Supporting Information

Mapping Sleep-Promoting Volatiles in Aromatic Plants with Machine Learning: A Comprehensive Survey of 2,300 Molecules

Peiqin Shi[a,d,†], Xing Huang[a,†], Qinfei Ke[a], Xingran Kou[a,*], Dachuan Zhang[b,c,*]

[a] Collaborative Innovation Center of Fragrance Flavour and Cosmetics, School of Perfume and Aroma Technology, Shanghai Institute of Technology, Shanghai 201418, China

[b] Department of Food Science and Technology, Faculty of Science, National University of Singapore, 117542, Singapore

[c] National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiangsu 215123, China

[d] School of Food Science and Technology, Jiangnan University, Wuxi, Jiangsu, 214122, China

† Contributed equally

* Correspondence:

Xingran Kou, kouxr@sit.edu.cn, +86-21-60877237

Dachuan Zhang, dachuan.zhang@nus.edu.sg, +65 8027 9362

Hyperparameters used for the knowledge graph-enhanced molecular contrastive learning with functional prompt (KANO) model

The batch size was set to 64, the number of epochs was set to 100, the number of runs was set to 5, the random seed was set to 43, the initial learning rate was set to 0.0001, the final learning rate was set to 0.0001, the maximum learning rate was set to 0.001, the depth was set to 3, the number of hidden neurons was set to 300, the number of hidden neurons in the feed-forward network was set to 300, the number of layers in the feed-forward network was set to 2, the ensemble size was set to 1, the number of learning rate schedules was set to 1, the number of warm-up epochs was set to 2, and the temperature was set to 0.1.

Hyperparameters used for the CHEM-BERT model

The number of epochs was set to 50, the batch size was set to 64, the learning rate was set to 0.0001, the random seed was set to 46, the number of layers was set to 4, the number of heads was set to 8, the embedding size was set to 1024, the model dimension was set to 1024, and the dropout rate was set to 0.2.

Hyperparameters used for the Attentive FP model

The batch size was set to 128, the number of epochs was set to 200, the dropout rate was set to 0.1, the fingerprint dimension was set to 150, the radius was set to 3, the value of T was set to 2, the weight decay was set to 2.9, and the learning rate was set to 3.5.

Hyperparameters used for the message-passing neural networks model

The batch size was set to 50, the number of epochs was set to 30, the initial learning rate was set to 0.0001, the maximum learning rate was set to 0.001, the hidden size was set to 300, the depth was set to 3, the number of hidden neurons in the feed-forward network was set to 300, the number of layers in the feed-forward network was set to 2, and the aggregation method was set to "mean".

Hyperparameters used for the random forest model combined with Molecular ACCess System Key fingerprints (RF-MACCS)

The criterion was set to 'gini', the maximum depth was set to 8, the maximum number of features was set to 'auto', the minimum number of samples per leaf was set to 1, the minimum number of samples per split was set to 2, the number of estimators was set to 60, and the random state was set to 42.

Hyperparameters used for the random forest model combined with RDKit fingerprints (RF-RDkit)

The criterion was set to 'gini', the maximum depth was set to 8, the maximum number of features was set to 'auto', the minimum number of samples per leaf was set to 1, the minimum number of samples per split was set to 2, the number of estimators was set to 260, and the random state was set to 42.

Hyperparameters used for the support vector machine model combined with Molecular ACCess System Key fingerprints (SVM-MACCS)

The parameter C was set to 1, the cache size was set to 200, the polynomial degree was set to 3, the gamma value was set to 0.1, the kernel function was set to 'rbf', the maximum number of iterations was set to -1, the probability was set to enabled, and the tolerance was set to 0.001.

Hyperparameters used for the extreme gradient boosting model combined with Molecular ACCess System Key fingerprints XGB-MACCS model

The column sample ratio per tree was set to 0.5, the learning rate was set to 0.01, the maximum depth was set to 9, and the number of estimators was set to 200.

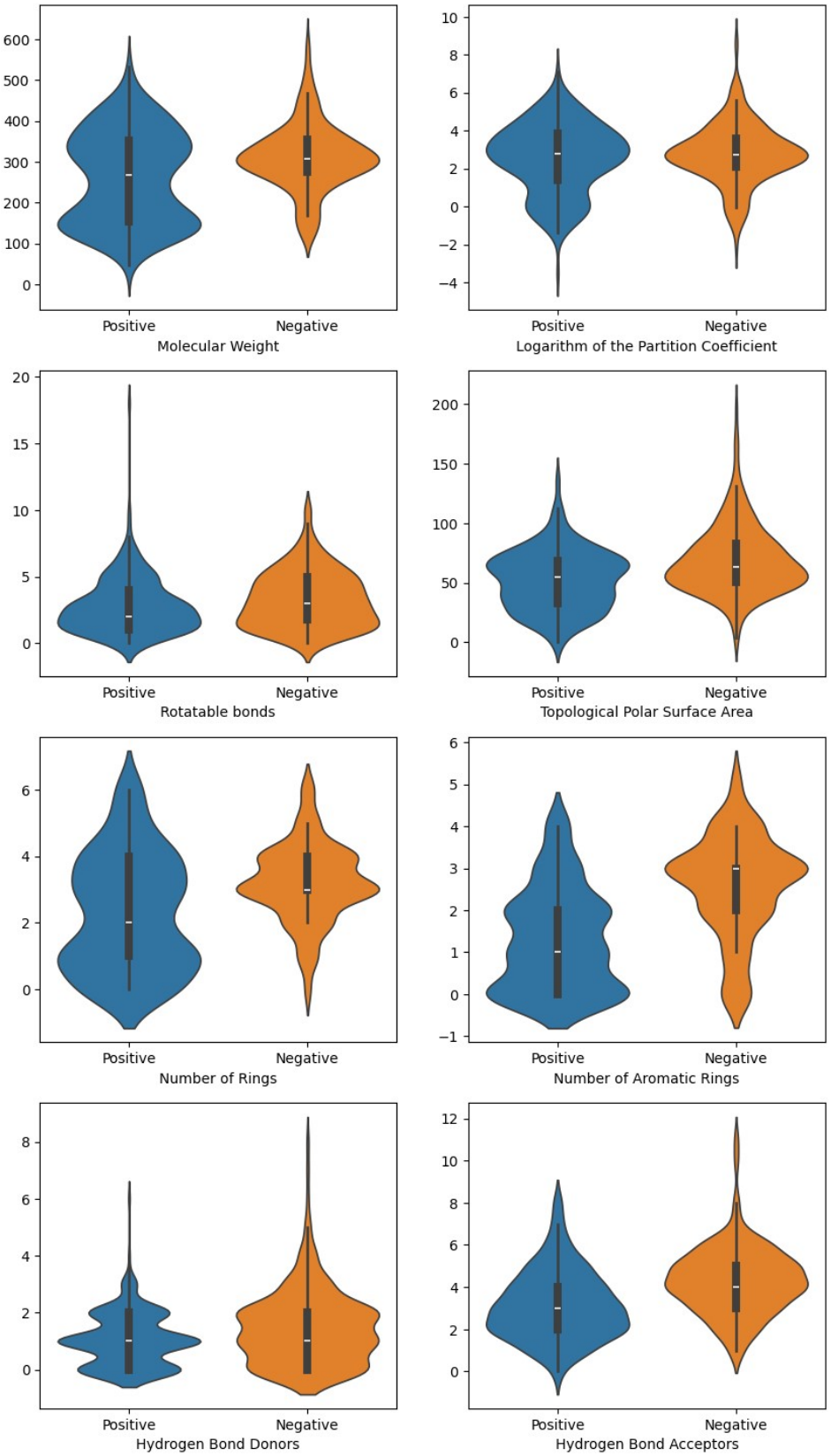


Figure S1. Comparison of physicochemical properties of positive and negative samples.

