symmetry

MDPI

Review

Chemometrics for Selection, Prediction, and Classification of Sustainable Solutions for Green Chemistry—A Review

Marta Bystrzanowska * and Marek Tobiszewski

Department of Analytical Chemistry, Faculty of Chemistry, Gdańsk University of Technology (GUT), 80-233 Gdańsk, Poland; marektobiszewski@wp.pl

* Correspondence: marbystr@student.pg.edu.pl

Received: 13 November 2020; Accepted: 9 December 2020; Published: 11 December 2020

Abstract: In this review, we present the applications of chemometric techniques for green and sustainable chemistry. The techniques, such as cluster analysis, principal component analysis, artificial neural networks, and multivariate ranking techniques, are applied for dealing with missing data, grouping or classification purposes, selection of green material, or processes. The areas of application are mainly finding sustainable solutions in terms of solvents, reagents, processes, or conditions of processes. Another important area is filling the data gaps in datasets to more fully characterize sustainable options. It is significant as many experiments are avoided, and the results are obtained with good approximation. Multivariate statistics are tools that support the application of quantitative structure–property relationships, a widely applied technique in green chemistry.

Keywords: multivariate statistics; sustainable chemistry; missing data; classification; grouping; solvents

1. Introduction

The term "chemometrics" was coined by the Swedish scientist Svante Wold in early 1970s while submitting a grant proposal for the application of statistical methods to chemical data [1]. It appeared as the word "kemometri," a combination of the forms "kemo-" for chemistry and "-metri" for measure [2].

Initially, chemometrics was defined as a "science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods." According to the name, the discipline of chemometrics originated from chemistry, where one of the first applications focused on improving the quantitative performance of analytical instruments, such as NIR (near infrared) calibration, HPLC (high-performance liquid chromatography) resolution, and UV–VIS deconvolution [3]. Chemometrics took the form of an interdisciplinary field that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments and to provide maximum chemical information by analysing chemical data. The numerous domains that are covered by chemometrics are presented by Santos et al. on a bibliometric map generated using more repeated words in the authors' search for the period 2014–2018 performed in the Science Citation Index Expanded [4]. However, the breakthrough in chemometrics is a response to various software and new high-dimensional hyphenated equipment appearance. These devices in chromatography have been allowed for the determination of various analytes in complex matrices with high resolution and precision. On the other hand, obtained results as large datasets become more difficult to interpret.

Due to rapid technological advances, the focus on multivariate methods is visible. Therefore, the distribution of multiple variables simultaneously provides more information than what could be obtained by considering each variable individually. Then some meaningful information may be chemometrically extracted. As mentioned above, chemometrics is a very important issue in fields

concerning environmental monitoring, forensics, chemical biology, food and nutrition, pharmaceutics, polymer, safety and healthcare diagnostics, fraud detection, green chemistry and sustainability, and omics sciences. The latter, together with some bioinformatics and cheminformatics, is becoming more and more popular recently (especially in an advanced data analysis).

However, the use of chemometrics is responsible not only for intelligent data analysis but more specifically for modelling, classification, selection, or searching for missing data. Due to the fact that chemical sciences are based on complex processes involving multistep chemical processes, with condition optimizations, selection of chemical reagents, and so forth, they are a great representative of a wide spectrum of chemometric utilization.

It is also worth noting that chemometric application may be an incredible approach to incorporating the green chemistry concept to chemical sciences via the usage of more environmentally friendly chemicals, analytical procedures, or chemical processes and their optimization (saving energy and materials) and prediction of properties to provide additional information and estimate environmental fate of chemical compounds and pollutants.

In the study, the application of chemometrics in green chemistry as a tool for selection (chemical substances, mainly solvents), classification (different types of organic solvents and ionic liquids), and property prediction (i.e., viscosity, density, carbon dioxide solubility, toxicity, partition coefficient, bioconcentration factor) is presented and discussed.

2. The Outline of Chemometric Tools

Chemometric tools may be divided into two groups: qualitative and quantitative methods. The first group is dedicated to solving problems of classification and pattern recognition. In other words, they allow for assigning an individual sample to a given group of samples or finding a sorting pattern in the underlying data structure of a set [5]. The idea of these methods is based on two philosophies dividing methods into unsupervised and supervised methods. The aim of unsupervised methods is to reveal the underlying data structure without the potential bias of knowing the group memberships beforehand. On the other hand, supervised methods are based on producing the best possible separation of the groups. Therefore, they maximize the capability of the classification method to predict the class membership of samples with unknown membership. Accordingly, it is worth bearing in mind that depending on the problem, one group of methods could be more suited for a given purpose. However, due to fact that it is not always an unambiguous choice, sometimes several chemometric tools are applied. In finding the connection between the detected signals and the exact concentration values, quantitative methods are used. As it is widely known, modern analytical devices generate huge datasets with thousands of spectral data (from Fourier transform infrared/near-infrared, mass spectrometry, nuclear magnetic resonance, etc.); therefore, finding a correlation is very often unclear and difficult. The quantitative analysis is based on regression techniques, whose concept involves exploration of a connection (linear or nonlinear) between one or several independent variables and one (or more, but usually one) dependent variable. If there is only one dependent and one independent variable, then the easiest case is presented—a univariate regression. However, sometimes, as in analytical chemistry problems, the situation is more complicated, including a greater number of dependent variables [6]. Taking the above into account, the selection of an appropriate chemometric tool is dictated by the purpose of the analysis and the characteristics of a given problem. Moreover, obtaining satisfactory results may require the use of several tools. The most commonly used chemometric tools in chemical analysis are briefly described below [7].

The most commonly used chemometric tools in chemical sciences are principal component analysis (PCA) [8,9] and cluster analysis (CA) [6,10]. These unsupervised techniques are very often applied for reducing the dimension of the original data [11], finding internal patterns in the dataset [12,13], or discovering the dominant factors [14,15]. In element classification, very popular are supervised techniques such as linear discriminant analysis (LDA) [16] and partial least squares (PLS) [17,18]. However, they may also be used for prediction [19,20]. An example of regression algorithms may

MOST WIEDZY Downloaded from mostwiedzy.pl

be similar to each other: multiple linear regression (MLR) [21] and principal component regression (PCR) [22]. They are mainly used in data analysis for finding the relationship among variables that effect the prediction of variable values (e.g., chemical compounds' properties). Nevertheless, the most widely used prediction tools are mathematical models from the quantitative structure–activity relationship (QSAR) family [23,24]. They allow for finding the physicochemical, biological, and environmental fate properties of compounds in reference to the knowledge of their chemical structure (new and existing chemical compounds) without animal use in, for example, toxicological testing. Nowadays, artificial neural network (ANN) and genetic algorithm (GA) are gaining more attention in the field of chemical sciences while identifying patterns in data, even complex ones. This is due to their structures and mechanisms, because both of them are comparable to evolutionary processes in nature, namely, equivalents of genes and chromosomes in GA [25] or the biological (human or animal) central nervous system (including neurons) in ANN [26]. They can be successfully used separately [27] or often as a combined tool [28,29]. It is worth noting that these are not all of the techniques that may be used for this purpose. Other approaches, for instance, sum of ranking differences (SRD) [30], *k*-nearest neighbours (KNN) method [31], and support vector machine, (SVM) [32,33], may also be successfully applied for alternative data treatment in the context of green chemistry. Details of the mentioned chemometric techniques are described elsewhere (some references given in brackets); therefore, they are not fully described in this review.

3. Selection

The problem of selection can be related to the solvents and other chemical reagents (for instance, derivatization agents) used in operations, such as extraction, clean-up, and derivatization. In these cases, the selection of appropriate solvents and chemical reagents for additional chemical activities is extremely important to obtain satisfactory results. Nevertheless, it is worth looking for substitutes for those chemicals mentioned above that are less hazardous to the environment, which correspond to the 5th and 8th of the 12 principles of green chemistry for solvents and derivatization agents, respectively. Considering the above, it is not surprising that the selection of appropriate chemical reagents is a topic of interest in chemometrics.

An approach for fast selection of solvents for a given industrial application with the use of chemometric tools is proposed by García et al. [34]. First, the QSPR (quantitative structure–property relationship) model is developed to find the relationship between the molecular structure and some fundamental solvent properties. Then MLR (multiple linear regression) and PLS (partial least squares) are used for the selection of 62 glycerol-based solvents with respect to three solvent features: the behaviour of the dissolution processes (solvatochromic parameter E_T^N), mechanical aspects (viscosity), and volatility aspects (closely related to safety, toxicity, and air pollution considered through the boiling point). A comparison of applied chemometric tools shows that both of them represent good results in the E_T^N solvation parameter. MLR is only appropriate in the E_T^N solvation parameter, whereas PLS offers better fitting of two of the three properties considered simultaneously. Viscosity and boiling point do not fit well enough to lead to a fully predictive model; however, PLS provides a higher value of determination coefficient for boiling point.

A solvent selection system based on a combination of chemometrics and multicriteria decision analysis is proposed by Tobiszewski et al. in line with the concept of green chemistry [35]. CA (cluster analysis), together with the TOPSIS (the technique for order of preference by similarity to ideal solution) algorithm, allows for, first, grouping and then ranking within groups of 151 solvents in respect to physicochemical, toxicological, and hazard parameters. Three clusters, as presented in Figure 1, are obtained: nonpolar and volatile (35 solvents), nonpolar and sparingly volatile (35 solvents), and polar (81 solvents). The results are compared with another SSG (solvent selection guide) developed by Pfizer [36], GlaxoSmithKline [37], AstraZeneca [38], Sanofi [39], and CHEM21 [40], which are well known in the pharmaceutical industry, confirming a general agreement of solvent rankings within each cluster.

