Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
International Journal of Pharmaceutics 325 (2006) 8–14
INTERNATIONAL JOURNAL OF PHARMACEUTICS
www.elsevier.com/locate/ijpharm

Prediction of solubility parameters using partial least square regression

Vimon Tantishaiyakul [a,*], Nimit Worakul [b], Wibul Wongpoowarak [b]

[a] Department of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Prince of Songkla University, Hat-Yai, Songkhla 90112, Thailand
[b] Department of Pharmaceutical Technology, Faculty of Pharmaceutical Sciences, Prince of Songkla University, Hat-Yai, Songkhla 90112, Thailand

Received 21 October 2005; received in revised form 23 May 2006; accepted 6 June 2006
Available online 9 June 2006

Abstract

The total solubility parameter (δ) values were effectively predicted by using computed molecular descriptors and multivariate partial least squares (PLS) statistics. The molecular descriptors in the derived models included heat of formation, dipole moment, molar refractivity, solvent-accessible surface area (SA), surface-bounded molecular volume (SV), unsaturated index (Ui), and hydrophilic index (Hy). The values of these descriptors were computed by the use of HyperChem 7.5, QSPR Properties module in HyperChem 7.5, and Dragon Web version. The other two descriptors, hydrogen bonding donor (HD), and hydrogen bond-forming ability (HB) were also included in the models. The final reduced model of the whole data set had R^2 of 0.853, Q^2 of 0.813, root mean squared error from the cross-validation of the training set (RMSEcv$^{\text{tr}}$) of 2.096 and RMSE of calibration (RMSE$^{\text{tr}}$) of 1.857. No outlier was observed from this data set of 51 diverse compounds. Additionally, the predictive power of the developed model was comparable to the well recognized systems of Hansen, van Krevelen and Hoftyzer, and Hoy.

Keywords: Solubility parameter; Partial least square; Molecular descriptors

1. Introduction

The solubility parameter (δ) is an intrinsic physicochemical property of a substance. It provides an easy numerical method of fast prediction the basic properties of materials (Bustamante et al., 1998), interaction between materials including drug-excipient, and drug-plasma protein (Forster et al., 2001), drug absorption (Roy and Flynn, 1998; Martini et al., 1999; Yen et al., 2005), formulating blends of solvents (Subrahmanyam and Suresh, 1999), solvent selection for organic reaction (Gani et al., 2005), and dosage form technology and design (Hancock et al., 1997; Minghetti et al., 1999; Wagner et al., 2005).

One of the essential functions of the solubility parameter is for evaluating the possibility of mixing between substances. Substances with similar values for δ are possibly miscible due to the balance of energy of mixing released by interactions within the substances and the energy released by interaction between the substances. Greenhalgh et al. (1999) categorized excipients based on the difference between the solubility parameters of excipients and drugs ($\Delta\delta$). It was concluded that substances with a $\Delta\delta < 7.0\,\text{MPa}^{1/2}$ were likely to be miscible, whereas those with $\Delta\delta > 10\,\text{MPa}^{1/2}$ were likely to be immiscible.

The solubility parameter concept was first proposed by Hildebrand as the square root of the cohesive energy density:

$$\delta = \sqrt{\frac{\Delta E_{\text{vap}}}{V}} \tag{1}$$

where ΔE_{vap} and V are the energy or heat of vaporization and molar volume of the liquid, respectively. This one-dimensional solubility parameter was applied primarily to a nonpolar liquid. The solubility parameter concept has since been extended to various systems such as polar, polymer–solvent, and polymer–polymer systems. Hansen (1967) extended the original Hildebrand parameter to three-dimensional solubility parameter for the polar and hydrogen bonding systems. According to this concept, the total solubility parameter (δ) is separated into three different types of partial solubility parameters relating to the specific intermolecular interactions:

$$\delta^2 = \delta_{\text{d}}^2 + \delta_{\text{p}}^2 + \delta_{\text{h}}^2 \tag{2}$$

where δ_{d}, δ_{p}, and δ_{h} are the dispersion, polar, and hydrogen bond partial solubility parmeters, respectively. For liquids, the δ_{d} value may be obtained by homomorph methods (Barton, 1975). The

[*] Corresponding author. Tel.: +66 7428 8864; fax: +66 7442 8239.
E-mail address: vimon.t@psu.ac.th (V. Tantishaiyakul).

0378-5173/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2006.06.009

δ_h $(\mathrm{cal/cm^3})^{1/2}$ was calculated directly from $\sqrt{5000N/V}$, where N is the number of alcohol groups in the molecule, V is the molar volume, and the number 5000 originates from the approximate value for the H···O bond energy of 5000 cal/mol (Hansen and Skaarup, 1967). The δ_p component was related to the cohesion energy of a fluid in terms of relative permittivity, refractive index, and dipole moment (Hansen and Skaarup, 1967). Hansen's total solubility parameter corresponds to the Hildebrand parameter, nevertheless these two quantities are probably different when they are obtained by different methods.

