ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

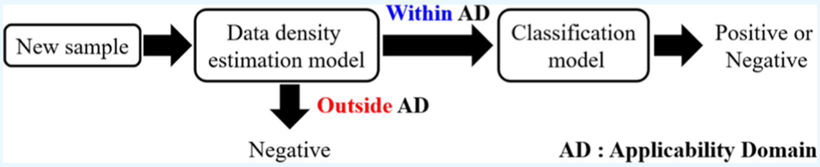
Molecular Design of Novel Herbicide and Insecticide Seed Compounds with Machine Learning

Yuki Nakayama, Saki Morishita, Hayato Doi, Tatsuya Hirano, and Hiromasa Kaneko*

Cite This: *ACS Omega* 2024, 9, 18488−18494

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Pesticides are widely used to improve crop productivity by eliminating weeds and pests. Conventional pesticide development involves synthesizing compounds, testing their activities, and studying their effects on the ecosystem. However, as pesticide discovery has an extremely low success rate, many compounds must be synthesized and tested. To overcome the high human, financial, and time costs of this process, machine learning is attracting increasing attention. In this study, we used machine learning for the molecular design of novel seed compounds for herbicides and insecticides. Classification models were constructed by using compounds that had been tested as herbicides and insecticides, and an inverse analysis of the constructed models was conducted. In the molecular design of herbicides, we proposed 186 new samples as herbicides using ensemble learning and a method for expressing explanatory variables that consider the relationships among eight weed species. For the molecular design of insecticides, we used undersampling and ensemble learning for the analysis of unbalanced data. Based on approximately 340,000 compounds, 12 potential insecticides were proposed, of which 2 exhibited actual activity when tested. These results demonstrate the potential of the developed machine-learning method for rapidly identifying novel herbicides and insecticides.

1. INTRODUCTION

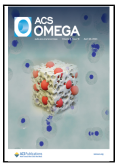
Pesticides are widely used to eliminate weeds and pests, thereby enhancing crop productivity. Traditionally, pesticides have been developed using an experimental approach in which various compounds are synthesized and tested, with a focus on their activity and effects on the ecosystem. Owing to its extremely low success rate, this process necessitates the synthesis and evaluation of numerous compounds. Computational science offers promising tools for accelerating the pesticide discovery process. For example, classification models can predict the presence or absence of activity based on chemical structures. Classification models are constructed using molecular descriptors or explanatory variables x, which are derived from chemical structures, and objective variables y, which represent the presence or absence of activity. Thus, activity can be predicted by inputting the x values of new chemical structures into the constructed model.[1−4] Machine-learning classification models based on multitasking models for quantitative structure−biological effect relationships (mtk-QSBERs) have made cutting-edge contributions to the field of molecular design.[5−10] In addition, the molecular design of potential pesticides has been performed using machine-learning classification models.[11−14]

Effective molecular design requires a model with high predictive ability, for which x is of particular importance. For example, if no information related to y is contained in x, a model with high predictive ability cannot be constructed regardless of the regression analysis or classification method used. In addition, any information contained in x that is unrelated to y will become noise in the constructed model, thereby decreasing the prediction performance for new data. Thus, constructing a model with high predictive ability necessitates the addition of information relating y to x or the extraction of information from the original x as well as the exclusion of information unrelated to y from x.

Furthermore, because the success rate of identifying active compounds during the development of new pesticides is extremely low, inactive compounds are much more common than active compounds. This imbalance in the data, where the negative class is much larger than the positive class, leads to prediction results that are biased toward the majority class. Consequently, almost all minority samples are predicted to be in the majority class.[15,16]

In this study, we used classification in machine learning for the molecular design of new seed compounds for herbicides and insecticides, which are the most frequently used pesticides.

Received: January 19, 2024
Revised: February 27, 2024
Accepted: March 27, 2024
Published: April 9, 2024



