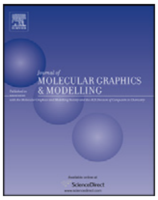
Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine

Chanin Nantasenamat [a], Chartchalerm Isarankura-Na-Ayudhya [a,*], Thanakorn Naenna [b], Virapong Prachayasittikul [a,*]

[a] Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Industrial Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand

ARTICLE INFO

Article history:
Received 10 January 2008
Received in revised form 8 April 2008
Accepted 8 April 2008
Available online 14 April 2008

Keywords:
Antioxidant
Bond dissociation enthalpy
Support vector machine
Phenols
QSPR

ABSTRACT

Antioxidants play crucial roles in scavenging oxidative damages arising from reactive oxygen species. Bond dissociation enthalpy (BDE) of phenolic O–H bond has well been accepted as an indicator of antioxidant activity since phenols donate the hydrogen atom to the free radicals thereby neutralizing its toxic effect. The BDEs from a data set of 39 antioxidant phenols were modeled using computationally inexpensive quantum chemical descriptors with multiple linear regression (MLR), partial least squares (PLS), and support vector machine (SVM). The molecular descriptors of the phenols were derived from calculations at the following theoretical levels: AM1, HF/3-21g(d), B3LYP/3-21g(d), and B3LYP/6-31g(d). Results indicated that when MLR and PLS were used as the regression methods, B3LYP/3-21g(d) gave the best performance with leave-one-out cross-validated correlation coefficients (*r*) of 0.917 and 0.921, respectively, while the semiempirical AM1 provided slightly lower *r* of 0.897 and 0.888, respectively. When SVM was used as the regression method no significant difference in the accuracy was observed for models using B3LYP/3-21g(d) and AM1 as indicated by *r* of 0.968 and 0.966, respectively. The quantitative structure–property relationship (QSPR) model of BDE discussed in this study offers great potential for the design of novel antioxidant phenols with robust properties.

1. Introduction

Reactive oxygen species (ROS) are natural byproducts of normal aerobic metabolism which include oxygen ions, free radicals, and peroxides [1]. Cells are normally safeguarded from the deleterious effects of ROS by antioxidant enzymes (e.g. superoxide dismutases and catalases) and small antioxidant molecules (e.g. ascorbic acid, vitamin E, and glutathione) [2]. A delicate balance in the production and elimination of ROS is crucial to normal cell and tissue function. However, perturbation to this equilibrium by changes in environmental factors give rise to a condition known as oxidative stress [3]. This oxidative damage affects various biological macromolecules such as DNA, RNA, proteins, and membrane lipids. The highly reactive nature of ROS stems from the presence of unpaired valence shell electrons, which can easily accept or transfer an electron.

Antioxidants are substances that play important roles in combating oxidative damages that are caused by ROS [2]. Antioxidants neutralize ROS by intercepting and interacting with reactive radicals as summarized as follows:

$$RO_2^{\bullet} + AOH \rightarrow ROOH + AO^{\bullet} \qquad (1)$$

where $RO_2^{\bullet}$ is the peroxy radical, AOH is the antioxidant, ROOH is the lipid hydroperoxide, and $AO^{\bullet}$ is the phenoxy radical.

Many naturally occurring antioxidants found thus far are based on phenolic compounds. Therefore, it is of great interest to be able to characterize the radical scavenging activities of these phenolic antioxidants. The use of O–H bond dissociation enthalpies (BDEs) in the characterization of phenolic antioxidants has well been documented [4–7]. Phenolic antioxidants with low BDEs are considered to have good antioxidative activity since they readily donate the hydrogen atom from the O–H bond to incoming ROS radicals. This is further supported by the fact that O–H BDE has been found to be well correlated with the activation energy of the H-abstraction reaction [8,9]. On the basis of these principles, O–H BDE has been widely used as a measure of the efficiency of radical scavenging activity [4–7,10,11]. The measurement of O–H BDE has

[*] Corresponding authors. Tel.: +66 2 418 0227; fax: +66 2 412 4110.
E-mail addresses: mtcis@mahidol.ac.th (C. Isarankura-Na-Ayudhya), mtvpr@mahidol.ac.th (V. Prachayasittikul).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.04.005

traditionally been made in solutions (such as in water or DMSO) [12] and subsequently performed in the gas phase [6,7]. The phenolic O–H BDE values has been found to differ significantly from different experimental studies ranging from 84.0 to 91.6 kcal mol^{-1} [13]. Such diversity in O–H BDE values implies the uncertainty of even experimentally derived values. Therefore, there is sufficient room for improving upon the calculation methodologies of O–H BDE.

Two theoretical approaches exist for the calculation of O–H BDE values. The first approach involves two series of calculations, one for each of the two forms of the antioxidant phenols: (1) the neutral state and (2) their respective radical states. To achieve accurate calculation of thermochemical properties, multideterminantal methods that thoroughly account for the effect of electron correlation or multilevel methods entailing numerous sequential high-level calculations coupled with energy corrections are often required [14–17]. The former could be obtained from coupled clustered [CCSD(T)] [18] or quadratic configuration [QCISD(T)] [14] methods in combination with very large basis sets. The calculated results are then extrapolated to the complete basis set (CBS) limit. However, such approach is limited to only small molecule due to its computationally intensive nature. In order to study larger molecular systems, an alternative approach such as the Gaussian-*n* series (e.g. Gaussian-3 and the latest Gaussian-4 theory) [19,20] and the CBS methods [21] are typically used. Gaussian-*n* methods are composite approaches that aim to approximate the result of a more expensive calculation by performing a set of high-level correlation calculations [CCSD(T), MP4, and QCISD(T)] with moderate sized basis sets. Likewise, CBS methods are sophisticated energy computations involving several pre-defined calculations for the purpose of generating very accurate energies [17,22–24].

However, such methods can be time-consuming which contradicts with our objective of a quick and inexpensive approach towards predicting the O–H BDE. Confronted with similar computational constraints of achieving accurate C–H BDEs at the expense of high computational cost, Lewin and Cramer tackled the problem by developing rapid quantum mechanical models for the accurate estimation of C–H BDEs [25]. To meet this challenge, this study utilizes quantitative structure–property relationship (QSPR) methods for correlating the physicochemical properties of phenolic antioxidants with their respective BDEs.

Herein, the application of computationally inexpensive quantum chemical descriptors in combination with robust supervised statistical and machine learning approaches was demonstrated for the development of QSPR models. To achieve that goal, the search for suitable theoretical level in calculating the molecular descriptors and the selection of reliable supervised learning methods for constructing the QSPR model was performed. Quantum chemical descriptors were derived from geometry optimizations at the following theoretical levels: (i) semiempirical AM1 method and at the *ab initio* levels, (ii) HF/3-21g(d), (iii) B3LYP/3-21g(d), and (iv) B3LYP/6-31g(d). Three supervised learning approaches, namely (i) multiple linear regression (MLR), (ii) partial least squares (PLS) regression, and (iii) support vector machine (SVM), were used for correlating the structures of antioxidant phenols with the hydroxyl BDEs. Two models, particularly descriptors derived from AM1 or B3LYP/3-21g(d) calculations, were identified to provide equally good predictive performance with support vector machine in an economical and efficient manner.

