Combinatorial QSAR of Ambergris Fragrance Compounds

Assia Kovatcheva,[†,#] Alexander Golbraikh,[‡] Scott Oloff,[‡] Yun-De Xiao,[§] Weifan Zheng,[||]
Peter Wolschann,[⊥] Gerhard Buchbauer,[†] and Alexander Tropsha*[,‡]

Department of Pharmaceutical Chemistry, University of Vienna, Althanstrasse 14,
A-1090 Vienna, Austria, Laboratory of Molecular Modeling, Division of Medicinal Chemistry and
Natural Products, School of Pharmacy, CB 7360, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599, Targacept, East First Street, Suite 300,
Winston-Salem, North Carolina, Eli Lilly and Company, 20 T. W. Alexander Drive,
P.O. Box 13951, Research Triangle Park, North Carolina 27709, and Department of Theoretical Chemistry
and Structural Biology, University of Vienna, Währinger Strasse 17, A-1090 Vienna, Austria

A combinatorial quantitative structure−activity relationships (Combi-QSAR) approach has been developed and applied to a data set of 98 ambergris fragrance compounds with complex stereochemistry. The Combi-QSAR approach explores all possible combinations of different independent descriptor collections and various individual correlation methods to obtain statistically significant models with high internal (for the training set) and external (for the test set) accuracy. Seven different descriptor collections were generated with commercially available MOE, CoMFA, CoMMA, Dragon, VolSurf, and MolconnZ programs; we also included chirality topological descriptors recently developed in our laboratory (Golbraikh, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147−158). CoMMA descriptors were used in combination with MOE descriptors. MolconnZ descriptors were used in combination with chirality descriptors. Each descriptor collection was combined individually with four correlation methods, including *k*-nearest neighbors (*k*NN) classification, Support Vector Machines (SVM), decision trees, and binary QSAR, giving rise to 28 different types of QSAR models. Multiple diverse and representative training and test sets were generated by the divisions of the original data set in two. Each model with high values of leave-one-out cross-validated correct classification rate for the training set was subjected to extensive internal and external validation to avoid overfitting and achieve reliable predictive power. Two validation techniques were employed, i.e., the randomization of the target property (in this case, odor intensity) also known as the Y-randomization test and the assessment of external prediction accuracy using test sets. We demonstrate that not every combination of the data modeling technique and the descriptor collection yields a validated and predictive QSAR model. *k*NN classification in combination with CoMFA descriptors was found to be the best QSAR approach overall since predictive models with correct classification rates for both training and test sets of 0.7 and higher were obtained for all divisions of the ambergris data set into the training and test sets. Many predictive QSAR models were also found using a combination of *k*NN classification method with other collections of descriptors. The combinatorial QSAR affords automation, computational efficiency, and higher probability of identifying significant QSAR models for experimental data sets than the traditional approaches that rely on a single QSAR method.

INTRODUCTION

The ambergris scent has been highly prized in the perfume industry due to its delicate note and good fixative properties. Originally, it has been extracted from ambergris, the natural product, which is released in the intestinal tract of the sperm whale (*Physeter macrocephalus L.*). According to the most popular theory,[1] ambergris is the pathological concretion of abscesses which arises from injures by incompletely digested food. It can be found in fragments of various weights[2] on the surface of the seawater after storms. Assuming that the real fragrance compounds are products of autooxidation of the principal ambergris' ingredient ambrein (Figure 1) which is odorless, Lederer[3] was the first to explain observations that the longer the fragments are floating in the water, the finer is the ambergris odor.

The commercial importance of the ambergris scent has stimulated the search for synthetic fragrance chemicals.[4] The synthesis of new odorants has been supported by several SAR studies. First SAR studies were carried out by Ohloff and co-workers and resulted in the "triaxial rule" of odor sensation.[5] According to this rule, ambergris fragrance compounds should have a *trans*-decalin skeleton with axial substituents in positions 1, 2, and 4 (Table 1, compounds **24** and **25** are a typical example). One of these substituents should be a functional group with an oxygen atom. However

* Corresponding author phone: (919)966-2955; fax: (919)966-0204; e-mail: alex_tropsha@unc.edu.
† Department of Pharmaceutical Chemistry, University of Vienna.
‡ University of North Carolina at Chapel Hill.
§ Targacept.
|| Eli Lilly and Company.
⊥ Department of Theoretical Chemistry and Structural Biology, University of Vienna.
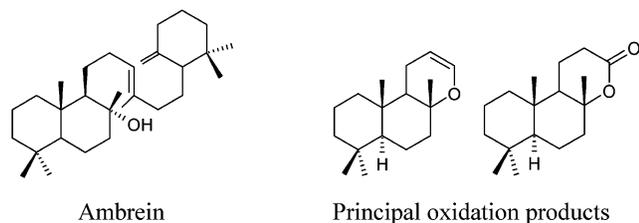# Visiting scientist at the Laboratory for Molecular Modeling, UNC-Chapel Hill.

AMBERGRIS FRAGRANCE COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004 **583**

Figure 1. Ambrein (odorless) and two of its principal oxidation products with ambergris-like odor.

the "triaxial rule" of odor sensation was postulated for the group of *trans*-decalin derivatives solely and did not explain the absence of the ambergris scent for other molecules which correspond to the structural requirements of the "triaxial rule" (Table 1, compounds **30**, **60**, **67**, etc.) or the presence of ambergris odor for molecules with a *cis*-decalin skeleton (Table 1, compounds **54**, **77**, **78**, etc.) or a skeleton of a different structural type (Table 1, compounds **1**, **81**, etc.).[6–8]

In the follow-up studies structural requirements were established for several chemical groups of ambergris odorants. Vlad et al.[9] found that in decalin systems the "ambergris triangle" formed by an oxygen atom and two hydrogen atoms should contribute to the lowest unoccupied molecular orbitals (LUMO) of the ambergris chemicals. They suggested that the LUMO takes part in the "orbital controlled electronic charge transfer" between the active molecules and the odorant receptor site.[9]

The theory about the presence of certain structural fragments which are responsible for the odor of ambergris was further advanced by Dimoglo[10] and Gorbachov and Rossiter,[8] who applied an electronic-topological approach.[11,12] Dimoglo defined two characteristic structural fragments. The first fragment included certain carbon atoms and an oxygen atom bound to a secondary or tertiary carbon. The second fragment consisted of two methyl groups with the same stereochemical orientation which are attached to a quaternary carbon atom. These two fragments could explain the disappearance of the ambergris odor when the five-member ring of compound **1** (Figure 2) is replaced by the six-member ring in compound **2** (Figure 2).

However, the same rule was insufficient to explain the presence of odor for compound **3** (Figure 2). Consequently, the already known correlation between the steric accessibility (SA) of the functional group (hydroxyl, ether, ester) and the ambergris odor[6,13] was appended by an additional term, SA of the oxygen atom.[10] This term helped to explain this particular case. For compounds with more complex structures the SA of the oxygen atom was appended by SA of a certain methyl group.[10]

As mentioned above, the presence or absence of the ambergris odor was correlated successfully only to the SA of the functional ether group in bicyclic ether derivatives of *Ambrox*.[6,13] Thus, the attempt to find additional structural fragments in combination with the SA terms was undertaken for compounds of particular structural types.[8,10] Due to the complexity and diversity of the ambergris odorants this methodology was insufficient to cover all compounds included in the data sets.

Recently Bajgrowicz et al.[14] synthesized six new camphor-derived stereoisomers which were found with the help of the olfactophore hypothesis[15] generated using CATALYST.[16] This hypothesis included one oriented hydrogen bond ac-

ceptor function (HBA)[17] and four hydrophobic functions. The hypothesis was able to explain the presence of the ambergris odor of these compounds very successfully. To improve the discrimination capacity of the hypothesis an additional search of excluded volumes was performed.

The olfactory response for different chemical classes was found to be sensitive to the compound stereochemistry. The response can change between stereomers from the presence to the absence of the ambergris odor, different level of odor intensity, or quality. Stereochemistry is a crucial factor which affects not only fragrances[18,19] but also many classes of biologically active compounds such as amino acids, carbohydrates, lipids,[20–23] many pharmaceuticals,[24–26] etc. Therefore, there is a challenge in the pharmaceutical, biochemical, and theoretical chemistry to develop predictive QSAR models for stereoisomers.[27–29]

Herein, we report on the development of robust QSAR models of 98 ambergris-type compounds of different structural types with known stereochemistry (Table 1). In the majority of previously reported studies the QSAR models are typically generated with a single modeling technique, frequently lacking external validation.[30] To achieve QSAR models of the highest quality, meaning both internal, and most importantly, external accuracy, we have developed and applied to the ambergris compounds a combinatorial QSAR approach, which explores all possible combinations of various collections of descriptors and optimization methods along with external model validation.

Different collections of descriptors were generated using MOE,[31] CoMFA,[32] CoMMA,[33] Dragon,[34] MolconnZ,[35] and Volsurf[36] programs as well as an in-house program that calculates chirality topological descriptors.[27] CoMMA descriptors were used in combination with MOE descriptors. MolconnZ descriptors were used in combination with chirality descriptors. Optimization methods included *k*-nearest neighbors (*k*NN) classification, decision tree,[31] binary classification,[31,37] and Support Vector Machines (SVM).[38] Every descriptor collection was explored individually in combination with every modeling technique resulting in 28 different types of QSAR models. Multiple predictive models with a correct classification rate for both the training set (CCR_{train}) and the test set (CCR_{test}) of at least 0.7 were found using the *k*NN classification method with different collections of descriptors. Several predictive models were found using other methods as well; however, not every combination of a descriptor set and the modeling technique afforded validated and predictive models.