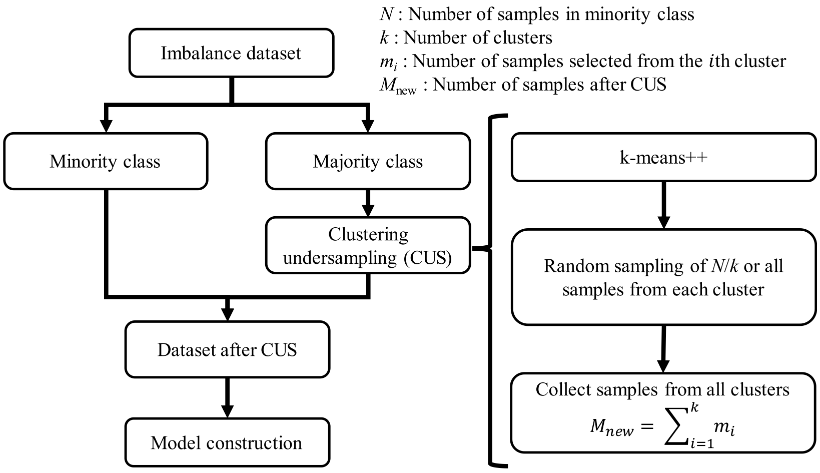

N : Number of samples in minority class
k : Number of clusters
m_i : Number of samples selected from the ith cluster
M_{new} : Number of samples after CUS
Imbalance dataset
Minority class
Majority class
Clustering undersampling (CUS)
Dataset after CUS
Model construction
k-means++
Random sampling of N/k or all samples from each cluster
Collect samples from all clusters
$M_{new} = \sum_{i=1}^{k} m_i$

Figure 1. Workflow for CUS. After the data set is divided into majority and minority classes, the majority samples are used for clustering with *k*-means++. Samples are randomly selected from each cluster divided using *k*-means++ and then integrated. The use of CUS to select the majority of samples results in an imbalance ratio approaching 1. Finally, the classification model is constructed.

The molecular design of new herbicides focused on four species (pre-emergence and early postemergence) and the use of various classification methods to develop an *x* for each *y*. For the molecular design of new insecticides, we proposed new seed compounds by constructing a classification model using undersampling and classification methods for approximately 20,000−30,000 unbalanced insecticidal activity test data points from four pest species.

2. METHODS

2.1. Data. A classification model of herbicidal activity was constructed by using bioassay results provided by Hokko Chemical Industry Co., Ltd. The assays targeted four weeds in rice paddy fields: *Echinochloa crus-galli* (E,c), *Lindernia procumbens* (L,p), *Monochoria vaginalis* (M,v), and *Schoenoplectus juncoides* (S,j). For each weed, the test compounds were applied pre-emergence (−) or early postemergence (+) at a dosage of 1200 g a.i./hectare. The herbicidal efficacy was determined by visual observation of the treated plants in comparison with the untreated plants 2 weeks after application. The herbicidal rating scores ranged from 0 to 100. A score of 0 indicates that the compound had no efficacy, whereas a score of 100 indicates that the weed was completely killed. Compounds scoring 90 points or more were considered positive, and those scoring less than 90 points were considered negative. The number of compounds analyzed for each weed is shown in Table S1.

To construct a classification model of insecticidal activity, we used bioassay results provided by Hokko Chemical Industry Co., Ltd. These assays targeted four pest species: cowpea aphid, *Aphis craccivora*; brown plant hopper, *Nilaparvata lugens*; common cutworm, *Spodoptera litura*; and two-spotted spider mite, *Tetranychus urticae*. The test compounds were diluted to 500 ppm and sprayed onto the test plants or leaf discs. Adult female *A. craccivora* and *T. urticae* were released on the plants or leaf discs 1 day before application. In contrast, second-instar larvae of *S. litura* and third-instar nymphs of *N. lugens* were released from the plants after spraying. The mortality count was performed 4−7 days after treatment. Compounds scoring mortality rates of 60% or more were considered positive, whereas those scoring mortality rates of less than 60% were considered negative. The number of compounds analyzed for each pest species is shown in Table S2.

2.2. Classification Models. For the molecular design of new herbicides, three classification methods were used: linear support vector machines using the linear kernel (LSVM),[17] nonlinear support vector machines using the Gaussian kernel (NLSVM),[17] and random forests (RF).[18] Owing to the small sample size, the predictive abilities of the classification models were evaluated using double cross-validation (DCV).[19] First, the samples were divided according to the number of outer folds. Then, one fold was used as the test data, and the remaining folds were used as training data, which were again divided by the number of inner folds for cross-validation to optimize the hyperparameters. This operation was repeated until all of the folds were used as test data. The predictive ability of each classification model was determined by comparing the actual and predicted classes in the outer cross-validation.

For the molecular design of new insecticides, we used two classification methods: extreme gradient boosting (XGB)[20] and light gradient boosting machine (LGBM).[21] The samples were divided into training and test data. The training data were used to construct the classification models, whereas the test data were used to validate the models. The predictive ability of each model was evaluated by comparing the actual and predicted classes of test data.

