



Transfer learning for predicting human skin sensitizers

Chun-Wei Tung^{1,2} · Yi-Hui Lin³ · Shan-Shan Wang³

Received: 28 September 2018 / Accepted: 21 February 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Computational prioritization of chemicals for potential skin sensitization risks plays essential roles in the risk assessment of environmental chemicals and drug development. Given the huge number of chemicals for testing, computational methods enable the fast identification of high-risk chemicals for experimental validation and design of safer alternatives. However, the development of robust prediction model requires a large dataset of tested chemicals that is usually not available for most toxicological endpoints, especially for human data. A small training dataset makes the development of effective models difficult with insufficient coverage and accuracy. In this study, an ensemble tree-based multitask learning method was developed incorporating three relevant tasks in the well-defined adverse outcome pathway (AOP) of skin sensitization to transfer shared knowledge to the major task of human sensitizers. The results show both largely improved coverage and accuracy compared with three state-of-the-art methods. A user-friendly prediction server was available at <https://cwtung.kmu.edu.tw/skinsensdb/predict>. As AOPs for various toxicity endpoints are being actively developed, the proposed method can be applied to develop prediction models for other endpoints.

Keywords Adverse outcome pathway · Allergic contact dermatitis · Alternative method · Multitask learning · Skin sensitization · ExtraTrees

Introduction

Allergic contact dermatitis is the second most common occupational illness accounting for 10–15% of all occupational disease worldwide (Basketter et al. 2015) that can be caused by exposure to skin sensitizers (Karlberg et al. 2008). The identification of hazardous chemicals and design of safer alternatives are essential tasks to improve human health. While animal testing remains the gold standard for

the preclinical assessment of skin sensitization, its results are not in full concordance with human responses (Roberts and Api 2018). In addition, the ban of animal testing on cosmetic ingredients by European Union stressed out the need of alternative methods with regard to human health concerns.

In recent years, the paradigm of skin sensitization assessment is shifting from traditional animal testing to adverse outcome pathway (AOP)-based integrated testing strategies (ITSs). The ITSs utilize simple weighting methods or complex machine learning methods to combine experimental results from in chemico and in vitro assays corresponding to key events of the skin sensitization AOP giving better predictive performance on human skin sensitizers over animal testing (Bauch et al. 2012; Jaworska and Hoffmann 2010; Jaworska et al. 2015; Roberts and Patlewicz 2018; Urbisch et al. 2015). However, given the vast number of chemicals for testing in drug development and risk assessment, computational methods such as machine-learning techniques could be better alternatives for fast prioritizing potential risks than resource- and time-consuming ITSs.

Structural alerts (SAs) representing the knowledge of functional groups with peptide reactivity were developed as

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00204-019-02420-x>) contains supplementary material, which is available to authorized users.

✉ Chun-Wei Tung
cwtung@livemail.tw

¹ Present Address: Graduate Institute of Data Science, College of Management, Taipei Medical University, 172-1, Sec. 2, Keelung Rd., Taipei 10675, Taiwan

² National Institute of Environmental Health Sciences, National Health Research Institutes, 35 Keyan Rd., Zhunan, Miaoli County 35053, Taiwan

³ School of Pharmacy, Kaohsiung Medical University, 100 Shihchuan 1st Rd., Kaohsiung 80708, Taiwan

softwares for predicting skin sensitizers (OECD 2018; Patlewicz et al. 2008; Sushko et al. 2012). Although SAs give useful mechanism information for peptide reactivity, their prediction performance was found to be worse than quantitative structure–activity relationship (QSAR) models (Alves et al. 2016a). QSAR models are statistics- or machine learning-based models constructed by identifying discriminative features and representations of chemicals from available experimental data. The size of training data is therefore critical to the prediction performance. Due to the lack of human data, previous methods mainly focused on the prediction of rodent data (Golla et al. 2009; Li et al. 2007a, b; Miller et al. 2005; Patlewicz et al. 2004; Sarath Kumar et al. 2016; Yuan et al. 2009). Currently available computational methods for predicting human skin sensitizers suffered from insufficient data yield a low coverage and moderate accuracy (Alves et al. 2016b; Braga et al. 2017; Golla et al. 2009). It is desirable to develop novel algorithms capable of generating effective prediction models based on a small data size that is a common issue for many computational works.

In this work, we present the first transfer learning-based method leveraging three relevant learning tasks for predicting human skin sensitizers with high prediction performance and coverage. The three relevant tasks correspond to the three key events of the well-defined AOP of skin sensitization (Ezendam et al. 2016; Kleinstreuer et al. 2018) including protein binding, keratinocyte activation and dendritic cell activation (OECD Testing Guidance 442C, 442D and 442E). In contrast to traditional machine learning algorithms that learn individual tasks separately, the proposed method utilized a tree-based ensemble multitask learning algorithm (Simm et al. 2014) that can deal with multiple tasks simultaneously and are able to uncover shared knowledge for knowledge transfer. The useful knowledge of SAs was also incorporated into the proposed method to improve prediction performance.

As the three AOP events are highly relevant to the major task, multitask learning boosts prediction performance with a high independent test accuracy of 0.909 for chemicals within the defined applicability domain (AD). A comparison to existing state-of-the-art methods shows the superior performance and coverage of the proposed method. Finally, a user-friendly web server named SkinSensPred has been implemented that is freely accessible at <https://cwtung.kmu.edu.tw/predict>. SkinSensPred is expected to be useful for the prioritization and identification of human skin sensitizers. The proposed multitask methods could be further applied to predict various toxicities with a well-defined AOP.

Materials and methods

Datasets

For developing prediction models for human skin sensitizers, a total of 203, 162, 166 and 121 organic chemicals with peptide reactivity, keratinocyte activation, dendritic cell activation and human skin sensitization data were extracted from SkinSensDB (<https://cwtung.kmu.edu.tw/skinsensdb>, version 2018-02-05) (Wang et al. 2017). Human skin sensitization data is the major learning task, while the other three key events of adverse outcome pathway (AOP) represent relevant tasks for knowledge transfer. As there are several inconsistent results for each AOP event curated from different studies, the final outcome of activation or inactivation for each key event is determined by a majority vote. In case of a tie, the corresponding assay results were excluded. To avoid over-estimation of prediction performance, the data were randomly divided into training datasets and independent test datasets in a 2:1 ratio. The training datasets were utilized to develop algorithms and train prediction models without the involvement of the independent test datasets. After the completion of prediction model, the independent test sets were utilized to test the model performance.

