Fast Approaches for Molecular Polarizability Calculations

Junmei Wang,* Xiang-Qun Xie,† Tingjun Hou,‡ and Xiaojie Xu§

Encysive Pharmaceuticals Inc., 7000 Fannin Street, Houston, Texas 77030, Pharmacological Sciences, College of Pharmacy, University of Houston, 4800 Calhoun Road, Houston, Texas 77204, Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92093, and College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, People's Republic of China

Received: December 7, 2006; In Final Form: March 9, 2007

Molecular polarizability of a molecule characterizes the capability of its electronic system to be distorted by the external field, and it plays an important role in modeling many molecular properties and biological activities. In this paper, a set of fast empirical models have been developed to predict molecular polarizability using two types of approaches. The first type of approaches is based on Slater's rules of calculating the effective atomic nuclear shielding constants. The best model (model 1A) of this category has achieved an average unsigned error (AUE), root-mean square error (RMSE), and average percent error (APE) of 2.23 au, 3.29 au, and 2.77%, respectively. The second type of model is based on an additive hypothesis of molecular polarizability. Five models have been constructed using different schemes of atom types. The best model that applies 14 atom types, model 2e, achieves AUE, RMSE, and APE of 0.99 au, 1.48 au, and 1.24%, respectively. This performance is much better than those of the models purely based upon chemical composition (model 2A and the Bosque and Sales model), for which errors are about 2-fold higher. It is expected that both model 1A and model 2E will have broad applications in QSAR and QSPR studies.

1. Introduction

The polarizability of an atom or molecule describes the response of its electron cloud to an external field. The polarization energy due to an external electric field E is proportional to E^2 for external fields that are weak compared to the internal electric fields between its nucleus and electron cloud. Technically, polarizability is a tensor quantity, but for spherically symmetric charge distributions it reduces to a single number. In many cases, an average polarizability is usually adequate in calculations. Polarizability appears in many formulas for low-energy processes involving the valence electrons of atoms or molecules. It is also widely used to describe the inductive and dispersive interactions of a molecule or molecular system.

Polarizability has been extensively applied in drug design. It is one of the descriptors that are extensively used in QSPR and QSAR studies. For example, we found that polarizability was obviously correlated to the logarithm of the *n*-octanol/water partition coefficient, log P, for a data set of 1904 molecules ($r^2 = 0.21$);[1] the correlation coefficient square between polarizability and aqueous solubility was 0.44 for a data set of 1708 molecules.[2] Polarizability has been successfully applied by many researchers in constructing QSPR models for many molecular properties, including Henry's law constant,[3] aqueous solubility,[4,5] subcooled liquid vapor pressures,[6] the partition coefficient of vaporous chemicals in a water-gas phase,[5] vapor pressure,[7] heat of vaporization,[7] diffusion coefficient,[7] etc. In a recent report, Verma, Kurup, and Hansch successfully applied polarizability in QSAR studies of 51 chemical−biological interactions.[8] In their studies, polarizability was simply calculated by adding up the number of valence electrons (*NVE*) in a molecule: $H = 1$, $C = 4$, $N = 5$, $P = 5$, $O = 6$, $S = 6$ and halogens $= 7$. The general form of those QSARs is illustrated by

$$\log 1/C = k \times NVE + c \tag{1}$$

In this equation, C is IC_{50} or EC_{50} or the molar contribution of a compound, k is a weight obtained by regression analysis, and c is a constant.

Even though the experimental polarizability is mostly determined by accurate measurements of a dielectric constant or refractive index (0.5% or better), one should treat many of the results with some caution if the data are obsolete and when the results are referred to optical frequencies instead of static ones.

In quantum mechanics, polarizability may be calculated by solving the coupled perturbed Hartree−Fock (CPHF) equations with electric field perturbations. In molecular mechanics, polarization is typically calculated with an atomic dipole interaction model, and atomic polarizabilities are the key parameters in those non-additive calculations. In a dipole interaction model, a molecule's polarizability is the trace of the inversion matrix of a $3N \times 3N$ matrix $\mathbf{R}$, which has diagonal elements being $1/\alpha$ and off-diagonal elements being the dipole field tensor T_{pq}, a function of distance between atoms p and q. The $3N \times 3N$ atomic representation of polarizability can be reduced to a normal 3×3 molecular representation tensor. More details are provided in our recent work[9] on atomic polarizability parametrization for three widely used dipole interaction models, namely, the Applequist,[10] the Thole exponential, and the Thole linear models.[11,12] In the following text, we give a brief review of some fast empirical approaches for estimating static molecular polarizabilities.

* To whom correspondence should be addressed at Encysive Pharmaceuticals Inc. Telephone: 713-578-6649. E-mail: jwang@encysive.com.
† University of Houston.
‡ Peking University.
§ University of California at San Diego.

