Prediction of Partial Molar Volumes of Amino Acids and Small Peptides: Counting Atoms versus Topological Indices

Suman Sirimulla, Maricarmen Lerma, and William C. Herndon*

Department of Chemistry, University of Texas at El Paso, 500 West University Avenue,
El Paso, Texas 79968

Received August 25, 2009

Experimental data of partial molar volumes of amino acids and small peptides were compiled from several publications and enabled us to perform a predicative analysis based on quantitative structure−property relationships (QSPR). Based on the simplest level of the descriptors, the new method has high accuracy and was found to be more reliable when compared to the latter QSPR method based on topological indexes. Incorporation of isoelectric pH and 3-D solvent-accessible surface area parameters increased the predictability of the equation to a small extent. Cross-validation studies show that this method is successful in predicting the partial molar volumes of other noncoded amino acids, dipeptides, and diketopiperazine derivatives. This method is the beginning of new studies for larger peptides and proteins. It also can be suggested to be used for molecules that contain the same type of atoms as an amino acid.

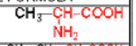
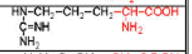
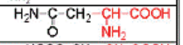
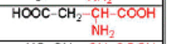
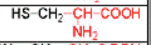
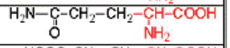
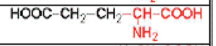
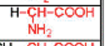
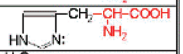
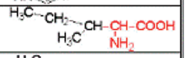
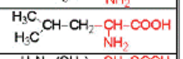
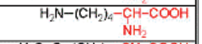
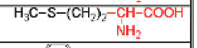
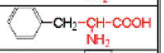
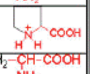
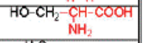
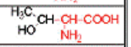
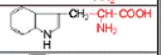
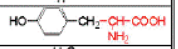
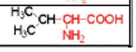
1. INTRODUCTION

The word “amino acids” usually refers to the 20 genetically encoded amino acids, either their neutral molecular forms, or their charged zwitterionic structures. Amino acid partial structures called residues, found in peptides and proteins, are also identified using the name of the parent acid.[1] These names, common abbreviations, used to identify the residual structures, and linear diagrams for the neutral molecular structures of each amino acid compound are given in Table 1. The word “peptide” generally denotes a small polymer made up of two or larger numbers of amino acid residues, linked by peptide bonds. Peptides can possess linear, cyclic, or branched molecular structures.[1]

Partial molar volumes (PMVs) of the zwitterionic α-amino acids and peptides in neutral aqueous solution are experimental properties, related to the thermodynamic properties by the equations of statistical mechanics (PMV thermochemical symbol, V2°, units, cm^3/mol). Experimental PMVs of aqueous amino acids and peptides to be considered in this Article are generally known to be very precise and accurate, derived from measurements of solution densities in a series of decreasing concentration, with extrapolation to infinite dilution.

A great deal of the interest pertaining to amino acids and larger peptides is coupled with their characterization as the basic building blocks of proteins. Proteins are macromolecules (polymers) that are constructed from one or more unbranched chains of amino acids. A typical protein contains 200−300 amino acids, but some are smaller (often called peptides), and some are much larger. The largest to date is titin, a protein found in skeleton and cardiac muscle, containing 26 926 amino acids in a single chain. Every function of the living cell depends in some way on proteins. In fact, the structure of cells, and the extracellular matrix in which they are embedded, is largely made of protein. Furthermore, the catalysis of a majority of biochemical reactions is carried out by enzymes, in which the essential components are mainly proteins; the receptors for hormones and other signaling molecules are also proteins. From a perusal of apropos protein literature, one gains the distinct impression that it is axiomatically assumed that experimental,

Table 1. Names, Abbreviations, and Linear Structures of 20 Amino Acids

NAME	ABBREVATIONS		LINEAR STRUCTURE FORMULA
Alanine	Ala	A	
Arginine	Arg	R	
Asparagine	Asn	N	
Aspartic acid	Asp	D	
Cysteine	Cys	C	
Glutamine	Gln	Q	
Glutamic Acid	Glu	E	
Glycine	Gly	G	
Histidine	His	H	
Isoleucine	Ile	I	
Leucine	Leu	L	
Lysine	Lys	K	
Methionine	Met	M	
Phenylalanine	Phe	F	
Proline	Pro	P	
Serine	Ser	S	
Threonine	Thr	T	
Tryptophan	Trp	W	
Tyrosine	Tyr	Y	
Valine	Val	V	

* Corresponding author e-mail: wherndon@utep.edu.

10.1021/ci900318c © 2010 American Chemical Society
Published on Web 01/08/2010

aqueous solution, infinite dilution partial molar volumes of the zwitterionic α-amino acids and peptides provide useful and important fundamental reference information related to studies of protein denaturation, hydration, intermolecular association, and other significant physical properties.

The research to be described in this Article primarily consists of development and evaluation of a fundamentally new and innovative method for predicting high accuracy values (as compared to experimental data) of the partial molar volumes of amino acids and peptides.

2. DISCUSSION

2.1. Partial Molar Volumes of Amino Acids. The aqueous solution, infinite dilution partial molar volumes of all 20 encoded amino acids have been known for some time.[27−29,33,34] They are included in a valuable key reference published in 1997 by Kharakoz et al.,[16] which summarizes and extensively evaluates all of the then existing experimental data, finally tabulating a set of recommended PMV values for 24 amino acids and 13 dipeptides. A table of these data, augmented with additional contemporaneous experimental data, and the analyzed results are presented. Kharakoz et al. found that the precision of the experimental data was generally excellent for amino acids, in every case actually better than 1.0 cm^3/mol, about 1% of the largest values of measured amino acid PMVs. One also notes that since the 1930s[4] the PMVs for many of the compounds have been reinvestigated in several different laboratories with good agreement between results, consistently better than 0.5 cm^3/mol.

2.2. Previous Correlation Studies. The initial impetus for the research described in this Article is a QSPR (quantitative structure−property relationship) of amino acid PMVs, which was previously carried out by Randic, Mills, and Basak,[26] published in 2000. They defined a new type of mathematical descriptor, a so-called "generalized topological index". The development of the index descriptor for each amino acid is carried out by fairly complex arithmetical calculations, and the full paper, cited above, should be consulted for a detailed explanation. The topological indices were then used as the independent variable parameters in a statistical QSPR study to obtain putative high-quality correlations of PMVs for 16 and 17 compound subsets of the 20 natural amino acids.

