ORIGINAL ARTICLE

Structure–property relationships for solubility of monosaccharides

Lorentz Jäntschi[1]

Abstract

A series of difficulties make inaccessible precise experimental determinations of solubilities in standard conditions for monosaccharides; in water, the monosaccharides may switch from the acyclic to cyclic form and also the cyclic forms can undergo mutarotation. There are many ways to express the structural information as structure descriptors, but the alternatives become fewer when looking for invariants with good identification abilities, and the characteristic polynomial is one of them. The disadvantage of the characteristic polynomial resides in the fact that is defined with disregard of the chemical information coming from the type of the element and the type of the bond. Here, an extension of the characteristic polynomial was used accounting for the chemical information. In water, monosaccharides exist in all forms, but only one is an invariant for all the acyclic form. If it is something in structure which associates the structure information with the solubility, then it is present in all its form including the acyclic one, and therefore, the acyclic forms can be used to derive structure–property relationships. A search for linear relationships expressing the solubility as a function of the structure of the acyclic forms of monosaccharides was conducted by using the extension of the characteristic polynomial. The search used the experimental data available to select the models that are able to estimate the solubility, with each different to the other in terms of the effects considered. Considering the obtained results, the extended characteristic polynomial provides a very good estimation capability for the solubilities of monosaccharides.

Keywords Monosaccharides · Solubilities · Extended characteristic polynomial · Structure–property relationships

Abbreviations

IUPAC SATP	International Union of Pure and Applied Chemistry Standard Ambient Temperature and Pressure
NMR	Nuclear Magnetic Resonance
CID number	PubChem Compound Identifier
eqs.	equations (when are more than one)

Introduction

Sugars are short-chain carbohydrates, with their molecule consisting of carbon (C), hydrogen (H) and oxygen (O) atoms with the general formula $C_m(H_2O)_n$ where $2 \leq m$ (and usually $3 \leq m \leq 7$) and $n \leq m$ (and usually $n = m$ or $n = m - 1$).

Monosaccharides are the simplest carbohydrates, with general formula $(CH_2O)_n$, where n ranges from 2 (diose,

H–(C=O)–(CH$_2$)–OH) to usually 7 (when $n = 3$ are trioses, when $n = 4$ are tetroses, when $n = 5$ pentoses, when $n = 6$ hexoses and when $n = 7$ heptoses). There are 24 monosaccharides (see Table 1) from diose ($n = 2$) to hexoses ($n = 6$). The monosaccharides with lower number of atoms (e.g., $n = 3$ and $n = 4$) may cyclize by dimerization leading to cyclic monosaccharides with $n = 6$ and $n = 8$, respectively, and also, the monosaccharides can join together to form disaccharides. A disaccharide is formed whenever two monosaccharides (identical or not) joined. Two identical monosaccharides can form up to eleven different disaccharides (Paulus and Klockow-Beck 1999), and the numbers increase even more abruptly when different monosaccharides are connected [Schmidt (1986) counted 720 trisaccharides, 34,560 tetrasaccharides and 2,144,640 pentasaccharides] and the consequence is an enormous diversity and complexity in carbohydrate structure and chemistry.

In Table 1, '$n =$' stands for the n from the general formula $(CH_2O)_n$ of the monosaccharides, while the split into aldoses and ketoses is based on the position of the double-bonded oxygen in their structure.

✉ Lorentz Jäntschi
  lorentz.jantschi@gmail.com

[1] Department of Physics and Chemistry, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Table 1 Monosaccharides from diose to hexoses in open-chain (acyclic) form

$n=$	Formula	Aldoses				Ketoses	
2	$C_2H_4O_2$	Glycolaldehyde (hydroxyacetaldehyde)					
3	$C_3H_6O_3$	D-glyceraldehyde				D-dihydroxyacetone	
4	$C_4H_8O_4$	D-erythose		D-threose		D-erythrulose	
5	$C_5H_{10}O_5$	D-ribose	D-arabinose	D-xylose	D-lyxose	D-xylulose	D-ribulose
6	$C_6H_{12}O_6$	D-talose	D-gulose	D-altrose	D-glucose	D-psicose	D-tagatose
		D-galactose	D-idose	D-mannose	D-allose	D-sorbose	D-fructose

When applies, the names are *ose for acyclic and *opyranose for cylic forms (e.g. Glucose - acyclic → Glucopyranose - cyclic)

The solubilities of monosaccharides are reported in both experimental and theoretical studies (Banerjee 1996; Briciu et al. 2010; Sârbu and Briciu 2010; Kot-Wasik et al. 2014), but unfortunately are at different experimental conditions (18 °C or 20 °C (ChemicalBook 2017) or other temperatures, 10 mmHg, 18 mmHg or other different barometric pressures (ChemSpider 2017) in place of IUPAC SATP of 25 °C and 1 bar (McNaught and Wilkinson 1997), or different solvents or solvent mixtures (van Putten et al. 2014)).

Actually, there are very few experimental data on solubilities of monosaccharides reported at IUPAC SATP conditions. Gray et al. (2003) reported solubilities for five monosaccharides, while (Teles et al. 2016) have the same five ones.

In addition, the recent literature is abundant of studies in connection with the solubility of monosaccharides and of a growing interest is not only their solubility in water, but also in water-based solvents and solvent mixtures. Thus, Ye et al. (2017) reported solubility data for butanol–water mixtures, while others reported for ionic aqueous solvent mixtures including water + NaCl (Hernández-Luis et al. 2003; Ghalami-Choobar et al. 2015), water + NaBr (Zhuo et al. 2005), water + NaI (Zhuo et al. 2008), water + LiCl, water + KH$_2$PO$_4$ and water + NaC$_6$H$_{11}$O$_7$ (Banipal et al. 2014a, b, 2015), water + 3-hydroxypropylammonium acetate (Singh et al. 2015), water + 1-hexyl-3-methyl imidazolium chloride (Zafarani-Moattar et al. 2017) and even in

a series of ionic aqueous solvent mixtures (Carneiro et al. 2013). Also, the solubility of other compounds in mixtures of water + monosaccharides is of interest as well—Nain (2016) reporting data for water + D-mannose solvent mixtures. Solubility-related recent studies include solid–liquid and vapor–liquid phases equilibrium data for some monosaccharides as well as some disaccharides (Jónsdóttir et al. 2002) and solid–liquid phase equilibrium data for binary and ternary systems of certain monosaccharides and water (Guo et al. 2017).