These individual solubility parameters can also be predicted from various methods including the group contribution calculations. According to van Krevelen (1990), each parameter can be estimated using these following equations:

$$\delta_d = \frac{\sum F_{di}}{\sum V_i} \tag{3}$$

$$\delta_p = \frac{\sqrt{\sum F_{pi}^2}}{\sum V_i} \tag{4}$$

$$\delta_h = \sqrt{\frac{\sum E_{hi}}{\sum V_i}} \tag{5}$$

where F_d is the dispersion component of giving δ_d, F_p the polar component, E_h the contribution of hydrogen bond forces to the cohesive energy, and i is a contributing group. The total solubility parameter is then estimated using Eq. (6):

$$\delta = \sqrt{\delta_d^2 + \delta_p^2 + \delta_h^2} \tag{6}$$

Another group contribution system used to estimate total and partial solubility parameters was proposed by Hoy (1970). In accordance with Hoy's system, the values of δ, δ_p and δ_h are initially evaluated, δ_d can be determined by the difference using Eq. (7):

$$\delta_d = \sqrt{\delta^2 - \delta_p^2 - \delta_h^2} \tag{7}$$

There are many methods for estimating δ experimentally, including inverse gas chromatographic (Choi et al., 1996; Zhao and Choi, 2001; Price and Shillcock, 2002), dissolution calorimetric measurements (Rey-Mermet et al., 1991), and solubility method (Martin et al., 1980; Rey-Mermet et al., 1991). In addition, computational methods have also been applied to the estimation of the solubility parameters of solvents, hydroxyethyl-, and hydroxypropyl cellulose, and alkyl phenol ethoxylates (Choi et al., 1992; Kavassalis et al., 1993; Choi et al., 1994; Suga and Takahama, 1996). The accuracy of the calculations, however, relies on the correct application of molecular force field parameters and the building of the bulk structure.

In this study, quantitative structure-property relationship (QSPR) was developed for predicting solubility parameter values. Molecular descriptors used were calculated directly from molecular structures, and partial least square (PLS) statistics was employed for building the predictive models.

2. Methods

2.1. Data set and δ values

The experimental results of δ values for 51 compounds were taken from literature (Hancock et al., 1997; Barra et al., 2000; Goharshadi and Hesabi, 2004). These δ values obtained from different sources were averaged and listed in Table 1. The well established Hansen solubility parameters were calculated using molecular modeling pro plus (ChemSW). Three different calculations were performed with the methods outlined by Hansen (proprietary algorithm of ChemSW), van Krevelen and Hoftyzer (van Krevelen, 1990), and Hoy (1970). These calculated values are presented in Table 1.

2.2. Molecular modeling and molecular descriptors

Molecular modeling calculations were performed using HyperChem 7.5 for Windows (Hypercube, FL, USA). Geometry optimization was performed initially by AMBER force field method of molecular mechanics and subsequently using the AM1 semi-empirical quantum chemistry. The obtained geometries were then optimized on the basis of the *ab initio* quantum mechanics method for single point calculation at the 3–21 G level. The advantage of the semi-empirical method over the *ab initio* method is that it is faster which is substantial for biomolecules; however, this may not be important for small molecules.

Molecular descriptors include binding energy and heat of formation were obtained from semi-empirical method calculation; van der Waals force (Vdw) was determined from molecular mechanics calculation; dipole moment and total energy were obtained from *ab initio* computation. The QSPR Properties module in HyperChem 7.5 was employed for further calculation of other descriptors such as solvent-accessible surface area (SA), surface-bounded molecular volume (SV), molar refractivity, polarizability, and molecular mass. The simplest one-dimensional (1D) descriptors include three empirical descriptors, unsaturated index (Ui), hydrophilic index (Hy), and aromatic ratio (ARR), as well as three properties, Ghose-crippen molar refractivity (MR), fragment-based polar surface area (PSA), and Moriguchi octanol-water partition coefficient (M log P), were calculated using Dragon Web version (Milano Chemometrics and QSAR group, http://www.disat.unimib.it/vhm). The hydrogen bonding acceptor (HA), hydrogen bonding donor (HD), and hydrogen bond-forming ability (HB), sum of HA and HD, were calculated as described by Xia et al. (1998). These calculated descriptors are listed in Table 2.

2.3. Statistical analysis

The relationship between the experimental δ values and descriptors was determined using PLS regression analysis. PLS is a bilinear modeling method where information in the descriptor matrix X is projected onto a small number of underlying ("latent") variables called PLS components, referred to as PCs.

