Chemical Research in Toxicology

pubs.acs.org/crt

Article

Applicability Domains Enhance Application of PPARγ Agonist Classifiers Trained by Drug-like Compounds to Environmental Chemicals

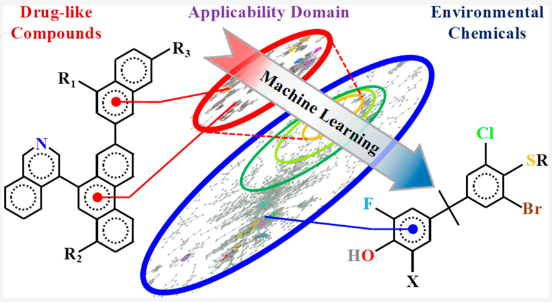
Zhongyu Wang, Jingwen Chen,* and Huixiao Hong

Cite This: *Chem. Res. Toxicol.* 2020, 33, 1382−1388

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Peroxisome proliferator activator receptor gamma (PPARγ) agonist activity of chemicals is an indicator of concerned health conditions such as fatty liver and obesity. *In silico* screening PPARγ agonists based on quantitative structure−activity relationship (QSAR) models could serve as an efficient and pragmatic strategy. Owing to the broad research interests in discovery of PPARγ-targeted drugs, a large amount of PPARγ agonist activity data has been produced in the field of medicinal chemistry, facilitating development of robust QSAR models. In this study, random forest classifiers were developed based on the binary-category data transformed from the heterogeneous PPARγ agonist activity data of drug-like compounds. Coupling with applicability domains, capability of the established classifiers for predicting environmental chemicals was evaluated using two external data sets. Our results demonstrated that applicability domains could enhance application of the developed classifiers to predict environmental PPARγ agonists.



INTRODUCTION

Peroxisome proliferator activator receptor gamma (PPARγ) is a member of nuclear receptor family that plays versatile roles in human physiology.[1,2] In the recent decade, transactivation of PPARγ has been found to be related to human fatty liver disease and endocrine-disrupting effects including undermined osteogenesis, promoted adipogenesis, and increased tendency toward later-life obesity.[3−5] PPARγ agonist activity of chemicals thus serves as an early indicator of the adverse health effects. Many chemicals such as organotins, phthalate esters, chlorinated polyfluorinated ether sulfonates, and brominated and organo-phosphorus flame retardants have been reported as PPARγ agonists in *in vitro* assays.[6−13,4] However, *in vitro* assays are time-consuming and expensive, thus not suitable for screening of the huge amount of environmental chemicals.

In vitro assays have provided valuable data for developing quantitative structure−activity relationship (QSAR) models.[14,15] QSAR modeling attempts to learn correlating patterns between the interested activity (e.g., the PPARγ agonist activity) and the structural description or features (e.g., molecular descriptors[16]) of a set of compounds (i.e., training set). Such patterns are expected to be valid for external compounds that fall inside the applicability domains of the models.[17] With advent of the big data era,[18,19] QSAR models have become a promising tool for filling up the data gap necessary for chemicals risk assessment. For example, the PPARγ agonist activity assay has been automated by the Tox21 project into a high-throughput screening (HTS) technology, which has screened thousands of chemicals (the so-called Tox21 10k library).[20,21] The Tox21 data set has been intensively exploited for QSAR modeling due to its public availability, broad coverage of chemicals, consistent bioassay protocols, and ready-made activity classes. Indeed, various machine learning algorithms have been used for establishing QSAR models by the Tox21 data with satisfactory performance.[22]

The dose gradients and cell lines in a typical HTS test are fixed for profiling many compounds,[23] which together with other problems can significantly affect the quality of HTS data and consequently influence the quality of established QSAR models. Indeed, Janesick et al. recently warned that PPARγ agonists identified by the Tox21 project contain some false positives.[24] Hence, exploring and integrating new data sets has become essential for improving QSAR modeling of PPARγ agonist activity.