Several indices, including accuracy, recall, precision, and *F*-score, were used to evaluate the predictive ability of the classification models quantitatively. For each of these indices, a value closer to 1 indicates a better performance of the classification model.

2.3. Explanatory Variables in the Molecular Design of Herbicides. For the molecular design of new herbicides, the classification focused on the representation of the *x* variables in an herbicide data set. In this study, models were constructed and evaluated using four methods for *x*: methods 1, 2, 3, and 4.

Method 1 calculated molecular descriptors from the chemical structures of the compounds as *x* to construct a classification model for each weed species. As eight weed species were investigated in this study, eight classification models were constructed.

Table 1. Evaluation Indexes in DCV for Method 1

weed species	descriptor	method	accuracy rate	recall	precision	*F*-score
E,c(−)	alvaDesc + MACCS	RF	0.837	0.929	0.813	0.867
E,c(+)	alvaDesc	NLSVM	0.776	0.926	0.735	0.820
L,p(−)	MACCS	LSVM	0.755	0.143	1.000	0.250
L,p(+)	MACCS	LSVM	0.918	0.000	0.000	0.000
M,v(−)	RDKit	RF	0.548	0.250	0.364	0.296
M,v(+)	MACCS	RF	0.643	0.385	0.417	0.400
S,j(−)	alvaDesc	RF	0.633	0.526	0.526	0.526
S,j(+)	RDKit	RF	0.878	0.500	0.667	0.571

In method 2, a single classification model was constructed by using compounds for all eight weed species to determine the relationships among them. Because descriptors calculated from the chemical structures of the compounds alone cannot represent differences among weed species, dummy variables that convert categorical data into numerical data of 0 and 1 were introduced to represent these relationships. In this study, eight dummy variables for the eight weed species were added to x.

Method 3 considered the differences between the weed species in terms of transfer learning. A matrix of descriptors for each weed species was connected vertically to a matrix of zero matrices for each weed species, except for the target weed species, which was connected horizontally to form a data set of x. In this representation, x is 9× (the number of descriptors).

Method 4 used ensemble learning, in which the majority vote of the three estimation results obtained from methods 1, 2, and 3 was used as the final estimated class. For ensemble learning, the results of the model with the highest predictive ability among the combinations of descriptor sets and classification methods were used. The predictive abilities of the models were evaluated based on the accuracy rate and recall of new data.

2.4. Molecular Design of Insecticides.

For the molecular design of new insecticide seed compounds, we used unbalanced data, in which the number of compounds that were classified as ineffective (negative) was very high compared to the number of compounds that were classified as effective (positive). In this study, we focused on undersampling. In addition to random undersampling (RUS), which randomly selects as many samples from the majority data as from the minority data, we applied clustering undersampling (CUS), which utilizes k-means++ clustering.[22]

The basic concept of CUS is illustrated in Figure 1. First, the samples are divided into majority and minority classes, and only the former are used for clustering using k-means++. Next, samples are randomly selected from each cluster and divided using k-means++. The number of samples to be selected (n) is given as

$$n = \frac{N}{k} \tag{1}$$

where N is the number of minority samples and k is the number of clusters. When a cluster does not have n samples, all of the samples in the cluster are selected. The number of samples selected from the ith cluster m_i is expressed as

$$1 < m_i \leq n \tag{2}$$

Finally, the selected samples from all of the clusters are integrated. The number of samples after CUS (M_{new}) is given by

$$M_{\mathrm{new}} = \sum_{i=1}^{k} m_i \tag{3}$$

By selecting the majority of samples through CUS, the imbalance ratio approaches 1, and the classification model is then constructed.

Furthermore, we combined undersampling and ensemble learning to improve the predictive ability of the classification model. Owing to the randomness of CUS, undersampling was repeated to obtain subdata sets, and a submodel was constructed for each subdata set. The prediction results of all of the submodels were combined to produce the final prediction result. In this study, 100 subdata sets were used. To reduce the number of samples that gave false positives, only samples that were predicted to be positive by all of the submodels were classified as positive.

To exclude samples with low confidence for positive predictions, we proposed a method that combines the applicability domain (AD) and classification. Before classifying a sample, the AD was set via the k-nearest neighbor (k-NN) algorithm[23] using only samples whose actual class was positive. Compounds that were newly estimated and outside the AD were considered unreliable, even when they were estimated as positive using the classification model; therefore, their results were negative. In contrast, compounds within the AD were classified, and the estimated class was used as the result. This method allows the reliability of the samples estimated as positive to be considered during classification.