Development of prediction models

To develop prediction models, the samples were firstly converted into two kinds of feature vectors of ECFP6 (extended connectivity fingerprint with a diameter of 6) and SAs (structural alerts). The ECFP6 feature consists of 1024 bits of ECFP6 fingerprint generated by using RCDK package (Guha 2007) which provides an interface of Chemistry Development Kit (CDK) (Willighagen et al. 2017) to R computing environment. The SAs features were generated by using ToxTree software (Patlewicz et al. 2008) consisting of 5 SAs representing alerts for acyl transfer agent, Michael acceptor, SN2, SNAr and Schiff base. Due to the high sparsity of the 5 SAs features that might lead to overfitting problems, the 5 SAs features were combined into a one-bit feature representing the occurrence of any of the 5 SAs. Feature vectors of ECFP6 (1024 bits) along and combination of ECFP6 and SAs (1025 bits) were then utilized to develop prediction models and determine the best feature encoding.

The development of single-task and multitask prediction models is based on binary tree-based ensemble learning algorithms of ExtraTrees (Geurts et al. 2006) and MT-ExtraTrees (Simm et al. 2014), respectively. ExtraTrees represents extremely randomized trees that each binary

tree is built based on randomly selected f features and s splitting values. MT-ExtraTrees is an extension to ExtraTrees by introducing an additional task-specific splitting criterion to divide samples into two branches consisting of samples of separated task subsets. The conventional feature-wise splitting criterion represents shared knowledge among learning tasks, while the task-wise splitting criterion represents task-specific knowledge. A default parameter of f/a was utilized as the probability to apply the task-wise splitting criterion instead of the feature-wise criterion, where a represents the number of all available features. In this study, the parameters of $f = \text{sqrt}(a)$ and $s = 2$ were applied to construct single-task and multitask models with 200 trees.

Performance measurement

To evaluate the prediction performance of the models, five commonly used threshold-dependent measurements were utilized including accuracy, sensitivity, specificity, precision, correct classification rate (CCR) and Matthews correlation coefficient (MCC) defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4)$$

$$\text{CCR} = \frac{\text{sensitivity} + \text{specificity}}{2}, \quad (5)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}, \quad (6)$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. For imbalanced dataset, both MCC and CCR are considered to be more appropriate for evaluating and comparing model performance. In addition, a non-parametric measurement of the area under the receiver operating characteristic curve (AUC) useful for model comparison was also evaluated. The receiver operating characteristic curve is a plot of sensitivity versus 1-specificity showing the trade-off between sensitivity and specificity for each decision threshold. The AUC has also been shown to be useful for model evaluation and comparison for both balanced and

imbalanced datasets (Huang and Tung 2017; Jin and Ling 2005; Tung 2013; Tung and Ho 2008).

Implementation of SkinSensPred prediction server

The prediction system and user interface were implemented using HTML, JavaScript, JAVA and R language as well as AngularJS framework on an Ubuntu operation system. Users can either input SMILES of chemicals or draw a chemical structure for prediction. For each input chemical structure, the R package of RCDK (Guha 2007) and ToxTree (Patlewicz et al. 2008) software will be utilized to generate its corresponding feature vector. The prediction model based on extraTrees (Simm et al. 2014) package will be applied to predict the skin sensitization potential with a probability score. A score value close to 1 indicates a strong probability of a human skin sensitizer. In contrast, a score value close to 0 indicates a strong probability of a non-sensitizer. An additional column of applicability domain (AD) gives information of whether the test chemical falls in the defined AD. The information of five SAs will also be shown in the result table to facilitate the interpretation of prediction results. All columns are sortable by clicking on corresponding column heads.

Results

Development of multitask prediction models

To develop the prediction method, curated chemical data and associated activation data of the four tasks were extracted from a curated SkinSensDB (Wang et al. 2017) database. In brief, chemicals were randomly divided into training and independent test datasets in a ratio of 2:1. Note that there are 121 chemicals in the major task with available human data. The extensively utilized extended connectivity fingerprint with a diameter of 6 (ECFP6) (Rogers and Hahn 2010) were applied to encode feature vectors. ECFP6 features represent the presence and absence of particular substructures whose occurrence patterns can be analyzed for discriminating skin sensitizers from non-sensitizers. In this study, ECFP6 features were mapped into a fixed-length of 1024-bit feature vector through a hashing procedure to reduce the required computation resource. As structure alerts (SAs) for peptide reactivity are widely utilized for identifying skin sensitizers, we also developed a novel combined feature vector integrating ECFP6 with SAs. As our purpose is to prioritize risk potentials for further validation, the measurement of prediction performance is mainly based on the area under the receiver operating characteristic curve (AUC) representing a trade-off between sensitivity and specificity for various

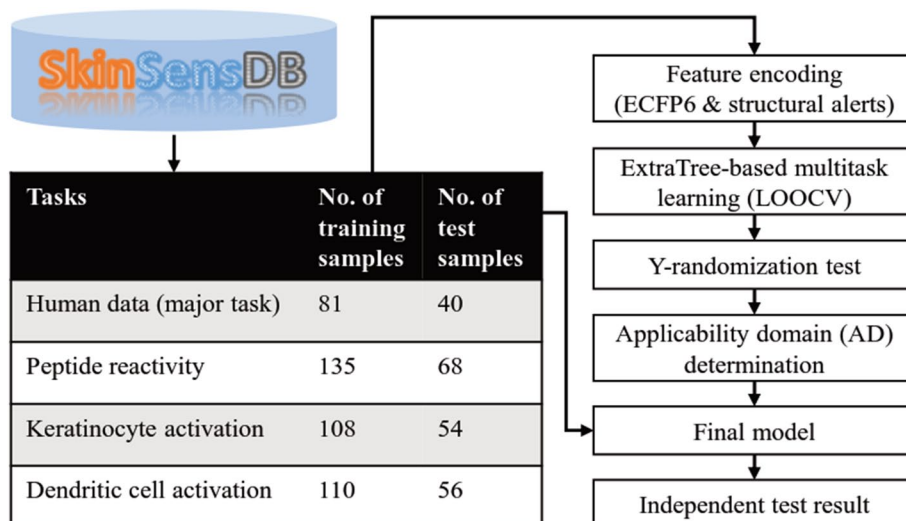


Fig. 1 System flow for developing and validating the proposed multitask model. Data were extracted from a curated database SkinSensDB and randomly divided into training and test datasets. The training dataset was utilized to develop prediction models and determine applicability domain. For the selection of models and features,

a leave-one-out cross-validation (LOOCV) technique and y-randomization test were applied to evaluate the robustness of the model. The independent test dataset was then utilized to test the prediction performance for unseen chemicals