MOST WIEDZY Downloaded from mostwiedzy.pl

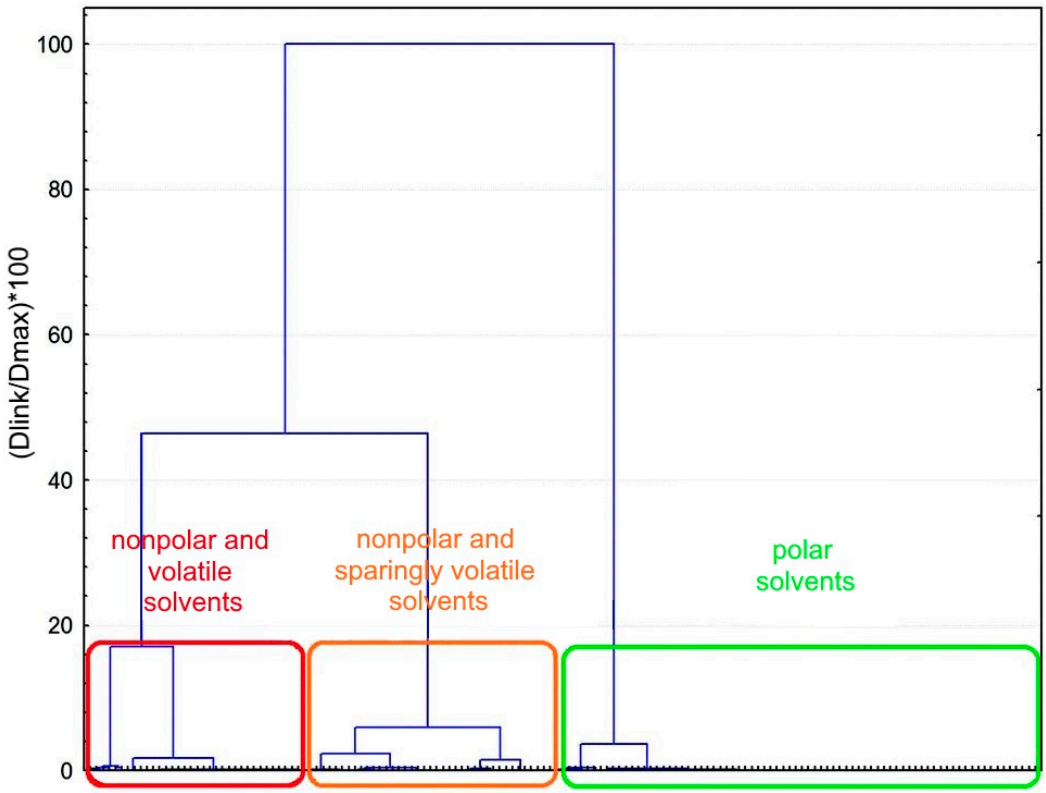

Figure 1. Clustering of the solvents based on their 9 physicochemical properties using CA (cluster analysis). Reproduced from Ref. A solvent selection guide based on chemometrics and multicriteria decision analysis (Tobiszewski et al. [35]) with permission from the Royal Society of Chemistry.

Similar results were recently presented by Sels et al. with the application of MDS (multidimensional scaling) [41]. Solvents were assigned to three groups based on their 22 physical properties according to safety, health, and environment scores: polar compounds, slightly water-soluble solvents, and hydrophobic solvents. In the MDS visualization, the solvents that were similar were plotted closer together in the 2D solvent space. However, it was noted that the relative influence of a functional group decreased with increasing chain length and molecular size. Then a straight line in the MDS visualization was not visible for homologous series from alcohols (due to drastic increase in boiling point and decrease in water solubility, vapour pressure, and relative evaporation rate). Moreover, the application of SUSSOL (Sustainable Solvents Selection and Substitution Software), a specially created software by applying artificial intelligence (AI), is presented for finding solvent replacements for *N*-methylpyrrolidone (NMP), toluene, and tetramethyl oxolane (TMO). The proposed alternative solvents are as follows: 10 candidate alternative solvents (including dimethyl sulfoxide, Cyrene, *N*-butyl pyrrolidone, pyridine, acetone, methyl acetoacetate, 1-ethyl pyrrolidone, dimethylacetamide, dimethylformamide, nicotine) for NMP; isobutylbenzene and p-cymene for toluene; and toluene, 1,1-dichloroethene, 1,1-dichloroethane, 1,1,1-trichloroethane, 1,1-dichloropropane, ethylene glycol diethyl ether (1,2-diethoxyethane), and so forth for TMO. An example of visualization dedicated to possible alternatives for NMP by SUSSOL software is presented in Figure 2.

A screening of potential PBT (persistent, bioaccumulative, and toxic) compounds (in an environment based on persistence, bioconcentration, and toxicity data) is another example of chemical selection, but different from solvents [42]. PCA is used to group chemicals representing many classes of pollutants of various chemical structures, such as dioxins, PCBs, PAHs, and pesticides, and various industrial chemicals according to their potential cumulative PBT behaviour. However, due to unavailability of experimental data, an approach combining multivariate analysis and QSAR/QSPR (quantitative structure–activity relationship) was applied, which allowed for the reduction of data gaps

MOST WIEDZY Downloaded from mostwiedzy.pl

in the dataset. The strength of the approach is validated in two sequential steps: first, performed on the available experimental dataset, including 54 chemicals, and then performed on the dataset of 180 chemicals (developed by QSPR). In Figure 3, the analysis of the latter dataset of organic compounds using PCA is presented.

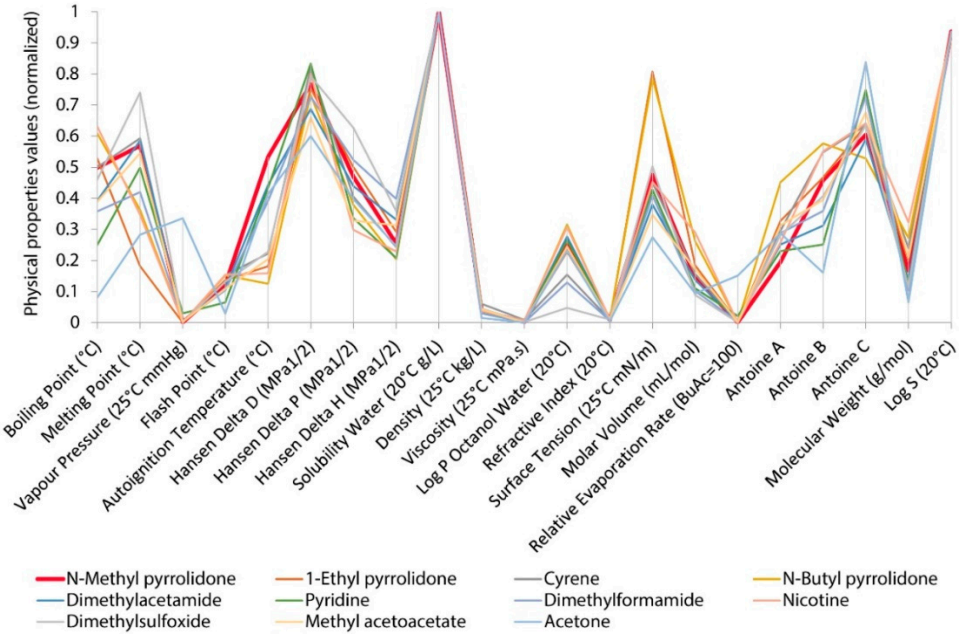

Figure 2. Visualization of the analysis results of substitution candidates for NMP in SUSSOL software. Reproduced from Ref. SUSSOL—Using Artificial Intelligence for Greener Solvent Selection and Substitution (Sels et al. [41]).

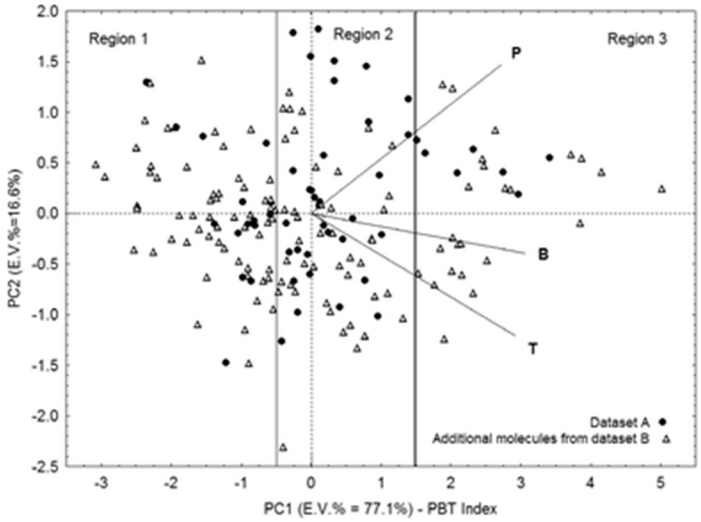

Figure 3. PCA (principal component analysis) on experimental and predicted PBT (persistent, bioaccumulative, and toxic) data for 180 organic compounds (dataset A – 54 comp. + dataset B – 126 comp.). Reproduced from Ref. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure (Papa and Gramatica [42]) with permission from the Royal Society of Chemistry.

According to PBT index values, chemicals are grouped into three regions: region 1—not PBT chemicals, region 2—chemicals with medium PBT properties, and region 3—PBT and vPvB (very persistent and very bioaccumulative) chemicals.

MOST WIEDZY Downloaded from mostwiedzy.pl

4. Classification

Classification as a systematic arrangement in groups or categories according to established criteria is sometimes very useful in designing a chemical process or reaction. It allows for recognizing some alternatives with corresponding characterization.

Translating the principle *similia similibus solvuntur* into the field of chemistry means solvents belonging to the same group demonstrate similar abilities to dissolve compounds. Therefore, chemometric classification of solvents according to the degree of polarity may provide information about possible substitutes. This kind of grouping addressed to organic solvents is one of the frequently undertaken problems in chemometrics, which is summarized in Table 1.

Table 1. Organic solvent classification according to the degree of polarity by chemometric application—summarized exemplary studies.

