Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE)

Yingqing Ran,* Neera Jain, and Samuel H. Yalkowsky

Department of Pharmaceutical Sciences, College of Pharmacy, The University of Arizona, 1703 East Mabel Street, Tucson, Arizona 85721

Received February 28, 2001

The revised general solubility equation (GSE) is used along with four different methods including Huuskonen's artificial neural network (ANN) and three multiple linear regression (MLR) methods to estimate the aqueous solubility of a test set of the 21 pharmaceutically and environmentally interesting compounds. For the selected test sets, it is clear that the GSE and ANN predictions are more accurate than MLR methods. The GSE has the advantages of being simple and thermodynamically sound. The only two inputs used in the GSE are the Celsius melting point (MP) and the octanol water partition coefficient (K_{ow}). No fitted parameters and no training data are used in the GSE, whereas other methods utilize a large number of parameters and require a training set. The GSE is also applied to a test set of 413 organic nonelectrolytes that were studied by Huuskonen. Although the GSE uses only two parameters and no training set, its average absolute errors is only 0.1 log units larger than that of the ANN, which requires many parameters and a large training set. The average absolute error AAE is 0.54 log units using the GSE and 0.43 log units using Huuskonen's ANN modeling. This study provides evidence for the GSE being a convenient and reliable method to predict aqueous solubilities of organic compounds.

INTRODUCTION

The aqueous solubility of a drug is an important factor that influences its release, transport, and absorption in the body. Poor aqueous solubility often causes a drug to appear inactive and may cause other biological problems. A fast and reliable approach to estimate aqueous solubility can be useful in predicting the biological activities of drugs. Numerous methods[1−12] to estimate aqueous solubilities of organic compounds have been reported. The general solubility equation (GSE) of Yalkowsky and Valvani[1] has been widely used in the pharmaceutical field. Recently, Jain and Yalkowsky[5] revised the GSE on thermodynamic grounds. Using a test set of 580 organic compounds, they also showed that the revised GSE is more accurate than original GSE. Ran and Yalkowsky[6] validated the revised GSE for 150 additional compounds.

The advantage of the GSE is that no training set or regression-generated coefficients are needed and only two parameters are used: the melting point (MP) and octanol−water partition coefficient (K_{ow}). Melting point data can be easily obtained from the CHEMFINDER, BEILSTEIN, and SCIFINDER websites as well as from a variety of handbooks. Partition coefficients can be either experimentally measured or calculated using software programs such as CLOGP.

Recently Huuskonen[7] described artificial neural network (ANN) and multiple linear regression (MLR) methods to estimate the aqueous solubilities of organic compounds. These two approaches were used along with two additional MLR methods to estimate of the solubilities of 21 pharmaceutically relevant compounds. Huuskonen[7] also calculated the aqueous solubility of a set of 413 organic compounds with the ANN method. In this study, we will compare the predictions of the general solubility equation of Jain and Yalkowsky[5] to those of the other four calculations.

Description of the Models. The first model for the estimation of aqueous solubility with a sound theoretical basis was proposed by Hansch[8] et al. They reasoned that solubility and partition coefficient are reciprocally related for organic liquids. The Hansch model is very similar to the general solubility equation for liquids. However, since it is not applicable to solids it will not be evaluated any further.

Artificial Neural Network, ANN, Model. More recently, Huuskonen[7] utilized molecular connectivity, shape factors, atom types, and electrotopological state indices in an artificial neural network.

Multiple Linear Regression, MLR, Models. There are a wide variety of multilinear regression approaches to aqueous solubility. Huuskonen[7] utilized the same parameters as used in the ANN model as parameters in a MLR model. Klopman[9] used another group contribution scheme in an MLR model, which does not include the melting point. Kühne[10] used an analogous group contribution scheme along with a melting point term.

General Solubility Equation, GSE. According to the revised general solubility equation the aqueous solubility of a nonelectrolyte is

$$\log S_w = 0.5 - \log K_{ow} - 0.01(\mathrm{MP} - 25) \qquad (1)$$

where K_{ow} is the octanol−water partition coefficient of the solute and MP is its melting point in °C. If the solute melts below 25 °C, its melting point is set equal to 25 so that the melting point term vanishes. There are no additional parameters or correction factors. (Note that in the original version of the GSE the constant in the above equation was set as 0.8 instead of 0.5 in the revised version.) The following is

* Corresponding author phone: (520)626-3847; fax: (520)626-4063; e-mail: ran@pharmacy.arizona.edu.

10.1021/ci010287z CCC: $20.00
Published on Web 07/18/2001

a brief description of the only assumptions of the general solubility equation:

1. The reduction in solubility due to the crystallinity of the solute (i.e., the ideal solubility) is described by the van't Hoff equation (which is similar in form to the Clausius-Clapyron equation)

$$\log \frac{S^C}{S^L} = \frac{\Delta S_m (T_m - T)}{2.303RT} \quad (2)$$

where S^C and S^L represent the solubilities of the crystal and the liquid, respectively, ΔS_m is the entropy of melting, and T_m and T are the melting point and study temperature (in Kelvin), respectively. At 25 °C (298 K) this becomes

$$\log \frac{S^C}{S^L} = \frac{\Delta S_m (\mathrm{MP} - 25)}{1364} \quad (3)$$

2. The entropy of melting of most organic compounds is approximated by Walden's rule, which states that

$$\Delta S_m = 13.5 \text{ cal/deg/mol} \quad (4)$$

which is analogous to Trouton's rule for the entropy of vaporization.

3. For liquid solutes the octanol−water partition coefficient is approximately equal to the octanol−water solubility ratio, S_o^L/S_w^L, so that

$$\log S_w^L = \log K_{ow} - \log S_o^L \quad (5)$$

4. Most organic liquids are completely miscible with octanol. (On the basis of regular solution theory this is true for all compounds which are less polar than ethanol and more polar than benzene.[1,5])

5. Pure octanol has a molarity of 6.3. If complete miscibility corresponds to 50 mole percent or 3.15 mol per liter, then

$$\log S_o^L = 0.50 \quad (6)$$

METHODS

The values of MP, ClogP, MlogP, and log S_{obs} of 21 organic compounds are given in Table 1. The log S_{obs} values were from the original paper of Yalkowsky.[2] Melting points were obtained from Merck Index, CHEMFIINDER and BEILSTEIN. Measured and calculated partition coefficients (MlogP and ClogP, respectively) were obtained from CLOGP software.