2. Methodology

2.1. Data set

The BDEs of 39 antioxidant phenols (Table 1) were obtained from the work of Bordwell and Cheng [26]. The BDEs of phenolic O–H bonds were estimated from the oxidation potentials and pK_{HA} values of phenoxide and naphthoxide ions in DMSO. The initial geometries of the antioxidant phenols were constructed using GaussView [27].

Table 1
Data set of the antioxidant phenols

No.	Name	CAS No.	Substituent	ΔBDE (kcal/mol)
1	Phenol	108-95-2	H	0
2	*o*-Cresol	95-48-7	2-Me	−1.65
3	*m*-Cresol	108-39-4	3-Me	−0.45
4	*p*-Cresol	106-44-5	4-Me	−1.15
5	3,5-Dimethylphenol	108-68-9	3,5-Me_2	−0.75
6	2,6-Dimethylphenol	576-26-1	2,6-Me_2	−4.35
7	4-*tert*-Butylphenol	98-54-4	4-*t*-Bu	−1.15
8	2,6-di-*tert*-Butylphenol	128-39-2	2,6-*t*-Bu_2	−7.75
9	2,4,6-tri-*tert*-Butylphenol	732-26-3	2,4,6-*t*-Bu_3	−7.65
10	4-Phenylphenol	92-69-3	4-Ph	−2.25
11	2-Methoxyphenol	90-05-1	2-MeO	−3.85
12	3-Methoxyphenol	150-19-6	3-MeO	0.35
13	Hydroquinone	123-31-9	4-OH	−8.35
14	3-Aminophenol	591-27-5	3-NH_2	−1.85
15	3-Dimethylaminophenol	99-07-0	3-Me_2N	−1.95
16	4-Aminophenol	123-30-8	4-NH_2	−12.55
17	4-Dimethylaminophenol	619-60-3	4-NMe_2	−9.55
18	2-Chlorophenol	95-57-8	2-Cl	0.15
19	3-Chlorophenol	108-43-0	3-Cl	1.95
20	4-Chlorophenol	106-48-9	4-Cl	0.45
21	3,5-Dichlorophenol	591-35-5	3,5-Cl_2	4.05
22	3,4,5-Trichlorophenol	609-19-8	3,4,5-Cl_3	3.25
23	4-Bromophenol	106-41-2	4-Br	0.85
24	*m*-Trifluoromethylphenol	98-17-9	3-CF_3	3.95
25	*p*-Trifluoromethylphenol	402-45-9	4-CF_3	5.45
26	*m*-Hydroxyacetophenone	121-71-1	3-MeCO	1.95
27	*p*-Hydroxyacetophenone	99-93-4	4-MeCO	2.95
28	3-Nitrophenol	554-84-7	3-NO_2	4.45
29	4-Nitrophenol	100-02-7	4-NO_2	4.85
30	4-Methoxyphenol	150-76-5	4-OMe	−5.25
31	4-Hydroxybenzophenone	1137-42-4	4-PhCO	2.65
32	3-Methylsulfonylphenol	14763-61-2	3-$MeSO_2$	2.45
33	4-Methylsulfonylphenol	14763-60-1	4-$MeSO_2$	5.15
34	3-Cyanophenol	873-62-1	3-CN	4.05
35	4-Cyanophenol	767-00-0	4-CN	4.35
36	1-Naphthol	90-15-3	1-NpOH	−5.85
37	4-Hydroxyphenolate	20217-26-9	4-O^-	−16.85
38	2-Naphthol	135-19-3	2-NpOH	−1.85
39	6-Bromo-2-naphthol	15231-91-1	6-Br-2-NpOH	−1.35

2.2. Descriptor generation

The molecular structures of antioxidant phenols were subjected to full geometry optimizations without symmetry constraints at the semiempirical level using the Austin Model 1 (AM1) method. Likewise, calculations at the *ab initio* levels were performed with Gaussian 03 [28] using the Hartree–Fock functional with the 3-21g(d) basis set (HF/3-21g(d)) and Becke's three-parameter hybrid Lee–Yang–Parr functional with the 3-21g(d) basis set (B3LYP/3-21g(d)) and with the 6-31g(d) basis set (B3LYP/6-31g(d)). The following quantum chemical descriptors were derived from these calculations: total energy (E_{Total}), dipole moment (μ), energy of the highest occupied molecular orbital (E_{HOMO}; negative value of E_{HOMO} was used as a measure of the ionization potential), energy of the lowest unoccupied molecular orbital (E_{LUMO}; negative value of E_{LUMO} was used as an indicator of the electron affinity), atomic charge on hydroxyl hydrogen (q_H), atomic charge on hydroxyl oxygen (q_O), and lengths of O–H bond (R_{O-H}). The atomic charges were derived from Mulliken population analysis. Ionization potential and electron affinity were calculated according to Koopmans' theorem [29].

2.3. Data pre-processing

Since the independent variables were of different range, they were adjusted to comparable scale by standardization using the following equation:

$$x_{ij}^{\sin} = \frac{x_{ij} - \bar{x}_j}{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2/N} \tag{2}$$

where $x_{ij}^{\sin}$ represents the standardized value, x_{ij} represents the value of each sample, $\bar{x}_j$ represents the mean of each descriptor, and N represents the sample size of the data set.

2.4. Multivariate regression

Multivariate regression was performed to correlate the independent variables, which in our case are the quantum chemical descriptors, with the dependent variable BDE. Three supervised learning approaches, namely MLR, PLS, and SVM, were used for correlating the structures of antioxidant phenols with the hydroxyl BDEs.

2.4.1. Multiple linear regression

MLR models were calculated using The Unscrambler 9.5 [30] software package to obtain equations of the following form:

$$Y = B_0 + \sum B_n X_n \tag{3}$$

where Y represents the BDEs of the antioxidant phenol compounds, B_0 represents the intercept, B_n represents the regression coefficients of descriptors X_n.

2.4.2. Partial least squares regression

PLS analysis [31,32] was also performed using The Unscrambler 9.5 software package with the PLS1 algorithm. The descriptors were subjected to pre-processing by mean-centering and auto-scaling to zero mean and unit variance according to Eq. (2). The amount of variables presented in the descriptor matrix was reduced to a small number of latent variables called PLS components (PCs), which retain the core information from the original data set. The PCs are few, orthogonal, and function as predictors of the dependent variable. The optimal number of PCs was determined according to the method of Haaland and Thomas [33] from a plot of PC versus the mean squared error (MSE) using leave-one-out cross-validation (LOO-CV). MSE was calculated using the following equation:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N}(p_i - a_i)^2}{n} \tag{4}$$

where p_i represents the predicted output, a_i represents the experimental value, and n represents the number of antioxidant phenols presented in the data set.