Table 1
Experimental, calculated and predicted solubility parameter values

	Solubility parameter (MPa$^{1/2}$)				Model 1	Model 2	
	Exp[a]	Hansen[b]	van Krevelen and Hoftyzer[b]	Hoy[b]		Train	Test
2-Hexanone	18.14	18.51	16.92	20.82	18.96		19.07
2-Pentanone	18.28	18.68	17.17	21.42	19.43	19.51	
Acetic acid	21.40	23.54	24.43	24.04	22.77		22.58
Acetone	20.05	19.98	18.29	23.54	20.97		20.97
Acetonitrile	24.10	24.42	25.06	29.24	23.14	23.01	
Barbital	27.60	26.06	23.81	29.26	26.53		26.14
Benzene	19.15	18.47	17.26	21.40	18.70		18.88
Benzocaine	31.70	22.77	22.92	22.69	28.46	27.93	
Benzoic acid	23.59	21.86	24.78	22.34	22.49	22.38	
Butylparaben	21.60	22.75	23.86	21.66	23.35	23.23	
Caffeine	26.70	24.84	30.12	[c]	28.00	27.83	
Carbamazepine	32.20	25.52	26.09	22.85	26.40		25.94
Carbon tetrachloride	18.10	17.80	[c]	[c]	18.22	18.34	
Cephalexin	33.45	26.35	28.09	24.77	35.00	33.93	
Chloroform	19.00	19.00	21.59	22.19	19.62	19.65	
Cyclohexane	17.20	16.63	15.85	19.75	16.06		16.28
Dibutyl phthalate	19.60	20.20	20.08	20.39	19.22	19.48	
Diclofenac	24.68	22.14	27.11	21.69	25.81		25.42
Diethyl phthalate	20.50	20.98	21.31	21.40	18.44	18.77	
Dimethyl phthalate	22.00	22.04	22.31	22.20	20.73	20.92	
Dimethylsulfoxide	26.00	26.67	[c]	[c]	23.15	23.05	
Dioctyl phthalate	18.20	19.21	18.69	19.06	16.91	17.30	
Ethanol	26.05	26.49	25.10	30.77	24.90		24.56
Ethyl acetate	18.70	18.13	18.23	21.77	19.24		19.38
Ethyl propionate	17.67	17.14	17.92	21.06	18.57	18.75	
Ethylene glycol	29.60	32.90	31.86	38.66	27.84	27.29	
Ibuprofen	19.46	19.25	20.56	19.28	20.32	20.32	
Isopropanol	23.00	23.51	22.97	25.36	23.39		23.12
Isopropyl acetate	17.12	17.01	17.65	20.46	18.65	18.82	
Methanol	29.50	29.61	28.31	31.78	27.83		27.35
Methylene chloride	19.50	20.18	20.83	22.93	20.50		20.48
n-Butyl acetate	17.58	17.41	17.69	20.55	18.22	18.41	
Neopentane	12.70	15.00	14.49	18.61	16.31	16.51	
n-Hexane	15.50	15.00	14.49	18.61	15.62	15.86	
N-Methylpyrrolidone	23.10	22.97	22.11	24.35	21.04	21.05	
n-Octane	15.80	15.48	14.89	18.31	14.89	15.17	
n-Octanol	19.55	21.04	19.76	21.97	19.90	19.81	
n-Pentane	14.80	14.50	14.21	18.84	16.13		16.34
n-Propanol	24.30	24.53	23.26	27.55	23.27	23.02	
n-Propyl acetate	18.00	17.49	17.92	21.06	18.70	18.87	
Oleic acid	15.95	17.54	17.74	18.54	17.56		17.66
Palmitic acid	16.10	17.63	18.00	18.51	17.22	17.32	
p-Aminobenzoic acid	26.67	24.89	28.64	23.94	31.60	30.76	
Phenylbutazone	25.10	20.67	22.70	22.10	20.03	20.25	
Propylene glycol	28.05	30.20	28.78	32.02	28.57	27.94	
Salicylic acid	23.72	25.50	26.71	23.66	23.38		23.25
t-Butanol	21.00	21.67	22.00	21.73	22.39	22.17	
Testosterone proprionate	19.40	17.69	19.46	19.09	19.59	19.72	
Tetrahydrofuran	19.13	19.41	17.76	23.76	19.13	19.20	
Theophylline	27.83	28.11	32.16	[c]	30.36	29.92	
Toluene	17.80	18.13	17.64	20.68	18.39		18.59
RMSE[d]		2.25	2.38[e]	3.80[f]	1.86	1.86	1.88

[a] Experimental values.
[b] Calculated using ChemSW.
[c] Incalculable due to missing fragment values for this method.
[d] Root mean square error.
[e] RMSE calculated based on 49 compounds.
[f] RMSE calculated based on 47 compounds.