Not all of the calculations could be performed on every compound. For each calculation performed on N compounds the average absolute error (AAE) and the root-mean-square error (RMSE) were determined by

$$\mathrm{AAE} = \frac{\sum |\log S_{calc} - \log S_{obs}|}{N} \quad (7)$$

and

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\log S_{calc} - \log S_{obs})^2}{N}} \quad (8)$$

Table 1. Physical Properties of the Drug Test Set Compounds

compd name	MP	ClogP	MlogP	log S_{obs}
antipyrine	111	0.20	0.23	0.39
aspirin	135	1.02	1.19	−1.61
atrazine	175	2.50	2.61	−3.55
benzocaine	89	1.92	1.86	−2.32
chlordane	25	5.80	6.00	−5.35
chlorpyriphos	43	4.49	4.82	−5.67
diazepam	125	3.16	2.99	−3.76
4,4′-DDT	109	6.76	6.91	−8.08
diazinon	>120	3.50	3.30	−3.76
diuron	159	2.68	2.68	−3.76
lindane	113	3.75	3.72	−4.60
malathion	25	2.70	2.38	−3.36
nitrofurantoin	>272	−0.47	−0.47	−3.38
parathion	25	3.47	3.83	−4.29
2,2′,4,5,5′-PCB	77	6.97	6.85	−6.77
phenobarbital	176	1.37	1.47	−2.34
phenolphthalein	264	2.63	2.63	−2.90
phenytoin	295	2.08	2.26	−3.99
prostaglandin E_2	67	2.01	2.82	−2.47
testosterone	155	3.22	3.32	−4.07
theophylline	272	−0.06	−0.02	−1.37

Table 2. Calculated Log Solubilities by Different Methods

	Huuskonen		Klopman	Kühne	Jain and Yalkowsky
compd name	ANN	MLR	MLR	MLR	GSE
antipyrine	−1.29	−1.20		−1.90	−0.59
aspirin	−1.69	−1.74	−1.52	−1.93	−1.79
atrazine	−3.51	−2.18	−3.05	−3.95	−3.61
benzocaine	−1.79	−1.85	−1.71		−2.00
chlordane	−7.29	−8.35	−7.55	−6.51	−5.50
chlorpyriphos	−5.61	−5.46	−5.77	−3.75	−4.50
4,4′-DDT	−7.67	−7.82	−8.00	−7.75	−7.25
diazepam	−4.05	−4.26		−4.51	−3.49
diazinon	−4.01	−4.10	−5.29	−4.98	<−3.75
diuron	−2.86	−3.20	−2.85	−3.38	−3.52
lindane	−4.71	−5.34	−4.88	−5.08	−4.10
malathion	−3.24	−3.63	−2.94	−3.48	−1.88
nitrofurantoin	−3.42	−3.03	−2.19	−2.62	<−1.50
parathion	−4.13	−3.98	−3.94	−4.59	−3.33
2,2′,4,5,5′-PCB	−7.21	−7.40	−7.90	−7.47	−6.87
phenobarbital	−2.97	−2.88	−2.08	−2.41	−2.48
phenolphthalein	−3.99	−4.05	−4.48	−4.61	−4.52
phenytoin	−3.40	−3.48	−3.47	−5.25	−4.46
prostaglandin E_2	−3.29	−4.35	−4.21		−2.74
testosterone	−3.98	−4.17	−5.17	−4.62	−4.12
theophylline	−1.71	−0.78	−1.07	0.54	−1.95

respectively. The predicted solubilities for the four methods discussed by Huuskonen for the 21 compounds are given in Table 2 along with those of the GSE (eq 1).

The values of MP, ClogP, MlogP, and log S_{obs} for Huuskonen's 413 compound test set are given in Table 3, along with the values of log S calculated by the GSE (eq 1) and by Huuskonen's ANN. Since the GSE uses the melting points to calculate log S, those compounds which decompose on melting or which do not have reported melting points were not used. The 22 compounds which decompose and the 3 with missing values are designated "dec" and "missing", respectively, in the comment column. Also, eight of the solubilities used by Huuskonen were calculated rather than experimental values. These are indicated by "calc" in the comment column and were not used in the comparison.

Of the 380 compounds used in the comparison, five are based on measured MlogP values because they were greatly different from the ClogP values. Finally the solubilities of seven compounds were corrected because the values used

Table 3. Comparison of GSE and Huuskonen's ANN[a]