DATA SET

98 compounds were selected from several publications (Table 1). The data set was compiled according to the following criteria: (i) diversity of the chemical structures and (ii) comprehensive data about the stereochemical configuration of each compound. The compounds were selected with respect to both their structural features and the qualitative description of the ambergris scent. Ambergris odor has been described as earthy, woody, camphor, fruity, rosy, marine, sandalwood, musky, cedarwood, ambergris with almond top notes, etc. Since the availability of quantitative data on the ambergris scent is very poor, only the presence or absence of the ambergris odor could be used to assign
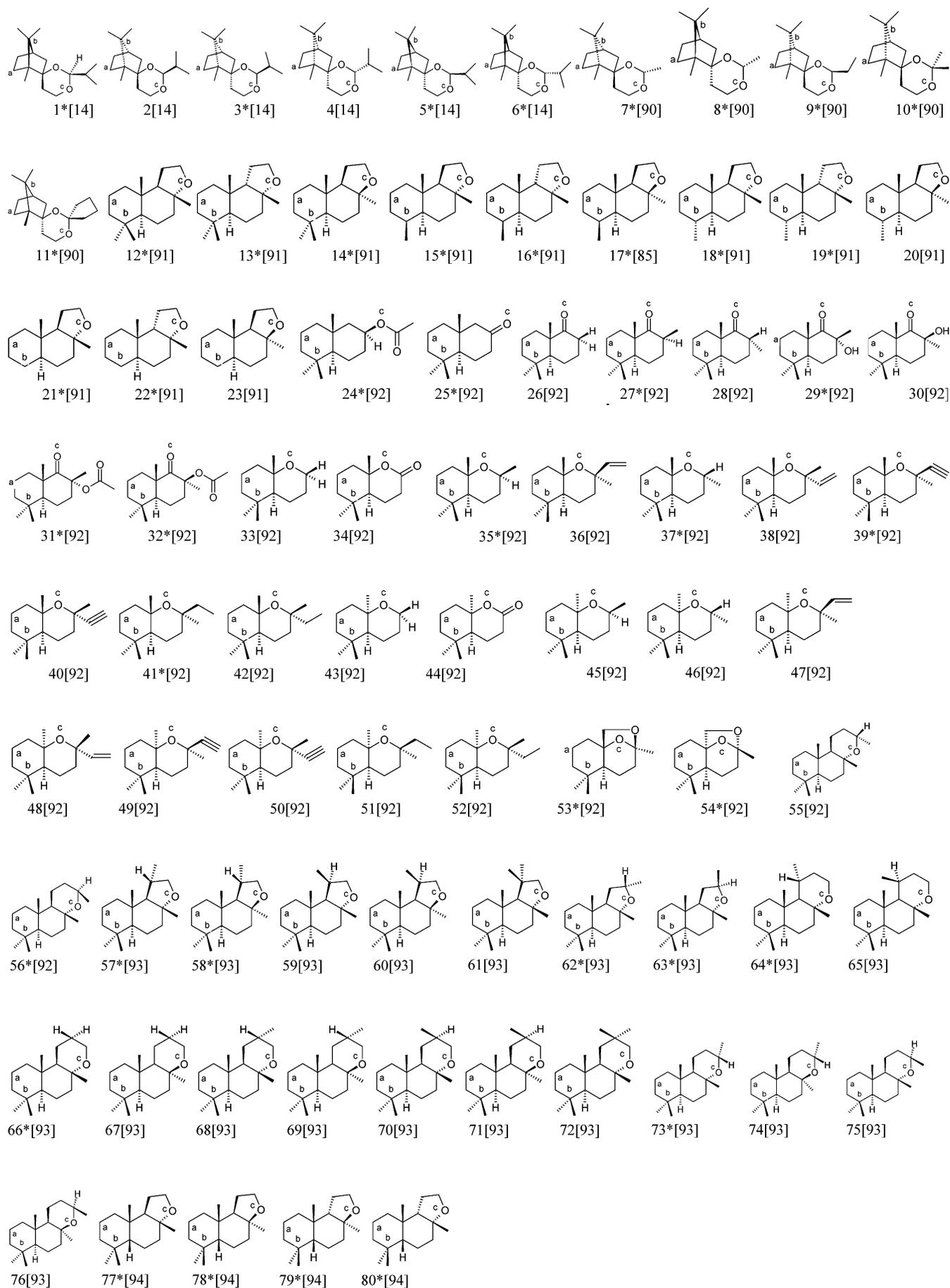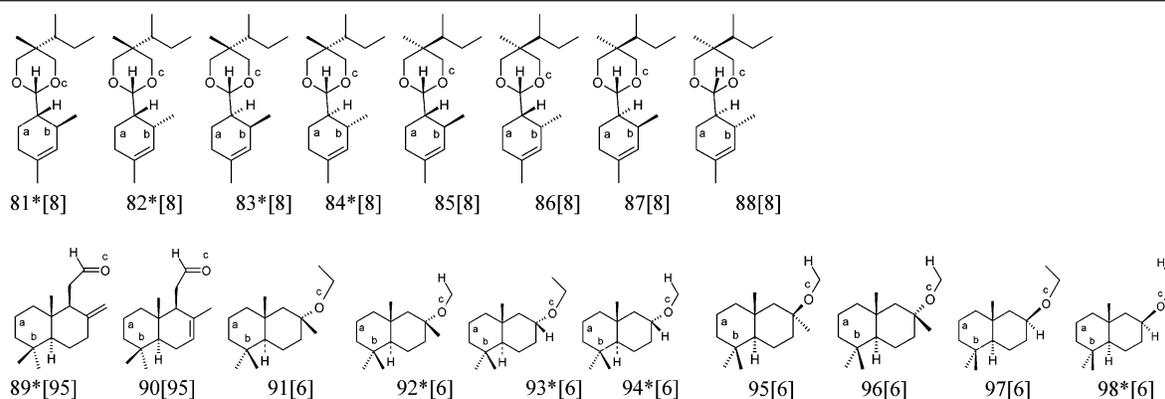
Table 1. Data Set of 98 Ambergris Fragrance Compounds[a]

AMBERGRIS FRAGRANCE COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004 **585**

Table 1. (Continued)

81*[8]   82*[8]   83*[8]   84*[8]   85[8]   86[8]   87[8]   88[8]

89*[95]   90[95]   91[6]   92*[6]   93*[6]   94*[6]   95[6]   96[6]   97[6]   98*[6]

[a] Compounds marked with an asterisk have ambergris odor. Atoms used for CoMFA alignment are marked by a letter.

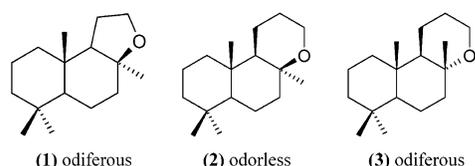**(1)** odiferous   **(2)** odorless   **(3)** odiferous

Figure 2. Structure and odor characteristics of several ambergris compounds.

the value of the odor intensity as the dependent variable for QSAR studies. Furthermore, only the ambergris feature was taken into account when assigning the odor intensity. Additional qualitative descriptions were neglected. All compounds which had at least some ambergris odor were given the value of odor intensity 1. Odorless compounds were given the value of 0.

Molecular Representation. Three-dimensional (3D) structures of molecules were built, and their geometries were optimized using Sybyl6.9.[39] Molecular mechanics calculations were performed using the Tripos force field with the Gasteiger–Hückel atomic charges. The optimized structures were then used for the descriptor generation.

DESCRIPTORS

CoMFA Descriptors. CoMFA descriptors were calculated after all molecules were aligned using carbon atoms *a* and *b* and oxygen atom *c* (Table 1). Since structures **1** and **15** (Table 1) show the most distinct ambergris odor, they were used as templates for the spatial alignment of all other molecules. First, molecule **15** was superimposed on molecule **1**. For the remaining molecules the distance constraints were imposed as follows. In molecules **2**–**12**, **26**–**32**, **81**–**88**, and **91**–**98** optimal distances between atoms *a*, *b*, and *c* were set to be equal to those in molecule **1**. In molecules **13**–**25**, **33**–**54**, **55**–**80**, **89**, and **90** optimal distances between atoms *a*, *b*, and *c* were set to be equal to those in molecule **15**. The force constant *k* was equal to 200. Molecules **2**–**14** and **16**–**98** were subjected to 10 fs molecular dynamics simulations with *T* = 300 K with subsequent minimization. Then molecules **2**–**12**, **26**–**32**, **81**–**88**, and **91**–**98** were superimposed on molecule **1**, and molecules **13**–**25**, **33**–**54**, **55**–**80**, **89**, and **90** were superimposed on molecule **15**.

A rectangular grid with step 2 Å was built around the aligned molecules; it was protruding for 4 Å outside of the region occupied by the molecules in each direction. Steric and electrostatic fields were calculated at each grid point using a carbon sp^3 probe atom with charge +1. The resulting field values were extracted from Sybyl6.9. Pairwise correlation analysis (see below) was performed on the field values, and 150 of them with pairwise correlation coefficients below 0.7 were selected and used as descriptors in combination with each of the modeling techniques.

Chirality Descriptors (CMTD). Chirality molecular topological descriptors (CMTD) defined in ref 27 included modified overall Zagreb indices,[27,40,41] molecular connectivity indices,[42-44] extended connectivity indices,[45] and overall connectivity indices.[46,47] All of the indices make use of the so-called chirality correction, which can be a real or imaginary number added to or subtracted from vertex degrees of a hydrogen-depleted molecular graph corresponding to atoms in R- and S-configurations, respectively. For example, the conventional index $^1\chi$ is defined as $^1\chi = \sum_{\text{All edges } ij} (a_i a_j)^{-0.5}$, where a_i and a_j are the vertex degrees of adjacent atoms i and j. The chirality index $^1\chi$ is defined as $^1\chi = \sum_{\text{All edges } ij} (a_i \pm c_i)^{-0.5}(a_j \pm c_j)^{-0.5}$, where c_i is the chirality correction for atom i. The plus sign is used, if the atom is in the R-configuration, and the minus sign is used, if the atom is in the S-configuration. For achiral atoms, the chirality correction is zero. Additional details can be found elsewhere.[27,28]

Chirality descriptors were used along with conventional chirality-insensitive overall Zagreb indices,[27,40,41] molecular connectivity indices,[42-44] extended connectivity indices,[45] and overall connectivity indices.[27,46] The chirality correction c was equal to 2. After applying complete correlation analysis (see below), the total number of descriptors was equal to 53. The descriptors were normalized by range-scaling, so that they had values within the interval [0,1].