Table 2
Molecular descriptors of compounds

	Vdw	BE	HF	*D*	*E*	SA	SV	RF	*P*	MW	HD	HA	HB	Ui	Hy	ARR	MR	PSA	M log P
2-Hexanone	1.31	−1779.26	−69.14	3.06	−192856.3	296.02	435.88	30.02	11.87	100.16	0	2	2	1.00	−0.80	0.00	30.54	17.07	1.44
2-Pentanone	1.07	−1497.32	−62.29	3.04	−168498.1	268.57	381.94	25.42	10.04	86.13	0	2	2	1.00	−0.77	0.00	25.94	17.07	1.06
Acetic acid	0.12	−772.38	−103.07	1.82	−142148.5	193.31	243.77	12.64	5.17	60.05	1	4	5	1.00	−0.43	0.00	12.64	17.07	−0.39
Acetone	0.29	−934.15	−49.31	3.23	−119780.0	209.79	275.29	16.19	6.37	58.08	0	2	2	1.00	−0.65	0.00	16.81	17.07	0.20
Acetonitrile	−0.15	−591.88	19.21	3.89	−82321.4	177.06	219.74	11.93	4.46	41.05	0	1	1	1.00	−0.53	0.00	11.93	23.79	−0.32
Barbital	2.40	−2520.80	−123.78	1.35	−401537.9	347.78	546.15	44.25	17.64	184.19	2	8	10	2.00	0.69	0.00	44.25	51.21	−0.16
Benzene	3.10	−1316.08	21.87	0.00	−143960.5	239.98	331.81	26.06	10.43	78.11	0	0	0	2.81	−0.92	1.00	26.06	0.00	2.26
Benzocaine	5.62	−2401.17	−57.92	3.88	−344090.8	363.93	556.34	47.03	18.01	165.19	2	5	7	3.00	0.60	0.50	47.03	26.30	1.78
Benzoic acid	4.67	−1696.09	−68.13	2.42	−261032.9	279.53	408.60	32.82	12.99	122.12	1	4	5	3.00	−0.74	0.67	32.82	17.07	1.70
Butylparaben	5.84	−2913.65	−125.75	3.48	−405177.8	422.15	654.82	53.15	20.97	194.23	1	6	7	3.00	−0.23	0.43	53.15	26.30	2.38
Caffeine	4.87	−2537.70	25.76	5.11	−422599.9	373.25	587.56	49.91	21.25	196.21	0	8	8	3.46	0.70	0.67	42.88	51.56	−0.29
Carbamazepine	5.89	−3424.70	49.44	3.92	−473483.9	421.90	698.13	71.89	27.42	236.27	2	4	6	3.91	−0.82	0.60	67.74	20.31	2.73
Carbon tetrachloride	0.00	−315.05	−28.20	0.00	−1171410.5	244.77	337.94	26.85	10.32	153.82	0	0	0	0.00	−0.18	0.00	2.30	0.00	2.23
Cephalexin	2.83	−4334.01	−70.40	5.51	−921230.6	574.32	944.39	88.98	34.99	347.39	4	13	17	3.46	0.38	0.23	87.01	79.75	0.21
Chloroform	0.00	−339.04	−29.07	1.60	−884833.6	225.08	296.75	21.36	8.39	119.38	0	0	0	0.00	−0.22	0.00	21.37	0.00	1.82
Cyclohexane	1.16	−1689.35	−38.78	0.00	−146150.4	263.25	379.19	27.61	11.01	84.16	0	0	0	0.00	−0.92	0.00	27.61	0.00	3.52
Dibutyl phthalate	11.55	−4295.36	−176.64	2.95	−572953.7	579.99	930.47	76.86	30.23	278.35	0	8	8	3.17	−0.79	0.30	76.86	52.60	3.62
Diclofenac	3.57	−3309.48	−53.80	3.06	−1036134.5	463.54	770.28	75.46	29.69	296.15	2	5	7	3.81	−0.24	0.60	76.95	17.07	3.99
Diethyl phthalate	15.53	−3158.41	−140.07	0.22	−475506.0	414.04	676.26	58.61	22.89	222.24	0	8	8	3.17	−0.73	0.38	58.61	52.60	2.58
Dimethyl phthalate	14.46	−2605.85	−137.69	3.02	−426797.2	392.83	599.48	49.11	19.22	194.19	0	8	8	3.17	−0.70	0.43	49.11	52.60	2.01
Dimethylsulfoxide	0.02	−819.85	−39.50	4.71	−344277.4	221.17	291.96	20.56		78.13	0	4	4	1.00	−0.43	0.00	20.56	36.28	−0.32
Dioctyl phthalate	12.71	−6537.50	−218.03	1.27	−767808.6	736.69	1280.70	113.41	44.91	390.56	0	8	8	3.17	−0.85	0.21	113.41	52.60	5.43
Ethanol	0.15	−776.72	−62.77	1.94	−96145.3	192.92	242.08	13.01	5.08	46.07	1	2	3	0.00	0.71	0.00	13.01	0.00	−0.17
Ethyl acetate	0.78	−1321.87	−102.38	1.90	−190864.2	260.38	361.52	22.16	8.84	88.11	0	4	4	1.00	−0.61	0.00	22.16	26.30	0.59
Ethyl propionate	1.33	−1603.02	−108.43	1.78	−215224.2	292.88	417.41	26.79	10.67	102.13	0	4	4	1.00	−0.67	0.00	26.79	26.30	1.00
Ethylene glycol	0.03	−881.12	−107.61	0.00	−142849.9	208.29	266.30	14.55	5.72	62.07	2	4	6	0.00	1.84	0.00	14.55	0.00	−1.05
Ibuprofen	4.09	−3381.77	−103.24	1.94	−407179.9	438.16	705.08	60.73	24.00	206.28	1	4	5	3.00	−0.85	0.40	60.73	17.07	3.23
Isopropanol	0.62	−1057.23	−68.19	1.92	−120508.0	221.41	293.23	17.43	6.92	60.10	1	2	3	0.00	0.37	0.00	17.43	0.00	0.35
Isopropyl acetate	1.21	−1601.05	−106.46	1.78	−215225.9	288.54	410.27	26.58	10.67	102.13	0	4	4	1.00	−0.67	0.00	26.58	26.30	1.00
Methanol	0.00	−495.95	−57.10	2.10	−71783.2	156.84	184.06	8.26	3.25	32.04	1	2	3	0.00	1.40	0.00	8.26	0.00	−0.81
Methylene chloride	0.00	−359.00	−25.92	2.19	−598253.6	199.60	252.79	16.44	6.47	84.93	0	0	0	0.00	−0.26	0.00	16.44	0.00	1.36
n-Butyl acetate	1.18	−1885.70	−116.02	1.93	−239580.7	320.61	469.90	31.29	12.51	116.16	0	4	4	1.00	−0.71	0.00	31.29	26.30	1.37
Neopentane	2.38	−1512.67	−33.00	0.00	−122520.2	254.59	363.61	24.63	9.95	72.15	0	0	0	0.00	−0.90	0.00	24.63	0.00	3.14
n-Hexane	1.28	−1799.85	−45.09	0.00	−146873.4	300.66	433.22	29.41	11.78	86.18	0	0	0	0.00	−0.92	0.00	29.41	0.00	3.52
N-Methylpyrrolidone	0.15	−1536.39	−40.47	4.05	−202117.6	266.95	382.25	27.15	10.61	99.13	0	3	3	1.00	−0.67	0.00	27.15	20.31	0.20
n-Octane	1.78	−2363.82	−58.86	0.00	−195589.5	362.65	540.69	38.61	15.45	114.23	0	0	0	0.00	−0.95	0.00	38.61	0.00	4.20
n-Octanol	1.65	−2468.57	−104.06	1.87	−242294.1	378.34	565.26	40.54	16.09	130.23	1	2	3	0.00	−0.20	0.00	40.54	0.00	2.27
n-Pentane	1.02	−1517.87	−38.20	0.05	−122515.4	270.55	379.59	24.81	9.95	72.15	0	0	0	0.00	−0.90	0.00	24.81	0.00	3.14
n-Propanol	0.38	−1058.73	−69.68	1.85	−120503.8	224.65	296.48	17.53	6.92	60.10	1	2	3	0.00	0.37	0.00	17.53	0.00	0.35
n-Propyl acetate	0.94	−1603.77	−109.18	1.96	−215222.7	291.24	416.27	26.69	10.67	102.13	0	4	4	1.00	−0.67	0.00	26.68	26.30	1.00
Oleic acid	3.00	−5150.39	−183.78	1.73	−531135.3	683.09	1095.46	87.40	34.33	282.47	1	4	5	1.59	−0.89	0.00	87.40	17.07	4.67
Palmitic acid	3.36	−4718.84	−198.22	1.73	−483162.8	620.30	992.30	77.08	30.86	256.43	1	4	5	1.00	−0.87	0.00	77.08	17.07	4.29
p-Aminobenzoic acid	4.65	−1864.02	−70.95	4.60	−295376.2	298.33	443.69	37.52	14.34	137.14	3	5	8	3.00	0.76	0.60	37.52	17.07	1.13
Phenylbutazone	4.42	−4611.51	22.52	0.69	−617609.0	563.02	943.42	88.76	35.04	308.38	0	6	6	3.91	−0.81	0.48	85.42	40.62	2.89
Propylene glycol	0.56	−1164.80	−116.19	3.12	−167215.9	232.31	314.87	18.97	7.55	76.10	2	4	6	0.00	1.46	0.00	18.97	0.00	−0.53
Salicylic acid	6.38	−1957.93	−54.88	1.76	−285370.3	299.70	450.03	38.15	14.83	136.15	1	4	5	3.00	−0.15	0.60	38.46	17.07	1.49
t-Butanol	1.09	−1335.93	−71.79	1.83	−144871.1	244.22	339.20	22.07	8.75	74.12	1	2	3	0.00	0.17	0.00	22.07	0.00	0.80
Testosterone proprionate	10.71	−5761.14	−155.62	5.26	−671643.6	570.19	1008.95	98.21	38.67	344.49	0	6	6	2.00	−0.87	0.00	93.71	43.37	4.64
Tetrahydrofuran	0.31	−1218.50	−58.57	2.35	−144128.7	230.61	311.60	20.55	7.98	72.11	0	2	2	0.00	−0.72	0.00	20.55	9.23	0.41
Theophylline	4.01	−2204.36	31.90	5.01	−397946.9	339.50	529.06	44.16	18.97	181.17	1	8	9	3.46	0.77	0.71	27.51	51.56	−0.63
Toluene	3.35	−1598.87	14.18	0.31	−168319.2	267.46	384.69	31.10	12.27	92.14	0	0	0	2.81	−0.94	0.86	31.10	0.00	2.61