2.5. Inverse Analysis of the Classification Model.

Inputting the values of the descriptors calculated from the chemical structures of compounds that had not been tested into the constructed classification model allowed estimation of the presence or absence of activity without conducting tests. By using these estimation results to determine which compounds should be tested, promising compounds were identified using a small number of tests.

The new herbicide candidates consisted of 165 compounds from the Hokko Chemical Industry. The number of samples used to estimate the presence or absence of activity for each weed species was 1320 (165 × 8). The classification model with the highest recall among methods 1, 2, 3, and 4 in Section 2.2 was used as the model for the inverse analysis. For the inverse analysis of the model from method 4, the AD set with k-NN was used for each method, and samples that were within the AD for all methods were considered to be within the AD for ensemble learning.

As new insecticide candidates, 499,724 compounds were obtained from the Namiki Shoji Co., Ltd. database.[24] Compounds containing salts or metal elements and those that had already been tested for activity were excluded, and an inverse analysis was performed using the remaining com-

Table 2. Evaluation Indexes in DCV for Methods 2 and 3

	descriptor	method	accuracy rate	recall	precision	*F*-score
method 2	RDKit + MACCS	RF	0.749	0.605	0.639	0.622
method 3	MACCS	RF	0.709	0.581	0.573	0.577

Table 3. Best Classification Models and Evaluation Indexes in Ensemble Learning

	based on accuracy rate				based on recall			
	accuracy rate	recall	precision	*F*-score	accuracy rate	recall	precision	*F*-score
method 1	0.765	0.550	0.696	0.615	0.720	0.612	0.585	0.598
method 2	0.749	0.605	0.639	0.622	0.749	0.605	0.639	0.622
method 3	0.709	0.581	0.573	0.577	0.688	0.612	0.537	0.572
method 4	0.746	0.581	0.641	0.610	0.743	0.620	0.625	0.623

pounds. This inverse analysis used the classification model with the largest *F*-value and the AD set with a *k*-NN.

3. RESULTS AND DISCUSSION

3.1. Molecular Design of Herbicides. For the molecular design of herbicides, five descriptor sets were used: RDKit, alvaDesc, and MACCS keys; RDKit + MACCS keys; and alvaDesc + MACCS keys. Three methods (RF, LSVM, and NLSVM) were used for classification, and DCV was conducted to validate the models using 5-fold cross-validation for outer cross-validation and leave-one-out for inner cross-validation.

In method 1, a classification model was constructed for each weed species to estimate the presence or absence of herbicidal activity. Table 1 shows the evaluation indices of the most predictive classification method for each weed species, and Table S3 shows the representative confusion matrix. Although the models for E,c(−) and E,c(+) could estimate classes with high accuracy, a recall greater than zero could not be obtained for L,p(+), likely because the number of positive compounds was much lower than the number of negative compounds.

Dummy variables were introduced into *x* in method 2, and variables combining zero matrices and descriptors were used as *x* in method 3 to construct a classification model that considered the relationship among the weed species. The presence or absence of herbicidal activity for the eight weed species was estimated by using each model. The evaluation indices of the most predictive classifiers in DCV are listed in Table 2, and the representative confusion matrices are shown in Table S4. Both introducing dummy variables and considering the relationships between weed species reduced the number of false negatives. In addition, some samples that gave false positives using method 1 were correctly classified using method 2. However, samples that were correctly classified using method 1 were misclassified using method 2. Although the number of false negatives with method 3 was similar to that with method 1, the number of false positives was higher. However, some samples that were false positives with method 1 were true positives with method 3, suggesting that method 3 is a valid representation method that can consider the relationships among weed species. Thus, samples that were misclassified by method 1 but correctly estimated using methods 2 and 3 can be identified. Similarly, the samples misclassified by method 2 were correctly estimated using methods 1 and 3. These results suggest that accuracy could be improved by employing ensemble learning.

Method 4 used the class obtained by determining the best model in methods 1, 2, and 3 based on the majority rule. Both an accuracy-rate-based approach and a recall-based approach were considered to select the best classification model for methods 1, 2, and 3. Table 3 shows the evaluation indices for each best model, and Figure 2 shows a histogram of the percentages predicted as positive using method 4. Ensemble learning with the best model selected based on the accuracy rate did not exceed the positive percentage obtained using method 1, whereas ensemble learning with the best model selected based on recall achieved improved recall. Furthermore, focusing on the samples in Figure 2 whose actual class was positive, the proportion of samples estimated to be positive at 0.67 was greater than that at 0.33. Similarly, for samples whose actual class was negative, classification by ensemble learning was also effective, with the proportion of samples estimated as negative at 0.33 exceeding that at 0.67.