One of the major presuppositions used by Dr. Herndon's research group at UTEP in numerous QSAR studies carried out during the last 20 years has been that optimum descriptors for analyzing molecular structure−property and structure−activity relationships are normally simple indicator variables. For these types of organic chemical systems, effective choices of indicator variables generally include direct counts of atoms by element, often extended to make use of the atom types sorted by hybridization and/or types of substituents, the bond types, or more complex functional groups and structural feature, all descriptors evaluated in order of increasing complexity. This additivity methodology, with the exception of the hierarchical extension, is, of course, one of the prosaic, common standards for legions of successful QSPR and QSAR studies.[2,6,31,35] A significant advantage is that the additive parameters of this type are local structural descriptors, bearing one-to-one correspondence to the conventional representations of molecular structure. As such, they normally allow precise structural interpretations of importance in determining the value of a physical chemical or biochemical property. In addition, the design of new molecules with altered desirable structures and properties is facilitated, and the additive predictive capabilities of derived QSPR and QSAR relationships are easily tested.

In addition, a previously published study in which Herndon and Radhakrishnan[13] discovered an exact mathematical algebraic equivalence between group additivity and topological index QSPR equations for the PMVs of saturated normal alkanes strongly suggested that a similar equivalence might be expected when these two different approaches were separately used to correlate the PMVs of the amino acids. However, we were interested in examining the capabilities of calculated methods to treat larger structures related to the amino acids such as polypeptides and possibly proteins. Extensions of the topological index methodology to these more complicated systems appeared to be much more difficult to implement than the simple additivity analysis. An improved level of accuracy was also desirable, and this did not seem to be attainable using the topological matrix method. Of course, the initial step in an extended investigation of the additivity approach required a detailed comparison of the two methodologies and a critical comparison of the two sets of correlation results. The differences between the two methods and the critical comparison of correlation results are outlined below.

2.3. General Topological Index versus Group Additivity. During the last 30 years, in addition to group additivity, the topological index[13] approach has evolved as a general protocol for developing quantitative structure−property and structure−additivity relationships, with large numbers of successful studies.[4,9−11,24] One of the adduced advantages is succinctness. Numerous topological index QSPR applications make use of only one topological structure index as an independent variable, exemplified in the first α-amino acid/PMV study of this type.[17] This may be the reason that the number of the descriptors in the Randic et al. topological index study[26] is emphasized to be unity, that is, a single, numerical generalized connectivity index value for each one of the α-amino acids. However, one should note that such emphasis is very misleading. In actuality, the number of parameters that must be optimized to establish the required individual generalized topological indices used in a final QSPR linear regression equation is a total of six: first, after several optimizations (by hand), four different weighting factors for C, N, O, and S variables, used in calculating the individual numerical values of the connectivity index for each compound (Table 2),[26] and, second, after regression analysis at every stage of optimization, the final independent variable index coefficient and a constant term (Table 4).[26]

Other recent topological index studies on molecular volumes of the α-amino acid have used a like number of adjustable parameters,[23,24] and for the research described in this Article, this number of the optimized parameters in the topological index study turned out to be useful, because it allowed direct comparison to the results in the present Article, and to several older examples of structure-PMV amino acid studies. Contrary to the first sentence of the Randic et al. paper, there are numerous such published additivity studies for α-amino acid. A 1934 ACS monograph edited by Cohn

Table 2. Experimental Partial Molar Volumes of Amino Acids (cm^3/mol)

encoded amino acids	PMV experimental	noncoded amino acids test set 1	PMV experimental
glycine	43.24	hydroxyproline	84.2
alanine	60.44	norleucine	107.75
valine	90.79	norvaline	91.7
leucine	107.66	B-alanine	58.5
isoleucine	105.6	isoserine	59.07
methionine	105.36	allothreonine	76.87
proline	82.2	β-phenylserine	124.69
phenylalanine	121.8	*S*-ethylcystein	104.2
tryptophan	143.9	ethionine	119.58
serine	60.69	*m*-tyrosine	123.11
threonine	76.86	3-aminotyrosine	129.61
asparagine	77.3	3,4-dihyroxyphenylalanine	125.76
glutamine	93.9	citrulline	115.94
tyrosine	123.7	α-aminobutyric acid	75.56
cysteine	73.45	β-aminobutyric acid	76.21
lysine	108.7	γ-aminobutyric acid	73.23
arginine	123.8	5-aminopentanoic acid	87.65
histidine	98.8	6-aminohexanoic acid	104.09
aspartic acid	74.3	7-aminoheptanoic acid	120
glutamic acid	89.5	8-aminooctanoic acid	136.03
		9-aminononanoic acid	151.3
		10-aminodecanoic acid	167.3
		11-aminoundecanoic acid	183

Table 3. Model-1 Multiple Linear Regression Analysis (PMV vs C, H, O, N, and S)[a]

variable	coefficient	std error	std coef	tolerance	*T*	*P* (2 tail)
carbon	7.783	0.238	1.830	0.076	32.673	0.000
hydrogen	3.736	0.181	1.546	0.042	20.669	0.000
nitrogen	4.220	0.513	0.285	0.198	8.230	0.000
oxygen	2.904	0.364	0.300	0.168	7.986	0.000
sulfur	14.624	1.282	0.188	0.871	11.404	0.000

Analysis of Variance

source	sum-of-squares	DF	mean-square	*F*-ratio	*P*
regression	12 028.416	4	3007.104	1050.095	0.000
residual	42.955	15	2.864		

[a] DEP VAR = PMV; $N = 20$; multiple $R = 0.998$; squared multiple $R = 0.996$; adjusted squared multiple $R = 0.995$; standard error of estimate = 1.692.

and Edsal[4] gives summaries of the earlier work, including the PMV work of Traube published in 1899. In general, both earlier and later experimental data and various model studies are uniformly characterized by concordant high-precision results,[4,8] and there are also numerous extensions to peptides and proteins.[7,12,14,15,19]

In contrast, many more recent additivity studies on peptides use a multitude of descriptors,[20,21,25,30] counting structural variables like the numbers of each of the 20 different residues, the types of end-groups, the number of amide linkages, etc. In the additivity results to be described later, we attempt a close comparison with the Randic et al. results by limiting a descriptor pool for regression to six elementary parameters, atom counts by (C, H, O, N, and S) and one additional descriptor, identified in the course of the investigation. Linear regression equations containing a constant were eschewed after finding that the value of an included constant was statistically insignificant in every attempted analysis, and also because of the ambiguity in defining a molecular structural interpretation for such a constant.

The two methodologies can be evaluated by comparison of the usual statistical parameters for correlated results (correlation coefficient, standard and mean deviations, and *F*-ratios). However, we were not able to formulate more stringent strategies for comparison of the two methods. This was principally because the several iterations for the nonlinear hand calculations of the generalized topological index, required before optimization in each case, precluded using cross-validation procedures.