Vdw, van der Waals force; BE, binding energy (kcal/mol); HF, heat of formation (kcal/mol); *D*, dipole moment (Debyes); *E*, total energy (kcal/mol); SA, solvent-accessible surface area; SV, surface-bounded molecular volume; RF, molar refractivity; *P*, polarizability; MW, molecular mass; HD, hydrogen bonding donor; HA, hydrogen boding acceptor; HB, hydrogen bond-forming ability; Ui, unsaturated index; Hy, hydrophilic index; ARR, aromatic ratio; MR, Ghose-crippen molar refractivity; PSA, fragment-based polar surface area; M log P, Moriguchi octanol-water partition coefficient.

The matrix Y is simultaneously used in estimating the “latent” variables in X that might be most relevant for predicting the Y variables.

The number of significant PCs for the PLS algorithm is determined using the cross-validation method. With cross-validation, some samples are kept out of the calibration and used for prediction. The process is repeated so that all samples are kept out once. The value for the left out compound is then predicted and compared with the known value. The prediction error sum of squares (PRESS) obtained in the cross-validation is calculated each time that a new PC is added to the model. The optimum number of PCs is concluded as the first local minimum in the PRESS versus PC plot. PRESS is defined as

$$\text{PRESS} = \sum_{i=1}^{n} (\hat{y} - y)^2 \tag{8}$$

where $\hat{y}$ is the estimated value of the ith object and y the corresponding reference value of this object. The goodness of fit is evaluated by root mean squared error (RMSE), which is defined as

$$\text{RMSE} = \sqrt{\frac{\text{PRESS}}{n}} \tag{9}$$

where n is the number of compounds.

The data analysis and multivariate calibrations were carried out using Unscrambler 6.1 (Computer-Aided Modelling A/S, Trondheim, Norway). All descriptor variables were preprocessed by autoscaling, using weights based on the variables’ standard deviation and the data were mean-centered.

A descriptors selection was performed in order to limit the amount of potentially irrelevant or redundant information. The selection was determined according to the magnitude of the absolute values of regression coefficients and the variable importance on the projection (VIP) obtained by the PLS regression (Chong and Jun, 2005). The insignificant descriptors were left out of the model and their importance for predictivity was determined by a cross-validation procedure. If the predictivity of the model increased, the descriptors in question were removed from the model otherwise the descriptors were kept in the model.