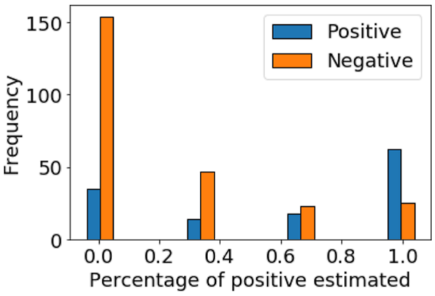


Figure 2. Histogram of the percentage of positive predictions using method 4.

An inverse analysis of the model was performed using a compound data set from the Hokko Chemical Industry. In the inverse analysis, the presence or absence of herbicidal activity was estimated, and herbicidal activity tests were prioritized on the basis of these results. A total of 165 compounds were subjected to inverse analysis, and 1320 (165 × 8) samples were used to estimate the presence or absence of activity for each weed species. Method 4, which had the highest recall among methods 1−4, was selected as the best model, and the inverse analysis was performed. The AD was set using *k*-NN with $k = 5$ and $\alpha = 0.80$. Any compounds that were located outside the AD, even in one model, were considered to be outside the AD. Table 4 summarizes the prediction results of the inverse analysis. At least one compound was predicted to be positive for all of the weed species. A histogram of the number of compounds predicted to be positive for each weed species is shown in Figure 3. Although the number of compounds predicted to be positive decreases as the number of weed species increases, some compounds are predicted to be positive for all eight weed species.

Table 4. Prediction Results for the Inverse Analysis of Herbicides

	positive	negative	outside AD
E,c(−)	73	70	22
E,c(+)	84	59	22
L,p(−)	23	120	22
L,p(+)	1	157	7
M,v(−)	42	88	35
M,v(+)	49	108	8
S,j(−)	28	115	22
S,j(+)	3	155	7

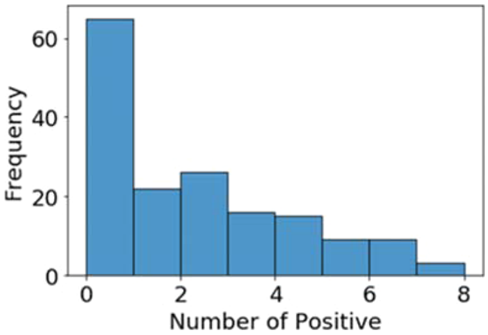


Figure 3. Histogram of the number of compounds estimated to be positive for 8 weed species.

Subsequently, activity tests were performed for the 177 samples that were predicted to be positive. In addition, 339 samples that were predicted to be negative were tested to validate the model. Table 5 presents the confusion matrix for the results of the inverse analysis; the accuracy rate, recall, precision, and *F*-score were 0.680, 0.559, 0.532, and 0.545, respectively. Positive test results were obtained for 99 of the 177 samples predicted to be positive and 78 of the 252 samples predicted to be negative, and 186 active compounds were proposed as herbicide seed compounds. Compared with the predictive performance of the model with DCV on the existing data set, that of the inverse analysis was not significantly decreased. Thus, the model constructed in this study could predict the presence or absence of herbicidal activity with high accuracy on an external data set. Among the proposed compounds, a few were predicted to be positive for all of the weed species. However, the activity test results were positive for only five of the eight weed species. Although three of the weed species were incorrectly proposed, these compounds showed activity for various targets.

Table 5. Confusion Matrix for the Inverse Analysis of Herbicides

	positive (prediction)	negative (prediction)
positive (activity test)	99	87
negative (activity test)	78	252

3.2. Molecular Design of Insecticides.

For this analysis, the data set of insecticide compounds was unbalanced, with imbalance ratios in the range of 5.34−11.75 (Table S5). In addition, the number of samples was greater than 20,000 for all pest species (Table S2). Therefore, undersampling by RUS and CUS was used, and the imbalance rate after undersampling for each pest was approximately 1.

For the molecular design of new insecticide seed compounds, eight combinations of descriptor sets and processing methods were used, as shown in Table 6. In addition, two classification methods were used: XGB and LGBM. The number of submodels used for ensemble training was 100, and α was set to 0.8 and k to 10 for k-NN. The samples were divided in a 7:3 ratio between the training and test data.