MolconnZ Descriptors. MolconnZ[35] descriptors were used along with chirality descriptors (MolconnZ/CMTD descriptors). Recently this combination of descriptors was successfully used in QSAR studies of several data sets containing chiral compounds.[28] MolconnZ descriptors included valence, path, cluster, path/cluster, and chain molecular connectivity indices,[42-44] kappa molecular shape indices,[48,49] topological[50] and electrotopological[51-54] state indices, differential connectivity indices,[43,55] graph's radius and diameter,[44,56] Wiener[57] and Platt[58] indices, Shannon,[59] and

Bonchev-Trinajstić[60] information indices, counts of different vertices,[23] and counts of paths and edges between different types of vertices.[23] In this case, after applying complete correlation analysis (see below), the total number of the *MolconnZ/CMTD* descriptors was equal to 64. Descriptors were normalized by range-scaling, so that they had values within the interval [0,1].

VolSurf Descriptors. VolSurf descriptors are obtained from 3D interaction energy grid maps.[36] Calculation of VolSurf descriptors includes the following steps. (i) Building a grid around a molecule. (ii) Calculation of an interaction field (with a water probe, or hydrophobic probe, etc.) in each grid point. (iii) Eight or more energy values are assigned, and for each energy value, the number of grid points inside the surface corresponding to this energy (volume descriptors) or belonging to this surface (surface descriptors) is calculated.

The main advantage of VolSurf descriptors is that they are alignment-free.[36] VolSurf descriptors include size and shape descriptors, hydrophilic and hydrophobic regions descriptors, interaction energy moments, and other descriptors.[36] The total number of *VolSurf* descriptors was 96.

MOE Descriptors. MOE descriptors[31] include both 2D and 3D molecular descriptors. 2D descriptors include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity[42−44] and kappa shape indices,[48,49] adjacency and distance matrix descriptors,[56,57,61−63] pharmacophore feature descriptors, and partial charge descriptors.[31,64] 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors.[31,65] In total, 191 *MOE* descriptors were calculated.

Comparative Molecular Moment Analysis (CoMMA) Descriptors. Thirteen alignment-independent CoMMA descriptors[33] were used in this study including three principal moments of inertia I_x, I_y, and I_z, dipole and quadrupole moments p and Q, three dipolar components, p_x, p_y, and p_z, and three components of displacement between the center of mass and center of dipole d_x, d_y, and d_z as well as two quadrupole moments Q_{xx} and Q_{yy}. All descriptors, except for the last two are calculated with respect to the principal axes of inertia. The last two descriptors are calculated with respect to the frame with the origin in the center of the dipole and with the axes having the same directions as the inertia axes.[33] Descriptors were calculated online.[66] All CoMMA descriptors were used in combination with the 191 MOE descriptors (see above). This collection of descriptors will be referred to as *CoMMA/MOE descriptors*.

Dragon Descriptors. Dragon descriptors[34] include different groups:[67] constitutional descriptors, topological indices, molecular walk counts,[45,68] BCUT descriptors,[69] Galvez topological charge indices,[70] 2D autocorrelations, charge indices, aromaticity indices,[71] Randic molecular profiles,[72] geometrical descriptors, RDF descriptors,[73,74] 3D-MoRSE descriptors,[75] Weighted Holistic Invariant Molecular (WHIM) descriptors,[76,77] empirical descriptors, GETAWAY descriptors,[78] functional groups, atom-centered fragments, empirical descriptors, and properties. The total number of descriptors was 641. Identical descriptors were discarded. For the remaining descriptors pairwise correlation analysis (see below) was performed. Thus, the number of *DRAGON* descriptors used in our calculations was reduced to 148.

DESCRIPTOR REDUCTION

The following descriptor exclusion methods were used to reduce the collinearity and correlation between descriptors.

Pairwise Correlation Analysis. The procedure consists of elimination of one of the descriptors from each pair with the modulus of the correlation coefficient higher than a predefined value R_{max}. The procedure must be carried out with care. Indeed, let $R_{ij} = R(d_i, d_j)$ be the correlation coefficient between descriptors d_i and d_j. Then from $R_{ij} > R_{max}$ and $R_{jk} > R_{max}$ does not follow that $R_{ik} > R_{max}$. So in this case, if d_j is eliminated, d_k must be retained.

In this work, we have used the following algorithm of the pairwise correlation analysis. (i) Sort descriptors by variance and exclude all descriptors with the variance lower than the predefined value. Let D be the descriptor with the highest variance. (ii) Calculate correlation coefficients between D and all other descriptors. (iii) Exclude descriptors having the modulus of the correlation coefficient with D higher than R_{max}. (iv) Let D be the next descriptor with the highest variance. Go to step (ii). If there are no descriptors left, stop.

Complete Correlation Analysis. The complete correlation analysis is used to select a subset of linearly independent descriptors. Descriptors are considered as vectors in N-dimensional space, where N is the number of compounds. Here we used the procedure similar to that described in ref 79. (i) Select a pair of descriptors with the lowest absolute value of the correlation coefficient. (ii) Select the next descriptor which has the lowest maximum correlation coefficient with all linear combinations of descriptors selected R_{min}. It can be done by projecting the next descriptor onto a subspace defined by the descriptors selected. R_{min} is equal to the ratio between the length of this projection and the length of the descriptor vector. (iii) Repeat (ii) until the maximum correlation coefficient with all linear combinations of descriptors selected would reach a predefined value R_{max}. In this work, $R_{max} = 0.99$.

DIVISION OF A DATA SET INTO TRAINING AND TEST SETS

A set of procedures for the division of a data set into training and test sets has been developed recently.[80,81] These procedures are based on sphere-exclusion algorithms (Figure 3).

The procedure implemented in this study starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Let D_{min} and D_{max} be the minimum and maximum elements of **D**, respectively. N probe sphere radii are defined by the following formulas. $R_{min} = R_1 = D_{min}$, $R_{max} = R_N = D_{max}/4$, $R_i = R_1 + (i-1)*(R_N-R_1)/(N-1)$, where $i = 2,...,N-1$. Each probe sphere radius corresponds to one division into the training and the test set. A sphere-exclusion algorithm used in this study consisted of the following steps. (i) Select randomly a compound. (ii) Include it in the training set. (iii) Construct a probe sphere around this compound. (iv) Select compounds from this sphere and include them alternatively into the test and the training sets. (v) Exclude all compounds from within this sphere from further consideration. (vi) If no more compounds are left, stop. Otherwise let m be the number of probe spheres constructed and n be the number of remaining compounds. Let d_{ij} ($i=1,...,m$; $j=1,...,n$) be the distances between the

AMBERGRIS FRAGRANCE COMPOUNDS

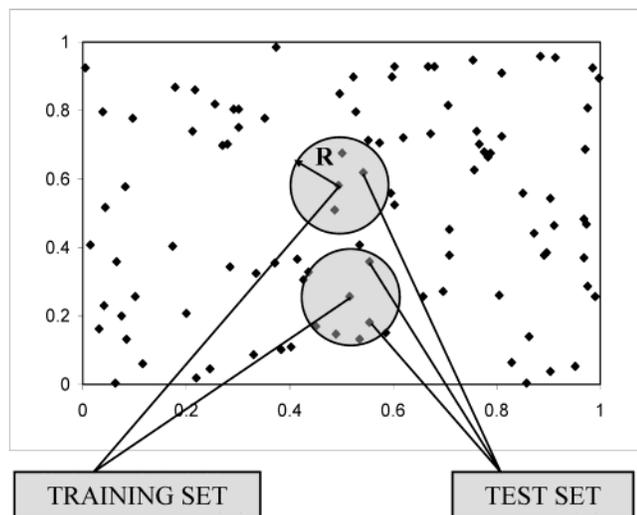*J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004* **587**

Figure 3. Sphere-exclusion algorithm.

remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest d_{ij} value and go to step (ii).

For each collection of descriptors, the data set was divided into 50 training and test sets of different relative sizes. Then three training and test sets were selected randomly for each collection of descriptors (Table 2); the number of compounds ranged between 58 and 91 for the training sets and between 40 and 7 for the corresponding test sets. For CoMMA/MOE descriptors, two divisions were selected because of the small number of CoMMA descriptors.

METHODS FOR QSAR ANALYSIS

The flowchart of the Combinatorial QSAR is given in Figure 4.

We describe briefly several data modeling algorithms that were employed in this work.

***k*NN-Classification Algorithm.** This approach has been implemented in our laboratory based on our earlier developments of the *k*NN QSAR methodology.[82,83] Let N be the number of compounds in a data set, and each compound belongs to one of several classes a, b, c, ... Classification *k*NN QSAR is a stochastic variable selection procedure based on the simulated annealing approach. The procedure is aimed at the development of a model with the highest fitness [correct classification rate (CCR) for the training set]. The parameters of the procedure are as follows: (1) the number of descriptors *nvar* to be selected from the entire set of descriptors; (2) the maximum number k of nearest neighbors; (3) the number of descriptors M that are changed at each step of the stochastic descriptor sampling procedure; (4) the starting T_{max} and ending T_{min} values of the simulation annealing "temperature", T, and the factor $d < 1$ to decrease T ($T_{next}=d \cdot T_{previous}$) at each step; (5) the number of times N the calculations must be performed before lowering T, if the CCR is not improved.

In all calculations reported in this work, $k = 5$, $T_{max} = 100$, $T_{min} = 10^{-9}$, $d = 0.95$, and $M = 3$. For all descriptor collections, D was varied from 10 to 50 with step 5. For each D, 10 models were built. Thus, the total number of models built for one division into the training and test set was 90.

