J Comput Aided Mol Des (2008) 22:345–366
DOI 10.1007/s10822-008-9190-y

Categorical QSAR models for skin sensitization based on local lymph node assay measures and both ground and excited state 4D-fingerprint descriptors

Jianzhong Liu · Petra S. Kern · G. Frank Gerberick · Osvaldo A. Santos-Filho · Emilio X. Esposito · Anton J. Hopfinger · Yufeng J. Tseng

Received: 28 September 2007 / Accepted: 30 January 2008 / Published online: 13 March 2008

Abstract In previous studies we have developed categorical QSAR models for predicting skin-sensitization potency based on 4D-fingerprint (4D-FP) descriptors and in vivo murine local lymph node assay (LLNA) measures. Only 4D-FP derived from the ground state (GMAX) structures of the molecules were used to build the QSAR models. In this study we have generated 4D-FP descriptors from the first excited state (EMAX) structures of the molecules. The GMAX, EMAX and the combined ground and excited state 4D-FP descriptors (GEMAX) were employed in building categorical QSAR models. Logistic regression (LR) and partial least square coupled logistic regression (PLS-CLR), found to be effective model building for the LLNA skin-sensitization measures in our previous studies, were used again in this study. This also permitted comparison of the prior ground state models to those involving first excited state 4D-FP descriptors. Three types of categorical QSAR models were constructed for each of the GMAX, EMAX and GEMAX datasets: a binary model (2-state), an ordinal model (3-state) and a binary-binary model (two-2-state). No significant differences exist among the LR 2-state model constructed for each of the three datasets. However, the PLS-CLR 3-state and 2-state models based on the EMAX and GEMAX datasets have higher predictivity than those constructed using only the GMAX dataset. These EMAX and GMAX categorical models are also more significant and predictive than corresponding models built in our previous QSAR studies of LLNA skin-sensitization measures.

Keywords Skin sensitization · Categorical QSAR models · Excited state structures

J. Liu · A. J. Hopfinger
College of Pharmacy, 1 University of New Mexico, MSC09 5360, Albuquerque, NM 87131-0001, USA

J. Liu · O. A. Santos-Filho · E. X. Esposito · A. J. Hopfinger · Y. J. Tseng
The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045, USA

P. S. Kern
Procter & Gamble Eurocor, Temselaan 100, 1853 Strombeek-Bever, Belgium

G. F. Gerberick
The Procter & Gamble Company, Miami Valley Innovation Center, P.O. Box 538707, Cincinnati, OH 45253-8707, USA

Y. J. Tseng (✉)
Department of Computer Science and Information Engineering, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No.1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan
e-mail: yjtseng@csie.ntu.edu.tw

Introduction

Allergic contact dermatitis (ACD) results from the T-lymphocyte mediated immune response to a chemical allergen coming into contact with the skin [1]. The small allergenic molecule (hapten) penetrates the skin and binds to a carrier protein, typically by covalent bond, to form an antigenic hapten-protein complex. This complex is then processed by antigen-presenting cells migrating to the draining lymph-nodes, where they introduce the haptens to the T-lymphocytes. A chemical compound's ability to behave as a contact allergen depends on how well it penetrates the stratum corneum and on its means to react, either directly or after metabolic activation, with skin proteins.

Thus, the chemical reactivity profile of a compound plays a major role in its propensity to be a chemical allergen.

Studies [2, 3] show that the photoreactive states of skin-sensitizing carcinogenic compounds are characterized by a substantial localization of the electronic excitation, thereby providing a logical basis for the structure-activity correlation of skin-sensitizing compounds. And in an analogous manner, the coumarins, which are one class of typical skin-sensitizing derivatives, also possess partially localized triplet states [4]. In contrast, the 3La state of benzo(a)pyrene is characterized more by electron delocalization than that of the non-carcinogenic benzo(e)pyrene [3]. Earlier studies [5] have also shown 5-fluorouracil to be much more reactive than thymine with respect to the excited state of some sensitizing carcinogenic compounds. Recently, other studies [6] have suggested the excited states of endogenous chromophores, such as porphyrins, melanin precursors and cross-link-fluorophores of skin collagen, exert skin photodamage by direct reaction with substrate molecules, leading to formation of reactive oxygen species. Further, the excited state cycloaddition of some skin-sensitizing carcinogenic compounds is predicted as the most favorable pathway [7], which agrees with the experimental findings that such a pathway provides the highest yield. Overall, these results suggest that the use of excited state molecular features in combination with ground state properties may be important in constructing models that accurately capture the overall mechanistic details of skin sensitization.

Quantitative structure activity relationships (QSARs) increasingly play a role in compound evaluation and screening, and are considered an important alternative for the estimation of toxicity effects, including skin sensitization. But QSARs have traditionally been developed using chemical properties and features derived from the ground state of a molecule [8–10]. Even for adverse biological responses, like skin sensitization, where the chemical reactivity of the molecule is known to play a role in the expression of the response, chemical reactivity has been only indirectly represented through ground state descriptors. These descriptors include properties derived explicitly from the electronic structure of a molecule, like molecular orbital energies (HOMO and LUMO), of the molecule or more empirical features, such as two-dimensional electrotopological descriptors [11, 12].

Using ground state descriptors of a molecule to characterize skin sensitization has had success. Presumably certain ground state descriptors provide glimpses of the reactivity behavior of a molecule. For example, we have carried out a two-state categorical QSAR modeling of skin sensitization [13] using a dataset constructed from the validated in vivo murine local lymph node assay (LLNA) [14]. In our initial study, ground state 4D-fingerprints (4D-FP) were used as a descriptor set to generate categorical QSAR models for two states: sensitizer and non-sensitizer [13]. This current study focused on exploring the usefulness of the statistical methodologies of logistic regression (LR) and partial-least square coupled logistic regression (PLS-CLR) to build two-state categorical models. The LR models have a cross-validated prediction accuracy range of 77.3–78.0% for the training set, while that for the PLS-CLR models ranges from 87.1 to 89.4%. For the test set, the prediction accuracy of the LR models ranges from 80.0 to 86.7%, while that for the PLS-CLR models range from 73.3 to 80.0%. These significant values show that the methods applied in this study are effective for separating non-sensitizers from sensitizers.

In a more recent study, we used the same LLNA dataset and both LR and PLS-CLR to build 3-state and two-2-state categorical models for skin sensitization potency [15]. The three-state QSAR model yields a classification accuracy of 73.4% for the training set and 63.6% for the test set, while the random average value of classification accuracy for any 3-state dataset is 33.3%. The two-2-state (four categories in total) QSAR model gives a classification accuracy of 83.2% for the training set and 54.6% for the test set, while the random average value of classification accuracy for any two-2-state dataset is 25%. A comparison of the results of the two-state modeling study described above suggests that including more than two categorical states in skin-sensitization modeling leads to a loss of accuracy and reliability. This may arise, in part, from the lack of including explicit descriptors derived from excited states of the molecules.

The two primary goals of this study are: (1) to develop a methodology to compute 4D-FP from the electronic and geometric structures of molecules in their excited states, and, (2) to use ground and/or excited state 4D-FP to build categorical QSAR models for skin sensitization that can be directly compared to equivalent models previously developed using only ground state 4D-FP. To facilitate the second goal—generating 4D-FP categorical QSAR models, three types of 4D-FP descriptors have been assigned to the descriptor pool: GMAX (ground state descriptors only), EMAX (excited state descriptors only), and GEMAX (the combination of ground and excited state descriptors).

The final results show that the constructed models based on EMAX and GEMAX datasets have a higher predictivity than those derived from the GMAX dataset. Interestingly, there are no obvious differences between the EMAX and GEMAX 4D-FP categorical QSAR models.

Material and methodology

Database construction

The LLNA training and test sets used in this study were pruned from a master skin-sensitization database, which itself was constructed from data provided by interested

organizations [16, 17]. This pruned database consists of 126 compounds which were originally categorized as non-, weak-, moderate-, strong- and extreme-skin sensitizers, according to each compound's EC3 value [18]. Depending on the size of the categorical model being built, the pruned database was reclassified into three main categories: 63 "non-weak" sensitizers, 35 "moderate" sensitizers and 28 "strong-extreme" sensitizers.

An initial 3D structure of each of the 126 compounds was constructed in its neutral form using HyperChem 7.05 software [19] to obtain its three dimensional coordinate and optimized geometry. For the ground state, each structure's energy was minimized using the AM1 molecular orbital method without any geometric constraints. For the excited state, each structure was energy minimized using both the AM1 [20] and the PM3 [21] molecular orbital methods, without any geometric constraints, for the lowest-energy excited state. In the ground and lowest-energy excited states, the energy-minimized structures and corresponding charge distributions were used as the initial structures in the 4D-FP module of the 4D-QSAR molecular modeling package [22]. Three 4D-FP trial descriptor matrices were employed in building the categorical QSAR models: GMAX (only the ground 4D-FP descriptors), EMAX (only the excited state 4D-FP descriptors), and GEMAX (the combination ground and excited state 4D-FP descriptors).

In this study, three types of categorical QSAR models are constructed for each of the three 4D-FP descriptor datasets: a binary model (2-state), ordinal model (3-state), and a binary-binary model (two-2-state). In the 2-state model, a "non-weak" sensitizer has a value of "0" and a "strong-extreme" sensitizer a value of "1". In the 3-state model, a "non-weak" sensitizer has a value of "0", a "moderate" sensitizer "1", and a "strong-extreme" sensitizer "2". In the two-2-state model, compounds are classified according to the following two steps: First, a "non-weak" sensitizer has a value of "0" and a "moderate-strong-extreme" sensitizer has a value of "1". In the second step, dealing with the selected "non-weak" sensitizers, a value of "0" is assigned to a non-sensitizer and a value of "1" for a weak-sensitizer. Similarly, for the selected "moderate-strong-extreme" sensitizers, the values of the dependent variable are assigned based on the skin-sensitization potency of each compound; "0" for a moderate-sensitizer and "1" for a "strong-extreme" sensitizer.

In building the 4D-FP QSAR models for this study, the training sets and the test sets members were randomly chosen from each skin-sensitization potency category of the parent 126-compound LLNA dataset. To compare the 2-state model constructed in this study with models developed in our previous studies, we chose a similar ratio for the number of compounds making up the training set and the number for the test set. This ratio is about 8:1, with 80 training set compounds and 11 test set compounds for this study. However, the predicted accuracies had no obvious differences among each study's 2-state model based on the GMAX, EMAX, and GEMAX 4D-FP trial descriptor matrices. Therefore, the final ratio was ultimately decreased, with fewer training set compounds and more test set compounds.

Universal 4D-fingerprints, 4D-FP, descriptors

The theory and methodology of the universal 4D-FP descriptors were developed using the 4D-QSAR paradigm and 4D molecular similarity analysis [23, 24]. The universal 4D-FP are the eigenvalues of the molecular similarity eigenvectors determined for a given molecule based on a set of absolute molecular similarity main distance-dependent matrices (MDDM). The eigenvectors capture the molecular information of a molecule's atom types, size, shape and conformational flexibility. The types of atoms composing a molecule are currently defined as eight interaction pharmacophore elements (IPE's), which were defined in previous papers [9, 25]. A cutoff value for the eigenvalues is applied, and those normalized eigenvalues below the cutoff are disregarded. For this study, the cutoff was set at 0.002. Our previous papers [9, 23–25] offer more details on deriving the 4D-FP descriptors. Construction of the trial descriptor matrix for all training set compounds is determined by maximizing its information content. For each compound in the training set, the number of significant eigenvalues in the eigenvector for a particular IPE pair (u, v) is computed. Then the maximum number of significant eigenvalues, nmax(u,v), across the training set is determined. Finally, the molecules in the training set are assigned nmax(u,v) eigenvalues from their corresponding eigenvectors for the IPE pair (u,v). Eigenvectors containing fewer than nmax(u,v) significant eigenvalues have these "missing" eigenvalues set to 0. For instance, if nmax(3,5) is 10, and the eigenvector for IPE pair (u,v) of a compound has only eight significant eigenvalues, the ninth and tenth eigenvalues for IPE pair(u,v) are set to 0.