The main problem in conducting studies relating the experimental measurements on carbohydrates is the scarcity of structural information from combined factors (difficulties to crystallize and the limitations in NMR analysis (Zwahlen and Vincent 2002)). Another challenge is the fact that usually the researchers conducting the structural determinations are not the same with the ones conducting the property measurements, and by this way, the reliability of the data sources is reduced, since very easily during the experimental treatment, the monosaccharides may switch from the acyclic to cyclic form and the cyclic forms can undergo mutarotation.

Other data which may be of interest reported for monosaccharides include equilibrium constants of their complexes, but also here the information available is scarce; Hacket et al. (1997) reports the equilibrium constants of

complexes between β-cyclodextrin and three out of the 24 monosaccharides listed in Table 1, obviously not enough paired data to do an analysis in the series.

The increasing interest of series-based data including the monosaccharides is confirmed by a recent study of Buttersack (2017) which reports data not only for monosaccharides (most of pentoses and all hexoses) which estimated hydrophobicity by direct measurement of the hydrophobic interaction of carbohydrates and other hydroxy compounds with a C18-modified silica gel column.

The solubilities in standard conditions for 8 out of the 24 monosaccharides listed in Table 1 were involved in this study to obtain relations expressing the solubility as a function of the structure of monosaccharides in their acyclic form.

Material and method

The available data about solubilities of 8 monosaccharides (listed in Table 2) were collected from the literature. For one (fructose) was necessary an extrapolation, while for other two (allose and psicose) a conversion of the units was made.

In water exists all forms but only one is 'unique'—e.g., does not have different conformation states—the acyclic form, and this is the reason for which it was used. If it is something in structure which explains its behavior in water, then it is present in all its forms including the acyclic one. The advantage of using the acyclic form is given by its uniqueness, which allows to do the desired inference in the whole set of structures.

The structural information as 3D geometries for the D-type isomers of acyclic forms was taken from PubChem database (CID numbers of the files given in Table 2). For one monosaccharide (CID 111123 corresponding to D-idose), the 3D geometry were built from its 2D geometry. On the 24 files containing different geometries of monosaccharides were calculated properties using Spartan'14 software in the following configuration: energy calculation with Hartree–Fock (HF) method, 6-31G* dual basis (Steele and Head-Gordon 2010); the infrared (IR) parameters (Pople et al. 1989) were computed too and thermodynamic entities were derived (C_V—molar heat capacity at constant volume H—enthalpy, S—entropy, G—free enthalpy, C_V^0, H^0, S^0, G^0 at 298.15 K and S^{0K}, C_V^{0K}, ZPE—zero point energy—all at 0 K).

Table 2 Monosaccharides experimental solubilities in mole fraction (mol/mol) at IUPAC STAP conditions along with identifiers of their chemical structure (PubChem CID)

No.	Name	PubChem CID	$(CH_2O)_n$	n	Solubility	Note
1	Glycolaldehyde	756	$C_2H_4O_2$	2		
2	D-dihydroxyacetone	670	$C_3H_6O_3$	3		
3	D-glyceraldehyde	751				
4	D-erythrose	94176	$C_4H_8O_4$	4		
5	D-erythrulose	5460177				
6	D-threose	439665				
7	D-arabinose	66308	$C_5H_{10}O_5$	5	0.0816	1
8	D-lyxose	65550				
9	D-ribose	5311110				
10	D-ribulose	151261				
11	D-xylose	644160			0.12953	1
12	D-xylulose	5289590				
13	D-allose	102288	$C_6H_{12}O_6$	6	0.04489	2
14	D-altrose	94780				
15	D-fructose	5984			0.0735	3
16	D-galactose	3037556			0.0432	1
17	D-glucose	107526			0.09447	1
18	D-gulose	167792				
19	D-idose	111123				
20	D-mannose	12305800			0.25884	1
21	D-psicose	90008			0.2266	4
22	D-sorbose	107428				
23	D-tagatose	92092				
24	D-talose	99459				

1—From (Gray et al. 2003); 2—converted from (Kozakai et al. 2015), 47 wt%; 3—extrapolated at 79.3 wt% (from 30°, 40° and 50° to 25°) from (Flood et al. 1996); 4—converted from (Fukada et al. 2010), 2.93 g/g; solubility column gives experimental solubilities in mole fraction (mol/mol)

On the CID files containing the structures listed in Table 2 were calculated the extended characteristic polynomials. Since the procedure of calculation for the extended polynomial is detailed elsewhere (Jäntschi and Bolboacă, 2016; Joiţa and Jäntschi, 2017), here it is given only in brief. The extended characteristic polynomial (ChPE) is calculated on the chemical structure containing only the heavy elements (without hydrogen atoms). The extended characteristic polynomial (ChPE) on a molecule (Mol) on the evaluation point λ is calculated depending on the choice of the atomic property (A_P) and of the metric operator (M_O) by using the identity (I) and the connectivity (C) matrices as the determinant ($|\bullet|$):

$$\text{ChPE}(A_P, M_O, \lambda, \text{Mol}) = \left| \lambda \cdot I(A_P, \text{Mol}) - C(M_O, \text{Mol}) \right| \quad (1)$$

As any other polynomial formula, the evaluation point λ (or the argument of the polynomial) is to be used to evaluate the polynomial in a certain point of the domain of the values; it may take any real value. Equation (1) resembles the classical definition of the characteristic polynomial (ChP = $|\lambda \bullet [\text{Id}] - [\text{Ad}]|$, in which the identity [Id] and adjacency [Ad] matrices of the graph of the Mol molecule are here changed into $[I(A_P, \text{Mol})]$ and $[C(MO, \text{Mol})]$, as it is detailed in Joiţa and Jäntschi (2017).

For a given molecule (Mol), each choice of the atomic property ($A_P \in$ {'A,' 'B,' 'C,' 'D,' 'E,' 'F,' 'G,' 'H'}) and of the metric operator ($M_O \in$ {'c,' 'C,' 'G,' 'G,' 't,' 'T'}) provides a polynomial formula for the molecule.

The encodings for the atomic property (A_P) provide specifications for the type of the element when the undistinguishable values of 1 from the main diagonal of the identity matrix [Id] are replaced with distinguishable values for each element as follows:

- 'A' provides relative (to the last element of period 7, Uuo, A(Uuo) = 294) atomic mass;
- 'B' provides the connection with the classical characteristic polynomial—always 1;
- 'C'—(atomic partial) charges, as available in the PubChem files;
- 'E'—electronegativity relative to Fluorine (4.00) on the Pauling (Pauling 1932) scale;
- 'F'—first ionization potential, relative to the potential of ionization for Hydrogen (1312 kJ/mol);
- 'G'—melting point temperature relative to diamond's allotrope of Carbon (3820 K);
- 'H'—number of attached hydrogen atoms relative to the same for CH_4 (4).