10.1021/jp068423w CCC: $37.00
Published on Web 04/27/2007

Besides the above two kinds of methods, a set of empirical algorithms have been developed to predict the average molecular polarizability efficiently. Most empirical models are based on a hypothesis that molecular polarizability is additive. In another word, the polarizability of a molecule can be obtained by summing up the contributions of a variety of atoms and/or functional groups in the molecule. Here is the reasoning: molar refraction (R) is found to be an additive property, molecular polarizability α is related to R by the Lorentz−Lorenz equation (eq 2), and α should therefore be an additive property. n is the

$$R = \left(\frac{n^2 - 1}{n^2 + 2}\right)\frac{M}{\rho} = \frac{4}{3}\pi N_0 \alpha \tag{2}$$

refractive index of a molecule (usually at 5893 Å, sodium D-line), and M and ρ are molecular weight and molar volume, respectively. N_0 is the Avogadro constant. The additive hypothesis has been extensively used by many researchers in fast calculation of average molecular polarizabilites. Bosque and Sales recently developed a model to calculate polarizabilities of 426 molecules from the chemical composition.[13] Although only 10 descriptors were used, the model achieved a good predictability (the average percent errors were 2.31% and 1.93% for the training set and test set, respectively). Stout and Dykstra extended the additive hypothesis to calculate tensor components of *xx*, *yy*, and *zz* by adding up the individual contribution of each atom.[14] The atomic polarizabilities of 13 atom types that covered C, N, O, and F were derived from a high-level ab initio calculations for over 30 organic molecules containing up to four non-hydrogen atoms. The average errors were around 10% and 3% for individual tensor elements and isotropic polarizability, respectively. Miller and Savchik proposed another empirical approach for calculating average molecular polarizability based on atomic hybrid components, τ_A(ahc), or by atomic hybrid polarizabilities, α_A(ahp).[15−16] They defined a set of 20 τ_A(ahc) and α_A(ahp) parameters based on atomic hybridizations for ten elements (C, H, O, N, S, P, and halogens). The average molecular polarizability relates to the two types of parameters by eqs 3 and 4, where N is the total number of electrons. Their models based on atomic hybrid components and atomic hybrid polarizabilities achieved average percent errors of 2.2% and 2.8% for about 400 compounds, respectively. It should be pointed out that the theoretic basis of Verma, Kurup, and Hansch's approach to estimate polarizability with NVE is also the additive property of molecular polarizability.

$$\alpha(\text{ahc}) = \frac{4}{N}\left(\sum_A \tau_A(\text{ahc})\right)^2 \tag{3}$$

$$\alpha(\text{ahp}) = \sum_A \alpha_A(\text{ahp}) \tag{4}$$

Besides those additive models, Glen provided an alternative way to calculate molecular polarizability emprically.[17] His model was based on Slater's rule for the calculation of effective atomic nuclear shielding constants. More details of the algorithm are presented in the Methods section (2.2.1).

In development of a set of atomic dipole interaction models with the Applequist and Thole schemes,[9] we found that the model performance could be significantly improved by using 14 atom types instead of 10 elements in parametrization. Inspired by this, we plan to develop a new empirical model according to eq 5, where N is the number of atom types, n_i is the number occurrence of atom type i in a molecule, and c_i is the weight of atom type i. The theoretical basis beyond this model is the additive characteristic of polarizability.

$$\alpha = \sum_{i=1}^{N} n_i c_i \tag{5}$$

Second, we will take advantage of a set of high-quality data used by Bosque and Sales to reexamine the Glen's model. A genetic algorithm developed by ourselves will be applied to optimize the effective quantum numbers n*, adjustable parameters in Glen's method.

Finally, we also plan to perform quantum mechanical calculations at the B3LYP/6-31G* level to calculate the CPHF polarizabilities and to make a comparison to the aforementioned empirical models as well as to experimental findings.

2. Methods

2.1. Data Sources. The Bosque and Sales data set[13] was used to reexamine the Glen's model[17] and to develop empirical models from the chemical composition using eq 5. Six molecules were dropped off due to duplication or apparent errors/typos. The 420 left molecules are very diverse in structure and include a variety of functional groups, including hydrocarbons (aliphatic and aromatic, cyclic and acyclic), alcohols, phenols, ethers, esters, proxides, aldehydes, ketones, carboxylic acids, amines, imines, amides, nitriles, nitro derivatives, disulfides, thiophenes, sulfides, sulfoxides, sulfones, phosphates, halides, etc.