When applying this methodology, the total number of 4D-FP descriptors (ntotal) for each compound in the training set will be the sum of the nmax(u,v) values for the 36 eigenvectors arising from the set of unique IPE pairs. The number of 4D-FP descriptors in the trial descriptor matrices is 290 each for the GMAX and EMAX, and 580 for the GEMAX matrix. A descriptor $\varepsilon_i(u,v)$ used in these matrices represents the ith eigenvalue in the eigenvector for the IPE pair (u,v). $\varepsilon_i(u,v)g$ indicates that this descriptor comes from ground state structure of a molecule while $\varepsilon_i(u,v)e$ denotes that the descriptor comes from the excited

state structure. The method involved in creating the trial descriptor matrix introduces a degree of "noise," with the need to add zero value eigenvalues. However, if the "noise" in a particular descriptor column renders the descriptor unfit for describing the variance of the dependant categorical variable, this descriptor will not show up in the optimized QSAR model. Thus, these matrices can be used directly in categorical QSAR-model construction.

Data reduction and model construction

As mentioned above, some degree of "noise" may exist within the trial descriptor matrices. Collinearity and/or multilinearity may also exist between/among the 4D-FP descriptors of all three trial descriptor matrices. Partial least square (PLS) regression [26] was used to reduce collinearity and noise. The reduced matrices for each of the trial descriptor matrices can be employed in the next step of model building, which is logistic regression (LR) analysis. This overall process is referred to as PLS-CLR. In this way, the original trial descriptor matrix will be reduced to a smaller matrix, where each column represents an extracted component from the original trial descriptor matrix. In this study, *xscri* denotes the *i*th PLS component extracted from the original trial descriptor matrix. Although more than 100 components are extracted from the original trial descriptor matrix, only the first 20 were selected for the three trial descriptor matrices, the reason being that the first set of components account for most of the variances in both the explanatory variables and the response logit. In this study, the first 20 components account for more than 80% of variance in all applications.

For the 2-state models, the response, Y—the skin-sensitization potency endpoint—can take on one of two possible values, denoted for convenience by 1 and 0. $Y = 1$ if a compound is a strong or extreme skin sensitizer and $Y = 0$ if the compound is a non- or weak-sensitizer. Assuming xi is a vector of explanatory independent variables, in this case the 4D-FP, and $P_1 = Pr(Y = 1)$ is the response probability to be modeled, then $P_0 = Pr(Y = 0)$. The linear logistic model has the form,

$$Logit(P_1) = \log\frac{P_1}{1-P_1} = \delta + b1 * X1 + b2 * X2 + \cdots + bk * Xk \quad (1)$$

or

$$P_1 = \frac{\exp(\delta + b1 * X1 + b2 * X2 + \cdots + bk * Xk)}{1 + \exp(\delta + b1 * X1 + b2 * X2 + \cdots + bk * Xk)} \quad (2)$$

$$P_0 = 1 - P_1 \quad (3)$$

where δ is the intercept parameter and b*i* is the vector of slope parameters in Eqs. 1 and 2. The linear LR models are fit to binary-response data using the maximum likelihood method, which is generated from either Fisher-scoring or a Newton-Raphson algorithm [27]. The predicted response probabilities are obtained by replacing the b parameter with its maximum likelihood estimate (MLE), $\hat{\rho}$.

For 3-state models, the response Y can take ordinal values, denoted for convenience in this study by 0, 1 and 2. In particular, the dependent variable Y takes values of 2 (strong or extreme skin sensitizer), 1 (moderate skin sensitizer), and 0 (non- or weak-sensitizer.)

Cumulative logits can be modeled with the proportional odds model. The proportional odds model assumes the cumulative logits can be represented as parallel linear functions of independent variables. That is, for each cumulative logit the parameters of the models are the same, except for the intercept. If $P_0 = P(Y = 0)$, $P_1 = P(Y = 1)$ and $P_2 = P(Y = 2)$, then ordinal logistic regression models the relationship between the cumulative logits of Y, or $\log(P_2/(1-P_2)) = \log(P_2/(P_{\ 1} + P_0))$ and $\log((P_2 + P_1)/(1-P_2-P_{\ 1})) = \log((P_2 + P_1)/P_0)$. The model assumes a linear relationship for each logit and parallel regression lines,

$$Logit(P_2) = \log\left(\frac{P_2}{1-P_2}\right) = \delta_2 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k \quad (4)$$

$$Logit(P_2 + P_1) = \log\left(\frac{P_2 + P_1}{1 - P_2 - P_1}\right) = \delta_1 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k \quad (5)$$

δ is the intercept parameter and b*i* is the vector of slope parameters. From the above equations, it is obvious that the intercepts are different, but the remaining regression parameters are the same. It is easy to see that the odds $P_2/(1-P_2)$ and $(P_2 + P_1)/P_0$ are proportional,

$$\frac{P_2}{1-P_2} = \exp(\delta_2) * \exp(b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k) \quad (6)$$

$$\frac{P_2 + P_1}{P_0} = \exp(\delta_1) * \exp(b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k) = C * \frac{P_2}{1-P_2} \quad (7)$$

where C is a constant and equals $exp(\delta_1 - \delta_2)$.

The maximum likelihood estimation is used to obtain the estimates of the model parameters. After estimators of $\delta_2, \delta_1, b_1, b_2, \ldots, b_k$ are computed, it is easy to compute

predicted probabilities using the following relationships derived from the above equations.

$$P_2 = \frac{\exp(\delta_2 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k)}{1 + \exp(\delta_2 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k)} \quad (8)$$

$$P_2 + P_1 = \frac{\exp(\delta_1 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k)}{1 + \exp(\delta_1 + b_1 * X_1 + b_2 * X_2 + \cdots + b_k * X_k)} \quad (9)$$

$$P_0 = 1 - (P_2 + P_1) \quad (10)$$

Goodness of fit and the predictivity of a model

The Hosmer–Lemeshow test [27] is widely used to evaluate the goodness of fit of a categorical model, and it can be applied in this application since the 4D-FP are continuous variables. This test involves dividing the observations into k groups of approximately the same size, and with similar estimated probabilities within one group. Then the Hosmer–Lemeshow goodness-of-fit statistic, χ^2, is defined as,

$$\chi^2 = \sum_{i=0}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (11)$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by,

$$E_i = N(F(Y_u) - F(Y_l)) \quad (12)$$

where F is the cumulative distribution function for the distribution being tested, Y_u is the upper limit for class i, Y_l is the lower limit for class i, and N is the sample size. Large values of χ^2 (and small p-values) indicate a lack of fit for the model.

Model predictivity was evaluated using both leave-one-out cross-validation and a compound test set. The higher the classification accuracy for the training and test sets, based on the leave-one-out measure, the greater the predictivity of a categorical QSAR model. Classification accuracy, sensitivity and specificity are defined, respectively, as,

$$Accuracy = \frac{tp + tn}{tp + fn + tn + fp} \quad (13a)$$

$$Sensitivity = \frac{tp}{tp + fn} \quad (13b)$$

$$Specificity = \frac{tn}{tn + fp} \quad (13c)$$

where tp and fn are the numbers of strong-extreme sensitizers for which the predicted probabilities are larger and less than the cutoff value of 0.5, respectively. Similarly, tn and fp are the numbers of non-weak sensitizers for which the predicted probabilities are less than and larger than the cutoff value of 0.5, respectively.

It should be noted that classification accuracy and goodness-of-fit have no direct inter-relationship. Accurate or inaccurate classification does not indicate goodness-of-fit, or vice-versa. However, use of classification accuracy is most appropriate when classification is a stated goal of the analysis [28], as is in this study.

The statistical methodology reported above can be implemented with a combination of standard statistical methods available in SAS [29].

Results

Logistic regression (LR) analysis for building 2-state models

The most significant and accurate categorical models were obtained by applying stepwise LR on 2-state models for all three types of trial descriptor matrices, GMAX, EMAX, and GEMAX. There are a total of 290 4D-fingerprints in both GMAX and EMAX. The number doubles in GEMAX. From the 126 selected compounds, 91 belong to "non-weak" and "strong-extreme" categories, with the remaining 35 compounds in the "moderate category". Eighty of the 91 "non-weak" and "strong-extreme" compounds were selected randomly as training set compounds to build the 2-state models. The significance level for a descriptor to enter or leave an evolving model was set at a default value of 0.05. The resulting optimal 2-state categorical QSAR models are shown below. For GMAX, the 2-state QSAR model is given by Eq. 14, which contains four 4D-FP descriptors. For EMAX, Eq. 15 defines the corresponding model, which contains five 4D-FP descriptors. Equation (16) defines the optimal 2-state categorical QSAR model for GEMAX, which contains four 4D-FP descriptors:

$$\begin{aligned} Logit(P_1) = &-14.80 + 20.69 * \varepsilon 1(any, np)g + 108.90 \\ &* \varepsilon 8(np, hs)g + 14.22 * \varepsilon 2(hba, hs)g \\ &+ 299.70 * \varepsilon 10(aro, hs)g \end{aligned} \quad (14)$$

$$\begin{aligned} Logit(P_1) = &-20.86 + 31.72 * \varepsilon 1(any, np)e \\ &- 202.40 * \varepsilon 5(np, hs)e + 394.70 * \varepsilon 7(np, hs)e \\ &+ 19.40 * \varepsilon 2(hba, hs)e + 321.70 * \varepsilon 10(aro, hs)e \end{aligned} \quad (15)$$

$$\begin{aligned} Logit(P_1) = &-15.11 + 112.00 * \varepsilon 8(np, hs)g \\ &+ 14.40 * \varepsilon 2(hba, hs)g + 303.60 \\ &* \varepsilon 10(aro, hs)g + 21.19 * \varepsilon 1(any, np)e \end{aligned} \quad (16$$

In Eqs. 14–16, by way of example, $\varepsilon 1(any, np)g$ represents the largest eigenvalue from the MDDM of

$u = (any)$ and $v = (np)$, of the ground state 4D-FP calculation. Likewise, $\varepsilon 5(np, hs)e$ represents the fifth largest eigenvalue from the MDDM of $u = (np)$ and $v = (hs)$ from excited state 4D-FP.

The GEMAX 2-state QSAR model is very similar to the GMAX 2-state QSAR model except that one 4D-FP descriptor changes from $\varepsilon 1(any, np)g$ to $\varepsilon 1(any, np)e$. Thus, in the GEMAX 2-state QSAR model, the largest 4D-FP (eigenvalue) with respect to the joint spatial distributions of *any* atoms and *non-polar* atoms across the training set compounds, comes from the excited state, not from the ground state. This particular 4D-FP is more significant when it comes from the excited state than when derived from the ground state. The increased statistical significance of eq(16) as compared to Eq. 14 is reflected in their respective Hosmer–Lemeshow goodness-of-fit measures. The GEMAX QSAR model has a smaller χ^2 value of 10.31 than the 11.89 in the GMAX QSAR model, indicating the GEMAX model is more significant than the GMAX model. However, the predicted accuracies of these two 2-state QSAR models, shown in Table 1, are the same. The predicted probability of each compound being a "strong-extreme sensitizer" and its corresponding predicted skin-sensitization potency category are also listed in Table 1.

The EMAX 2-state categorical QSAR model has an additional descriptor, $\varepsilon 5(np, hs)e$, compared to both the GMAX and GEMAX models. The LR analysis using the EMAX trial descriptor matrix found an additional significant descriptor compared to both the GMAX and GEMAX trial descriptor matrices. One 4D-FP, ε $8(np, hs)g$, appearing in the GMAX and GEMAX QSAR models was found to change slightly to $\varepsilon 7(np, hs)e$ in the EMAX. Overall, a comparison of the regression coefficients and constant of the EMAX QSAR model (Eq. 15), as compared to Eqs. 14 and 16, indicates the EMAX QSAR model captures additional and somewhat different attributes with respect to the spatial distribution of hydrophobic atoms. However, there are similar spatial features in regards to hydrogen bond acceptors over the training set compounds. The highly negative regression constant and negative regression coefficient for $\varepsilon 5(np, hs)e$ in Eq. 15 causes the final EMAX model to have highly positive regression coefficients for the other 4D-FP in order to compensate. That is, for descriptors with positive regression coefficients in Eq. 15 that also appear in Eq. 14 or Eq. 16, an increase value for each descriptor yields a greater contribution to the probability of the compound falling into the "strong-extreme" category using Eq. 15 than with Eq. 14 or Eq. 16.