The encodings for the metric operator provides specifications for the bonds when undistinguishable values of 1 from the adjacency matrix ([Ad]) for the {'c', 'g', 't'} alternatives and from the distance matrix ([Di]) for the {'C', 'G', 'T'} alternatives are replaced with distinguishable values as follows:

- Inverse (in Å^{-1}) of the geometrical distances (in Å) replaces 1's in [Ad] (when M_O = 'g') and in [Di] (when M_O = 'G');
- Inverse of the bond order replaces 1's in [Ad] (when M_O = 'c') and in [Di] (when M_O = 'C');
- Provides the connection with the classical characteristic polynomial (on [Ad], when M_O = 't') and on its extension on the distance matrix (on [Di], using its inversed values, when M_O = 'T').

Thus, when $M_O \in T$ {'c,' 'g,' 't'}, Eq. 1 can be rewritten as ChPE(A_P, M_O, λ, Mol) = $|\lambda \bullet \text{Id}(A_P, \text{Mol}) - \text{Ad}(M_O, \text{Mol})|$ and when $M_O \in$ {'C', 'G', 'T'}, then ChPE(A_P, M_O, λ, Mol) = $|\lambda \bullet \text{Id}(A_P, \text{Mol}) - \text{Di}(M_O, \text{Mol})|$, where $I(A_P, \text{Mol})$, $\text{Ad}(M_O, \text{Mol})$ and $\text{Di}(M_O, \text{Mol})$ are functions which replace the values of 1 from identity ([Id]), adjacency ([Ad]) and distance ([Di]) matrices depending on the selected alternatives of A_P and M_O.

The extended characteristic polynomials can be computed on any value of the argument (λ), but like for the replacements of the values of 1, only the values from [−1, 1] range provide contraction mappings (Jäntschi et al. 2016). The evaluation of the polynomial was made in 2001 equally spaced points from [−1, 1], including thus −1, 0 and 1 in this series of evaluation points. The name of the evaluated extended characteristic polynomial was given with eight characters, $L_1 L_2 L_3 L_4 d_1 d_2 d_3 d_4$, where d_1, d_2, d_3, and d_4 are the digits of the representation of the number ranging from 0 to 1000 used to provide the equally spaced points of evaluation. The letters L_1 to L_4 have the following assignment: L_4 is 'N' for negative λ arguments (from −1.000 to −0.001) and 'P' for nonnegative λ arguments (from 0.000 to 1.000), L_3 encodes the connectivity (M_O) alternatives, L_2 encodes the identity (A_P) alternatives, while L_1 encodes a micro-linearization to macro-linearization alternative ('I' leaves the values unchanged, $f(x) = x$, 'R' provides reciprocal values, $f(x) = 1/x$, while 'L' provides the logarithm of the absolute values, $f(x) = \ln(|x|)$). A total number of 288,144 (2001•8•6•3) possible value-based representations of the structure result by this way. For a set of molecules, each valid (non-null) representation provides a structure descriptor.

Two alternatives were considered: to obtain a relationship expressing the solubility from calculated properties (the ones calculated with the Spartan'14 software) and to obtain a relationship expressing the solubility from the evaluations of the extended characteristic polynomial.

Linear regressions were involved, when the search was conducted with the following alternative models [Eq. 2 with one (x) descriptor, Eqs. 3 to 5 with two (x_1 and x_2) descriptors].

$$y \sim \hat{y} = a_0 + a_1 \cdot x \text{ or } \hat{y} = a_1 \cdot x \tag{2}$$

$$y \sim \hat{y} = a_0 + a_2 \cdot x_1 \cdot x_2 \text{ or } \hat{y} = a_2 \cdot x_1 \cdot x_2 \tag{3}$$

$$y \sim \hat{y} = a_0 + a_3 \cdot x_1 + a_4 \cdot x_2 \text{ or } \hat{y} = a_3 \cdot x_1 + a_4 \cdot x_2 \tag{4}$$

$$y \sim \hat{y} = a_0 + a_3 \cdot x_1 + a_4 \cdot x_2 + a_5 \cdot x_1 \cdot x_2$$
$$\text{or } \hat{y} = a_3 \cdot x_1 + a_4 \cdot x_2 + a_5 \cdot x_1 \cdot x_2 \tag{5}$$

Equation (2) is simple linear regression, Eq. (4) is multiple (bi-varied) linear regression with only additive effects among descriptors, and Eq. (3) is multiple (bi-varied) linear regression with only multiplicative effects among descriptors, while Eq. (5) is multiple (bi-varied) linear regression quantifying for both additive and multiplicative effects among descriptors.

Adjusted determination coefficients allow selection of the best explanatory models, and their Fisher Z transformations (Fisher 1915, 1921) allow comparison among them.

The condition to use Eqs. (2)–(5) is that the dependent variable reconstitute a normal distribution. Otherwise, a series of transformations of the data are required (Bolboacă and Jäntschi 2013).

Results and discussion

On the 24 files containing different geometries of monosaccharides, the series of calculated properties are listed in Table 3 (Spartan'14 software, energy calculation with HF 6-31G* dual basis), and their associated thermodynamic quantities are listed in Table 4 (Spartan'14 software, computing IR parameters).

Unfortunately, the solubility listed in Table 2 is a little departed from normality. A histogram of the values easily proves this (see Fig. 1).

The logarithm of the solubilities reconstitute the normality much better (see Fig. 2).

Because of very few data (only 8 values), it is difficult to do a test for normality. The alternatives are Anderson–Darling (AD) and Kolmogorov–Smirnov (KS) (Bolboacă and Jäntschi 2009). The KS test provides a value of 0.24358 ($p = 0.644$) when solubility is tested for normality and a value of 0.14593 ($p = 0.985$) when ln(solubility) is tested for normality. The AD test provides a value of 0.56634 ($p = 0.656$) when solubility is tested for normality and a value of 0.2579 ($p = 0.949$) when ln(solubility) is tested for normality. The p values show that the ln(solubility) is much

closer to normality than the solubility. Therefore, the logarithm transformation was applied to the solubility.

Searching for regressions of type Eqs. (2)–(5) with ln(solubility) (as well as with solubility) as dependent variable and whole pool of properties listed in Tables 3 and 4 as independent variables with potential explanatory power was unsuccessful. Only a very poor association between ln(solubility) and SolvE (solvation energy) was identified ($r = 0.32$, $r^2 = 0.10$, $r_{adj}^2 = 1 - (1 - r^2)(8 - 1)/(8 - 2) = -0.05$), definitely insufficient to be taken into account. Only its intercept is statistically significantly different from zero. A possible explanation is the fact that the solubility (and its logarithm) is almost orthogonal on the other properties (average of the correlation of ln(solubility) with the properties listed in Tables 3 and 4 is 0.05).