To establish the predictive power of a model, one needs to divide the available data set into the training and test sets. In general, a training set should contain 60–80% of the full data. For assigning compounds to training and test sets, compounds were ordered by δ values, and every third was selected for the test set, the remaining compounds were used as a training set. The test and training sets comprised 17 and 34 compounds, respectively, and indicated in Table 1 (model 2).

3. Results and discussion

The data set used consists of 51 compounds with experimental δ values ranging from 12.70 to 33.45 MPa$^{1/2}$ (Table 1). All descriptors generated from their molecular structures are simple to calculate with suitable software. These molecular descriptors were selected according to the relationship between compounds’ physicochemical properties/interactions and solubility parameters (Hildebrand, 1936; Hansen and Skaarup, 1967) and are listed in Table 2. Since δ values are correlated with cohesive energy density, the reciprocals of SA and SV were used in this PLS analysis accordingly.

PLS analysis was initially applied to the complete data set of 51 compounds and 19 descriptors. The preliminary analysis yielded a model containing three PCs with a squared correlation coefficient (R^2) of 0.865 and a squared correlation coefficient for cross-validation (Q^2) of 0.799. According to the comparison between the R^2 and Q^2, as well as the predicted RMSE value (RMSEcv$^{\text{tr}}$ of 2.180) and the calibration RMSE value (RMSE$^{\text{tr}}$ of 1.776), it may specify that the model is slightly overfitted. In accordance with the PLS analysis, no outlier was observed from this data set.

Although collinearity is not a problem for PLS, the use of the reduced number of significant descriptors may be able to improve the performance of the model (Andersson et al., 2002; Seggiani and Pannocchia, 2003; Tantishaiyakul and Wongpuwarak, 2005). To obtain a model containing fewer descriptors, the insignificant or redundant descriptors were removed individually. Such descriptors include Vdw, binding energy, total energy, polarizability, MW, HA, ARR, MR, PSA, and M log P. According to the descriptor selection, MR is removed while the molar refractivity obtained from HyperChem is retained. This may indicate that the former (Ghose-crippen molar refractivity) from Dragon is probably less informative than the latter.

As can be seen in Table 3, reducing a number of descriptors certainly does not produce negative effects on the predictive abilities of PLS models. The resulting PLS model provides a comparable but slightly more balanced model (model 1) than the original one with RMSEcv$^{\text{tr}}$ of 2.096 and RMSE$^{\text{tr}}$ of 1.857. This demonstrates that the information contained in the descriptors is successfully used in this new derived model. The final reduced model, thus, involved 9 descriptors including heat of formation, dipole moment, molar refractivity, 1/SA, 1/SV,

Table 3
PLS statistics of the derived PLS models

	R^2	Q^2	N^{PC}	N^{tr}	F	RMSE$^{\text{tr}}$	P	RMSEcv$^{\text{tr}}$	N^{ts}	RMSEp$^{\text{ts}}$
Model 1	0.853	0.813	2	51	138.88	1.857	<0.001	2.096		
Model 2	0.854	0.801	2	34	90.70	1.855	<0.001	2.172	17	1.883

R^2, squared correlation coefficient; Q^2, squared cross-validated correlation coefficient; N^{pc}, number of PLS components; N^{tr}, number of compounds in the training set; F, ordinary F value; RMSE$^{\text{tr}}$, root mean squared error for the dependent variable of the training set; P, level of significance; RMSEcv$^{\text{tr}}$, root mean squared error for the dependent variable from the cross-validation procedure of the training set; N^{ts}, number of compounds in the test set; RMSEp$^{\text{ts}}$, root mean squared error for the dependent variable of the test set.

HD, HB, Ui, and Hy; these variables can undoubtedly supply an adequate amount of significant information for the solubility parameter predictive model.

To estimate the external predictive ability of this multivariate PLS model, 17 compounds were selected as the external test set and the remaining 34 compounds were employed as a training set. The developed model 2 based on the molecular descriptors using in model 1 shows a good predictive ability as shown in Table 3. In addition, Fig. 1 presents that models 1 and 2 give a good fit to the one-to-one correlation line.

The types of descriptors providing the best predictive results in these PLS models are related to energy, surface area, surface volume, hydrogen bond and polar component of compounds. In general, HyperChem accurately calculates heat of formation by subtracting atomic heats of formation from the binding energy. This descriptor is somewhat related to cohesive energy (Yingling et al., 2001; Campbell and Starr, 2002), therefore, it is selected in the models. SV is also an important parameter and associated with δ calculation as indicated in Eq. (1). SA is interrelated to SV, it is significant for predicting δ values as well. Additionally, dipole moment, which obviously describes the polar component of the compounds is one of the most important descriptors of the derived models. Two descriptors HD and HB are precisely the hydrogen bonding components, and thus they were selected for generating the models. The HA descriptor may be redundant for the models which comprise both HD, and HB descriptors as well as other additional related descriptors such as Hy. This Hy descriptor directly reflects the hydrophilicity of a molecule. Another employed descriptor is Ui, which is calculated based on a number of multiple bonds in the molecule including double bonds, triple bonds and aromatic bonds. Regarding the molar refractivity descriptor, this parameter represents the volume of the molecules and also accounts for their dispersion and polar components (Barton, 1975). The scaled regression coefficients of models 1 and 2 are presented in Table 4. Accordingly, this suggests that with increasing heat of formation, dipole moment, HD, HB, Ui and Hy, the compounds get higher δ values. Meanwhile, the increase of molar refractivity, SA and SV reduces the compounds' δ values.