The predicted probabilities, classifications and accuracies for models (14), (15) and (16), based on the training set compounds, are shown in Table 1. A summary of classification accuracy for the test set predictions is given in Table 2. This study, and our previous 2-state modeling study [13], yield very similar predicted accuracies, when based solely upon the ground state, for the training set: 87.5% for this study and 85.6% in the previous study. However, the predicted accuracy in this study (63.6%) for the test set is lower than the last study (86.7%). The reason is that both the size and choice of the training and test set compounds differ somewhat in the studies. However, the predictive accuracy for the test set of this study is still larger than that for a random blind test, which is 47.4% based on the cutoff value used in this study. Thus, the 2-state models, Eqs. 14–16, have a predictive capacity of around 17% (64–47%) above that of random chance.

PLS-CLR analysis for building 2-state models

The previous study to develop 2-state categorical QSAR modeling of the LLNA dataset, using only ground state 4D-FP, showed that PLS-CLR analysis led to models of higher prediction accuracy than the LR models [15]. This finding encouraged us to perform a PLS-CLR analysis in this study to improve the prediction accuracy of the resulting 2-state models. PLS-CLR was used to construct 2-state categorical QSAR models for the trial descriptor matrices GMAX, EMAX and GEMAX. The GMAX and EMAX matrices yielded 124 components. As noted in the Material and methodology section, the first 20 components are only used for QSAR model construction. The PLS-CLR 4D-FP QSAR model for GMAX is Eq. 17, which contains four extracted components. For EMAX, the optimal QSAR model is Eq. 18 with five components, while for GEMAX, Eq. 19 having four extracted components is the optimal PLS-CLR 2-state categorical QSAR model.

$$Logit(P_1) = -2.710 + 1.384 * xscr1 + 1.158 * xscr5 + 1.463 * xscr9 + 2.499 * xscr13 \quad (17)$$

$$Logit(P_1) = -2.497 + 1.716 * xscr1 + 1.531 * xscr5 + 0.720 * xscr8 + 1.096 * xscr9 + 1.861 * xscr13 \quad (18)$$

$$Logit(P_1) = -2.942 + 1.200 * xscr1 + 1.044 * xscr5 + 1.107 * xscr9 + 1.847 * xscr13 \quad (19)$$

The three models, Eqs. 17–19, are very similar to one another. In fact, the extracted components are exactly the same for the GMAX and GEMAX QSAR models. The only modest differences between the two are found in the regression coefficients and regression constants. But like the LR 2-state models, the EMAX QSAR model has one more descriptor, *xscr*8, compared to the other two models. This suggests that information extracted in the PLS-CLR QSAR model-building process involves an additional PLS

Table 1 The predicted probabilities and corresponding classifications of the training set compounds using LR 2-state QSAR models, Eqs. 14–16, for the ground state descriptors (GMAX), excited state descriptors (EMAX). and the combination of ground and excited state descriptors (GEMAX)

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted probability	Predicted class	Predicted probability
4-Methoxyacetophenone	0	0.302	0	0.389	0	0.295	0
4-Nitrobenzyl bromide	1	0.837	1	0.825	1	0.842	1
Benzyl bromide	1	0.083	0	0.150	0	0.087	0
Benzaldehyde	0	0.206	0	0.338	0	0.210	0
Cyclamen aldehyde	0	0.014	0	0.004	0	0.014	0
Cinnamic alcohol	0	0.026	0	0.000	0	0.025	0
1, 4-Phenylenediamine	1	0.895	1	0.842	1	0.899	1
p-Benzoquinone	1	0.845	1	0.872	1	0.848	1
Hydroxycitronellal	0	0.102	0	0.149	0	0.099	0
Maleic anhydride	1	0.951	1	0.997	1	0.954	1
3-Phenylenediamine	1	0.756	1	0.996	1	0.757	1
1-Chloromethylpyrene	1	0.992	1	0.993	1	0.993	1
2,4,6-Trichloro-1,3,5-triazine (cyanuric chloride)	1	0.963	1	0.998	1	0.966	1
Chlorobenzene	0	0.029	0	0.132	0	0.028	0
1-Bromobutane	0	0.020	0	0.000	0	0.020	0
Hexane	0	0.007	0	0.000	0	0.007	0
2,2,6,6-Tetramethyl-heptane-3,5-dione	0	0.168	0	0.080	0	0.164	0
1-Bromooctadecane	0	0.001	0	0.000	0	0.001	0
Benzyl benzoate	0	0.540	1	0.312	0	0.523	1
Ethyl vanillin	0	0.254	0	0.006	0	0.226	0
Propyl gallate	1	0.298	0	0.312	0	0.283	0
a-Amyl cinnamic aldehyde	0	0.006	0	0.001	0	0.006	0
Hydroquinone	1	0.932	1	0.920	1	0.936	1
Octanoic acid	0	0.158	0	0.002	0	0.130	0
Dodecylthiosulphonate	1	0.033	0	0.010	0	0.038	0
1-(3′,4′,5′-Trimethoxyphenyl)-4-dimethylpentane-1, 3,-dione	0	0.024	0	0.017	0	0.029	0
5-Methyl-2,3-hexanedione	0	0.117	0	0.308	0	0.118	0
4-Nitroso-*N*,*N*-dimethylaniline	1	0.702	1	0.875	1	0.701	1
4-Allylanisole	0	0.008	0	0.000	0	0.007	0
Ethyl acrylate	0	0.062	0	0.108	0	0.054	0
1-Bromododecane	0	0.001	0	0.000	0	0.001	0
Oxalic acid	0	0.000	0	0.000	0	0.000	0
2-Mercaptobenzothiazole	0	0.046	0	0.019	0	0.045	0
5,5-Dimethyl-3-thiocyanatomethyl-2(3H)-furanone	1	0.524	1	0.838	1	0.532	1
3-Ethoxy-1-(2′,3′,4′,5′-tetramethylphenyl)propane-1, 3-dione	0	0.208	0	0.255	0	0.189	0
C11-Azlactone	0	0.011	0	0.000	0	0.012	0
6-Methyleugenol	0	0.453	0	0.258	0	0.453	0
5-Methyleugenol	0	0.493	0	0.472	0	0.487	0
3-Methyleugenol	0	0.179	0	0.066	0	0.178	0
5-Methylisoeugenol	1	0.365	0	0.492	0	0.367	0
cis-6-Nonenal	0	0.007	0	0.000	0	0.008	0

Table 1 continued

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted probability	Predicted class	Predicted probability
1-Chlorotetradecane	0	0.001	0	0.000	0	0.001	0
Butyl glycidyl ether	0	0.108	0	0.163	0	0.103	0
4-Amino-m-cresol	1	0.925	1	0.996	1	0.930	1
Lyral	0	0.246	0	0.226	0	0.242	0
4,4,4-Trifluro-1-phenylbutane-1,3-dione	0	0.099	0	0.011	0	0.097	0
Oleyl methane sulphonate	0	0.014	0	0.000	0	0.013	0
Methyl hexadecyl sulphonate	0	0.007	0	0.000	0	0.006	0
1-Iodododecane	0	0.001	0	0.000	0	0.001	0
Furil	0	0.236	0	0.406	0	0.238	0
Chlorpromazine hydrochloride	1	0.971	1	0.994	1	0.975	1
Piperonyl butoxide	0	0.012	0	0.001	0	0.008	0
Abietic acid	0	0.040	0	0.006	0	0.039	0
2-Nitro-p-phenylenediamine	1	0.853	1	0.449	0	0.859	1
p-Methylhydrocinnamic aldehyde	0	0.010	0	0.000	0	0.010	0
1-Phenyloctane-1,3-dione	0	0.100	0	0.073	0	0.116	0
1-(2′,5′-Dimethylphenyl)butane-1,3-dione	0	0.296	0	0.432	0	0.335	0
Glycerol	0	0.738	1	0.653	1	0.740	1
Propylene glycol	0	0.158	0	0.088	0	0.160	0
b-Propriolactone	1	0.822	1	0.710	1	0.828	1
7,12-Dimethylbenz(a)anthracene	1	0.979	1	0.986	1	0.980	1
R(+)-Limonene	0	0.158	0	0.088	0	0.160	0
Sulphanilamide	0	0.784	1	0.169	0	0.785	1
1-Iodohexane	1	0.902	1	0.992	1	0.905	1
1-Bromononane	0	0.002	0	0.000	0	0.003	0
1-Bromoundecane	0	0.002	0	0.000	0	0.002	0
1-Methyl-3-nitro-1-nitrosoguanidine	1	0.782	1	0.968	1	0.788	1
1-Butanol	0	0.016	0	0.000	0	0.015	0
7-Bromotetradecane	0	0.001	0	0.000	0	0.001	0
a-Butyl cinnamic aldehyde	0	0.004	0	0.000	0	0.004	0
Isopropyl isoeugenol	1	0.216	0	0.312	0	0.212	0
1-Bromotridecane	0	0.001	0	0.000	0	0.001	0
Linalool	0	0.053	0	0.003	0	0.054	0
Lilial(*p*-tert-butyl-a-ethyl hydrocinnamal	0	0.012	0	0.001	0	0.012	0
Diethylphthalate	0	0.156	0	0.366	0	0.156	0
Hexahydrophthalic anhydride	1	0.880	1	0.979	1	0.888	1
Phthalic anhydride	1	0.278	0	0.654	1	0.276	0
2-Acetylcyclohexanenone	0	0.220	0	0.014	0	0.217	0
Diphenylcyclopropenone	1	0.675	1	0.803	1	0.680	1
Coumarin	0	0.624	1	0.424	0	0.630	1
Predicted accuracy (%)		87.5		92.5		87.5	

component formed from the EMAX matrix compared to the PLS components in the GMAX and GEMAX matrices.

Overall, the prediction accuracies are exactly the same for the three datasets; 95.0% for the training set and 90.9% for the test set. Both values are higher than those of the LR 2-state model. This suggests that the extracted PLS components from each of the original GMAX, EMAX and GEMAX matrices *do* include more comprehensive 4D-fingerprint information than realized in a LR model of individual single 4D-fingerprints. Because the training and

Table 2 Predicted probabilities and corresponding classifications for the test set using LR 2-state categorical QSAR models for the calculated data from the ground state (GMAX), excited state (EMAX) and the combination of ground and excited state(GEMAX) descriptor pools

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted class	Predicted probability	Predicted class
2-Hydroxypropyl methacrylate	0	0.173	0	0.041	0	0.162	0
6-Methylcoumarin	0	0.381	0	0.031	0	0.384	0
Phenyl Benzoate	0	0.849	1	0.916	1	0.856	1
Ethyl benzoylacetate	0	0.158	0	0.216	0	0.179	0
Benzocaine	0	0.171	0	0.111	0	0.164	0
Benzoyl peroxide	1	0.461	0	0.002	0	0.476	0
2-Aminophenol	1	0.726	1	0.929	1	0.734	1
Eugenol	0	0.391	0	0.140	0	0.384	0
2-Ethyl butaldehyde	0	0.005	0	0.005	0	0.004	0
Benzoyl chloride	1	0.269	0	0.775	1	0.276	0
4-Hydrobenzoic acid	0	0.683	1	0.813	1	0.688	1
Predicted accuracy (%)		63.6		72.7		63.6	

test sets are randomly selected from the original dataset, the best way to compare the significance of the models from the GMAX, EMAX and GEMAX matrices is to make test set predictions by employing additional compounds.