Therefore, seeking relationships estimating solubility from structure information provided by the extended characteristic polynomial seems appropriate and is definitely helpful. By applying Eqs. (2)–(5) were selected the best explanatory structure-based descriptors as the ones providing the best association defined by the models of the equations; their names and values for the compounds with known solubility are given in Table 5.

In Table 5, *Selection from* column indicates which alternative (one of the equations: 2a, 2b, 3a, 3b, 4a, 4b, 5a, 5b) identified the descriptor as able to explain the ln(solubility).

As can be seen in Table 5, all selected descriptors came from geometry-based approach ('G' letter in the third position, when the adjacency matrix ([Ad]) is replaced with the distance matrix ([Di]) in the former definition of the characteristic polynomial ($|\lambda \bullet \text{Id}(A_P, \text{Mol}) - \text{Di}(M_O, \text{Mol})|$). It may seem surprising, but is not, since the distance matrix brings further knowledge than the adjacency (let us remember that the 1's are in the same position in the distance matrix as are in the adjacency; the change is on 0's, and some of 0's from adjacency are replaced with nonzero values based on the distances between the atoms). Also this selection of the geometry as the predominant metric explains why the properties derived from energetic calculations given in Tables 3 and 4 fail to estimate the solubility, and there are many geometrical arrangements sharing same energy. Regarding the atomic property, two of them were not selected in the best explanatory descriptors: number of the hydrogen atoms ('H') and relative atomic mass ('A'), and it seems by this way that those have a little influence on the solubility. It may seem not quite logical, when we think about the dissociation process accompanying the dissolving, but if it is taken into account that each hydrogen is differently involved in this dissociation process, then it may seem reasonable that their number does not play an important role while the partial charges ('C') do.

Melting point temperature of the elements ('G') provides the first level of approximation for the association between

Table 3 Calculated molecular properties

Crt	PubChem CID	n	Conf	LUMO	HOMO	D_M	Solv_E	Energy	Energy aq.
1	756	2	9	2.89	−12.40	3.38	−35.051	−227.742725	−227.756075
2	670	3	81	2.67	−12.25	3.28	−52.645	−341.623769	−341.643820
3	751	3	81	2.62	−12.23	3.14	−52.017	−341.618389	−341.638201
4	94176	4	729	2.37	−12.57	2.85	−48.998	−455.499452	−455.518114
5	5460177	4	729	2.43	−12.06	5.55	−71.961	−455.494973	−455.522381
6	439665	4	729	2.07	−12.49	3.37	−57.712	−455.494156	−455.516137
7	66308	5	6561	2.13	−12.53	2.28	−64.690	−569.378402	−569.403042
8	65550	5	6561	2.57	−12.32	2.65	−63.846	−569.380459	−569.404777
9	5311110	5	6561	2.47	−12.41	1.81	−58.726	−569.380968	−569.403336
10	151261	5	6561	2.28	−12.48	2.00	−70.455	−569.374340	−569.401175
11	644160	5	6561	2.41	−12.62	4.19	−68.046	−569.373118	−569.399035
12	5289590	5	6561	2.64	−12.20	5.54	−64.704	−569.382252	−569.406896
13	102288	6	59049	2.29	−12.30	5.05	−73.659	−683.237679	−683.265734
14	94780	6	59049	1.93	−12.16	5.53	−84.332	−683.229153	−683.261273
15	5984	6	59049	2.46	−12.36	7.42	−101.82	−683.244672	−683.283451
16	3037556	6	59049	2.34	−12.53	2.62	−80.301	−683.247123	−683.277708
17	107526	6	59049	2.24	−12.46	3.06	−93.406	−683.236483	−683.272060
18	167792	6	59049	2.13	−12.74	2.77	−79.442	−683.240252	−683.270510
19	111123	6	59049	2.24	−12.56	4.37	−70.993	−683.257349	−683.284389
20	12305800	6	59049	2.27	−12.39	4.74	−93.210	−683.241742	−683.277244
21	90008	6	59049	2.43	−12.15	5.56	−92.166	−683.248260	−683.283365
22	107428	6	59049	2.37	−12.40	5.76	−92.456	−683.248805	−683.284019
23	92092	6	59049	2.20	−12.37	5.61	−90.195	−683.246039	−683.280392
24	99459	6	59049	2.21	−12.30	4.86	−79.783	−683.238285	−683.268673

CID compound identifier from PubChem; *Conf* number of conformers ($=9^{n-1}$); *LUMO* lowest unoccupied molecular orbital energy (in eV); *HOMO* highest occupied molecular orbital energy (in eV); *DM* dipole moment (in Debye); *Energy* total energy (in a.u.; a.u. = Hartrees); *Energy aq.* total energy in solvated form at infinite dilution (in a.u.; a.u. = Hartrees); *Solv_E* solvation energy (from SM5.4 Cramer–Truhlar method– see (Chambers et al. 1996), in kJ/mol). Please note that the solvation energy gives only an estimate, and thus, a possibly better alternative is to account for the number of surrounding water molecules

the chemical structure and solubility (RGGN0477 = 1/GG(−0.477) evaluated polynomial selected by Eq. 2a model in Table 5).

Splitting effects models [Eq. (3), additive, and Eq. (4), multiplicative] select both the same two atomic properties explaining the best association topology of the molecule ('B') and the first ionization potential ('F'), and while taking into consideration both additive and multiplicative effects [in Eq. (5)], the pair of explanatory atomic properties is changed to electronegativity ('E') and partial charges ('C'). Because the same pair of atomic properties was selected for both additive and multiplicative effects models, it seems reasonable to account for both effects, and therefore, the model selected by Eqs. (5) should be used for predictions [see the models listed in Table 6, where UV(λ) is the value of the extended characteristic polynomial UV in λ, when U encodes the atomic property (A_P) and V encodes the distance metric (M_O)].

