BT	RMSE (cal; pred) R^2 (cal; pred) Complexity Excluded compounds	0.625; 2.58 0.941; 0.08 NT = 187, *TS* = 5; 16, 32, 36	1.421; 2.230 0.940; 0.522 NT = 70, TS = 3 1	1.284; 4.34 0.931; 0.420 NT = 192, TS = 4 2, 19, 20, 37, 44
RF	RMSE (cal; pred) R^2 (cal; pred) Complexity Excluded compounds	2.15; 2.37 0.587; 0.193 NT = 70; 16, 32, 36	5.01; 3.04 0.551; 0.521 NT = 100 1	4.16; 3.03 0.387; 0.420 NT = 100 2, 19, 20, 37, 44

The artificial neural network (ANN) structure is given as *i*-*h*-*o*, where *i*, *h*, and *o* are numbers of input, hidden, and output neurons, followed by the activation function in hidden and output layers (logistic, tanh, exponential, and identity); regression tree (RT) structure is described by tree size (TS); boosted tree model structure is defined by number of trees (NT) and TS; random forest complexity is described by NT included in the model; for stopping conditions and TS, see section 2.2.

δ_h, respectively. In our case, the lowest errors were 0.42, 0.41, and 1.89. In all cases, a removal of small number of compounds ($1 < n < 5$) from the training set significantly improved the performance of the models. A consistency in the compound outlying effect is observed with each Hansen's solubility parameter, although we could not find a reasonable structural or physicochemical explanation for such a behavior of the compounds considered. The remaining substances resulted in good models with residuals following normal distribution. Detailed reports on the scrutinized models can be found in the Supporting Information (Data sheets 2-4).

To select the best models, we decided to use the nonparametric SRD comparison, because of its robustness, no requirements for normal data distribution, but most importantly, because the 2 main criteria that we imposed on the selection of the preferential models are easily implemented in the SRD. Namely, the first criterion was that the preferable training model should provide the values (obtained from the predicted or the CV subsets) as close as possible to the reference input data (the lowest SRD scores). The second condition imposes that such a model applied to the unknowns should result in the predictions as close as possible to the consensus of the values defined by all compared models. Essentially, in a consensus-based comparison, the systematic and random errors associated with each model should be at least partially eliminated, resulting in the consensus estimates that are better than any estimate based on a single model.

In the case of δ_d, the lowest SRD value was obtained for the ANN2 model followed by PLS and the rest of the ANN models, if comparison is done with the known reference values. In a consensus-based comparison of the unknowns, the best ranked is PLS (Figure 2A), closely followed by ANN5, ANN3, ANN4, and ANN1. Therefore, the preferential values of δ_d were calculated by the consensus of PLS and the ANNs.

In the case of a comparison of δ_h with the reference values, the best ranked model is ANN5, closely followed by the rest of the ANN models. The consensus-based comparison of the unknowns yields PLS as the best option, which is