Based on the above results, a smaller training set of 63 compounds was randomly selected from the original training set. This then led to an augmented test set having an additional 28 compounds. The same PLS-CLR method was applied to the new training set using the corresponding new GMAX, EMAX and GEMAX matrices. The resulting PLS-CLR QSAR models are Eqs. 20–22 for GMAX, EMAX and GEMAX, respectively. All three QSAR models contain four PLS components.

$$Logit(P_1) = -3.531 + 1.406 * xscr1 + 1.488 * xscr5 + 1.384 * xscr9 + 2.856 * xscr13 \quad (20)$$

$$Logit(P_1) = -1.391 + 0.567 * xscr1 + 0.404 * xscr3 + 0.565 * xscr9 + 0.664 * xscr11 \quad (21)$$

$$Logit(P_1) = -2.633 + 0.593 * xscr1 + 0.504 * xscr3 + 0.752 * xscr11 + 1.274 * xscr17 \quad (22)$$

Comparing the PLS-CLR models above to the PLS-CLR models given by Eqs. 17–19 reveals that the PLS-CLR GMAX QSAR model (Eq. 20) shares the same descriptors as Eq. 17, and the corresponding regression coefficients of the two models differ only slightly. This suggests that reducing the size of the training set has little impact on QSAR model construction. The corresponding difference in the prediction accuracy of the two training sets indicates a slight decrease for the smaller training set, 89.2%, compared to 90.9%.

The PLS-CLR EMAX QSAR model (Eq. 21) contains one less descriptor than Eq. 18. Both QSAR models have two identical descriptors, *xscr*1 and *xscr*9. Also, fewer descriptors in Eq. 21 seemingly lead to a lower prediction accuracy for the training set: 92.1% (Eq. 21) compared to 95.0% (Eq. 18). Still the PLS-CLR QSAR model yields higher predictivity for the test set: 96.4% compared to 90.9% for Eq. 18.

The 4D-FP QSAR model generated from the excited state data has better predictivity than the corresponding ground state model. The GEMAX QSAR model (Eq. 22) is distinct from Eq. 19 and also has better predictivity for both the training and test sets than in Eq. 19. The predicted probability of being a "strong-extreme sensitizer" for both the training and test sets, as well as the predicted accuracy measures for Eqs. 20–22, are listed in Tables 3 and 4, respectively. The higher prediction accuracy found for the augmented test set for the models built from the EMAX and GEMAX datasets indicates that the 4D-FP information derived from the excited state geometry and charge distribution are significant in constructing optimal 2-state models. This also suggests that when downsizing the training set from 91 to 63 compounds there is no correspondingly diminished capacity to construct QSAR models that meaningfully differentiate "sensitizers" from "non-sensitizers" for the GAMX, EMAX and GEMAX datasets.

PLS-CLR analysis for building 3-state models

With success at distinguishing the predictive behaviors of the EMAX and GEMAX PLS-CLR 2-state models using the augmented test set, this set (and corresponding training set) was also employed in the 3-state modeling

Table 3 Predicted probabilities and corresponding classifications for training set using the PLS-CLR 2-state categorical QSAR models for the calculated data from the ground state (GMAX), excited state (EMAX), and the ground and excited state(GEMAX) descriptors based on the diminished training sets

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted probability	Predicted class	Predicted probability
4-Methoxyacetophenone	0	0.000	0	0.029	0	0.020	0
4-Nitrobenzyl bromide	1	0.997	1	0.952	1	0.963	1
Benzyl bromide	1	0.760	1	0.435	0	0.656	1
Benzaldehyde	0	0.000	0	0.063	0	0.001	0
Cyclamen aldehyde	0	0.029	0	0.110	0	0.003	0
Cinnamic alcohol	0	0.000	0	0.058	0	0.000	0
1, 4-Phenylenediamine	1	0.998	1	0.997	1	1.000	1
p-Benzoquinone	1	1.000	1	0.996	1	0.991	1
Hydroxycitronellal	0	0.250	0	0.012	0	0.004	0
Maleic anhydride	1	1.000	1	0.981	1	0.934	1
3-Phenylenediamine	1	0.789	1	0.980	1	1.000	1
1-Chloromethylpyrene	1	1.000	1	0.966	1	0.999	1
2,4,6-Trichloro-1,3,5-triazine (cyanuric chloride)	1	0.983	1	0.999	1	0.930	1
Chlorobenzene	0	0.000	0	0.023	0	0.000	0
1-Bromobutane	0	0.000	0	0.015	0	0.000	0
Hexane	0	0.087	0	0.228	0	0.009	0
2,2,6,6-Tetramethyl-heptane-3,5-dione	0	0.019	0	0.145	0	0.445	0
1-Bromooctadecane	0	0.000	0	0.001	0	0.000	0
Benzyl benzoate	0	0.446	0	0.151	0	0.030	0
Ethyl vanillin	0	0.003	0	0.007	0	0.001	0
Propyl gallate	1	1.000	1	0.822	1	0.879	1
a-Amyl cinnamic aldehyde	0	0.001	0	0.203	0	0.002	0
Hydroquinone	1	1.000	1	0.998	1	0.972	1
Octanoic acid	0	0.002	0	0.006	0	0.001	0
Dodecylthiosulphonate	1	0.106	0	0.028	0	0.016	0
1-(3′,4′,5′-Trimethoxyphenyl)-4-dimethylpentane-1, 3,-dione	0	0.005	0	0.003	0	0.000	0
5-Methyl-2,3-hexanedione	0	0.127	0	0.421	0	0.724	1
4-Nitroso-*N*,*N*-dimethylaniline	1	0.984	1	0.553	1	0.969	1
4-Allylanisole	0	0.000	0	0.089	0	0.038	0
Ethyl acrylate	0	0.000	0	0.042	0	0.002	0
1-Bromododecane	0	0.076	0	0.005	0	0.000	0
Oxalic acid	0	0.009	0	0.533	1	0.130	0
2-Mercaptobenzothiazole	0	0.001	0	0.084	0	0.000	0
5,5-Dimethyl-3-thiocyanatomethyl-2(3H)-furanone	1	0.341	0	0.285	0	0.963	1
3-Ethoxy-1-(2',3',4',5'-tetramethylphenyl)propane-1, 3-dione	0	0.128	0	0.270	0	0.000	0
C11-Azlactone	0	0.000	0	0.001	0	0.000	0
6-Methyleugenol	0	0.000	0	0.017	0	0.013	0
5-Methyleugenol	0	0.383	0	0.069	0	0.073	0
3-Methyleugenol	0	0.021	0	0.040	0	0.011	0
5-Methylisoeugenol	1	0.126	0	0.062	0	0.598	1
cis-6-Nonenal	0	0.000	0	0.036	0	0.006	0
1-Chlorotetradecane	0	0.335	0	0.007	0	0.000	0
Butyl glycidyl ether	0	0.422	0	0.060	0	0.138	0

Table 3 continued

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted probability	Predicted class	Predicted probability
4-Amino-m-cresol	1	1.000	1	0.964	1	0.999	1
Lyral	0	0.372	0	0.013	0	0.000	0
4,4,4-Trifluro-1-phenylbutane-1,3-dione	0	0.000	0	0.323	0	0.000	0
Oleyl methane sulphonate	0	0.000	0	0.001	0	0.000	0
Methyl hexadecyl sulphonate	0	0.000	0	0.000	0	0.000	0
1-Iodododecane	0	0.089	0	0.005	0	0.000	0
Furil	0	0.000	0	0.032	0	0.469	0
Chlorpromazine hydrochloride	1	0.974	1	0.512	1	0.605	1
Piperonyl butoxide	0	0.000	0	0.000	0	0.000	0
Abietic acid	0	0.000	0	0.000	0	0.000	0
2-Nitro-p-phenylenediamine	1	0.954	1	0.892	1	1.000	1
p-Methylhydrocinnamic aldehyde	0	0.002	0	0.226	0	0.097	0
1-Phenyloctane-1,3-dione	0	0.113	0	0.437	0	0.098	0
1-(2′,5′-Dimethylphenyl)butane-1,3-dione	0	0.010	0	0.378	0	0.211	0
Glycerol	0	0.059	0	0.301	0	0.175	0
Propylene glycol	0	0.002	0	0.084	0	0.013	0
b-Propriolactone	1	0.996	1	0.995	1	0.662	1
7,12-Dimethylbenz(a)anthracene	1	1.000	1	0.942	1	0.980	1
R(+)-Limonene	0	0.002	0	0.084	0	0.013	0
Sulphanilamide	0	0.000	0	0.032	0	0.153	0
Predicted accuracy (%)		95.2		92.1		96.8	

analysis. The third state, as compared to the 2-state models, includes "moderate" skin-sensitizers in the GMAX, EMAX and GEMAX datasets. The PLS-CLR methodology was used to build the 3-state models in this study.

The optimal 3-state models are given by Eqs. 23–28 for GMAX, EMAX and GEMAX datasets, respectively.

GMAX

$$\begin{aligned} Logit(P_2) = &-2.425 + 0.216 * xscr1 + 0.289 * xscr4 \\ &+ 0.432 * xscr5 + 0.446 * xscr10 \\ &+ 0.379 * xscr12 \end{aligned} \quad (23)$$

$$\begin{aligned} Logit(P_2 + P_1) = &\ 0.215 + 0.216 * xscr1 + 0.289 * xscr4 \\ &+ 0.432 * xscr5 + 0.446 * xscr10 \\ &+ 0.379 * xscr12 \end{aligned} \quad (24)$$

EMAX

$$\begin{aligned} Logit(P_2) = &-2.450 + 0.342 * xscr1 + 0.249 * xscr2 \\ &+ 0.366 * xscr3 + 0.187 * xscr4 \\ &+ 0.193 * xscr7 + 0.625 * xscr9 \\ &+ 0.356 * xscr12 + 0.392 * xscr13 \\ &+ 0.312 * xscr14 \end{aligned} \quad (25)$$

$$\begin{aligned} Logit(P_2 + P_1) = &\ 0.639 + 0.342 * xscr1 + 0.249 * xscr2 \\ &+ 0.366 * xscr3 + 0.187 * xscr4 \\ &+ 0.193 * xscr7 + 0.625 * xscr9 \\ &+ 0.356 * xscr12 + 0.392 * xscr13 \\ &+ 0.312 * xscr14 \end{aligned} \quad (26)$$

GEMAX

$$\begin{aligned} Logit(P_2) = &-3.772 + 0.187 * xscr1 + 0.394 * xscr4 \\ &+ 0.637 * xscr5 + 0.745 * xscr10 \\ &- 0.609 * xscr11 + 0.570 * xscr12 \\ &+ 0.330 * xscr13 + 0.505 * xscr16 \\ &+ 0.606 * xscr18 \end{aligned} \quad (27)$$

$$\begin{aligned} Logit(P_2 + P_1) = &\ 0.326 + 0.187 * xscr1 + 0.394 * xscr4 \\ &+ 0.637 * xscr5 + 0.745 * xscr10 \\ &- 0.609 * xscr11 + 0.570 * xscr12 \\ &+ 0.330 * xscr13 + 0.505 * xscr16 \\ &+ 0.606 * xscr18 \end{aligned} \quad (28)$$

As described in the Material and methodology section, in a given dataset, $Logit(P_2)$ and $Logit(P_2 + P_1)$ have the

Table 4 Predicted probabilities and corresponding test set classifications using PLS-CLR 2-state QSAR models for the calculated data from GMAX, EMAX and GEMAX based on the augmented test set