compd name	comment	MP	ClogP	MlogP	log S_{obs}	log S_{ANN}	log S_{GSE}
abietic acid	calc	173	6.28		−3.80	−4.40	
acetazoleamide	MLOGP	258	−1.25	−0.26	−2.36	−2.60	−1.57
acridine		107	3.41	3.40	−3.67	−3.53	−3.73
acrylamide		84	−0.61	−0.67	0.96	0.53	0.515
adipic acid		152	−0.02	0.08	−0.82	−0.35	−0.75
aldosterone	MLOGP	164	−0.06	1.08	−3.85	−3.40	−1.97
aldrin	MLOGP & corr	104	5.41	6.50	−6.31	−6.37	−6.79
alizarin		287	2.38		−2.78	−2.63	−4.50
allantoin		238	−2.23		−1.60	−1.25	0.60
allobarbital		171	0.75	1.15	−2.06	−1.84	−1.71
alloxanthin	dec	253	−3.92		−2.23	−3.08	
alpha-estradiol		220	3.78	3.86	−4.84	−4.11	−5.23
4-aminobenzoic acid		189	0.98	0.83	−0.40	−1.06	−2.12
aminocarb		93	1.98	1.90	−2.36	−1.78	−2.16
aminophenazone		108	−0.75		−0.62	−1.23	0.42
2-(4-aminophenyl)−6-methylbenzothiazole		191	3.64		−3.68	−4.22	−4.80
2-aminothiazole	dec	93	0.22	0.38	−0.36	0.19	
amitriptyline		193	4.85	4.92	−4.46	−5.24	−6.03
aniline		−6	0.91	0.90	−0.41	−0.57	−0.41
1-anthranol		158	3.82		−4.73	−4.14	−4.65
asulam		153	−0.31	−0.27	−1.66	−2.01	−0.47
1,4-benzenediol		170	0.81	0.59	−0.17	−0.13	−1.76
benzenesulfonamide		151	0.30	0.31	−1.56	−1.39	−1.06
benzhydrol		66	2.45	2.67	−2.55	−3.15	−2.36
benzo[*b*]fluorene	MLOGP	208	5.25	5.77	−8.04	−6.43	−7.10
benzo[*e*]pyrene		178	6.12	6.44	−7.80	−8.22	−7.15
benzo(*ghi*)perylene		278	6.58	6.22	−9.03	−9.28	−8.61
benzo[*k*]fluoranthene		216	6.12	6.11	−8.49	−8.20	−7.53
benzoin		134	2.38	2.13	−2.85	−3.77	−2.97
benzonitrile		−13	1.57	1.56	−1.00	−1.76	−1.07
benzophenone		49	3.18	3.18	−3.12	−3.83	−2.92
benzothiazole		2	2.08	2.01	−1.50	−1.68	−1.58
benzylurea		148	0.73	0.73	−0.95	−1.60	−1.46
beta-iodopropionic acid		82	0.93		−0.43	−1.18	−1.00
betamethasone	dec	231	1.75	2.01	−3.77	−3.74	
bibenzyl		50	4.59	4.79	−4.62	−4.91	−4.34
bis-(2-chloroethyl)ether	corr	−52	1.00	1.29	−0.92	−1.90	−0.50
bis-(*p*-aminophenyl)methane		93	1.75	1.59	−2.30	−3.34	−1.93
4-bromoacetanilide		165	2.28	2.29	−3.08	−2.36	−3.18
bromobenzene		−31	3.01	2.99	−2.55	−2.37	−2.51
1-bromo-2-chlorobenzene		−12	3.58		−3.19	−3.30	−3.08
1-bromo-3-chloropropane		−59	1.85		−1.85	−2.18	−1.35
bromocyclohexane		−57	3.33	3.20	−2.30	−2.69	−2.83
bromodichloromethane		−57	2.09	2.10	−1.54	−1.54	−1.59
1-bromoheptane		−58	4.25	4.36	−4.43	−4.14	−3.75
bromophos		51	5.09	5.21	−6.09	−5.43	−4.85
2-bromopropane		−89	2.13	2.14	−1.59	−1.19	−1.63
3-bromopropylene		−50	1.65	1.79	−1.50	−0.84	−1.15
brompyrazone		220	1.40		−3.12	−2.64	−2.85
1,3-butadiene		−109	1.90	1.99	−1.87	−1.31	−1.40
butanethiol		−116	2.23	2.28	−2.18	−0.81	−1.73
1-butanol		−90	0.82	0.88	0.00	−0.03	−0.32
2-butanol		−115	0.60	0.61	0.43	0.22	−0.10
2-butoxyethane		−70	0.84	0.83	−0.42	−0.89	−0.34
buturon		143	2.66	3.00	−3.90	−2.82	−3.34
bytylate		<25	4.01	4.15	−3.68	−2.01	−3.51
camphor		177	2.18	2.38	−1.99	−2.26	−3.20
caproic acid		−3	1.92	1.92	−1.06	−0.94	−1.42
carbetamide	missing		1.35		−1.83	−2.66	
carbophenothion		<25	5.94	5.33	−5.74	−6.06	−5.44
carbromal		118	1.62	1.54	−2.68	−2.16	−2.05
cetyl alcohol		56	7.17		−7.26	−6.18	−6.98
chloralose		183	0.41	1.02	−1.84	−1.71	−1.49
chloramphenicol		151	1.28	1.14	−2.11	−3.41	−2.04
chlordimeform		32	2.79	2.89	−2.86	−2.97	−2.36
chlorfenac		161	3.43	3.20	−3.08	−3.57	−4.29
chloroacetamide		117	−0.50	−0.53	−0.02	−0.17	0.08
2-chloroaniline		−2	1.91	1.90	−1.52	−1.68	−1.41
4-chlorobiphenyl		78	4.74	4.61	−5.20	−5.12	−4.77
1-chloroheptane		−70	4.11	4.15	−3.99	−3.58	−3.61
2-chloro-2-methylbutane		−74	2.92	2.52	−2.51	−1.95	−2.42
1-chloro-2-methylpropane		−131	2.39		−2.00	−1.36	−1.89
2-chlorophenol		7	2.15		−1.06	−1.34	−1.65
2-chloropropane		−117	1.99	1.90	−1.41	−0.69	−1.49

Table 3 (Continued)

4-chlorophenoxyacetic acid		157	2.20	2.25	−2.29	−2.12	−3.02
3-chloropentane		−106	3.05		−2.63	−2.10	−2.55
chloropropylate		73	4.87		−4.53	−5.22	−4.85
3-chloropropylene		−135	1.51		−1.36	−0.42	−1.01
chlorothalonil		250	3.47	2.90	−5.64	−5.92	−5.22
chloroxuron		151	3.79		−4.89	−3.72	−4.55
chlorpropamide		127	2.35	2.27	−3.03	−3.05	−2.87
chlorpropham		41	3.37	3.51	−3.38	−2.91	−3.03
chlorpyriphos-methyl	MLOGP	46	3.81	4.31	−4.82	−5.42	−4.02
chrysene		256	5.66	5.81	−8.06	−7.18	−7.47
cimetidine		142	0.35	0.40	−1.35	−3.34	−1.02
citral		<10	2.95		−2.06	−2.39	−2.45
citric acid		153	−1.56	−1.72	0.51	0.43	0.78
cocaine		98	2.57	2.30	−2.23	−2.59	−2.80
coniine		−18	1.98		−1.50	−0.35	−1.48
cortisone acetate		235	1.83	2.10	−4.00	−4.06	−3.43
coumaphos		91	4.33	4.13	−5.38	−5.79	−4.49
coumarin		71	1.41	1.39	−1.89	−2.22	−1.37
cycloate		12	3.66	3.97	−3.40	−2.53	−3.03
cyclohexane		7	3.35	3.44	−3.10	−3.15	−2.85
cyclohexanenecarboxylic acid		30	1.84		−1.45	−0.97	−1.39
cyclooctane		15	4.47	4.45	−4.15	−4.44	−3.97
cyclopentene		−135	2.31		−2.10	−2.14	−1.81
d-camphoric acid		184	1.75		−1.42	−1.26	−2.84
deoxycholic acid		174	4.51		−3.95	−4.43	−5.50
deoxycorticosterone		136	3.25	2.83	−3.75	−4.27	−3.86
deoxycorticosterone acetate	calc	−136	3.79	3.08	−4.63	−5.14	
d-fenchone		6	2.18		−1.85	−2.24	−1.68
di-(2-ethylhexyl) phthalate		−50	8.71	7.45	−6.96	−7.60	−8.21
dibenz(*a*,*h*)anthracene		266	6.84	6.50	−8.66	−8.76	−8.75
dibenzamid	calc	130	0.65	0.64	−2.27	−3.57	
dibenzothiophene		97	4.56	4.49	−4.38	−5.22	−4.78
1,3-dibromobenzene		−7	3.87	3.75	−3.54	−3.68	−3.37
1,4-dibromobenzene		87	3.87	3.79	−4.07	−3.68	−3.99
1,2-dibromoethylene		<25	1.95		−1.32	−1.66	−1.45
dibromomethane		−53	1.53	1.88	−1.17	−1.37	−1.03
1,2-dibromopropane		−55	2.27		−2.15	−2.46	−1.77
1,3-dibromopropane		−34	1.99	2.37	−2.08	−2.64	−1.49
dibucaine		64	5.34	4.40	−3.70	−3.95	−5.23
1,3-dichlorobenzene		−25	3.57	3.53	−3.04	−2.98	−3.07
1,4-dichlorobenzene		53	3.57	3.44	−3.27	−2.97	−3.35
1,3-dichloro-5,5-dimethylhydantoin		145	0.47		−2.60	−2.33	−1.17
1,3-dichloro-2-propanol		−4	0.20		−0.11	−1.02	0.30
dicaphton		53	3.55	3.88	−4.31	−4.75	−3.33
1,2-dichloroethane		−35	1.46	1.47	−1.06	−1.16	−0.96
1,2-dichloroethylene		−81	1.77	1.86	−1.30	−1.09	−1.27
1-dihydrocarvone	calc				−2.18	−1.83	
1,1-diethoxyethane		−100	0.93	0.84	−0.43	−0.86	−0.43
dimetan		46	1.83		−0.85	−1.34	−1.54
2,2-dimethyl-1-butanol		−35	1.62		−1.04	−0.67	−1.12
2,3-dimethyl-2-butanol		−14	1.40	1.48	−0.41	−0.41	−0.90
5,5-dimethyl-2,4-hexadione	calc	224	1.00		−1.63	−1.13	
2,3-dimethylnaphthalene		102	4.26	4.40	−4.72	−4.64	−4.53
2,6-dimethyl-4-pyrimidiamine		180	0.63	0.39	−1.28	0.00	−1.68
2,2-dimethyl-3-pentanol		−5	1.93		−1.15	−0.99	−1.43
2,3-dimethyl-2-pentanol		<25	1.93		−0.89	−0.98	−1.43
2,3-dimethyl-3-pentanol		<−30	1.93		−0.85	−0.96	−1.43
dimethyl-phathalate		2	1.56	1.56	−1.66	−2.45	−1.06
2,2-dimethyl-1-propanol		50	1.09	1.31	−0.40	−0.08	−0.84
3,4-dimethylpyridine		−12	1.59		0.36	−0.33	−1.09
3,5-dimethylpyridine		−7	1.64	1.78	0.38	−0.30	−1.14
2,4-dimethylquinoline		<25	3.03		−1.94	−2.37	−2.53
1,3-dimethylthiourea	calc	78	0.79	0.57	−1.46	−1.64	
dinoseb		56	3.54	3.56	−3.38	−3.61	−3.35
diphenic acid		227	1.80	2.07	−2.28	−3.12	−3.32
diphenyl ether		86	4.24	4.21	−3.96	−3.45	−4.35
diphenylpropane	corr	150	3.67	3.32	−3.28	−3.31	−4.42
dipropropylamine		−63	1.60	1.67	−0.46	−0.87	−1.10
dl-1,2-diphenylethanol		67	2.98		−2.52	−3.55	−2.90
dl-tropic acid		119	0.43	0.77	−0.93	−1.40	−0.87
endrin	dec	200	3.70	4.55	−6.18	−6.49	
equilenin		258	3.27		−5.24	−4.73	−5.10
eriodictyol	dec	257	1.84	2.02	−3.62	−2.04	
erythriol		122	−1.71	−2.29	0.70	0.81	1.24
ethanethiol		−148	1.17		−0.60	0.13	−0.67