The experimental polarizability, α, was obtained from the measurements of refractive index R using the Lorentz−Lorenz equation (eq 2). The experimental refractive indexes were measured at 20 or 25 °C at the D-line of sodium wavelength (5893 Å). The whole data set was randomly divided into a training set (nos. 1−335 in Table S1) and a test set (nos. 336−420 in Table S1) by Bosque and Sales for model validation. However, all the models were generated with the whole data set except those indicated explicitly. The compound names, the SMILES strings as well as the experimental values are listed in Table S1 of the Supporting Information.

2.2. Model Construction. *2.2.1. Glen's Approach Reexamination.* Theoretically, polarizability may be expressed in terms of atomic radius (r) of maximum electron density.[17]

$$\alpha = \frac{4}{9}\sum_i (r_i^2)^2 \tag{6}$$

$$r_i^2 = \left(\frac{n_i^*}{2(Z - s_i)}\right)^2 (2n_i^* + 1)(2n_i^* + 2)a_0^2 \tag{7}$$

$$\alpha = \frac{4}{9}\sum_j N_j (r_j^2)^2 \tag{8}$$

where r_i is the atomic radius for electron i, $a_0 = 0.5292$ Å is the Bohr radius. n* is the effective quantum number, which is 1, 2, 3, 3.7, 4.0, and 4.2 for principal quantum numbers 1, 2, 3, 4, 5, and 6, respectively, Z is the nuclear charge, and s is a screening constant that is empirically determined by the following Slater rules:[18]

Rule 1: Electrons are divided into groups and each group has a different shielding constant: (1) 1s; (2) 2s, 2p; (3) 3s, 3p; (4) 3d; (5) 4s, 4p; (6) 4d; (7) 4f; (8) 5s, 5p; (9) 5d; etc.

Rule 2: Electrons in a higher group do not shield those in lower groups.

Rule 3: For s and p valence electrons, electrons in the highest group contribute 0.35 except the first group (1s), where 0.30 is used instead, electrons in the one lower group contribute 0.85, and electrons in other lower groups contribute 1.0.

Rule 4: For d and f valence electrons, electrons in the highest group contribute 0.35, and electrons in other lower groups contribute 1.0. The effective nuclear charge is $Z - s$. Equation 6 is simplified to eq 8 because electrons in the same electron group have the same radius r_j, where N_j is the number of electrons in the jth electron group.

The above scheme and parameters perform well in reproducing many properties for atoms that include atomic radii, diamagnetic susceptibilities, X-ray levels, and ionization potentials. However, it may not be used directly for molecules without modification because electron density distribution around atoms is significantly changed when a molecule is formed. Glen suggested that after reoptimizing the effective quantum number the calculated molecular polarizability according to eq 6 or 8 could reproduce the experimental values. Although Glen's model ($n^* = 0.94$, 1.88, 2.65, 3.15, and 3.35 for principal quantum numbers of 1, 2, 3, 4, and 5, respectively) could explain 96.07% of the variation and the standard error of prediction was 5.28 for 28 molecules in his training set, the model performed poorly for a 64-molecule test set, for which the average percent error was 12.5%. Because more than 10 years have passed, it is time to revise the model to incorporate newly emerged high-quality polarizability data using more promising optimization methods.

Genetic algorithm, an efficient heuristic optimization method, has been widely used to solve optimization problems such as conformational searches, molecular docking, and QSPR model generation.[19−24] The power of GA lies in its abilities to efficiently deal with multiple dimension problems, no matter whether the variables are coupled or not. Therefore, GA should also be suitable to optimize the effective quantum number parameters in Glen's model. The following is a brief explanation on how a genetic algorithm works. A target function is minimized by a genetic algorithm through three basic operations that mimic natural evolution and selection, which are mutation, crossover, and selection. First of all, a set of "chromosomes", which encode answers to a question, are randomly generated. The "genes" in a "chromosome" correspond to the descriptors in question. For each "chromosome", the fitness is evaluated by a scoring function. The higher the score, the better the fitness and the closer to the real answer it is. New "chromosomes" are then generated through swapping certain "genes" between multiple "chromosomes" and mutating "genes" to other values. In the subsequent selection operation, "chromosomes" with high fitness are evolved to the next generation and those having low fitness are allowed to perish. The three operations are iteratively performed until a termination criterion is met.

A real number-encoded genetic algorithm program developed by ourselves was applied to optimize the effective quantum number parameters to minimize the average percent error of the 420 molecular polarizabilities.