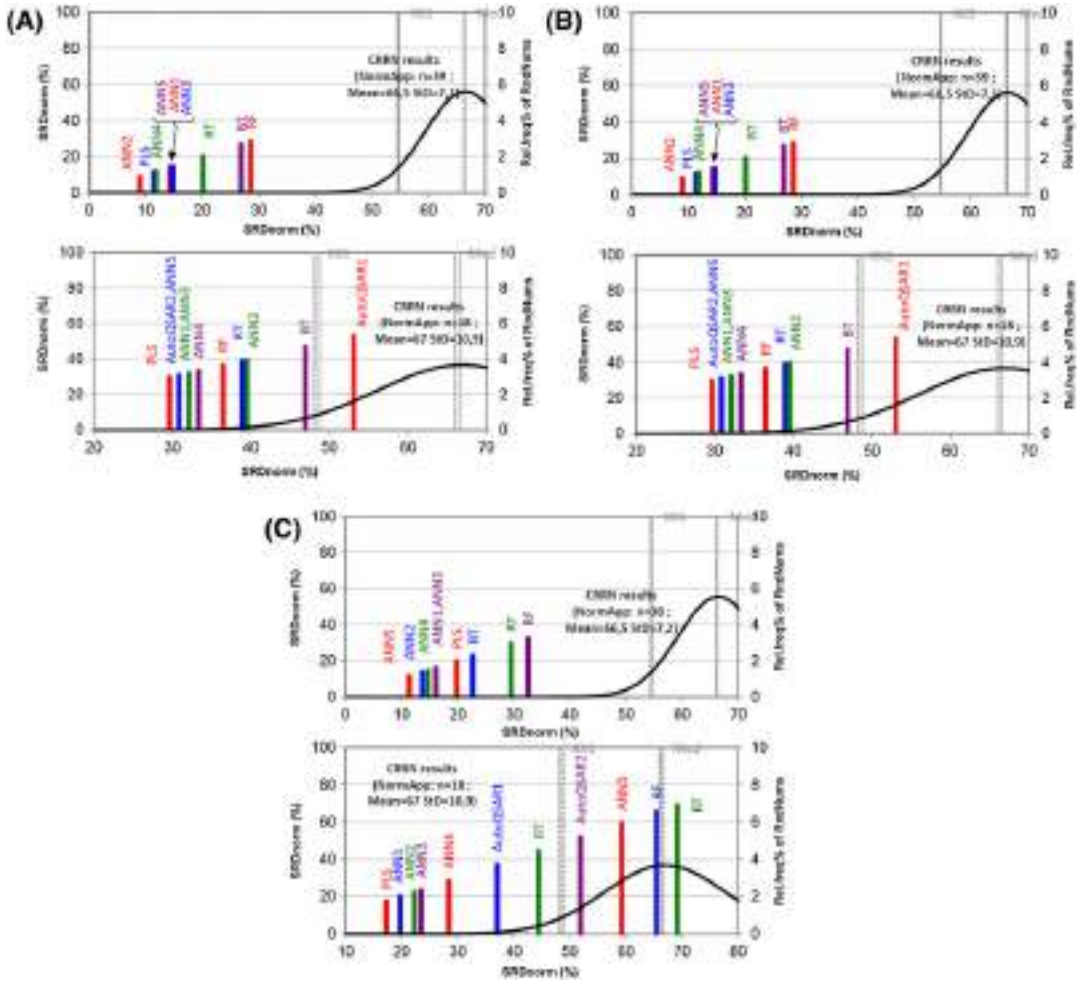

FIGURE 2 Sum of ranking difference (SRD) ranking of models for prediction of Hansen's solubility parameters: (A) δ_d, (B) δ_p, and (C) δ_h; the upper subplots correspond to comparison with the reference values; the lower subplots correspond to consensus ranking; normalized SRD values (%) are on the *x*-axis and left side *y*-axis; right-side *y*-axis represents relative frequencies of random numbers (%)

closely followed by ANN1 to ANN3 (Figure 2B). All of them are equally suitable, although ANN2 demonstrates a slightly better statistical performance (Table 2), and therefore, it was selected as a preferable model.

In the case of δ_p, a comparison of SRD with the reference values points out to ANN1 to ANN4 as to the best models, while the consensus-based SRD of the unknowns identifies ANN1 and ANN3 at the second position (Figure 2C). Considering slightly better performance parameters of ANN3 over the other ones (Table 2), this model was selected as the preferential one.

In most cases, the RFs and BTs were ranked as the worst, while the general RTs demonstrated an intermediate to low proficiency. Considering the small-sized trees as slow learners, such outperformance of the general regression tress over the boosted or random ensembles was not expected. The AutoQSAR models also resulted in the overall intermediate scores.

3.2 | Prediction of the absorption sites for target molecules based on the best estimates of the Hansen's solubility parameters

The Hansen's solubility parameters have been calculated based on the best ranked models in the SRD comparison, as described in section 3.1. The values are summarized in Table 3.

According to the BSC classification and following the Bagley's grouping criteria, all target compounds have been assigned to Groups 2 and 3 (Figure 3). As described in the literature (BSC), ziprasidone[14] and aripiprazole[15] have the absorption profiles of Group 2 (Class 2)[7,16] as compounds characterized by traveling through GIT, which remains in agreement with our prediction profile. Aripiprazole belongs to Group 3 as a compound that travels along an entire GIT with the duration of the absorption process higher than 10 hours. However, ziprasidone and all of its impurities are classified as the Group 2 members, ie, as the compounds readily absorbed from the upper parts of small intestine, with the absorption lasting between 4 and 9 hours. All aripiprazole impurities, with an exception of Imp A4, belong to Group 3.