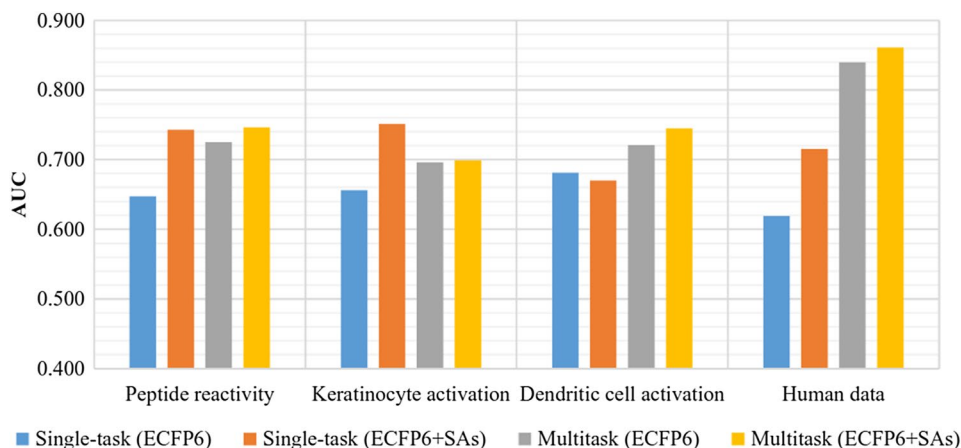
decision threshold. An AUC value close to 1 indicates a good prioritizing ability. The system flow is shown in Fig. 1.

To determine the effect of different feature encodings on the prediction performance, a leave-one-out cross-validation (LOOCV) method was applied to the training dataset. The first experiment utilized the ECFP6 to encode chemicals into feature vectors with 1024 bits for developing prediction models. The proposed multitask learning method achieves a high AUC value of 0.840 based on LOOCV that is much higher than a conventional single-task learning method (AUC = 0.619). The second experiment utilized the multitask learning method with a combined feature vector consisting of ECFP6 and SAs that further improved its prediction performance with a higher AUC value of 0.861. The inclusion of SAs as features also improved the performance

of single-task learning method with an AUC of 0.715. The multitask learning method with the combined feature vector was then adopted to build the final prediction model for further evaluation. To further validate the model, 100 runs of y-randomization were conducted based on modified training datasets with randomly permuted target values and compared with the original model. The original model with an AUC value of 0.861 is much better than that of y-randomized models ($AUC = 0.5505 \pm 0.058$) showing the significance and robustness of the developed model. The detailed comparison of LOOCV performance is shown in Fig. 2 (Supplementary Table S1 and S2).

Since the key to successful application of transfer learning is the relevance of learning tasks, it is interesting to know whether the multitask learning method can be applied to

Fig. 2 Comparison of LOOCV performance based on the training dataset. The proposed multitask learning methods achieve superior performance over conventional single-task methods. The feature of structure alerts (SAs) offers useful information for improving LOOCV performance



predict data obtained from rodent-based assay (local lymph node assay, LLNA). Additional experiments were conducted to evaluate LOOCV performance of multitask and single-task methods using LLNA and the three AOP tasks. Similarly, 483 chemicals with LLNA data were extracted from SkinSensDB and divided into training and independent test datasets in a ratio of 2:1. Based on the combined feature vector of ECFP6 and SAs, a poor AUC performance of 0.686 for the multitask learning method was obtained that is only slightly higher than that of the single-task method with an AUC value of 0.651. The results indicate that LLNA might not be highly relevant to the three AOP tasks that are consistent with previous studies (Roberts and Patlewicz 2018; Tung et al. 2018).

Independent test and the determination of applicability domain (AD)

The independent test dataset of human data was utilized to further estimate the prediction performance for unseen chemicals. A good AUC performance of 0.858 was obtained showing the robustness of the proposed model. Since the chemical space is too big to cover, only models with a well-defined AD could be potentially useful in real-world applications. For the adjustment of AD, a strict method was applied that the determination of fingerprints outside the AD depends solely on the LOOCV performance for the training dataset. The independent test dataset was again used for evaluating the prediction performance. The selection of fingerprints for the exclusion of chemicals outside of the AD is based on the decrease of AUC performance. First, the decrease of AUC values for each fingerprint was calculated using chemicals with and without the corresponding fingerprint. Subsequently, the fingerprints whose occurrence giving a performance decrease larger than or equal to a threshold value $t \in \{0.5, 0.6, \dots, 0.9\}$ were identified as outside of AD. The evaluation results are shown in Table 1. For the threshold $t=0.9$ without any fingerprints selected,

the LOOCV performance is the same as the model based on all chemicals. For $t=0.8$, only two fingerprints were selected and its corresponding AUC value is 0.913 showing a large improvement. For $t=0.7$, 79 fingerprints were selected as outside of AD and all performance measurements reach a high value with an excellent AUC value of 0.984. For $t=0.5$ and $t=0.6$, the performance is only slightly increased with two and one additionally selected fingerprints, respectively, and could potentially overfit the training set. The threshold value of 0.7 is therefore considered the best parameter for AD determination. A total of 79 fingerprints whose occurrence shows at least 70% worse prediction performance were selected based on the training dataset for the exclusion of chemicals outside of the AD. After the adjustment of AD, the prediction model shows promising AUC performance of 0.984 and 0.915 for LOOCV and independent test, respectively, with a coverage of 72.73%. The corresponding accuracies are 0.955 and 0.909 for LOOCV and independent test, respectively.