Table 6. Descriptor and Processing Methods for the Molecular Design of Insecticides

descriptor	processing method
alvaDesc	no processing
RDKit	*k*-NN
MACCS keys fingerprint	RUS
ECFP4	RUS + *k*-NN
alva + MACCS	RUS + ensemble learning
RDKit + MACCS	RUS + ensemble learning + *k*-NN
alva + ECFP4	CUS + ensemble learning
RDKit + ECFP4	CUS + ensemble learning + *k*-NN

The evaluation indices for each pest species and treatment are shown in Table 7, and the confusion matrices are shown in Tables S6−S9. The RUS and CUS resulted in fewer false negatives than the other treatments. Without treatment, the number of false negatives increased, likely due to the unbalanced training data used in the analysis. In contrast, with RUS + ensemble learning, the number of false positives was lower than that when only RUS was used.

Figure 4 shows a histogram of the percentage of samples estimated as positive when the actual class was negative using the *A. craccivora* results as an example. In the RUS + ensemble learning method used in this study, only samples for which all 100 submodels were estimated to be positive were considered positive. For example, in the samples at approximately 0.20 in Figure 4, approximately 20% of the submodels were estimated to be positive, but the results were negative. In particular, a large number of the submodels were estimated to be positive for up to approximately 30% of the samples, likely because many samples were randomly estimated to be positive in certain submodels. The ability to correctly estimate such samples as negative is an advantage of incorporating ensemble learning. However, because the result was not considered positive unless all submodels were estimated to be positive, the number of false negatives was higher for the model using RUS + ensemble learning than for that using only RUS. However, the *F*-score was higher for RUS + ensemble learning than for RUS for all of the pests, confirming that ensemble learning effectively reduced the number of false positives rather than the number of false negatives.

For all of the pests, the AD was set for the positive samples, which reduced the number of false-positive samples and improved the predictive ability of the classification model. However, the number of false negatives increased, and the recall decreased because samples whose actual class was positive were present outside the AD.

Inverse analysis was performed using approximately 340,000 compounds from the Namiki Shoji Co., Ltd. database.[24] The model with the largest F-score was used for inverse analysis. The AD was set using k-NN with α = 0.8 and k = 10. Table 8 presents the estimated results of the inverse analysis. Among the compounds that were estimated to be positive, 12 were

Table 7. Evaluation Indices of the Best Class Classification Results for Each Process

insect name	processing method	method	descriptor	accuracy rate	recall rate	precision rate	*F*-score
Aphis craccivora	no processing	XGB	alva + ECFP4	0.875	0.403	0.685	0.508
	k-NN	XGB	alva + ECFP4	0.876	0.375	0.711	0.491
	RUS	LGBM	RDKit + ECFP4	0.749	0.704	0.356	0.473
	RUS + *k*-NN	LGBM	ECFP4	0.787	0.591	0.391	0.470
	RUS + ensemble	LGBM	ECFP4	0.871	0.439	0.643	0.522
	RUS + ensemble + *k*-NN	LGBM	ECFP4	0.871	0.413	0.652	0.505
	CUS + ensemble	LGBM	alva + ECFP4	0.864	0.457	0.599	0.519
	CUS + ensemble + *k*-NN	LGBM	alva + ECFP4	0.864	0.431	0.607	0.504
Nilaparvata lugens	no processing	XGB	alva + ECFP4	0.899	0.472	0.734	0.575
	k-NN	XGB	alva + ECFP4	0.896	0.436	0.736	0.548
	RUS	LGBM	alva + ECFP4	0.802	0.784	0.406	0.535
	RUS + *k*-NN	LGBM	alva + ECFP4	0.835	0.673	0.453	0.541
	RUS + ensemble	LGBM	alvaDesc	0.896	0.504	0.694	0.584
	RUS + ensemble + *k*-NN	LGBM	RDKit + MACCS	0.891	0.475	0.678	0.559
	CUS + ensemble	LGBM	alvaDesc	0.889	0.544	0.636	0.586
	CUS + ensemble + *k*-NN	LGBM	alvaDesc	0.887	0.506	0.641	0.566
Spodoptera litura	no processing	XGB	alva + MACCS	0.940	0.384	0.747	0.507
	k-NN	XGB	alva + MACCS	0.939	0.357	0.770	0.488
	RUS	LGBM	alva + MACCS	0.816	0.751	0.270	0.397
	RUS + *k*-NN	LGBM	alva + ECFP4	0.867	0.638	0.332	0.436
	RUS + ensemble	LGBM	alvaDesc	0.939	0.474	0.671	0.555
	RUS + ensemble + *k*-NN	LGBM	alva + ECFP4	0.938	0.447	0.681	0.540
	CUS + ensemble	LGBM	alvaDesc	0.930	0.532	0.569	0.550
	CUS + ensemble + *k*-NN	LGBM	alvaDesc	0.930	0.490	0.575	0.529
Tetranychus urticae	no processing	XGB	alva + MACCS	0.920	0.591	0.825	0.689
	k-NN	XGB	RDKit + ECFP4	0.918	0.547	0.847	0.664
	RUS	LGBM	alvaDesc	0.866	0.774	0.535	0.633
	RUS + *k*-NN	LGBM	RDKit + ECFP4	0.886	0.685	0.603	0.641
	RUS + ensemble	LGBM	RDKit + ECFP4	0.919	0.603	0.804	0.689
	RUS + ensemble + *k*-NN	LGBM	RDKit + ECFP4	0.916	0.572	0.810	0.670
	CUS + ensemble	LGBM	RDKit + ECFP4	0.913	0.638	0.743	0.686
	CUS + ensemble + *k*-NN	LGBM	RDKit + ECFP4	0.913	0.594	0.770	0.670