The procedure starts with the calculation of the similarity between each pair of classes. Let n_a and n_b be the number of compounds in classes a and b, respectively, and m be the number of descriptors selected by the *k*NN classification procedure. The Tanimoto coefficient[84] was used as a similarity measure between two classes

$$T(a,b) = \frac{\sum_{i=1}^{m} \bar{D}_i^a \bar{D}_i^b}{\sum_{i=1}^{m}(\bar{D}_i^a)^2 + \sum_{i=1}^{m}(\bar{D}_i^b)^2 - \sum_{i=1}^{m}\bar{D}_i^a \bar{D}_i^b} \quad (1)$$

where $\bar{D}_i^a$ and $\bar{D}_i^b$ are average values of descriptor i for classes a and b, respectively

$$\bar{D}_i^a = \frac{\sum_{j=1}^{n_a} D_{ij}^a}{n_a} \quad \text{and} \quad \bar{D}_i^b = \frac{\sum_{j=1}^{n_b} D_{ij}^b}{n_b}$$

where D_{ij}^a is the descriptor value for compound j of class a. Evidently, $T(a,a) = 1$. Then weighted similarities $S_{i,C}$ between each compound i and each class C (a, or b, or c...) are calculated. Let k be the number of nearest neighbors of compound i.

$$S_{i,C} = \sum_{p=1}^{k}\left[\frac{\exp(-\alpha d_{ip}/\sum_{p'=1}^{k} d_{ip'})}{\sum_{q=1}^{k} \exp(-\alpha d_{iq}/\sum_{p'=1}^{k} d_{ip'})}T(a_p,C)\right] \quad (2)$$

Then where a_p in $T(a_p,C)$ is the class of compound p, and α is a parameter, which in this study was set to 2, and d_{ip} is the distance between compound i and its pth nearest neighbor.

In the leave-one-out cross-validation procedure, every compound i of the training set is classified according to the classes of its nearest neighbors as follows. First, the similarity $S'_{i,C}$ between compound i and each class C is calculated as follows.

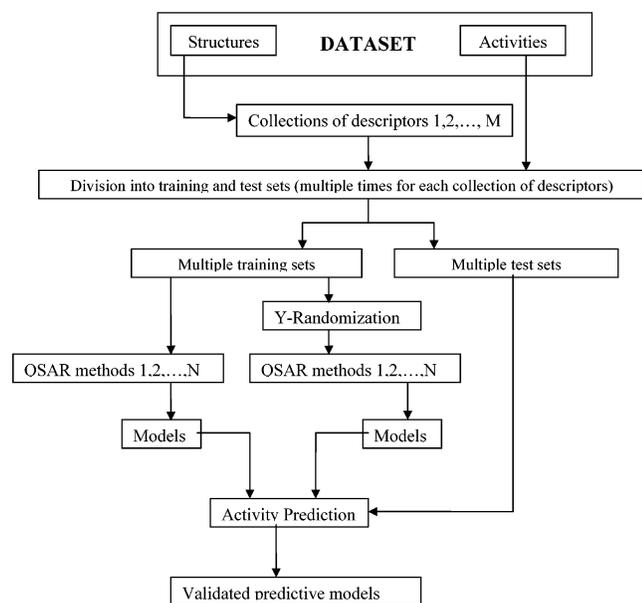$$S'_{i,C} = \sum_{j=1}^{k}\left[\frac{\exp(-d_{ij})}{\sum_{j'=1}^{k} \exp(-d_{ij'})}S_{i,C}\right] \quad (3)$$

Then compound i is assigned the class which corresponds to the highest value of $S'_{i,C}$. The CCR is equal to N_{corr}/N, where N and N_{corr} are the total number of compounds and the number of compounds classified correctly, respectively.

Classification *k*NN algorithm works as follows.

1. Set $T = T_{max}$.

2. Select randomly a subset of D descriptors.

3. For each compound, predict its activity using expression 1.

4. Select the number of nearest neighbors, which gives the highest CCR.

Table 2. Size of the Training and Test Sets Generated with the Sphere Exclusion Algorithm

	division 1			division 2			division 3		
collection of descriptors	no. of compds in the train./ test set	no. of active/ inactive compds (train. set)	no. of active/ inactive compds (test set)	no. of compds in the train./ test set	no. of active/ inactive compds (train. set)	no. of active/ inactive compds (test set)	no. of compds in the train./ test set	no. of active/ inactive compds (train. set)	no. of active/ inactive compds (test set)
Dragon	70/28	36/34	16/12	76/22	40/36	12/10	73/25	39/34	13/12
CMTD	67/31	37/30	15/16	72/26	39/33	13/13	58/40	30/28	22/18
MolconnZ/CMTD	73/25	41/32	11/14	68/30	32/36	20/10	78/20	41/37	11/9
MOE	61/37	38/23	14/23	91/7	49/42	3/4	70/28	36/34	16/12
CoMFA	82/16	43/39	9/7	77/21	41/36	11/10	71/27	38/33	14/13
VolSurf	80/18	44/36	8/10	73/25	40/33	12/13	68/30	37/31	15/15
CoMMA/MOE	66/32	36/30	16/16	73/25	39/34	13/12	-	-	-

Figure 4. The flowchart of the combinatorial QSAR methodology including validation.

5. Change number $M < D$ of descriptors to the same number of descriptors selected randomly out of all descriptors.

6. Repeat steps 3 and 4 with the modified descriptor subset obtained in step 5.

7. If the new CCR (CCR$_{new}$) is higher than the previous one (CCR$_{old}$), accept the new set of descriptors and go to step 5. Otherwise, accept it with the probability $p = \exp[-(CCR_{old} - CCR_{new})/T]$ and go to step 5, or reject it with the probability $(1-p)$, and go to step 8.

8. If CCR does not change after step 5 is performed N times for the current T, and if $T > T_{min}$, decrease T and go to step 5, and if $T \leq T_{min}$ stop. If step 5 has been performed less than N times for the current CCR, go to step 5.

Thus, the output from the procedure is a QSAR model, which is characterized by the set of D descriptors selected, the number k of nearest neighbors, and the value of CCR for the training set (CCR$_{train}$).

Z-cutoff value characterizes the maximum distance between a compound for which the prediction is made and its closest nearest neighbor of the training set in the descriptor space. The square of this distance can be represented as a sum of the average distance square between nearest neighbors within the training set and a number of Z of this distance variance: $D^2_{max} = \langle D^2_{near\cdot neighb} \rangle + Z \cdot Var_{near\cdot neighb}$.

Classification accuracy of the model is estimated using the test set.

1. For each compound of the test set, k-nearest neighbors from the training set are found.

2. All compounds of the test set, the distances of which to their closest nearest neighbor are within the defined Z-cutoff, are selected.

3. Similarity of each compound chosen in step 2 to each class is calculated using expression (3). The compound is assigned a class, to which it has higher similarity.

4. Classification accuracy of the model is characterized by the CCR for the test set (CCR$_{test}$).

In this study, Z was equal to 2 by default. The maximum Z-cutoff value, for which a reliable prediction of new compounds can be obtained, is a characteristic of the applicability domain[30] of a QSAR model. As we shall see, using $Z = 2$, high CCR$_{test}$ values have been obtained for several descriptor collections. We also searched for the lowest possible Z value by decreasing Z below 2 with step 0.1 until at least one test set compound was found outside the corresponding applicability domain.

Binary Tree Classification. We have used a binary tree classification algorithm as implemented in the Molecular Operating Environment (MOE) package.[31] The method consists of two parts: tree growing and tree pruning. Tree growing is carried out by splitting the nodes according to the rules in the form $x \leq c$ (if descriptor x is a continuous variable) or $x = c$ (if it is a categorical data), where c is the best value for splitting the node. Splitting is based on the Gini index of diversity[85]

$$G(t) = 1 - \sum_{i=1}^{K} P_i^2(t) \qquad (4)$$

where $P_i^2(t)$ are the fractions of compounds of each class i ($i=1,...,K$) in node t. $G(t)$ is used as the node t impurity measure. The goodness of a split is measured by the change C of the impurity of the node by splitting it

$$C = G(t) - P_L G(t_L) - P_R G(t_R) \qquad (5)$$

where P_L and P_R are the proportions of cases going to the left t_L and right t_R child nodes.[31] In each step of the tree growing, the node is split which gives the greatest decrease of impurity. A node cannot be split, if all compounds in it belong to the same class or if the number of compounds in it is lower than a predefined limit.[31]

Each leaf in the tree is assigned to a class maximally represented in this leaf. The misclassification rate in node t is calculated as $r(t) = 1 - n_j/n_t$, where n_j is the number of

compounds of class j, and n_t is the total number of compounds in the node. The total misclassification rate $R(T)$ = $N_{misclass}/N_{tot}$, where $N_{misclass}$ and N_{tot} are the total number of misclassified compounds and the total number of compounds in the data set, respectively. If classes have different sizes, the misclassification rate is multiplied by weights defined as $w_j = N_{tot}/N_j$, where N_j is the number of compounds in class j.[31]

An initially grown tree is very large and has a very high correct classification rate for the training set. However, usually it performs poorly for the test set.[31] Tree pruning is a procedure used to decrease the size of the tree and increase its classification accuracy for the test set. This procedure was performed using the test set. By pruning branches of the tree, the accuracy of classification for the training set and the size of the tree are decreased. A modified tree misclassification rate $R_a(T) = R(T) + aL(T)$ is defined, where $L(T)$ is the number of leaves in the tree, and $a > 0$ is a parameter. According to this equation, the size of the tree and the misclassification rate are balanced.[31] By increasing a, smaller trees can be found for which $R'_a(T) = R_a(T)$. Pruning is performed by finding a sequence of successively smaller trees T_i, starting from the initially grown tree. The smallest tree T_N is just the root node. The misclassification rate for each T_i is calculated using this test set. The output of the procedure is the tree with $R(T)$ within a specified number of standard errors of the minimum of all subtree $R(T)$ values. The standard error is defined as $\sigma = \sqrt{p(1-p)N_{test}}$, where p is the proportion of correctly classified cases. This subtree is referred to as the best subtree.[31]

The class of a new compound is predicted by assigning it to the class of a leaf this compound belongs to. Classes assigned to compounds of the external test set (obtained with the sphere-exclusion algorithm) were used for the estimation of the model classification accuracy. The following parameters have been used: minimum node split size 10, ordered threshold 6, and best tree threshold 0.5.

Binary QSAR. Binary QSAR is a new technique developed by P. Labute[37] and implemented in the MOE package.[31] This approach can be applied, if the activities y_i of compounds take only two values, zero and one, which correspond to inactive and active compounds, respectively. Binary QSAR is based on the Bayesian inference technique, which is used to classify a compound as an active or inactive one. Let m be the total number of compounds, and m_0 and m_1 are the number of inactive and active compounds ($m=m_0+m_1$). Then if descriptors $X_1, X_2,..., X_n$ are not correlated, then the conditional probability that a compound with descriptor values $X_1 = x_1, X_2 = x_2, ... X_n = x_n$ is active, i.e., $p(x)$ = $Pr(Y=1|X_1=x_1,X_2=x_2...X_n=x_n)$ can be estimated as[37]

$$p(x) = \left[1 + \frac{m_0 + 1}{m_1 + 1}\prod_{j=1}^{n}\frac{f_j(x_j,0)}{f_j(x_j,1)}\right]^{-1} \quad (6)$$

where $f(x,y) = Pr(X_j=x_j|Y=y)$. Without loss of generality, it is assumed that descriptors $X_1, X_2,..., X_n$ have the mean value of zero and variance one.