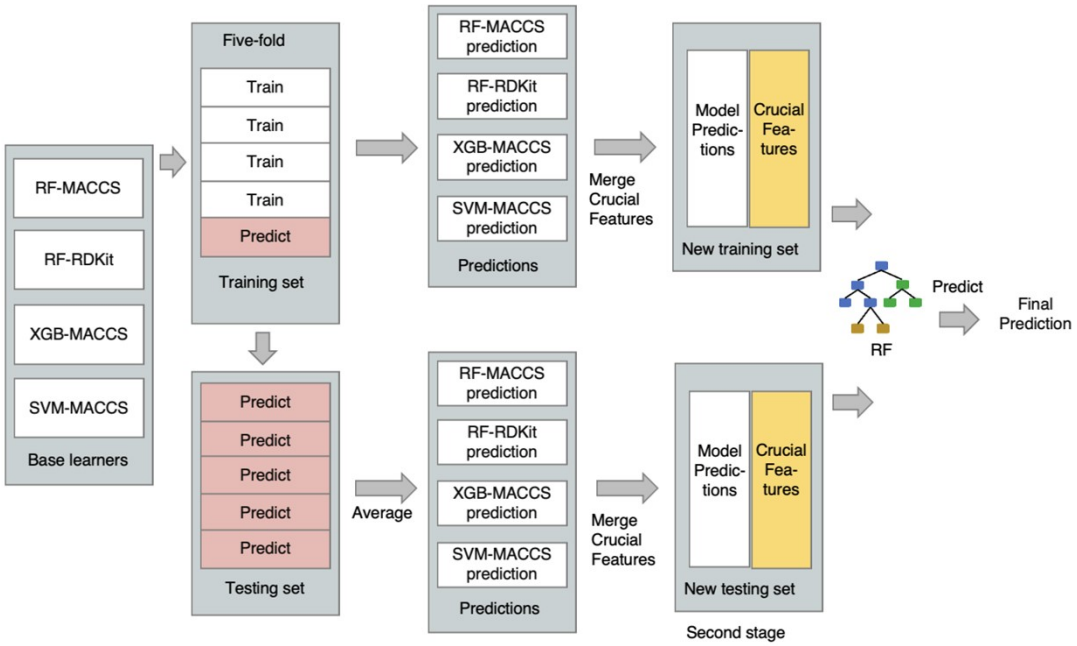


Figure S2. Stacking model framework. The stacking model operates in two stages: base learner training and prediction aggregation. First, the dataset is divided into training and testing sets. The training set undergoes five-fold cross-validation, splitting it into five subsets. The four base learners (RF-MACCS, RF-RDKit, XGB-MACCS, and SVM-MACCS) are each trained on four of these subsets and validated on the remaining subset, ensuring that each learner is trained and tested on distinct data partitions. The base learner's predictions on the training set are then combined to form an integrated "meta-training" set. In the second stage, a random forest model is trained on this meta-training set, learning to integrate the outputs from the base learners. For the testing set, each base learner independently generates predictions that are merged into a "meta-testing" set, which is finally processed by the trained random forest model to produce the final predictions.

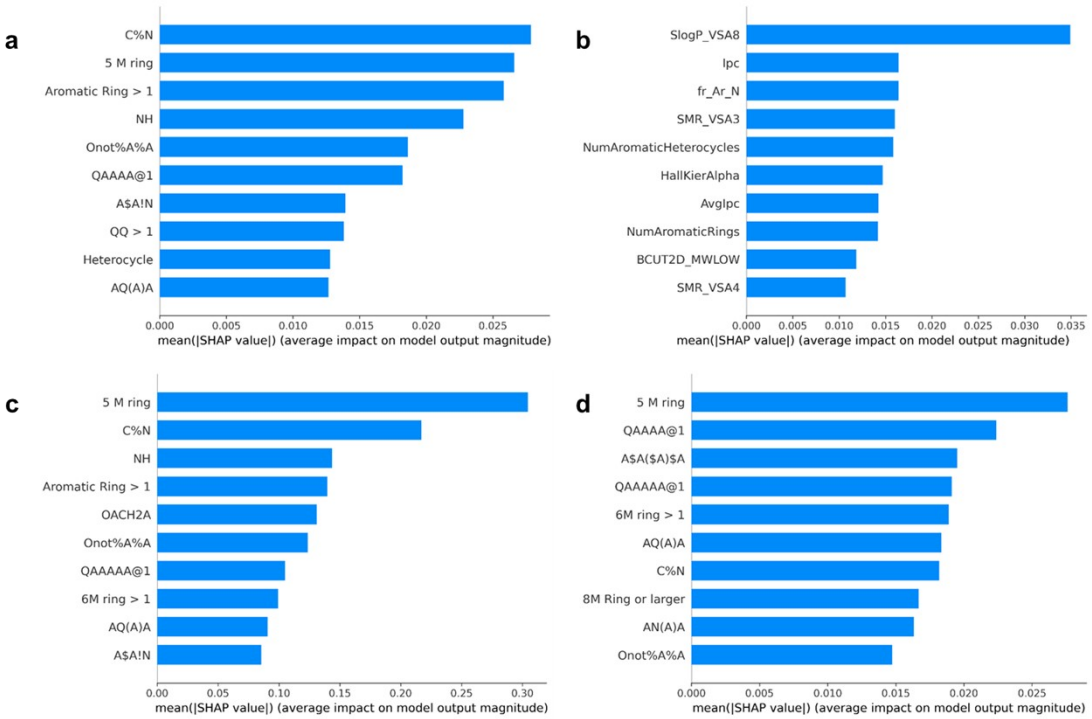


Figure S3. Top 10 key molecular descriptors across various models. (a) RF-MACCS model. (b) RF-RDKit model. (c) XGBoost-MACCS model. (d) SVM-MACCS model.

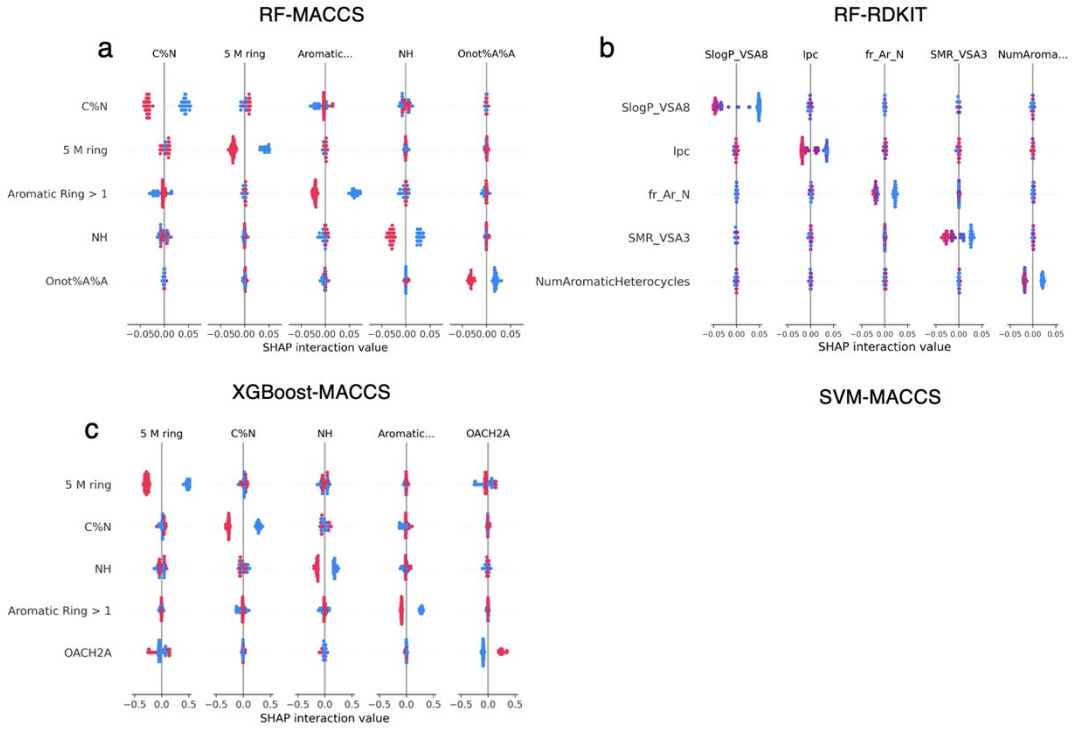


Figure S4. Interaction summary plot of decision tree-based models. (a) RF-MACCS model. (b) RF-RDKit model. (c) XGBoost-MACCS model. Each point in the plot represents the interaction value of a specific feature in the sample. A higher interaction value indicates that the interaction between this feature and another feature has a greater impact on the model's prediction.