Classification Object	Chemometric Tool	Evaluated Parameters	Results—Groups of Solvents	Ref.
83 organic solvents	PCA	• the Kirkwood function (K) • molecular refraction (MR) • molecular dipole moment (μ) • the parameter of Hildebrand • index of refraction (n) • boiling point (bp) • energies of HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital)	9 groups of solvents: • aprotic dipolar: acetonitrile, acetone, ethyl acetate, dichloromethane • aprotic highly dipolar: dimethyl sulfoxide, *N*,*N*-dimethyl formamide, pyridine • aprotic highly polarizable dipolar: hexamethylphosphotriamide • aromatic apolar: toluene, benzene • aromatic polar: chlorobenzene, o-dichlorobenzene • electron-pair donor: triethylamine, diethyl ether, dioxane • hydrogen bonding: methanol, ethanol, pentan-2-ol • hydrogen bonding strongly associated: formamide, water, ethylene glycol • miscellaneous: carbon disulphide, chloroform, aniline	Chastrette et al. (1985) [43]
101 organic solvents	Parker–Reichardt classification	correlation between dielectric β parameter and empirical solvent polarity parameter E_T^N	4 groups (and 2 subgroups) of solvents: • weakly dipolar nonhydrogen bonding donor: ethers, carboxylic esters, tertiary amines, halogen-substituted hydrocarbons • dipolar nonhydrogen bonding donor: ketones, *N*,*N*-disubstituted amides, nitro-substituted hydrocarbons, nitriles, sulphoxides, sulphones, cyclic carbonates, pyridine • hydrogen bonding donor: water, alcohols, carboxylic acid, glycols ○ nonprimary alcohols and aniline ○ phenol and its derivatives • *N*-monosubstituted amides and formamide	Dutkiewicz (1990) [44]

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 1. *Cont.*

Classification Object	Chemometric Tool	Evaluated Parameters	Results—Groups of Solvents	Ref.
51 solvents	KNN	Empirical scale parameters: • PAC (polarity/acidity) • PBC (polarity/basicity) • PPC (polarity/polarizability)	8 groups of solvents: • Nonpolar inert solvents: aliphatic hydrocarbons) • nonpolar-polarizable: aromatic hydrocarbons, tetrachloromethane, carbon disulphide • nonpolar-basic: ethers, triethylamine • little polar-polarizable: aliphatic halogen derivatives, substituted benzenes with heteroatom-containing substituents • little polar-basic: cyclic ethers, ketones, esters, pyridine • polar-aprotic: acetanhydride, dialkylamides, acetonitrile, nitromethane, dimethyl sulfoxide, sulfolane • polar-protic: alcohols, acetic acid • exceptional solvents: water, formamide, glycol, hexamethylphosphoric triamide	Pytela (1989) [45]
152 organic solvents	KNN, CP-ANN, QSPR	4 molecular descriptors (theoretical descriptions of the molecular structure)	5 groups of solvents: • aprotic polar • aromatic apolar or lightly polar • electron-pair donors • hydrogen bonding donors • aliphatic aprotic apolar	Gramatica et al. (1999) [46]
76 solvents	ANN	9 characteristics (application in a field of C60 fullerene solubility)	9 groups of solvents: • apolar and slightly polar: n-pentane, n-hexane, n-octane, n-decane • apolar and slightly polar: n-dodecane, benzene, m/o/*p*-xylene, toluene, ethylbenzene, cumene • apolar and slightly polar: carbon disulphide, tetrachloroethylene • weakly polar: fluorobenzene, dichloromethane, o-cresol • weakly polar: chlorobenzene, pyridine • weakly polar: bromobenzene, bromoform • hydrogen bond donors and others: methanol, ethanol, 1-propanol, 1-butanol, acetone • hydrogen bond donors and others: 1-pentanol, 1-hexanol, 1-octanol, 1-decanol • highly polar: nitrobenzene, benzonitrile • highly polar: 1,2-ethanediol, water, *N*-methylformamide, acetonitrile, *N*,*N*-dimethylformamide • miscellaneous: chloroform, 1-aminobutane	Pushkarova and Kholin (2014) [47]

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 1. *Cont.*

Classification Object	Chemometric Tool	Evaluated Parameters	Results—Groups of Solvents	Ref.
236 industrial solvents	PCA, CA	quantum and experimental parameters	10 groups of solvents: • hydrogen bond donor: short-chain alcohols, phenols, acetic acid, butyric acid • hydrogen bond donor with high polarizability: tributylamine, glycols, long-chain alcohols • hydrogen bonds acceptor/electron-pair donor: amines, pyridines, aniline, anisole, dioxane • aprotic dipolar: ethyl acetate, cyclohexanone, acetophenone, acetone • aprotic dipolar-polarizable: sulfolane, ketones with at less C7, hexamethylphosphoramide • aprotic very strongly dipolar: nitro/nitrile compounds • aprotic apolar: linear or cyclic alkanes • aprotic apolar with pi bonds: aromatics, xylenes, cyclohexane • halogenated hydrocarbons: dichloromethane, carbon disulphide, halogenated derivatives of benzene, carbon tetrachloride	Levet et al. (2016) [48]
72 solvents	FCM, FLDA	Chemical parameters connected with polarity and selectivity developed by Snyder (related to different polar interactions): • proton acceptor (x_e) • proton donor (x_d) • dipole (x_n) • chromatographic strength (P') derived from gas–liquid partition coefficient • toluene similitudes (x_t) • methylethylketone similitudes (x_m)	FCM—8 groups (selected examples): • cyclohexanone, ethylmethylketone, dioxane, acetophenone, benzonitrile, ethyl acetate, nitrobenzene • dimethyl sulfoxide, ethyleneglycol, m-cresol, m-methylpyrrolidone • *p*-xylene, toluene, benzene, bromobenzene • aniline, dimethylformamide, propylene carbonate, *N*,*N*-dimethyl acetamide, acetic acid • 1-propanol, 2-propanol, tetrahydrofuran, 1-butanol, tert-butanol, anisole, ethanol • fluorobenzene, 1-octanol • pyridine, triethylene glycol, benzyl alcohol, acetonitrile, methanol, acetone • formamide, water, dodecafluoroheptanol FLDA—8 groups (selected examples): • diethylether, triethylamine • propanol, 1-octanol, 2-propanol, 1-butanol, ethanol, tert-butanol, methanol • pyridine, methylformamide, triethylene glycol, *N*,*N*- dimethyl acetamide, dimethyl sulfoxide • acetic acid, ethylene glycol, formamide • methylene chloride, ethylene chloride • acetophenone, dioxane, acetonitrile, acetone, tetrahydrofuran, aniline, ethyl acetate • chlorobenzene, *p*-xylene, benzene, anisole, toluene, chloroform • dodecafluoroheptanol, water, m-cresol	Guidea and Sârbu(2020) [49]

Interestingly, these classifications are carried out for various objects (types of solvents) using different chemometric tools, for instance, PCA, KNN (*k*-nearest neighbours method), Parker-Reichardt classification, CP-ANN (counter-propagation artificial neural network), ANN (artificial neural network),

MOST WIEDZY Downloaded from mostwiedzy.pl

PCA, and CA, obtaining similar results. An example may be the study performed by Dutkiewicz [44] using the Parker–Reichardt classification, whose results highly correspond to those obtained by a more complex multivariate statistical method presented by M. Chastrette et al. [43]. Moreover, there are applications with few tools applied. The idea is to improve the results of classification, for instance, by making them more chemically interpretable, as in organic solvent classification based on molecular descriptors (theoretical descriptions of the molecular structure), where KNN application is followed by CP-ANN [46].

One of the latest works considers a classification of 72 solvents according to polarity and selectivity issues based on the Snyder approach (related to different polar interactions), performed using FCM (fuzzy *c*-means) and FLDA (fuzzy linear discriminant analysis) [49]. The used fuzzy chemometric techniques show high efficiency and information power methods in solvent characterization and classification (an approach for rationalchoosing of a good solvent). The obtained results (division into eight groups of solvents) are in good agreement with the Snyder classification, especially using FLDA (the highest value of 100% for the solvents corresponding to groups II and V and the lowest value of 66.67% for the solvents of group I).

However, the classification does not always take into account a large number of groups/classes. Salahinejad [50] proposed a division of solvents for single-walled carbon nanotube dispersion into two groups: solvents and nonsolvents (solvents with effectively zero of nanotube dispersibility). The classification is conducted separately with several tools, such as RF (random forest), SVM (support vector machine), MLP (multilayer perceptron), and QDA (quadratic discriminant analysis). According to the results of the sum of ranking difference (SRD) procedure, the RF classifier based on selected descriptors is the best classification model, while the SVM, MLP, and QDA are ranked as good models.

Moreover, another classification of solvents based on a chemical group of compounds was performed by Katritzky et al. [51] and Tobiszewski et al. [52]. In the first case, a classification of the theoretical molecular descriptors, derived from the chemical structure alone (QSPR model), according to their relevance to specific types of intermolecular interaction (including cavity formation, electrostatic polarization, dispersion, and hydrogen bonding) in liquid media is presented. According to the PCA results, 11 classes of solvents were formed: hydrocarbons; halo-hydrocarbons; saturated, unsaturated, and cyclic ethers; esters and polyesters; aldehydes, ketones, and amides; nitriles and nitro hydrocarbons; hydroxylic compounds; amines and pyridines; thiols, sulphides, sulfoxides, and thio compounds; phosphorus compounds; and compounds with vastly different chemical functionalities. In the latter case, CA and PCA were used to group around 130 potentially green organic solvents according to their similarity based on physiochemical parameters, as well as to assess and identify variables from which properties missing values such as bioconcentration factors, water–octanol, and octanol–air partitioning constants can be predicted. The CA results show that polar solvents are divided into three major groups: (a) less volatile solvents, slightly water soluble with high values of logKOW and logBCF (alcohols with ether functional groups, aromatic alcohols, and short-chain organic acids apart from formic and acetic); (b) less volatile and very highly water-soluble solvents (lactate esters, formic and acetic acids, glycerol, and some alcohols with other functional groups); and (c) highly volatile, low-boiling-point, high vapour pressure, and Henry's law constant solvents ("traditional" polar solvents, like short-chain alcohols, ketones, aldehydes, and esters). On the other hand, nonpolar solvents were divided into volatile, water-nonsoluble, and slightly water-soluble solvents. According to a chemometric analysis connected with finding the internal relationship between bioconcentration factors and physiochemical parameters, in polar solvents, the variable logBCF forms a separate latent factor not directly correlated with other variables (specific importance of this parameter as a discriminant for the dataset). Unlike in nonpolar solvents, the relationship between parameters like logBCF and logKOW and Henry's law constant and the correlation of logKOA with a whole group of physicochemical parameters, like surface tension, density, boiling, and melting point, is visible.