3. METHODS AND RESULTS

3.1. PMVs of Encoded Amino Acids. The information to be analyzed and evaluated in this part of the Article consists primarily of results for the linear regression correlation of encoded amino acid experimental PMVs data, based on an atomistic additivity model. Additional data for 13 amino acids with unusual substituents or side-chain functionality[22] and 10 straight-chain aliphatic α,ω amino acids[3] have been added to the set of amino acids recommended by Kharakoz et al.[16] to compile our amino acid PMV data set, listed in Table 2. The original journal citation for

Table 4. Comparison of Results from Model-1 and Generalized Topological Index Models (cm^3/mol)

amino acid	PMV exp	model-1 calculated	model-1 residuals	topolog-17 residuals	topolog-16 residuals
glycine	43.24	44.272	−1.032	−0.24	−0.13
alanine	60.44	59.527	+0.913	−0.46	−0.70
valine	90.79	90.035	+0.755	−3.48	−4.39
leucine	107.66	105.290	+2.370	−1.73	−2.94
isoleucine[a]	105.60	105.290	+0.310	no data	no data
methionine	105.36	104.659	+0.701	+4.87	+3.84
proline	82.20	82.564	−0.364	+1.72	+1.08
phenylalanine	121.80	121.169	+0.631	+9.42	+8.15
tryptophan	143.90	144.690	−0.790	−0.70	−2.62
serine	60.69	62.431	−1.741	+3.99	+3.84
threonine[a]	76.86	77.685	−0.825	no data	no data
asparagine	77.30	78.169	−0.869	−5.76	−6.45
glutamine[b]	93.90	93.424	+0.476	−13.65	no data
tyrosine	123.70	124.073	−0.373	−0.47	−1.98
cysteine	73.45	74.151	−0.701	−6.56	−7.18
lysine[a]	108.70	113.245	−4.545	no data	no data
arginine	123.80	121.684	+2.116	+8.51	+7.11
histidine	98.80	98.787	+0.013	−1.90	−2.94
aspartic acid	74.30	73.118	+1.182	+5.28	+4.89
glutamic acid	89.50	88.372	+1.128	+1.17	+0.45

[a] Not used in Randic et al.[26] [b] Defined as an outliner in Randic et al.,[26] to be ignored.

each of the listed compounds was examined, and the PMV values that are given for each compound have been compared to the data in the original articles.

As far as can be ascertained, these 43 amino acids comprise the complete set of amino acids with known experimental PMVs, where the PMVs were determined at 25 °C and several appropriate series of low concentrations. The partition of the data in the table (Table 2) into a reference set comprising the 20 uncoded amino acids, used for development of parameters to correlate their experimental PMVs, and a 23 compound test set to evaluate predictions, is an obvious natural consequence of the composition of the table.

3.2. PMV Data Analysis: Model-1. The starting point for the data analysis is the PMV data listed in the second column of Table 2. A single parameter, the molecular weights of the 20 coded amino acids, gives what might be considered as an acceptable correlation with the aqueous solution PMVs (correlation coefficient 0.919). Taking the numbers of atoms by element as descriptors (C, H, N, O, and S), defined as model-1, must necessarily lead to an improved correlation because of the additional flexibility in the independent variable descriptor set. In addition, the reader may recall that the atom level is also the first and simplest level for descriptors in our preferred hierarchical variant of QSPR methodology. The statistical results of model-1 regression are tabulated in Table 3, and the expectation of obtaining an improved correlation is unquestionably fulfilled, generating the astonishing results illustrated in Figure 1.

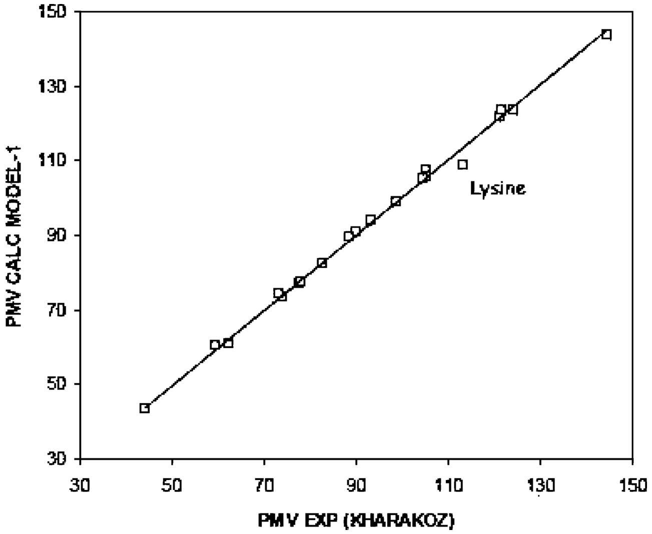


Figure 1. Model-1 PMVs for encoded amino acids: calculated versus experimental (cm^3/mol).

Calculated PMVs and residual errors for model-1 and the topological model results for the 17 and 16 compounds subsets studied by Randic et al.[26] are compared in Table 4. The 16 compound Randic et al. results were obtained after disregarding the largest outlier (glutamine) in the 17-compounds model. Any benefits arising from this modification are not readily apparent.

Randic et al. in their amino acid PMV study justified the removal of outliers with an error greater than twice the standard error, on the supposition that the ensuing higher quality regression would have greater usefulness for the data set in hand.

However, in the present study, it is preferable to retain all of the experimental data, because the ultimate objective in this present research is to test the capability of the amino acid atom count parameters to actually predict accurate values for partial molar volumes of related compounds, that is, molecular systems with distinctly different structural motifs, but composed of atoms of the same five elements. Examples include noncoded amino acids, whose PMVs are given in Table 2 (test set 1), and a sizable number of small polypeptides whose PMVs will be treated and discussed later. Additional examples (not treated in this Article) with various types of known thermodynamic properties include a few large polypeptides and proteins, and over 100 other known compounds of biological interest containing one or more of each of the specified elemental atoms.

3.3. PMV Data Analysis: Model-2. The model-1 correlation of the amino acid data is certainly remarkable,

Table 5. Model-2 Multiple Linear Regression Analysis (PMV vs C, H, O, N, S, SIDE-CHN)[a]

variable	coefficient	std error	std coef	tolerance	*T*	*P* (2 tail)
carbon	7.867	0.201	1.849	0.074	39.106	0.000
hydrogen	3.722	0.151	1.540	0.042	24.654	0.000
nitrogen	5.514	0.637	0.372	0.089	8.656	0.000
oxygen	2.219	0.393	0.229	0.100	5.645	0.000
sulfur	14.492	1.072	0.187	0.869	13.524	0.000
SCHAIN	−2.541	0.926	−0.065	0.291	−2.743	0.016

Analysis of Variance

source	sum-of-squares	DF	mean-square	*F*-ratio	*P*
regression	12 043.434	5	2408.687	1207.059	0.000
residual	27.937	14	1.996		

[a] DEP VAR = PMV; N = 20; multiple R = 0.999; squared multiple R = 0.998; adjusted squared multiple R = 0.997; standard error of estimate = 1.413.