2.4.3. Support vector machine

SVM calculations were performed using John Platt's Sequential Minimal Optimization (SMO) algorithm [34] with the Waikato Environment for Knowledge Analysis (Weka) version 3.4.7 [35]. SVM is a learning approach developed by Vapnik and co-workers based on the Statistical Learning Theory [36,37]. The technique has demonstrated much success in modeling biological and chemical properties as demonstrated in literatures [38–40]. Detailed accounts of SVM theory can be found in several excellent books and tutorials [36,41–46]. Here we briefly delve into the main concepts of SVM.

A training set of m compounds with known biological activity y_i (e.g. O–H BDE) and structurally derived descriptors x_i are represented as $\{(x_i, y_i)\}_{i=1}^{m}$, where correlations between structure and activities are defined by $y_i = f(x_i)$. The term $f(x_i)$ can be represented by a linear function of the form

$$f(x_i) = \langle w_i, x_i \rangle + b \tag{5}$$

where w designates the weight vector of the linear function and b corresponds to the threshold coefficient. SVM approximates the set of data with a linear function that is formulated in the high dimensional feature space with the following function:

$$y = \sum_{i=1}^{m} w_i \Phi(x_i) + b \tag{6}$$

where $\{\Phi(x_i)\}_{i=1}^{m}$ represents the features of input variables subjected to kernel transformation while $\{w_i\}_{i=1}^{m}$ and b are coefficients.

SVM is essentially a linear learning approach that was originally devised for classification problems. However, it is also amenable to regression problems through the use of ε-insensitive loss function. SVM can handle data possessing non-linear relationships via the so-called kernel trick. Kernel transformation is essentially a projection of the descriptor matrix from the input space into the higher dimensional feature space (Fig. 1). This can be described by the following equation:

$$K(x, y) = \langle \phi(x) \cdot \phi(y) \rangle \tag{7}$$

where K is a kernel function and ϕ is a mapping from input space $X \in x$, y to the feature space F.

Several kernel functions are available for non-linear transformation of the input space. Popular kernel functions used in SVM include the variance–covariance based linear and polynomial

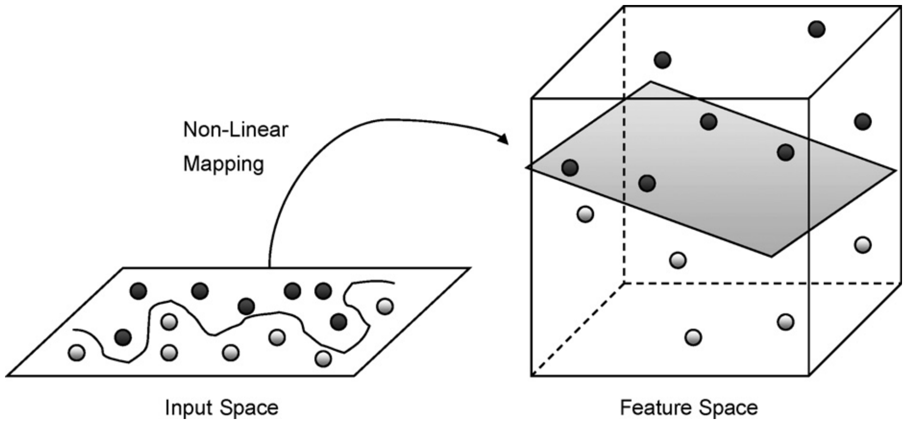


Fig. 1. Schematic of non-linear mapping of input space onto higher dimensional feature space by kernel transformation.

kernels and the Euclidean distance based radial basis function kernels.

Radial basis function kernel was used to perform this non-linear mapping as described by the following equation:

$$K(x, y) = \exp(-\gamma\|x - y\|^2) \tag{8}$$

After kernel transformation, the new feature space allows the data to be linearly separable by hyperplanes where hyperplane that maximizes the distance between the data samples was selected by the algorithm as the maximal hyperplane.

Minimization of the regularized risk function satisfies two essential properties of SVMs by means of estimating coefficients w and b: (i) define regression estimation by performing risk minimization with respect to the ε-insensitive loss function and (ii) perform risk minimization based on the SRM principle in which elements of the structure are defined by the inequality $\|w\|^2 \leq$ constant.

The regularized risk function is defined as

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(y_i, f(x_i, w)) + \frac{1}{2}\|w\|^2 \tag{9}$$

where $(C/N)\sum_{i=1}^{N} L_\varepsilon(y_i, f(x_i, w))$ is the empirical error (risk) and $(1/2)\|w\|^2$ is a measure of function flatness. The empirical error is measured by the ε-insensitive loss function $L_\varepsilon(y, f(x, w))$ in which errors below ε would not be penalized. The penalty parameter C is a regularized constant responsible for determining the trade-off between the empirical error and the model complexity.

The estimation performance of SVM regression models is determined by the ε-insensitive loss function as follows:

$$L_\varepsilon(y, f(x, w)) = \begin{cases} |y - f(x, w)| - \varepsilon & \text{for} |y - f(x, w)| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The parameter ε is referred to as the tube size, and it is defined as the approximation accuracy placed on the training data points. Essentially, the goal of support vector regression is to select a function $f(x)$ such that there is at most ε deviation from the actual value y_i for all training data while being as flat as possible. In other words, the loss function ignores errors as long as it is less than ε but would accept no significant deviation from it.

The generalization performance of the SVM model relies on the proper selection of parameters. For RBF kernels, there are two parameters involved: the complexity parameter C and the RBF kernel width γ. As the parameter C and gamma are not universally optimal to all problems an empirical parameter search is required. Therefore, determination of the optimal configuration of the SVM model was performed via a two-level grid search [47,48]. Initially, a coarse grid search was performed by exponentially adjusting the values of (C, γ). Regions affording good performance are identified and then subjected to further refinement via a local grid search.

2.5. Variable selection

MLR regression coefficients were used as a measure of the variable's importance toward the regression model. Particularly, the magnitude of the regression coefficients indicates the relative perturbation each independent variable has on the dependent variable.

2.6. Internal validation procedure

LOO-CV is an internal validation procedure used to provide estimates of the predictivity of the data sets in a cost-effective manner as all data samples were used in model development [35]. LOO-CV initiates by leaving one data sample out as the testing set while using $n - 1$ samples as the training set. This procedure was carried out iteratively until all data samples were given the chance to be left out as the testing set.

2.7. Statistical analysis

The adjusted value of R^2 accounts for the number of predictors (independent variables) in the model. It is calculated according to the following equation:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \cdot \left(\frac{n - 1}{n - p}\right) \tag{11}$$

where n represents sample size and p represents the number of predictors.

The F ratio between explained (R^2) and unexplained ($1 - R^2$) variance with m and $n - m - 1$ degrees of freedom is calculated according to the following equation:

$$F_{(m,n-m-1)} = \frac{R^2/m}{(1 - R^2)/(n - m - 1)} \tag{12}$$

where m is the number of independent variables and n is the number of compounds presented in the data set.