In Table 6, Equation indicates the type of the equation and it is one of the alternatives given as Eqs. (2)–(5), and

Equation for ln(Solubility) gives the selected polynomials (GG selected by the equation of type 2a, BG and FG selected by Eq. 3a, and so on), their evaluation points (λ) which provides the best explanatory power for the model ($\lambda = −0.477$ for GG polynomial in Eq. 2a, $\lambda = 0.987$ for BG and $\lambda = 0.285$ for FG polynomial for the selections made using Eq. 3a, and so on), as well as the meaning of the encodings for the names of the polynomials (GG polynomial was obtained using 'G' alternative for atomic properties—melting point temperature relative to diamond's allotrope of Carbon, 'G' alternative for metric operator—calculations made using the Cartesian coordinates from geometric 3D model of the molecule) as well as the selected alternative for the connectivity ([C] = [Di], the distance matrix in the case of the model obtained using the equation of type 2a. Also *Equation for ln(Solubility)* gives the coefficients of the equations, as it is $−2.35_{\pm 0.24}$ for Eq. (2a), where the value of −2.35 is the intercept for the model of type 2a, given along with their 95% confidence intervals and at 5% risk being in error, the value of −2.35

Table 4 Calculated thermodynamic properties

Crt	PubChem CID	ZPE	H^0	G^0	S^0	C_V^0	S^{0K}	C_V^{0K}
1	756	162.83	−227.675997	−227.704780	276.66	55.38	29.10	24.94
2	670	244.43	−341.523364	−341.560820	329.83	101.2	45.73	41.57
3	751	244.26	−341.518036	−341.555586	330.66	100.2	45.73	41.57
4	94176	325.24	−455.366208	−455.408600	373.31	142.3	62.36	58.20
5	5460177	322.53	−455.362553	−455.405398	377.30	145.8	70.67	66.52
6	439665	325.25	−455.360904	−455.403355	373.82	142.5	54.04	49.89
7	66308	401.87	−569.213708	−569.261211	418.31	188.0	78.99	74.83
8	65550	401.39	−569.215924	−569.263520	419.12	188.3	78.99	74.83
9	5311110	405.34	−569.215060	−569.262349	416.42	185.1	78.99	74.83
10	151261	400.21	−569.209971	−569.258022	423.13	191.4	95.62	91.46
11	644160	401.19	−569.288653	−569.256244	419.08	188.2	78.99	74.83
12	5289590	401.61	−569.217487	−569.265243	420.54	189.9	87.30	83.14
13	102288	475.89	−683.042335	−683.095083	464.50	236.5	95.62	91.46
14	94780	474.85	−683.034178	−683.086872	464.02	237.3	95.62	91.46
15	5984	470.36	−683.050948	−683.104632	472.74	242.8	112.2	108.1
16	3037556	476.41	−683.051591	−683.104278	463.96	236.3	87.30	83.14
17	107526	472.20	−683.042318	−683.095383	467.29	240.0	103.9	99.77
18	167792	473.93	−683.045484	−683.098501	466.87	238.3	103.9	99.77
19	111123	484.82	−683.059242	−683.110401	450.51	224.9	78.99	74.83
20	12305800	473.14	−683.047310	−683.100329	466.88	238.6	95.62	91.46
21	90008	472.52	−683.053845	−683.107398	471.59	241.0	112.2	108.1
22	107428	471.30	−683.054745	−683.108414	472.61	242.0	112.2	108.1
23	92092	467.84	−683.053104	−683.107233	476.66	245.3	120.6	116.4
24	99459	474.96	−683.043226	−683.096071	465.35	237.4	95.62	91.46

CID Compound identifier from PubChem; *ZPE* zero point energy (in kJ/mol)

The thermodynamic quantities from Table 4 are given for IUPAC STAP ($T_0 = 298.15$ K): H^0 (in a.u.; a.u. = Hartrees), G^0 (in a.u.; a.u. = Hartrees), S^0 (in J mol^{-1} K^{-1}), C_V^0 (in J•mol^{-1}•K^{-1}) and at 0 K: S^{0K} (in J•mol^{-1}•K^{-1}), C_V^{0K} (in J•mol^{-1}•K^{-1})

Fig. 1 Histogram of the solubilities from Table 2
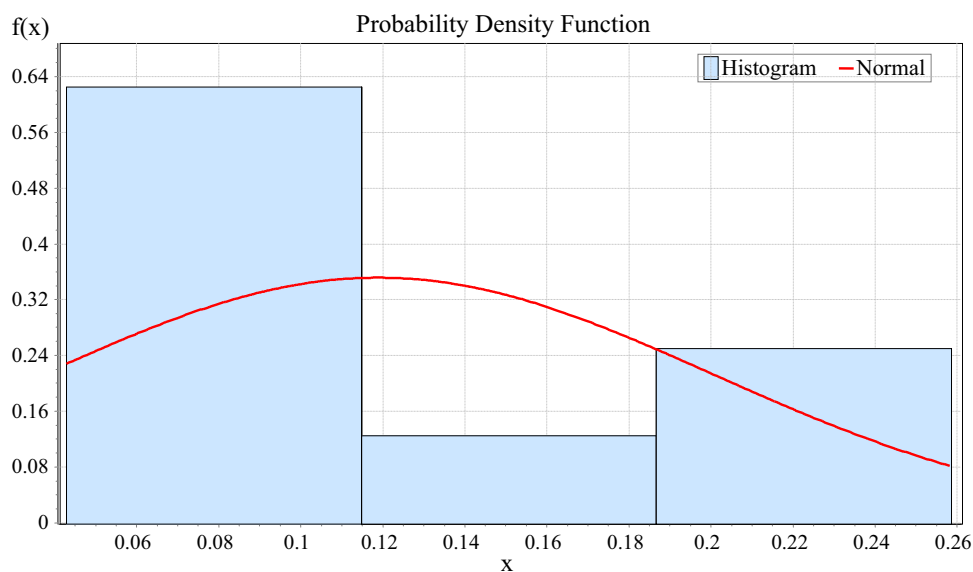
Fig. 2 Histogram of the logarithm of the solubilities from Table 2
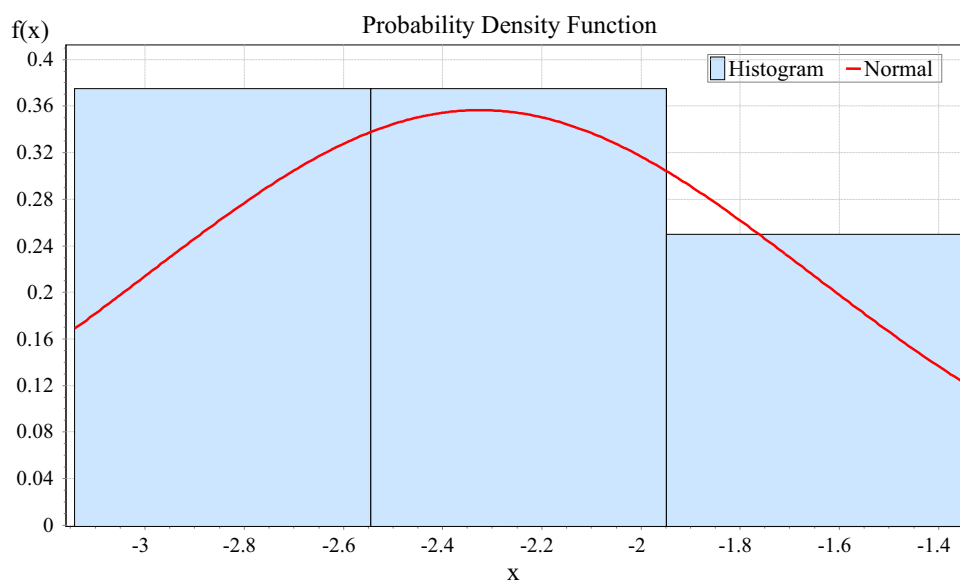
Table 5 Selected descriptors with highest explanatory power