Table 6. Comparisons of Experimental PMVs versus Calculated PMVs of Model-1, Model-2, Model-3, and Model-4

amino acids	PMV exp	model-1 predicted	model-1 residuals	model-2 predicted	model-2 residuals	model-3 predicted	model-3 residuals	model-4 predicted	model-4 residuals
alanine	60.44	59.527	0.913	59.603	0.837	58.796	1.644	59.646	0.794
arginine	123.8	121.684	2.116	123.255	0.545	123.506	0.294	123.364	0.436
asparagine	77.3	78.169	−0.869	76.383	0.917	76.919	0.381	76.911	0.389
asparticacid	74.3	73.118	1.182	74.449	−0.149	74.635	−0.335	74.685	−0.385
cysteine	73.45	74.151	−0.701	74.095	−0.645	73.734	−0.284	73.423	0.027
glutamine	93.9	93.424	0.476	91.693	2.207	92.412	1.488	92.431	1.469
glutamicacid	89.5	88.372	1.128	89.759	−0.259	90.708	−1.208	90.691	−1.191
glycine	43.24	44.272	−1.032	44.293	−1.053	43.139	0.101	42.548	0.692
histidine	98.8	98.787	0.013	99.133	−0.333	98.778	0.022	98.997	−0.197
isoleucine	105.6	105.29	0.31	105.532	0.068	105.962	−0.362	105.814	−0.214
leucine	107.66	105.29	2.37	105.532	2.128	105.988	1.672	105.925	1.735
lysine	108.7	113.245	−4.545	112.227	−3.527	111.872	−3.172	111.866	−3.166
methionine	105.36	104.659	0.701	104.715	0.645	105.076	0.284	105.387	−0.027
phenyalanine	121.8	121.169	0.631	121.689	0.111	121.243	0.557	121.392	0.408
proline	82.2	82.564	−0.364	82.779	−0.579	81.709	0.491	81.6	0.6
serine	60.69	62.431	−1.741	61.822	−1.132	61.188	−0.498	61.355	−0.665
threonine	76.86	77.685	−0.825	77.132	−0.272	76.992	−0.132	76.874	−0.014
tryptophan	143.9	144.69	−0.79	144.116	−0.216	144.455	−0.555	144.407	−0.507
tyrosine	123.7	124.073	−0.373	123.908	−0.208	123.718	−0.018	123.529	0.171
valine	90.79	90.035	0.755	90.223	0.567	90.225	0.565	90.242	0.548

particularly if compared to the generalized topological index results, in which a larger number of adjustable parameters is employed, applied to a fewer number of molecules. However, it should be noted that the mean deviation for model-1 (1.09 cm^3/mol), although small, is still more than twice as large as the usual reported precision of amino acid PMV experimental data. However, it is readily apparent that approximately 20% of the total mean error for model-1 is due to the single compound, lysine, with a negative 4.545 cm^3/mol error, approximately 4 times the mean error, and 2.7 times the calculated standard error (see Figure 1 and Table 2).

Of course, the statistical regression results can be substantially improved by simply omitting lysine from the analysis, as was done in the Randic et al. topological index study. However, a better structure-related approach is suggested by the fact that lysine happens to belong to a small subgroup of the encoded α-amino acid in which the predominating structures at neutral pH in aqueous solution do not possess the customary α-ammonium carboxylate zwitterion. Instead, ionized side-chain functional groups are partially involved, aliphatic ammonium group in the cases of lysine, arginine, and histidine, and the side-chain carboxylate for aspartic and glutamic acids.

The most simplistic way to model these side chains without exceeding six parameters is to define a single indicator-type variable: +1 for side-chain amino group structures (which thus must include tryptophan), −1 for side-chain carboxyl group, and a value of zero otherwise. A physical interpretation of this construct could be that one is postulating that effects of anionic and cationic side chains on the measured PMVs are close to equal in magnitude and opposite in sign. The results of this atom type with side-chain parameter regression model (model-2), tabulated in Tables 5 and 6, provide a practical justification for this six-parameter model. There are dramatic improvements in regression statistics, including a 20% improvement in the important *F*-ratio statistic, and a significant reduction in standard error of estimate from 1.6 to 1.2. The self-evident high quality of the correlation is illustrated below in Figure 2.

However, the practical usefulness of a QSPR analysis does not really lie in the ability of a model equation to correlate data for one specific property, even for a large set of dissimilar compounds. The ultimate goals of QSAR activity and QSPR property studies are prediction of biological potencies or physical properties of yet uninvestigated compounds. Obtaining a really excellent QSPR correlation equation, as is found for model-2, should not end an

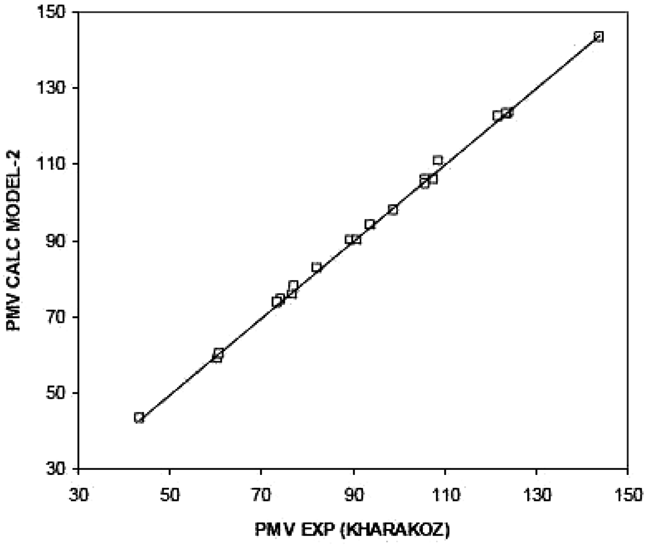


Figure 2. Model-2 linear regression: calculated PMV values (abscissa) versus experimental values (ordinate) (cm^3/mol).

investigation, and the correlation results should not be considered as proof that a particular methodology will have general predictive usefulness.

Thus, the additivity methodology presented in this model, and the correlation equation represented by Figure 2, need to be evaluated for the ability to predict PMVs, rather than just correlate already known experimental values. For the additivity model, this predictive capability will be demonstrated and discussed next, but such a corresponding detailed analysis is not feasible for the topological index model. However, one comparison of results from the atom count and topological models is possible, which does address the question of prediction accuracy, at least for three of the 20 encoded amino acids. The results of this comparison, in the next paragraph, will terminate both our criticisms of the generalized topological index, and also bring this section to a germane conclusion.

Randic et al.[26] used their best QSPR regression (six optimized parameters, $N = 16$) to, in fact, predict values for the PMVs of isoleucine, threonine, and lysine. As it was mentioned previously, these compounds were not used in developing their QSPR equations. We used model-2 (also six optimized parameters) to develop a QSPR equation, leaving out the experimental data for the same three amino acids. The model-2 errors are −1.22, −0.84, and −5.35 cm^3/mol, respectively, to be compared to the Randic et al. respective errors, +9.9, +10.05, and −13.7 cm^3/mol. Thus, the true predictions are comparable to those obtained for correlations, very good for the atom additivity model and unacceptable large errors for the generalized topological index model.