Table 3 (Continued)

ethinyl estradiol		142	3.86	3.67	−4.30	−4.22	−4.53
ethiofencarb		<25	2.20	2.04	−2.09	−3.14	−1.70
ethyl acetate		−84	0.71	0.73	−0.04	0.07	−0.21
ethylbenzene		−95	3.17	3.15	−2.77	−2.26	−2.67
ethyl biscoumacetate		177	3.21		−3.66	−4.41	−4.23
2-ethylbutric acid		−14	1.92	1.68	−0.81	−0.73	−1.42
ethyl caproate		−68	2.83		−2.31	−2.09	−2.33
ethyl caprylate		−47	3.88		−3.39	−3.22	−3.38
ethylenethiourea		203	−0.66	−0.66	−0.71	−0.30	−0.62
2-ethylnaphthalene		−7	4.34	4.38	−4.29	−4.99	−3.84
2-ethylhexylmine		−76	2.91	2.82	−1.71	−2.57	−2.41
ethyl isopropyl ether		<25	1.18		−0.55	−0.53	−0.68
ethyl nonanoate		−36	4.41		−3.80	−3.75	−3.91
etryptamine		97	2.26		−2.57	−2.73	−2.48
fenarimol		117	2.86	3.60	−4.38	−4.58	−3.28
fenitrothion		3	3.21	3.30	−4.04	−4.02	−2.71
fenthion		8	3.84	4.09	−4.57	−4.55	−3.34
formetanate	missing		0.56		−2.34	−2.20	
2-furoic acid		130	1.06		−0.48	−1.15	−1.61
furosemide		206	1.87	2.03	−3.66	−3.46	−3.18
gallic acid	dec	222	0.43	0.70	−1.16	−1.14	
gamma-butyrolatone		−45	−0.39		1.07	0.41	0.89
glucose		146	−2.21	−2.41	0.74	0.57	1.50
glybuthiazole		221	1.54		−3.74	−3.43	−3.00
guaiacol	corr	28	1.32	1.32	−0.68	−0.74	−0.85
haloperidol		148	3.85	4.28	−4.43	−4.44	−4.58
hematein	dec	250	−0.18		−2.70	−2.08	
heptachlor		95	4.92		−6.32	−6.55	−5.12
heptachlor epoxide	MLOGP	160	3.87	4.98	−6.29	−6.72	−5.83
3-heptanol		−70	2.19	2.24	−1.47	−1.50	−1.69
4-heptanone		−33	1.91	2.04	−1.30	−1.61	−1.41
heptylamine		−23	2.51	2.57	−1.85	−2.29	−2.01
hexachlorobutadiene		−21	4.90	4.78	−4.91	−4.08	−4.40
1,2,3,6,7,8-hexahydropyrene		133	5.34		−5.96	−6.21	−5.92
2-hexanol		<25	1.66	1.76	−0.89	−0.94	−1.16
2-hexanone		−57	1.38	1.38	−0.80	−1.07	−0.88
hexestrol		185	5.11		−4.43	−4.40	−6.21
hexyl acetate		−112	3.75	4.13	−2.46	−2.25	−3.25
hexylamine		−23	1.98	2.06	−1.10	−1.65	−1.48
4-hexylresorcinol		65	3.95	3.45	−2.59	−2.79	−3.85
hippuric acid		189	0.65	0.31	−1.68	−1.94	−1.79
hydantoin		222	−1.69	−1.69	−0.40	0.27	0.22
hydrastine		132	2.08		−4.11	−3.25	−2.65
hydrazobenzene		131	2.97	2.94	−2.92	−2.48	−3.53
hydrocinchonine		268	2.97		−2.63	−3.61	−4.90
hydrocortisone	dec	215	1.70	1.61	−2.97	−3.23	
1-hydroxychlordene		201	3.36		−5.46	−5.25	−4.62
hydroxyisoandrosterone		193	2.63		−3.59	−3.46	−3.81
4-hydroxy-2-methylquinoline		235	2.81		−1.20	−1.57	−4.41
3-hydroxytetrahydrofuran		<25	−0.78		1.05	0.69	1.28
hypoxanthine	dec	150	−1.20	−1.11	−2.29	−1.36	
indane		−51	3.15	3.18	−3.04	−2.59	−2.65
indazole		147	1.63	1.77	−2.16	−0.67	−2.35
inosine	dec	212	−3.16	−2.10	−1.23	−2.60	
iodobenzene		−29	3.27	3.25	−2.78	−3.20	−2.77
iodoethane		−108	2.00		−1.60	−1.71	−1.50
isoamyl salicylate		<25	4.32		−3.16	−2.66	−3.82
isopentanol		−117	1.22	1.16	−0.52	−0.31	−0.72
isophrone		−8	2.09		−1.06	−1.64	−1.59
isopropyl acetate		−73	1.24	1.22	−0.55	−0.24	−0.74
isoproturon		158	2.40	2.50	−3.54	−2.39	−3.23
kasugamycin		169	−1.47		−2.93	−2.69	0.53
kebuzone		128	0.90		−3.27	−3.59	−1.43
khellin		154	2.57		−2.40	−4.03	−3.36
levodopa		295	−2.82	−2.74	−1.60	−1.36	0.62
linalool		<25	2.75		−1.99	−1.85	−2.25
l-menthone		<25	2.83		−2.49	−2.44	−2.33
L-tyrosine	dec	342	−2.22	−2.26	−2.57	−1.47	
m-difluorobenzene		−59	2.43	2.21	−2.00	−1.48	−1.93
mebendazole		289	3.06	2.83	−3.88	−3.77	−5.20
medinoterb acetate		86	3.25		−4.47	−4.56	−3.36
mefenamic acid	corr	230	4.94	5.12	−4.08	−3.33	−6.49
menadione		105	2.45	2.20	−3.03	−2.97	−2.75
menthol		45	3.23	3.23	−2.53	−2.21	−2.93
7-mercaptopteridine	missing		−0.21		−2.71	−2.70	