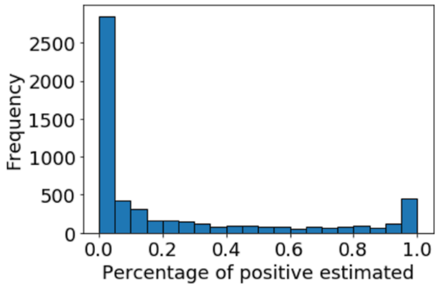


Figure 4. Histogram of the percentage of samples estimated as positive when the actual class is negative.

Table 8. Estimated Results for the Inverse Analysis of the Molecular Design of Insecticides

	positive	negative	outside AD
Aphis craccivora	448	91,417	246,872
Nilaparvata lugens	4374	75,210	259,182
Spodoptera litura	51	81,850	256,830
Tetranychus urticae	47	82,361	256,931

tested for insecticidal activity. Notably, one of these compounds was active against *N. lugens* and another was active against *S. litura*. Thus, we successfully developed new insecticide seed compounds using a classification model with undersampling and ensemble learning.

4. CONCLUSIONS

In this study, machine learning was employed for the molecular design of novel pesticides. For the molecular design of new herbicides, we targeted eight weed species and estimated the presence or absence of activity for each weed species. Two methods (incorporating dummy variables and combining molecular descriptors and zero matrices) were used to correctly classify samples that had been misclassified by the classifiers estimated for each weed species. Ensemble learning of these estimation results further improved the predictive ability of the model, and inverse analysis using the model with ensemble learning revealed 186 potential new herbicides.

For the molecular design of new insecticides, the predictive ability of the model was improved by applying an undersampling method to unbalanced data sets. Inverse analysis of the model allowed us to predict the active compounds. Subsequent experimental validation of the proposed compounds revealed two new insecticide seed compounds.

Physicochemical or structural interpretation of the machine-learning classification model can be performed based on feature importance, such as that revealed by cross-validated permutations, and local interpretation, including the local slopes of model predictions. Thus, the proposed method is expected to accelerate the molecular design of novel pesticides.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c00655.

Number of compounds analyzed, confusion matrices for the molecular design of herbicides, imbalance ratios for the molecular design of insecticides, and confusion matrices for the molecular design of insecticides (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Hiromasa Kaneko − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan;* orcid.org/0000-0001-8367-6476; Phone: +81-44-934-7197; Email: hkaneko@meiji.ac.jp

Authors

Yuki Nakayama − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan*

Saki Morishita − *Hokko Chemical Industry Co., Ltd., Atsugi-shi, Kanagawa 243-0023, Japan*

Hayato Doi − *Hokko Chemical Industry Co., Ltd., Atsugi-shi, Kanagawa 243-0023, Japan*

Tatsuya Hirano − *Hokko Chemical Industry Co., Ltd., Atsugi-shi, Kanagawa 243-0023, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c00655

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Fan, T.; Sun, G.; Zhao, L.; Cui, X.; Zhong, R. QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds. *Int. J. Mol. Sci.* **2018**, *19*, 3015.

(2) Chiddarwar, R. K.; Rohrer, S. G.; Wolf, A.; Tresch, S.; Wollenhaupt, S.; Bender, A. *In Silico* Target Prediction for Elucidating the Mode of Action of Herbicides Including Prospective Validation. *J. Mol. Graphics Modell.* **2017**, *71*, 70−79.