3. Results and discussion

The present investigation employs readily available quantum chemical descriptors in the construction of QSPR models that reliably calculates the bond dissociation enthalpies of antioxidant phenols. To achieve robust model, series of calculations were performed by looking into the performance difference as afforded by descriptors calculated at various theoretical levels followed by regression analysis with different multivariate methods. The molecular structures of the antioxidant phenols were drawn into the computer according to the structures reported by Bordwell and Cheng [26]. The substituent effect arising from variations of the functional moieties of a library of phenolic compounds was accounted for through the use of quantum chemical descriptors. Essentially, these independent variables provided quantitative description of the molecular features as well as information of their distinct functional variations. Multivariate analysis was then performed using these independent variables as input for the development of predictive models of BDEs. Prior to performing regression studies, the data was standardized according to Eq. (2) as to scale the values to comparable levels where the mean and standard deviation were adjusted to 0 and 1, respectively.

Evidently, the quality of the descriptors as well as the type of multivariate approach used for constructing the QSPR model exerts great influence on its predictive performance. Our previous investigations have demonstrated the usefulness of quantum chemical descriptors in modeling the spectral properties of the green fluorescent protein [49] and the imprinting factor of molecularly imprinted polymers [50]. The optimal theoretical level required for generating molecular descriptors that provide good predictivity was examined at the following theoretical approaches: AM1, HF/3-21g(d), B3LYP/3-21g(d), and B3LYP/6-31g(d). Of the four theoretical levels tested, it was observed that AM1 and B3LYP/3-21g(d) exhibited better accuracy than the other theoretical levels as observed from r of 0.814 and 0.895, respectively (Table 2). Therefore, these two theoretical approaches were selected for further investigations.

It was noticed that the B3LYP functional in combination with the smaller basis set 3-21g(d) unexpectedly outperformed that of the higher basis set 6-31g(d). This technical anomaly could be attributed to possible error that may arise as a result of calculations

Table 2
Predictive performance of MLR models using all compounds of the data set

Descriptors	r_{Training}	$r_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV(adj)}}$	F ratio[a]	$\text{RMS}_{\text{Training}}$	$\text{RMS}_{\text{LOO-CV}}$
AM1	0.932	0.814	0.663	0.633	16.201	1.834	4.687
HF/3-21g(d)	0.889	0.761	0.579	0.542	11.352	2.316	4.624
B3LYP/3-21g(d)	0.950	0.895	0.801	0.783	33.212	1.583	2.743
B3LYP/6-31g(d)	0.914	0.730	0.533	0.492	9.412	2.059	4.773

[a] Critical F value with 4 and 33 degrees of freedom is 2.659.

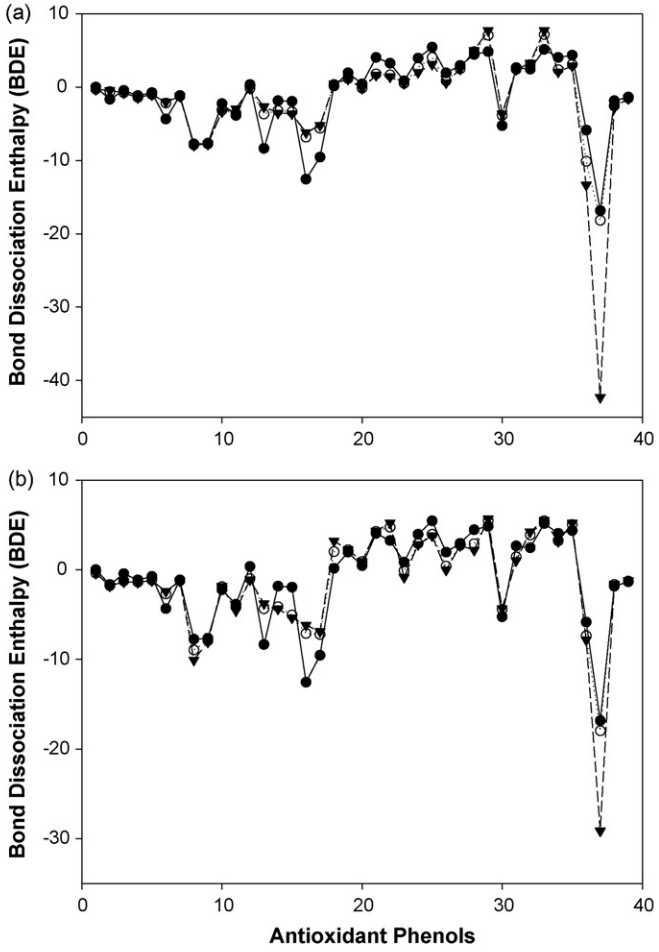


Fig. 2. Plot of the phenolic antioxidant compounds along with the experimental BDE (●), and the predicted BDE of the training (○) and cross-validated testing (▼) set. Compound 37 was identified as outlier and removed from QSPR models using molecular descriptors derived from AM1 (a) and B3LYP/3-21g(d) (b) calculations.

with the B3LYP functional. Considerable evidences in the literature strongly suggest the inability of B3LYP to accurately calculate the values of R–X BDEs as well as its failure to model hydrogen abstraction reactions [51–54]. The B3LYP functional generally suffers from systematic underestimation of the classical barrier height. Guner et al. observed similar unusual effect of performance degradation as a function of increasing basis set size in their study of hydrocarbon pericyclic reactions using the B3LYP functional [55,56].

In order to enhance the predictivity of the QSPR model, closer examination of the results indicated that one of the compound exerted great influence on the overall performance. Particularly, compound **37** was identified by The Unscrambler software package to be an outlier as also illustrated in Fig. 2 where the predicted value of compound **37** is shown to greatly differ from the experimental value. Removal of this outlying sample improved the predictive performance as illustrated in Table 3 by the boost of r from 0.895 to 0.903 for MLR model using descriptors calculated at the B3LYP/3-21g(d) level. Likewise, similar improvements in the predictivity were also observed for descriptors generated at the AM1 level where r rose from 0.814 to 0.874. Analogously, Sun et al. compared the effectiveness of molecular geometries of phenolic antioxidants calculated at the semiempirical AM1 with density functional B3LYP levels and found that AM1 geometry gave comparable accuracy with that derived from B3LYP calculations [57]. Those findings are in line with our results as AM1 shows comparable degree of performance with that of B3LYP/3-21g(d) descriptors.

To further refine the QSPR models, independent variables possessing low regression coefficient values were removed as they exert minimal perturbation to the dependent variable (Fig. 3). Such variables subjected for removal in QSPR model using B3LYP/3-21g(d) derived descriptors included E_{Total}, μ, and q_{O}, while the remaining descriptors, comprising of IP, EA, q_{H}, and $R_{\text{O–H}}$, were used for further investigations. This further increased the predictive accuracy to r of 0.917. Likewise, the performance for QSPR model using descriptors generated from AM1 calculations rose to 0.897 upon removal of E_{Total}, μ, and EA.