Chemical name	Observed class	GMAX		EMAX		GEMAX	
		Predicted probability	Predicted class	Predicted probability	Predicted class	Predicted probability	Predicted class
1-Iodohexane	1	0.075	0	0.999	1	1.000	1
1-Bromononane	0	0.001	0	0.004	0	0.001	0
1-Bromoundecane	0	0.000	0	0.003	0	0.003	0
1-Methyl-3-nitro-1-nitrosoguanidine	1	1.000	1	1.000	1	0.002	0
1-Butanol	0	0.000	0	0.034	0	0.000	0
7-Bromotetradecane	0	0.000	0	0.004	0	0.001	0
a-Butyl cinnamic aldehyde	0	0.102	0	0.054	0	0.000	0
Isopropyl isoeugenol	1	0.048	0	0.455	0	0.934	1
1-Bromotridecane	0	0.000	0	0.007	0	0.149	0
Linalool	0	0.470	0	0.237	0	0.000	0
Lilial(*p*-tert-butyl-a-ethyl hydrocinnamal	0	0.000	0	0.122	0	0.001	0
Diethylphthalate	0	0.001	0	0.091	0	0.004	0
Hexahydrophthalic anhydride	1	0.998	1	0.635	1	0.999	1
Phthalic anhydride	1	0.434	0	0.861	1	0.585	1
2-Acetylcyclohexanenone	0	0.092	0	0.063	0	0.001	0
Diphenylcyclopropenone	1	0.997	1	0.992	1	1.000	1
Coumarin	0	0.151	0	0.117	0	0.000	0
2-Hydroxypropyl methacrylate	0	0.005	0	0.005	0	0.005	0
6-Methylcoumarin	0	0.130	0	0.259	0	0.000	0
Phenyl Benzoate	0	0.062	0	0.473	0	0.987	1
Ethyl benzoylacetate	0	0.151	0	0.184	0	0.153	0
Benzocaine	0	0.010	0	0.166	0	0.041	0
Benzoyl peroxide	1	0.834	1	0.967	1	1.000	1
2-Aminophenol	1	0.735	1	0.951	1	0.996	1
Eugenol	0	0.005	0	0.132	0	0.292	0
2-Ethyl butaldehyde	0	0.006	0	0.102	0	0.000	0
Benzoyl chloride	1	0.864	1	0.845	1	0.954	1
4-Hydrobenzoic acid	0	0.021	0	0.125	0	0.028	0
Predicted accuracy (%)		89.2		96.4		92.8	

same 4D-FP set and regression coefficient value for each PLS-component descriptor. The only difference between the equations is the regression constant. Moreover, the GMAX models (Eqs. 23, 24) are very similar to the GEMAX models (Eqs. 27, 28), and all descriptors in the GMAX models also appear in the GEMAX models. This suggests that the final GEMAX models not only contain all 4D-FP information from the GMAX dataset, but also 4D-FP information from EMAX. The EMAX models are not very similar to GMAX models, which may mean that the 4D-FP information inherent to the EMAX models is different from that contained in the GMAX model. The extracted PLS components in the GEMAX dataset are hybrids of both the GMAX and EMAX datasets. This hybrid composition of the EMAX PLS components is also present in the PLS-CLR 2-state models (Eqs. 17–22), but it is more obvious in the 3-state categorical QSAR models.

Prediction accuracy for the training set using Eqs. 23–28 for GMAX, EMAX and GEMAX are 63.4, 72.04 and 78.5%, respectively. Table 5 lists the predicted probabilities for being a "non-weaker" sensitizer, "moderate" sensitizer, and "strong-extreme" sensitizer for each test set compound. The prediction accuracy values for the test set using GMAX, EMAX and GEMAX, are 48.5, 87.9 and 72.7%, respectively. The GEMAX QSAR model generates the highest predicted accuracy for the training set of the three, but the EMAX QSAR model is most predictive for the test set. Both the EMAX and GEMAX models have greater prediction accuracy values than the

Table 5 Predicted probabilities and corresponding test set classifications using the PLS-CLR 3-state QSAR models for the calculated 4D-FP from GMAX, EMAX and the GEMAX based on augmented test set

Chemical name	Obs. Class	GMAX				EMAX				GEMAX			
		Pred. probability to be			Pred. class	Pred. probability to be			Pred. Class	Pred. probability to be			Pred. class
		2	1	0		2	1	0		2	1	0	
N-Methyl-*N*-nitrosourea	2	0.487	0.443	0.070	2	0.998	0.002	0.000	2	0.130	0.770	0.100	1
1-Bromononane	0	0.005	0.059	0.936	0	0.002	0.041	0.957	0	0.000	0.020	0.979	0
1-Bromoundecane	0	0.007	0.082	0.911	0	0.005	0.093	0.903	0	0.004	0.202	0.793	0
1-Bromodocosane	1	0.008	0.096	0.896	0	0.099	0.608	0.294	1	0.004	0.178	0.819	0
1-Methyl-3-nitro-1-nitrosoguanidine	2	0.000	0.000	1.000	0	1.000	0.000	0.000	2	0.000	0.000	1.000	0
1-Butanol	0	0.003	0.033	0.964	0	0.002	0.044	0.953	0	0.000	0.007	0.993	0
7-Bromotetradecane	0	0.017	0.182	0.801	0	0.013	0.212	0.775	0	0.001	0.034	0.965	0
a-Butyl cinnamic aldehyde	0	0.026	0.249	0.724	0	0.004	0.077	0.919	0	0.004	0.178	0.818	0
Isopropyl isoeugenol	2	0.415	0.494	0.091	1	0.375	0.555	0.071	1	0.406	0.570	0.024	1
N-ethyl-*N*-nitrosourea	1	0.029	0.264	0.708	0	0.835	0.156	0.009	2	0.194	0.741	0.065	1
1-Bromotridecane	0	0.023	0.223	0.754	0	0.029	0.367	0.604	0	0.013	0.434	0.553	0
Linalool	0	0.034	0.294	0.672	0	0.005	0.101	0.894	0	0.009	0.332	0.660	0
Lilial(*p-tert*-butyl-a-ethyl hydrocinnamal	0	0.123	0.539	0.338	1	0.043	0.454	0.503	0	0.030	0.622	0.348	1
Diethylphthalate	0	0.103	0.514	0.382	1	0.005	0.091	0.904	0	0.004	0.180	0.816	0
Hexahydrophthalic anhydride	2	0.362	0.526	0.112	1	0.490	0.464	0.045	2	0.343	0.626	0.031	1
Phthalic anhydride	2	0.465	0.459	0.076	2	0.573	0.394	0.033	2	0.674	0.318	0.008	2
2-Acetylcyclohexanenone	0	0.076	0.459	0.465	0	0.005	0.100	0.895	0	0.000	0.022	0.978	0
Diphenylcyclopropenone	2	0.959	0.038	0.003	2	0.995	0.005	0.000	2	0.973	0.027	0.000	2
1-Naphthol	1	0.055	0.393	0.552	0	0.088	0.591	0.321	1	0.005	0.221	0.774	0
Coumarin	0	0.085	0.480	0.435	1	0.012	0.202	0.786	0	0.006	0.272	0.722	0
2-Hydroxypropyl methacrylate	0	0.001	0.015	0.984	0	0.003	0.057	0.940	0	0.000	0.001	0.999	0
6-Methylcoumarin	0	0.072	0.448	0.481	0	0.012	0.201	0.787	0	0.001	0.052	0.947	0
2-methoxy-4-methylphenol	1	0.029	0.266	0.705	0	0.077	0.569	0.354	1	0.008	0.319	0.673	0
2-Phenyl Propionaldehyde	1	0.129	0.546	0.325	1	0.182	0.648	0.170	1	0.216	0.727	0.057	1
Phenyl Benzoate	0	0.559	0.387	0.053	2	0.515	0.444	0.041	2	0.009	0.351	0.640	0
Ethyl benzoylacetate	0	0.079	0.466	0.456	1	0.042	0.448	0.510	0	0.004	0.194	0.801	0
Benzocaine	0	0.011	0.128	0.860	0	0.017	0.258	0.725	0	0.000	0.003	0.997	0
Benzoyl peroxide	2	0.944	0.052	0.004	2	0.970	0.029	0.001	2	0.825	0.172	0.004	2
2-Aminophenol	2	0.410	0.497	0.093	1	0.865	0.128	0.007	2	0.999	0.001	0.000	2
Eugenol	0	0.049	0.368	0.583	0	0.170	0.648	0.182	1	0.027	0.601	0.371	1
2-Ethyl butaldehyde	0	0.050	0.374	0.576	0	0.014	0.228	0.757	0	0.000	0.021	0.979	0
Benzoyl chloride	2	0.650	0.313	0.037	2	0.887	0.107	0.006	2	0.934	0.065	0.001	2
4-Hydrobenzoic acid	0	0.006	0.076	0.917	0	0.027	0.352	0.621	0	0.001	0.040	0.959	0
Predicted accuracy (%)		60.6				87.9				72.7			

The definitions in the "*Pred. Probability to be"* columns are: 2 means "strong-extreme" sensitizer, 1 means "moderate" sensitizer, 0 means "non-weak" sensitizer

GMAX model yields for the training and test sets. These results indicate that the extracted PLS components from the excited state 4D-FP provide significant information for the reliable categorical classification of skin-sensitization potency across a structurally diverse set of compounds.

PLS-CLR analysis for building two-2-state models

As described in the Material and methodology section, the two-2-state models were constructed using a two-step process. The first step uses the entire training set divided into two sets: "non-weak sensitizers" assigned a 0 classification

value and "moderate-strong-extreme sensitizers" with a value of 1. A PLS-CLR 2-state model is then constructed from each of the GMAX, EMAX, and GEMAX datasets. If the predicted class of a training set molecule correctly fits the observed class using the PLS-CLR 2-state model, this molecule is then employed in the second step of the model-building process. The PLS-CLR 2-state models of the first step for GMAX, EMAX and GEMAX datasets are, respectively,

$$\begin{aligned} Logit(P_1) = {} & 1.035 + 0.699 * xscr1 + 0.398 * xscr2 \\ & + 0.366 * xscr3 + 0.234 * xscr6 \\ & + 0.291 * xscr7 + 0.589 * xscr8 \\ & + 0.826 * xscr11 + 0.851 * xscr17 \end{aligned} \quad (29)$$

$$\begin{aligned} Logit(P_1) = {} & 1.263 + 0.734 * xscr1 + 0.618 * xscr3 \\ & + 0.601 * xscr4 + 0.338 * xscr6 \\ & + 0.825 * xscr8 + 0.534 * xscr9 \\ & + 0.626 * xscr12 + 0.630 * xscr16 \\ & + 0.981 * xscr17 + 0.811 * xscr20 \end{aligned} \quad (30)$$

$$\begin{aligned} Logit(P_1) = {} & 1.084 + 0.488 * xscr1 + 0.298 * xscr2 \\ & + 0.310 * xscr3 + 0.148 * xscr6 \\ & + 0.190 * xscr7 + 0.499 * xscr8 \\ & + 0.449 * xscr11 + 0.748 * xscr17 \end{aligned} \quad (31)$$

A total of 108, 111 and 108 compounds were predicted correctly by Eqs. 29, 30 and 31, respectively. From the correctly predicted compounds, random selection is used to create the GMAX, EMAX and GEMAX training and test sets for the second step in the model-building process. In order to compare the two-2-state models built from the three datasets, the same number of training set compounds is used in all three datasets, and the remainder of the compounds are put into the test sets. Table 6 shows the distributions of "non-weak sensitizers" and "moderate-strong-extreme sensitizers" across the three datasets, as well as the total numbers of training set and test set compounds. The "non-weak sensitizer" compound sets and the "moderate-strong-extreme sensitizer" compound sets were used to perform individual PLS-CLR 2-state analyses.

Step-2 PLS-CLR analysis of the "non" and "weak" sensitizer sets

PLS-CLR analysis was performed using the correctly predicted "non-weak sensitizer" compounds in the first step of building the two-2-step QSAR models for each of the three datasets. The second-step, 2-state QSAR models for the GMAX, EMAX and GEMAX datasets are given by Eqs.(32–34), respectively,

$$\begin{aligned} Logit(P_1) = {} & 2.175 + 0.405 * xscr3 + 0.341 * xscr4 \\ & + 0.456 * xscr8 \end{aligned} \quad (32)$$

$$\begin{aligned} Logit(P_1) = {} & 2.300 + 0.362 * xscr3 + 0.295 * xscr4 \\ & + 0.521 * xscr8 \end{aligned} \quad (33)$$

$$\begin{aligned} Logit(P_1) = {} & 3.933 + 0.399 * xscr1 + 0.399 * xscr2 \\ & + 0.212 * xscr4 + 0.579 * xscr8 \end{aligned} \quad (34)$$

The GMAX model has he exact same descriptors as the EMAX model, but there are slight differences in the regression coefficient values and the regression constant. QSAR models Eqs. 32 and 33 also show similar prediction accuracies for the training set, which are reported in Table 6. The predicted probability and corresponding classification of each compound in the training and test set are listed in Tables 7a and 8a.