Table 3 (Continued)

methacrylic acid		16	0.66	0.93	0.01	0.43	−0.16
methoxsalen		143	2.30	1.93	−3.66	−3.84	−2.98
methoxychlor		89	5.17	5.08	−6.89	−7.28	−5.31
methyl acetate		−98	0.18	0.18	0.52	0.53	0.32
2-methylaniline		−15	1.36	1.32	−0.85	−1.14	−0.86
4-methylaniline		43	1.41	1.39	−1.21	−1.15	−1.09
9-methylanthracene		76	4.99	5.07	−5.89	−5.91	−5.00
2-methylbenzimidazole	calc	176	1.83		−1.96	−1.17	
methyl benzoate		−12	2.11	2.12	−1.85	−1.85	−1.61
2-methyl-1,3-butadiene		−120	2.30		−2.03	−1.62	−1.80
2-methyl-1-butanol		−70	1.22	1.29	−0.47	−0.33	−0.72
methylcapronate		−71	2.30	2.42	−2.00	−1.66	−1.80
methylcyclohexane		−126	3.87	3.61	−3.85	−3.77	−3.37
2-methylcyclohexanone		−14	1.38		−0.94	−1.06	−0.88
3-methyl-3-heptanol		−83	2.59		−1.60	−1.77	−2.09
methyl gallate		200	0.93		−1.24	−1.12	−2.18
methyl isopropyl ether		<25	0.65		−0.06	0.05	−0.15
1-methylnaphthalene		−22	3.81	3.87	−3.70	−4.08	−3.31
3-methylpentane		−118	3.74	3.60	−3.68	−3.67	−3.24
2-methyl-1-pentanol		<25	1.75		−1.11	−0.92	−1.25
2-methyl-2-pentanol		−107	1.53		−0.49	−0.63	−1.03
4-methyl-2-pentanol		−90	1.53		−0.80	−0.64	−1.03
4-methyl-2-pentanone	corr	−80	1.25	1.31	−0.72	−0.77	−0.75
4-methyl-1-pentene		−154	3.25		−3.24	−2.88	−2.75
6-methylprednisolone	dec	228	1.70		−2.99	−3.43	
methyl propionate		−88	0.71	0.82	−0.14	0.05	−0.21
7-methylpteridine		133	−0.36		0.06	−0.79	−0.22
methyl *tert*-butyl ether		−109	1.05	0.94	−0.24	−0.31	−0.55
3-methylthiophene		−69	2.29	2.45	−2.39	−1.38	−1.79
17-methyltestosterone		162	3.74	3.63	−3.99	−4.39	−4.61
5-methyluracil		316	−0.56	−0.62	−1.52	−0.38	−1.85
6-methyluracil		318	−0.56	−0.77	−1.26	−0.31	−1.87
metiazinic acid		146	3.90		−3.94	−3.34	−4.61
metronidazole		160	−0.46	−0.02	−1.26	−1.53	−0.39
monolinuron		80	2.31	2.30	−2.57	−2.45	−2.36
morphine	dec	254	0.59	0.76	−3.28	−2.01	
morpholine		−5	−0.41	−0.86	1.06	0.87	0.91
n-amyl carbamte		94	1.41	1.35	−1.47	−1.28	−1.60
naphthacene		341	5.66	5.90	−8.60	−7.26	−8.32
1-naphthylamine		49	2.09	2.25	−1.92	−2.91	−1.83
natamycin	dec	200	−4.55		−3.21	−3.83	
n-butylbenzene		−88	4.23	4.38	−4.06	−3.81	−3.73
n-hexane		−95	3.87	3.90	−3.84	−3.90	−3.37
n-hexylbenzene		−61	5.29	5.52	−5.21	−5.23	−4.79
nicotine		−8	0.90	1.17	0.79	−1.30	−0.40
niflumic acid		204	3.79	4.43	−4.17	−4.00	−5.08
niridazole		260	0.77	0.95	−3.22	−2.77	−2.62
nitrapyrin		63	3.42	3.41	−3.76	−3.65	−3.30
nitrazepam		224	2.31	2.13	−3.80	−3.63	−3.80
2-nitroacetanilide		93	1.00	1.00	−1.91	−2.37	−1.18
4-nitroaniline		146	1.26	1.39	−2.37	−1.62	−1.97
4-nirobenzoic acid		242	1.84	1.69	−2.80	−2.09	−3.51
3-nitrophenol		98	1.85	2.00	−1.01	−1.48	−2.08
1-nitropropane		−108	0.77	0.87	−0.80	−0.23	−0.27
2-nitropropane		−93	0.55	0.80	−0.62	−0.13	−0.05
2-nitrotoluene		−9	2.30	2.30	−2.33	−2.20	−1.80
4-nitrotoluene		55	2.38	2.37	−2.49	−2.23	−2.18
1-nitroso-1-ethylurea		103	0.32	0.23	−0.96	−0.51	−0.60
n-methylacetanilide		102	1.11	1.12	−0.95	−1.59	−1.38
n-methylaniline		−57	1.64	1.66	−1.28	−0.89	−1.14
n-methylanthranilic acid		172	2.36		−2.88	−1.52	−3.33
n-methylpyrrolidone		−24	−0.40	−0.54	1.00	0.45	0.90
n-methylurea		130	−1.30	−1.40	1.13	0.52	0.75
n-octyl carbamate		<25	3.00	2.84	−3.30	−2.84	−2.50
n-octylamine		8	3.04	3.03	−2.75	−2.91	−2.54
nonanal		63	3.48		−3.17	−3.29	−3.36
5-nonanone		−50	2.97	2.88	−2.59	−2.79	−2.47
norethisterone acetate		161	3.74		−4.79	−4.85	−4.60
O-benzyl carbamate		88	1.20	1.20	−0.35	−1.78	−1.33
1-octanol		−15	2.94	3.00	−2.39	−2.43	−2.44
o,p′-DDT		74	6.76		−6.62	−7.74	−6.75
o-chlorobenzoic acid		142	2.10	2.05	−1.89	−2.05	−2.77
o-cresol		31	1.97	1.99	−0.62	−0.84	−1.53
oryzalin	corr	141	2.92		−3.61	−4.04	−3.58
oxalic acid	dec	189	−3.03		0.38	0.45	