TABLE 3 The estimates of Hansen's solubility parameters based on the best ranked models by the SRD comparison

No.	Compound	Hansen's Solubility Parameter			Bagley's Combined Volume Term
		δ_d	δ_p	δ_h	δ_v
1	Ziprasidone	21.61	6.04	8.54	22.43
2	Imp Z1	21.79	10.49	11.65	24.19
3	Imp Z2	22.22	6.09	8.82	23.04
4	Imp Z3	22.14	4.90	8.16	22.68
5	Imp Z4	21.49	10.49	8.80	23.91
6	Imp Z5	22.04	7.16	8.46	23.18
7	Imp Z6	22.35	10.37	8.59	24.64
8	Imp Z7	22.38	6.33	8.28	23.26
9	Imp Z8	21.65	6.21	9.87	22.52
10	Aripiprazole	20.94	5.43	10.04	21.63
11	Imp A1	20.76	8.64	9.30	22.49
12	Imp A2	20.84	7.40	10.29	22.12
13	Imp A3	20.85	7.53	10.03	22.17
14	Imp A4	22.04	7.83	14.16	23.39
15	Imp A5	21.12	4.25	10.35	21.54
16	Imp A6	20.97	4.99	8.77	21.56
17	Imp A7	20.68	4.70	9.29	21.21
18	Imp A8	20.97	6.55	8.99	21.97

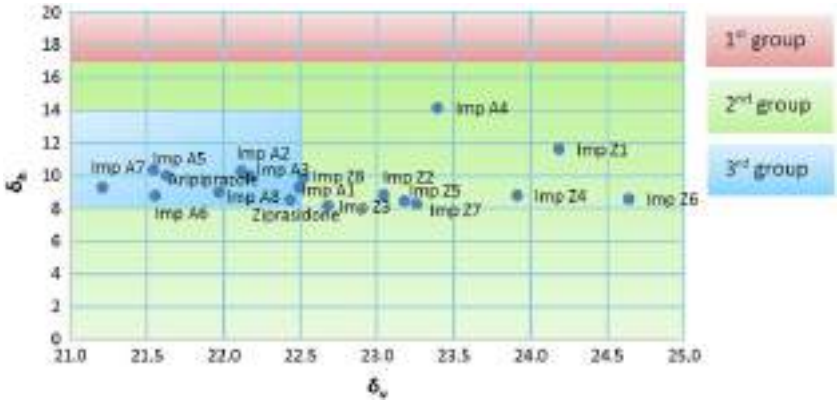

FIGURE 3 Bagley's classification diagram for ziprasidone, aripiprazole, and their impurities; the first group corresponds to the drugs absorbed from the upper gastrointestinal tract (GIT), the second group compiles compounds absorbed mainly from small intestine (absorption duration between 4 and 9 h), and the third group involves substances absorbed along an entire GIT (absorption duration >10 h)

The pharmaceutical regulatory agencies such as the Food and Drug Administration and the European Medicines Agency have raised the concerns regarding the presence of genotoxic impurities in the APIs that could exert a negative impact on human health. The term "genotoxicity" covers a wider range of genetic damages, regardless if such damage is or is not corrected through the cell DNA-repairing mechanism.[33] It is worth mentioning that the ziprasidone impurity, Imp Z4, possesses genotoxic potential due to the presence of an alkylating group (ethyl chloride).[34] From the absorption prediction model developed in this study, it comes out that the absorption of Imp Z4 takes place in small intestine and it lasts between 4 and 9 hours.