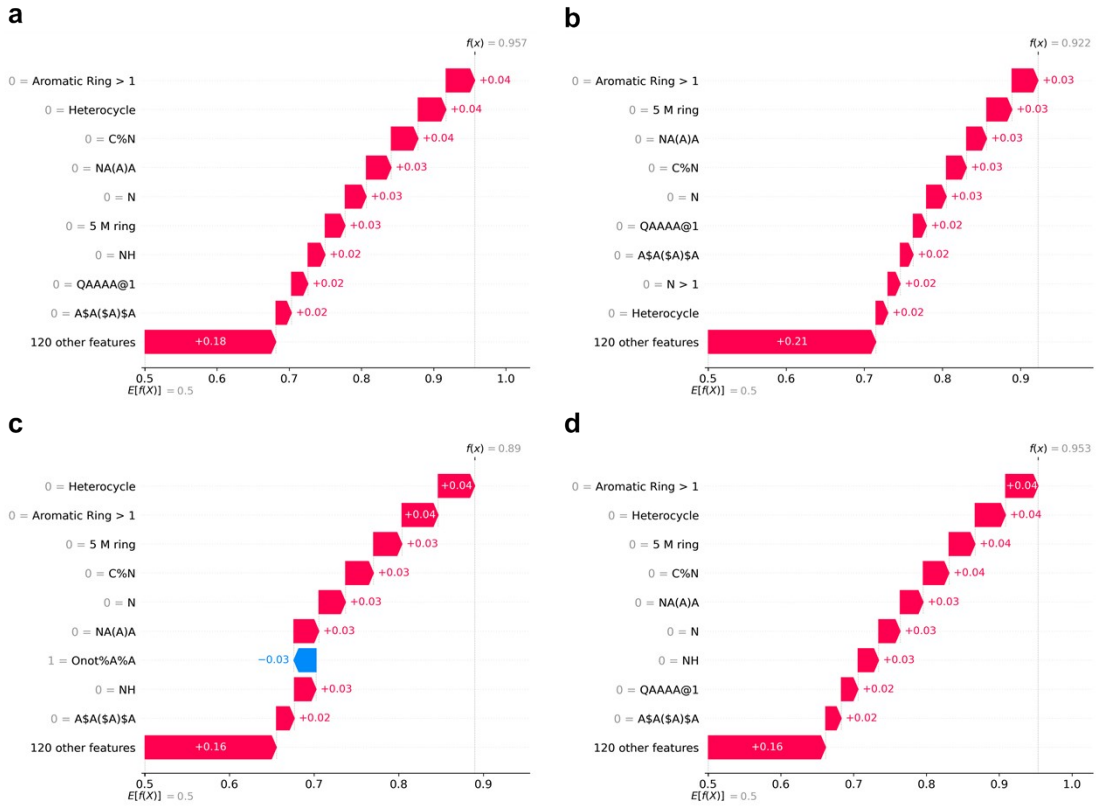


Figure S5. Critical features of selected VOCs identified by the RF-MACSS model. (a) Carvacrol, (b) Safranal, (c) Vanillin, and (d) Methyl eugenol.

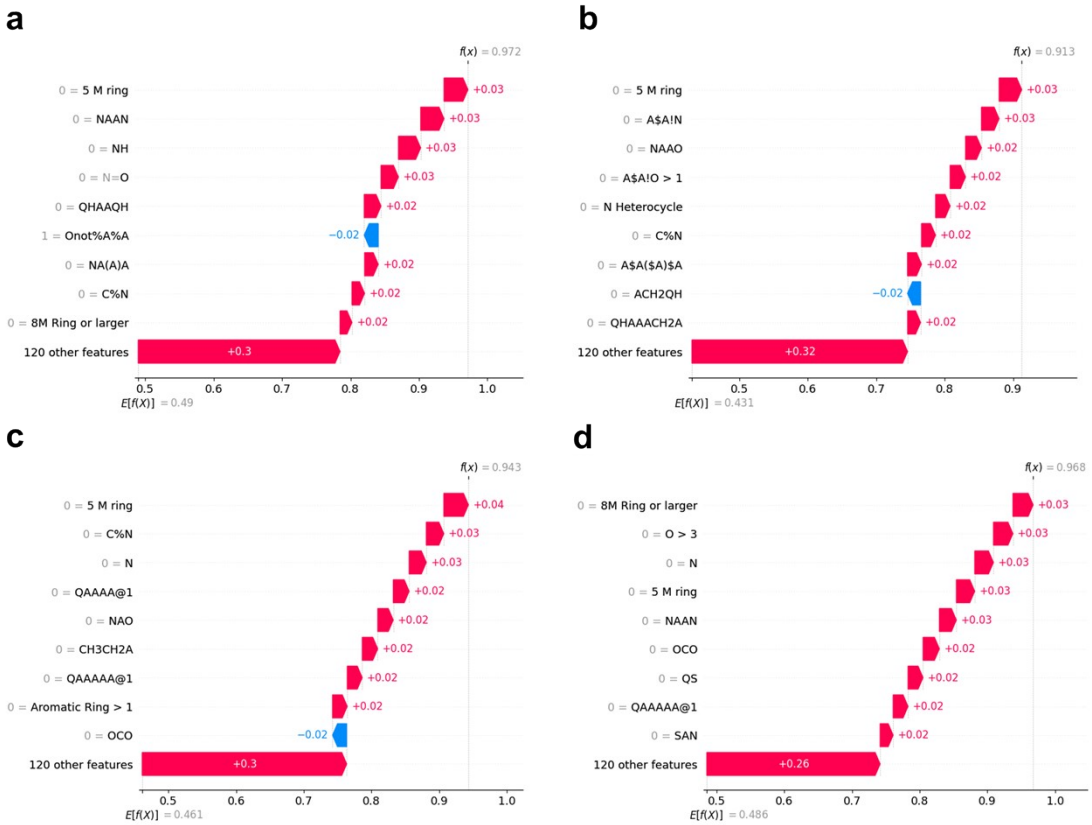


Figure S6. Critical features of selected VOCs identified by the SVM-MACCS model. (a) Carvacrol, (b) Safranal, (c) Vanillin, and (d) Methyl eugenol.