Compared to the GMAX dataset, three more compounds are correctly categorized for the EMAX dataset in the first

Table 6 Total number of compounds in the training test sets and the predicted accuracies for the training and test sets based on Eqs. (32–37) using the 4D-FP data from the GMAX, EMAX and GEMAX datasets

		Non sensitizer	Weak sensitizer	Moderate sensitizer	Strong-extreme sensitizer
GMAX	Training set	15	24	20	19
	Pred. accuracy	76.90%		79.4%	
	Test set	9	6	7	8
	Pred. accuracy	66.7%		73.3%	
EMAX	Training set	14	25	20	19
	Pred. accuracy	79.4%		76.9%	
	Test set	9	9	6	9
	Pred. accuracy	72.2%		86.7%	
GEMAX	Training set	13	26	20	19
	Pred. accuracy	82.1%		87.8%	
	Test set	10	8	4	8
	Pred. accuracy	77.8%		91.7%	

step of the two-2-state model-building process. But this time the same number of training set compounds was used for all three datasets in the second step of the two-2-state model construction. Thus, the EMAX test set contains three additional compounds. Two of these compounds, phenyl benzoate and eugenol, are found incorrectly categorized by Eq. 29 in the GMAX PLS-CLR analysis, but are correctly predicted by Eq. 30 in the EMAX PLS-CLR analysis. Therefore, these two compounds do not appear in GMAX test set but are found in the EMAX test set. Further, each compound is also correctly categorized by Eq. 33, of the second step in building the EMAX two-2-state model.

One compound, coumarin, is correctly categorized by Eq. 29 in the first step, but is incorrectly predicted by Eq. 30. Therefore, this coumarin appears in the GMAX test set but not the EMAX test set. Also, this compound is incorrectly categorized by Eq. 32. Such behavior leads to an overall lower prediction accuracy for the GMAX test set compared to the test set prediction accuracies for both EMAX and GEMAX. Coumarin is correctly categorized by Eq. 31 in the first step, but not by Eq. 34. However, Eq. 34 does correctly categorize ethyl benzoylacetate, which is correctly categorized in the first step for all three datasets, but both Eqs. 32 and 33 incorrectly predict it. 1-bromononane appears in both the EMAX and GEMAX test sets. Only the GEMAX QSAR model (Eq. 33) correctly categorizes it. As seen in Fig. 1a, the prediction accuracy of the training set is higher in EMAX and GEMAX, and an 18% increase in specificity for GEMAX compared with GMAX models, although the sensitivity is decreased by 2.5% when comparing GEMAX with GMAX. These results, in composite, suggest that the EMAX and GEMAX QSAR models have a better predictivity than the GMAX QSAR model, based on both the training and test set findings for the "non" and "weak" sensitizer classifications.

Step-2 PLS-CLR analysis for the "moderate" and "strong-extreme" sensitizer sets

PLS-CLR analysis was applied to each of the three datasets based on the correctly predicted "moderate-strong-extreme sensitizer" compounds of the first step to building the two-2-state models. The resulting categorical models for discriminating "moderate" and "strong-extreme" sensitizers are listed below for the GMAX, EMAX and GEMAX datasets, respectively.

$$Logit(P_1) = -2.802 + 0.932 * xscr1 + 1.504 * xscr5 - 0.608 * xscr6 + 1.403 * xscr13 \quad (35)$$

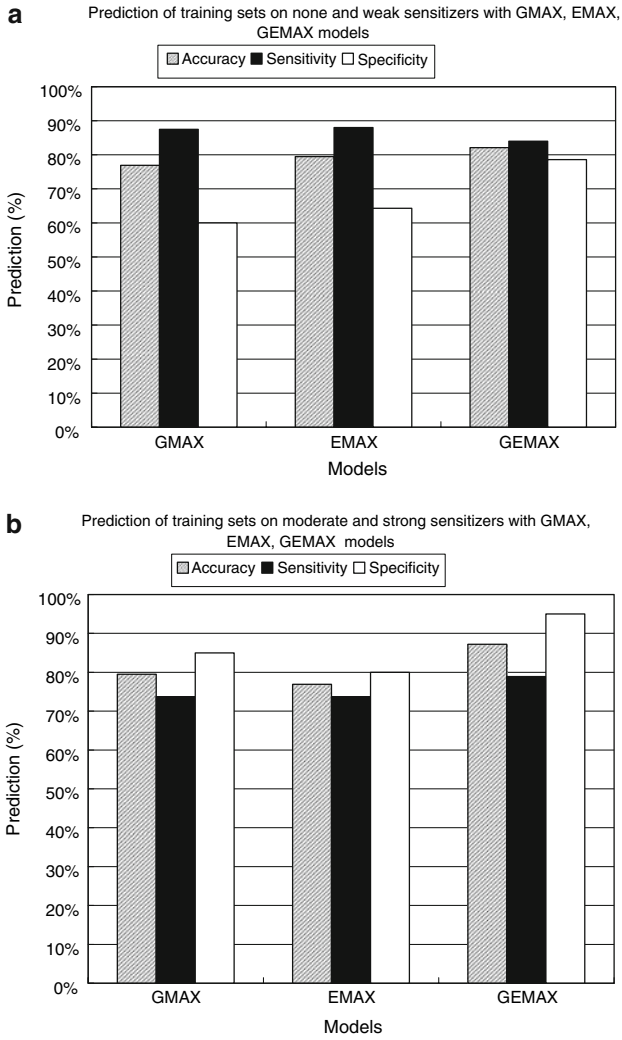

Fig. 1 Predicted accuracy, sensitivity, and specificity for (**a**) the non-weak sensitizers and (**b**) the moderate and strong sensitizers in training sets based on PLS-CLR two-2-state Eqs. 31–33 using the 4D-FP data from the GMAX, EMAX and GEMAX datasets

$$Logit(P_1) = -0.250 + 0.347 * xscr1 + 0.393 * xscr2 + 0.789 * xscr13 \quad (36)$$

$$Logit(P_1) = -0.282 + 0.215 * xscr1 + 0.247 * xscr2 + 0.512 * xscr13 \quad (37)$$

In this case, the EMAX model, Eq. 36, is quite similar to the GEMAX model, Eq. 37. Both have the same descriptors but slightly different corresponding regression coefficients and regression constants. They also have similar prediction accuracies for the training set, listed in Table 6. The predicted probability values of each compound in the training and test sets, using Eqs. 35–37, are listed in Tables 7b, and 8b respectively.

In general, most compounds in each of the three test sets for the GMAX, EMAX and GEMAX datasets are correctly

classified by Eqs. 35–37. 1-Methyl-3-nitro-1-nitrosoguanidine appears in all test sets of the three datasets. The GMAX categorical QSAR model, Eq. 35, incorrectly categorizes this compound, and, actually, predicts the probability of this compound being a strong-extreme sensitizer as 0. Conversely, the QSAR categorical models from the EMAX and GEMAX datasets correctly predict 1-methyl-3-nitro-1-nitrosoguanidine as being a strong-extreme sensitizer, with a probability of 1. Phthalic anhydride also appears in all three dataset test sets, and none of the models, Eqs. 35–37, correctly classify this compound.

Both 3-aminophenol and diethyl sulfate appear in the GMAX dataset test set, but not in the GEMAX dataset test set. 3-aminophenol also does not appear in the EMAX dataset test set. These two compounds are selected into their respective training sets. N-ethyl-N-nitrosourea appears in both the GMAX and EMAX test sets, but is incorrectly classified by Eqs. 35 and 36, respectively. This compound is not selected in the first step of PLS-CLR analysis for the GEMAX dataset.

It appears that the first-step PLS-CLR analysis in building the two-2-state models serves as a partial filter to eliminate some of the more difficult to characterize and predict, or "bad", sample compounds, thus enhancing the second-step PLS-analysis to yield models with high prediction accuracies for the "non" and "weak" sensitizers. Moreover, higher prediction accuracies realized for both EMAX and GEMAX training and test sets, as compared to the GMAX training and test set, for "non-" and "weak-" sensitizer classifications again suggests that both of the EMAX and GEMAX datasets include more significant 4D-FP information than is found in the GMAX dataset (see Figs. 1, 2). This result is consistent with previous studies, suggesting that the excited state of skin-sensitizing compounds can better access its "receptor" than the ground state structure of the compound [7].

Discussion

The novel aspect of this study has been to introduce the excited state 4D-FP descriptors, along with the ground state 4D-FP, into the initial descriptor pool to build categorical skin-sensitization potency QSAR models. By finding a way to compute 4D-FP for any state of a molecule provides (as far as we are aware) the only means to create a set of common and comparable universal QSAR descriptors across the electronic states available to a molecule. However, as to which electronic state(s) to include in a QSAR study cannot be addressed by the 4D-QSAR paradigm. In performing this study of skin sensitization we took into consideration that transport of the sensitizer to its target and its chemical reactivity are each integral components to the overall skin-sensitization mechanism of action. The choice of using the ground state to model transport is relatively clear, but the selection of the electronic state(s) to model reactivity is less apparent. Our working assumption was that the first excited state of each compound would be relevant in this study. We have felt justified in this approach because across the 2-state, 3-state and two-2-state categorical QSAR models, the inclusion of the EMAX and/or GEMAX 4D-FP descriptor sets with the GMAX dataset, or in some cases alone, generate better models than those produced solely from the GMAX 4D-FP descriptor sets.

The ground and excited state 4D-FP descriptors differ from one another because of the changes in the geometry and the electron charge density distribution that result when going from the ground state to an excited state. The actual distribution of IPE types across a molecule can also change for an excited state relative to the ground state because of the electron redistribution in the molecule. But in this study, the ground state and first excited states both have very similar 4D-FP descriptor sets, $\{\varepsilon i(u, v)\}$; this is illustrated for all test sets in Table 2, and the corresponding calculated $\varepsilon 1(any, np)$ and $\varepsilon 2(any, np)$ listed in Table 9. The correlation coefficient between $\varepsilon 1(any, np)g$ and $\varepsilon 1(any, np)e$ is as high as 0.997, and the correlation coefficient between $\varepsilon 2(any, np)g$ and $\varepsilon 2(any, np)e$ is as high as 0.992. High cross-correlations occur across the entire set of LLNA compounds for many ground and corresponding excited state 4D-FP descriptors, and this observation is not restricted to only the $\varepsilon 1(any, np)$ or $\varepsilon 2(any, np)$ 4D-FPs of the test set compounds. Nevertheless, the similarities and differences contained in the ground and excited state 4D-FP models (GMAX and EMAX) combine to lead to an optimized GEMAX model with a unique set of descriptor and coefficients that contains essential information from both the GMAX and the EMAX models. The 4D-FP descriptors can also differ in the LR models for the GMAX and EMAX datasets as is seen between Eqs. 14 and 15.

Still the difference between Eq. 14 and Eq. 15 is only the addition of a term, *+394.70*e7(np, hs)e*, and the ranking of importance and coefficient for the *(np, hs)* term. It is interesting that a nonpolar and non-hydrogen eigenvalue are constructive for the GMAX model (Eq. 14), yet destructive for the EMAX model (Eq. 15). Additionally, this eigenvalue is ranked in the top half (ranked 8th, 5th, and 8th for GMAX, EMAX, and GEMAX, respectively) for all model types. While these IPE types should not be influenced by shifts in electron density, or other electronic structure properties, it is interesting to see their markedly different use in the GMAX and EMAX models.