MOST WIEDZY Downloaded from mostwiedzy.pl

A different approach for the classification of 259 solvents according to the experimentally found and theoretically predicted physicochemical parameters presented by 15 specific descriptors is proposed by Nedyalkova et al. (2020) [53]. The variables involved parameters such as melting point, boiling point, density, water solubility, vapour pressure, Henry's law constant, octanol–water and octanol–air partition coefficients, and bioconcentration factor, some of which are implemented within the modules of EPI Suite or by the SMILES codes (simplified molecular input line entry system). The fuzzy hierarchical clustering methods allow for checking whether the experimental values of the respective variables correspond to the calculated ones, and the partitioning procedure could determine stable groups of similarity between the variables with highly different degrees of membership. The performed partitioning with respect to specific descriptors divides solvents into 10 classes (some examples of solvents within each class are presented in brackets) (i.e., chlorinated solvents—class 1 (iodoethane, n-butyl acetate, m-cresol, diethyl carbonate, chloroform), nonpolar and volatile solvents—class 2 (bromoethane, benzonitrile, isobutyl acetate, carbon disulphide), polar and nonpolar solvents mixed—class 3 (benzene, dichloromethane, diethyl ether, triethylene glycol, polyethyleneglycol 200), polar solvents—classes 4–7 (dioctylsuccinate, oleic acid, 2-pyrrolidone, glycerol, water, 1-octanol, nitrobenzene, methyl stearate), high molecular weight polar solvents—class 8 (ethyl laurate, anisole), large group of mostly polar solvents with some exceptions—class 9 (triethylamine, ethanol, 1-butanol, formamide, toluene, o-xylene, aniline, n-heptane, d-limonene, styrene, acetone, phenol, acetonitrile), and outlier—class 10 (perfluorooctane 20). The relationships between solvents of various natures (polar, nonpolar, volatile, etc.) and the physicochemical variables are found, despite the fact that missing data of specific descriptors are fulfilled via theoretical calculation. Moreover, applied chemometric techniques allow for partitioning solvents with more or less similar characteristics in terms of higher, smallest, or intermediate values of considered descriptors.

One of the most interesting groups of solvents are ionic liquids (ILs) due to their desired feature—designing of solvents with particular properties (within certain ranges) by a combination of selected cation and anion. Therefore, characterization of their types is very important for finding an appropriate alternative, for instance, in phases for gas chromatography. This aspect is discussed by González-Álvarez et al. in the classification of three ILs with hexacationic imidazolium, polymeric imidazolium, and phosphonium as cations and halogens, thiocyanate, boron anions, triflate, and bistriflimide as anions [54]. The application of CA, LDA (linear discriminant analysis), D-PLS (discriminant partial least squares), and MLR shows that two main groups of phases may be distinguished: ILs with acidic and basic characterization. After the identification of the two natural groups of ILs by CA, several supervised chemometric techniques, such as LDA, D-PLS, and MLR were used to construct models of pattern recognition and classification rules for ILs. All tools showed high prediction capacity and were successfully used for characterizing IL classes. The best results were obtained via LDA with >96% for classification and >92% for prediction, followed by MLR with 96.7% and 92% in the prediction for classes A and B, respectively.

In another study, 227 ionic liquids and their related salts were also classified based on their toxicities towards rat cell lines [55]. Regardless of the used chemometric method (LDA, CA, SVM (support vector machine), or CP-ANNs (counter-propagation artificial neural networks)), ILs were classified into four categories: low, moderate, high, and very high toxicity. In this study, CP-ANN turned out to be more favourable over other methods in terms of accuracy of classification, underlining that CP-ANNs may extract actual information and knowledge from the dataset.

An interesting approach with a classification map called the Σpider diagram was proposed by Lesellier [56]. Solvents were classified based on physiochemical properties encountered with other visual presentations, such as Snyder triangle, Hansen parameters, LSER (linear solvation energy relationships), Abraham descriptors, COSMO-RS (Conductor like Screening Model for Real Solvents) parameters, and solvatochromic solvent selectivity. Visualization of the last solvent classification is presented in Figure 4.

MOST WIEDZY Downloaded from mostwiedzy.pl

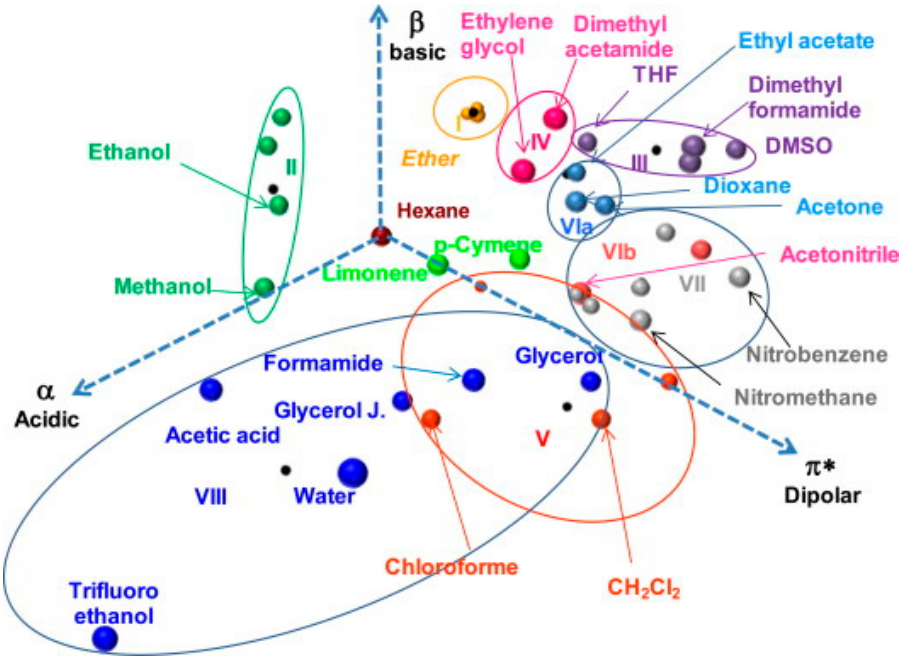

Figure 4. Spider diagram based on solvatochromic parameters π^*, α, β. Reprinted from Journal of Chromatography A, 1389, E. Lesellier, Σpider diagram: A universal and versatile approach for system comparison and classification: Application to solvent properties, 49–64, Copyright 2015, with permission from Elsevier.

This diagram shows many advantages of solvent classification through a better view of solvents having no acidic character (for the solvatochromic solvent selectivity), easier usage due to the "flattening" of the spherical view down to a single plane (for Hansen parameters), more subtle classification due to the use of five parameters instead of three (for COSMO-RS), and simple view of the solvent groups having similar or different properties (for Abraham descriptors). An approach may be useful not only for selecting suitable solvents for extraction, separation, or purification approaches and for solubility studies but also for choosing greener solvents.

There are also other fields of interest apart from solvents, for instance, pharmaceutical excipients in reference to their solubility parameters [57]. PCA is used to predict a behaviour of materials in a multicomponent system (e.g., for the selection of the best materials to form stable pharmaceutical liquid mixtures or stable coating formulation). It is significantly important because similarity between the values of the respective components of the solubility parameter allows for the estimation of the compatibility between different materials (solvents, colorants, lubricants, coating components, and powder blends).

5. Properties (Prediction and Correlation)

Knowledge of the physicochemical properties of compounds is necessary to predict their behaviour under various conditions or factors during chemical reactions, and their behaviour in various media or compartments in the environment (environmental fate). Therefore, this explains the need to obtain information on the solvents' and other chemical reagents' properties. Unfortunately, sometimes there are missing points in chemical characteristics. Thus, some prediction and computational methods for filling the gaps are highly required and successfully applied.

An example of the most popular advanced and computational modelling approaches may be QSAR (quantitative structure–activity relationship) and EPI Suite (Estimation Programs Interface Suite). QSAR models allow for the prediction of the physicochemical, biological, and environmental fate properties of compounds in reference to knowledge of their chemical structure. The concept is based on establishing quantitative relationships between descriptors (referring to the chemical

MOST WIEDZY Downloaded from mostwiedzy.pl

structure) and the target property capable of predicting activities of novel compounds [58]. On the other hand, EPI Suite may estimate physical/chemical and environmental fate properties such as water solubility, octanol–water partition coefficient, Henry's law constant, melting point, boiling point, and aquatic toxicity, taking into account chemical structure as input data (depending on the chosen estimation model program) [59]. However, the easiest manner is chemical predictive modelling, which is based on an observation of some patterns, correlations between variables in dataset. In this respect, the chemometric tools play an important role.

As mentioned in Section 3, the use solvents in chemistry is one of the most important issues with respect to environmental aspects. In this manner, the type of solvent and its amount are of great importance. ILs are very often described in the context of solvents with incredible features, such as negligible vapour pressure, high chemical and thermal stability, low flammability, large liquidus range, high ionic conductivity, large electrochemical window, excellent solvation ability of a wide range of compounds, and most of all, possibility of designing for specific demands (due to an appropriate selection of cation and anion). However, there are also numerous studies where the authors pay attention to the environmental problem due to poor biodegradability, toxicity, and methods of preparation and degradation after use [60–65]. Nevertheless, the lack of data for IL characterization in the context of greenness assessment is a serious problem. It may make the evaluation difficult and in some sense inaccurate and inappropriate in flat assertions on ILs as alternative green solvents [66]. Hence, a large number of publications on predicting the properties of ionic liquids have been performed, as shown in Table 2.

Table 2. Prediction of ionic liquid properties by applying chemometric tools—summarized exemplary studies.