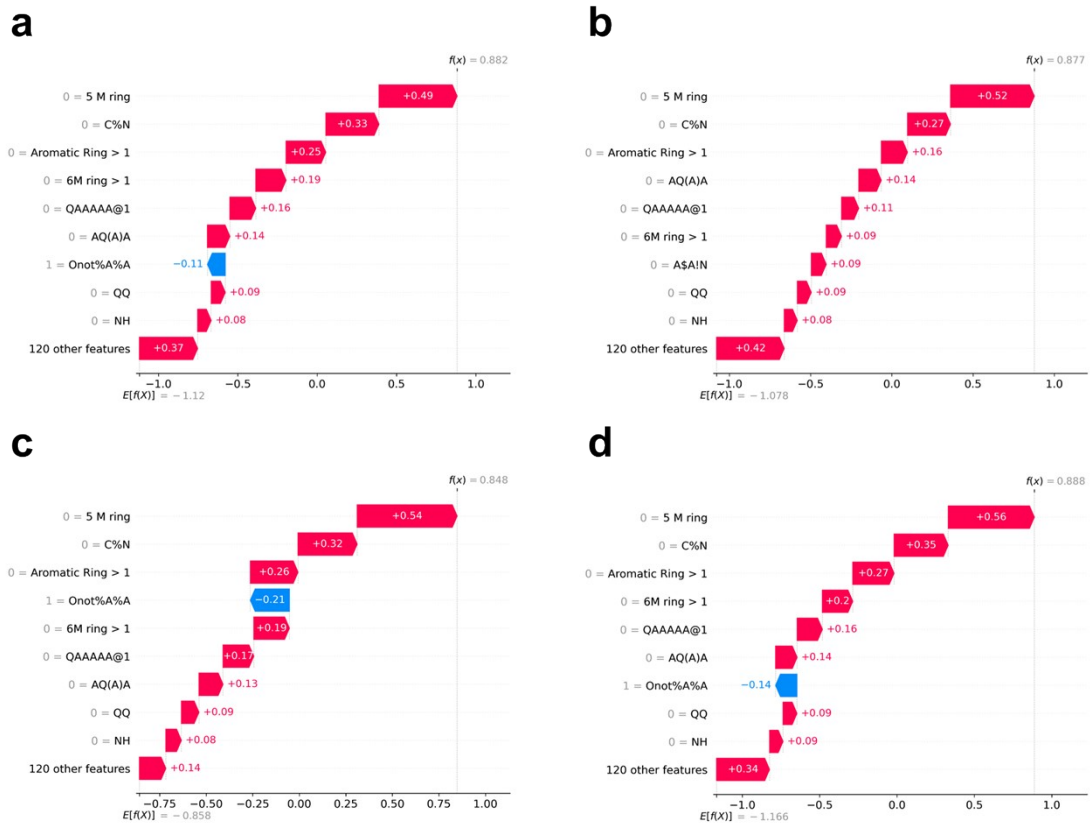


Figure S7. Critical features of selected VOCs identified by the XGBoost-MACCS model.

(a) Carvacrol, (b) Safranal, (c) Vanillin, and (d) Methyl eugenol.

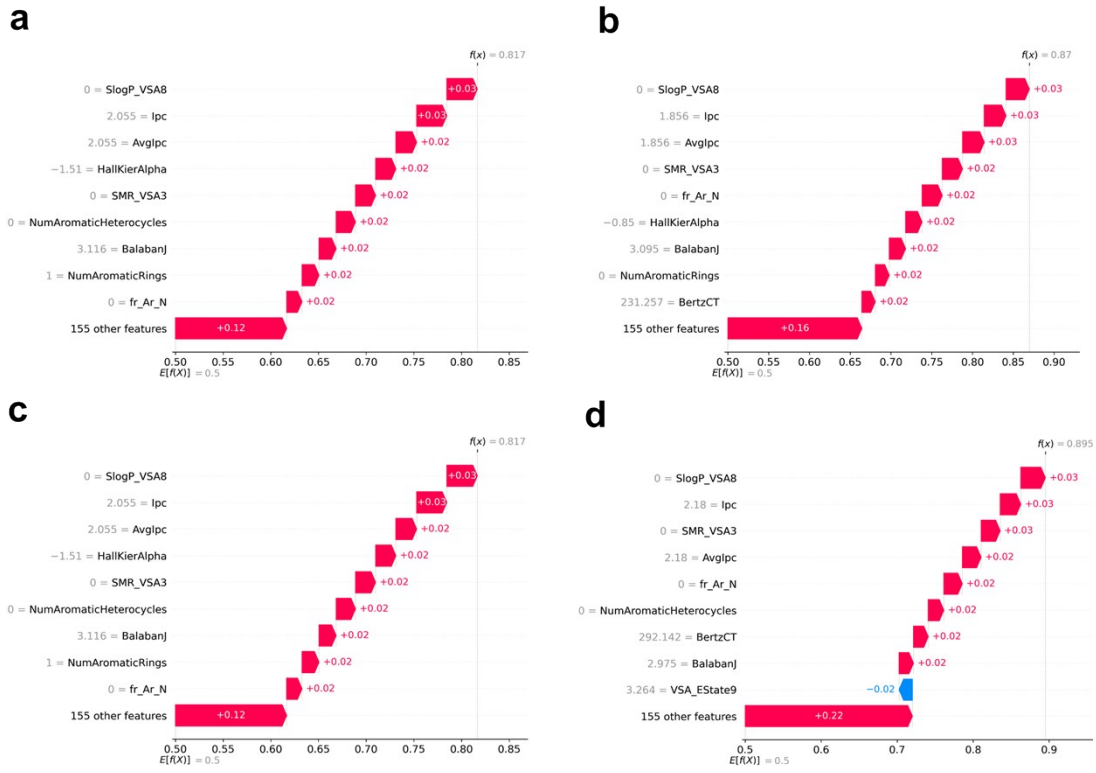


Figure S8. Critical features of selected VOCs identified by the RF-RDKit model. (a) Carvacrol, (b) Safranal, (c) Vanillin, and (d) Methyl eugenol.

