Pencil and Paper Estimation of Hansen Solubility Parameters

Didier Mathieu[*]

CEA, DAM, Le Ripault, F-37260 Monts, France

Tel. (+33) 02 47 34 41 85

E-mail: didier.mathieu@cea.fr

S1 Systematic fragmentation

S1.1 Details of the fragmentation schemes

This section describes systematic algorithms to decompose a molecule into additive fragments in view of predicting HSP components. These procedures were used in this work to assess the value of conventional additivity schemes, by contrast to models based on tailored fragments specifically designed for each property considered and described in the main article. A rigorous comparison of the various approaches, based on exactly the same dataset, is reported below. The following hierarchical multi-levels scheme based on systematic fragment definitions was introduced:

1. At this most rudimentary level, the compound is described as the set of its atoms, characterized only by their atomic symbols. The present data set include nine different elements: C, H, N, O, F, Cl, Br, S and P.

2. At this level, the parameters associated with a given atom depend not only on its symbol X but also on its coordination number n. They are thus denoted Xn. For instance, carbon atoms are represented by three types of fragments denoted C4, C3 or C2 according to coordination numbers. This fragmentation level yields 15 different fragments for the present data set.

3. This level corresponds to the so-called geometrical fragment (GF) approach that proves highly successful in reliably predicting properties determined by intermolecular interaction, like sublimation enthalpies.[1] For the present data set, it yields 22 different fragments. According to this procedure, every non-hydrogen atom is assigned a label X$n-n_H$ where n_H is the number of hydrogen atoms attached. For instance, O2-0 and O2-1 stand for sp3 oxygen atoms in ether and hydroxyl groups, respectively. Obviously, this scheme does not involve any specific fragment associated with hydrogen atoms, as the latter are implicitly included into the definition of their neighbors. The rationale

behind this fragmentation procedure lies in the critical role of available atomic surface areas and the specificity of hydrogen resulting from its small size.[1]

4. This level is obtained through a straightforward extension of the previous one, with atom labels denoted $Xn-n_H n_L$ where n_L denotes the number of neighbors with atomic number $Z > 9$. For instance, C3-01 defines the fragment associated with the carbon atom attached to bromine in bromobenzene. This level yields a number of fragments (35) similar to the Fedors method.[2]

5. At this last level, the symbol of every neighbor is taken into account in the definition of atom labels. Bond orders are taken into account as this yields some improvement. Thus, every atom is labeled after its symbol followed by the symbols of its neighbors and associated bond orders. For instance, C(#N)(−C) denotes a carbon atom in a cyano group, and C(=O)(−C)(−O) a carbon such as found in esters. This level obviously leads to a very large number (62) of adjustable parameters.

6. For completeness, it should be mentioned that we also considered a further level corresponding to so-called group bonds.[3] As many as 120 different group bonds prove necessary for the present data set. For instance, ethane and methanol exhibit their own unique group bonds, denoted respectively C(−H)₃−C(−H)₃ and C(−H)₃−O(−H). Owing to the fact that the number of parameters required is not significantly smaller than the number of compounds in the data set, it proved impossible to fit any useful relationship on this basis. Therefore, group bonds are not further discussed in the sequel.

S1.2 HSP evaluation methods

The two approaches common encountered to estimate HSP from additive fragments were considered in this study:

- **Method 1.** Every HSP component δ_k is directly decomposed as a linear function of the number N_G of any group G according to the following equation:

$$\delta_k = \delta_k^0 + \sum_G N_G \delta_k(G) \tag{1}$$

- **Method 2.** Every HSP component δ_k is evaluated as $(E_k/V_m)^{1/2}$ where V_m is the molar volume and the HSP energy component as obtained as a sum of fragment contributions as follows:

$$E_k = \sum_G N_G E_k(G) \tag{2}$$

S1.3 Comparison of various approaches

This section summarizes the results obtained using the systematic fragmentation schemes described in section S1.1 combined with either method 1 or method 2 described in section S1.2. These results are compared not only to those obtained using the procedures based on tailored fragments described in the main article, but also to the previously published Stefanis and Panayiotou (SP) and Fedors procedures, that were specifically re-implemented for the present study.