Table 4
Scaled PLS regression coefficients of models 1 and 2

Descriptor	Model 1	Model 2
Constant	16.039	16.321
Heat of formation	0.845	0.810
Dipole moment	1.104	1.119
Molar refractivity	−0.068	−0.079
Hydrogen bonding donor	1.476	1.353
Hydrogen bond-forming ability	1.170	1.200
Unsatuarated index	0.723	0.752
Hydrophilic index	1.673	1.638
1/solvent accessible surface area	0.469	0.393
1/surface-bounded molecular volume	0.496	0.421

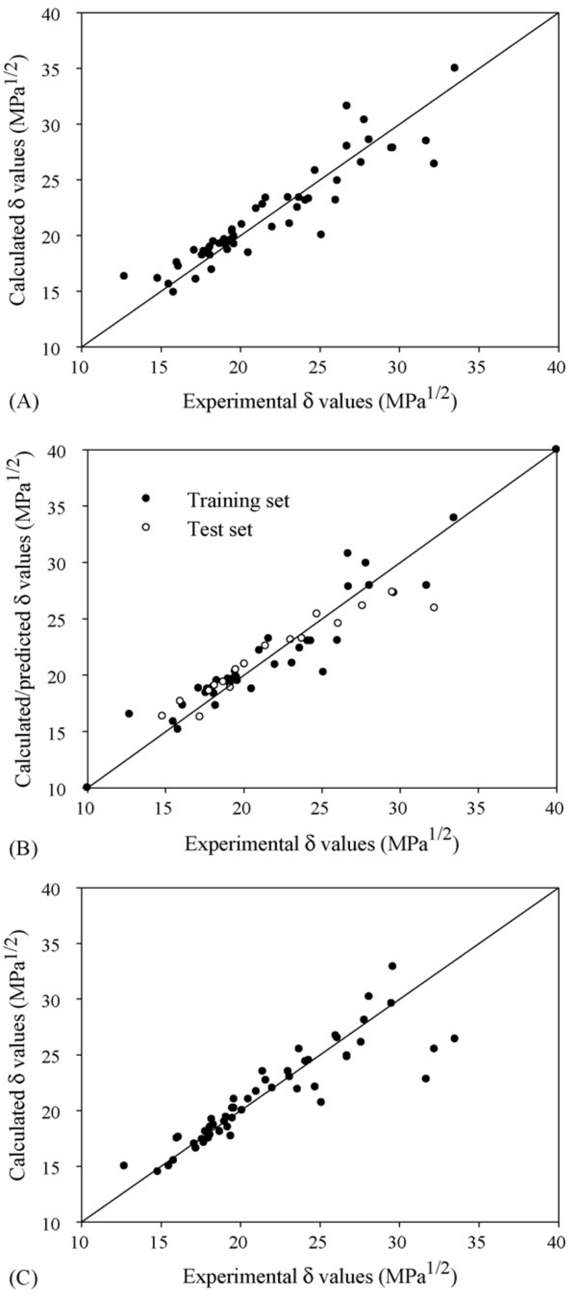

Fig. 1. Relationship between experimental vs. calculated and/or predicted solubility parameters of (A) model 1, (B) model 2 and (C) Hansen's method.

Furthermore, the calculations of δ values of these 51 compounds using model 1 were compared to the well established Hansen's, van Krevelen and Hoftyzer's, and Hoy's approaches. As shown in Table 1, RMSE of model 1 is slightly less than those obtained from the established methods, indicating the somewhat higher prediction ability of the derived PLS model. This relationship between experimental and calculated δ values using Hansen's approach is also presented in Fig. 1.

In conclusion, the PLS models with good predictability for δ values were developed in this study. These reliable models are based on simple calculated descriptors from structures. Such method of descriptor calculation has advantages over a group contribution approach in terms of its accounting for the interactions between neighboring groups, and also its applicability for the new molecule containing a novel functional group not previously reported. As such, these models are beneficial for design of new compounds with the required δ values.

References

Andersson, P.L., Maran, U., Fara, D., Karelson, M., Hermens, J.L.M., 2002. General and class specific models for prediction of soil sorption using various physicochemical descriptors. J. Chem. Inf. Comput. Sci. 42, 1450–1459.

Barra, J., Pena, M.-A., Bustamante, P., 2000. Proposition of group molar constants for sodium to calculate the partial solubility parameters of sodium salts using the van Krevelen group contribution method. Eur. J. Pharm. Sci. 10, 153–161.

Barton, A.F.M., 1975. Solubility parameters. Chem. Rev. 75, 731–753.

Bustamante, P., Pena, M.A., Barra, J., 1998. Partial-solubility parameters of naproxen and sodium diclofenac. J. Pharm. Pharmacol. 50, 975–982.

Campbell, C.T., Starr, D.E., 2002. Metal adsorption and adhesion energies on MgO (100). J. Am. Chem. Soc. 124, 9212–9218.

Choi, P., Kavassalis, T.A., Rudin, A., 1992. Estimation of three-dimensional solubility parameters of alkyl phenol ethoxylates using molecular dynamics simulations. J. Colloid Interfece Sci. 150, 386–393.