4 | CONCLUSION

Modeling of the Hansen's solubility parameters by linear and nonlinear approaches resulted in the well-established PLS and the artificial neural networks (ANNs) models. In the SRD comparison, the ANN and the PLS models scored the best, while RTs, BTs, and RFs performed significantly worse. The AutoQSAR, a new machine learning module available from the Schrödinger LLC, demonstrated an intermediate performance. Based on the Hansen's parameters predicted from the best scored models, the most probable absorption sites for each drug and a corresponding degradation product were estimated. While aripiprazole and all its impurities with an exception of one are absorbed along an entire length of the GIT, ziprasidone and all ziprasidone impurities are predicted to be absorbed in an upper part of small intestine. Statistical performance of the best ranked models proved to remain in agreement with similar models based on the quantum-mechanical DFT computations and the ANN modeling.

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia grant nos. 172017 and 172033. F.A. would like to thank Dr. Dávid Bajusz for providing an additional expertise on the quantum-mechanical computations and carrying out an independent trial of the Auto QSAR modeling.

ORCID

Filip Andrić http://orcid.org/0000-0001-7932-833X
Mario Zlatović http://orcid.org/0000-0003-4311-1731
Danica Agbaba http://orcid.org/0000-0001-5907-9823

REFERENCES

1. Marcus RN, McQuade RD, Carson WH, et al. The efficacy and safety of aripiprazole as adjunctive therapy in major depressive disorder: A second multicenter, randomized, double-blind, placebo-controlled study. *J Clin Psychopharmacol.* 2008;28(2):156-165.

2. Lemke TL, Williams DA, Roche VF, Zito SW. *Foye's Principles of. Medicinal Chemistry, 7th Ed.* Lippincot Williams & Wilkins: Baltimore; 2012:449-484.