The performances of the various methods are summarized on Figure S1. They are characterized on the basis of the value of the determination coefficient $R^2(\text{LOO})$ obtained as outcome of a leave-one-out cross-validation. These values are reported along the y-axis. The indexes $0-8$ on the x-axis refer respectively to the following fragmentation schemes:

1. Indexes 0-4 refer to the fragmentation schemes numbered 1-5 in section S1.1.

2. Index 5 refers to the Fedors method.[2]

3. Index 6 refers to the SP model restricted to first-order groups.[4]

4. Index 7 refers to the full SP model usinn both first- and second-order groups.
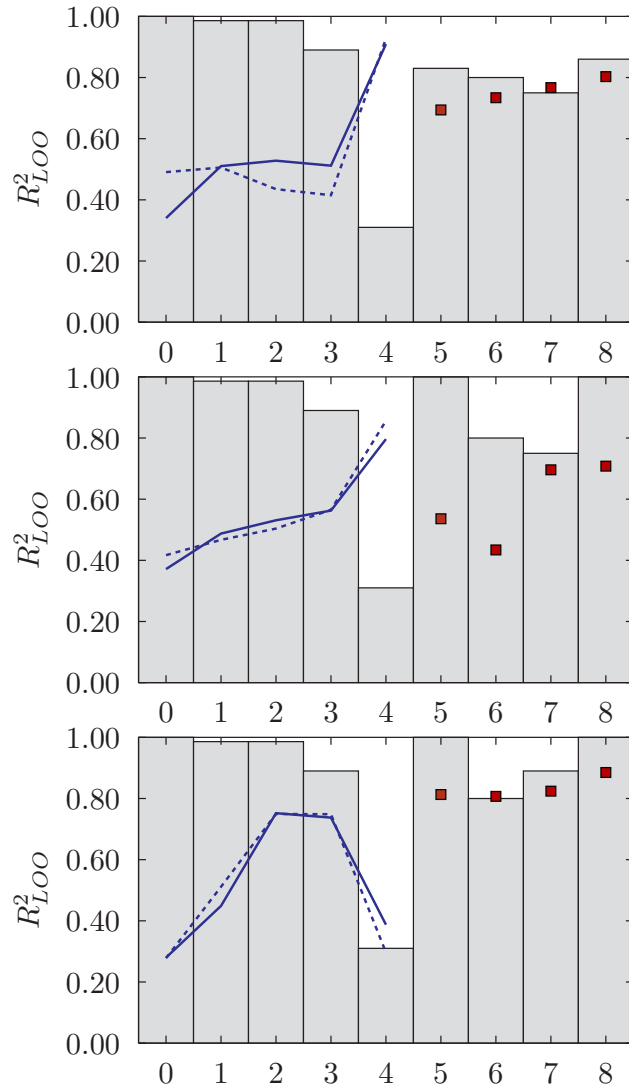
Figure S1: Performances of the models for δ_d (top), δ_p (middle) and δ_h (bottom). See text for details.

5. Finally, index 8 refers to the additivity schemes based on tailored fragments described in the main text of the article.

Regarding the kind of equation employed to derive HSP from additive contributions, both methods 1 and 2 described in S1.2 are used in combination with the systematic fragmentation schemes (indexes 0-4), method 1 is used for the SP models (indexes 6 and 7), and method 2 is used in other cases.

The score of each procedure is represented as a red square for models associated with indexes 5-8, and blue lines for the new procedures defined in this Supporting Information on the basis of systematic fragments (indexes 0-4). The straight line represents the scores obtained using method 1, while the dashed line is for the scores obtained using method 2. For every fragmentation scheme, the fraction of the data set for which all required parameters could be obtained is shown as a gray background. Although the systematic fragmentation schemes were introduced in the present work, not all required parameters could be obtained owing to linear dependencies. Thus, the scope of the fragmentation schemes range from a 100% coverage of the data set for index 0 to $\simeq 30\%$ of index 4.