Choi, P., Kanavassalis, T.A., Rudin, A., 1994. Estimation of Hansen solubility parameters for (hydroxyethyl)- and (hydroxypropyl) cellulose through molecular simulation. Ind. Eng. Chem. Res. 33, 3154–3159.

Choi, P., Kavassalis, T., Rudin, A., 1996. Measurement of three-dimensional solubility parameters of nonyl phenol ethoxylates using inverse gas chromatography. J Colloid Interface Sci. 180, 1–8.

Chong, I.-G., Jun, C.-H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemom. Intell. Lab. Syst. 78, 103–112.

Forster, A., Hempenstall, J., Tucker, I., Rades, T., 2001. Selection of excipients for melt extrusion with two poorly water-soluble drugs by solubility parameter calculation and thermal analysis. Int. J. Pharm. 226, 147–161.

Gani, R., Jimenez-Gonzalez, C., Constable, D.J.C., 2005. Method for selection of solvents for promotion of organic reactions. Comput. Chem. Eng. 29, 1661–1676.

Goharshadi, E.K., Hesabi, M., 2004. Estimation of solubility parameter using equations of state. J. Mol. Liq. 113, 125–132.

Greenhalgh, D.J., Williams, A.C., Timmins, P., York, P., 1999. Solubility parameters as predictors of miscibility in solid dispersions. J. Pharm. Sci. 88, 1182–1190.

Hancock, B.C., York, P., Rowe, R.C., 1997. The use of solubility parameters in pharmaceutical dosage form design. Int. J. Pharm. 148, 1–21.

Hansen, C.M., 1967. The three-dimensional solubility parameters. Key to paint component affinities. II. Dyes, emulsifiers, mutual solubility and compatibility and pigments. J. Paint Technol. 39, 505–511.

Hansen, C.M., Skaarup, K., 1967. The three dimensional solubility parameter—key to paint component affinities. III. Independent calculation of the parameter components. J. Paint Technol. 39, 511–520.

Hildebrand, J.H., 1936. The Solubility of Non-Electrolytes. Reinhold, New York.

Hoy, K.L., 1970. New values of the solubility parameters from vapor pressure data. J. Paint Technol. 42, 76–80.

Kavassalis, T.A., Choi, P., Rudin, A., 1993. The calculation of 3D solubility parameters using molecular models. Mol. Simul. 11, 229–241.

Martin, A., Newburger, J., Adjei, A., 1980. Extended Hildebrand solubility approach: solubility of theophylline in polar binary solvents. J. Pharm. Sci. 69, 487–491.

Martini, L.G., Avontuur, P., George, A., Willson, R.J., Crowley, P.J., 1999. Solubility parameter and oral absorption. Eur. J. Pharm. Biopharm. 48, 259–263.

Minghetti, P., Cilurzo, F., Casiraghi, A., Montanari, L., 1999. Application of viscometry and solubility parameters in miconazole patches development. Int. J. Pharm. 190, 91–101.

Price, J.A., Shillcock, I.M., 2002. Inverse gas chromatographic measurement of solubility parameters in liquid crystalline systems. J. Chromatogr. A 964, 199–204.

Rey-Mermet, C., Ruelle, P., Nam-Tran, H., Buchmann, M., Kesselring, U.W., 1991. Significance of partial and total cohesion parameters of pharmaceutical solids determined from dissolution calorimetric measurements. Pharm. Res. 8, 636–642.

Roy, S.D., Flynn, G.L., 1998. Solubility and related physicochemical properties of narcotic analgesics. Pharm. Res. 15, 1370–1375.

Seggiani, M., Pannocchia, G., 2003. Prediction of coal ash thermal properties using partial least-squares regression. Ind. Eng. Chem. Res. 42, 4919–4926.

Subrahmanyam, C.V.S., Suresh, S., 1999. Solubility behaviour of haloperidol in individual solvents determination of partial solubility parameters. Eur. J. Pharm. Biopharm. 47, 289–294.

Suga, Y., Takahama, T., 1996. Application of molecular simulation to prediction of solubility parameter. Chem. Lett., 281–282.

Tantishaiyakul, V., Wongpuwarak, W., 2005. Prediction of Pgp–ATPase interaction and rhodamine 123 efflux inhibitory activities of propafenone analogs using PLS statistics. J. Mol. Struct. (THEOCHEM) 718, 183–189.

van Krevelen, D.W., 1990. Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions, 3rd ed. Elsevier, Amsterdam.

Wagner, K.G., Dowe, U., Zadnik, J., 2005. Highly loaded interactive mixtures for dry powder inhalers: prediction of the adhesion capacity using surface energy and solubility parameters. Pharmazie 60, 339–344.

Xia, C.Q., Yang, J.J., Ren, S., Lien, E.J., 1998. QSAR analysis of polyamine transport inhibitors in L1210 cells. J. Drug Target 6, 65–77.

Yen, T.E., Agatonovic-Kustrin, S., Evans, A.M., Nation, R.L., Ryand, J., 2005. Prediction of drug absorption based on immobilized artificial membrane (IAM) chromatography separation and calculated molecular descriptors. J. Pharm. Biomed. Anal. 38, 472–478.

Yingling, Y.G., Zhigilei, L.V., Garrison, B.J., 2001. The role of the photochemical fragmentation in laser ablation: a molecular dynamics study. J. Photochem. Photobiol. A 145, 173–181.

Zhao, L., Choi, P., 2001. Determination of solvent-independent polymer–polymer interaction parameter by an improved inverse gas chromatographic approach. Polymer 42, 1075–1081.