Two models based on the Glen approach[17] were investigated. In model 1A, five effective quantum numbers for $n = 1$, 2, 3, 4, and 5 were optimized; in model 1B, we used more than one effective quantum number for different elements even for the same quantum number. In total, 15 parameters were optimized. We intended to find out if the fitting performance could be improved by using more parameters. It should be pointed out that in model 1B, the effective quantum numbers somehow lose their physical meaning and become pure variables. The nuclear charges Z, the screening factors s, the number of electrons in each group N, and the effective quantum numbers n^* of ten elements (H, C, N, O, F, P, S, Cl, Br, I), are listed in Table 1. The Gasteiger−Marsili charges were assigned for the 420 molecules using Sybyl7.0[25] and the numbers of electrons in the highest groups were adjusted accordingly. Take methanol as an example, the point charge of oxygen is −0.398; therefore, the $2s^22p^4$ group has 6.398 electrons, and the $2s^22p^2$ group of carbon has 3.967 electrons because its charge is 0.033. To investigate the reliability of the models, model 1A_validation was generated by using the 335 molecules in the training set. All the other settings of model 1A_validation were as the same as those of model 1A.

TABLE 1: List of Electron Groups, Nuclear Charge (Z), Number of Electrons in Echo Charge Group (N), Screening Factor (s), Effective Quantum Number (n^*) and Parameter Names of Both Models 1A and 1B for Ten Elements

element	electron group	Z	N	s	n^*	parameter (model 1A)	parameter (model 1B)
H	$1s^1$	1	1	0	1	n1	n1
C	$1s^2$	6	2	0.3	1	n1	n2
C	$2s^22p^2$	6	4	2.69	2	n2	n3
N	$1s^2$	7	2	0.3	1	n1	n4
N	$2s^22p^3$	7	5	3.04	2	n2	n5
O	$1s^2$	8	2	0.3	1	n1	n6
O	$2s^22p^4$	8	6	3.39	2	n2	n7
F	$1s^2$	9	2	0.3	1	n1	n8
F	$2s^22p^5$	9	7	3.74	2	n2	n9
P	$1s^2$	15	2	0.3	1	n1	n10
P	$2s^22p^6$	15	8	4.09	2	n2	n11
P	$3s^23p^3$	15	5	10.2	3	n3	n12
S	$1s^2$	16	2	0.3	1	n1	n10
S	$2s^22p^6$	16	8	4.09	2	n2	n11
S	$3s^23p^4$	16	6	10.55	3	n3	n12
C1	$1s^2$	17	2	0.3	1	n1	n8
C1	$2s^22p^6$	17	8	4.09	2	n2	n9
C1	$3s^23p^5$	17	7	10.9	3	n3	n13
Br	$1s^2$	35	2	0.3	1	n1	n8
Br	$2s^22p^6$	35	8	4.09	2	n2	n9
Br	$3s^23p^6$	35	8	11.25	3	n3	n13
Br	$3d^{10}$	35	10	19.95	3	n3	n13
Br	$4s^24p^5$	35	7	30.1	4	n4	n14
I	$1s^2$	53	2	0.3	1	n1	n8
I	$2s^22p^6$	53	8	4.09	2	n2	n9
I	$3s^23p^6$	53	8	11.25	3	n3	n13
I	$3d^{10}$	53	10	19.95	3	n3	n13
I	$4s^24p^6$	53	8	30.45	4	n4	n14
I	$4f^{14}$	53	14	38.31	4	n4	n14
I	$5s^25p^5$	53	7	48.1	5	n5	n15

Important parameters that controlled the GA performance are listed as follows: (1) *Population Size*: the number of chromosomes in one generation (100). (2) *Chromosome Size*: the number of variables in question (5 for model 1A and 15 for model 1B). (3) *Elite Size*: the number of "Elite" chromosomes, which entered the next generation directly (5). (4) *Mutation Probability*: the probability of performing mutation on each gene of each chromosome (0.05). (5) *Crossover Probability*: the probability of performing crossover on each "chromosome" in a population (0.40). (6) *Selection Methods:* tournament. (7) *Tournament Number*: number of "chromosomes" that participate in selection each time (3). (8) *Maximum Iteration:* the Maximum iteration of optimization (50 000). GA optimizations were performed at least four times and the best parameter set was selected as the final models.

2.2.2. Static Molecular Polarizability Models Based on Summation of Atomic Polarizabilities. As mentioned in the Introduction, Bosque and Sales developed a model by summing up the contribution of each element.[13] In this work, we planned to further improve their model by introducing several more atom

TABLE 2: List of the Performance of Molecular Polarizabilities Prediction by a Set of Eleven Models for the Bosque and Sales' Data Set (Ref 13)[a]

models[b]	AUE (au)	RMSE (au)	APE (%)
M1A	2.23	3.29	2.77
M1AV	2.25	3.31	2.78
M1AV for training set	2.22	3.29	2.81
M1AV for test set	2.37	3.42	2.66
M1AP	26.04	32.01	31.65
M1B	1.95	2.88	2.46
M2A	1.68	2.29	2.20
M2AP	1.68	2.29	2.21
M2B	1.32	1.96	1.72
M2C	1.09	1.59	1.42
M2D	1.08	1.56	1.38
M2E	0.99	1.48	1.24
M2EV	1.04	1.52	1.30
M2EV for training set	1.01	1.52	1.31
M2EV for test set	1.13	1.51	1.29