To avoid overestimation of prediction performance solely based on the independent test dataset, the training dataset was further randomly divided into two sets with similar data size. Subsequently, a threefold cross-validation scheme using the original parameters and AD definition was applied to iteratively test the robustness of our model. For each run, one fold is utilized to test the prediction model trained on the other two folds. The test AUC values of the three folds are 0.858, 0.807 and 0.882 without AD adjustment and 0.915, 0.912 and 0.923 with AD adjustment, respectively. The small difference among the AUC values with AD adjustment shows the robustness of our model.

Comparison to existing methods

To demonstrate the effectiveness of the multitask model, we compared its prediction performance with four existing methods including conventional SAs, recently published two read-across methods (Tung et al. 2018) and Pred-Skin (Braga et al. 2017). Since Pred-Skin reports only accuracy values based on cross-validation, the comparison is based on cross-validation accuracy except for read-across methods and SAs without a training process. ToxTree (Patlewicz et al. 2008) was utilized to generate SAs for predicting 121 human data and the chemicals with fired alerts were considered in its AD. The previously published mechanism-based read-across methods (Tung et al. 2018) of majority vote and decision tree were implemented to evaluate its prediction performance whose AD was determined by applying various similarity threshold. The published fivefold cross-validation performance of Pred-Skin based on a similar human dataset was utilized for comparison (Alves et al. 2016b; Braga et al. 2017). Please note that the AD determination for Pred-Skin was based on all available data. The AD determination

Table 1 LOOCV performance on the training dataset for various threshold values

Threshold	0.5	0.6	0.7	0.8	0.9
#Features outside of AD	81	80	79	2	0
Accuracy	0.968	0.969	0.955	0.840	0.825
Sensitivity	0.973	0.974	0.950	0.864	0.833
Specificity	0.962	0.962	0.962	0.806	0.812
Precision	0.973	0.974	0.974	0.864	0.870
AUC	0.993	0.993	0.984	0.913	0.858
MCC	0.935	0.935	0.906	0.670	0.640
CCR	0.968	0.968	0.956	0.835	0.823

for the proposed multitask model was solely based on the training dataset suggesting a more reliable estimation of prediction performance. The recalculation of AD based on the whole dataset can further improve the coverage and accuracy, however, there will be no external data to independently test the AD. For comparison, LOOCV performance was calculated for the proposed multitask method using the combined feature vector on the whole dataset (Supplementary Table S3). The performance comparison of the proposed method and publicly available methods is shown in Fig. 3. Please note that only performance with a coverage above 70% were considered except for Pred-Skin and SAs with fewer data points for comparison.

The proposed multitask model provides the highest accuracy and coverage for predicting chemical sensitizers. The utilization of the default similarity threshold of 0.7 (Tung et al. 2018) generated a good CCR performance of 0.834 for majority vote-based read-across methods, while the decision tree-based models failed to generate good performance. It is interesting to note that the CCR of SAs is higher than Pred-Skin and read-across methods without considering their AD. For comparison, the CCR of the LLNA method for predicting human data was also evaluated based on curated data in

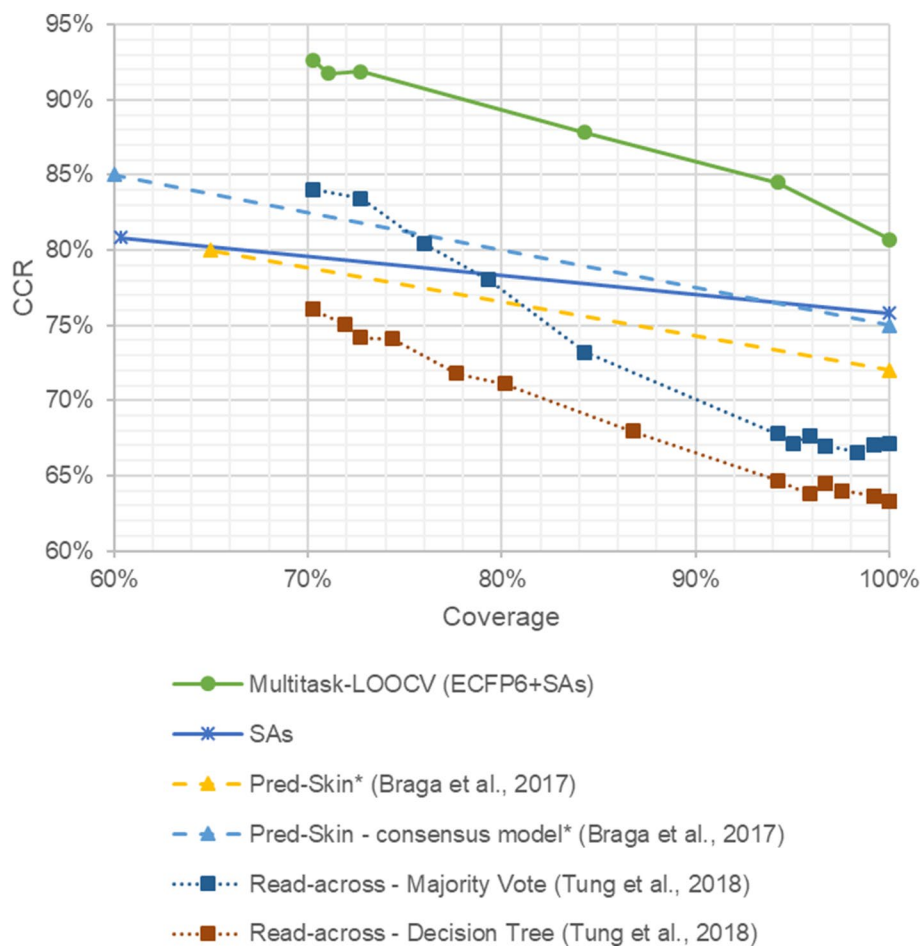
SkinSensDB. A CCR performance of 0.741 was obtained based on a total of 86 chemicals with both LLNA and human data. The results show that the proposed multitask model is promising for the identification of human skin sensitizers.

Estimation of prediction performance for unseen chemicals without any assay information

While the proposed model with defined AD performs well in the previous experiments, the success of prediction could benefit from the data of the three relevant tasks associated with the same test chemicals. To further test the robustness of our method for unseen chemicals without any data of the four tasks, a modified LOOCV (MLOOCV) test was conducted on all available chemicals to simulate the situation. For each holdout test chemical in the major task, only data not associated with the test chemical were utilized to construct the prediction model for evaluating the holdout test chemical, i.e., there is no data in the model associated with the test chemical.