Table 3 (Continued)

palmitic acid		62	7.21		−6.81	−5.44	−7.08
p-aminoacetophenone		106	0.86	0.83	−1.61	−1.52	−1.17
parabanic acid		230	−2.54		−0.40	−0.54	0.89
parathion-methyl		36	2.79	2.86	−3.68	−3.74	−2.40
pathalic anhydride		131	1.60	1.60	−1.39	−2.49	−2.16
p-bromotoluene		29	3.50	3.42	−3.19	−3.10	−3.04
2,2′,3,3′,4,4′,5,5′,6-PCB		205	9.34	9.14	−10.26	−10.32	−10.64
2,2′,3,3′,4,4′,6-PCB		122	7.91		−8.30	−8.80	−8.38
2,2′,3,3′,4,5,5′,6,6′-PCB		204	9.21	8.16	−10.41	−10.31	−10.50
2,2′,3,3′,4,5-PCB		101	7.45	7.32	−8.42	−8.10	−7.71
2,2′,3,3′,4-PCB		119	6.73		−7.05	−7.41	−7.17
2,2′,3,3′,6,6′-PCB		114	7.07	7.12	−8.65	−8.02	−7.46
2,2′,3,4,4′,5′-PCB		135	7.45		−8.32	−8.05	−8.05
2,2′,3,4,5′-PCB		112	6.85	6.85	−7.91	−7.38	−7.22
2,2′,4,4′-PCB		42	6.38	6.29	−6.51	−6.77	−6.05
2,2′,5,5′-PCB		87	6.38	6.26	−7.00	−6.76	−6.50
2,2′,5-PCB		44	5.67	5.60	−6.02	−6.24	−5.36
2,3,4,5-PCB		92	6.39	6.41	−7.16	−6.88	−6.56
2,3,6-PCB		49	5.67		−6.29	−6.28	−5.41
2,4,4′-PCB		57	5.92	5.62	−6.21	−6.24	−5.74
3,3′-PCB		29	5.46	5.30	−5.80	−5.71	−5.00
p-chlorotoluene		7	3.35	3.33	−3.08	−2.66	−2.85
p-difluorobenzene		88	2.43	2.13	−1.97	−1.46	−2.56
pebulate		<25	3.74	3.84	−3.35	−2.30	−3.24
pentachlorobutadiene		<25	4.05		−4.23	−3.49	−3.55
pentachlorophenol		174	4.68	5.12	−4.28	−4.34	−5.67
pentamethylbenzene		54	4.49	4.56	−4.00	−4.46	−4.28
1-pentanol		−108	1.35	1.56	−0.60	−0.62	−0.85
1-pentyne		−106	1.98	1.98	−1.64	−2.27	−1.48
perfluidone		142	3.78		−3.80	−4.07	−4.45
permethrin		34	7.38	6.50	−6.29	−7.22	−6.97
phenantherene		100	4.49	4.47	−5.26	−5.52	−4.74
phenetole		−30	2.59	2.51	−2.33	−1.96	−2.09
phenothrin		<25	7.20		−5.24	−6.27	−6.70
2-phenyl-3,1-benzoxazin-4-one		124	2.55		−4.61	−4.06	−3.04
2-phenylethanol		−14	1.33	1.36	−0.74	−1.30	−0.83
phorate		−43	3.84	3.83	−4.11	−4.03	−3.34
phthalimide		238	1.15	1.15	−2.61	−2.19	−2.78
p-hydroxybenzoic acid		217	1.56	1.58	−1.41	−0.91	−2.98
pipemidic acid		253	−2.73	−2.15	−2.98	−3.81	0.95
2,5-piperazinedione	dec	233	−1.72		−0.83	0.17	
piperine		132	2.91		−3.46	−3.23	−3.48
piroxicam		210	1.89	1.98	−4.16	−3.42	−3.24
p-isopropyltoluene		−67	4.07	4.10	−3.77	−3.56	−3.57
prednisolone	dec	240	1.38	1.62	−3.21	−3.23	
prednisonone-21-trimethyl acetate		233	3.78		−4.58	−4.51	−5.36
pregnenolone		193	4.03	4.22	−4.65	−4.68	−5.21
primidone		281	0.88	0.91	−2.64	−2.60	−2.94
promazine		30	4.90	4.55	−4.30	−3.77	−4.45
promethazine		60	4.90	4.81	−4.26	−3.68	−4.75
prometryn		118	3.29	3.51	−4.10	−3.87	−3.72
propyl benzoate		−52	3.17	3.01	−2.67	−2.85	−2.67
propyl isopropyl ether		<25	1.71		−1.34	−1.19	−1.21
propylthiouracil		219	−0.33		−2.15	−2.03	−1.11
p−*tert*-butylphenol		98	3.30	3.31	−2.41	−2.51	−3.53
p-toluic acid		182	2.38	2.27	−2.60	−1.59	−3.45
p-xylene		13	3.14	3.15	−2.77	−2.21	−2.64
2-pyrazinecarboxamide		190	−0.71	−0.60	−0.91	−0.24	−0.44
3-pyridinemethanol		−8	−0.39	−0.02	0.96	0.17	0.89
pyrrolidone		24	−0.97	−0.85	1.07	0.62	1.47
quinethazone		250	0.55		−3.29	−2.90	−2.30
8-quinolinol		72	2.08	2.02	−2.42	−1.23	−2.05
reposal		213	2.19		−2.64	−2.61	−3.57
rolitetracycline	dec	162	0.48		−1.42	−2.74	
rotenone		165	4.19	4.10	−4.42	−4.86	−5.09
spironolactone	dec	201	2.25	2.26	−4.28	−5.66	
styrene		−30	2.87	2.95	−2.82	−1.74	−2.37
sulfaguanidine		190	−1.24	−1.22	−1.99	−2.13	0.09
sulfamethazine		198	1.07	0.28	−2.27	−2.88	−2.30
sulfanilamide		165	−0.57	−0.62	−1.36	−1.27	−0.33
sulfaperine		262	0.57	0.34	−2.82	−2.69	−2.44
sulfapyridine		191	0.84	0.00	−2.70	−2.33	−2.00
2,4,5-T		158	3.33	3.31	−2.96	−3.42	−4.16
tert-butylbenzene		−85	3.97	4.11	−3.66	−3.65	−3.47
tetrachloroethylene		−22	3.48	3.40	−2.54	−2.42	−2.98