[a] The whole data set is divided into training set (no. 1-335) and test set (no. 336-420). AUE, RMSE, and APE are average unsigned error, root-mean-square error, and average percent error, respectively. [b] M1A = model 1A; M1AV = model 1A validation; M1B = model 1B; M2A = model 2A; M2AP = model 2A paper; M2B = model 2B; M2C = model 2C; M2D = model 2D; M2E = model 2E; M2EV = model 2E validation.

types. In model 2A, ten atom types, that is to say, each element has only one atom type, are applied. This model was slightly different from the Bosque and Sales model because six molecules were dropped off in our data set.

Inspired by the significant performance improvement by introducing three atom types for sp^1, sp^2, and sp^3 carbons in our work of atomic polarizability parametrization for a set of dipole interaction models,[9] we decided to allow the three hybridized carbons to have different coefficients in regression analysis. This resulted in model 2B.

On the basis of model 2B, we analyzed the error sources of the prediction and tried to introduce more atom types to reduce the prediction error. More models, model 2C, model 2D, ..., were introduced by this way. We want to emphasize again that an atom type was introduced only when it could significantly deduce the prediction errors. For the best model, we would generate a corresponding validation model only using the training set and make predictions for the test set.

There are many approaches to build up correlations between a molecular property and its descriptors, which include multiple linear regression (MLR), partial least-squares (PLS) fitting, artificial neural networks (ANN), genetic algorithm (GA), etc. On the basis of the fact that molecular polarizability is additive, MLR was applied to build up those models (models 1A, 2A, 2B, 2C, and 2D).

Ab initio optimizations at the B3LYP/6-31G* level were performed for all the 420 molecules with the Jaguar package of Schrodinger LLC.[26] Molecular polarizability was calculated by solving the coupled perturbed Hartree−Fock (CPHF) equations with electric field perturbations. The ab initio polarizability was compared not only to the experimental value but also to that predicted by empirical models.

3. Results and Discussion

3.1. Models Based on Glen's Approach. A genetic algorithm was applied to optimize the effective quantum numbers, n^*, and to make the calculated molecular polarizabilities by eqs 6−8 reproduce the experimental values. For model 1A, GA was performed five times and very similar results were obtained (APE = 2.7736, 2.7742, 2.7739, 2.7735, 2.7734 for five runs),

TABLE 3: List of Effective Quantum Number Parameters for the Four Models Based on Glen's Scheme

parameter	model 1A	model 1A validation[a]	model 1A paper[b]	model 1B
n1	0.90488	0.90830	0.94000	0.88643
n2	2.01297	2.01180	1.88000	2.81320
n3	2.81108	2.80943	2.65000	1.79504
n4	2.62613	2.62596	3.15000	1.28852
n5	2.81449	2.81454	3.35000	2.07080
n6				3.25160
n7				0.65683
n8				2.77171
n9				1.72830
n10				2.80851
n11				2.62696
n12				2.81902
n13				4.26174
n14				3.73676
n15				2.44175

[a] Same to model 1A except being constructed only with the 335 training set molecules [b] Glen's model (ref 17).

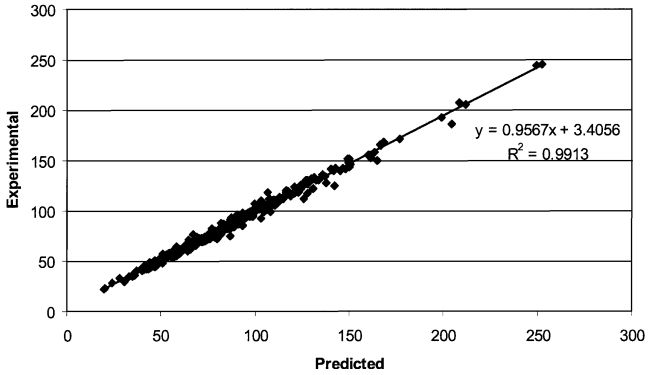

Figure 1. Plot of calculated versus experimental polarizability (au) of the 420-molecule set using model 1A with the corresponding regression equation.

indicating that the GA settings were sufficient to lead GA to converge. The fifth model was adopted as the last model. The calculated average unsigned error, root-mean-square error and average percent error were 2.23 au, 3.29 au, and 2.77%, respectively.