3.4. PMV Data Analysis: Model-3 and Mode-4 (Effect of pH and 3D Solvent-Accessible Surface Area). As all of the amino acids are comprised of an acidic carboxylic group and a basic amine group, there is always an exchange of hydrogen ion from the carboxylic group to amine group, which makes the amino acids have both positive charge and negative charge. The charge of amino acids is highly dependent on pH. To make the comparisons more meaningful, all of the amino acids are adjusted to their neutral form, and their pH at this point, which is also called as isoelectric pH (abbreviated as PI), is calculated and introduced as a descriptor in the equation. Table 9 and Figure 5 summarize the statistical results of this model (model-3). Although there is not a significant change in the statistical parameters from model-2, the residuals were decreased to a small extent.

To further examine the effect of three-dimensional descriptors on this equation, we calculated the three-dimensional solvent-accessible surface area (ASA) of amino acids present at their isoelectric pH. Both isoelectric pH and their ASA are calculated using JChem[18] molecular modeling software. The major 3D conformer of each amino acid at their isoelectric pH is considered for the solvent-accessible surface area calculation, and water is used as the solvent (solvent radius: 1.4 Å). The statistical results are shown in Table 10, and the fitting graph is shown in Figure 6. The results look equally good as model-3 except the *F*-ratio is

Table 7. Model-2 Predictions of PMVs for 23 Noncoded Amino Acids (cm^3/mol)

noncoded amino acids	PMV experimental	PMV calculated	calc errors
hydroxyproline	84.2	83.846	0.354
norleucine	107.75	106.043	1.707
norvaline	91.7	90.4	1.3
B-alanine	58.5	59.114	−0.614
isoserine	59.07	60.169	−1.099
allothreonine	76.87	75.812	1.058
β-phenylserine	124.69	123.59	1.1
S-ethylcystein	104.2	105.048	−0.848
ethionine	119.58	120.691	−1.111
m-tyrosine	123.11	123.59	−0.48
3-aminotyrosine	129.61	128.454	1.156
3,4-dihyroxyphenylalanine	125.76	124.644	1.116
citrulline	115.94	114.429	1.511
α-aminobutyric acid	75.56	74.757	0.803
β-aminobutyric acid	76.21	74.757	1.453
γ-aminobutyric acid	73.23	74.757	−1.527
5-aminopentanoic acid	87.65	90.4	−2.75
6-aminohexanoic acid	104.09	106.043	−1.953
7-aminoheptanoic acid	120	121.686	−1.686
8-aminooctanoic acid	136.03	137.329	−1.299
9-aminononanoic acid	151.3	152.972	−1.672
10-aminodecanoic acid	167.3	168.615	−1.315
11-aminoundecanoic acid	183	184.258	−1.258

Table 8. Experimental and Prediction Data of PMVs of Dipeptides (cm^3/mol)

test set 2 dipeptides	PMV experimental	PMV model-2 predicted	PMV error
GlyGly	76.30^{a}	78.279	−1.979
GlyLeu	139.30^{a}	140.851	−1.551
LeuGly	143.70^{a}	140.851	2.849
GlyAla	92.80^{a}	93.922	−1.122
AlaGly	94.80^{a}	93.922	0.878
GlyPhe	155.54^{a}	157.343	−1.803
PheGly	160.30^{d}	157.343	2.657
GlyVal	122.30^{a}	125.208	−2.908
ValGly	126.00^{a}	125.208	0.792
GlySer	92.93^{a}	94.977	−2.047
GlyThr	108.50^{a}	110.620	−2.120
GlyAsn	110.11^{a}	113.087	−2.977
AlaAla	110.60^{a}	109.565	1.035
SerSer	111.80^{a}	111.674	0.126
Gly(α-aminobutane)	107.81^{b}	109.565	−1.755
diketopiperazines			
c-GlyGly	76.85^{c}	69.616	7.234
c-AlaAla	112.58^{c}	100.902	11.678
c-SarSar	113.36^{c}	100.902	12.458

[a] Kharakoz, ref 16. [b] Mishra and Ahluwalia, ref 22. [c] Hakin et al., ref 9. [d] Greenstein and Wyman, ref 5.

Table 9. Model-3 Multiple Linear Regression Analysis (PMV vs C, H, O, N, S, SIDE-CHN, PI)[a]

variable	coefficient	std error	std coef	tolerance	*T*	*P* (2 tail)
carbon	7.730	0.195	1.817	0.065	39.730	0.000
hydrogen	4.014	0.200	1.661	0.020	20.091	0.000
nitrogen	6.058	0.638	0.409	0.073	9.494	0.000
oxygen	2.245	0.357	0.232	0.100	6.296	0.000
sulfur	14.780	0.982	0.190	0.850	15.049	0.000
SCHAIN	−2.316	0.847	−0.060	0.286	−2.733	0.017
PI	−0.509	0.253	−0.133	0.031	−2.007	0.066

Analysis of Variance

source	sum-of-squares	DF	mean-square	*F*-ratio	*P*
regression	12 050.043	6	2008.341	1224.163	0.000
residual	21.328	13	1.641		

[a] DEP VAR = PMV; $N = 20$; multiple $R = 0.999$; squared multiple $R = 0.998$; adjusted squared multiple $R = 0.997$; standard error of estimate = 1.281.

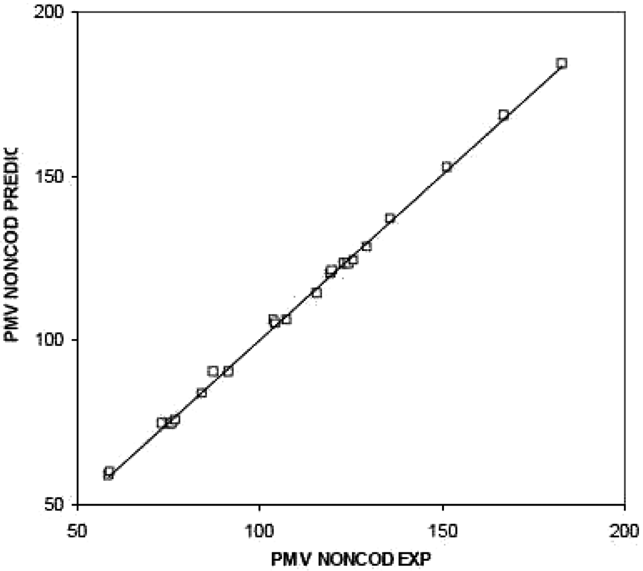


Figure 3. Predicted versus experimental PMVs for noncoded amino acids. The data points are superposed on a single line drawn with unit slope (cm^3/mol).