Table 3 (Continued)

1,2,4,5-tetrafluorobenzene		4	2.57		−2.38	−2.11	−2.07
2,3,4,5-tetraiodopyrrol	dec	140	5.56		−3.46	−4.89	
tetramethylurea		−1	−0.11	0.19	0.94	0.85	0.61
tetroxoprim		153	0.63	0.56	−2.10	−2.74	−1.41
thiourea		175	−1.02	−1.02	0.32	−0.54	0.015
thiram		155	1.76	1.73	−3.90	−4.31	−2.56
tranid		159	0.89		−2.08	−2.72	−1.73
trans-2-heptene		−109	3.91		−3.82	−3.87	−3.41
triallate		29	4.73	4.60	−4.88	−3.73	−4.27
triamcinolone acetonide		274	2.17	2.53	−4.32	−4.21	−4.16
triamcinolone diacetate	calc	235	1.86	1.92	−4.13	−4.52	
trichlorfon		83	0.68	0.51	−0.22	−2.03	−0.76
trichlormethiazide	dec	266	0.85	0.56	−2.68	−3.63	
1,2,3-trichlorobenzene		53	4.04	4.14	−4.00	−3.85	−3.82
1,3,5-trichlorobenzene		65	4.28	4.19	−4.48	−3.81	−4.18
trichloroehylene		−86	2.63	2.61	−1.96	−1.74	−2.13
2,4,5-trichlorophenol		68	3.58	3.72	−2.21	−2.95	−3.51
1,2,3-trichloropropane		−15	1.98		−1.92	−2.24	−1.48
1,1,2-trichlorotrifluoroethane		−36	3.29	3.16	−3.04	−2.89	−2.79
tricresyl phosphate	corr	−33	5.95		−6.01	−7.00	−5.45
tricyclazole		187	1.83	1.70	−2.07	−3.11	−2.95
triethyl phosphate		−56	0.28	0.80	0.43	−0.63	0.22
trifluoperazine		232	5.21	5.03	−4.52	−5.08	−6.78
3-trifluoromethylaniline		5	2.29	2.29	−1.47	−2.02	−1.79
1,1,1-trifluoro-2-propanol		−52	0.83	0.71	0.30	−0.48	−0.33
trifluralin		48	5.29	5.07	−5.68	−5.51	−5.02
1,1,3-trimethycyclopentane		−142	4.35		−4.48	−4.42	−3.85
trimethylamine		−117	0.02	0.16	0.84	1.20	0.48
tripropylamine		−94	3.19	2.79	−2.28	−1.64	−2.69
tubercidin	dec	247	−1.47		−1.95	−2.47	
undecanoic acid		28	4.57		−3.55	−3.43	−4.10
uracil		335	−1.06	−1.07	−1.48	−0.17	−1.54
urea		135	−1.66	−1.66	0.96	0.89	1.06
ursodeoxycholic acid		203	4.51		−4.29	−4.45	−5.79
vanillic acid		209	1.35	1.43	−2.05	−1.14	−2.69
warfarin		161	2.89	2.70	−3.89	−4.14	−3.75
2,6-xylenol		49	2.47	2.36	−1.31	−1.36	−2.21
3,4-xylenol		63	2.42	2.23	−1.41	−1.39	−2.30

[a] Missing: melting point is missing; dec: decomposed; calc: log S_{obs} is calculated value; corr: aqueous solubility has been corrected; MlogP: MlogP is used in GSE to calculate aqueous solubility.

by Huuskonen[7] do not correspond to those of the original literature source. The above are designated by a "MlogP" and "corr", respectively, in the comment column.

RESULTS AND DISCUSSION

Comparison of the Models. This manuscript compares the general solubility equation (GSE) of Jain and Yalkowsky[5] with four different methods to test 21 compounds described by Huuskonen and to a larger test set of 413 compounds for which Huuskonen only applied his ANN calculation. Among the criteria that can be used to select one model over another are fit, applicability, parsimony, convenience, and elegance. The general solubility equation will be compared to the four models described above on the basis of each of these criteria.

Fit. The most commonly used criterion for the evaluation of a scientific model is the goodness of its fit to the available data. The difference between the observed and predicted values for each data point is defined as the error for that point. Average absolute error (AAE) and root-mean-square error (RMSE) are generally used to evaluate fit.

Various statistical techniques (such as regression, neural networking, Monte Carlo simulations, or principal component analysis) generate coefficients for input parameters that minimize these measures of error. The data used to validate the model (i.e., the test set) must be independent of the data used to generate the model (i.e., the training set). In general, increasing the number of input parameters will reduce errors and improve fit. However, it reduces the precision of each of the coefficients generated.

Table 4. Comparison of Predictions for 21 Drug Test Set

	Huuskonen		Klopman	Kühne	Jain and Yalkowsky
	ANN	MLR	MLR	MLR	GSE
N	21	21	19	19	19
AAE	0.51	0.74	0.78	0.88	0.53
RMSE	0.72	1.01	1	1.09	0.72

Table 4 contains the number of compounds for which each of the five calculations calculation is applicable, along with the corresponding AAE and RMSE values. The error calculations based upon the original experimental data reported by Yalkowsky.[2] Since Huuskonen[7] used different experimental solubility values the errors for his ANN calculation are somewhat lower than those in the table.

It is clear that the artificial neural network (ANN) and the general solubility equation (GSE) give the best fit to the small set of drug solubility data. Both the average absolute error and the root-mean-square error average are about 0.3 log units higher for the multilinear regression (MLR) models of Huuskonen, Klopman, and Kühne. This corresponds to about a factor of 2.0 in the accuracy of the predictions.

Table 5. Comparison of Predictions for the 413 Compound Test Set

	Huuskonen	Jain and Yalkowsky
	ANN	GSE
N	380	380
AAE	0.431	0.546
RMSE	0.569	0.764

The average absolute errors and root-mean square errors for the larger test set calculated by the ANN method of Huuskonen[7] and by the GSE of Jain and Yalkowsky[5] are given in Table 5. Based upon Huuskonen's data set the ANN model gives predictions that are about 21% more accurate than those of the GSE. Furthermore, it is applicable to compounds for which melting point data are not available. However, it is clear that, despite its limited input and lack of training set data, the GSE gives reasonable predictions for the compounds.

It is worth noting that in cases for which GSE and ANN methods give similar predictions, they are always in good agreement with the experimentally determined values. Thus, if the two very different methods are in agreement with each other, there is a high degree of certainty that the calculated values are close to the true value.