By having more than one n^* parameter for a quantum number n, model 1B was expected to have a better performance. It is true that the errors of model 1B were marginally smaller, which were 1.95 au, 2.88 au, and 2.46% for AUE, RMSE, and APE, respectively. However, the slightly better performance could not compensate the use of 10 more parameters. Therefore, we think model 1A is a better model in practice. The molecular polarizabilities of the 420 molecules predicted by both models are listed in Table S2 of the Supporting Information. The AUE, RMSE, and APE are summarized in Table 2. The effective quantum number parameters of both models are listed in Table 3. The plot of experimental versus calculated polarizabilities predicted by model 1A for the whole data set is shown in Figure 1. Although the performance of model 1A is slightly worse than that of the Bosque and Sales model that had an APE of 2.23%,[13] model 1A applies only half of the descriptors of the Bosque and Sales model. Therefore, we believe model 1A is more reliable and may perform better for novel compounds not covered by the Bosque and Sales model.

To test the reliability of model 1A, model 1A_validation was constructed just using the training set molecules and then applied to predict molecular polarizabilities for the molecules in the

TABLE 4: List of Atomic Polarizability Coefficients for the Seven Models Based on the Additive Hypothesis of Steric Molecular Polarizability

elem	atom type	model 2A	model 2B	model 2C	model 2D	model 2E	model 2E validation[a]	model 2A paper[b]
C	C1 (sp^1)	10.175	10.768	10.079	10.257	10.152	10.253	10.19
C	C2 (sp^2)	10.175	8.803	8.757	8.76	8.765	8.865	10.19
C	C3 (sp^3)	10.175	5.594	5.679	5.669	5.702	5.814	10.19
H		1.177	3.391	3.386	3.402	3.391	3.365	1.174
F		1.447	3.742	3.383	3.847	3.833	3.794	1.498
CI		14.643	16.068	16.473	16.417	16.557	16.394	14.576
Br		22.303	24.171	24.367	24.626	24.123	24.693	22.202
I		36.692	38.554	38.896	39.065	38.506	39.409	36.778
N	NO N in nitro	7.093	5.698	6.341	5.917	10.488	11.55	6.951
N	NA sp2 N with three bonded atoms	7.093	5.698	6.341	7.903	6.335	6.162	6.951
N	N	7.093	5.698	6.341	5.917	6.335	6.162	6.951
O		3.829	4.2	4.482	4.432	4.307	4.259	3.853
S	SO S in sulfone	19.757	20.265	15.515	15.385	15.726	16.109	20.178
S	S	19.757	20.265	22.481	22.326	22.366	22.351	
P		15.892	14.89	10.968	11.509	11.173	10.813	16.736
constant		2.252	−0.837	−1.46	−1.548	−1.529	−1.758	2.146

[a] Same as model 2E except being constructed only with the 335 training set molecules. [b] Bosque and Sale's model (ref 13).

TABLE 5: List of the Performance of the Six Models Based on the Additive Hypothesis of Steric Molecular Polarizability (in au)

	model 2A	model 2B	model 2C	model 2D	model 2E	model 2E validation
			Leave-One-Out Analysis			
components	6	4	5	5	6	5
standard error	2.42	2.13	1.68	1.71	1.57	1.68
q^2	0.994	0.996	0.997	0.997	0.998	0.997
			Full Component Analysis			
standard error	2.29	1.96	1.59	1.56	1.48	1.52
r^2	0.995	0.996	0.997	0.997	0.998	0.998
n1	6	4	5	5	6	5
n2	413	415	414	414	413	329
F	13092.7	26770.2	32647.6	32163.6	31466.1	26497.5
prob of $r^2 = 0$	0	0	0	0	0	0

test set. The results were very encouraging: the AUE, RMSE, and APE were 2.22 au, 3.29 au, and 2.81% for the 335 molecules in the training set, respectively; and 2.37 au, 3.42 au, and 2.66% for the 85 molecules in the test set. Interestingly, although the AUE and RMSE of the test set were marginally higher, the APE was not.

We also tested how Glen's parameter set performed in predicting molecular polarizability for the 420-molecule data set. The three errors were much larger than that of both model 1A and model 1B, which were 26.04 au, 32.01 au, and 31.65% for AUE, RMSE, and APE, respectively. It should be pointed out that it is somewhat unfair to make such comparisons, considering Glen's data set, which probably included many low-quality data, was quite different from the one we used.