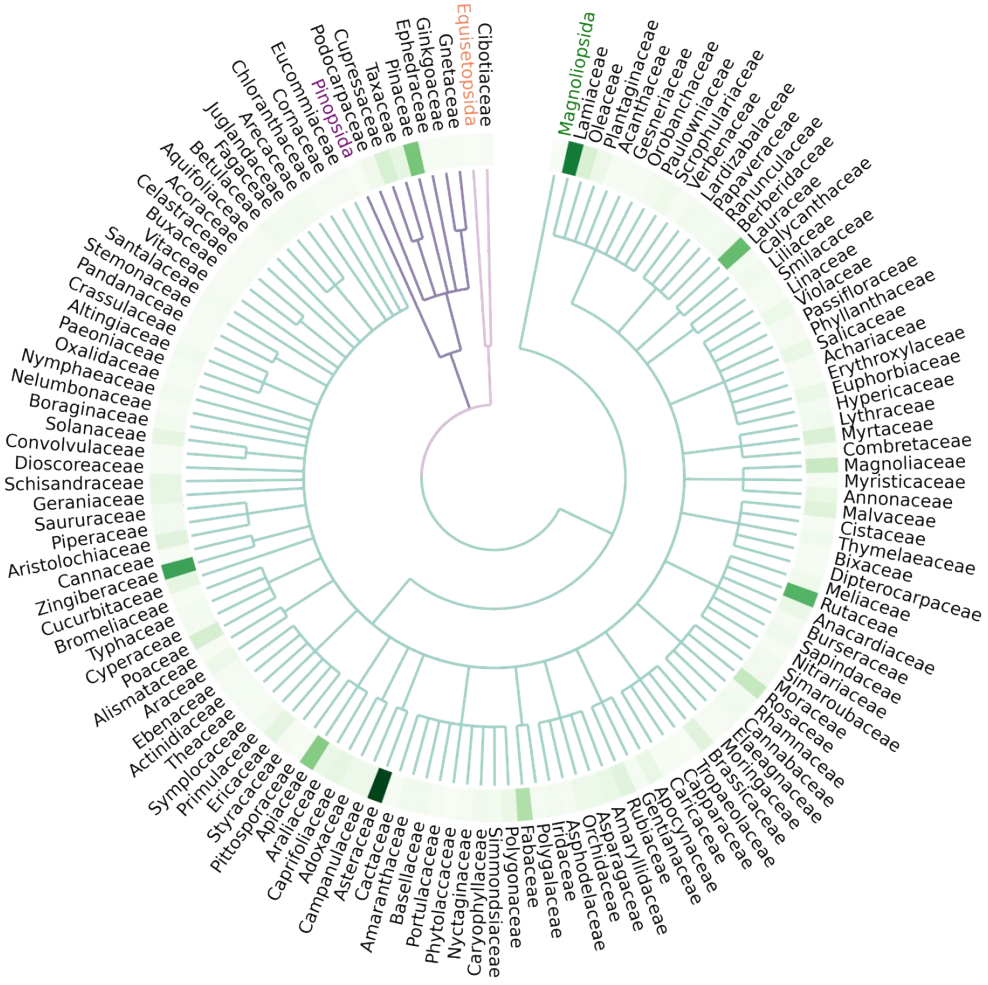


Figure S9. Phylogenetic tree of aromatic plants presenting the diversity (number of sleep-promoting VOCs) and content of sleep-promoting VOCs identified in these plants. Branch colors indicate major plant classes: green for Angiospermae, purple for Gymnospermae, and pink for Pteridophyta. The branching structure reflects their evolutionary divergence, with the outermost labels representing plant families. Each node corresponds to a taxonomic unit, progressing from classes to orders and families from the center outward. The accompanying heatmap illustrates the diversity and content of sleep-promoting VOCs identified in these plants based on the sum of their predictive scores and content in the plant extracts. Darker colors indicate a higher potential for sleep promotion.

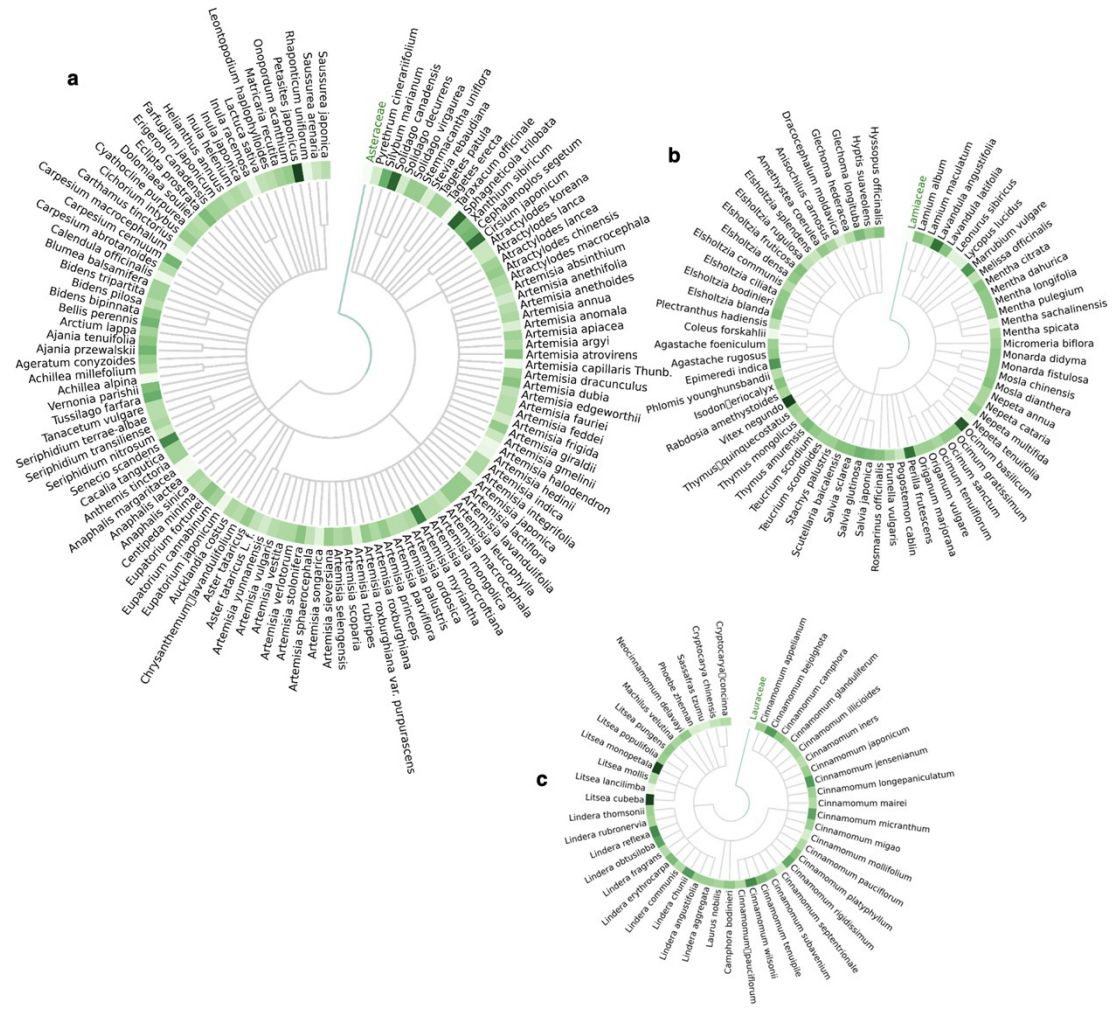


Figure S10. Phylogenetic tree of Asteraceae (a), Lamiaceae (b), and Lauraceae (c) plants from Magnoliopsida class presenting the diversity of sleep-promoting VOCs identified in these plants.