Applicability. Not all of the calculations could be performed on every compound. The applicability of all MLR methods to a particular compound or test set is based solely upon the availability of data for the parameters corresponding to its constituent groups. This is dependent upon the size and quality of the training set. Similarly, the applicability of Huuskonen's ANN model is dependent upon the size, accuracy, and structural diversity of its training set.

The applicability of the GSE is dependent upon the availability of melting point and partition coefficient data. It cannot be used without values for both of these parameters. Fortunately, melting point data are available for most solid compounds that do not decompose before melting. Also, because liquid solutes require no crystal term, they are assigned a melting point value of 25 °C so that the term (MP-25) in eq 1 is equal to zero. Therefore, actual melting point values are not needed for liquids. If the compound decomposes before melting, the GSE gives a maximum solubility prediction. Partition coefficients can be readily calculated by CLOGP for most compounds. However, the GSE will perpetuate errors generated by CLOGP unless experimental values are available.

As can be seen from Tables 4 and 5 the GSE is applicable to 19 of the 21 drugs in the small test set and to 380 of the 413 compounds chosen by Huuskonen[7] for the large test set. All of the large test set compounds can be treated by the ANN method since they were chosen to be compatible with their training set.

Parsimony. Just as there are an infinite number of curves that can be drawn through any set of points, there are an infinite number of theories that can explain any set of data. Occum's razor[11] states "It is vain to do with more what can be done with less". This is the basis for the principle of parsimony, or economy of modeling, i.e., the principle that the smaller the number of assumptions, parameters, or mathematical steps, the better the model.

All three MLR models utilize a large number of group contribution parameters that are evaluated on the basis of the test set data. The ANN model is effectively a multiparameter equation that is too complex to be described in a publication. It is essentially a polynomial in exponential terms. The GSE on the other hand utilizes only two parameters. In the present study, the ANN model was trained on a set of 884 compounds, whereas the GSE uses neither a training set nor computer modeling. Instead it uses two well-defined parameters: melting point and octanol−water partition coefficient.

Convenience. An important determinant of the success of a property estimation method is its convenience and user friendliness. A convenient, user-friendly method requires a minimum of user training and can ideally be without the aid of a pencil, let alone a computer. This is one of the major advantages of the general solubility equation. If the melting point and partition coefficient of the solute are known, as is often the case, the estimation of its aqueous solubility by the GSE is a trivial task. This is clearly not the case for the MLR or ANN methods, which are of little value if the user is far from a computer.

Elegance. The elegance of a model or theory is often the key factor in determining its acceptability by the scientific community. An elegant model is pleasing to the mind. It, and each assumption, upon which it is based, must be intuitive as well as consistent with both experience and with other accepted theories. Ideally, a model should be mathematically derived from generally accepted basic principles. This enables the model to be validated on all available data because there is no need to have data for a separate training set and test set. Although elegance is somewhat subjective, it should not be ignored as a criterion for model selection.

Each of the MLR models as well as the ANN model utilizes a large number of regression generated coefficients that are based upon data from a training set. Adding or deleting a few compounds to the training set could change the coefficients and thus the predicted values for the test set compounds. There is no independently verifiable or theoretical relationship between any of the parameters and solubility. The only relationship between each input parameter and the dependent variable is the one generated by the computer. The GSE model on the other hand does not rely upon a training set and thus contains no computer fitted parameters; it relies only upon two experimentally determined values for the test set. The final equation was derived from basic principles and a few intuitively acceptable assumptions. These assumptions are described by Jain and Yalkowsky.[5]

Since both MP and logP are dependent upon structure they can, in principle, be predicted from structure. In fact the logP values used in the calculations were generated through the use of the software program, CLOGP. Unfortunately, a reliable means of calculating melting point from structure is not yet available, so that only experimental values are used. Efforts are currently aimed at improving the accuracy of melting point prediction. Note that since an experimental melting point is required for the GSE approach, it cannot be used to calculate accurate solubility for compounds that decompose before melting.

OTHER MODELS

In a previous report, Ran and Yalkowsky[6] showed that the general solubility equation gives better prediction of the

solubility of 150 compounds than a Monte Carlo simulation proposed by Jorgenson and Duffy.[12] The solvatochromatic, mobile order, and UNIFAC models each offer useful insight into the complexity of the solubilization process.

SUMMARY

The revised GSE was derived from basic principles and the following assumptions as summarized above and described by Jain and Yalkowsky:[5] The constant 0.5 is based upon regular solution theory and the complete miscibility of organic compounds with octanol. The logP term is based upon the assumptions that the partition coefficient is equal to the solubility ratio and independent of concentration. The MP-25 term is based upon the validities of both the van't Hoff equation and Walden's rule for the entropy of melting.

The GSE provides reasonably accurate predictions of the solubilities of a wide variety of organic compounds with a minimum of input data. Because it requires no training set and includes no fitted parameters, the GSE can be used without modification for all nonionized organic compounds. The ANN and MLR methods, on the other hand, require large training sets, the selection of a group of parameters that are related to the structure of the molecule, and the fitting of those parameters to the experimental data. However, if the parameters are well selected the computationally generated parameters the ANN and MLR methods can be more accurate than any fixed value calculation.

This study supports the use of the GSE as an independent means of estimating aqueous solubility. Because of its simplicity and accuracy and because it is different from most other methods, it should be considered, either alone or in combination with a more complex method, for all compounds that have melting point data available.

REFERENCES AND NOTES

(1) Yalkowsky, S. H.; Valvani, S. C. Solubility and Partitioning I: Solubility of Nonelectrolytes in Water. *I. Pharm Sci.* **1980**, *69*, 912−922.
(2) Yalkowsky, S. H. Estimation of the Aqueous Solubility of Complex Organic Compounds. *Chemosphere* **1993**, *26*, 1239−1261.
(3) Yalkowsky, S. H. *Solubility and solubilization in Aqueous Media*; Oxford University: Oxford, 1999.
(4) Yalkowsky, S. H.; Sinkula, A. A.; Valvani, S. C. *Physical Chemical Properties of drugs*; Marcel Dekker: New York, 1980.
(5) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I:Application to Organic Non-Electrolytes. *J. Pharm. Sci.* **2000**, *90*, 234−252.
(6) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.
(7) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(8) Hansch, C.; Quninlan J. E.; Lawrence, G. L. The Linear Free-Energy Relationshipbetween Partition Coefficient and the Aqueous Solubility of Organic Liquids. *J. Org. Chem*. **1968**, *33*, 347−350.
(9) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility ofOrganic Molecules by the Group Contribution approach. Application to the Study of Biodegredation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(10) Kühne, R.; Ebert, R.-U.; Kleint, F.; Scmidt, G.; Schüürmann, G. Group ContributionMethods to Estimate water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.
(11) Occum, R. B. *Barletts Quatations* 1300−1349.
(12) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from MonteCarlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

CI010287Z