Since the objective of this study is to calculate the dissociation enthalpies of O–H bonds as a measure of their relative

Table 3
Summary of the predictive performance of the initial and refined MLR models

Descriptors	r_{Training}	$r_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV(adj)}}$	F ratio	$\text{RMS}_{\text{Training}}$	$\text{RMS}_{\text{LOO-CV}}$
AM1[a]	0.932	0.814	0.663	0.597	8.697[b]	1.834	4.687
AM1[c]	0.932	0.874	0.764	0.718	13.865[d]	1.604	2.188
AM1[e]	0.923	0.897	0.805	0.787	33.973[f]	1.637	1.974
B3LYP/3-21g(d)[a]	0.950	0.895	0.801	0.763	17.828[b]	1.583	2.743
B3LYP/3-21g(d)[c]	0.942	0.903	0.815	0.780	18.932[d]	1.486	1.912
B3LYP/3-21g(d)[e]	0.939	0.917	0.841	0.827	43.601[f]	1.524	1.777

[a] Initial data set.
[b] Critical F value at the 95% confidence level with 4 and 33 degrees of freedom is 2.659.
[c] Remove compound 37 as outlier.
[d] Critical F value at the 95% confidence level with 7 and 30 degrees of freedom is 2.334.
[e] Exclude three descriptors from QSPR model.
[f] Critical F value at the 95% confidence level with 7 and 31 degrees of freedom is 2.323.

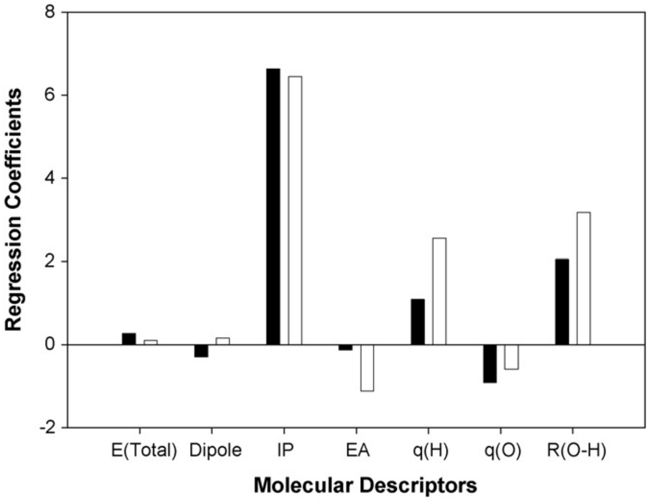


Fig. 3. Regression coefficient values of molecular descriptors derived from AM1 (■) and B3LYP/3-21g(d) (□) calculations.

antioxidative activity, therefore, it is worthy to observe whether the lengths of O–H bonds are correlated with O–H BDEs or not. Table 4 summarizes the O–H bond lengths of the phenolic antioxidants geometrically optimized at various theoretical levels. Results indicated that there is a general positive correlation of O–H bond length with their respective O–H BDE where AM1 gave the best performance of the four theoretical levels tested. Particularly, phenolic antioxidant with longer O–H bond indicates that there is weaker attraction of the hydrogen atom toward the oxygen atom. As such the hydroxyl group easily loses the hydrogen atom to ROS radicals as also indicated by its lower BDE values. It is also interesting to note that R_{O-H} was retained as molecular descriptor in both QSPR models: (1) model using AM1 descriptor and (2) model using B3LYP/3-21g(d) descriptor.

Next, the most suitable multivariate regression approach for this endeavor was investigated using two other methods besides MLR, which included PLS and SVM. It was observed that the use of molecular descriptors derived from calculation at the B3LYP/3-21g(d) level yield good predictive performance with *r* in excess of 0.9. Particularly, the use of typical regression methods, such as MLR and PLS, resulted in *r* of 0.917 and 0.921, respectively. On the other hand, models using descriptors generated at the semiempirical AM1 level provided relatively good precision but to a lesser extent than those calculated at the density functional level. This is illustrated by *r* of 0.897 and 0.888 when regression analysis was performed using MLR and PLS, respectively.

SVM is a supervised learning method capable of solving many complex classification problems. SVM has demonstrated success in many perplexing biological and chemical problems, such as the classification of protein folds [58] and secondary structures [59], prediction of DNA splice junction sites [60], and classification of the activity/inactivity of potential lead compounds [61]. In this

Table 4
Summary of O–H bond lengths (R_{O-H}) in phenolic antioxidants geometrically optimized at different levels of theory

Cpd. no.	AM1	HF/3-21g(d)	B3LYP/3-21g(d)	B3LYP/6-31g(d)	ΔBDE (kcal/mol)
16	0.9672	0.9639	0.9919	0.9691	−12.55
17	0.9675	0.9639	0.9921	0.9693	−9.55
13	0.9678	0.9640	0.9920	0.9694	−8.35
8	0.9655	0.9660	0.9859	0.9645	−7.75
9	0.9657	0.9660	0.9873	0.9644	−7.65
36	0.9660	0.9606	0.9895	0.9679	−5.85
30	0.9677	0.9639	0.9920	0.9693	−5.25
6	0.9678	0.9621	0.9904	0.9687	−4.35
11	0.9688	0.9640	0.9923	0.9697	−3.85
10	0.9683	0.9642	0.9924	0.9699	−2.25
15	0.9679	0.9640	0.9920	0.9695	−1.95
14	0.9680	0.9641	0.9920	0.9695	−1.85
38	0.9683	0.9644	0.9925	0.9701	−1.85
2	0.9681	0.9638	0.9916	0.9694	−1.65
39	0.9686	0.9645	0.9925	0.9702	−1.35
4	0.9680	0.9642	0.9923	0.9697	−1.15
7	0.9680	0.9642	0.9924	0.9697	−1.15
5	0.9679	0.9642	0.9924	0.9698	−0.75
3	0.9680	0.9641	0.9922	0.9697	−0.45
1	0.9680	0.9642	0.9923	0.9698	0
18	0.9689	0.9641	0.9921	0.9699	0.15
12	0.9683	0.9641	0.9921	0.9695	0.35
20	0.9684	0.9643	0.9922	0.9698	0.45
23	0.9686	0.9643	0.9922	0.9698	0.85
19	0.9684	0.9643	0.9922	0.9699	1.95
26	0.9682	0.9642	0.9924	0.9698	1.95
32	0.9684	0.9643	0.9923	0.9700	2.45
31	0.9687	0.9645	0.9927	0.9704	2.65
27	0.9688	0.9644	0.9926	0.9703	2.95
22	0.9691	0.9645	0.9922	0.9701	3.25
24	0.9686	0.9642	0.9922	0.9698	3.95
21	0.9688	0.9645	0.9922	0.9701	4.05
34	0.9686	0.9644	0.9923	0.9701	4.05
35	0.9689	0.9645	0.9925	0.9703	4.35
28	0.9689	0.9642	0.9921	0.9700	4.45
29	0.9697	0.9647	0.9925	0.9705	4.85
33	0.9697	0.9645	0.9925	0.9703	5.15
25	0.9691	0.9645	0.9924	0.9701	5.45
r with ΔBDE	0.778	0.181	0.503	0.596	