3.2. Models Based on the Summation of Atomic Polarizabilities. The theoretical basis of models 2A, 2B, 2C, 2D, and 2E is the additive property of polarizability. Molecular polarizability is calculated with eq 5. The atomic polarizabilities, c_i in this equation, were obtained by regression analysis. Same as the Bosque and Sales' model, model 2A applied ten descriptors (each element has one atom type) to create a model. The performance (AUE = 1.68 au, RMSE = 2.29 au, APE = 2.20%) was only slightly different from that of the Bosque and Sales model due to the exclusion of six compounds. In model 2B, sp^1, sp^2, and sp^3 carbons were allowed to take different parameters and the performance of the fitting was significantly improved (AUE = 1.32 au, RMSE = 1.96 au, APE = 1.72%).

We examined the error sources of model 2B and found that many sulfur-containing compounds had large prediction errors. Therefore, we decided to make the sulfur in sulfone functional group be differentiated from the other kinds of sulfurs. This resulted in another model, model 2C. The performance of model 2C was further improved and AUE, RMSE, and APE were 1.09 au, 1.59 au, and 1.42%, respectively.

For model 2C, we found that some nitrogen-containing compounds, especially nitro derivatives had large prediction errors. Two schemes were designed to introduce more atom types. In the first scheme, the discrimination of three atoms bonded sp^2 nitrogen from the other nitrogen atoms leads to model 2D; in the second scheme, nitrogen in the nitro functional group was separated from the other nitrogen atoms (model 2E). The performance of model 2D (AUE = 1.08 au, RMSE = 1.56 au, APE = 1.38%) was only marginally better than that of model 2C, indicating that this atom type scheme was not effective. In contrast, model 2E achieved a much better performance, which had 0.99 au, 1.48 au, and 1.24% for AUE, RMSE, and APE, respectively. This was a very encouraging result because both AUE and APE of model 2E are reduced about 50% compared to those of models based on chemical composition (model 2A and the Bosque and Sales model), with a small price of adding four more atom types.

Based on model 2E, a validation model (model 2E_validation) was generated by fitting atomic polarizability parameters only using the 335-molecule training set. The AUE, RMSE, and APE for the training set molecules were 1.01 au, 1.52 au, and 1.31%, respectively, and the corresponding errors were 1.13 au, 1.51 au, and 1.29% for the 85 molecules in the test set. The quite

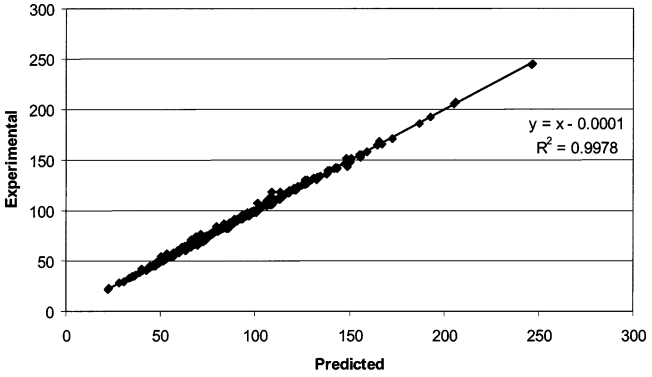

Figure 2. Plot of calculated versus experimental polarizability (au) of the 420-molecule set using model 2E with the corresponding regression equation.

similar prediction performance of both the training and test sets implied that this series of models was reliable.

The molecular polarizabilities of the 420 molecules calculated by all the six models (models 2A, 2B, 2C, 2D, 2E, and 2E_validation) are listed in Table S2 of the Supporting Information and the AUE, RMSE, and APE are summarized in Table 2. The atomic polarizability parameters of those models are listed in Table 4. The plot of experimental versus calculated polarizabilities predicted by model 2E, the best model, for the whole data set is shown in Figure 2.

Quantum mechanical molecular polarizabilities of the 420 molecules, obtained by solving the CPHF equations, are listed in Table S1. It is clear that the B3LYP/6-31G* polarizabilities are systematically smaller than those experimental values. The AUE, APE, and RMSE were 16.03 au, 16.67 au, and 19.97%, respectively. A very good correlation between QM and experimental polarizabilities was identified ($R^2 = 0.9901$, $\alpha_{qm}^{corr} = 1.1176\alpha_{qm} + 8.0311$). After corrections with this linear regression model, the AUE, RMSE, and APE now are 2.11 au, 31.5 au, and 2.95%, respectively. It is quite obvious that the performance is still inferior to those of the empirical models developed in this work.

We also calculated molecular refraction for the 420 molecules with the CMR program implemented in Sybyl7.0.[25] Interestingly, CMR correlated very well to the experimental polarizability ($R^2 = 0.997$, RMSE = 1.708, $F = 143\,462.113$). It was not a surprising result at all, given the relationship between molecular polarizability and molecular refraction revealed by the Lorentz−Lorenz equation (eq 2).