Predicted Property	Chemometric Tools	Evaluated Objects	Way of Estimation	Ref.
Carbon dioxide solubility	RB, MLP, MQR, MPE	• [emim][PF6] • [hmim][PF6] • [bmim][BF4] • [hmim][BF4] • [omim][BF4]	experimental thermodynamic data and molecular structure information	Torrecilla et al. (2008) [67]
Melting point	ANN	97 imidazolium salts with varied anions	14 molecular descriptors	Torrecilla et al. (2008) [68]
Viscosity	ANN	58 ionic liquids at several temperatures	molecular mass of the anion and cation, the mass connectivity index, and the density at 298 K	Valderrama et al. (2011) [69]
Electric conductivity	MLR, BP-ANN	35 ILs at different temperatures	structural descriptors	Cao et al. (2013) [70]
Density	ER, ANN	mixtures of ionic liquids and molecular solvents (water, alcohols, ketones, ethers, hydrocarbons, esters, and acetonitrile)	molar mass, critical volume, temperature, acentric factor of each component of the IL mixtures	Huang et al. (2014) [71]
Design of ionic liquids	PCA, CA	172 ILs	structural similarity and identification of structure aspects responsible for a given IL physicochemical properties (viscosity, n-octanol–water partition coefficient, solubility and enthalpy of fusion via ILPC predictor)	Barycki et al. (2016) [72]
Lipophilicity	QSPR, PCA	selected ionic liquid (only imidazolium-based cations)	comparison of hydrophobic or hydrophilic character according to some methods: chromatographic analysis, statistical, and chemometric approach	Studzińska et al. (2007) [73]
Toxicity	PCR, PLS, decision tree(s) model	various combinations of cations (imidazole, pyridinium, quinolinium, ammonium, phosphonium) and anions (BF_4, Cl, PF_6, Br, CFNOS, NCN_2, $C_6F_{18}PBF_4$, $C_6F_{18}P$)	molecular descriptors and EC_{50} concentrations for inhibition of acetylcholinesterase	Ž. Kurtanjek (2014) [74]

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 2. *Cont.*

Predicted Property	Chemometric Tools	Evaluated Objects	Way of Estimation	Ref.
Toxicity	PCA	375 ILs with six different types of cations namely, imidazolium, ammonium, phosphinium, pyridinium, pyrolidinium, and sulfonium	multiple endpoints for various organisms based on WHIM descriptors	Sosnowska et al. (2014) [75]
Toxicity	QSAR, MLR, ELM	160 ILs with 57 cations and 21 anions	toxicity towards AChE based on theSEP area and the screening charge density distribution area (S σ) descriptors	Zhu et al. (2019) [76]
Toxicity	QSPR, MLR	304 ILs of different combinations of 8 cations (ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium, quinolinium) and 12 anions (chloride, bis(trifluoromethylsulfonyl) amide, bromide, iodide ion, sulfonate, borate, phosphate, fatty acid, dicyanamide, formate, thiocyanate, acetate, etc.)	toxicity against leukaemia rat cell line IPC-81 (logEC_{50}) based on 33 descriptors describing the structural features of ionic liquids related to toxicity (i.e., chain length of the cationic head group)	Wu et al. (2020) [77]

Abbreviations: AChE—Acetylcholinesterase; BP-ANN—Back Propagation Artificial Neural Network; ELM—Extreme Learning Machine; ILPC—Ionic Liquid PhysicoChemical; MPE—Mean Prediction Error; SEP—Surface Electrostatic Potential; WHIM descriptors—Weighted Holistic Invariant Molecular descriptors

The prediction of IL properties may be successfully conducted using different chemometric tools. It is mostly proved by a comparison of predicted values with experimental/literature ones, such as in estimation melting point [68] or viscosity [69]. Moreover, it sometimes happens that one technique is applied to select appropriate descriptors; then another one is used for the prediction of a particular feature. In some cases, the applications of several chemometric methods are compared, as presented with the example of carbon dioxide solubility [67], electric conductivity [70], density [71], and toxicity [74]. In first case, nonlinear models, such as RB (radial basis network) and MLP (multilayer perceptron) turned out to be more adequate when the mathematical complexity of the model is not important or a high accuracy is necessary. On the other hand, MQR (multiple quadratic regression) is recommended for faster computation if the operating conditions are stable. Prediction of electric conductivity using an ANN model is more favourable than using an MLR model due to more rational nonlinear modelling. An interesting approach is presented for the latter case—toxicity prediction based on molecular descriptors and EC_{50} concentrations for the inhibition of acetylcholinesterase using a decision tree(s) model. Decision tree(s) models (R = 0.992) significantly outperform other models, such as PCR (principal component regression) and PLS (R = 0.62 and 0.64), for numerical predictions of EC_{50} concentrations and the classification of ILs into four levels of toxicity. The visualization of this division into four classes is presented in Figure 5.

It is not always the rule that one of the models used is clearly better than the others. Very often, all of them or some of them lead to satisfactory results, which is described by Huang et al. [71] for density prediction. ER (extended Riedel) and ANN proved to be accurate in a wide range of compositions and temperatures. However, the ER model is a better alternative because it can be used directly without any adjustable parameter and computer-aided program. Sometimes satisfactory results may be obtained by the application several chemometric tools, one by one. Barycki et al. proposed the application of PCA for the definition of the distribution trends of four IL properties dependently on their structures. Then CA is used to provide some detailed information concerning IL distribution [72]. It is also worth noting that chemometrics may be the basis for developing other tools. According to the observed strong relationship between the variance in the observed toxicity and the cations' descriptors, a toxicity ranking index based on the structural similarity of cations (TRIC) for initial toxicity screening studies

MOST WIEDZY Downloaded from mostwiedzy.pl

of ILs has been developed [75]. However, the use of TRIC cannot be individual. It is limited to the prediction of toxicity endpoints used in its development.

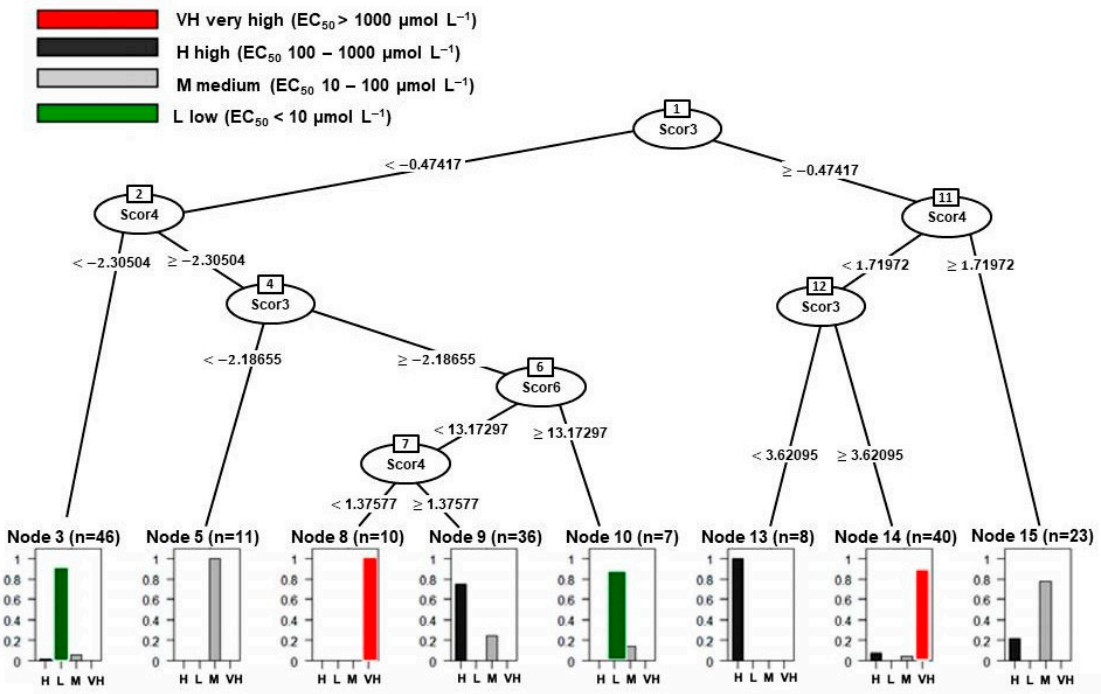

Figure 5. Decision tree model for predictions based on the IL classification of ln(EC_{50}) into four toxicity categories. Reproduced from Ref. Chemometric versus Random Forest Predictors of Ionic Liquid Toxicity (Kurtanjek [74]).

One of the most frequently predicted environmental parameters is toxicity, which may be noticed due to the visible trend in IL properties' prediction analysis as summarized above [74–77]. It is expressed by different endpoints towards various organisms. Toxicity assessment is very important from green chemistry's point of view. Some examples of studies concerning the prediction of toxicity for selected chemicals as potential pollutants are summarized in Table 3.

Table 3. Examples of toxicity prediction for different groups of chemical compounds by applying chemometric tools.

Chemical Compound	Chemometric Tool	Organism	Toxicity Results	Ref.
Metals as: TI, Cd, and Ag	RSM	growth of cabbage seedlings	Ag is observed to be the most toxic, while Tl and Cd, although toxic, exhibited fairly similar effects.	Allus et al. (1988) [78]
Nitrobenzenes	LS-SVM, QSPR, PLS, PCA, GA-PLS, MLR	*Tetrahymena pyriformis* [79]	n/a	Niazi et al. (2007) [80]
Organic compounds (including some pharmaceuticals)	QSTR, PLS	human (human lethal concentration)	The ETA models suggest that the toxicity increases with bulk, chloro (hydrophobic) functionality, presence of heteroatoms within a chain or a ring and unsaturation, and decreases with hydroxyl (polar) functionality and branching.	Roy and Ghosh (2008) [81]
Chemical compounds	SVM, ANN	*Pimephales promelas*	n/a	Tan et al. (2010) [82]