3. Greenberg WM, Citrome L. Ziprasidone for schizophrenia and bipolar disorder: A review of the clinical trials. *CNS Drug Rev.* 2007;13(2):137-177.
4. Keck PE, Marcus R, Tourkodimitris S, et al. A placebo-controlled, double-blind study of the efficacy and safety of aripiprazole in patients with acute bipolar mania. *Am J Psychiatry.* 2003;160(9):1651-1658.
5. Sayeh HG, Morganti C, Adams CE. Aripiprazole for schizophrenia: Systematic review. *Br J Psychiatry.* 2006;189(2):102-108.
6. Sakai JB. *Practical Pharmacology for the Pharmacy Technician*. New York, NY: Lippincott Williams & Wilkins; 2008.
7. Daousani C, Macheras P. Biopharmaceutical classification of drugs revisited. *Eur J Pharm Sci.* 2016;95:82-87.
8. Rohrschneider L. Solvent characterization by gas-liquid partition coefficient of selected solutes. *Anal Chem.* 1973;45(7):1241-1247.
9. Genty M, Gonzalez G, Clere C, Desangle-Gouty V, Legendre JY. Determination of the passive absorption through the rat intestine using chromatographic indices and molar volume. *Eur J Pharm Sci.* 2001;12(3):223-229.
10. Waterbeemd H, Gifford E. ADMET in silico modelling: Towards prediction paradise? *Nat Rev Drug Discov.* 2003;2(3):192-204.
11. Terada M, Marchessault RH. Determination of solubility parameters for poly(3-hydroxyalkanoates). *Int J Biol Macromol.* 1999;25(1-3):207-215.
12. Breitkreutz J. Prediction of intestinal drug absorption properties by three-dimensional solubility parameters. *Pharm Res.* 1998;15(9):1370-1375.
13. Martini LG, Avontuur P, George A, Willson RJ, Crowley PJ. Solubility parameter and oral absorption. *Eur J Pharm Biopharm.* 1999;48(3):259-263.
14. Patil JSC, Korachagaon AV, Shiralashetti SS, Marapur SC. Enhancing dissolution rate of ziprasidone via co-grinding technique with highly hydrophilic carriers. *RGUHS J Pharm Sci.* 2012;2(2):26-32.
15. Lentz AK, Quitko M, Morgan DG, Grace JE Jr, Gleason C, Marathe PH. Development and validation of a preclinical food effect model. *J Pharm Sci.* 2007;96(3):459-472.
16. Li M. *Organic Chemistry of Drug Degradation*. Cambridge, UK: RSC Publishing; 2012:124-126.
17. Hong J, Shah JC, Mcgonagle MD. Effect of cyclodextrin derivation and amorphous state of complex on accelerated degradation of ziprasidone. *J Pharm Sci.* 2011;100(7):2703-2716.
18. Djordjević Filijović N, Pavlović A, Nikolić K, Agbaba D. Validation of an HPLC method for determination of aripiprazole and its impurities in pharmaceuticals. *Acta Chromatogr.* 2014;26(1):13-28.
19. Hansen CM. *Hansen Solubility Parameters: A User's Handbook*. 2nd ed. New York, NY: CRC Press; 2007.
20. Kitak T, Dimičić A, Planinšek O, Šibanc R, Srčič R. Determination of solubility parameters of ibuprofen and ibuprofen lysinate. *Molecules.* 2015;20(12):21,549-21,568.
21. Heberger K. Sum of ranking differences compares methods fairly. *TrAC, Trends Anal Chem.* 2010;29(1):101-109.
22. Kovačević SZ, Podunavac-Kuzmanović SO, Jevrić LR, et al. How to rank and discriminate artificial neural networks? Case study: Prediction of anticancer activity of 17-picolyl and 17-picolinylidene androstane derivatives. *J Iran Chem Soc.* 2016;13(3):499-507.
23. Bielicka-Daszkiewicz K, Voelkel A, Pietrzynska M, Héberger K. Role of Hansen solubility parameters in solid phase extraction. *J Chromatogr A.* 2010;1217(35):5564-5570.
24. Brownfield B, Kalivas JH. Consensus outlier detection using sum of ranking differences of common and new outlier measures without tuning parameter selections. *Anal Chem.* 2017;89(9):5087-5094.
25. Šegan S, Božinović N, Opsenica I, Andrić F. Consensus-based comparison of chromatographic and computationally estimated lipophilicity of benzothiepino[3,2-c]pyridine derivatives as potential antifungal drugs. *J Sep Sci.* 2017;40(10):2089-2096.
26. Kovačević SZ, Tepić AN, Jevrić LR, et al. Chemometric guidelines for selection of cultivation conditions influencing the antioxidant potential of beetroot extracts. *Comput Electron Agric.* 2015;118:332-339.
27. Sipos L, Bernhardt B, Gere A, et al. Multicriteria optimization to evaluate the performance of Ocimum basilicum L. varieties. *Ind Crop Prod.* 2016;94:514-519.
28. Alhalaweh A, Alzghoul A, Kaialy W. Data mining of solubility parameters for computational prediction of drug-excipient miscibility. *Drug Dev Ind Pharm.* 2014;40(7):904-909.
29. Hildebrand JH, Scott RL. Solution of nonelectrolytes. *Annu Rev Phys Chem.* 1950;1(1):75-92.
30. Grassi G, Lapasin R, Grassi M, Lapasin R, Colombo I. *Understanding Drug Release and Absorption Mechanism: A Physical and Mathematical Approach*. New York, NY: CRC Press; 2006:568-660.
31. Perea JD, Langner S, Savador M, et al. Combined computational approach based on density functional theory and artificial neural networks for predicting the solubility parameters of fullerenes. *J Phys Chem B.* 2016;120(19):4431-4438.
32. Járvás G, Quellet C, Dallos A. Estimation of Hansen solubility parameters using multivariate nonlinear QSPR modeling with COSMO screening charge density moments. *Fluid Phase Equilib.* 2011;309(1):8-14.

33. Szekely MG, Amores de Sousa C, Gil M, Castelo Ferreira F, Heggie W. Genotoxic impurities in pharmaceutical manufacturing: Sources, regulations, and mitigation. *Chem Rev.* 2015;115(16):8182-8229.

34. Boulton DW, Kollia G, Mallikaarju S et al. Pharmacokinetics and tolerability of intramuscular, oral and intravenous aripiprazole in healthy subjects and in patients with schizophrenia. *Clin Pharmacokinet.* 2008;47(7):475-485.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Obradović D, Andrić F, Zlatović M, Agbaba D. Modeling of Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities: A nonparametric comparison of models for prediction of drug absorption sites. *Journal of Chemometrics.* 2018;32:e2996. https://doi.org/10.1002/cem.2996