As expected, the lack of relevant data in the model led to a lower AUC of 0.781 for the MLOOCV test without incorporating AD. However, with the defined AD by excluding

Fig. 3 Comparison of the proposed multitask models with existing models. The proposed multitask model provides good performance compared with other approaches. For structural alerts (SAs), the performance value within AD is the accuracy considering only chemicals with alerts. *The reported performance data of Pred-Skin for predicting human data is based on fivefold cross-validation and the determination of applicability domain is based on its whole dataset



chemicals having the 79 selected fingerprints, a high AUC value of 0.887 was obtained for the MLOOCV test (Supplementary Table S4). The CCR value of MLOOCV test with the defined AD is 0.843 (coverage = 72.73%) that is still better than the Pred-Skin whose CCR is 0.8 for a lower coverage of 65% (Braga et al. 2017). Please note that the consensus model of Pred-Skin utilized both LLNA and human data for prediction that is similar to the LOOCV test without excluding chemicals with LLNA data and is therefore excluded in the comparison.

The read-across methods (Tung et al. 2018) were also re-implemented to evaluate their MLOOCV performance. As the predictivity of read-across methods rely on the size of database, a moderate CCR value of 0.791 was obtained for the majority vote model with a very low coverage of 35.54%. The results showed that the proposed multitask learning method with combined feature vectors can be applied to unseen chemicals without any data in the training dataset when the chemicals fall within the defined AD.

SkinSensPred server

To facilitate the prioritization and identification of human skin sensitizers, we have implemented a user-friendly web tool named SkinSensPred that is freely available at <https://cwtung.kmu.edu.tw/skinsensdb/predict>. All available data were utilized for training the final prediction model. Users can use SkinSensPred by either entering SMILES

or drawing chemical structures. Given test chemicals, the server will automatically encode feature vectors and apply the prediction model to generate prediction results with additional information of probabilities, applicability domain and structure alerts. A snapshot of the web server is shown in Fig. 4.

Assessment of pro- and pre-haptens

Indirectly acting sensitizers including pro- and pre-haptens need activation to be sensitizing (Lepoittevin 2006). It is interesting to test SkinSensPred for its ability to identify pro- and pre-haptens as human skin sensitizers. A total of 20 chemicals obtained from a previous study (Patlewicz et al. 2016) were applied to test SkinSensPred. Table 2 shows the prediction results with scores, AD, SAs and human data. Among the 20 tested chemicals, 18 chemicals were correctly predicted (accuracy = 90.00%). By excluding chemicals outside of AD, a 94.75% accuracy (15/16) was obtained. A total of six chemicals of 1,4-phenylenediamine, 2,5-diaminotoluene sulphate, isoeugenol, D-limonene, cinnamyl alcohol, and geraniol were not curated in SkinSensDB yet and only D-limonene was incorrectly predicted. A sensitizer linalool was incorrectly predicted as a non-sensitizer with a low score of 0.06 but it does not fall in the AD. D-limonene is the only incorrectly predicted chemical within the defined AD with a score of 0.57 that is close to the decision threshold representing a less confident prediction. D-limonene was also

The screenshot displays the SkinSensPred web interface. The top navigation bar includes links for HOME, DATABASE, STATISTICS, DOWNLOAD, and HELP. Below this, there are tabs for SEARCH, PREDICT, and BROWSE. The main interface is divided into two sections: 'Input chemical(s)' and 'Result(s)'. The 'Input chemical(s)' section contains a 'BATCH ADD' button and a text area for pasting SMILES strings, with an example provided: COC(=O)C=CC(=O)OC, CC1=CC(=C(C(=C1)O)C=O)O, and CC(CCO)O. Below the text area is an 'ADD FROM MOLECULAR EDITOR' button and a molecular editor tool. The 'Result(s)' section shows a table of predictions for the three input chemicals. The table includes columns for SMILES, Human sensitizer (Score), AD, Acyl transfer agent, Michael acceptor, SN2, SNAr, Schiff base, and Read across. The results are as follows:

SMILES	Human sensitizer (Score)	AD	Acyl transfer agent	Michael acceptor	SN2	SNAr	Schiff base	Read across
<chem>COC(=O)C=CC(=O)OC</chem>	1 (1.00)	Y	✓	1	✓	✓	✓	✗
<chem>CC1=CC(=C(C(=C1)O)C=O)O</chem>	1 (0.98)	N	✓	1	✓	✓	1	✗
<chem>CC(CCO)O</chem>	✓ (0.61)	Y	✓	✓	✓	✓	✓	✗

At the bottom right of the results table, it indicates '1 - 3 of 3' with navigation arrows.

Fig. 4 User interface of SkinSensPred. Users can either input SMILES of chemicals or draw a chemical structure for predicting human skin sensitizers and analyzing AD and SAs

Table 2 Prediction results for pro- and pre-haptens

Chemical	Skin-SensPred (score)	AD	SA	Human data
Ethylenediamine	P (1.00)	Y	Y	P
Resorcinol	P (0.99)	N	Y	P
3-Dimethylamino propylamine	P (1.00)	N	Y	P
1,2-Dibromo-2,4-dicyanobutane	P (1.00)	Y	Y	P
1,4-Phenylenediamine	P (1.00)	Y	Y	P
2,5-Diaminotoluene sulphate	P (0.57)	N	N	P
2-Aminophenol	P (1.00)	Y	Y	P
Abietic acid	P (0.94)	Y	N	P
Hydroquinone	P (1.00)	Y	Y	P
Isoeugenol	P (1.00)	Y	Y	P
Lauryl gallate	P (1.00)	Y	Y	P
Propyl gallate	P (1.00)	Y	Y	P
d-Limonene	P (0.57)	Y	N	N
Linalool	N (0.06)	N	N	P
2-Nitro-1,4-phenylenediamine	P (1.00)	Y	Y	P
Aniline	P (1.00)	Y	N	P
Chlorpromazine hydrochloride	P (1.00)	Y	Y	P
Cinnamyl alcohol	P (1.00)	Y	Y	P
Geraniol	P (1.00)	Y	Y	P
Eugenol	P (1.00)	Y	Y	P

Human data were obtained from Patlewicz et al. (2016)

reported to be a sensitizer by LLNA, DPRA and h-CLAT assays with an exception of the KeratinoSens™ assay (Patlewicz et al. 2016). Notably, d-limonene and linalool listed as Category 5 and 4, respectively (Basketter et al. 2014), are both in the categories close to the decision boundary for categorizing human sensitizers whose determination of human sensitizers is less confident. Altogether, SkinSensPred is expected to be useful for identifying sensitizers of pro- and pre-haptens.