Owing to the need for discovering PPARγ-targeted drugs for diseases such as diabetes,[25,26] PPARγ agonist activity data of a considerable number of drug-like compounds have been reported in the medicinal chemistry literature and collected in public databases such as ChEMBL.[27−29] In this study, the PPARγ agonist activity data of drug-like compounds in

Special Issue: Environmental Toxicology

Received: December 12, 2019
Published: February 4, 2020

https://dx.doi.org/10.1021/acs.chemrestox.9b00498

ChEMBL were curated and explored. An evidence-based scheme was developed to discretize the heterogeneous PPARγ agonist activity data into categorized data, based on which classifiers were established. Random forest (RF) algorithm[30] was employed for its convenience, competence, and efficiency for learning patterns from large data sets.[31,22,32] The ChEMBL compounds represent only a subset of the vast number of structurally diverse environmental chemicals faced by the regulators, implying the necessity of applicability domains (ADs) for the developed classifiers.[33] The established classifiers with associated ADs were then applied to environmental chemicals so as to demonstrate their usefulness.

■ MATERIALS AND METHODS

Data Curation. From the ChEMBL database, the *Homo sapiens* PPARγ bioactivity data were retrieved and then integrated into a single data sheet where each activity record is a continuous value or a comment from an assay for a compound. Each assay has a brief summary on its purpose and function. Cell line information on the transactivation assays was extracted from the "Assay Description" in the database or retrieved from literature. Additionally, for any efficacy-related transactivation assay, the positive reference compound (e.g., rosiglitazone) and its efficacy value (e.g., 100% or 120-fold) were also extracted from the "Assay Description" or retrieved from literature. Erroneous records (e.g., wrongly handled units, incorrectly transcribed data) were corrected or discarded. The final ChEMBL data set is provided in Table S1.

As for typical environmental chemicals, the chemicals with reported PPARγ agonist activities in the field of environmental health and toxicology (EHT) were collected as the first external set (Table S2). Quantitative records and conclusive descriptions of the transactivation activities in associated literature were referenced to categorize the PPARγ agonist activities of the EHT compounds.

The Tox21 chemicals[34] were employed as the second external set. The PPARγ agonist bioassay data were obtained from PubChem[35] (AID: 743140).

Processing Compounds. RDKit (version 2019.03.1, www.rdkit.org) was employed to process the simplified molecular input line entry system (SMILES) codes of the compounds. For any compound with disconnected SMILES code, if a monocomponent compound remains after stripping off simple inorganic or organic components/salts (Table S3), the compounds were then neutralized and their canonicalized SMILES codes were used as their structure identities. Finally, the SMILES codes with metal or metalloid were removed, which resulted in only organic compounds. The same processing was applied to chemicals in the external sets. For the Tox21 chemicals, the chemicals with duplicate or inconsistent data were removed. Furthermore, for EHT and Tox21 data sets, the compounds that were also found in the ChEMBL set were excluded, thus providing truly "unseen" chemicals for validating classifiers.

Processing PPARγ Agonist Activity Data. Transactivation and binding data were divided into five types: (a) quantitative transactivation efficacy values, (b) quantitative transactivation EC_{50} values, (c) qualitative transactivation activity comments, (d) quantitative binding affinity values, and (e) qualitative binding activity comments. The efficacy values were calibrated as percentage (calibrated efficacy values, *CEV*s, %) relative to known full agonists. The EC_{50} and binding affinity values (units converted to mol/L) were transformed into negative logarithm values (*NLV*s, e.g., $-\log EC_{50}$, $-\log IC_{50}$). Using *CEV*s, the efficacy data were defined as active ($CEV \geq 10\%$), inconclusive ($2\% \leq CEV < 10\%$), and inactive ($-2\% < CEV < 2\%$). Similarly, the EC_{50} and binding affinity data were categorized by *NLV*s as active ($NLV \geq 5$), inconclusive ($3 \leq NLV < 5$), and inactive ($NLV < 3$). The above criteria cover an activity range of the most reported environmental chemicals,[10] meanwhile excluding a small fraction of compounds with weak binding affinities or transactivation efficacies that may be tagged as inconclusive agonists by different bioassays. For each compound, active, inconclusive, and inactive data were counted for each of the aforementioned five data types. A compound is considered active if it has at least one active efficacy record, at least one active EC_{50} or binding affinity record, and no inactive records across all five record types; inactive if it has at least one inactive records across all five types and no active or inconclusive transactivation records; inconclusive otherwise. The categorized data set is provided in Table S4.

Molecular Descriptors. Molecular descriptors were generated from the neutralized SMILES codes of the compounds using Mordred (v.1.2.0).[36] Three dimensional (3D) descriptors were not considered to avoid inconsistency in multiple 3D conformations of flexible molecules. Descriptors that cannot be correctly calculated for at least one compound or are constant across all training-set compounds are removed. To guarantee that the external data sets can be predicted using the model developed on the training data set, the descriptors in both the training set and the external sets were kept.