4. Conclusions

In this work, taking the advantage of a high-quality 420-molecule data set, we developed a set of empirical models to predict molecules' static polarizabilities. The best model based on Glen's approach is model 1A, which has AUE, RMSE, and APE of 2.23 au, 3.29 au, and 2.77%, respectively. Although its performance is slightly poorer than the Bosque and Sales model based on chemical composition, model 1A is attractive because it only applies five effective quantum numbers as descriptors. GA has once again been proved to be able to handle some nonlinear problems like this one.

Five other models rooted on the additive hypothesis of molecular polarizability were developed through linear regression. It is encouraging that by adding four more atom types, model 2E achieves a much better performance than model 2A, which is purely based on chemical composition (each element has only one atom type), indicated by AUE, RMSE, and APE of 0.99 au, 1.48 au, and 1.24%, respectively. We believe that both model 1A and model 2E will have great applications on QSAR and QSPR studies.

Acknowledgment. We are grateful to acknowledge the research support from NCSA (MCB000013N (J.W.)).

Supporting Information Available: The compound names, the SMILES as well as the experimental values of the 420-molecule data set are listed in Table S1. The calculated molecular polarizabilities of 420 molecules by the 11 models are listed in Table S2. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible surface area. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058−1067.

(2) Wang, J. M.; Krudy, G.; Hou, T. J.; Holland, G.; Xu, X. J. Development of reliable aqueous solubility models and their application in drug-like analysis. *J. Chem. Inf. Mod.* submitted for publication.

(3) Dearaden, J. C.; Schüürmann, G. Quantitative structure-property relationships for predicting Henry's law constant form molecular structure. *Environ. Toxicol. Chem.* **2003**, *22*, 1755−1770.

(4) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.

(5) Raevsky, O.; Andreeva, E.; Raevskaja, O.; Skvortsov, V.; Schaper, K. QSAR analysis of the partitioning of vaporous chemicals in a water-gas phase system and the water solubility of liquid and solid chemicals on the basis of fragment and physicochemical similarity and descriptors. *SAR QSAR Environ. Res.* **2005**, *16*, 191−202.

(6) Yang, P.; Chen, J.; Chen, S.; Yuan, X.; Schramm, K. W.; Kettrup, A. QSPR models for physicochemical properties of polychlorinated diphenyl ethers. *Sci. Total Environ.* **2003**, *305*, 65−76.

(7) Hilal, S. H.; Karickhoff, S. W. Prediction of the vapor pressure boiling point, heat of vaporization and diffusion coefficient of organic molecules. *QSAR Comb. Sci.* **2003**, *22*, 565−574.

(8) Verma, R. P.; Kurup, A.; Hansch, C. On the rule of polarizability in QSAR. *Bioorg. Med. Chem.*, **2005**, *13*, 237−255.

(9) Wang, J.; Hou, T.; Duan, Y. Molecular polarizability calculations with dipole interaction models. *J. Phys. Chem. A*, submitted for publication.

(10) Applequist, J.; Carl, J. R.; Fung, K. K. *J. Am. Chem. Soc.* **1972**, *94*, 2952- 2960.

(11) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341−350.

(12) van Duijnen, P. T.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399−2407.

(13) Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1154−1163.

(14) Stout, J. M.; Dykstra, C. E. *J. Am. Chem. Soc.* **1995**, *117*, 5127−5132.

(15) Miller, K. J.; Savchik, J. A. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.

(16) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8642.

(17) Glen, R. C. *J. Comput-Aided Mol. Des.* **1994**, *8*, 457−466.

(18) Slater, J. C. Atomic shielding constants. *Phys. Rev. 36*, 57−64.

(19) Wang, J.; Krudy, G.; Xie, X.-Q.; Wu, C.; Holland, G. *J. Chem. Inf. Mod.* **2006**, *46*, 2674−2683.

(20) Wang, J.; Kollman, P. A. Automatic parameterization of force field by systematic search and genetic algorithms. *J. Comput. Chem.* **2001**, *22*, 1219−1228.

(21) Hou, T.; Wang, J. M.; Niao, N.; Xu, X. J. Applications of genetic algorithms on the structure-activity relationships analysis of some cinnamamides. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775−781.

(22) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(23) Xiao, Y. L.; Williams, D. E. Genetic algorithms for docking of actinomycin d and deoxyguanosine molecules with comparison to the crystal structure of actinomycin d-deoxyguanosine complex. *J. Phys. Chem.* **1994**, *98*, 7191−7200.

(24) Bowie, J. U.; Eisenberg, D. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 4436−4440.

(25) Sybyl user manual, Tripos Inc., St. Louis, MO, 1995.

(26) Jaguar 5.5 user manual, Schrodinger LLC, New York, NY, 2004.