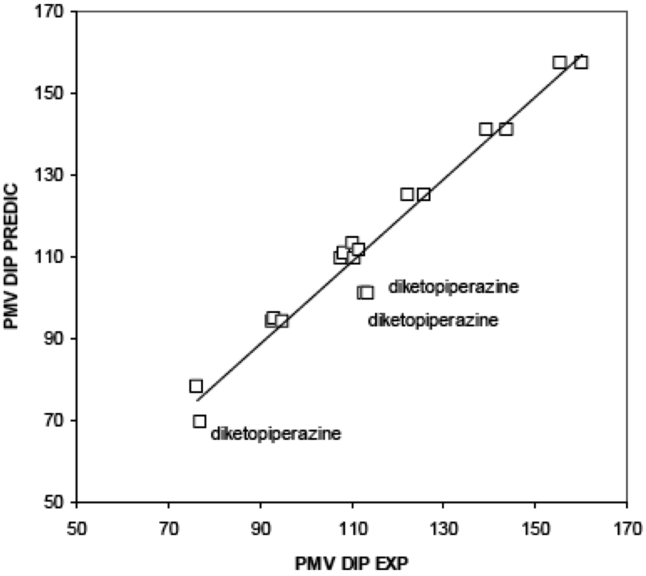


Figure 4. PMVs for dipeptides: predicted data versus experimental data (cm^3/mol).

decreased from 1207 to 1040. All calculated PMV values from this model regression are now within 1−3% of the experimental values in every case, close to the limits of precision of the experimental measurements.

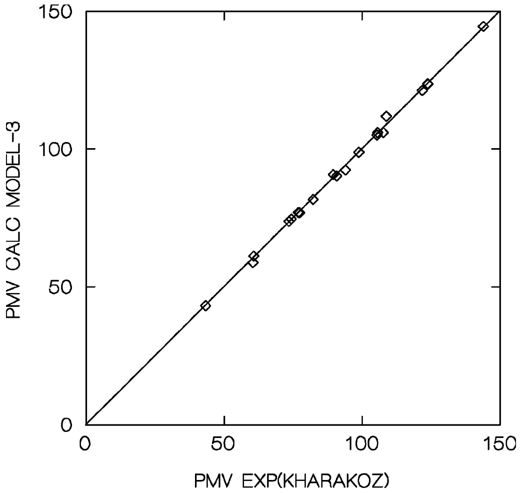


Figure 5. Model-3 linear regression: calculated PMV values (abscissa) versus experimental values (ordinate) (cm^3/mol).

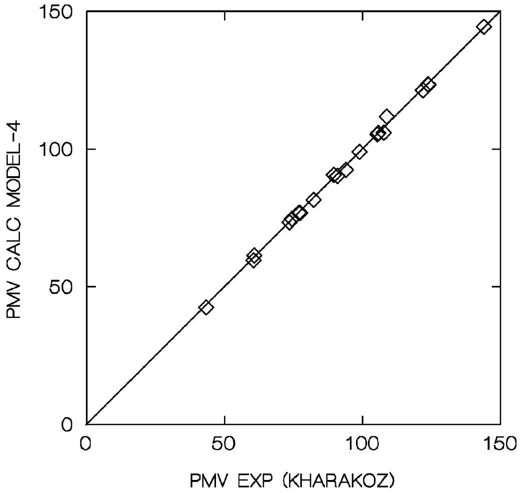


Figure 6. Model-4 linear regression: calculated PMV values (abscissa) versus experimental values (ordinate) (cm^3/mol).

4. CROSS-VALIDATION/VERIFIABLE PREDICTIONS FOR PMVS OF NONCODED AMINO ACIDS AND DIPEPTIDES

4.1. Pragmatic Aspects of QSAR/QSPR Predictions. The work to be described in this section concerns the capability of the additivity model correlation calculations to provide easily obtained and, most importantly, verifiable predictions, not just correlations of the PMV data under consideration. One of the most common of the contemporary procedures used to validate the capability to predict in medicinal chemistry QSPR and QSAR studies has been to exclude a small random sample of the compounds to be used as a test set for prediction.[32] As far as can be determined, no significant failures of this validation procedure have ever been reported in the scientific literature, particularly for the sets of similar, related, so-called congeneric, molecules normally considered in biochemical or biophysical studies. One always seems to obtain acceptable “predicted” values of the target property for the compounds of the test set. Of course, however, one has to suspend any belief in the value of random sampling as an effective statistical analysis tool to interpret such results as bona fide predictions.

In general, regarding prediction, Herndon and co-workers have previously proposed that the quality and usefulness of QSAR/QSPR equations can only be truly appraised by requiring demonstrations of predictive capabilities with concrete verifiable examples. For a sizable large fraction of the studies carried out in biophysical and physical chemical systems, this requirement must be met by using the limited amounts of data in hand. In such cases, it is necessary to treat the available data so as to allow one or both of two different types of predictive results to be generated and verified: (1) predictions for compounds with biological activities or physical properties lying outside of the range of such properties for corresponding training sets, or (2) predictions for properties of compounds with nontrivial molecular structural features, not existing in the training structures. These precepts are illustrated and enforced in the two examples given next.

4.2. Verifiable Predictions of PMVs for Noncoded Amino Acids. The sources and general accuracy of the data for the work to be described in this section were discussed in section 3.1. The data consist of experimentally obtained PMVs for the 20 encoded α-amino acid and the corresponding experimental results for 23 additional noncoded amino acids. Thus, the number of the noncoded amino acids with experimentally measured PMV values is somewhat larger than the number of the natural α-amino acid. In addition, several types of functionality are present in this test set, not found in the coded acids, and 10 of the noncoded compounds

Table 10. Model-2 Multiple Linear Regression Analysis (PMV vs C, H, O, N, S, SIDE-CHN, PI, ASA)[a]

variable	coefficient	std error	std coef	tolerance	*T*	*P* (2 tail)
carbon	7.448	0.357	1.751	0.019	20.842	0.000
hydrogen	3.768	0.328	1.559	0.007	11.490	0.000
nitrogen	5.992	0.645	0.404	0.072	9.295	0.000
oxygen	1.670	0.707	0.173	0.026	2.363	0.036
sulfur	13.350	1.806	0.172	0.254	7.391	0.000
SCHAIN	−2.372	0.853	−0.061	0.284	−2.781	0.017
PI	−0.617	0.279	−0.161	0.026	−2.211	0.047
ASA	0.023	0.025	0.259	0.002	0.945	0.363

Analysis of Variance

sourve	sum-of-squares	DF	mean-square	*F*-ratio	*P*
regression	12 051.521	7	1721.646	1040.821	0.000
residual	19.849	12	1.654		

[a] DEP VAR = PMV; $N = 20$; multiple $R = 0.999$; squared multiple $R = 0.998$; adjusted squared multiple $R = 0.997$; standard error of estimate = 1.286.

are aliphatic α,ω amino acids in which only one of the 10 compounds has the basic structure of an α amino acid.

The model-2 regression analysis of the PMVs of the 20 coded amino acids provides the optimized coefficients of the six additivity parameters, given in Table 5. These coefficients are used to obtain predicted values of the PMVs for the 23 noncoded amino acids. The results, of course, can then be verified by comparison with the actual experimental data. These data and the predicted PMV values are listed in columns two and three of Table 7, respectively, with the prediction errors given in column four of the table. Each of the compounds listed in Table 7 has a molecular structure with distinct individual structural differences, differing from the molecular structures of the 20 encoded amino acids. Thus, accurate predictions of the PMVs for the compounds in Table 7 using the optimized model-2 atom parameters (Table 5) present a stringent validation challenge for the atomistic additive protocol.