Fingerprints of Compounds and Similarity Measurement. The "GetMorganFingerprintAsBitVect" function in the RDKit package was employed to generate extended-connectivity fingerprints (ECFPs).[37] A diameter of 4 was adopted (i.e., ECFP4), and fingerprints were folded into 1024 bits. The "GenMACCSKeys" function was employed to generate molecular access system (MACCS) structural keys.[38] For each compound, the Boolean vectors of its ECFP4 fingerprints or MACCS keys were employed as features that are different from the non-3D molecular descriptors for training classifiers. For any two compounds, the Tanimoto coefficient (also known as Jaccard index[39]) was calculated as their similarity based on either ECFP4 fingerprints or MACCS keys.

Classifier Training and Cross Validation. Compounds that were categorized as active and inactive were used as the training set. Scikit-learn[40] was employed to build and validate RF classifiers. A "RandomForestClassifier" instance was initialized with 500 trees (see Table S5 for other parameters) and fed with the activity categories and the non-3D Mordred descriptors or the fingerprints for the training.

Given the imbalance between the active and inactive classes of the training set, stratified five-fold cross validation (CV) was performed three times to estimate the performance of established classifiers. The stratified folds were randomly sampled by preserving the active-inactive ratio. Area under the receiver operating characteristic curve (AUC) was adopted for measuring performance of the classifiers because it decouples classifier performance from class skew.[41,42] AUCs were calculated by the "roc_auc_score" function in Scikit-learn. Besides, additional classifiers were also trained from randomly permuted feature matrices. The CV performance was compared between the classifiers based on the original training set and the permuted ones to check if the performance of the original training set based classifiers is obtained by chance. Given that the performance of the trained classifiers can be influenced by random seeds that determine actual instances and features sampled by trees of the forest,[30] 10 classifiers were generated from different random seeds and statistics of their CV AUCs were reported.

AD Characterization and Application of RF Classifiers to External Sets. A first-type AD was calculated based on the Euclidean distance in the descriptor space.[17] Briefly, the center of the training set was calculated using the standardized descriptors (mean-centered and standard deviation-scaled by "preprocessing.StandardScaler") as a vector of mean values of respective standardized descriptors. The maximum of the Euclidean distances (calculated by "DistanceMetric.pairwise" function) between the center and the training set compounds is taken as a threshold. If the Euclidean distance between a query compound (from the external set) and the center is larger than the threshold, the query compound is considered outside the AD.

Similarity to the molecules in training set is a good indicator for prediction accuracy of the trained QSAR model.[43] Thus, the ADs measured by similarity in ECFP4 fingerprints and MACCS keys were calculated to investigate the impact of ADs on prediction performance. If the number of compounds with similarity greater than a cutoff S_{cutoff} in the training set is larger than a defined number N_{min}, the query is inside the AD. We examined different ADs defined by combinations of N_{min} values (1 to 50, with the step of 1) and S_{cutoff} values (0.2 to 0.6 for ECFP4 fingerprints and 0.6 to 0.9 for MACCS keys, with the step of 0.01). These ADs were noted by their N_{min} and S_{cutoff} values as $AD_{fingerprint}\{N_{min}, S_{cutoff}\}$.

When considering ADs, only compounds inside the AD were predicted by the RF classifiers and corresponding AUCs were calculated. By comparing the AUCs before and after the consideration of the ADs, the functionality of different ADs was evaluated.

■ RESULTS AND DISCUSSION

Influence of Cell Lines on PPARγ Agonist Activity of ChEMBL Compounds. Most of the compounds curated from ChEMBL were tested in one unique cell line. Most of the compounds with PPARγ agonist activity from two or more cell lines have large ranges of *CEV*s or pEC_{50} (−logEC_{50}) values as depicted in Figure 1. Except for cell lines, other experimental conditions (e.g., molecular biological reporter systems, signal detection methods) may also contribute to the discordance of the *CEV* or pEC_{50} values.[44] Thus, accurate prediction of these values collected from heterogeneous data sources is not necessarily meaningful.