Each function $f(x)$ can be estimated by considering a histogram of observed descriptor values on a set of B bins $(b_0b_1],..., (b_{B-1}b_B)$, where $b_0 = -\infty$ and $b_B = +\infty$. The number of compounds within bin k

$$B_k = \sum_{i=1}^{m}\int_{b_{k-1}}^{b_k}\delta(x - x_i)dx$$

can be smoothed by approximating each δ-function with a Gaussian with variance σ^2:[37]

$$B_k = E_k - E_{k-1}, E_k = \frac{1}{2}\sum_{i=1}^{m}erf\left(\frac{b_k - x}{\sigma\sqrt{2}}\right)$$

σ is referred to as the smoothing parameter. Finally, $f(x)$ can be estimated as[37]

$$\hat{f}(x) = \sum_{k=1}^{B}\frac{B_k + 1/c}{c + B/c}[E_k - E_{k-1}] \quad (7)$$

where

$$c = \sum_{k=1}^{B}B_k$$

In the same way, all $f_j(x_j,0)$ and $f_j(x_j,1)$ can be estimated and

$$p(x) = \left[1 + \frac{m_0 + 1}{m_1 + 1}\prod_{j=1}^{n}\frac{\hat{f}_j(x_j,0)}{\hat{f}_j(x_j,1)}\right]^{-1} \quad (8)$$

Thus, the whole binary QSAR procedure consists of the following steps.[37] (i) The principal component analysis of the descriptor matrix to produce a variance-covariance matrix of $x_i = Q(d_i-u)$ equal to the identity matrix, where x_i are principal components and $d_i = (d_{i1},..., d_{in})$ are descriptor values for compound i. (ii) Estimate the binary QSAR model $p(x)$ parameters. The probability that a compound with descriptors d_i^{new} is active can be estimated as $p(Q(d_i^{new}-u))$.

Support Vector Machines (SVM). The SVM method was developed by V. Vapnik.[38] The application of SVM to the binary classification problem was implemented in our group as follows. Let m be a number of representative points of compounds scattered in an n-dimensional descriptor space. Compounds can be active (activity is equal to 1) or inactive (activity is equal to -1). The problem is to divide active and inactive compounds by a hyperplane in the descriptor space. If the solution of this problem is possible, the data set is referred to as separable. Otherwise it is nonseparable. If a data set is separable, the solution can be found as follows. Equation of any hyperplane in the descriptor space can be represented as $(\mathbf{wx}) - b = 0$, where $\mathbf{w}$ is normal to the hyperplane, $\mathbf{x}$ is a vector with the beginning in the origin and end on the hyperplane, and $(\mathbf{wx})$ is the dot product of w and x. Let it be the dividing hyperplane. Without loss of generality, we can assume that for any point x_i with activity $y_i = 1$ $(\mathbf{wx}_i) + b \geq +1$, and for all points x_i with $y_i = -1$ $(\mathbf{wx}_i) + b \leq -1$. These two inequalities can be combined in one:

$$y_i[(\mathbf{wx}_i) + b] - 1 \geq 0 \quad (9)$$

The distance between the hyperplane and the closest to it data set points is equal to $1/||\mathbf{w}||$, where $||\mathbf{w}||$ is the norm of $\mathbf{w}$. Thus, by minimizing $||\mathbf{w}||$ or $||\mathbf{w}||^2$ with constraints (9), the optimal dividing hyperplane can be found. This optimi-

zation problem can be solved by minimizing the Lagrangian

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m}\alpha_i[y_i((\mathbf{wx}_i) + b) - 1] \quad (10)$$

where $\alpha_i \geq 0 \; \forall \; i$ are the Lagrange multipliers. Methods of solving this problem and finding $\mathbf{w}$ and b are described in ref 86. For all $y_i[(\mathbf{wx}_i)+b]-1 > 0$, $\alpha_i = 0$, and for all $y_i-[(\mathbf{wx}_i)+b] - 1 = 0$, $\alpha_i > 0$. Points for which $\alpha_i > 0$ are called support vectors. These points belong to hyperplanes $y_i[(\mathbf{wx}_i)+b] - 1 = 0$. In fact, only these points are necessary to build the optimal dividing hyperplane. Assigning compounds to a class of actives or inactives can be carried out by finding y_i from inequality (9).

In practice, if the number of points is lower than the number of descriptors minus one and no $K+2$ points belong to a K-dimensional hyperplane, the data set is always separable. So if the number of descriptors is higher than the number of compounds minus one, there is a high risk of overfitting. The hyperplane will perfectly separate points of the training set while there will be poor separation of the test set. In this case the same approach is applied as in the case when the solution does not exist, namely, such a hyperplane is sought, which divides active and inactive compounds with the classification error minimized. Constraints (9) are replaced by the inequalities

$$y_i[(\mathbf{wx}_i) + b] \geq 1 - \xi_i, \; \xi_i \geq 0 \quad (11)$$

where ξ_i ($i=1,...,m$) are slack variables. The optimal hyperplane can be found by minimizing the Lagrangian

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m}\alpha_i[y_i((\mathbf{wx}_i) + b) - 1 + \xi_i] +$$

$$Cf(\sum_{i=1}^{m}\xi_i) - \sum_{i=1}^{m}\mu_i\xi_i \quad (12)$$

where $0 \leq \alpha_i \leq C \; \forall \; i$ and μ_i are the Lagrange multipliers for the constraint $\xi_i \geq 0$. A penalty function $f(\sum_{i=1}^{m}\xi_i)$ is a positive monotonically increasing function of each parameter ξ_i. We have used the penalty function in the following form

$$f = \begin{cases} 0, \text{if } \sum_{i=1}^{m}\xi_i \leq \epsilon \\ \sum_{i=1}^{m}\xi_i - \epsilon, \text{if } \sum_{i=1}^{m}\xi_i > \epsilon \end{cases} \quad (13)$$

where ϵ is a parameter. Support vectors are defined by the condition $y_i[(\mathbf{wx}_i)+b] - 1 + \xi_i = 0$.

The hyperplane parameters depend on C and ϵ. These parameters are also varied to reduce the overfitting. Thus the SVM procedure was run multiple times to find the optimum values of C and ϵ.

CONFUSION MATRICES AND CLASSIFICATION ACCURACY OF QSAR MODELS

Confusion matrices are used to estimate the classification accuracy of a QSAR model. In the case when compounds belong to two classes (active and inactive compounds), a 2 × 2 confusion matrix can be defined as in Table 3a, where

Table 3. a. 2 × 2 Confusion Matrix and b. Normalized 2 × 2 Confusion Matrix

Section a			
	observed		
predicted	active	inactive	total
active	TP	FP	TP + FP
inactive	FN	TN	FN + TN
total	$N_{act} = TP + FN$	$N_{inact} = FP + TN$	$N = N_{act} + N_{inact}$

Section b			
	observed		
predicted	active	inactive	total
active	TP/N_{act}	FP/N_{inact}	$TP/N_{act} + FP/N_{inact}$
inactive	FN/N_{act}	TN/N_{inact}	$FN/N_{act} + TN/N_{inact}$
total	$N_{act}/N_{act} = 1$	$N_{inact}/N_{inact} = 1$	2

N_{act} and N_{inact} are the number of active and inactive compounds in the data set, TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives. The following classification accuracy characteristics associated with confusion matrices are widely used in QSAR studies: sensitivity ($S=TP/N_{act}$), specificity ($SP=TN/N_{inact}$), and enrichment $E = TP*N/[(TP+FP)*N_{act}]$. For all QSAR models developed in this project, CCR was defined as $N_{corr}/N = (TP+TN)/N$, where N and N_{corr} were the total number of compounds and the number of correctly classified compounds.

In this paper, we have employed normalized confusion matrices (Table 3b). A normalized confusion matrix can be obtained from the nonnormalized one by dividing the first column by N_{act} and the second column by N_{inact}. Normalized enrichment is defined in the same way as E but is calculated using a normalized confusion matrix: $E_n = 2TP*N_{inact}/[TP*N_{inact}+FP*N_{act}]$. E_n takes values within the interval [0, 2].

Y-RANDOMIZATION

Y-randomization (randomization of response, i.e., in our case, activities) is a widely used approach to establish the model robustness.[87] It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower values of CCR for the test set than the models built using the training set with real activities. If this condition is not satisfied, real models built for this training set are not reliable and should be discarded.