However, as one can see from the tabulated results, the challenge is simply met by a straightforward application of an additivity procedure. In fact, the mean and standard deviations for the prediction errors are only ±1.252 and ±1.393 cm^3/mol, respectively, and the correlation coefficient comparing experimental and predicted values is larger than 0.9999. In addition, only one compound of this test set of noncoded compounds has a prediction error larger than 1 cm^3/mol, a 3% error for 5-aminopentanoic acid ($C_5H_{11}NO_2$). The upper limit of error, observed for the encoded amino acid correlation, is about the same, a 2% discrepancy for lysine ($C_6H_{14}N_2O_2$).

The results of this study are indisputable and are accurate predictions of PMVs as compared to actual precise experimental results. It is possible that one might think that graphs illustrating these types of results might be considered to be superfluous. Even so, such a graph illustrating the results summarized in Table 8 is presented in Figure 3. This figure simply adds graphical emphasis to the overall aspects of the extraordinary high quality of this prediction study. It is also interesting to note the distinct similarity of Figure 3 to the previous results presented in Figures 1 and 2 in which optimized results of regression correlations have been illustrated.

4.3. Verifiable Predictions of PMVs for Dipeptides. The next logical step in the development of the research described in this Article is an extension of the successful amino acid methodology to a consideration of the PMVs of peptides, beginning, of course, with dipeptides. The data that will be examined consist of PMVs for the 18 dipeptides listed in Table 7. Thirteen of the dipeptides are taken from the Kharakoz compilation,[16] and the other five are from other sources.[5,9,22] Three of these latter compounds[9] are diketopiperazine derivatives, that is, neutral cyclic dipeptides, for which chemical intuition projects a systematic error in predicted PMVs if calculated with the open zwitterionic amino acids parameters.

The data for PMVs of dipeptides are obviously not extensive, nor as diverse as the amino acid data. Only four of the 18 known values are for compounds that are not derivatives of glycine, and three of the 18 are derivatives of 2,5-diketopiperazine (the cyclic GlyGly peptide). Be that as it may, each of the dipeptides is distinctly different from the amino acids from section 4.2, containing either one or two un-ionized peptide linkages, in addition to possessing an increased distance separating the normal zwitterionic charges. Thus, the calculated dipeptide PMVs are significant extrapolations from the model-2 correlation of PMVs for the monomeric amino acids.

The overall result clearly identifies the cyclic dipeptides as outliers, and small overall decreased accuracy of the predicted PMV values is found, demonstrated in the graph shown in Figure 4. A graph of predicted versus experimental values, excluding the three cyclic dipeptides, is actually very similar to the parametrization graph (Figure 2), and to the model-2 prediction graph (Figure 3), that is, quite uninformative, just data points lying close to a straight line drawn with close to unit slope. However, such a comparison does emphasize the fact that the errors for predicted monomeric amino acids and for open-chain dipeptide compounds are the same order of precision as the errors of the correlation in the investigation of the parent monomeric acids.

The large errors in the predicted values for the three cyclic dipeptides were not unexpected. These compounds are not dissimilar to the other dipeptides in possessing small cyclic structures, but the structures of the three molecules also differ by not existing as zwitterionic at the pH of neutral water. Because they are neutral organic compounds, the surrounding water structure is anticipated to be highly dissimilar to that of the normal ionic open-chain amino acids and dipeptides. Of course, it is possible to mitigate a large fraction of the errors for the three cyclic dipeptides if the independent variable list for the statistical regression analysis made use of a parameter denoting the presence or absence of the cyclic peptide structures. Similar indicator variables are commonly used in QSAR and QSPR studies, and, in fact, our model-2 parameters include just such a descriptor for the three types of amino acid side chains. The parametrization of this type of descriptor for cyclic dipeptides in the present study is not justifiable due to the absence of a set of experimental data that could be used to evaluate both correlative and, more importantly, the predictive utility of the cyclicity parameter.

5. SUMMARY AND CONCLUDING REMARKS

The water solubility of an amino acid or peptide is one of its most important physical properties, for both practical and scientific reasons. Some of the main scientific uses involve aqueous measurements of partial molar quantities such as the partial molar heat capacity or the partial molar volume, key parameters in understanding the thermodynamics of aqueous solutions. The most important of these partial molar quantities is the partial molar free energy or chemical potential. However, as explained in the Introduction, the easiest such property to measure precisely is the partial molar volume. This is actually a very fortunate circumstance because it can be coupled with the fact that experimental, aqueous solution, infinite dilution partial molar volumes of amino acids and peptides provide useful and important fundamental reference information related to studies of protein denaturation, hydration, intermolecular association, and other significant physical properties.

Generally, quantitative structure/property studies (QSPR) of amino acids, peptides, and even proteins have made use of the encoded α-amino acid residues as the basic units for understanding the physical properties of such biomolecules.

The main initial goal of the research for this Article was to examine the possibility that experimental partial molar volumes of amino acids might be understandable on an atomistic basis (five parameters for C, H, N, O, S) rather than 20 parameters (for the 20 different residues).

All of the most significant findings and conclusions related to the work reported in this Article are linked to the initial, completely unexpected, results, strongly supporting this hypothesis. These results are presented earlier, tabulated in Tables 3−6, and illustrated in Figures 1 and 2. The tables and figures, based on multilinear regression studies, demonstrate that coded amino acid PMV data are precisely determined by the number of atoms of each type, with one additional indicator variable designating types of side chains. Some might consider the ensuing studies and results, reported in section 4, to be even more intriguing and potentially useful. In this section, the atom parameters from the 20 amino acid correlation are used to provide accurate, verifiable predictions of the PMVs in a dipeptide data set. (See Tables 7 and 8 for the relevant data and the graphs in Figures 3 and 4.). Although the incorporation of pH and three-dimensional solvent-accessible surface area parameters does not change the statistics drastically, they improve the predictability to a small extent.

Future research along the lines described in the present Article could include studies incorporating PMV data for additional amino acids, tripeptides, tetrapeptides, and perhaps even larger peptides. PMVs for several larger peptides are known, and it might be possible to bracket the peptide size where conformational effects make a linear model ineffective. Extending this idea, it is possible to imagine that the difference between the calculated atomistic PMVs of a series of denatured proteins (as compared to the experimental PMV values) might be a useful measure of conformational and folding properties.

6. GENERALIZATION OF METHOD

Finally, it should be noted that an atomistic model is easily applicable to nearly all types of molecular chemical systems and can be used to correlate and predict nearly all types of physicochemical properties and biological properties. In the Herndon group, we believe that the atomistic hierarchical methodology for QSAR and QSPR studies provides an optimum starting point for such studies.