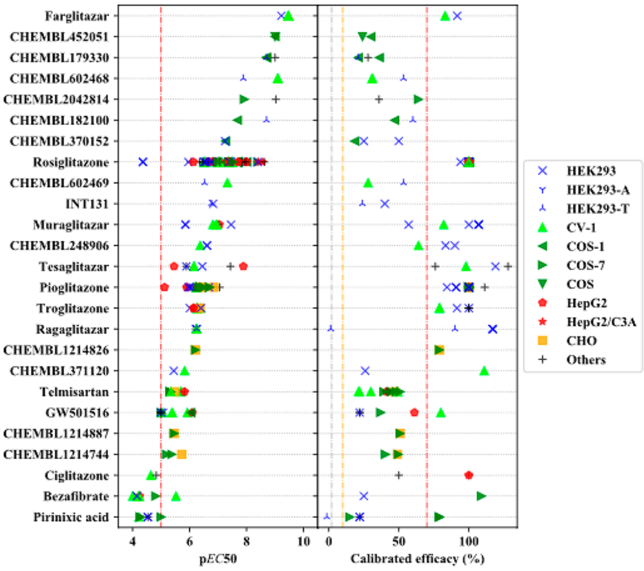


Figure 1. Distribution of efficacy and binding affinity of the PPARγ agonists. ChEMBL IDs are given for compounds without preferred common names. Cell line annotation: HEK293 and HEK293-A/T, human embryonic kidney cells; CV-1 and COS, green monkey kidney cells; HepG2 and HepG2/C3A, human hepatoblastoma cells; CHO, Chinese hamster ovary cells. Dashed lines: thresholds for discretizing the continuous values.

Considering the qualitative activities, if a compound has both active and inactive data, there is a major inconsistency in its activity. Otherwise, if a compound has both active (or inactive) and inconclusive evidence, there is a minor inconsistency in its activity. As to the transactivation efficacy, 10% chemicals have minor inconsistency, and 6% have major inconsistency. As to the pEC_{50}, 3% chemicals have minor inconsistency. On the basis of the above statistics, even though a drug-like compound has been tested only in a unique experimental condition (such as a certain cell line), it is generally confident to extrapolate its activity conclusion to other experimental conditions.

Discrepancy in Activity Data between Drug-like and Environmental PPARγ Agonists. As shown in Figure 2, the Tox21 chemicals with detected agonist activities (active or inconclusive) are on average weaker in both efficacy and potency/affinity (reflected by the *NLV*s) than the drug-like agonists. It can be reasonably assumed that the ChEMBL compounds represent typical drug-like compounds, and the Tox21 compounds represent a subset of the concerned environmental chemicals. In the big data era of toxicology, although integration and utilization of the legacy data from the medicinal chemistry and drug discovery is emphasized,[19] less effort has been made to characterize essential differences between data landscapes of the drug-like legacy and the typical environmental chemicals. Herein, obvious discrepancy in the efficacy and potency/affinity has been observed between the PPARγ agonists from the drug design industry and from HTS projects. Such discrepancy is a reminder that prudence (such as strict applicability domains) is required for models based on drug-like compounds to be reliably applied to environmental chemicals.

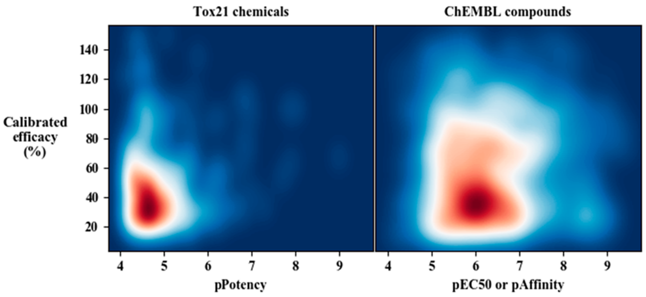


Figure 2. Distribution of reported efficacy and potency/binding affinity (as negative logarithm values) of the Tox21 and ChEMBL compounds with detected PPARγ agonist activity (Red, white, blue color map: high to low density)

Classifiers and CV Performance. Classifiers based on the non-3D Mordred descriptors, ECFP4 fingerprints, and MACCS keys achieved average stratified five-fold CV AUCs of 0.956 ± 0.002, 0.969 ± 0.001, and 0.960 ± 0.002, respectively, demonstrating that these classifiers were superb at identifying PPARγ agonists out of the documented drug-like compounds. Meanwhile, CV AUCs of the classifiers based on the randomly permuted data sets dropped to around 0.5, indicating that the high CV performance was not obtained by chance. Previous QSAR studies on PPARγ agonist activity typically cover a limited number of structural congeners,[45,46] thus restricting their applicability. This study curated more than 1700 compounds, yielding a larger and more inclusive data set for modeling PPARγ agonist activity, thus allowing much broader applicability.