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 3. *Cont.*

Chemical Compound	Chemometric Tool	Organism	Toxicity Results	Ref.
Organic chemicals	QSAR, MLR, PLS, GFA, G/PLS	*Daphnia magna*	Higher lipophilicity and electrophilicity, less negative charge surface area and presence of ether linkage, hydrogen bond donor groups and acetylenic carbons are responsible for greater toxicity of chemicals. Diversity in chemically different compounds in mechanisms of toxic actions is observed.	Kar and Roy (2010) [83]
Per- and polyfluorinated (PFCs) chemicals	PCA, QSAR, MLR, GA,	rodents (oral)	The importance of negative hydrophobicity and positive electronegativity for the overall toxicity of PFCs for rodents.	Bhhatarai and Gramatica (2011) [84]
Herbicides	ANN, QSAR	rat (oral)	n/a	Hamadache et al. (2016) [85]
Agrochemicals (fungicides, herbicides, insecticides, and microbiocides)	QSAR	*Daphnia magna*	The toxicity increases with lipophilicity and decreases with polarity.	Khan et al. (2019) [86]
Silver nanoparticles	CA, PCA	links between ecotoxicity and physicochemical features (*Daphnia magna*, *Thamnocephalus platyurus*, and *Daphnia galeata*)	n/a	Nedyalkova et al. (2017) [87]
Silver nanoparticles	PCA, CA, *k*-means clustering, MLR	*Daphnia magna*, *Thamnocephalus platyurus*, *Escherichia coli*, *Pseudomonas fluorescens*, *Pseudokirchneriella subcapitata*, *Pseudomonas putida*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, mammalian cells, algae, yeast, and fungi	The relation AT/ZP (acute toxicity measure, EC_{50}/LC_{50}/zeta potential of nanomaterial in the test) is not very indicative for the toxic impact of the AgNPs studied.	Nedyalkova et al. (2019) [88]

Abbreviations: ETA—Extended Topochemical Atom; GA-PLS—Genetic Algorithm-Partial Least Square; GFA—Genetic Function Approximation; G/PLS—Genetic Partial Least Squares; QSTR—Quantitative Structure Toxicity Relationship; RSM—Response Surface Methodology

Based on the above studies, the methods from the family of QSAR models are willingly used for toxicity prediction. They allow for the achievement of good results and provide more than 95% predictions for agrochemical toxicity towards *Daphnia magna* [86]. QSAR models are often supported by chemometrics; however, there is no dominant chemometric tool that ensures the best prediction ability. In nitrobenzene toxicity prediction, LS-SVM (least squares-support vector machines) turned out to be the more powerful method than the rest [80]. The reason is fact that LS-SVM (for quantum chemical descriptors) drastically enhances the ability of prediction in QSAR (prediction of IGC_{50} toxicity) studies superior to MLR and PLS.

Other parameters of great importance for the assessment of the environmental risk associated with the use of chemical compounds are the partition coefficients towards different media. They allow for the estimation of the affinity of a particular chemical compound to a selected phase system. Octanol–air or octanol–water partition coefficients may be applied as the predictors of the partitioning of semivolatile organic chemicals to aerosols or a chemical compound to dissolve in fats, oils, lipids, and nonpolar solvents, respectively. Moreover, the value of the latter coefficient could provide information on the potential for bioaccumulation as well as in persistent compounds undergoing biomagnification [89,90]. In Table 4, a list of studies on the chemometric prediction of partition coefficients in presented.

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 4. Prediction of partition coefficients by applying chemometric tools—summarized exemplary studies.

Partition Coefficient	Chemometrics Tool	Evaluated Objects	Way of Estimation	Ref.
n-octanol-ir partition coefficient	QSAR/QSPR, PCA, PCR	chloronaphthalene congeners	190 different quantum-chemical, thermodynamical, and topological characteristics of chloronaphthalenes as descriptors	Puzyn and Falandysz (2005) [91]
Water-polydimethylsiloxane partition coefficient	QSPR, GA, MLR, ANN	organic compounds	molecular descriptors: minimum atomic orbital electronic population, Kier shape index, polarity parameter/square distance, and complementary information content	Golmohammadi and Dashtbozorgi (2010) [92]
n-octanol-water partition coefficient	LS-SVM, QSPR, MLR, SVR, ANN	organic compounds (derivative phenolic compounds)	n/a	Goudarzi and Goodarzi (2008) [21]
n-octanol-water partition coefficient	QSPR, mRMR-GA-SVR	aromatic compounds	68 molecular descriptors derived solely from the structures of the aromatic compounds	Yang et al. (2008) [93]
n-octanol-water partition coefficient	QSPR, MLR/PLS/RBF-PLS	organic compounds		Goudarzi and Goodarzi (2010) [94]
n-octanol-water partition coefficient	QSAR, CoMFA, CoMSIA	21 polychlorinated naphthalenes (PCNs) congener	3D descriptors according to the experimental values of logKOW for 21 PCNs	Gu et al. (2017) [95]
polyurethane foam-air partition coefficients	QSPR, MLR, ANN, SVM	170 organic compounds comprising 9 distinct classes (PAHs, benzenes, esters, aliphatic and cyclic hydrocarbons, polychlorinated biphenyls, musk, nitrogen and sulphur compounds, pesticides, other compounds)	368 molecular descriptors	Zhu et al. (2020) [96]

The information summarized in Table 4 shows that the application of the combination the QSPR model and chemometric methods is common. In the estimation of the water–polydimethylsiloxane [92] and n-octanol–water [21] partition coefficients of organic compounds, the best techniques turned out to be ANN and LS–SVM, respectively. This results in a significant improvement in prediction quality. Two years later, Goudarzi and Goodarzi [94] conducted a prediction of the n-octanol–water partition coefficient for the same dataset of organic compounds but using different techniques, namely, MLR, PLS, and RBF-PLS (radial basic function-partial least squares). This time, due to flexible mapping of the selected features by manipulating their functional dependence implicitly unlike regression analysis, RBF-PLS is considered to be better than MLR and PLS models.

An interesting approach for the n-octanol–water partition coefficient for polychlorinated naphthalenes (PCNs) congener is proposed by Gu et al. [95], where QSAR is combined with comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). These two models are dedicated to 3D-QSAR approaches, where the 3D conformation property of compounds has to be taken into account (possibility of exploring, visualizing a structural information, and designing new compounds with particular properties). Although the results of both models show good prediction ability, the CoMSIA model is better in designing new types of compound molecules due to the higher number of descriptors. The readiness of chemicals to concentrate in organisms when the compounds are present in the environment may also be defined by bioconcentration factor (BCF). Prediction of this environmental property for some organic compounds using QSAR combined with GA-ANN (for the selection of appropriate descriptors) is proposed by Fatemi et al. [29].

MOST WIEDZY Downloaded from mostwiedzy.pl

6. Conclusions

There are various chemometric tools that can give benefits in terms of green chemistry. Application of even the simplest and well-known techniques for dimensionality reduction and grouping of objects or variables, such as CA or PCA, may result in significant advantages. These are the treatments for missing data, so chemical parameters are predicted without performing problematic, time-consuming, and material-demanding measurements. Even finding correlations in the dataset can give clues on the selection of proper materials. In this way, there is a possibility of estimation of the environmental fate of chemical compounds if the predicted datapoints refer to their behaviour in the environment. Reducing the number of elements in the dataset by grouping objects according to similarities leads to a preselection of objects for further consideration by more detailed studies. Selection of chemical compounds with similar characteristics by chemometric techniques is helpful in finding greener alternatives, compounds that are less problematic but retain their desired features. Multivariate statistics are successfully applied in green chemistry studies, and their significance is expected to be growing.

Author Contributions: Conceptualization, M.B. and M.T.; data curation, M.B.; writing—original draft preparation, M.B. and M.T.; writing—review and editing, M.T.; supervision, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brereton, R.G. Chemometrics in analytical chemistry. A review. *Analyst* **1987**, *112*, 1635–1657. [CrossRef]
2. Kiralj, R.; Ferreira, M.M.C. The past, present, and future of chemometrics worldwide: Some etymological, linguistic, and bibliometric investigations. *J. Chemom.* **2006**, *20*, 247–272. [CrossRef]
3. Brereton, R.G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J.M.; Walczak, B.; Tauler, R. Chemometrics in analytical chemistry—Part I: History, experimental design and data analysis tools. *Anal. Bioanal. Chem.* **2017**, *409*, 5891–5899. [CrossRef] [PubMed]
4. Santos, M.C.; Nascimento, P.; Guedes, W.N.; Pereira-Filho, E.R.; Filletti, É.R.; Pereira, F.V. Chemometrics in analytical chemistry—An overview of applications from 2014 to 2018. *Eclética Química J.* **2019**, *44*, 11–25. [CrossRef]
5. Defernez, M.; Kemsley, E. The use and misuse of chemometrics for treating classification problems. *TrAC Trends Anal. Chem.* **1997**, *16*, 216–221. [CrossRef]
6. Rácz, A.; Bajusz, D.; Héberger, K. Chemometrics in Analytical Chemistry. In *Applied Chemoinformatics: Achievements and Future Opportunities*; Engel, T., Gasteiger, J., Eds.; Wiley-VCH: Weinheim, Germany, 2018; pp. 471–499. [CrossRef]
7. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.
8. Vončina, D.B. Chemometrics in analytical chemistry. *Nova Biotechnol.* **2009**, *211*, 211–216.
9. Camacho, J.; Picó, J.; Ferrer, A. Data understanding with PCA: Structural and Variance Information plots. *Chemom. Intell. Lab. Syst.* **2010**, *100*, 48–56. [CrossRef]
10. Huang, B.K.; Huang, L.Q.; Qin, L.P. Cluster analysis of NIR fingerprint of four species plants in *Valeriana officinalis* L. *J. Chin. Med. Mater.* **2008**, *31*, 1494–1496.
11. Mohsin, G.F.; Schmitt, F.-J.; Kanzler, C.; Hoehl, A.; Hornemann, A. PCA-based identification and differentiation of FTIR data from model melanoidins with specific molecular compositions. *Food Chem.* **2019**, *281*, 106–113. [CrossRef]
12. Guo, Y.; Ni, Y.; Kokot, S. Evaluation of chemical components and properties of the jujube fruit using near infrared spectroscopy and chemometrics. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2016**, *153*, 79–86. [CrossRef]
13. Massart, D.L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*; John Wiley & Sons: New York, NY, USA, 1983.