Discussion

In conclusion, a novel prediction method SkinSensPred was developed for human skin sensitizers with a defined AD. The multitask learning-based model benefits from shared knowledge transferred from relevant learning tasks defined by AOP generate promising performance with high accuracy and coverage. The integration of SAs further enhanced its prediction performance. Rigorous tests including LOOCV, y-randomization and independent test as well as the comparison to existing state-of-the-art methods show the robustness of SkinSensPred. The ability of SkinSensPred for identifying sensitizers of pro- and pre-haptens was also evaluated. A user-friendly web server was implemented to facilitate

the study of human skin sensitizers and is freely available at <https://cwtung.kmu.edu.tw/skinsensdb/predict>. To the best of our knowledge, this is the first transfer learning work based on AOP knowledge. SkinSensPred is expected to be useful for prioritizing human skin sensitization potentials for further testing. As many novel AOPs is under active development, the proposed method could be applied to build predictive models for various chemical-induced toxicities by transferring shared knowledge from relevant learning tasks in AOPs.

The collection and generation of sufficient data for machine learning is the major obstacle to the development of effective models for various human toxicity endpoints. Also, conflicting results of the same assay from different laboratories are commonly observed. In this study, both SkinSensDB and the transfer learning algorithm contribute to the superior performance of SkinSensPred. SkinSensDB provides the largest and standardized data source by collecting data from different sources and eliminating outliers based on a majority vote scheme. According to the Chemical and Product Categories (CPCat) Database of US EPA (Dionisio et al. 2015), the chemicals curated in SkinSensDB are highly diverse covering 13 out of 15 ACToR UseDB categories, including antimicrobial, colorant, consumer use, drug, food additive, fragrance, industrial manufacturing, inert, personal care, pesticide, petrochemical, fertilizer and chemical warfare. SkinSensPred based on a diverse chemical set is expected to have broad applicability domain. While the key events and associated assays of AOP for skin sensitization are still under study, the high relevance of the key events to the adverse outcome provides useful shared knowledge to improve the prediction of skin sensitizers.

While the OECD defines AD as the response and chemical structure space in which the model makes predictions with given reliability, the implementation of AD can be varied to address three different points of applicability, reliability and decidability (Hanser et al. 2016; Netzeva et al. 2005). For applicability, the proposed model is a machine learning-based model built on the ECFP6 feature space of a diverse chemical dataset that is generally considered applicable to all chemicals. Therefore, the focus of the AD determination in this manuscript is on the reliability by fingerprint selection with a prediction score for the decidability. While the largely improved performance shows the effectiveness of the AD determined by the selection of fingerprints proposed in this study, other methods such as similarity, metabolic transformation and physicochemical properties could be further conducted to characterize the AD for various considerations.

The trade-off between interpretability and accuracy is a well-known issue for choosing a computational model. In the regulatory aspect, neither rule-based or statistically based method alone is considered sufficient for decision-making. Rule-based methods with interpretable mechanistic

rules assuming a direct correlation between structure features/properties and a given endpoint provide only moderate accuracy due to the complex nature of biological systems. In contrast, statistically or machine learning-based methods capable of capturing high-level nonlinear correlation between input features and the endpoint can generate high accuracy and are usually considered as a black-box method. In this study, the utilized ECFP6 features are hashed fingerprints without predefined substructure for each bit that are computationally efficient, but further decrease the interpretability of the constructed model. ICH M7 Guideline for accessing mutagenic impurities provides a potential example for adopting computational methods to rule out mutagenicity concern of chemicals by applying both the two complementary methods in the absence of adequate experimental data (Bossà et al. 2018). SkinSensPred prediction system represents a unique machine learning method incorporating rules of structural alerts into feature vectors generating high accuracy and the prediction server also generates rule-based mechanistic information as a complementary method to improve regulatory acceptance.

Acknowledgements This work was supported by Ministry of Science and Technology of Taiwan [MOST104-2221-E-037-001-MY3, MOST107-2221-E-037-005-MY3]; National Health Research Institutes [NHRI-107A1-EMCO-0318184]; and Research Center for Environmental Medicine in Kaohsiung Medical University from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. This work was initiated in Kaohsiung Medical University and completed in Taipei Medical University. The funding agencies play no role in the study design, data analysis and manuscript preparation.

Data availability SkinSensPred and datasets are freely available at <https://cwtung.kmu.edu.tw/skinsensdb>.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