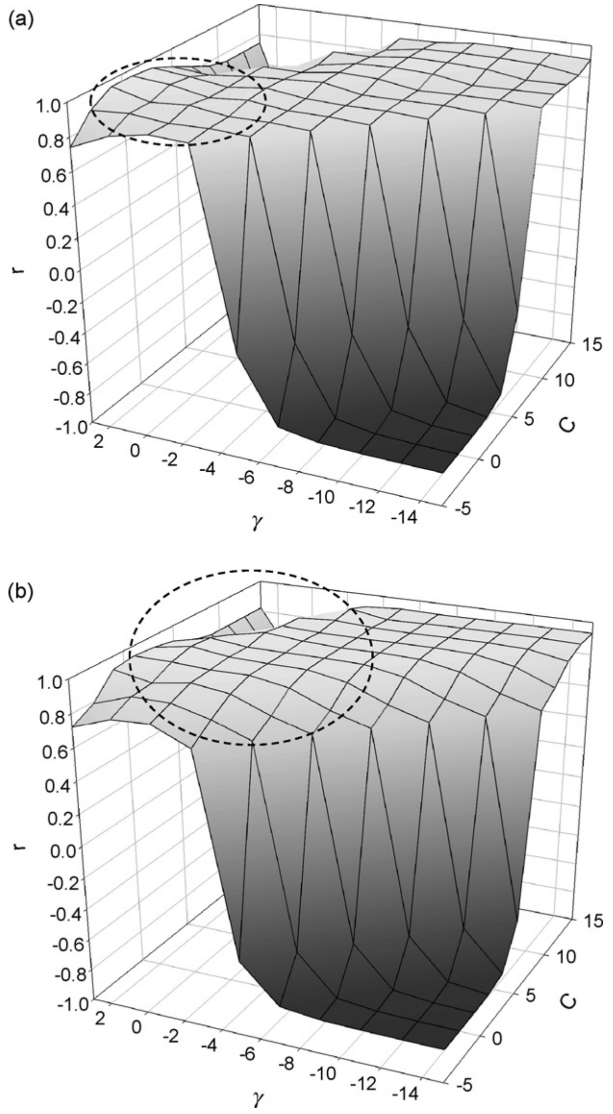


Fig. 4. Three-dimensional mesh plot of the coarse grid search of the SVM parameters for QSPR models using AM1 (a) and B3LYP/3-21g(d) (b) descriptors. The correlation coefficient (r) was plotted as a function of the SVM parameters, C and γ. Regions giving rise to good prediction accuracy is shown in the circled area.

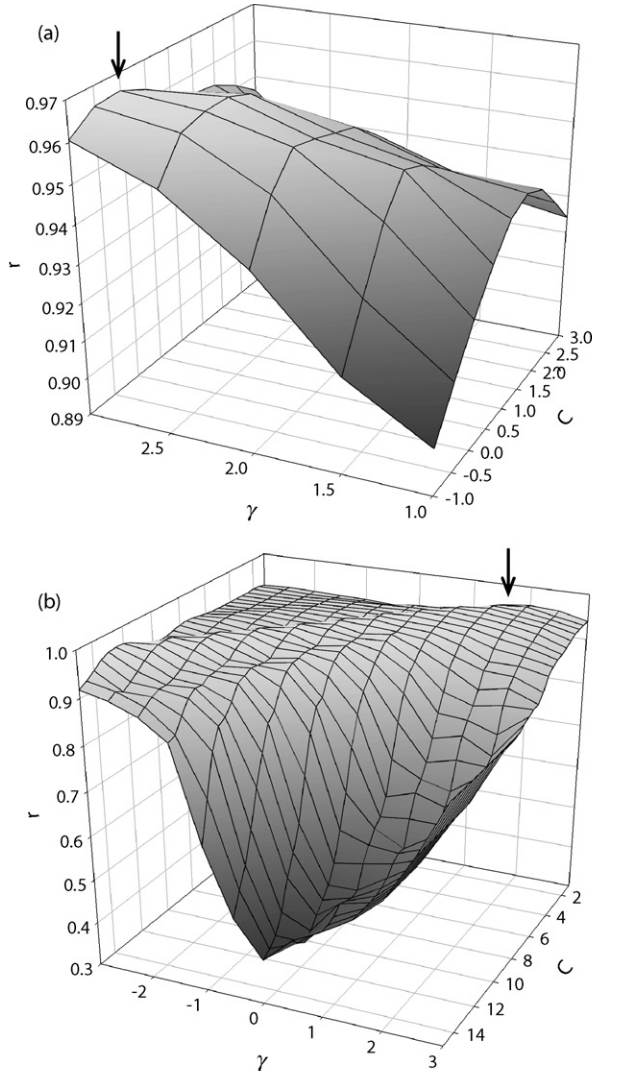


Fig. 5. Three-dimensional mesh plot of the local grid search of the SVM parameters for QSPR models using AM1 (a) and B3LYP/3-21g(d) (b) descriptors. The correlation coefficient (r) is plotted as a function of the SVM parameters, C and γ. Regions giving rise to good prediction accuracy is indicated by the arrow.

study, SVM was employed for correlating the physicochemical properties of the antioxidant phenols with their BDEs. Since SVM is a linear learning machine, non-linear data must be represented in a way that allows the algorithm to classify them in the new feature space [36]. This is achieved by first converting the descriptor matrix onto a richer feature space while conserving the original information via radial basis function kernel transformation.

To achieve an optimal set of parameters for the construction of robust SVM models, two sequential stepwise parameter searches were performed. An initial global grid search for the optimal SVM parameters, C and γ, was carried out to identify regions affording good accuracy. In this parameter search, the exponential n value of 2^n for C was varied from -5 to 15 by increments of 2 while γ was varied from -15 to 3 by increments of 2. As illustrated in Fig. 4, the optimal region was identified as indicated within the circled area. The selected region was then subjected to a more thorough search in order to locate the optimal set of parameters. Results from Fig. 5 indicated that the optimal value of C and γ was 2^0 and 2^3, respectively, for models using descriptors generated at the AM1

Table 5
Summary of the predictive performance of the various QSPR models

Descriptors	Multivariate method	r_{Training}	$r_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV}}$	$R^2_{\text{LOO-CV(adj)}}$	F ratio[a]	$\text{RMS}_{\text{Training}}$	$\text{RMS}_{\text{LOO-CV}}$
B3LYP/3-21g(d)	MLR	0.939	0.917	0.841	0.827	43.601	1.524	1.777
B3LYP/3-21g(d)	PLS	0.933	0.921	0.848	0.835	46.113	2.016	2.175
B3LYP/3-21g(d)	SVM	0.987	0.968	0.937	0.931	122.752	0.752	1.122
AM1	MLR	0.929	0.897	0.805	0.787	33.973	1.637	1.974
AM1	PLS	0.915	0.888	0.789	0.770	30.765	2.243	2.491
AM1	SVM	0.984	0.966	0.933	0.927	115.172	0.813	1.247

[a] Critical F value at the 95% confidence level with 4 and 33 degrees of freedom is 2.659.