PubChem CID	5984	66308	90008	102288	107526	644160	3037556	12305800	Selection from
RGGN0477$\cdot 10^4$	2.252	0	-2.656	3.135	-2.102	0	3.482	-5.108	Equation 2a
RFGP0285$\cdot 10^1$	8.034	0	-7.632	4.927	-4.533	-1.095	5.302	-9.684	Equation 3a
LBGP0987	0	2.589	2.317	1.246	0	2.668	2.050	1.782	Equation 3a
IFGN0873$\cdot 10^{-6}$	5.687	-3.452	6.554	6.398	6.942	1.954	-7.914	7.595	Equation 4a
RBGN0735$\cdot 10^6$	1.083	0	0	1.609	1.007	0	0	0	Equation 4a
REGP0207$\cdot 10^2$	1.783	0	0	3.332	-4.447	0	1.790	1.325	Equation 5a
RCGP0496$\cdot 10^4$	0	-1.942	2.138	0	-9.849	0	-1.914	4.211	Equation 5a

Table 6 Models with the best explanatory power for solubility of monosaccharides

Equation	Equation for ln(Solubility) with ChPE(A_P, M_O, λ, Mol) $= \|\lambda \cdot I(A_P, \text{Mol}) - C(M_O, \text{Mol})\|$	Statistics
2a	$-2.35_{\pm 0.24} - 2.03_{\pm 0.86} \cdot \dfrac{10^{-5}}{GG(-0.477)}$ GG(-0.477): $A_P = $ G, $M_O = $ 'G', [C] $=$ [Di], $\lambda = -0.477$	$r^2 = 0.848$; $r^2_{\text{adj}} = 0.822$
3a	$-2.49_{\pm 0.15} - 6.40_{\pm 1.52} \cdot \dfrac{\ln(BG(0.987))}{10^3 \cdot FG(0.285)}$ BG(0.987): $A_P = $ 'B', $M_O = $ 'G', [C] $=$ [Di], $\lambda = 0.987$ FG(0.285): $A_P = $ 'F', $M_O = $ 'G', [C] $=$ [Di], $\lambda = 0.285$	$r^2 = 0.947$; $r^2_{\text{adj}} = 0.925$
4a	$-2.21_{\pm 0.07} + 1.12_{\pm 0.12} \cdot \dfrac{FG(-0.873)}{10^7} - 9.78_{\pm 1.0} \cdot \dfrac{10^5}{BG(-0.735)}$ FG(-0.873): $A_P = $ 'F', $M_O = $ 'G', [C] $=$ [Di], $\lambda = -0.873$ BG(-0.735): $A_P = $ 'B', $M_O = $ 'G', [C] $=$ [Di], $\lambda = -0.735$	$r^2 = 0.994$; $r^2_{\text{adj}} = 0.992$
5a	$-2.03_{\pm 0.02} - \dfrac{32.50_{\pm 1.28}}{EG(0.207)} + \dfrac{2462_{\pm 61}}{CG(0.496)} + \dfrac{14848_{\pm 2112}}{EG(0.207) \cdot CG(0.496)}$ EG(0.207): $A_P = $ 'E', $M_O = $ 'G', [C] $=$ [Di], $\lambda = 0.207$ CG(0.496): $A_P = $ 'C', $M_O = $ 'G', [C] $=$ [Di], $\lambda = 0.496$	$r^2 = 0.9997$; $r^2_{\text{adj}} = 0.9996$

is subject to change with ± 0.24. The *Statistics* column gives the determination coefficient (r^2) and adjusted (due to the size of the sample) one (r^2_{adj}).

The equation listed as entry 5a in Table 6 was used for estimation (on the compounds with known solubility) and for prediction (for the rest of the compounds). The results are given in Table 7. As can be seen in Table 7, the estimated

values are close to the measured values for the solubility. The greatest departure is at CID 90008 (corresponding to D-psicose), and it is below 2%, which is safely in the range of the experimental error of the measurements on relative scales (5%).

In Table 7, *ln(Solubility)* column contains the values of the ln(Solubility) as were estimated (for the first 8 monosaccharides) and, respectively, predicted (for the rest of them) by the equation listed as entry 5a in Table 6, while *Solubility* column contains the exponent of those values (exp(ln(solubility) = solubility) ready to be used.

The estimated values show a solubility of D-idose (CID 111123) and D-ribose (CID 5311110) similar to the solubility of D-psicose (CID 90008), while D-dihydroxyacetone (CID 670) seems to have the highest solubility (about 0.41 mol/mol, see Table 7). According to the literature, at 20 °C (not at 25 °C), D-dihydroxyacetone solubility is greater than 930 g/l (SCCS SCCS 2010; more than 10 mol/l) sustaining thus this estimation of its solubility for 25 °C. The second solubility among the selected monosaccharides seems to have D-altrose (about 0.35 mol/mol,

see Table 7), but unfortunately no experimental data are available for comparison.

For glycolaldehyde (CID 756), D-erythrose (CID 94176), D-glyceraldehyde (CID 751), D-erythrulose (CID 5460177), D-threose (CID 439665), D-lyxose (CID 65550), D-ribulose (CID 151261) and D-xylulose (CID 5289590), the estimation provides a solubility similar to the solubility of D-xylose (CID 644160), about 0.13 mol/mol, which is likely since in water the monosaccharides suffer a complex process of mutarotation, leading to a series of different forms, as were shown in (Curtius et al. 1968).

Please note that in the absence of experimental data available, the results provided as predicted solubilities for monosaccharides (the last 16 entries in Table 7) are not validated. In order to be validated, further measurements on solubilities of monosaccharides must be conducted.