- Alves V, Muratov E, Capuzzi S et al (2016a) Alarms about structural alerts. *Green Chem* 18(16):4348–4360. <https://doi.org/10.1039/C6GC01492E>
- Alves VM, Capuzzi SJ, Muratov E et al (2016b) QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. *Green Chem* 18(24):6501–6515. <https://doi.org/10.1039/C6GC01836J>
- Basketter DA, Alepee N, Ashikaga T et al (2014) Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis* 25(1):11–21. <https://doi.org/10.1097/DER.0000000000000003>
- Basketter DA, White IR, McFadden JP, Kimber I (2015) Skin sensitization: Implications for integration of clinical data into hazard identification and risk assessment. *Hum Exp Toxicol* 34(12):1222–1230. <https://doi.org/10.1177/0960327115601760>
- Bauch C, Kolle SN, Ramirez T et al (2012) Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol* 63(3):489–504. <https://doi.org/10.1016/j.yrtph.2012.05.013>
- Bossà C, Benigni R, Tcheremenskaia O, Battistelli CL (2018) (Q)SAR methods for predicting genotoxicity and carcinogenicity: scientific rationale and regulatory frameworks. *Methods Mol Biol* 1800:447–473. https://doi.org/10.1007/978-1-4939-7899-1_20
- Braga RC, Alves VM, Muratov EN et al (2017) Pred-Skin: A Fast and Reliable Web Application to Assess Skin Sensitization Effect of Chemicals. *J Chem Inf Model* 57(5):1013–1017. <https://doi.org/10.1021/acs.jcim.7b00194>
- Dionisio KL, Frame AM, Goldsmith MR et al (2015) Exploring consumer exposure pathways and patterns of use for chemicals in the environment. *Toxicol Rep* 2:228–237. <https://doi.org/10.1016/j.toxrep.2014.12.009>
- Ezendam J, Braakhuis HM, Vandebriel RJ (2016) State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. *Arch Toxicol* 90(12):2861–2883. <https://doi.org/10.1007/s00204-016-1842-4>
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63(1):3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Golla S, Madhally S, Robinson RL Jr, Gasem KA (2009) Quantitative structure-property relationship modeling of skin sensitization: a quantitative prediction. *Toxicol In Vitro* 23(3):454–465. <https://doi.org/10.1016/j.tiv.2008.12.025>
- Guha R (2007) Chemical informatics functionality in R. *J Stat Softw* 18(5):1–16
- Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. *SAR QSAR Environ Res* 27(11):893–909. <https://doi.org/10.1080/1062936X.2016.1250229>
- Huang SH, Tung CW (2017) Identification of consensus biomarkers for predicting non-genotoxic hepatocarcinogens. *Sci Rep* 7:41176. <https://doi.org/10.1038/srep41176>
- Jaworska J, Hoffmann S (2010) Integrated Testing Strategy (ITS) - Opportunities to better use existing data and guide future testing in toxicology. *ALTEX* 27(4):231–242
- Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, Miyazawa M (2015) Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch Toxicol* 89(12):2355–2383. <https://doi.org/10.1007/s00204-015-1634-2>
- Jin H, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 17(3):299–310. <https://doi.org/10.1109/TKDE.2005.50>
- Karlberg AT, Bergstrom MA, Borje A, Luthman K, Nilsson JL (2008) Allergic contact dermatitis—formation, structural requirements, and reactivity of skin sensitizers. *Chem Res Toxicol* 21(1):53–69. <https://doi.org/10.1021/tx7002239>
- Kleinstreuer NC, Hoffmann S, Alepee N et al (2018) Non-animal methods to predict skin sensitization (II): an assessment of defined approaches (*). *Crit Rev Toxicol* 48(5):359–374. <https://doi.org/10.1080/10408444.2018.1429386>
- Lepoittevin JP (2006) Metabolism versus chemical transformation or pro- versus prehapens? *Contact Dermatitis* 54(2):73–74. <https://doi.org/10.1111/j.0105-1873.2006.00795.x>
- Li Y, Pan D, Liu J et al (2007a) Categorical QSAR Models for skin sensitization based upon local lymph node assay classification measures part 2: 4D-fingerprint three-state and two-2-state logistic regression models. *Toxicol Sci* 99(2):532–544. <https://doi.org/10.1093/toxsci/kfm185>
- Li Y, Tseng YJ, Pan D et al (2007b) 4D-fingerprint categorical QSAR models for skin sensitization based on the classification of local

- lymph node assay measures. *Chem Res Toxicol* 20(1):114–128. <https://doi.org/10.1021/tx6002535>
- Miller MD, Yourtee DM, Glaros AG, Chappelow CC, Eick JD, Holder AJ (2005) Quantum mechanical structure-activity relationship analyses for skin sensitization. *J Chem Inf Model* 45(4):924–929. <https://doi.org/10.1021/ci050018z>
- Netzeva TI, Worth A, Aldenberg T et al (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern Lab Anim* 33(2):155–173
- OECD (2018) OECD QSAR Toolbox. <https://www.qsartoolbox.org/>. Accessed 13 Sep 2018
- Patlewicz GY, Basketter DA, Pease CK et al (2004) Further evaluation of quantitative structure-activity relationship models for the prediction of the skin sensitization potency of selected fragrance allergens. *Contact Dermatitis* 50(2):91–97. <https://doi.org/10.1111/j.0105-1873.2004.00322.x>
- Patlewicz G, Jeliakova N, Safford RJ, Worth AP, Aleksiev B (2008) An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* 19(5–6):495–524. <https://doi.org/10.1080/10629360802083871>
- Patlewicz G, Casati S, Basketter DA et al (2016) Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul Toxicol Pharmacol* 82:147–155. <https://doi.org/10.1016/j.yrtph.2016.08.007>
- Roberts DW, Api AM (2018) Chemical applicability domain of the local lymph node assay (LLNA) for skin sensitisation potency. Part 4. Quantitative correlation of LLNA potency with human potency. *Regul Toxicol Pharmacol* 96:76–84. <https://doi.org/10.1016/j.yrtph.2018.04.022>
- Roberts DW, Patlewicz G (2018) Non-animal assessment of skin sensitization hazard: Is an integrated testing strategy needed, and if so what should be integrated? *J Appl Toxicol* 38(1):41–50. <https://doi.org/10.1002/jat.3479>
- Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754. <https://doi.org/10.1021/ci100050t>
- Sarath Kumar KL, Tangadpalliwar SR, Desai A, Singh VK, Jere A (2016) Integrated Computational Solution for Predicting Skin Sensitization Potential of Molecules. *PLoS One* 11(6):e0155419. <https://doi.org/10.1371/journal.pone.0155419>
- Simm J, Magrans De Abril I, Sugiyama M (2014) Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression. *IEICE Transactions on Information Systems* E97D(6):1677–1681. <https://doi.org/10.1587/transinf.E97.D.1677>
- Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV (2012) ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J Chem Inf Model* 52(8):2310–2316. <https://doi.org/10.1021/ci300245q>
- Tung CW (2013) Prediction of pupylation sites using the composition of k-spaced amino acid pairs. *J Theor Biol* 336:11–17. <https://doi.org/10.1016/j.jtbi.2013.07.009>
- Tung CW, Ho SY (2008) Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics* 9:310. <https://doi.org/10.1186/1471-2105-9-310>
- Tung CW, Wang CC, Wang SS (2018) Mechanism-informed read-across assessment of skin sensitizers based on SkinSensDB. *Regul Toxicol Pharmacol* 94:276–282. <https://doi.org/10.1016/j.yrtph.2018.02.014>
- Urbisch D, Mehling A, Guth K et al (2015) Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul Toxicol Pharmacol* 71(2):337–351. <https://doi.org/10.1016/j.yrtph.2014.12.008>
- Wang CC, Lin YC, Wang SS, Shih C, Lin YH, Tung CW (2017) SkinSensDB: a curated database for skin sensitization assays. *J Cheminform* 9:5. <https://doi.org/10.1186/s13321-017-0194-2>
- Willighagen EL, Mayfield JW, Alvarsson J et al (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9(1):33. <https://doi.org/10.1186/s13321-017-0220-4>
- Yuan H, Huang J, Cao C (2009) Prediction of skin sensitization with a particle swarm optimized support vector machine. *Int J Mol Sci* 10(7):3237–3254. <https://doi.org/10.3390/ijms10073237>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Contents