First of all, it may be observed that method 1 and method 2 yield similar performances, in line with the fact that Stefanis et al. found no advantage in using the former equation.[4] The only exception is for δ_d, for which method 2 sometimes performs worse, in spite of the fact that it is more physically consistent. In fact, for this HSP component, systematic decomposition schemes do not provide any method clearly superior to the straigthforward decomposition of E_d as a sum of environment-independant atomic contributions (index 0).

For systematic models, $R^2(\text{LOO})$ usually increases with the corresponding level of fragmentation represented by the corresponding index ranging from 0 to 4. This is clearly to be expected. However, this is not true for δ_h for which a dramatic loss of accuracy is observed on going to the highest level. This can be easily explained as this component depends essentially on groups bearing either labile protons or proton acceptors. Therefore, increasing the level of fragmentation in the context of a systematic approach introduces a growing number

of irrelevant parameters.

In any case, the cost for increasing the level of fragmentation is a loss of generality which becomes unacceptable for the highest level (index 0) as the model is then applicable to only about 30% of the compounds in the data set. Consequently, it is clear from Figure S1 that such systematic fragmentation schemes cannot compete with conventional group contribution (GC) methods or procedures based on tailored fragments.

For δ_p, a small but steady improvement is observed on going from the Fedors to the SP method to the present model. As pointed out in the main article, δ_p proves especially ill-suited for additivity methods as the Fedors method and the SP method perform poorly. Only a very detailed level of fragmentation (full SP method) or the use of physically-motivated fragments (present approach) yield acceptable results. In contrast, δ_h proves relatively easy to estimate on the basis of an additivity approach. The fair results obtained using the Fedors method is consistent with present findings that only a small set of additive contribution is sufficient to obtain fairly good predictions.

S2   Assessment of the Hansen-Beerbower equation

It is well-known that δ_p may be approximated on the basis of the Hansen-Beerbower (HB) equation, ie. by assuming that it is proportional to $\mu/V^{1/2}$ where μ is the dipole moment of the molecule and V the corresponding molar volume:[5]

$$\delta_p = k_{HB} * \mu/V^{1/2} \tag{3}$$

To check this relation, initial 3D geometries are built for the present database with the help of the BALLOON program.[6] They are subsequently optimized at the B3LYP/6-31G** level of theory, using the Gaussian92 software package,[7] and μ is taken as the resulting dipole moment. An average ratio of 39 is obtained between δ_p in MPa$^{1/2}$ and $\mu/V^{1/2}$ in Debye.(cc/mol)$^{-1/2}$. The corresponding determination coefficient is $R^2 = 0.635$. This rel-

atively low value is to be expected owing to the occurence of flexible molecules for which different values of μ may be calculated depending on the conformation used.

Nonetheless, this procedure proves more successful that the Fedors method, the SP method without second order contributions and all systematic models presently developed. In fact, the most successful method for δ_p appears to be the full SP scheme, with $R^2_{LOO} = 0.696$. However, it is applicable only to 75% of the present data set. A more reliable and general method is thus clearly desirable.

S3   Worked-out examples

This section illustrates the application of the models eventually put forward in the main article to two molecules: 2-chloroacetamide and paracetamol. For brevity, units are implicit. They are cm^3/mol for molar volume and molar refractivity, J/mol for cohesive energies and $MPa^{1/2}$ for HSP components. The dispersion component of the HDPs is obtained from the relationship of δ_d^2 to R_D and V_m:

$$\delta_d^2 = 93.8 + 2016 \left(\frac{R_D}{V_m}\right)^2 + \frac{75044}{V_m}\left(\frac{R_D}{V_m}\right)^2 \tag{4}$$

2-chloroacetamide (CASRN: 79-07-2)

The molecule is made of the five geometrical fragments: $>C=$, $>CH_2$, $-Cl$, $=O$ and $-NH_2$. Their contributions to the molar volume V_m and molar refraction R_D are obtained respectively from ref 8 and ref 9. As a result, their values are estimated as follows:

$$V_m = 0.00 + 16.57 + 24.74 + 14.89 + 17.81 = 74.0 \tag{5}$$

$$R_D = 3.15 + 4.60 + 5.87 + 1.84 + 4.60 = 20.1 \tag{6}$$

The dispersion component of HSPs is then obtained from eq 4 where $V_m = 74.0$ and $R_D = 20.1$. The value thus obtained is $\delta_d = 17.8$. The two other HSP components δ_p and δ_h are obtained from the sum of additive contributions to the corresponding cohesive energies E_p and E_h:

$$E_p = \text{Cl(H0)} + \text{N(H2)} + \text{O(H0)} + \text{C=O(amide)} \tag{7}$$