The Y-randomization test was performed for training sets which afforded the models with the highest CCR values. The calculations were performed five times for each collection of descriptors and each optimization method, with the input parameters identical to those used for building models with real activities. Models built with randomized activities were used to predict activities of the corresponding test set. CCR_{train} and CCR_{test} values for models built with real and randomized activities were compared with each other. Using kNN classification QSAR, multiple models were built. To estimate the robustness of the classification QSAR models we used the following criterion. Let N_{real} and N_{rand} be the total number of models built with real and randomized activities, respectively, and n_{real} and n_{rand} be the corresponding number of models with CCR ≥ 0.7 (which we considered

AMBERGRIS FRAGRANCE COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004 **591**

Table 4. Combinatorial QSAR[a]

	method							
	*k*NN		decision tree		binary QSAR		SVM	
descriptors[b]	CCR_{train}[c]	CCR_{test}	CCR_{train}	CCR_{test}	CCR_{train}	CCR_{test}	CCR_{train}	CCR_{test}
Dragon (148)	**0.70 (45)**	**0.86**	**0.70**	**0.78**	**0.72**	**0.76**	0.83	0.68
CMTD (53)	0.72 (25)	0.65	0.67	0.74	0.76	0.50	0.81	0.58
MolconnZ/CMTD (64)	0.67 (50)	0.60	0.62	0.53	0.85	0.47	0.87	0.53
MOE (191)	0.79 (35)	0.65	**0.74**	**0.71**	**0.74**	**0.86**	0.77	0.65
CoMFA (150)	**0.76 (15)**	**0.89**	0.75	0.62	0.71	0.65	**0.83**	**0.75**
Volsurf (96)	**0.77 (25)**	**0.85**	0.77	0.60	**0.74**	**0.70**	0.94	0.53
COMMA/MOE (204)	**0.77 (15)**	**0.75**	**0.74**	**0.72**	**0.73**	**0.70**	0.73	0.69

[a] Correct classification rate (CCR) for models with the highest prediction accuracy for each combination of the method and collection of descriptors. CCR values for the best models (when both training and test sets have these values greater than 0.7) are shown in bold. Although some models have higher CCR_{test} values for the test set (see Table 6: MolconnZ/CMTD, MOE), they did not pass the Y-randomization test and were not accepted. [b] The total number of descriptors is given in parentheses. [c] The number of descriptors selected by the variable selection *k*NN classification procedure is given in parentheses.

acceptable). The fractions of models with CCR $\geq$ 0.7 are $F_{real} = n_{real}/N_{real}$ and $F_{rand} = n_{rand}/N_{rand}$, respectively. The robustness of predictive models (i.e., with CCR$\geq$0.7) built with real activities of the training set was defined as $R = 1 - F_{rand}/F_{real}$. R takes values from minus infinity to 1. If $R \geq 0.9$, predictive models are considered reliable. The R values were calculated for the fitness function, i.e., CCR_{train}, obtained in the LOO cross-validation procedure incorporated in the *k*NN classification QSAR, and for prediction of activities of the test set (CCR_{test}).

RESULTS AND DISCUSSION

The best models for all possible combinations of descriptor collections and QSAR analysis methods are shown in Table 4. We discuss below the results for every QSAR method in detail.

***k*NN Modeling.** For each descriptor collection and each division into the training and the test set, 90 variable selection QSAR models were built. The number of descriptors selected by *k*NN classification procedure varied from 10 to 50 with step 5. All models with $CCR_{train} \geq 0.70$ were validated using Y-randomization and external prediction for the corresponding test sets.

The Y-randomization tests were carried out for all models. For each collection of descriptors, activities of training sets corresponding to models with the highest CCR_{test} values (division 1, Table 2) were randomized five times, and all calculations were repeated for each randomization. Thus, for each training set the number of all models built with randomized activities was 450. Highest CCR_{train} values for models built with real and randomized activities appeared to have similar values (Figure 5a).

This phenomenon can be explained as follows. Regardless of the number of active (1) and inactive (0) compounds in the data set, if we randomly assign 1 and 0 to each compound with a probability of 0.5, the expected CCR value will be equal to 0.5. If the model is trained to "correctly" predict even randomized activities, the CCR_{train} for the training set can significantly exceed 0.5. This is exactly what we observe. Thus a CCR_{train} is not a good characteristic of a binary classification QSAR model. We will demonstrate this conclusion on the models built with other methods considered in this paper. We have also calculated the number and the fraction of models with $CCR_{train} \geq 0.7$ for the training sets.
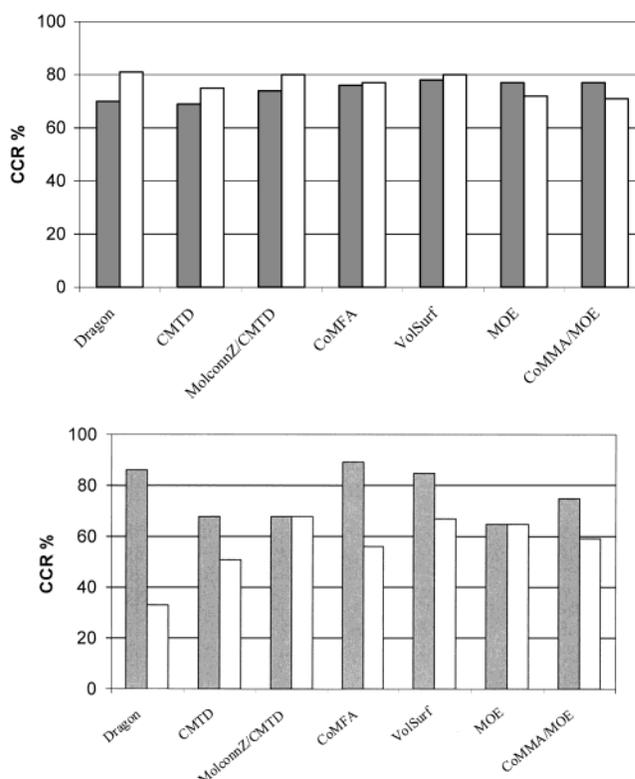

Figure 5. a. *k*NN classification modeling of the training sets with different descriptors. Highest CCR_{train} values for models built with real (gray) and randomized (white) activities of the training sets (division 1) for all collections of descriptors are shown. b. *k*NN classification modeling of the test sets with different descriptors. Highest CCR_{test} values for models built with real (gray) and randomized (white) activities of the training sets (division 1) for all collections of descriptors are shown.

(Results for division 1 are presented in Table 5a.) These results give a deeper insight into this problem. The robustness of a model with a high CCR value can be estimated by the robustness parameter R introduced above. The robustness of all models with high CCR_{train} values was very low (Table 5a). In some cases, the highest CCR_{test} values obtained with models built with real and randomized activities of the training sets also have similar values (Figure 5b) demonstrating the low robustness of such models (Table 5b).

However, the robustness of those models with high CCR_{test} values was also high (Table 5b). All models with $CCR_{test} \geq 0.7$ appeared to have $CCR_{train} \geq 0.7$, but the opposite was

592 *J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004*
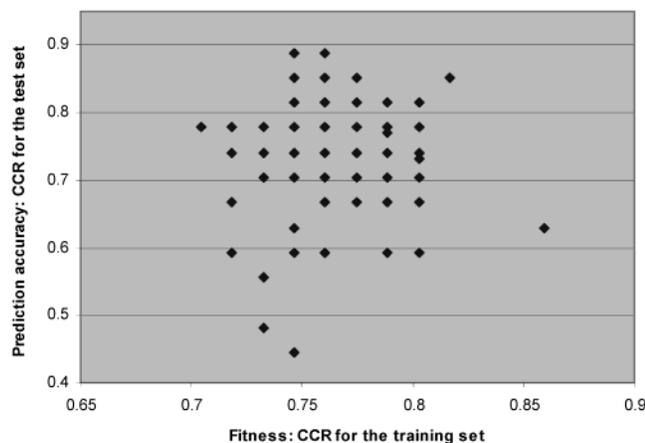
KOVATCHEVA ET AL.

Figure 6. CCR_{test} versus CCR_{train} of 90 kNN classification models built with CoMFA descriptors (division 1).

Table 5. a. Robustness R of Models with $CCR_{train} \geq 0.7^a$ and b. Robustness R of Models with $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7^b$

	n_{real}	$F_{real} =$ n_{real}/N_{real}	n_{rand}	$F_{rand} =$ n_{rand}/N_{rand}	$R =$ $1 - F_{rand}/F_{real}$
			a.		
CMTD	18	0.20	337	0.75	−2.75
CoMFA	90	1.00	437	0.97	0.03
CoMMA/MOE	84	0.93	249	0.55	0.41
Dragon	78	0.87	298	0.66	0.24
MOE	90	1.00	339	0.75	0.25
MolconnZ/CMTD	47	0.52	295	0.66	−0.27
VolSurf	90	1.00	338	0.75	0.25
			b.		
CMTD	0	0	0	0	
CoMFA	50	0.56	13	0.029	0.95
CoMMA/MOE	9	0.10	23	0.051	0.49
Dragon	33	0.37	0	0	1.00
MOE					
MolconnZ/CMTD	0	0	12	0.027	
VolSurf	30	0.33	15	0.033	0.9

[a] N_{real} and N_{rand} are the number of models built with real and randomized activities of the training set. n_{real} and n_{rand} are the number of corresponding models with $CCR_{train} \geq 0.7$. [b] N_{real} and N_{rand} are the number of models built with real and randomized activities of the training set, respectively. n_{real} and n_{rand} are the number of models with $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7$, respectively.

not true. Thus, the models with high values of both CCR_{train} and CCR_{test} and a high robustness of the test set predictions were considered acceptable.

We shall emphasize that there exists no correlation between CCR_{train} and CCR_{test} values (Figure 6). Thus, the validation of a classification QSAR model must include comparison of predicted and observed activities for the external test set (i.e., compounds which were not used in model building). This conclusion was made previously by several authors who reported QSAR models with the continuous dependent variable.[88,89] Figure 6 is the classification QSAR analogue of the so-called Kubinyi paradox.[88,89] Despite a high value of CCR_{train}, many models in Figure 6 have a low prediction accuracy ($CCR_{test} < 0.7$).

Similar results have been obtained for divisions 2 and 3 (Table 2). For division 2, 21 models built with CoMFA descriptors, 5 models built with CoMMA/MOE descriptors, 75 models built with MOE descriptors, and 20 models built with Dragon descriptors had both CCR_{train} and CCR_{test} equal or exceeding 0.7. For division 3, 66 models built with

CoMFA descriptors, 48 models built with Dragon descriptors, and 3 models built with MOE descriptors had both CCR_{train} and CCR_{test} equal or exceeding 0.7. For all these sets, except for division 2 of the data set using Dragon descriptors, the models had robustness $R > 0.9$. Statistical characteristics of predictive models with the highest CCR_{test} obtained with the kNN classification method are given in Table 6. For all descriptor collections, except for VolSurf descriptors, all compounds of the test set were within the cutoff distance with $Z = 2$ (see kNN classification section above). For VolSurf descriptors, five compounds of the test set of division 1 and four compounds of the test sets of divisions 2 and 3 were outside of the cutoff distance with $Z = 2$ threshold from compounds of the training set and were not classified. Lowest Z values characterizing the applicability domain are given in Table 6.