REFERENCES AND NOTES

(1) Barret, G. C.; Elmore, D. T. Introduction. *Amino Acids and Peptides*, 1st ed.; Press syndicate of University of Cambridge: Cambridge, UK, 1998: pp 1−3.

(2) Zefirov, N. S.; Palyulin, V. A. QSAR for boiling points of "small" sulfides. Are the "high-quality-structure-property-activity regressions" the real high quality QSAR models. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1022–1027.

(3) Chalikian, T. V.; Sarvazyan, A. P.; Breslauer, K. J. Partial molar volumes, expansibilities, and compressibilities of α,ω-aminocarboxylic acids in aqueous solutions between 18 and 55 °C. *J. Phys. Chem.* **1993**, *97*, 13017–13026.

(4) Cohn, E. J.; McMeekin, T. L.; Edsall, J. T.; Weare, J. H. Studies in the physical chemistry of amino acids, peptides and related substances. II. The solubility of α-amino acids in water and in alcohol−water mixtures. *J. Am. Chem. Soc.* **1934**, *56*, 2270–2282.

(5) Greenstein, J. P.; Wyman, J.; Cohn, E. J. Studies of multivalent amino acids and peptides. III. The dielectric constants and electrostriction of the solvent in solutions of tetrapoles. *J. Am. Chem. Soc.* **1935**, *57*, 637–642.

(6) Rum, G.; Herndon, W. C. Three-Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids. *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Application*; Proceedings of the 10th European Symposium on Structure-Activity Relationships, QSAR and Molecular Modeling: Barcelona, Spain, Sept. 4−9, 1994.

(7) Häckel, M.; Hinz, H.; Hedwig, G. R. Partial molar volumes of proteins: amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10−90 °C. *Biophys. Chem.* **1999**, *82*, 35–50.

(8) Hakin, A. W.; Hedwig, G. R. Group additivity calculations of the thermodynamic properties of unfolded proteins in aqueous solution: a critical comparison of peptide-based and HFK models. *Biophys. Chem.* **2001**, *89*, 253–264.

(9) Hakin, A. W.; Kowalchuck, M. G.; Liu, J. L.; Marriott, R. A. Thermodynamics of protein model compounds: apparent and partial molar heat capacities and volumes of several cyclic dipeptides in water. *J. Solution Chem.* **2000**, *29*, 131–151.

(10) Hall, L. H.; Dailey, R. S.; Kier, L. B. Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: path 3. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 598–603.

(11) Hall, L. H.; Kier, L. B. Determination of topological equivalence in molecular graphs from the topological state. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115–131.

(12) Harano, Y.; Imai, T.; Kovalenko, A.; Kinoshita, M.; Hirata, F. Theoretical study for partial molar volume of amino acids and polypeptides by the three-dimensional reference site model. *J. Chem. Phys.* **2001**, *114*, 9506–9511.

(13) Herndon, W. C.; Radhakrishnan, T. P.; Zivkovic, T. P. Characteristic and matching polynomials of chemical graphs. *Chem. Phys. Lett.* **1998**, *152*, 233–238.

(14) Imai, T.; Kinoshita, M.; Hirata, F. Theoretical study for partial molar volume of amino acids in aqueous solution: Implication of ideal fluctuation volume. *J. Chem. Phys.* **2000**, *112*, 9469–9478.

(15) Amend, J. P.; Helgeson, H. C. Calculation of the standard molal thermodynamic properties of aqueous biomolecules at elevated temperatures and pressures II. Unfolded proteins. *Biophys. Chem.* **2000**, *84*, 105–136.

(16) Kharakoz, D. P. Partial volumes and compressibilities of extended polypeptide chains in aqueous solution: additivity scheme and implication of protein unfolding at normal and high pressure. *Biochemistry* **1997**, *36*, 10276–10285.

(17) Kier, L. B.; Hall, L. H. Deviation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583–589.

(18) JChem, version 5.2; ChemAxon: Budapest, Hungary, 2009.

(19) Makhatadze, G. I.; Medvedkin, V. N.; Privalov, P. L. Partial molar volumes of polypeptides and their constituent groups in aqueous solution over the broad temperature range. *Biopolymers* **1990**, *30*, 1001–1010.

(20) Matta, C. F.; Bader, R. F. W. Atoms-in-molecules study of the genetically encoded amino acids. II. Computational study of molecular geometries. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 519–538.

(21) Matta, C. F.; Bader, R. F. W. Atoms-in-molecules study of the genetically encoded amino acids III. Bond and atomic properties and their correlations with experiment including mutation-induced changes in protein stability and genetic coding. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 360–399.

(22) Mishra, A. K.; Ahluwalia, J. C. Apparent molal volumes of amino acids, N-acetylamino acids, and peptides in aqueous solutions. *J. Phys. Chem.* **1984**, *88*, 86–92.

(23) Pogliani, L. Molecular connectivity model for determination of physicochemical properties of alpha-amino acids. *J. Phys. Chem.* **1993**, *97*, 6731–6736.

(24) Pogliani, L. Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J. Phys. Chem.* **1996**, *100*, 18065–18077.

(25) Popelier, P. L. A.; Aicken, F. M. Atomic properties of selected biomolecules: quantum topological atom types of hydrogen, oxygen, nitrogen and sulfur occurring in natural amino acids and their derivatives. *Chem.-Eur. J.* **2003**, *9*, 1207–1216.

(26) Randic, M.; Mills, D.; Basak, S. C. On characterization of physical properties of amino acids. *Int. J. Quantum Chem.* **2000**, *80*, 1199–1209.

(27) Rao, M. V. R.; Atreyi, M.; Rajeswari, M. R. Partial molar volumes of α-amino acid with ionogenic side chains in water. *J. Phys. Chem.* **1984**, *88*, 3129–3131.

(28) Rellick, L. M.; Beckel, W. J. Comparison of van der Waals and semiempirical calculations of the molecular volumes of small molecules and proteins. *Biopolymers* **1997**, *42*, 191–202.

(29) Shahidi, F. A.; Farrell, P. G. Partial molar volumes of some α-aminocarboxylic acids in water. *J. Chem. Soc., Faraday Trans. I* **1981**, *77*, 963–968.

(30) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.

(31) Sotomatsu-Niwa, T.; Ogino, A. Evaluation of the hydrophobic parameters of the amino acid chains of peptides and their application in QSAR and conformational studies. *J. Mol. Struct.* **1997**, *392*, 43–54.

(32) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.

(33) Yasuda, Y.; Tochio, N.; Sakurai, M.; Nitta, K. Partial molar volumes and isentropic comprenssibilities of amino acids in dilute aqueous solutions. *J. Chem. Eng. Data* **1998**, *43*, 205–214.

(34) Wang, J.; Yan, Z.; Zhuo, K.; Lu, J. Partial molar volumes of some α-amino acid in aqueous sodium acetate solutions at 308.15 K. *Biophys. Chem.* **1999**, *80*, 179–188.

(35) Zefirov, N. S.; Palyulin, V. A. Fragmental approach in QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112–1122.

CI900318C