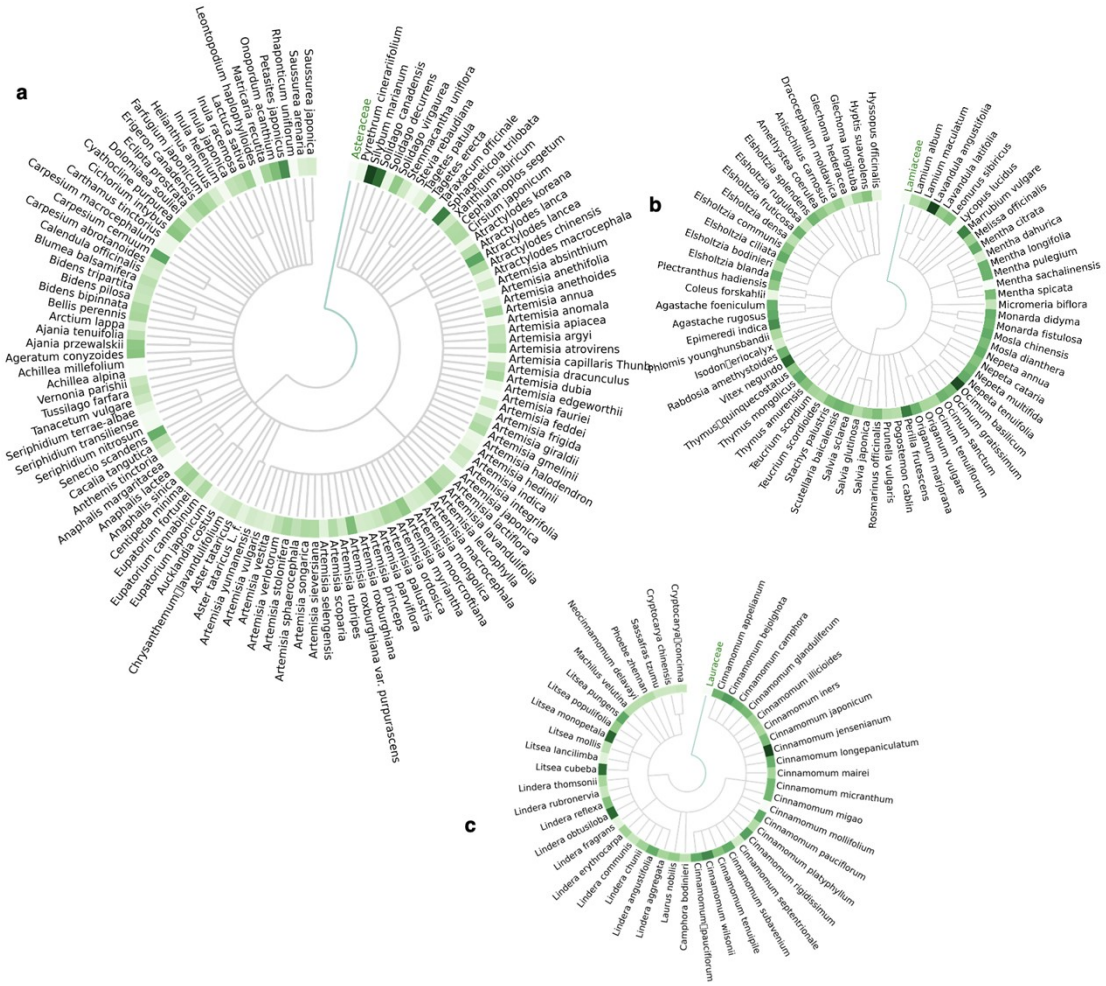


Figure S11. Phylogenetic tree of Asteraceae (a), Lamiaceae (b), and Lauraceae (c) plants from Magnoliopsida class presenting the content (content of individual sleep-promoting VOCs) and diversity (number of sleep-promoting VOCs) of VOCs identified in these plants.

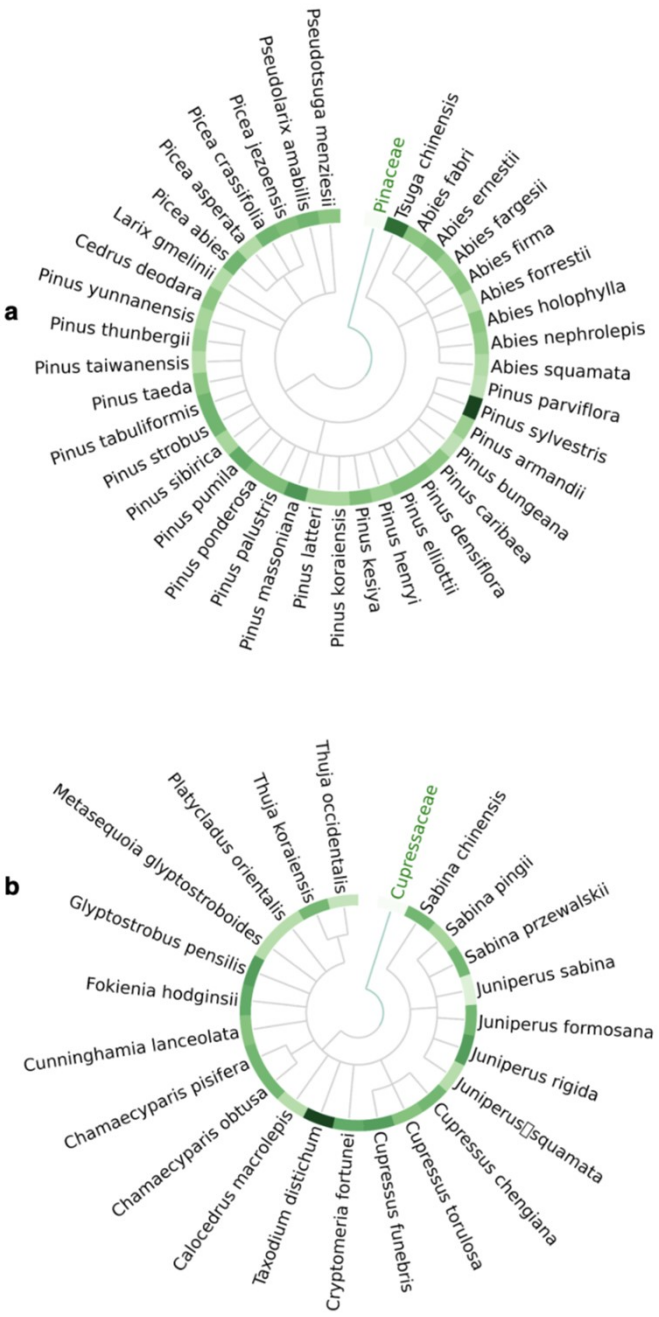


Figure S12. Phylogenetic tree of Pinaceae (a) and Cupressaceae (b) plants from Pinopsida class presenting the diversity of sleep-promoting VOCs identified in these plants.

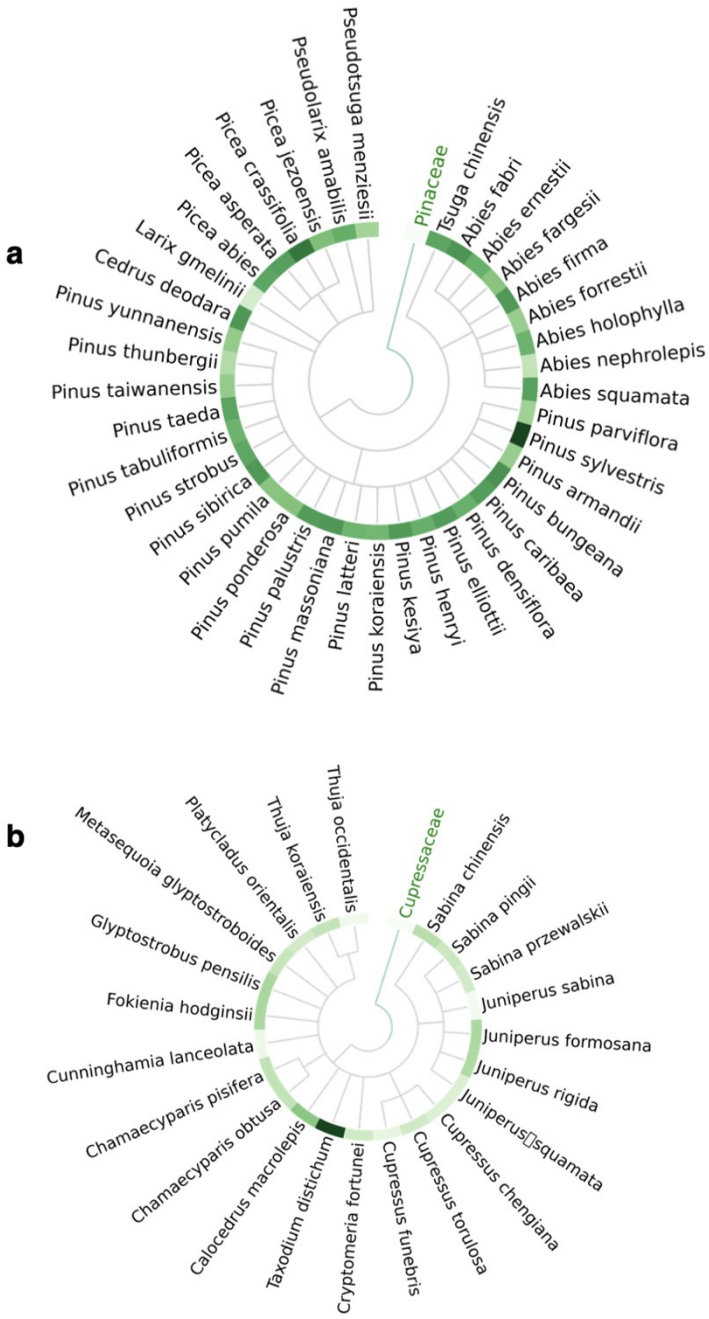


Figure S13. Phylogenetic tree of Pinaceae (a) and Cupressaceae (b) plants from Pinopsida class presenting the content and diversity of sleep-promoting VOCs identified in these plants.