Conclusions

The experimental measurements of the solubilities of monosaccharides are scarce and rarely reported at standard conditions. Since solubility is of great importance for the

Table 7 Experimental, estimated and predicted solubilities of monosaccharides

Name	PubChem CID	Ln(Solubility)		Solubility	
		Experimental	Estimated	Experimental	Estimated
D-fructose	5984	− 2.6105	− 2.6095	0.0735	0.07357
D-arabinose	66308	− 2.5059	− 2.5081	0.0816	0.08142
D-psicose	90008	− 1.4846	− 1.5036	0.2266	0.22232
D-allose	102288	− 3.1035	− 3.1129	0.04489	0.044472
D-glucose	107526	− 2.3595	− 2.3592	0.09447	0.094493
D-xylose	644160	− 2.0438	− 2.0300	0.12953	0.131336
D-galactose	3037556	− 3.1419	− 3.1338	0.0432	0.04355
D-mannose	12305800	− 1.3515	− 1.3410	0.25884	0.261576
			Predicted		Predicted
Glycolaldehyde	756		− 2.0300		0.13134
D-dihydroxyacetone	670		− 0.8795		0.41499
D-erythrose	94176		− 2.0300		0.13134
D-glyceraldehyde	751		− 2.0300		0.13134
D-erythrulose	5460177		− 2.0300		0.13134
D-threose	439665		− 2.0300		0.13134
D-lyxose	65550		− 2.0300		0.13134
D-ribose	5311110		− 1.4962		0.22397
D-ribulose	151261		− 2.0300		0.13134
D-xylulose	5289590		− 2.0300		0.13134
D-altrose	94780		− 1.0455		0.35152
D-gulose	167792		− 1.6565		0.19080
D-idose	111123		− 1.4951		0.22424
D-sorbose	107428		− 3.0879		0.04560
D-tagatose	92092		− 2.4668		0.08486
D-talose	99459		− 2.9196		0.05395

biological role of the monosaccharides and the search for property–property relationships expressing the solubility was unsuccessful, a search for a structure–property relationship was conducted.

By using the extended characteristic polynomial, a structure–property relationship was found with great capacity of estimation of the solubility for the monosaccharides ($r^2_{adj} = 0.9996$). The relation suggests that the solubility of the monosaccharides is strongly dependent on the geometry and the atomic partial charges and the electronegativities of the elements play the main role in its expression. The relation was used to predict the solubility for 16 monosaccharides, when plausible solubilities were obtained.

References

Banerjee S (1996) Estimating water solubilities of organics as a function of temperature. Water Res 30(9):2222–2225. https://doi.org/10.1016/0043-1354(96)00306-5

Banipal PK, Aggarwal N, Banipal TS (2014a) Study on interactions of saccharides and their derivatives with potassium phosphate monobasic (1:1 electrolyte) in aqueous solutions at different temperatures. J Mol Liq 196:291–299

Banipal PK, Hundal AK, Aggarwal N, Banipal TS (2014b) Studies on the interactions of saccharides and methyl glycosides with lithium chloride in aqueous solutions at (288.15 to 318.15) K. J Chem Eng Data 59(8):2437–2455. https://doi.org/10.1021/je5001523

Banipal PK, Singh V, Aggarwal N, Banipal TS (2015) Hydration behaviour of some mono-, di-, and tri-saccharides in aqueous sodium gluconate solutions at (288.15, 298.15, 308.15 and 318.15) K: volumetric and rheological approach. Food Chem 168:142–150. https://doi.org/10.1016/j.foodchem.2014.06.104

Bolboacă S-D, Jäntschi L (2009) Distribution fitting 3. Analysis under normality assumption. BUASVMCN Hortic 66(2):698–705

Bolboacă S-D, Jäntschi L (2013) Quantitative structure–activity relationships: linear regression modelling and validation strategies by example. Biomath 2(1):a1309089. https://doi.org/10.11145/j.biomath.2013.09.089

Briciu RD, Kot-Wasik A, Wasik A, Namiesnik J, Sarbu C (2010) The lipophilicity of artificial and natural sweeteners estimated by reversed-phase thin-layer chromatography and computed by various methods. J Chromatogr A 1217:3702–3706. https://doi.org/10.1016/j.chroma.2010.03.057

Buttersack C (2017) Hydrophobicity of carbohydrates and related hydroxy compounds. Carbohyd Res 446–447:101–112. https://doi.org/10.1016/j.carres.2017.04.019

Carneiro AP, Held C, Rodríguez O, Sadowski G, Macedo EA (2013) Solubility of sugars and sugar alcohols in ionic liquids: measurement and PC-SAFT modelling. J Phys Chem B 117(34):9980–9995. https://doi.org/10.1021/jp404864c

Chambers CC, Hawkins CD, Cramer CJ, Truhlar DG (1996) Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. J Phys Chem 100(40):16385–16398. https://doi.org/10.1021/jp9610776

ChemicalBook (2017) D(-)-Fructose. Accessed January 12, 2017. http://chemicalbook.com/chemicalproductproperty_en_cb6139083.htm

ChemSpider (2017) D-(-)-Lyxose. Accessed January 12, 2017. http://chemspider.com/Chemical-Structure.58993.html

Curtius H-C, Völlmin JA, Müller M (1968) Determination of the mutarotation of monosaccharides by gas chromatography - Elucidation of different forms of fructose and sorbose by gas chromatography, infrared and mass spectroscopy. Fresenius' Zeitschrift für Analytische Chemie 243(1):341–349. https://doi.org/10.1007/BF00530708

Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika 10(4):507–521. https://doi.org/10.2307/2331838

Fisher RA (1921) On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron 1:3–32

Flood AE, Addai-Mensah J, Johns MR, White ET (1996) Refractive index, viscosity, density, and solubility in the system Fructose + Ethanol + Water at 30, 40, and 50 C. J Chem Eng Data 41(3):418–421. https://doi.org/10.1021/je950188f

Fukada K, Ishii T, Tanaka K, Yamaji M, Yamaoka Y, Kobashi K-I, Izumori K (2010) Crystal structure, solubility, and mutarotation of the rare monosaccharide D-Psicose. Bull Chem Soc Jpn 83(10):1193–1197. https://doi.org/10.1246/bcsj.20110107

Ghalami-Choobar B, Shafaghat-Lonbar M, Mossayyebzadeh-Shalkoohi P (2015) Activity coefficients determination and thermodynamic modeling of (NaCl + Na₂HCit + glucose + H₂O) system at T = (298.2 and 308.2) K. J Mol Liq 212:922–929. https://doi.org/10.1016/j.molliq.2015.10.051

Gray MC, Converse AO, Wyman CE (2003) Sugar monomer and oligomer solubility. data and predictions for application to biomass hydrolysis. Appl Biochem Biotechnol 105–108:179–193. https://doi.org/10.1385/ABAB:105:1-3:179

Guo L, Wu L, Zhang W, Liang C, Hu Y (2017) Experimental measurement and thermodynamic modeling of binary and ternary solid–liquid phase equilibrium for the systems formed by L-arabinose, D-xylose and water. Chin J Chem Eng 25(10):1467–1472. https://doi.org/10.1016/j.cjche.2017.02.005