DECISION TREE

Decision trees were built using the MOE package.[31] Complete descriptor collections were used in all calculations because the package does not provide the option of automatic variable selection. The following predictive models were obtained using Dragon descriptors for division 1 (Table 2): $CCR_{train} = 0.74$, $CCR_{test} = 0.75$ (S=0.56, SP=1.0, E=1.75, and E_n=2.0); Dragon descriptors for division 2 $CCR_{train} = 0.70$, $CCR_{test} = 0.78$ (S=0.67, SP=0.90, E=1.63 and E_n=1.74), MOE descriptors (division 2) $CCR_{train} = 0.74$, $CCR_{test} = 0.71$ (S=0.67, SP=0.75, E=1.55 and E_n=1.45), and CoMMA/MOE descriptors (division 2, Table 2) CCR_{train} = 0.74, $CCR_{test} = 0.72$ (SE=0.62, SP=0.83, E=1.54, and E_n=1.57). The Y-randomization test performed five times for each of these divisions into training and test sets gave the following results. Only for two models built with Dragon descriptors (division 2) and MOE descriptors (division 2) both CCR_{train} and CCR_{test} values were higher than 0.7. Only one model built with Dragon descriptors (division 1) had both CCR_{train} and CCR_{test} higher than 0.7. For that reason, these models cannot be regarded as accurate.

BINARY QSAR

Binary QSAR calculations were carried out using MOE.[31] Calculations were performed using complete descriptor collections because the package does not provide the option of automated selection of variables. The maximum number of principal components was 10, the smoothing parameter was equal to 0.25, and the binary threshold was set to 0.5. The following predictive models were obtained using MOE descriptors (division 2, Table 2) $CCR_{train} = 0.74$, $CCR_{test} = 0.86$ (S=0.67, SP=1, E=2.33, and E_n=2), Dragon descriptors (division 3, Table 2) $CCR_{train} = 0.72$, $CCR_{test} = 0.76$ (S=0.69, SP=0.82, E=1.57, and E_n=1.61), and VolSurf descriptors (division 3, Table 2) $CCR_{train} = 0.74$, $CCR_{test} = 0.70$ (S=0.60, SP=0.80, E=1.50, and E_n=1.50). The Y-randomization test performed five times for these divisions into training and test sets gave $CCR_{test} < 0.6$, except for one value of $CCR_{test} = 0.86$ for MOE descriptors (division 2). CCR_{train} for this model was equal to 0.7. The test set in this division contains only seven compounds. If the activities of 0 and 1 are assigned randomly with a probability of 0.5, the probability that for six out of seven compounds the activities will be predicted correctly and $CCR_{test} = 0.86$ is

Table 6. *k*NN Classification Models with Highest CCR$_{test}$ Values for All Divisions into Training and Test Sets and All Collections of Descriptors: TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative, S-Sensitivity, SP-Specificity, E-Enrichment, E_n-the normalized Enrichment, CCR-Correct Classification Rate[a]

| collection of descriptors | divisions | confusion matrix (test set) | | | | statistics for the test set | | | | | CCR$_{train}$ | applicability domain |
		TP	TN	FP	FN	S	SP	E	En	CCR		Lowest Z values
Dragon	**1**	**13**	**11**	**1**	**3**	**0.81**	**0.92**	**1.63**	**1.81**	**0.86**	**0.70**	1.9
	2	8	10	0	4	0.67	1.00	1.83	2.00	0.82	0.76	0.9
	3	**11**	**10**	**2**	**2**	**0.85**	**0.83**	**1.63**	**1.67**	**0.84**	**0.77**	0.4
CMTD	1	9	12	4	6	0.60	0.75	1.43	1.41	0.68	0.66	1.4
	2	6	11	2	7	0.46	0.85	1.50	1.50	0.65	0.72	1.3
	3	13	10	8	9	0.59	0.56	1.13	1.14	0.58	0.78	1.4
MolconnZ/CMTD	1	9	8	6	2	0.81	0.57	1.36	1.31	0.68	0.74	1.6
	2	9	9	1	11	0.45	0.90	1.35	1.64	0.60	0.67	1.1
	3	7	6	3	4	0.64	0.67	1.27	1.31	0.65	0.73	1.2
MOE	1	12	12	10	3	0.80	0.55	1.35	1.28	0.65	0.77	0.2
	2	**3**	**4**	**0**	**0**	**1.00**	**1.00**	**2.33**	**2.00**	**1.00**	**0.70**	0.1
	3	11	12	11	3	0.63	0.67	1.25	1.30	0.64	0.79	0.3
CoMFA	**1**	**9**	**5**	**2**	**0**	**1.00**	**0.71**	**1.45**	**1.56**	**0.88**	**0.81**	0.6
	2	**10**	**6**	**4**	**0**	**1.00**	**0.60**	**1.43**	**1.43**	**0.80**	**0.84**	2
	3	**12**	**12**	**1**	**2**	**0.86**	**0.92**	**1.78**	**1.84**	**0.89**	**0.76**	0.7
VolSurf[b]	**1**	**4**	**7**	**1**	**1**	**0.80**	**0.86**	**2.08**	**1.73**	**0.85**	**0.77**	2
	2	6	7	5	3	0.67	0.58	1.27	1.23	0.62	0.86	2
	3	8	11	3	4	0.67	0.79	1.58	1.52	0.73	0.76	2
CoMMA/MOE	1	12	12	4	4	0.75	0.75	1.50	1.50	0.75	0.77	0.4
	2	8	10	2	5	0.62	0.83	1.54	1.57	0.72	0.74	0.2

[a] Statistics for acceptable predictive models are printed in bold. [b] Five compounds of test set of division 1 and four compounds of test sets of divisions 2 and 3 were beyond the applicability domain with the cutoff distance with Z = 2 from compounds of the training set and were not classified.

$7/128 = 0.055$. Nevertheless, we cannot consider this model acceptable. At the same time, CCR$_{train}$ values were equal or higher than 0.7 two times for MOE descriptors, three times for VolSurf descriptors, and five times for Dragon descriptors. Thus, again we can make a conclusion that CCR$_{train}$ alone is not a good characteristic of the classification accuracy.

SUPPORT VECTOR MACHINES

In this study we have used SVM with a linear kernel as described above. As in the previous cases, for models built with real and randomized activities most CCR values for the training sets were similar and high. At the same time, all CCR values for the test sets were lower than 0.7, except for division 2 for CoMFA descriptors (Table 2) for which CCR$_{test}$ = 0.75 (S=0.67, SP=0.86, E=1.52, and E_n=1.65). For this model, CCR$_{train}$ = 0.83. The Y-randomization test performed five times gave the highest CCR$_{test}$ = 0.56. However, for one of the randomizations CCR$_{train}$ = 0.84. Again, in this case, CCR$_{train}$ was not a good characteristic of the classification accuracy.

CONCLUSIONS

The objective of this work was to conduct the most comprehensive QSAR analysis of a data set of 98 ambergris fragrance compounds with complex stereochemistry. This data set consists of compounds of several structural types. Within each group, all compounds have almost identical structures, with only differences in chiralities of some atoms. As we have shown, a standard approach to QSAR studies, when only one method and one collection of descriptors is used, has a high chance to fail. Herein we have investigated a combinatorial QSAR approach, which considers all possible independent kNN models that can be built with various optimization methods and different descriptor collections.

This methodology became feasible due to the rapid development of computer technologies, which resulted in a dramatic increase of the speed of calculations. The following four QSAR methods have been included: classification *k*NN QSAR, decision tree,[31] binary QSAR,[31,37] and Support Vector Machines (SVM).[38] Currently, *k*NN classification QSAR is the only method which is fully automated in our laboratory. The following seven collections of descriptors have been calculated: CoMFA,[32] CoMMA,[33] MOE,[31] chirality descriptors,[27] MolcoonZ,[35] Dragon,[34] and VolSurf.[36] CoMMA descriptors were used in combination with MOE descriptors. MolconnZ descriptors were used in combination with chirality descriptors. A sphere exclusion algorithm was used to divide a data set into diverse and representative training and test sets. QSAR models were built using all possible combinations of data modeling techniques, collections of descriptors, and corresponding training and test sets. It was found that not all combinations of modeling methods and descriptor collections produce valid QSAR models. This fact itself corroborates the necessity for an automated combinatorial QSAR procedure in order to generate and mine the space of QSAR models to identify all validated models. Using the combi-QSAR approach, we were able to obtain several predictive QSAR models for this data set.

*k*NN classification method in combination with CoMFA descriptors gave predictive models for all divisions of a data set into training and test sets (Table 6). Thus, for our data set, the combination of *k*NN classification with CoMFA descriptors is the best combination of a QSAR method and the descriptor collection. Multiple predictive QSAR models have been obtained using the *k*NN classification with Dragon, MOE, VolSurf, and CoMMA/MOE descriptors but only for one or two of the divisions of a data set into training and test sets. We showed that statistical significance of QSAR classification models can be evaluated by a robustness

parameter, which was defined in the Y-randomization section. We suggest that this simple parameter should be used in classification QSAR studies.

Several predictive models were obtained for one of the divisions into training and test sets using other methods. Thus, a combination of SVM with CoMFA descriptors gave one predictive model; binary QSAR gave predictive models in combination with Dragon descriptors and in combination with VolSurf descriptors. The Decision Tree gave one predictive model in combination with CoMMA/MOE descriptors. The Decision tree and Binary QSAR were used as implemented in the MOE package. Relative failure of these methods can be partially explained by the fact that their direct use does not allow variable selection. In the future, we will develop automatic variable selection procedures for these methods, similar to those used in *k*NN. Low prediction accuracy of almost all SVM models is probably the consequence of using its linear version. Low prediction accuracy of all models built with CMTD descriptors can be explained by the fact that they require exhaustive calculations with different chirality correction values and different subclasses of descriptors, as it was described in our previous papers.[27,28]

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) Anonis, D. P. Ambergris. *Perfume Flavorist* **1995**, *20*, 7−11.

(2) Cambie, R. C. Perfumes related to Ambergris. *Chem. Ind. New Zealand* **1967**, *November*, 4−8.

(3) Lederer, E. Chemistry and biochemistry of some mammalian secretions and excretions. *J. Chem. Soc.* **1949**, 2115−2125.

(4) Buchbauer, G.; Heneis, V. M.; Krejci, V.; Talsky, C.; Wunderer, H. Über eine neue Synthese von Desmethylambraoxid. *Monatsh-Chem.* **1985**, *116*, 1345−1358.