Supplementary Tables	2
Table S1. Comparison of LOOCV performance on the training dataset for single-task and multitask models using ECFP6.....	2
Table S2. Comparison of LOOCV performance on the training dataset for single-task and multitask models using ECFP6 and SAs.	2
Table S3. Conventional LOOCV performance on the whole dataset.	3
Table S4. Modified LOOCV (MLOOCV) performance on the whole dataset.	3

Supplementary Tables

Table S1. Comparison of LOOCV performance on the training dataset for single-task and multitask models using ECFP6.

Task	Single-task				Multitask			
	Peptide reactivity	Keratinocyte activation	Activation of dendritic cells	Human data	Peptide reactivity	Keratinocyte activation	Activation of dendritic cells	Human data
TP	90	57	50	35	73	56	47	38
TN	10	14	18	14	23	19	25	20
FP	26	23	27	18	13	18	20	12
FN	9	14	15	14	26	15	18	11
Accuracy	0.741	0.657	0.618	0.605	0.711	0.694	0.655	0.716
Sensitivity	0.909	0.803	0.769	0.714	0.737	0.789	0.723	0.776
Specificity	0.278	0.378	0.400	0.438	0.639	0.514	0.556	0.625
Precision	0.776	0.712	0.649	0.660	0.849	0.757	0.701	0.760
AUC	0.647	0.656	0.681	0.619	0.725	0.696	0.721	0.840
MCC	0.238	0.196	0.182	0.156	0.346	0.309	0.281	0.403
CCR	0.594	0.591	0.585	0.576	0.688	0.652	0.640	0.701

Table S2. Comparison of LOOCV performance on the training dataset for single-task and multitask models using ECFP6 and SAs.

Task	Single-task				Multitask			
	Peptide reactivity	Keratinocyte activation	Activation of dendritic cells	Human data	Peptide reactivity	Keratinocyte activation	Activation of dendritic cells	Human data
TP	90	61	49	40	73	55	48	38
TN	13	17	19	17	23	19	26	25
FP	23	20	26	15	13	18	19	7
FN	9	10	16	9	26	16	17	11
Accuracy	0.763	0.722	0.618	0.704	0.711	0.685	0.673	0.778
Sensitivity	0.909	0.859	0.754	0.816	0.737	0.775	0.738	0.776
Specificity	0.361	0.459	0.422	0.531	0.639	0.514	0.578	0.781
Precision	0.796	0.753	0.653	0.727	0.849	0.753	0.716	0.844
AUC	0.743	0.751	0.670	0.715	0.746	0.699	0.745	0.861
MCC	0.324	0.349	0.186	0.364	0.346	0.292	0.319	0.548
CCR	0.635	0.659	0.588	0.674	0.688	0.645	0.658	0.779

Table S3. Conventional LOOCV performance on the whole dataset.

#features for exclusion	102	97	96	95	81	80	79	11	2	0
#samples in AD	55	63	65	66	85	86	88	102	114	121
coverage	45.45%	52.07%	53.72%	54.55%	70.25%	71.07%	72.73%	84.30%	94.21%	100%
TP	36	38	38	39	48	48	50	54	59	60
TN	16	20	22	22	31	31	31	36	38	38
FP	2	3	3	3	3	3	3	6	9	10
FN	1	2	2	2	3	4	4	6	8	13
Accuracy	0.945	0.921	0.923	0.924	0.929	0.919	0.920	0.882	0.851	0.810
Sensitivity	0.973	0.950	0.950	0.951	0.941	0.923	0.926	0.900	0.881	0.822
Specificity	0.889	0.870	0.880	0.880	0.912	0.912	0.912	0.857	0.809	0.792
Precision	0.947	0.927	0.927	0.929	0.941	0.941	0.943	0.900	0.868	0.857
AUC	0.953	0.941	0.945	0.945	0.940	0.933	0.934	0.919	0.898	0.870
MCC	0.875	0.828	0.837	0.838	0.853	0.831	0.833	0.757	0.691	0.608
CCR	0.931	0.910	0.915	0.916	0.927	0.918	0.919	0.879	0.845	0.807

Table S4. Modified LOOCV (MLOOCV) performance on the whole dataset.

#features for exclusion	102	97	96	95	81	80	79	11	2	0
#samples in AD	55	63	65	66	85	86	88	102	114	121
coverage	45.45%	52.07%	53.72%	54.55%	70.25%	71.07%	72.73%	84.30%	94.21%	100%
TP	32	34	34	34	42	43	45	49	53	55
TN	15	19	21	21	29	29	29	33	34	34
FP	3	4	4	4	5	5	5	9	13	14
FN	5	6	6	7	9	9	9	11	14	18
Accuracy	0.855	0.841	0.846	0.833	0.835	0.837	0.841	0.804	0.763	0.736
Sensitivity	0.865	0.850	0.850	0.829	0.824	0.827	0.833	0.817	0.791	0.753
Specificity	0.833	0.826	0.840	0.840	0.853	0.853	0.853	0.786	0.723	0.708
Precision	0.914	0.895	0.895	0.895	0.894	0.896	0.900	0.845	0.803	0.797
AUC	0.923	0.878	0.887	0.882	0.884	0.884	0.887	0.841	0.812	0.781
MCC	0.681	0.665	0.681	0.657	0.667	0.669	0.675	0.599	0.513	0.456
CCR	0.849	0.838	0.845	0.835	0.839	0.840	0.843	0.802	0.763	0.753