Hacket F, Coteron J-M, Schneider H-J, Kazachenko VP (1997) The complexation of glucose and other monosaccharides with cyclodextrins. Can J Chem 75(1):52–55. https://doi.org/10.1139/v97-007

Hernández-Luis F, Amado-González E, Esteso MA (2003) Activity coefficients of NaCl in trehalose-water and maltose-water mixtures at 298.15 K. Carbohydr Res 338(13):1415–1424. https://doi.org/10.1016/s0008-6215(03)00177-0

Jäntschi L, Bolboacă S-D (2016) Extending characteristic polynomial from graphs to molecules. In: International conference on sciences, University of Oradea, Faculty of Sciences, May 13–14, 2016, Oral presentation, Mathematics section, (Friday) May 13, 2016, pp 1500–1520

Jäntschi L, Bálint D, Bolboacă S-D (2016) Multiple linear regressions by maximizing the likelihood under assumption of generalized Gauss-Laplace distribution of the error. Comput Math Methods Med 2016:8578156. https://doi.org/10.1155/2016/8578156

Joiţa DM, Jäntschi L (2017) Extending the characteristic polynomial for characterization of C₂₀ fullerene congeners. Mathematics 5(4):84. https://doi.org/10.3390/math5040084

Jónsdóttir SÓ, Cooke SA, Macedo EA (2002) Modeling and measurements of solid-liquid and vapor-liquid equilibria of polyols and carbohydrates in aqueous solution. Carbohydr Res 337(17):1563–1571. https://doi.org/10.1016/S0008-6215(02)00213-6

Kot-Wasik A, Wasik A, Namiesńik J, Sârbu C, Naşcu-Briciu RD (2014) Retention modeling of some saccharides separated on an amino column. J Liq Chromatogr Relat Technol 37(10):1383–1396. https://doi.org/10.1080/10826076.2013.789806

Kozakai T, Fukada K, Kuwatori R, Ishii T, Senoo T, Izumori K (2015) Aqueous phase behavior of the rare monosaccharide D-Allose and X-ray crystallographic analysis of D-Allose dihydrate. Bull Chem Soc Jpn 88(3):465–470. https://doi.org/10.1246/bcsj.20140337

McNaught AD, Wilkinson A (1997) IUPAC. Compendium of chemical terminology, 2nd edn. Blackwell Scientific Publications, Oxford

Nain AK (2016) Physicochemical study of solute-solute and solute-solvent interactions of glycine, L-alanine, L-valine and L-isoleucine in aqueous-D-mannose solutions at temperatures from 293.15 K to 318.15 K. J Chem Thermodyn 98:338–352. https://doi.org/10.1016/j.jct.2016.03.012

Pauling L (1932) The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. J Am Chem Soc 54(9):3570–3582. https://doi.org/10.1021/ja01348a011

Paulus A, Klockow-Beck A (1999) Structures and properties of carbohydrates. In: Paulus A, Klockow-Beck A (eds) Analysis of carbohydrates by capillary electrophoresis. Vieweg Teubner, Wiesbaden, pp 28–48

Pople JA, Head-Gordon M, Fox DJ (1989) Gaussian-1 theory: a general procedure for prediction of molecular energies. J Chem Phys 90(10):5622–5629. https://doi.org/10.1063/1.456415

Sârbu C, Briciu RD (2010) Lipophilicity of natural sweeteners estimated on various oils and fats impregnated thin-layer chromatography plates. J Liq Chromatogr Relat Technol 33(7–8):903–921. https://doi.org/10.1080/10826071003766021

SCCS (2010) SCCS/1347/10. Scientific committee on consumer safety, opinion on dihydroxyacetone, European commission, directorate-general for health & consumers. SCCS, Cambridge

Schmidt RR (1986) New methods for the synthesis of glycosides and oligosaccharides—are there alternatives to the Koenigs–Knorr method? Angew Chem, Int Ed Engl 25:212–235. https://doi.org/10.1002/anie.198602121

Singh V, Chhotaray PK, Gardas RL (2015) Effect of protic ionic liquid on the volumetric properties of ribose in aqueous

solutions. Thermochim Acta 610:69–77. https://doi.org/10.1016/j.tca.2015.04.023

Steele RP, Head-Gordon M (2010) Dual-basis self-consistent field methods: 6-31G* calculations with a minimal 6-4G primary basis. Mol Phys 105(19–22):2455–2473. https://doi.org/10.1080/00268970701519754

Teles ARR, Dinis TBV, Capela EV, Santos LMNBF, Pinho SP, Freire MG, Coutinho JAP (2016) Solubility and solvation of monosaccharides in ionic liquids. Phys Chem Chem Phys 18(29):19722–19730. https://doi.org/10.1039/C6CP03495K

van Putten RJ, Winkelman JGM, Keihan F, van der Waal JC, de Jong E, Heeres HJ (2014) Experimental and modeling studies on the solubility of d-Arabinose, D-Fructose, D-Glucose, D-Mannose, Sucrose and d-Xylose in methanol and methanol-water mixtures. Ind Eng Chem Res 53(19):8285–8290. https://doi.org/10.1021/ie500576q

Ye T, Qu H, Gong X (2017) Measurement and correlation of liquid-liquid equilibria for the ternary systems of Water + d -Fructose + 1-Butanol, Water + d -Glucose + 1-Butanol, and Water + d -Galactose + 1-Butanol at (288.2, 303.2 and 318.2) K. J Chem Eng Data 62(8):2392–2399. https://doi.org/10.1021/acs.jced.7b00302

Zafarani-Moattar MT, Shekaari H, Mazaher Haji Agha E (2017) Thermodynamic studies on the phase equilibria of ternary ionic liquid, 1-hexyl-3-methyl imidazolium chloride + D-fructose or sucrose + water systems at 298.15 K. Fluid Ph Equilibria 436:38–46. https://doi.org/10.1016/j.fluid.2016.12.024

Zhuo K, Zhang H, Wang Y, Liu Q, Wang J (2005) Activity coefficients and volumetric properties for the NaBr + maltose + water system at 298.15 K. J Chem Eng Data 50(5):1589–1595. https://doi.org/10.1021/je050064v

Zhuo K, Liu H, Zhang H, Liu Y, Wang J (2008) Activity coefficients and volumetric properties for the NaI + maltose + water system at 298.15 K. J Chem Eng Data 53(1):57–62. https://doi.org/10.1021/je700366w

Zwahlen C, Vincent SJF (2002) Determination of 1H homonuclear scalar couplings in unlabeled carbohydrates. J Am Chem Soc 124(24):7235–7239. https://doi.org/10.1021/ja017358v