(5) Ohloff, G.; Naef, F.; Decorzant, R.; Thommen, W.; Sund, E. Synthesis of potential ambra odorants. 5,5,9-Trimethyldecalyl derivatives. *Helv. Chim. Acta* **1973**, *56*, 1414−1448.

(6) Winter, B. Ring-opened Analogues of *Ambrox*: Synthesis and Structure-Odor Relationships. *Helv. Chim. Acta* **1989**, *72*, 1278−1283.

(7) Vial, C.; Thommen, W.; Naef, F. Structure−activity relationship in ambergris-type woody odorants possessing a hydronaphthalene skeleton. *Helv. Chim. Acta* **1989**, *72*, 1390−1399.

(8) Gorbachov, M. Yu.; Rossiter, K. J. A New Electronic-Topological Investigation on the Relationship between Chemical Structure and Ambergris Odour. *Chem. Senses* **1999**, *24*, 171−178.

(9) Bersuker, I. B.; Dimoglo, A. S.; Gorbachov, M. Yu.; Koltsa, M. N.; Vlad, P. F. Structural and electronic origin of ambergris odor of cyclic compounds. *Nouv. J. Chim.* **1985**, *9*, 211−218.

(10) Dimoglo, A. S.; Vlad, P. F.; Shvets, N. M.; Coltsa, M. N. Electronic-topological investigations of the relationship between chemical structure and ambergris odor. *New. J. Chem.* **1995**, *19*, 1217−1226.

(11) Bersuker, I. B.; Dimoglo, A. S. In *Reviews in Computational chemistry*; Lipkowitz, K., Boyd, D., Eds.; VCH: New York, 1991; Chapter 10.

(12) Shvets, N. Applied program system for the prognosis of biological activity of chemical compounds: development and use. *Comput. J. Moldova (Kishinev)* **1993**, *1*, 101−110.

(13) Winter, B. In *QSAR: Quantitative Structure Activity Relationships in Drug Design (QSAR in olfaction: Ambergris type odorants)*; Alan R. Liss, Inc.: 1989; pp 401−405.

(14) Bajgrowicz, J. A.; Frank, I. Camphor-derived ambergris/woody odorants: 1,7,7-trimethyl-2′-iso-propylspiro [bicyclo[2.2.1]heptane-2,4′-(1,3-dioxanes)]. *Tetrahedron: Asymmetry* **2001**, *12*, 2049−2057.

(15) Bajgrowicz, J. A.; Broger, C. *Molecular Modelling in the Design of New Odorants: Scope and Limitations*; Baser, K. H. C., Ed.; Proceedings of the 13th International Congress of Flavours, Fragrances and Essential Oils; AREP: Istanbul, 1995; Vol. 3, pp 1−15.

(16) CATALYST 4.6; Molecular Simulations Inc. (Accelrys): San Diego, CA, 2000.

(17) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(18) Bilke, S.; Mosandl, A. Enantioselective analysis of 2-methyl-4-(2,2,3-trimethylcyclopent-3-en-1-yl)-but-2-enol, 2-methyl-4-(2,2,3-trimethylcyclopent-3-en-1-yl)-but-2-enal and α-campholene aldehyde by capillary gas chromatography. *J. Sep. Sci.* **2001**, *24* (10), 819−822.

(19) Buchbauer, G.; Lebada, Ph.; Wiesinger, L.; Weiss-Greiler, P.; Wolschann, P. On the odor of the enantiomers of Madrol. *Chirality* **1997**, *9*, 380−385.

(20) Horton, H. R.; Moran, L. A.; Ochs, R. S.; Rawn, J. D.; Scrimgeour, K. G. *Principles of Biochemistry*; Neil Patterson Publishers Prentice Hall: Englewood Cliffs, NJ, 2002.

(21) Potapov, V. M. *Stereochemistry*; Khimia, Moscow, 1988.

(22) Solms, J.; Vuataz, L.; Egli, R. H. The taste of L- and D-amino acids. *Experientia* **1965**, *21*, 692−694.

(23) Schiffman, S. S.; Clark, T. B.; Gagnon, J. 3D Influence of chirality of amino acids on the growth of perceived taste intensity with concentration. *Physiol. Behav.* **1982**, *28*, 457−465.

(24) DeCamp, W. H. The FDA Perspective on the development of stereoisomers. *Chirality* **1989**, *1*, 2−6.

(25) Hutt, A. J.; Tan, S. C. Drug chirality and its clinical significance. *Drugs* **1996**, *52*, 1−12.

(26) Wnendt, S.; Zwingenberger, K. Thalidomide's chirality. *Nature* **1997**, *385*, 303−304.

(27) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147−158.

(28) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 144−154.

(29) Kovatcheva, A.; Buchbauer, G.; Golbraikh, A.; Wolschann, P. QSAR Modeling of α-campholenic derivatives with sandalwood odor. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 259−266.

(30) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, *22*, 69−77.

(31) http://www.chemcomp.com/fdept/prodinfo.htm#Cheminformatics.

(32) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Efect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(33) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.

(34) http://www.disat.unimib.it/chm/Dragon.htm.

(35) http://www.eslc.vabiotech.com.

(36) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *THEOCHEM* **2000**, *503*, 17−30.

(37) http://www.smi.stanford.edu/projects/helix/psb99/Labute.pdf.

(38) Vapnik, V. N. In *The nature of statistical learning theory*; Springer: NewYork, 2000.

(39) http://www.tripos.com.

(40) Bonchev, D. Overall connectivity − a next generation molecular connectivity. *J. Mol. Graph. Model.* **2001**, *20 Sp. Iss. SI*, 65−75.

(41) Gutman, I.; Ruscić, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys.* **1975**, *62*, 3399.

(42) Randić, M. On Characterization on Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(43) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.

(44) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Wiley: New York, 1986.

(45) Rücker, G.; Rücker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695.

AMBERGRIS FRAGRANCE COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004 **595**

(46) Bonchev, D. Overall connectivity and molecular complexity. In *Topological indices and related descriptors*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 361−401.

(47) Bonchev, D. Novel indices for the topological complexity of molecules. *SAR/QSAR Environ. Res.* **1997**, *7*, 23−43.

(48) Kier, L. B. A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(49) Kier, L. B. Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant. Struct-Act. Relat.* **1987**, *6*, 8−12.

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

(51) Hall, L. H.; Mohney, B. K.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43−51.

(52) Hall, L. H.; Mohney, B. K.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(53) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: 1999.

(54) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. The E-State Fields. Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.

(55) Kier, L. B.; Hall, L. H. A Differential Molecular Connectivity Index. *Quant. Struct.-Act. Relat.* **1991**, *10*, 134−140.

(56) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

(57) Wiener, H. J. Structural determination of paraffin boiling points. *Am. Chem. Soc.* **1947**, *69*, 17−20.

(58) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **1947**, *15*, 419−420.

(59) Shannon, C.; Weaver, W. In *Mathematical theory of Communication*; University of Illinois: Urbana, 1949.

(60) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer discrimination by topological information approach. *J. Comput. Chem.* **1981**, *2*, 127−148.

(61) Wiener, H. Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636−2638.

(62) Balaban, A. T. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.

(63) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(64) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(65) Stanton, D.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(66) http://www.research.ibm.com/comma.

(67) Todeschini, R.; Consonni, V. In *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(68) Gutman, I.; Rücker, C.; Rücker, G. On walks in molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739−745.

(69) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(70) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520−525.

(71) Randic, M. Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron* **1975**, *31*, 1477−1481.

(72) Randic, M. Molecular profiles. Novel geometry-dependent molecular descriptors. *New J. Chem.* **1995**, *19*, 781−791.

(73) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrat. Spectrosc.* **1999**, *19*, 151−164.

(74) Hemmer, M. C.; Gasteiger, J. Prediction of three-dimensional molecular structures using information from infrared spectra. *Anal. Chim. Acta* **2000**, *420*, 145−154.

(75) Schuur, J. H.; Setzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(76) Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. Modeling and prediction by using WHIM descriptors in QSAR studies: toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere* **1996**, *32*, 1527−1545.

(77) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS−WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.

(78) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.

(79) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.

(80) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 357−369.

(81) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 241−253.

(82) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and k-nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(83) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the *k*-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(84) Willett, P.; Winterman, V. A. Comparison of some measures for the determination of intermolecular structural similarities. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(85) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth International Group: 1984.

(86) Schölkopf, B.; Smola, J. A. Learning with Kernels. In *Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*; Ed.; MIT Press: Cambridge, 2002.

(87) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: 1995; pp 309−318.

(88) Kubinyi, H.; Hamprecht, F. A., Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(89) Golbraikh, A.; Tropsha, A. Beware of q²! *J. Mol. Graphics Model.* **2002**, *20*, 269−276.

(90) Bajgrowicz, J. A. Camphor derivatives as new odorants. Eur. Pat. 7616641, 1997; *Chem. Abstr.* **1997**, 126:225430.

(91) Ohloff, G.; Giersch, W.; Pickenhagen, W.; Furrer, A.; Frei, B. Significance of the Geminal Dimethyl Group in the Odor Principle of Ambrox. *Helv. Chim. Acta* **1985**, *68*, 2022−2029.

(92) Ohloff, G.; Giersch, W.; Schulte-Elte, K. H.; Vial, Ch. Stereochemistry-odor relations of 1-decalone derivatives and their oxygen analogues. *Helv. Chim. Acta* **1976**, *59*, 1140−1157.

(93) Ohloff, G.; Vial, Ch.; Demole, E.; Enggist, P.; Giersch, W.; Jegou, E.; Caruso, A. J.; Polonsky, J.; Lederer, E. Conformation-Odor Relationships in Norlabdane Oxides. *Helv. Chim. Acta* **1986**, *69*, 163−173.

(94) Escher, S.; Giersch, W.; Niclass, Y.; Bernardinelli, G.; Ohloff, G. Configuration-odor relationships in 5β-ambrox. *Helv. Chim. Acta* **1990**, *73*, 1935−1947.

(95) Torre, M. C.; Garcia, I.; Sierra, M. A. Straightforward synthesis of the strong ambergris odorant bicyclohomofarnesal and its endo-isomer from R−(+)-sclareolide. *Tetrahedron Lett.* **2002**, *43*, 6351−6353.

CI034203T