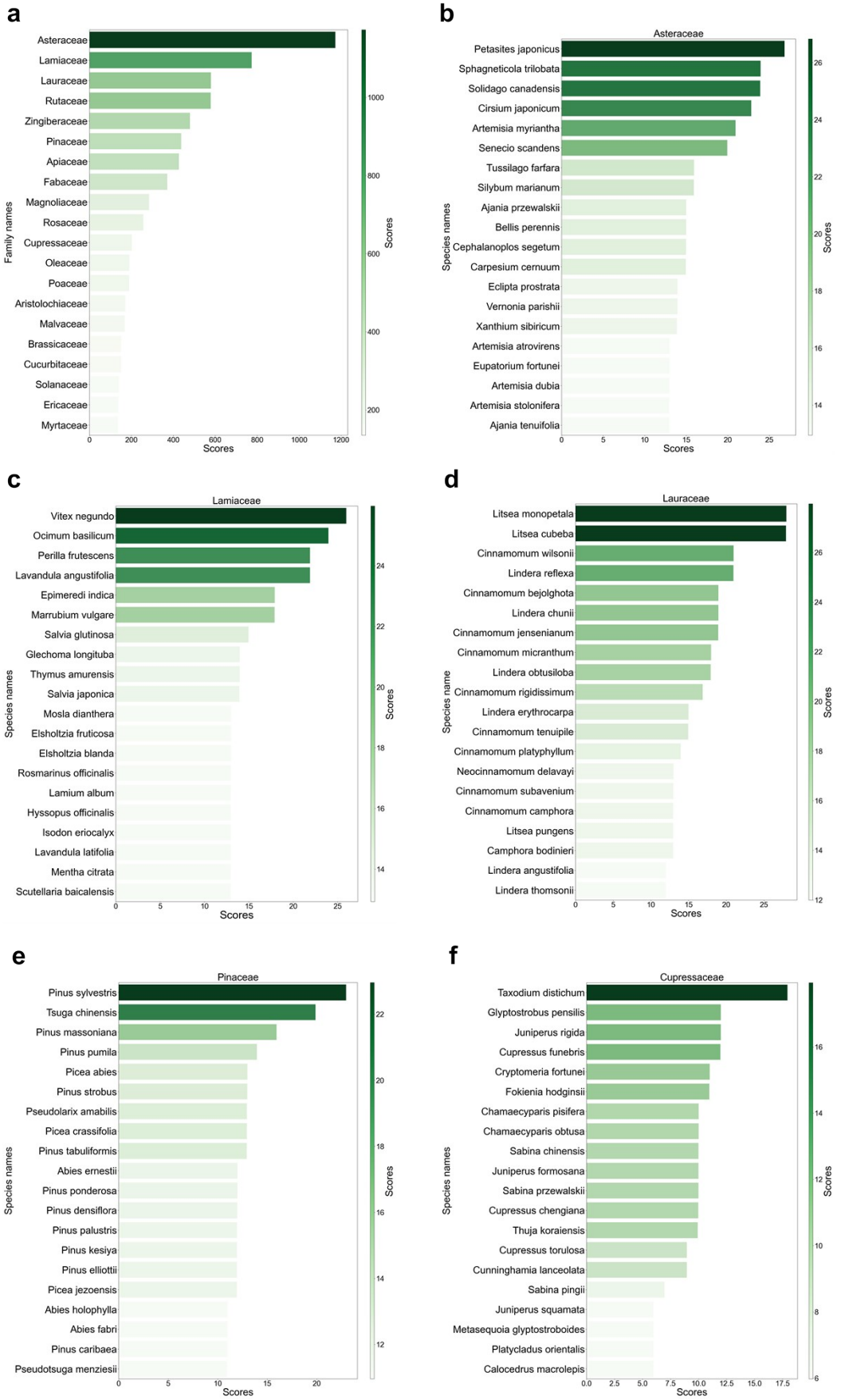


Figure S14. Aromatic plants with the highest diversity of sleep-promoting VOCs.

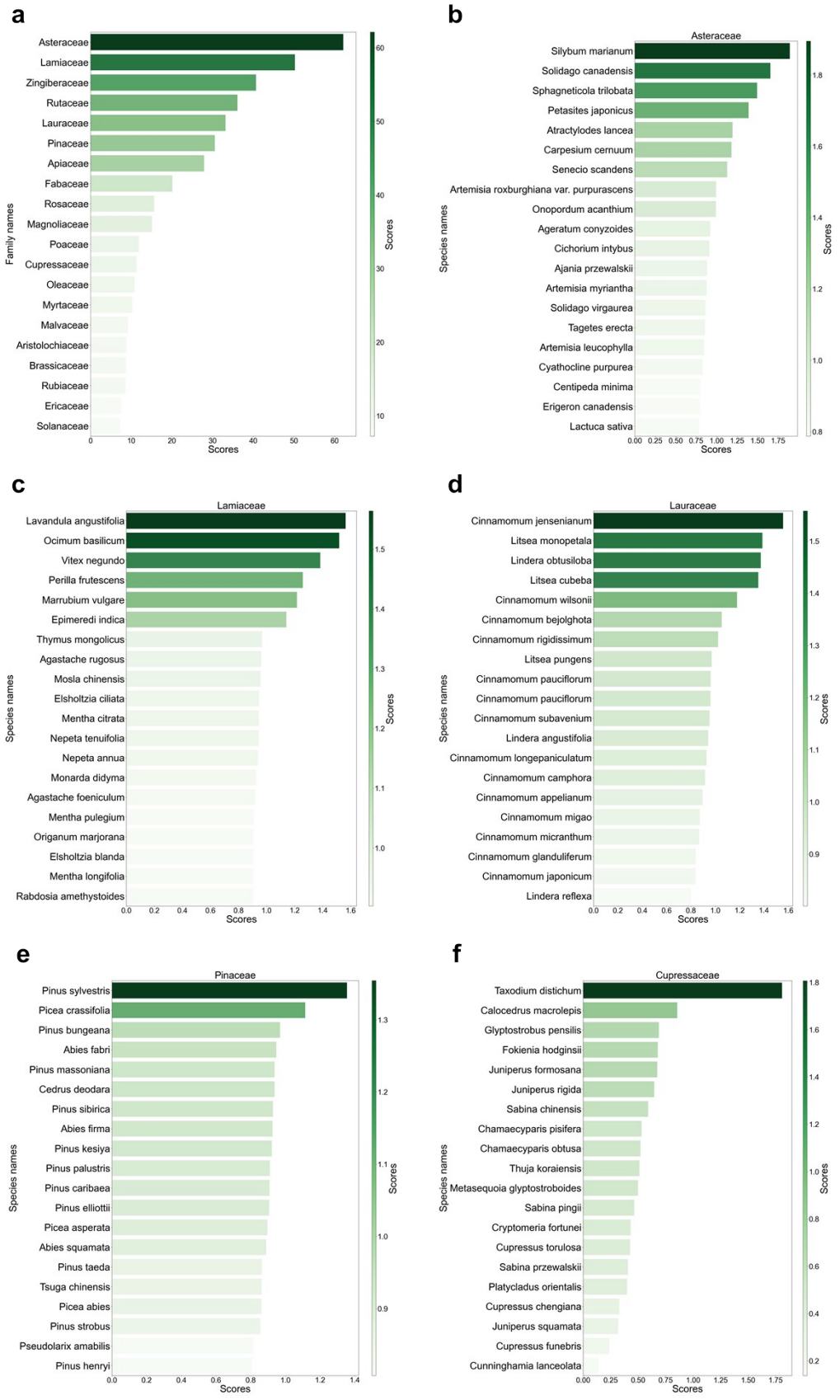
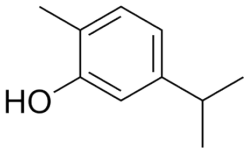
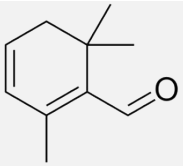
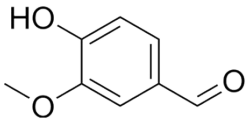
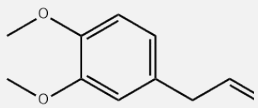


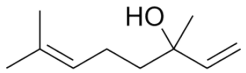
Figure S15. Aromatic plants with the highest diversity and content of sleep-promoting VOCs.