Application of Classifiers to EHT Set. As shown in Table 1, applying the non-3D Mordred descriptors-based classifiers to all the EHT compounds resulted in AUCs approximating 0.8. Compared with the non-3D Mordred descriptors-based classifiers, the ECFP4 fingerprints or MACCS keys based classifiers performed poorly on the EHT set. Interestingly, the fingerprints based classifiers slightly outperformed the non-3D Mordred descriptors-based classifiers in the CV in terms of AUCs. However, their performances on the environmental compounds were reversed. The most contributed non-3D Mordred descriptors are topochemical indices such as BCUT

Table 1. Performance [Area under Receiver Operating Characteristic Curve (AUC)] of Classifiers on Environmental Health and Toxicology (EHT) Compounds

compounds selected by		classifiers based on different descriptors		
		non-3D Mordred descriptors	ECFP4	MACCS keys
no ADs	(*n* = 40)	0.792 ± 0.019	0.691 ± 0.015	0.698 ± 0.014
Euclidean distance AD	(*n* = 25)	0.876 ± 0.020		
AD_{ECFP4} {1, 0.26}	(*n* = 21)	0.883 ± 0.023	0.740 ± 0.014	0.867 ± 0.018
AD_{MACCS_keys} {1, 0.69}	(*n* = 13)	0.888 ± 0.015	0.524 ± 0.032	0.881 ± 0.018

Table 2. Performance [Area under Receiver Operating Characteristic Curve (AUC)] of Classifiers on Tox21 Compounds

compounds selected by		AUCs of classifiers based on		
		non-3D Mordred descriptors	ECFP4	MACCS keys
no ADs	(*n* = 6342)	0.693 ± 0.005	0.648 ± 0.003	0.569 ± 0.003
Euclidean distance AD	(*n* = 5824)	0.681 ± 0.005		
AD_{ECFP4} {1, 0.26}	(*n* = 2858)	0.695 ± 0.004	0.643 ± 0.005	0.612 ± 0.002
AD_{ECFP4} {10, 0.26}	(*n* = 1165)	0.752 ± 0.005	0.663 ± 0.006	0.685 ± 0.006
AD_{ECFP4} {1, 0.55}	(*n* = 47)	0.862 ± 0.007	0.914 ± 0.004	0.917 ± 0.006
AD_{ECFP4} {2, 0.55}	(*n* = 23)	0.907 ± 0.008	0.976 ± 0.004	0.971 ± 0.008
AD_{MACCS_keys} {1, 0.69}	(*n* = 1186)	0.684 ± 0.005	0.680 ± 0.007	0.550 ± 0.004
AD_{MACCS_keys} {10, 0.69}	(*n* = 337)	0.823 ± 0.005	0.832 ± 0.011	0.758 ± 0.007
AD_{MACCS_keys} {1, 0.85}	(*n* = 62)	0.907 ± 0.007	0.882 ± 0.008	0.826 ± 0.011
AD_{MACCS_keys} {2, 0.85}	(*n* = 40)	0.950 ± 0.006	0.946 ± 0.009	0.946 ± 0.015

(Burden - CAS - University of Texas eigenvalues)[47] and centered Moreau-Broto autocorrelation[48] descriptors. These descriptors give numeric properties of molecular topology.[16] On the other hand, the fingerprints describe the presence of local fragment patterns or specific substructures.[16,37] The results suggest that when the CV AUCs of classifiers constructed using different descriptor systems are comparable, the descriptors representing the whole molecule properties provide better generalizing capability than the local binary fingerprints.