(3) Hammann, F.; Schöning, V.; Drewe, J. Prediction of Clinically Relevant Drug-Induced Liver Injury from Structure Using Machine Learning. *J. Appl. Toxicol.* **2019**, *39*, 412−419.

(4) Yang, L.; Sang, C.; Wang, Y.; Liu, W.; Hao, W.; Chang, J.; Li, J. Development of QSAR Models for Evaluating Pesticide Toxicity against *Skeletonema costatum*. *Chemosphere* **2021**, *285*, 131456.

(5) Speck-Planche, A.; Cordeiro, M. N. D. S. De Novo Computational Design of Compounds Virtually Displaying Potent Antibacterial Activity and Desirable in Vitro ADMET Profiles. *Med. Chem. Res.* **2017**, *26*, 2345−2356.

(6) Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb. Sci.* **2017**, *19*, 501−512.

(7) Kleandrova, V. V.; Speck-Planche, A. Multitasking Model for Computer-Aided Design and Virtual Screening of Compounds with High Anti-HIV Activity and Desirable ADMET Properties. In *Multi-Scale Approaches in Drug Discovery*; Speck Planche, A., Ed.; Elsevier, 2017, pp 55−81..

(8) Speck-Planche, A.; Kleandrova, V. V. Multi-Condition QSAR Model for the Virtual Design of Chemicals with Dual Pan-Antiviral and Anti-Cytokine Storm Profiles. *ACS Omega* **2022**, *7*, 32119−32130.

(9) Kleandrova, V. V.; Speck-Planche, A. PTML Modeling for Pancreatic Cancer Research: In Silico Design of Simultaneous Multi-Protein and Multi-Cell Inhibitors. *Biomedicines* **2022**, *10*, 491.

(10) Speck-Planche, A.; Kleandrova, V. V. Demystifying Artificial Neural Networks as Generators of New Chemical Knowledge: Antimalarial Drug Discovery as a Case Study. In *Machine Learning in Chemistry: The Impact of Artificial Intelligence*; Cartwright, H. M., Ed.; *Theoretical and Computational Chemistry Series*; The Royal Society of Chemistry, 2020; Vol. *17*, pp 398−423.

(11) Speck-Planche, A.; Natalia Dias Soeiro Cordeiro, M.; Guilarte-Montero, L.; Yera-Bueno, R. Current Computational Approaches towards the Rational Design of New Insecticidal Agents. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 304−314.

(12) Speck-Planche, A.; Guilarte-Montero, L.; Yera-Bueno, R.; Rojas-Vargas, J. A.; García-López, A.; Uriarte, E.; Molina-Pérez, E. Rational Design of New Agrochemical Fungicides Using Substructural Descriptors. *Pest Manage. Sci.* **2011**, *67*, 438−445.

(13) Speck-Planche, A.; Kleandrova, V. V.; Rojas-Vargas, J. A. QSAR Model toward the Rational Design of New Agrochemical Fungicides with a Defined Resistance Risk Using Substructural Descriptors. *Mol. Diversity* **2011**, *15*, 901−909.

(14) Speck-Planche, A.; Kleandrova, V. V.; Scotti, M. T. Fragment-Based Approach for the in Silico Discovery of Multi-Target Insecticides. *Chemom. Intell. Lab. Syst.* **2012**, *111*, 39−45.

(15) Zhu, H.; Liu, G.; Zhou, M.; Xie, Y.; Kang, Q. A Noisy-Sample-Removed Under-Sampling Scheme for Imbalanced Classification of Public Datasets. *IFAC-PapersOnLine* **2020**, *53*, 624−629.

(16) Kumari, P.; Nath, A.; Chaube, R. Identification of Human Drug Targets Using Machine-Learning Algorithms. *Comput. Biol. Med.* **2015**, *56*, 175−181.

(17) Vapnik, V. Pattern Recognition Using Generalized Portrait Method. *Autom. Remote Control* **1963**, *24*, 774−780.

(18) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(19) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated Double Cross Validation. *J. Chemom.* **2009**, *23*, 160−171.

(20) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA*, 2016; pp 785..

(21) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017.

(22) Arthur, D.; Vassilvitskii, S. V. k-means++: The Advantages of Careful Seeding. *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA*, 2007; pp 1027−1035.

(23) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21−27.

(24) Namiki Shoji Co., Ltd. https://www.namiki-s.co.jp/ (accessed 2023-01-19).