The extracted PLS components also exhibit high similarities and differences across the ground and excited state models and corresponding predictions; the predicted probability and class for models constructed from GMAX,

Table 7 Predicted probabilities and corresponding training set classifications using the PLS-CLR two-2-state QSAR models for the 4D-FP data from the GMAX, EMAX and GEMAX datasets for (a) "non" and "weak" sensitizers and (b) "moderate" and "strong-extreme" sensitizers

Chemical Name	Obs. class	GMAX		EMAX		GEMAX	
		Pred. probability	Pred. class	Pred. probability	Pred. class	Pred. probability	Pred. class
(a) *"Non" and "weak" sensitizers*							
4-Metoxyacetophenone	0	0.876	1	0.846	1	0.71	1
Benzaldehyde	0	0.697	1	0.729	1	0.793	1
Cyclamen aldehyde	1	–	–	0.978	1	–	–
Cinnamic alcohol	1	–	–	0.937	1	0.955	1
Hydroxycitronellal	1	0.819	1	0.884	1	0.705	1
Chlorobenzene	0	0.505	1	0.535	1	0.748	1
1-Bromobutane	0	0.038	0	0.068	0	0.081	0
Hexane	0	0.028	0	0.044	0	0.074	0
1-Bromooctadecane	1	0.981	1	0.987	1	0.999	1
Benzyl benzoate	1	0.74	1	0.594	1	1	1
Ethyl vanillin	0	0.295	0	0.441	0	0.173	0
a-Amyl cinnamic aldehyde	1	0.983	1	0.979	1	0.991	1
Octanoic acid	0	0.632	1	0.571	1	0.244	0
1-(3′,4′,5′-Trimethoxyphenyl)-4-dimethylpentane-1, 3,-dione	0	0.077	0	0.178	0	0.272	0
4-Allylanisole	1	0.915	1	0.919	1	0.921	1
Ethyl acrylate	1	0.698	1	0.687	1	0.428	0
1-Bromododecane	1	0.946	1	0.959	1	0.979	1
Oxalic acid	1	0.129	0	0.122	0	0.37	0
2-Mercaptobenzothiazole	1	0.903	1	0.952	1	0.974	1
3-Ethoxy-1-(2′,3′,4′,5′-tetramethylphenyl)propane-1, 3-dione	1	0.852	1	0.789	1	0.913	1
C15 Azlactone	1	0.881	1	0.832	1	0.96	1
6-Methyleugenol	1	0.892	1	0.934	1	0.901	1
3-Methyleugenol	1	0.918	1	0.951	1	0.947	1
cis-6-Nonenal	1	0.682	1	0.72	1	0.566	1
1-Chlorotetradecane	1	0.966	1	0.98	1	0.994	1
Butyl glycidyl ether	1	–	–	–	–	0.789	1
Lyral	1	0.965	1	0.966	1	0.955	1
4,4,4-Trifluro-1-phenylbutane-1,3-dione	1	0.428	0	0.438	0	0.194	0
Oleyl methane sulphonate	1	0.946	1	0.948	1	0.991	1
Methyl hexadecyl sulphonate	0	0.639	1	0.582	1	0.455	0
1-Iododododecane	1	0.944	1	0.96	1	0.978	1
Furil	0	0.319	0	0.247	0	0.097	0
Abietic acid	1	0.767	1	0.849	1	0.988	1
Piperonyl butoxide	0	0.011	0	0.028	0	0.039	0
1-Phenyloctane-1,3-dione	1	0.908	1	0.855	1	0.875	1
1-(2′,5′-Dimethylphenyl)butane-1,3-dione	1	0.907	1	0.874	1	0.813	1
Glycerol	0	0.135	0	0.177	0	0.368	0
Propylene glycol	0	0.158	0	0.184	0	0.266	0
R(+)- Limonene	1	0.158	0	0.184	0	0.266	0
Sulphanilamide	0	0.092	0	0.091	0	0.126	0
1-Bromononane	0	0.58	1	–	–	–	–
1-Bromoundecane	1	0.846	1	–	–	–	–

Table 7 continued

Chemical Name	Obs. class	GMAX		EMAX		GEMAX	
		Pred. probability	Pred. class	Pred. probability	Pred. class	Pred. probability	Pred. class
(b) *"Moderate" and "strong-extreme" sensitizers*							
4-Nitrobenzyl bromide	1	0.89	1	0.274	0	0.214	0
Benzyl bromide	1	0.464	0	0.149	0	0.216	0
a-Methyl cinnamic aldehyde	0	0.017	0	0.024	0	0.037	0
1, 4-Phenylenediamine	1	0.977	1	0.975	1	0.917	1
p-Benzoquinone	1	0.996	1	0.971	1	0.958	1
1-Chloromethylpyrene	1	1	1	1	1	1	1
Maleic anhydride	1	0.993	1	0.979	1	0.967	1
meta-Phenylene diamine	1	0.945	1	0.959	1	0.86	1
2,4,6-Trichloro-1,3,5-triazine (cyanuric chloride)	1	0.318	0	0.367	0	0.262	0
3-Dimethylaminopropylamine	0	0.045	0	0.589	1	0.456	0
Diethylenetriamine	0	0.032	0	0.396	0	0.38	0
Palmitoyl chloride	0	–	–	0.515	1	0.496	0
3,4-Dihydrocoumarin	0	0.384	0	0.167	0	0.189	0
Propyl gallate	1	0.997	1	0.874	1	0.928	1
Benzylidene acetone (4-phenyl-3-buten-2-one)	0	0.045	0	0.046	0	0.067	0
Phenylacetaldehyde	0	0.27	0	0.053	0	0.086	0
Hydroquinone	1	0.775	1	0.964	1	0.954	1
Dodecylthiosulphonate	1	0.236	0	0.685	1	0.686	1
Vinyl pyridine	0	0.154	0	0.045	0	0.09	0
Tetramethylthiuram disulfide	0	0.481	0	0.241	0	0.185	0
4-Nitroso-N,N-dimethylaniline	1	0.903	1	0.428	0	0.502	1
Diethyl maleate	0	0.157	0	0.593	1	0.488	0
Trans-cinnamaldehyde	0	0.268	0	0.109	0	0.173	0
5,5-Dimethyl-3-thiocyanatomethyl-2(3H)-furanone	1	0.701	1	0.291	0	0.295	0
Bisphenol A-diglycidyl ether	0	0	0	0.118	0	0.229	0
1-(2′,5′-Diethylphenyl)butane-1,3-dione	0	–	–	0.371	0	–	–
3-Methylisoeugenol	0	0.252	0	–	–	–	–
5-Methylisoeugenol	1	0.32	0	0.687	1	0.6	1
perilla aldehyde	0	0.277	0	0.129	0	0.138	0
1,2-Benzisothiazolin-3-one (Proxel active)	0	0.025	0	–	–	–	–
2-Methyl-2H-Isothiazol-3-one	0	0.008	0	0.216	0	0.265	0
4-Amino-m-cresol	1	0.884	1	0.795	1	0.848	1
2-(4-Amino-2-nitroanilino)-ethanol	0	0.826	1	0.113	0	0.218	0
5,5-Dimethyl-3-methylene-dihydro-2(3H)-furanone	0	0.507	1	0.285	0	0.311	0
12-Bromo-1-dodecanol	0	0.245	0	–	–	–	–
3,5,5-Trimethylhexanoyl chloride	0	0.543	1	0.21	0	0.202	0
1-Bromoeicosane	0	0	0	0.034	0	0.048	0
Chlorpromazine hydrochloride	1	0.892	1	0.992	1	0.988	1
2-Nitro-p-phenylenediamine	1	0.976	1	0.943	1	0.834	1
b-Propriolactone	1	0.197	0	0.578	1	0.604	1
7,12-Dimethylbenz(a)anthracene	1	1	1	1	1	1	1
3-Aminophenol	0	–	–	0.838	1	0.881	1
Diethyl sulfate	0	–	–	–	–	0.288	0

A "–" in the table means the compound is not part of this test set

Table 8 Predicted probabilities and corresponding test set classifications using the PLS-CLR two-2-state QSAR models for the 4D-FP data from the GMAX, EMAX and GEMAX datasets for (a) "non" and "weak" sensitizers and (b) "moderate" and "strong-extreme" sensitizers

Chemical Name	Obs. class	GMAX		EMAX		GEMAX	
		Pred. probability	Pred. class	Pred. probability	Pred. class	Pred. probability	Pred. class
(a) *"Non" and "weak" sensitizers*							
1-Bromononane	0	–	–	0.643	1	0.477	0
1-Bromoundecane	1	–	–	0.853	1	0.881	1
1-Butanol	0	0.129	0	0.344	0	0.275	0
7-Bromotetradecane	1	0.987	1	0.992	1	0.998	1
a-Butyl cinnamic aldehyde	1	0.929	1	0.925	1	0.921	1
1-Bromotridecane	1	0.968	1	0.979	1	0.992	1
Linalool	1	0.872	1	0.876	1	0.894	1
Lilial(p-*tert*-butyl-a-ethyl hydrocinnamal	1	0.989	1	0.985	1	0.997	1
Diethylphthalate	0	0.496	0	0.437	0	0.339	0
2-Acetylcyclohexanenone	0	0.871	1	0.842	1	0.650	1
Coumarin	0	0.958	1	–	–	0.918	1
2-Hydroxypropyl methacrylate	0	0.051	0	0.102	0	0.015	0
6-Methylcoumarin	0	0.898	1	0.877	1	0.831	1
Phenyl Benzoate	1	–	–	0.767	1	1.000	1
Ethyl benzoylacetate	0	0.639	1	0.598	1	0.342	0
Benzocaine	0	0.610	1	0.697	1	0.690	1
Eugenol	1	–	–	0.918	1	–	–
2-Ethyl butaldehyde	1	0.891	1	0.921	1	0.889	1
4-Hydrobenzoic acid	0	0.361	0	0.415	0	0.428	0
(b) *"Moderate" and "strong-extreme" sensitizers*							
3-Aminophenol	0	0.947	1	–	–	–	–
Diethyl sulfate	0	0.158	0	0.352	0	–	–
3-Bromomethyl-5,5-dimethyl-dihydro-2(3H)-furanone	0	0.342	0	0.258	0	0.251	0
N-Methyl-*N*-nitrosourea	1	0.829	1	0.979	1	0.964	1
1-Bromodocosane	0	0.000	0	0.452	0	0.428	0
1-Methyl-3-nitro-1-nitrosoguanidine	1	0.000	0	1.000	1	1.000	1
Isopropyl isoeugenol	1	–	–	0.785	1	–	–
N-ethyl-*N*-nitrosourea	0	0.998	1	0.992	1	–	–
Hexahydrophthalic anhydride	1	0.749	1	0.594	1	0.519	1
Phthalic anhydride	1	0.138	0	0.073	0	0.096	0
Diphenylcyclopropenone	1	1.000	1	0.974	1	0.972	1
1-Naphthol	0	0.329	0	0.300	0	0.314	0
2-Phenylpropionaldehyde	0	0.338	0	0.115	0	0.182	0
Benzoyl peroxide	1	0.995	1	0.935	1	0.866	1
2-Aminophenol	1	0.886	1	0.791	1	0.848	1
Benzoyl chloride	1	0.978	1	0.985	1	0.539	1

A "–" in the table means the compound is not part of this test set

EMAX and GEMAX descriptors are shown in Table 2 and the corresponding PLS components (*xscr*1, *xscr*2, *xscr*7 and *xscr*8) are listed in Table 10. The first 6 PLS components for the GMAX and EMAX datasets are similar with divergence of the components beginning with the seventh component. Moreover, the GEMAX has some extracted PLS components that differ from those of both GMAX and EMAX, while at the same time sharing similar PLS components. This dual nature among the 4D-FP may explain why Eqs. 32 and 33 and Eqs. 36 and 37 are similar.