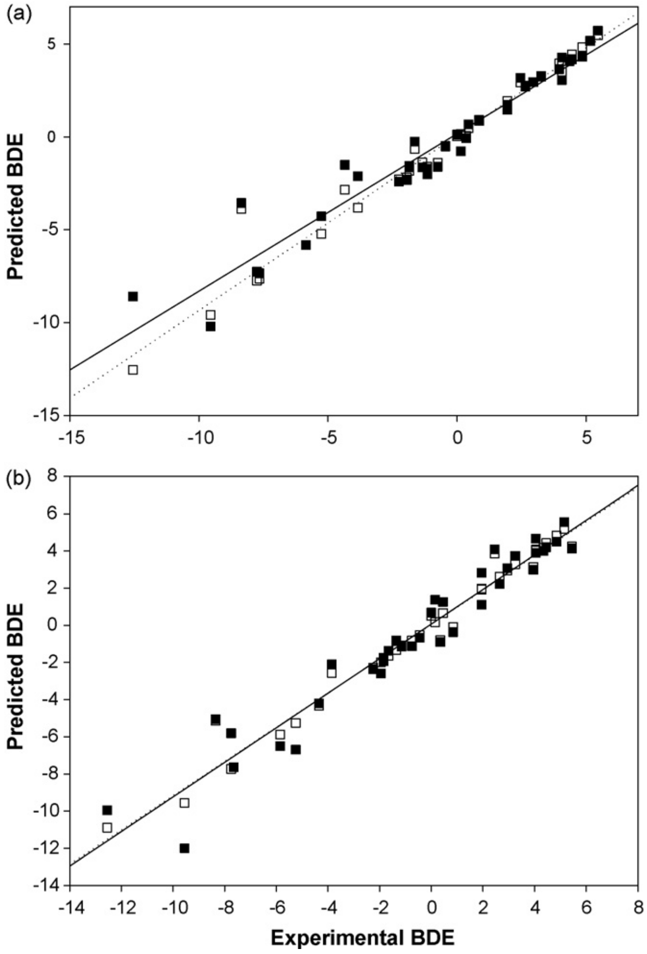


Fig. 6. Plot of the predicted BDE versus the experimental BDE for the training set (□; regression line is represented as dotted line) and testing set (■; regression line is represented as solid line).

level. Furthermore, it can also be seen that the optimal C and γ was $2^{2.5}$ and $2^{1.5}$, respectively, for models using B3LYP/3-21g(d) descriptors.

A predictive model was then constructed using the empirically derived parameters. The experimental BDEs of the antioxidant phenols were well correlated with the predicted BDEs as presented in Fig. 6. From this plot it can be seen that the SVM models using descriptors generated from the semiempirical AM1 or the density functional B3LYP method in combination with the 3-21g(d) basis set afforded similar level of high predictivity as judged by r of 0.966 and 0.968, respectively (Table 5). Furthermore, the observed F values for all models are much greater (30.765–122.752) than the critical F value (2.659) at the 95% confidence level with 4 and 33 degrees of freedom. Therefore, we can reject the null hypothesis and conclude that the independent variables correlate well with the O–H BDE values.

4. Conclusion

This study proposes the combined use of quantum chemical descriptors with supervised learning methods for the development of robust QSPR models of antioxidant phenol BDEs. The substituent effects arising from variation of the functional groups at various positions of the phenol ring could be accounted for by the QSPR models described in this study. The predicted BDEs of antioxidant phenols were found to be in good agreement with the experimental values. Of the three multivariate regression methods used in this study, SVM was found to outperform the traditional regression methods such as MLR and PLS. The methodology proposed in this study facilitates rapid evaluation of antioxidant properties using readily available quantum chemical descriptors that are inexpensive to calculate in combination with robust modeling methods such as SVM.

Acknowledgements

This research is financially supported in part by the Young Research Scholar grant (no. MRG5080450) from The Thailand Research Fund to C.N. and the annual budget of Mahidol University (B.E. 2551).

References

[1] T.E. Andreoli, Free radicals and oxidative stress, Am. J. Med. 108 (2000) 650–651.

[2] P.G. Winyard, C.J. Moody, C. Jacob, Oxidative activation of antioxidant defence, Trends Biochem. Sci. 30 (2005) 453–461.

[3] J.M. McCord, The evolution of free radicals and oxidative stress, Am. J. Med. 108 (2000) 652–659.

[4] N.M. Katarina, Mechanistic studies of phenolic antioxidants in reaction with nitrogen- and oxygen-centered radicals, J. Mol. Struct. (Theochem.) 818 (2007) 141–150.

[5] K.M. Nikolic, Theoretical study of phenolic antioxidants properties in reaction with oxygen-centered radicals, J. Mol. Struct. (Theochem.) 774 (2006) 95–105.

[6] C. Giacomelli, S. Miranda Fda, N.S. Goncalves, A. Spinelli, Antioxidant activity of phenolic and related compounds: a density functional theory study on the O–H bond dissociation enthalpy, Redox Rep. 9 (2004) 263–269.

[7] N. Nenadis, L.F. Wang, M.Z. Tsimidou, H.Y. Zhang, Radical scavenging potential of phenolic compounds encountered in *O. europaea* products as indicated by calculation of bond dissociation enthalpy and ionization potential values, J. Agric. Food Chem. 53 (2005) 295–299.

[8] K. Tanaka, S. Sakai, S. Tomiyama, T. Nishiyama, F. Yamada, Molecular orbital approach to antioxidant mechanisms of phenols by an ab initio study, Bull. Chem. Soc. Jpn. 64 (1991) 2677–2680.

[9] S. Tomiyama, S. Sakai, T. Nishiyama, F. Yamada, Factors influencing the antioxidant activities of phenols by an ab initio study, Bull. Chem. Soc. Jpn. 66 (1993) 299–304.

[10] D. Shanks, H. Frisell, H. Ottosson, L. Engman, Design principles for alpha-tocopherol analogues, Org. Biomol. Chem. 4 (2006) 846–852.

[11] Z. Velkov, E. Balabanova, A. Tadjer, Radical scavenging activity prediction of *o*-coumaric acid thioamide, J. Mol. Struct. (Theochem.) 821 (2007) 133–138.

[12] J. Lind, X. Shen, T.E. Eriksen, G. Merenyi, The one-electron reduction potential of 4-substituted phenoxyl radicals in water, J. Am. Chem. Soc. 112 (1990) 479–482.

[13] P. Mulder, H.G. Korth, D.A. Pratt, G.A. DiLabio, L. Valgimigli, G.F. Pedulli, K.U. Ingold, Critical re-evaluation of the O–H bond dissociation enthalpy in phenol, J. Phys. Chem. A 109 (2005) 2647–2655.

[14] J.A. Pople, M. Head-Gordon, K. Raghavachari, Quadratic configuration interaction. A general technique for determining electron correlation energies, J. Chem. Phys. 87 (1987) 5968–5975.

[15] K. Raghavachari, G.W. Trucks, J.A. Pople, M. Head-Gordon, A fifth-order perturbation comparison of electron correlation theories, Chem. Phys. Lett. 157 (1989) 479–483.

[16] J.W. Ochterski, G.A. Petersson, J.A. Montgomery, A complete basis set model chemistry. V. Extensions to six or more heavy atoms, J. Chem. Phys. 104 (1996) 2598–2619.