Table S1. Machine learning models' performance on the validation and test sets.

Algorithms	Descriptors	Validation set				Test set			
		AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall
KANO	Graph	0.94±0.032	0.845±0.047	0.853±0.053	0.810±0.161	0.901±0.037	0.829±0.047	0.805±0.091	0.881±0.089
CHEM-BERT	Graph	0.641±0.086	0.518±0.016	0.650±0.189	0.565±0.060	0.482±0.045	0.528±0.107	0.528±0.107	0.900±0.200
Attentive FP	Graph	0.823±0.043	0.715±0.103	0.700±0.131	0.844±0.095	0.876±0.025	0.780±0.042	0.756±0.100	0.873±0.085
MPNN	Graph	0.890±0.038	0.824±0.030	0.808±0.031	0.848±0.055	0.893±0.033	0.824±0.015	0.809±0.048	0.839±0.057
RF	ECFP4	0.938±0.005	0.858±0.006	0.847±0.007	0.875±0.014	0.948±0.023	0.886±0.031	0.883±0.037	0.905±0.031
	MACCS	0.946±0.002	0.869±0.008	0.866±0.009	0.873±0.019	**0.964±0.026**	**0.896±0.040**	**0.906±0.026**	**0.898±0.065**
	RDKit	0.937±0.007	0.864±0.009	0.849±0.011	0.886±0.017	**0.957±0.020**	**0.892±0.025**	**0.900±0.017**	**0.893±0.051**
GBDT	ECFP4	0.918±0.011	0.839±0.010	0.821±0.016	0.864±0.013	0.938±0.032	0.863±0.037	0.864±0.045	0.875±0.029
	MACCS	0.923±0.008	0.850±0.018	0.846±0.009	0.854±0.037	0.940±0.032	0.876±0.037	0.874±0.038	0.894±0.063
	RDKit	0.931±0.003	0.866±0.006	0.855±0.012	0.879±0.028	0.944±0.017	0.873±0.034	0.880±0.020	0.878±0.061
XGBoost	ECFP4	0.928±0.003	0.852±0.006	0.841±0.019	0.866±0.016	0.941±0.035	0.863±0.059	0.867±0.038	0.874±0.073
	MACCS	0.931±0.007	0.852±0.012	0.850±0.010	0.855±0.026	**0.954±0.028**	**0.886±0.027**	**0.890±0.027**	**0.893±0.051**
	RDKit	0.937±0.005	0.867±0.010	0.862±0.018	0.875±0.024	0.950±0.019	0.880±0.023	0.875±0.030	0.895±0.051
SVM	ECFP4	0.924±0.004	0.865±0.012	0.860±0.022	0.870±0.028	0.932±0.020	0.869±0.023	0.874±0.038	0.877±0.062
	MACCS	0.946±0.004	0.874±0.005	0.867±0.008	0.883±0.016	**0.967±0.022**	**0.886±0.031**	**0.900±0.027**	**0.883±0.070**
	RDKit	0.907±0.042	0.752±0.030	0.816±0.013	0.745±0.184	0.923±0.031	0.776±0.090	0.843±0.149	0.768±0.206
KNN	ECFP4	0.869±0.034	0.735±0.067	0.668±0.089	0.971±0.042	0.906±0.028	0.782±0.073	0.714±0.088	0.989±0.010
	MACCS	0.923±0.008	0.854±0.014	0.844±0.014	0.865±0.018	0.938±0.029	0.876±0.037	0.876±0.018	0.890±0.067
	RDKit	0.916±0.007	0.851±0.011	0.830±0.008	0.879±0.017	0.933±0.019	0.888±0.051	0.882±0.038	0.910±0.070
Stacking model						**0.994±0.008**	**0.961±0.024**	**0.957±0.033**	**0.967±0.024**

Table S2. Predicted volatile compounds from aroma plants with sleep-promoting activity.

Names	Structure	Plant sources	Prediction scores
Carvacrol		*Lycopus lucidus, Monarda fistulosa, Mosla chinensis, Nepeta annua, Ocimum sanctum, Origanum vulgare, Scutellaria baicalensis, Thymus amurensis, Thymus mongolicus, Cinnamomum glanduliferum, Hibiscus rosa-sinensis, Salvia sclarea, Origanum marjorana, Elsholtzia rugulosa, Ledum palustre, Centipeda minima, Citrus macroptera, Sauropus androgynus, Urtica dioica, Alpinia zerumbet, Boenninghausenia sessilicarpa, Cotinus coggygria, Seriphidium nitrosum, Solidago canadensis, Cynoglossum lanceolatum, Carthamus tinctorius, Cichorium intybus, Stellaria media, Melilotus albus, Anisochilus carnosus, Elsholtzia bodinieri, Hyssopus officinalis*	1.0
Safranal		*Nelumbo nucifera, Portulaca oleracea, Cucurbita pepo*	1.0
Vanillin		*Vanilla planifolia, Ceiba pentandra*	0.96
Methyl eugenol		*Cinnamomum bejolghota, Cinnamomum platyphyllum, Cinnamomum rigidissimum, Laurus nobilis, Lindera angustifolia, Myristica fragrans, Taxus chinensis, Agastache rugosus, Ocimum tenuiflorum, Aster tataricus L. f., Asarum cardiophyllum, Acorus illicioides, Myrtus communis, Piper betle, Cymbopogon goeringii, Actaea cimicifuga, Rosa banksiae, Rosa mairei, Rosa odorata, Clausena dunniana, Zanthoxylum dimorphophyllum, Populus alba, Salix matsudana, Illicium difengpi, Sparganium stoloniferum, Curcuma longa, Asarum heterotropoides, Asarum sieboldii, Catharanthus roseus, Lithospermum erythrorhizon,*	1.0

		Asarum caudigerellum, Asarum forbesii, Asarum inflatum, Asarum macranthum, Artemisia fauriei, Bidens tripartita, Lepidium apetalum, Vaccinium myrtillus, Oxytropis kansuensis	
Linalool		*Lavandula angustifolia, Lavandula latifolia, Mentha citrata, Mentha dahurica, Monarda didyma, Ocimum basilicum, Origanum vulgare, Perilla frutescens, Rosmarinus officinalis, Salvia japonica, Scutellaria baicalensis, Stachys palustris, Thymus amurensis, Thymus mongolicus, Akebia quinata, Cinnamomum appelianum, Cinnamomum bejolghota, Cinnamomum camphora, Cinnamomum glanduliferum, Cinnamomum iners, Cinnamomum jensenianum, Cinnamomum micranthum, Cinnamomum migao, Cinnamomum pauciflorum, Cinnamomum platyphyllum, Cinnamomum tenuipile, Cinnamomum wilsonii, Laurus nobilis, Lindera angustifolia, Lindera erythrocarpa, Lindera reflexa, Litsea cubeba*	1.0