Therefore, using the parameters provided in Table 2:

$$E_p = 1637 + 8235 + 1603 + 15972 = 27447 \tag{8}$$

Similarly:

$$E_h = 2 \times \text{HC} + 2 \times \text{HN(amide)} + \text{N} + \text{O} + \text{X} \tag{9}$$

Using the parameters in Table 3:

$$E_h = 2 \times 24.5 + 2 \times 5060 + 3252 + 1980 + 412 = 15813 \tag{10}$$

From the cohesive energies E_p and E_h and using $V_m = 74.0$, we obtain $\delta_p = 19.3$ and $\delta_h = 14.6$.

Paracetamol (CASRN: 103-90-2)

The molecule is made of the following geometrical fragments (where lower case symbols denote aromatic atoms): >cH (four fragments), >c= (two fragments), >C=, $-\text{CH}_3$, >NH, $-\text{OH}$, and =O. The molar volume and refractivity are therefore obtained as follows:

$$V_m = 4 \times 13.23 + 2 \times 0.00 + 0.00 + 29.58 + 7.74 + 11.78 + 14.89 + 6.89 + 1.82 = 125.6 \tag{11}$$

where the last two terms 6.89 and 1.82 in the sum are volume corrections associated respectivement with any 6-membered ring and with aromaticity (see ref 8 for details).

$$R_D = 4 \times 4.46 + 2 \times 3.48 + 3.15 + 5.74 + 3.69 + 2.51 + 1.84 = 41.73 \tag{12}$$

Using eq 4 with $V_m = 125.6$ and $R_D = 41.73$, a value $\delta_d = 19.5$ is obtained.

$$E_p = N(H1) + O(H0) + C{=}O(\text{amide}) + O(H1) \tag{13}$$

Using the parameters provided in Table 2:

$$E_p = 2783 + 1603 + 15972 + 4125 = 24483 \tag{14}$$

Similarly:
$$E_h = 7 \times HC + HN(\text{amide}) + HO + O + N \tag{15}$$

In other words, using parameters provided in Table 3:

$$E_h = 7 \times 24.5 + 5060 + 16945 + 2 \times 1980 + 3252 = 29388.5 \tag{16}$$

These values of E_p and E_h correspond to $\delta_p = 14.0$ and $\delta_h = 15.3$.

References

(1) Mathieu, D. Simple alternative to neural networks for predicting sublimation enthalpies from fragment contributions. *Ind. Eng. Chem. Res.* **2012**, *51*, 2814–2819.

(2) Fedors, R. F. A method for estimating both the solubility parameters and molar volumes of liquids. *Polym. Eng. Sci.* **1974**, *14*, 147–154.

(3) Pan, Y.; Jiang, J.; Wang, W. Quantitative structure-property relationship studies for

predicting flash points of alkanes using group bond contribution method with back-propagation neural network. *J. Haz. Mater.* **2007**, *147*, 424–430.

(4) Stefanis, E.; Panayiotou, C. Prediction of Hansen Solubility Parameters with a New Group-Contribution Method. *Int. J. Thermophys.* **2008**, *29*, 568–585.

(5) Hansen, C. M.; Beerbower, A. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd ed.; Standen, A., Ed.; Interscience: New-York, 1971; pp 889–910, Suppl. Vol.

(6) Vaino, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

(7) Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

(8) Mathieu, D.; Bouteloup, R. Reliable and Versatile Model for the Density of Liquids Based on Additive Volume Increments. *Ind. Eng. Chem. Res.* **2016**, *55*, 12970–12980.

(9) Bouteloup, R.; Mathieu, D. Improved model for the refractive index: application to potential components of ambient aerosol. *Phys. Chem. Chem. Phys.* **2018**, *20*, 22017–22026.