MOST WIEDZY Downloaded from mostwiedzy.pl

14. Liang, Y.Z.; Xie, P.; Chan, K. Quality control of herbal medicines. *J. Chromatogr. B* **2004**, *812*, 53–70. [CrossRef]
15. Wesołowski, M.; Konieczynski, P. Thermoanalytical, chemical and principal component analysis of plant drugs. *Int. J. Pharm.* **2003**, *262*, 29–37. [CrossRef]
16. Bansal, A.; Chhabra, V.; Rawal, R.K.; Sharma, S. Chemometrics: A new scenario in herbal drug standardization. *J. Pharm. Anal.* **2014**, *4*, 223–233. [CrossRef]
17. Li, S.; Ng, T.-T.; Yao, Z.P. Quantitative analysis of blended oils by matrix-assisted laser desorption/ionization mass spectrometry and partial least squares regression. *Food Chem.* **2020**, *334*, 127601. [CrossRef] [PubMed]
18. Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemom.* **1996**, *10*, 463–482. [CrossRef]
19. Denham, M.C. Choosing the number of factors in partial least squares regression: Estimating and minimizing the mean squared error of prediction. *J. Chemom.* **2000**, *14*, 351–361. [CrossRef]
20. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
21. Goudarzi, N.; Goodarzi, M. Prediction of the logarithmic of partition coefficients (log P) of some organic compounds by least square-support vector machine (LS-SVM). *Mol. Phys.* **2008**, *106*, 2525–2535. [CrossRef]
22. Lee, H.; Park, Y.M.; Lee, S. Principal Component Regression by Principal Component Selection. *Commun. Stat. Appl. Methods* **2015**, *22*, 173–180. [CrossRef]
23. Tong, W.; Hong, H.; Xie, Q.; Shi, L.; Fang, H.; Perkins, R. Assessing QSAR Limitations—A Regulatory Perspective. *Curr. Comput. Aided Drug Des.* **2005**, *1*, 195–205. [CrossRef]
24. Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504. [CrossRef] [PubMed]
25. Hibbert, D.B. Genetic algorithms in chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277–293. [CrossRef]
26. Mehrotra, K.; Mohan, C.K.; Ranka, S. *Elements of Artificial Neural Networks*; MIT Press: Cambridge, MA, USA, 2000.
27. Attia, K.A.; Nassar, M.W.; El-Zeiny, M.B.; Serag, A. Effect of genetic algorithm as a variable selection method on different chemometric models applied for the analysis of binary mixture of amoxicillin and flucloxacillin: A comparative study. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2016**, *156*, 54–62. [CrossRef] [PubMed]
28. Golmohammadi, H.; Dashtbozorgi, Z. Quantitative structure–property relationship studies of gas-to-wet butyl acetate partition coefficient of some organic compounds using genetic algorithm and artificial neural network. *Struct. Chem.* **2010**, *21*, 1241–1252. [CrossRef]
29. Fatemi, M.H.; Jalali-Heravi, M.; Konuze, E. Prediction of bioconcentration factor using genetic algorithm and artificial neural network. *Anal. Chim. Acta* **2003**, *486*, 101–108. [CrossRef]
30. Gere, A.; Rácz, A.; Bajusz, D.; Károly, H. Multicriteria decision making for evergreen problems in food science by sum of ranking differences. *Food Chem.* **2020**, 128617. [CrossRef]
31. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
32. Cao, D.-S.; Dong, J.; Wang, N.-N.; Wen, M.; Deng, B.-C.; Zeng, W.-B.; Xu, Q.-S.; Liang, Y.-Z.; Lu, A.-P.; Chen, A.F. In silico toxicity prediction of chemicals from EPA toxicity database by kernel fusion-based support vector machines. *Chemom. Intell. Lab. Syst.* **2015**, *146*, 494–502. [CrossRef]
33. Li, X.; Kong, W.; Shi, W.; Shen, Q. A combination of chemometrics methods and GC–MS for the classification of edible vegetable oils. *Chemom. Intell. Lab. Syst.* **2016**, *155*, 145–150. [CrossRef]
34. García, J.I.; Garcia-Marin, H.; Mayoral, J.A.; Pérez, P. Quantitative structure–property relationships prediction of some physico-chemical properties of glycerol based solvents. *Green Chem.* **2013**, *15*, 2283–2293. [CrossRef]
35. Tobiszewski, M.; Tsakovski, S.; Simeonov, V.; Namieśnik, J.; Pena-Pereira, F. A solvent selection guide based on chemometrics and multicriteria decision analysis. *Green Chem.* **2015**, *17*, 4773–4785. [CrossRef]
36. Alfonsi, K.; Colberg, J.; Dunn, P.J.; Fevig, T.; Jennings, S.; Johnson, T.A.; Kleine, H.P.; Knight, C.; Nagy, M.A.; Perry, D.A.; et al. Green chemistry tools to influence a medicinal chemistry and research chemistry based organisation. *Green Chem.* **2008**, *10*, 31–36. [CrossRef]
37. Henderson, R.K.; Jiménez-González, C.; Constable, D.J.C.; Alston, S.R.; Inglis, G.G.A.; Fisher, G.; Sherwood, J.; Binks, S.P.; Curzons, A.D. Expanding GSK's solvent selection guide–embedding sustainability into solvent selection starting at medicinal chemistry. *Green Chem.* **2011**, *13*, 854–862. [CrossRef]

MOST WIEDZY Downloaded from mostwiedzy.pl

38. Hargreaves, C.R.; Manley, J.B. ACS GCI Pharmaceutical Roundtable–Collaboration to Deliver a Solvent Selection Guide for the Pharmaceutical Industry. 2008. Available online: http://www.acs.org/content/dam/acsorg/greenchemistry/industriainnovation/roundtable/solvent-selection-guide.pdf (accessed on 3 August 2020).
39. Prat, D.; Pardigon, O.; Flemming, H.-W.; Letestu, S.; Ducandas, V.; Isnard, P.; Guntrum, E.; Senac, T.; Ruisseau, S.; Cruciani, P.; et al. Sanofi's Solvent Selection Guide: A Step Toward More Sustainable Processes. *Org. Process. Res. Dev.* **2013**, *17*, 1517–1525. [CrossRef]
40. Prat, D.; Hayler, J.; Wells, A. A survey of solvent selection guides. *Green Chem.* **2014**, *16*, 4546–4551. [CrossRef]
41. Sels, H.; De Smet, H.; Geuens, J. SUSSOL—Using Artificial Intelligence for Greener Solvent Selection and Substitution. *Molecules* **2020**, *25*, 3037. [CrossRef] [PubMed]
42. Papa, E.; Gramatica, P. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. *Green Chem.* **2010**, *12*, 836–843. [CrossRef]
43. Chastrette, M.; Rajzmann, M.; Chanon, M.; Purcell, K.F. Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters. *J. Am. Chem. Soc.* **1985**, *107*, 1–11. [CrossRef]
44. Dutkiewicz, M. Classification of organic solvents based on correlation between dielectric β parameter and empirical solvent polarity parameter ENT. *J. Chem. Soc. Faraday Trans.* **1990**, *86*, 2237–2241. [CrossRef]
45. Pytela, O. A new classification of solvents based on chemometric empirical scale of parameters. *Collect. Czechoslov. Chem. Commun.* **1990**, *55*, 644–652. [CrossRef]
46. Gramatica, P.; Navas, N.; Todeschini, R. Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of molecular descriptors. *TrAC Trends Anal. Chem.* **1999**, *18*, 461–471. [CrossRef]
47. Pushkarova, Y.; Kholin, Y.V. A procedure for meaningful unsupervised clustering and its application for solvent classification. *Cent. Eur. J. Chem.* **2014**, *12*, 594–603. [CrossRef]
48. Levet, A.; Bordes, C.; Clément, Y.; Mignon, P.; Chermette, H.; Forquet, V.; Morell, C.; Lantéri, P. Solvent database and in silico classification: A new methodology for solvent substitution and its application for microencapsulation process. *Int. J. Pharm.* **2016**, *509*, 454–464. [CrossRef] [PubMed]
49. Guidea, A.; Sârbu, C. Fuzzy characterization and classification of solvents according to their polarity and selectivity. A comparison with the Snyder approach. *J. Liq. Chromatogr. Relat. Technol.* **2020**, *43*, 336–343. [CrossRef]
50. Salahinejad, M. Application of classification models to identify solvents for single-walled carbon nanotubes dispersion. *RSC Adv.* **2015**, *5*, 22391–22398. [CrossRef]
51. Katritzky, A.R.; Fara, D.C.; Kuanar, M.; Hür, E.; Karelson, M. The Classification of Solvents by Combining Classical QSPR Methodology with Principal Component Analysis. *J. Phys. Chem. A* **2005**, *109*, 10323–10341. [CrossRef]
52. Tobiszewski, M.; Nedyalkova, M.; Madurga, S.; Pena-Pereira, F.; Namieśnik, J.; Simeonov, V. Pre-selection and assessment of green organic solvents by clustering chemometric tools. *Ecotoxicol. Environ. Saf.* **2018**, *147*, 292–298. [CrossRef]
53. Nedyalkova, M.; Sârbu, C.; Tobiszewski, M.; Simeonov, V. Fuzzy Divisive Hierarchical Clustering of Solvents According to Their Experimentally and Theoretically Predicted Descriptors. *Symmetry* **2020**, *12*, 1763. [CrossRef]
54. González-Álvarez, J.; Mangas-Alonso, J.J.; Arias-Abrodo, P.; Gutiérrez-Álvarez, M.D. A chemometric approach to characterization of ionic liquids for gas chromatography. *Anal. Bioanal. Chem.* **2014**, *406*, 3149–3155. [CrossRef]
55. Izadiyan, P.; Fatemi, M. Chemometric classification of 227 Ionic Liquids and their related salts according to their toxicities to Rat Cell Lines. In Proceedings of the Iranian Biennial Chemometrics Seminar, Tabriz, Iran, 9–10 November 2011.
56. Lesellier, E. Σpider diagram: A universal and versatile approach for system comparison and classification: Application to solvent properties. *J. Chromatogr. A* **2015**, *1389*, 49–64. [CrossRef]
57. Adamska, K.; Voelkel, A.; Héberger, K. Selection of solubility parameters for characterization of pharmaceutical excipients. *J. Chromatogr. A* **2007**, *1171*, 90–97. [CrossRef] [PubMed]
58. Sild, S.; Piir, G.; Neagu, D.; Maran, U. Storing and Using Qualitative and Quantitative Structure–Activity Relationships in the Era of Toxicological and Chemical Data Expansion. In *Big Data in Predictive Toxicology*; Neagu, D., Richarz, A.N., Eds.; Royal Society of Chemistry: London, UK, 2020; pp. 185–213. [CrossRef]