Table 9 The calculated $\varepsilon 1(any, np)$ and $\varepsilon 2(any, np)$ 4D-FP values for all test set compounds listed in Table 2 for both the GMAX and EMAX datasets

Chemical name	$\varepsilon 1(any, np)g$	$\varepsilon 1(any, np)e$	$\varepsilon 2(any, np)g$	$\varepsilon 2(any, np)e$
2-Hydroxypropyl methacrylate	0.4160	0.4190	0.1224	0.1236
6-Methylcoumarin	0.4310	0.4310	0.0994	0.0994
Phenyl Benzoate	0.3391	0.3375	0.1170	0.1177
Ethyl benzoylacetate	0.3499	0.3422	0.0957	0.0999
Benzocaine	0.3697	0.3705	0.0894	0.0890
Benzoyl peroxide	0.3065	0.3028	0.0916	0.0920
2-Aminophenol	0.5614	0.5607	0.1126	0.1123
Eugenol	0.3848	0.3864	0.0940	0.0948
2-Ethyl butaldehyde	0.4557	0.4611	0.0799	0.0797
Benzoyl chloride	0.5400	0.5398	0.1063	0.1064
4-Hydrobenzoic acid	0.5045	0.5048	0.0829	0.0826

Table 10 The extracted PLS components *xscr*1, *xscr*2, *xscr*7 and *xscr*8 from the 4D-FP descriptors for all test set compounds listed in Table 2 for the GMAX, EMAX and GEMAX datasets

Chemical name	*xscr*1			*xscr*2			*xscr*7			*xscr*8		
	GMAX	EMAX	GEMAX	GMAX	EMAX	GEMAX	GMAX	EMAX	GEMAX	GMAX	EMAX	GEMAX
2-Hydroxypropyl methacrylate	−1.435	−1.483	−2.063	−4.608	−4.513	−6.450	−0.445	−0.012	−0.360	−3.420	−2.556	−4.251
6-Methylcoumarin	0.938	0.959	1.341	−2.796	−2.717	−3.898	−1.144	−0.920	−1.448	−2.220	−2.368	−3.212
Phenyl benzoate	6.815	6.770	9.607	6.826	6.451	9.390	−2.633	−2.396	−3.532	−2.759	−3.286	−4.282
Ethyl benzoylacetate	−1.071	−1.044	−1.495	−3.508	−3.438	−4.912	−0.698	−0.582	−0.926	−2.983	−3.219	−4.414
Benzocaine	2.215	2.162	3.095	−7.554	−7.513	−10.656	−0.378	−0.280	−0.517	−0.123	0.007	−0.062
Benzoyl peroxide	6.143	6.160	8.700	2.419	2.314	3.348	2.355	2.640	3.512	−1.535	−2.048	−2.640
2-Aminophenol	7.992	7.946	11.271	−1.138	−1.072	−1.560	0.177	0.005	0.127	1.873	2.041	2.871
Eugenol	−0.287	−0.295	−0.412	−4.361	−4.289	−6.117	4.587	4.816	6.624	1.161	2.037	2.274
2-Ethyl butaldehyde	−2.003	−2.024	−2.848	0.149	0.192	0.241	−0.293	−0.190	−0.330	0.179	0.553	0.509
Benzoyl chloride	2.140	2.197	3.067	0.440	0.551	0.703	2.155	2.116	3.036	−0.512	−0.660	−0.833
4-Hydrobenzoic acid	4.553	4.515	6.413	−7.542	−7.424	−10.583	−1.494	−1.235	−1.986	−2.528	−2.400	−3.477

While building the two-2-state models, PLS-CLR analysis in the first step filtered out "bad" sample compounds, those incorrectly classified skin sensitizers of the LLNA dataset. If all of the "non" and "weak" sensitizers, including those incorrectly classified in this first step, are used to build a second step model to separate "non" and "weak" sensitizers–again, by performing PLS-CLR analysis–the resulting 2-state ground, excited and combined QSAR models are as follows:

$$Logit(P_1) = 3.578 + 0.536 * xscr2 + 1.019 * xscr5 + 0.663 * xscr7 + 1.129 * xscr10 \quad (38)$$

$$Logit(P_1) = 2.039 + 0.520 * xscr5 + 1.167 * xscr11 - 1.158 * xscr17 \quad (39)$$

$$Logit(P_1) = 2.694 + 0.198 * xscr4 + 0.551 * xscr8 + 0.434 * xscr15 \quad (40)$$

These models differ from their corresponding models, namely Eqs. 32–34, when the "bad" compounds are excluded in the second step of model construction. Moreover, the predicted accuracies for these three models, Eqs. 38–40, for the GMAX, EMAX and GEMAX test sets are 66.7, 66.7, and 70.8%, respectively. Each prediction accuracy is obviously lower than the corresponding prediction accuracy when the "bad" compounds are excluded (see Table 6), especially for GEMAX. The predicted sensitivity for these three models, Eqs. 38–40, for the GMAX, EMAX and GEMAX test sets are 100%, and the predicted specificity is 44.4% for both and GMAX and EMAX and 60.0% for GEMAX (see Fig. 2a).

Similarly, if all "moderate" and all "strong-extreme" sensitizers, including the "bad" compounds found in the first step, are used to build the second step two-2-state

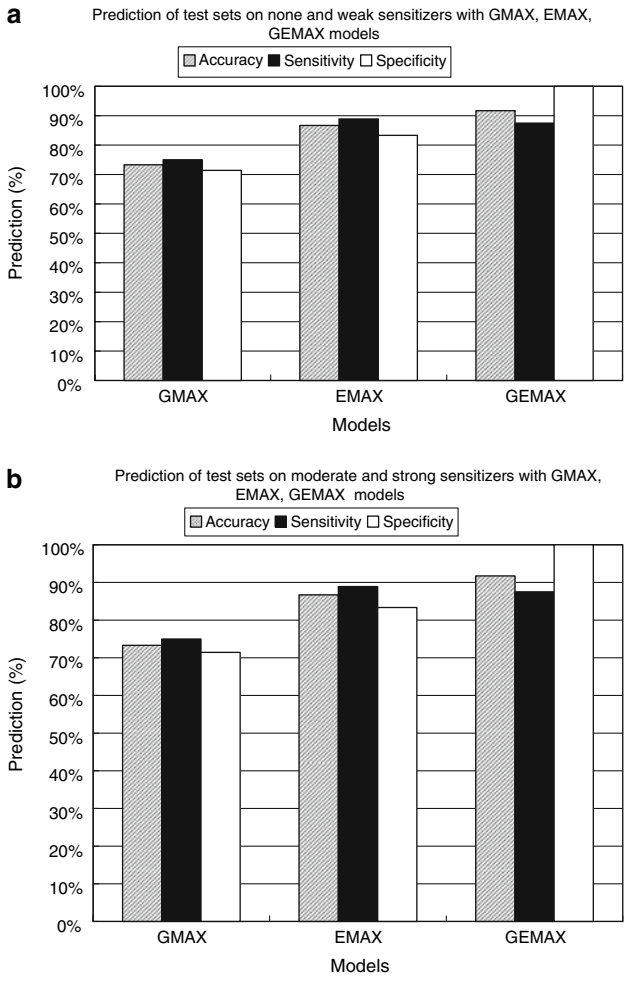

Fig. 2 (**a**) Predicted accuracy, sensitivity, and specificity for none and weak sensitizers in the test sets based on PLS-CLR two-2-state Eqs. 34–37 using the 4D-FP data from the GMAX, EMAX and GEMAX datasets. (**b**) Predicted accuracy, sensitivity, and specificity for moderate and strong sensitizers in the test sets based on PLS-CLR two-2-state Eqs. 34–36 using the 4D-FP data from the GMAX, EMAX and GEMAX datasets

models with PLS-CLR analysis, the following GMAX, EMAX and GEMAX models result:

$$Logit(P_1) = -3.675 + 1.028 * xscr1 + 1.163 * xscr5 + 1.654 * xscr18 + 1.497 * xscr20 \quad (41)$$

$$Logit(P_1) = -2.045 + 0.377 * xscr2 + 0.405 * xscr3 + 0.401 * xscr4 + 0.767 * xscr14 + 0.982 * xscr17] \quad (42)$$

$$Logit(P_1) = -3.365 + 0.788 * xscr1 + 1.114 * xscr5 - 0.503 * xscr6 + 0.985 * xscr13 \quad (43)$$

All of these models differ from the corresponding models, namely Eqs. 35–37, when the "bad" compounds are excluded in the second step of construction. Prediction accuracy values for the test sets of each model given by Eqs. 35–37 are 66.7, 70.8 and 79.1%, respectively. These prediction accuracies are also significantly lower than those in Table 6. The predicted sensitivity for these three models, Eqs. 38–40 for the GMAX, EMAX and GEMAX test sets are 75.0, 88.9, and 87.5%, and predicted specificity are 71.4, 88.9% for both GMAX and EMAX and 100.0% for GEMAX, (Fig. 2b). These findings suggest that the first step PLS-CLR analysis, in addition to building a categorical QSAR model, is also an effective method for filtering out "bad" compounds from the training sets used in step two of the two-2-state PLS-CLR analysis. Also, GEMAX provides significantly better prediction accuracy and specificity compared to use of ground states descriptors or excited states descriptors alone.

We point out that it is not clear that the GEMAX models are better than corresponding GMAX and/or EMAX models if we limit such an evaluation to predictions of individual test compounds. But one must consider the composite picture of the overall results as measured by predicitivity, accuracy, specificity and sensitivity, across all the models developed in the study to make a meaningful comparison. Under these encompassing criteria, one concludes that the GEMAX models are somewhat more significant than the EMAX and GMAX models. Thus, the 4D-FP information resulting from the combination of the ground and first excited states of the LLNA molecules leads to better categorical QSAR models than those built using only ground state 4D-FP descriptors. Still, when building skin-sensitization QSAR models it remains unclear whether the first excited state is best, or whether one, or more, other non-ground state(s) should be considered in deriving the 4D-FP for the initial descriptor pool.

Acknowledgements This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 R21 GM075775-01. Information on Novel Preclinical Tools for Predictive ADME-Toxicology can be found at http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-023.html. Links to nine initiatives are found at http://nihroadmap.nih.gov/initiatives.asp. This work was also supported in part by The Procter & Gamble Company. Resources of the Laboratory of Molecular Modeling and Design at UIC and The Chem21 Group, Inc. were used in performing these studies.

References

1. Engelhard VH (1994) Sci Am 271:54
2. Song PS, Chin CA, Yamazaki I, Baba H (1975) Int J Quantum Chem 1
3. Moore TA, Mantulin WW, Song PS (1973) Photochem Photobiol 18:185
4. Mantulin WW, Song PS (1973) J Am Chem Soc 95:5122
5. Ou C-N, Tsai C-H, Song P-S (1977) In: Castellani A (ed) Research in photobiology. Plenum Press, New York, pp 257–265
6. Wondrak GT, Jacobson MK, Jacobson EL (2005) J Pharmacol Exp Ther 312:482
7. Li XY, Eriksson LA (2005) Photochem Photobiol 81:1153

8. Pan DH, Iyer M, Liu JZ, Li Y, Hopfinger AJ (2004) J Chem Inf Comput Sci 44:2083
9. Hopfinger AJ, Wang S, Tokarski JS, Jin BQ, Albuquerque M, Madhav PJ, Duraiswami C (1997) J Am Chem Soc 119:10509
10. Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ (2004) J Chem Inf Comput Sci 44:1526
11. Kulkarni A, Hopfinger AJ, Osborne R, Bruner LH, Thompson ED (2001) Toxicol Sci 59:335
12. Kodithala K, Hopfinger AJ, Thompson ED, Robinson MK (2002) Toxicol Sci 66:336
13. Li Y, Tseng YJ, Pan DH, Liu JZ, Kern PS, Gerberick GF, Hopfinger AJ (2007) Chem Res Toxicol 20:114
14. Kimber I, Basketter DA (1992) Food Chem Toxicol 30:165
15. Li Y, Pan D, Liu J, Kern PS, Gerberick GF, Hopfinger AJ, Tseng YJ (2007) Toxicol Sci 99:532
16. Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA (2005) Dermatitis 16:157
17. Fedorowicz A, Zheng LY, Singh H, Demchuk E (2004) Int J Mol Sci 5:56
18. Contact Sensitisation: Classification According to Potency., European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) (2003) Brussels, Belgium
19. HyperChem Program Release 7.05 for Windows (2005) In: Hypercube, Inc.
20. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:2338
21. Stewart JJP (1989) J Comput Chem 10:209
22. Pearlstein RA (1998) CHEMLAB-II users guide. CHEMLAB Inc., Lake Forest, IL
23. Duca JS, Hopfinger AJ (2001) J Chem Inf Comp Sci 41:1367
24. Liu JZ, Yang L, Li Y, Pan DH, Hopfinger AJ (2005) J Comput Aid Mol Des 19:567
25. Liu JZ, Pan DH, Tseng YF, Hopfinger AJ (2003) J Chem Inf Comput Sci 43:2170
26. Glen WG, Dunn WJ, Scott DR (1989) Tetrahedron Comput Methodol 2:349
27. Daganzo CF (1999) Logistics systems analysis. Springer
28. Hosmer DW, Lemeshow S (2000) Applied logistic regression. John Wiley & Sons, Inc., New York
29. Institute, S., SAS/STAT User's Guide, SAS Institute Inc., Cary, NC, 1999