[17] G.A. Petersson, T.G. Tensfeldt, J.A. Montgomery, A complete basis set model chemistry. III. The complete basis set-quadratic configuration interaction family of methods, J. Chem. Phys. 94 (1991) 6091–6101.

[18] M.R. Hoffmann, H.F. Schaefer Iii, The treatment of triple excitations within the coupled cluster description of molecular electronic structure, J. Chem. Phys. 83 (1985) 703–712.

[19] L.A. Curtiss, K. Raghavachari, P.C. Redfern, V. Rassolov, J.A. Pople, Gaussian-3 (G3) theory for molecules containing first and second-row atoms, J. Chem. Phys. 109 (1998) 7764–7776.

[20] L.A. Curtiss, P.C. Redfern, K. Raghavachari, Gaussian-4 theory, J. Chem. Phys. 126 (2007) 084108–084112.

[21] K.A. Peterson, T.H. Dunning Jr., Benchmark calculations with correlated molecular wave functions. VIII. Bond energies and equilibrium geometries of the CH_n and C_2H_n (n = 1–4) series, J. Chem. Phys. 106 (1997) 4119–4140.

[22] M.R. Nyden, G.A. Petersson, Complete basis set correlation energies. I. The asymptotic convergence of pair natural orbital expansions, J. Chem. Phys. 75 (1981) 1843–1862.

[23] G.A. Petersson, M.A. Al-Laham, A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms, J. Chem. Phys. 94 (1991) 6081–6090.

[24] J.A. Montgomery, J.W. Ochterski, G.A. Petersson, A complete basis set model chemistry. IV. An improved atomic pair natural orbital method, J. Chem. Phys. 101 (1994) 5900–5909.

[25] J.L. Lewin, C.J. Cramer, Rapid quantum mechanical models for the computational estimation of C–H bond dissociation energies as a measure of metabolic stability, Mol. Pharm. 1 (2004) 128–135.

[26] F.G. Bordwell, J.-P. Cheng, Substituent effects on the stabilities of phenoxyl radicals and the acidities of phenoxyl radical cations, J. Am. Chem. Soc. 113 (1991) 1736–1743.

[27] R. Dennington II, T. Keith, J. Millam, K. Eppinnett, W.L. Hovell, R. Gilliland, GaussView, version 3.09, Semichem, Inc., Shawnee Mission, KS, 2003.

[28] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.

[29] T. Koopmans, Uber die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms, Physica 1 (1934) 104–113.

[30] Camo Process AS, The Unscrambler, version 9.5, Norway, 2006.

[31] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[32] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. 58 (2001) 109–130.

[33] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.

[34] J.C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research, Technical Report MSR-TR-98-14, 1998.

[35] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[36] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.

[37] S. Abe, Support Vector Machines for Pattern Classification, Springer-Verlag London Limited, New York, 2005.

[38] Y. Sakiyama, H. Yuki, T. Moriya, K. Hattori, M. Suzuki, K. Shimada, T. Honma, Predicting human liver microsomal stability with machine learning techniques, J. Mol. Graph. Model. 26 (2008) 907–915.

[39] H.H. Lin, L.Y. Han, C.W. Yap, Y. Xue, X.H. Liu, F. Zhu, Y.Z. Chen, Prediction of factor Xa inhibitors by machine learning methods, J. Mol. Graph. Model. 26 (2007) 505–518.

[40] G. Liang, Z. Li, Scores of generalized base properties for quantitative sequence-activity modelings for *E. coli* promoters based on support vector machine, J. Mol. Graph. Model. 26 (2007) 269–281.

[41] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, The MIT Press, Cambridge, 2002.

[42] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, The MIT Press, Cambridge, 2002.

[43] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[44] L. Wang, Support Vector Machines: Theory and Applications, Springer-Verlag, Berlin, 2005.

[45] S. Abe, Support Vector Machines for Pattern Classification, Springer-Verlag, London, 2005.

[46] N. Chen, W. Lu, J. Yang, G. Li, Support Vector Machine in Chemistry, World Scientific Publishing Co. Pte. Ltd., Singapore, 2004.

[47] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, Technical Report, 2007.

[48] C. Staelin, Parameter Selection for Support Vector Machines, Hewlett-Packard Company, Technical Report HPL-2002-354 (R.1), 2003.

[49] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. Tansila, T. Naenna, V. Prachayasittikul, Prediction of GFP spectral properties using artificial neural network, J. Comput. Chem. 28 (2007) 1275–1289.

[50] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Quantitative structure-imprinting factor relationship of molecularly imprinted polymers, Biosens. Bioelectron. 22 (2007) 3309–3317.

[51] B.J. Lynch, D.G. Truhlar, How well can hybrid density functional methods predict transition state geometries and barrier heights? J. Phys. Chem. A 105 (2001) 2936–2941.

[52] J.K. Kang, C.B. Musgrave, Prediction of transition state barriers and enthalpies of reaction by a new hybrid density-functional approximation, J. Chem. Phys. 115 (2001) 11040–11051.

[53] H. Basch, S. Hoz, *Ab initio* study of hydrogen abstraction reactions, J. Phys. Chem. A 101 (1997) 4416–4431.

[54] M.L. Coote, Reliable theoretical procedures for the calculation of electronic-structure information in hydrogen abstraction reactions, J. Phys. Chem. A 108 (2004) 3865–3872.

[55] V.A. Guner, K.S. Khuong, K.N. Houk, A. Chuma, P. Pulay, The performance of the Handy/Cohen functionals, OLYP and O3LYP, for the computation of hydrocarbon pericyclic reaction activation barriers, J. Phys. Chem. A 108 (2004) 2959–2965.

[56] V. Guner, K.S. Khuong, A.G. Leach, P.S. Lee, M.D. Bartberger, K.N. Houk, A standard set of pericyclic reactions of hydrocarbons for the benchmarking of computational methods: the performance of ab initio, density functional, CASSCF, CASPT2, and CBS-QB3 methods for the prediction of activation barriers, reaction energetics, and transition state geometries, J. Phys. Chem. A 107 (2003) 11445–11459.

[57] Y. Sun, D. Chen, C. Liu, Evaluation of the effectiveness of AM1 geometry used in calculating O–H bond dissociation enthalpy, J. Mol. Struct. (Theochem.) 618 (2002) 181–189.

[58] M.T. Shamim, M. Anwaruddin, H.A. Nagarajaram, Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs, Bioinformatics 23 (2007) 3320–3327.

[59] J.J. Ward, L.J. McGuffin, B.F. Buxton, D.T. Jones, Secondary structure prediction with support vector machines, Bioinformatics 19 (2003) 1650–1655.

[60] C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Recognition of DNA splice junction via machine learning approaches, EXCLI J. 4 (2005) 114–129.

[61] J.C. Saeh, P.D. Lyne, B.K. Takasaki, D.A. Cosgrove, Lead hopping using SVM and 3D pharmacophore fingerprints, J. Am. Chem. Soc. 45 (2005) 1122–1133.