MOST WIEDZY Downloaded from mostwiedzy.pl

59. EPA Website. Available online: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface (accessed on 30 January 2020).
60. Gerrity, D.; Stanford, B.D.; Trenholm, R.A.; Snyder, S.A. An evaluation of a pilot-scale nonthermal plasma advanced oxidation process for trace organic compound degradation. *Water Res.* **2010**, *44*, 493–504. [CrossRef] [PubMed]
61. Coleman, D.; Gathergood, N. Biodegradation studies of ionic liquids. *Chem. Soc. Rev.* **2010**, *39*, 600–637. [CrossRef] [PubMed]
62. Siedlecka, E.M.; Czerwicka, M.; Neumann, J.; Stepnowski, P.; Fernández, J.F.; Thöming, J. Ionic Liquids: Methods of Degradation and Recovery. In *Ionic Liquids: Theory, Properties, New Approaches*; Kokorin, A., Ed.; IntechOpen: Rijeka, Croatia, 2011; pp. 701–722. [CrossRef]
63. Matzke, M.; Thiele, K.; Müller, A.; Filser, J. Sorption and desorption of imidazolium based ionic liquids in different soil types. *Chemosphere* **2009**, *74*, 568–574. [CrossRef] [PubMed]
64. Stepnowski, P.; Mrozik, W.; Nichthauser, J. Adsorption of Alkylimidazolium and Alkylpyridinium Ionic Liquids onto Natural Soils. *Environ. Sci. Technol.* **2007**, *41*, 511–516. [CrossRef]
65. Stolte, S.; Arning, J.; Bottin-Weber, U.; Müller, A.; Pitner, W.-R.; Welz-Biermann, U.; Jastorff, B.; Ranke, J. Effects of different head groups and functionalised side chains on the cytotoxicity of ionic liquids. *Green Chem.* **2007**, *9*, 760–767. [CrossRef]
66. Bystrzanowska, M.; Pena-Pereira, F.; Marcinkowski, Ł.; Tobiszewski, M. How green are ionic liquids?—A multicriteria decision analysis approach. *Ecotoxicol. Environ. Saf.* **2019**, *174*, 455–458. [CrossRef]
67. Torrecilla, J.S.; Palomar, J.; García, J.; Rojo, E.; Rodríguez, F. Modelling of carbon dioxide solubility in ionic liquids at sub and supercritical conditions by neural networks and mathematical regressions. *Chemom. Intell. Lab. Syst.* **2008**, *93*, 149–159. [CrossRef]
68. Torrecilla, J.S.; Rodríguez, F.; Bravo, J.L.; Rothenberg, G.; Seddon, K.R.; López-Martin, I. Optimising an artificial neural network for predicting the melting point of ionic liquids. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5826–5831. [CrossRef]
69. Valderrama, J.O.; Muñoz, J.M.; Rojas, R.E. Viscosity of ionic liquids using the concept of mass connectivity and artificial neural networks. *Korean J. Chem. Eng.* **2011**, *28*, 1451–1457. [CrossRef]
70. Cao, Y.; Yu, J.; Song, H.; Wang, X.; Yao, S. Prediction of electric conductivity for ionic liquids by two chemometrics methods. *J. Serbian Chem. Soc.* **2013**, *78*, 653–667. [CrossRef]
71. Huang, Y.; Zhao, Y.; Zeng, S.; Zhang, S.; Zhang, S. Density Prediction of Mixtures of Ionic Liquids and Molecular Solvents Using Two New Generalized Models. *Ind. Eng. Chem. Res.* **2014**, *53*, 15270–15277. [CrossRef]
72. Barycki, M.; Sosnowska, A.; Piotrowska, M.; Urbaszek, P.; Rybińska, A.; Grzonkowska, M.; Puzyn, T. ILPC: Simple chemometric tool supporting the design of ionic liquids. *J. Cheminform.* **2016**, *8*, 40. [CrossRef] [PubMed]
73. Studzińska, S.; Stepnowski, P.; Buszewski, B. Application of Chromatography and Chemometrics to Estimate Lipophilicity of Ionic Liquid Cations. *QSAR Comb. Sci.* **2007**, *26*, 963–972. [CrossRef]
74. Kurtanjek, Ž. Chemometric versus Random Forest Predictors of Ionic Liquid Toxicity. *Chem. Biochem. Eng. Q.* **2014**, *28*, 459–463. [CrossRef]
75. Sosnowska, A.; Barycki, M.; Zaborowska, M.; Rybińska-Fryca, A.; Puzyn, T. Towards designing environmentally safe ionic liquids: The influence of the cation structure. *Green Chem.* **2014**, *16*, 4749–4757. [CrossRef]
76. Zhu, P.; Kang, X.; Zhao, Y.; Latif, U.; Zhang, H. Predicting the Toxicity of Ionic Liquids toward Acetylcholinesterase Enzymes Using Novel QSAR Models. *Int. J. Mol. Sci.* **2019**, *20*, 2186. [CrossRef]
77. Wu, T.; Li, W.; Chen, M.; Zhou, Y.; Zhang, Q. Estimation of Ionic Liquids Toxicity against Leukemia Rat Cell Line IPC-81 based on the Empirical-like Models using Intuitive and Explainable Fingerprint Descriptors. *Mol. Inform.* **2020**, *39*, 2000102. [CrossRef]
78. Allus, M.A.; Brereton, R.G.; Nickless, G. Chemometric studies of the effect of toxic metals on plants: The use of response surface methodology to investigate the influence of Tl, Cd and Ag on the growth of cabbage seedlings. *Environ. Pollut.* **1988**, *52*, 169–181. [CrossRef]
79. Dearden, J.C.; Cronin, M.T.D.; Schultz, T.W.; Lin, D.T. QSAR Study of the Toxicity of Nitrobenzenes toTetrahymena pyriformis. *Quant. Struct. Relatsh.* **1995**, *14*, 427–432. [CrossRef]
80. Niazi, A.; Jameh-Bozorghi, S.; Nori-Shargh, D. Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines. *J. Hazard. Mater.* **2008**, *151*, 603–609. [CrossRef] [PubMed]

MOST WIEDZY Downloaded from mostwiedzy.pl

81. Roy, K.; Ghosh, G. QSTR with Extended Topochemical Atom Indices. 10. Modeling of Toxicity of Organic Chemicals to Humans Using Different Chemometric Tools. *Chem. Biol. Drug Des.* **2008**, *72*, 383–394. [CrossRef] [PubMed]
82. Tan, N.X.; Li, P.; Rao, H.B.; Li, Z.-R.; Li, X.-Y. Prediction of the acute toxicity of chemical compounds to the fathead minnow by machine learning approaches. *Chemom. Intell. Lab. Syst.* **2010**, *100*, 66–73. [CrossRef]
83. Kar, S.; Roy, K. QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors. *J. Hazard. Mater.* **2010**, *177*, 344–351. [CrossRef] [PubMed]
84. Bhhatarai, B.; Gramatica, P. Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. *Mol. Divers.* **2011**, *15*, 467–476. [CrossRef]
85. Hamadache, M.; Hanini, S.; Benkortbi, O.; Amrane, A.; Khaouane, L.; Moussa, C.S. Artificial neural network-based equation to predict the toxicity of herbicides on rats. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 7–15. [CrossRef]
86. Khan, P.M.; Roy, K.; Benfenati, E. Chemometric modeling of Daphnia magna toxicity of agrochemicals. *Chemosphere* **2019**, *224*, 470–479. [CrossRef]
87. Nedyalkova, M.; Donkova, B.V.; Simeonov, V. Chemometrics Expertise in the Links Between Ecotoxicity and Physicochemical Features of Silver Nanoparticles: Environmental Aspects. *J. AOAC Int.* **2017**, *100*, 359–364. [CrossRef]
88. Nedyalkova, M.; Dimitrov, D.; Donkova, B.; Simeonov, V. Chemometric Evaluation of the Link between Acute Toxicity, Health Issues and Physicochemical Properties of Silver Nanoparticles. *Symmetry* **2019**, *11*, 1159. [CrossRef]
89. Waring, M.J. Lipophilicity in drug discovery. *Expert Opin. Drug Discov.* **2010**, *5*, 235–248. [CrossRef]
90. Chen, M.; Borlak, J.; Tong, W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology* **2013**, *58*, 388–396. [CrossRef] [PubMed]
91. Puzyn, T.; Falandysz, J. Computational estimation of logarithm of n-octanol/air partition coefficient and subcooled vapor pressures of 75 chloronaphthalene congeners. *Atmos. Environ.* **2005**, *39*, 1439–1446. [CrossRef]
92. Golmohammadi, H.; Dashtbozorgi, Z. Prediction of water-to-polydimethylsiloxane partition coefficient for some organic compounds using QSPR approaches. *J. Struct. Chem.* **2010**, *51*, 833–846. [CrossRef]
93. Yang, S.-S.; Lu, W.C.; Gu, T.-H.; Yan, L.-M.; Li, G.-Z. QSPR Study of n -Octanol/Water Partition Coefficient of Some Aromatic Compounds Using Support Vector Regression. *QSAR Comb. Sci.* **2008**, *28*, 175–182. [CrossRef]
94. Goudarzi, N.; Goodarzi, M. QSPR study of partition coefficient (Ko/w) of some organic compounds using radial basic function-partial least square (RBF-PLS). *J. Braz. Chem. Soc.* **2010**, *21*, 1776–1783. [CrossRef]
95. Gu, W.; Chen, Y.; Zhang, L.; Li, Y. Prediction of octanol-water partition coefficient for polychlorinated naphthalenes through three-dimensional QSAR models. *Hum. Ecol. Risk Assess. Int. J.* **2017**, *23*, 40–55. [CrossRef]
96. Zhu, T.; Gu, L.; Chen, M.; Sun, F. Exploring QSPR models for predicting PUF-air partition coefficients of organic compounds with linear and nonlinear approaches. *Chemosphere* **2020**, 128962. [CrossRef]

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

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

MOST WIEDZY Downloaded from mostwiedzy.pl