With different ADs, the non-3D Mordred descriptors-based and MACCS keys based classifiers achieved average AUCs over 0.86 (Table 1), which is comparable to the CV AUCs of previously reported Tox21-databased classifiers.[32,49,50] Exceptionally, the ECFP4 fingerprints-based classifiers performed much worse on the compounds selected by AD_{MACCS_Keys}{1, 0.69} than on all 40 compounds of the EHT set. ECFP4 fingerprints theoretically can describe an unlimited number of specific fragment patterns, which were developed specifically for structure−activity modeling.[37] However, drug-like PPARγ agonists rarely have heavily halogenated fragments that are commonly found in environmental chemicals. Thus, the poly halogenated compounds tend to be predicted as nonagonists. On the other hand, ECFP4 patterns from phthalate frames are commonly found in drug-like PPARγ agonists. For example, 59 out of 66 ChEMBL compounds similar to (Tanimoto coefficient >0.26) di(2-ethylhexyl)phthalate (CASRN 117−81−7) are PPARγ agonists. Indeed, the phthalate derivatives tend to be predicted as PPARγ agonists. Such biases in training lead to false predictions by the ECFP4 fingerprints-based classifiers. On the contrary, MACCS keys identify only 167 general structural patterns, which are coarser than the ECFP4 fingerprints. However, this coarseness in turn endows MACCS keys with better generalizing capacity than ECFP4 fingerprints.

ADs are essential for the regulatory applications of QSAR models.[33] Although classic AD characterization methods[17] have been proved useful for regression tasks, it is largely obscure that whether these ADs would function well for classification tasks. In fact, a few studies on the Tox21-based classifiers temporarily omitted the discussion on the associated ADs. Meanwhile, some studies found that applying ADs did not show improvement in classification accuracy.[51] Here, the obvious discrepancy between the drug-like compounds and the typical environmental chemicals provides a realistic application scenario for investigating the functionality of ADs. Our results demonstrated the compatibility between the descriptor systems and the AD characterization methods: improper combination of them may result in unexpected bad performance of the classifiers.

Application of Classifiers to Tox21 Set. Applying the non-3D Mordred descriptors based classifiers on the Tox21 compounds resulted in AUCs approximating 0.7, which is significantly lower than those on EHT compounds. As the size of the Tox21 set is large, it is expected that performance of the classifiers was improved with the ADs. On the other hand, the fingerprints-based classifiers showed worse performance than the non-3D Mordred descriptors-based classifiers (Table 2), which is consistent with the case of applying the classifiers to the EHT set.

The AD in Euclidean distance of the non-3D Mordred descriptors did not improve the performance of the classifiers. Euclidean distance integrates the descriptors of a compound into a single distance from the center of the training-set descriptor space. Chemicals are considered inside the AD if they locate inside the hyper-dimensional sphere defined by the cutoff radius. Thus, the non-3D Mordred descriptors-based AD in Euclidean distance, unlike the fingerprints-based ADs, lacks the power to distinguish delicate feature patterns of chemicals.

Unexpectedly, prediction performance on the chemicals in AD_{ECFP4}{1, 0.26} and AD_{MACCS_keys}{1, 0.69} were not satisfactory for the Tox21 set (Table 2). However, both ADs demonstrated their usefulness for the EHT set. There are 39 common compounds in the EHT and Tox21 compounds sets (Table S6), of which 15% have major inconsistency and 36% have minor inconsistency in the activity. Examples of the major inconsistency include 2,4,6-tribromophenol (CASRN 118−79−6), 2,2′,4,4′-tetrabromodiphenyl ether (CASRN 5436−43−1), diisobutyl phthalate (CASRN 84−69−5), and benzyl butyl phthalate (CASRN 85−68−7). These compounds demonstrated PPARγ activation in a reporter gene assay using the HEK (human embryonic kidney)-293H cell line[10] but inactive in the Tox21 HTS assay using the same cell line. Examples of the minor inconsistency include tetrabromobisphenol A and tetrachlorobisphenol A, which were reported as PPARγ activators in HEK-293H and HGELN cell lines, and induced adipogenesis-related gene expression in 3T3-L1 cells.[10,11,52] The Tox21 project, however, categorized their activity as inconclusive. The inconsistency in the experimental PPARγ agonist activity data for the same compounds indicates possible quality issues in the assays and introduces uncertainty in performance evaluation of the established classifiers.

Nonetheless, it is believed that qualitative activity conclusions for the majority of the Tox21 chemicals can be consistently extrapolated to the general experimental systems. Indeed, performances of the classifiers were significantly boosted by using stricter ADs (such as increasing $N_{\min}$ or S_{cutoff}). Especially, utilizing 0.55 for similarity in ECFP4 fingerprints or 0.85 for similarity in MACCS keys as S_{cutoff}, which was suggested by Stumpfe and Bajorath for visibly similar structures for pairwise comparison,[53] resulted in extremely accurate prediction of the PPARγ agonists (Table 2). Thus, the established classifiers could do a superb job at predicting environmental compounds whose chemical structures are highly similar to the drug-like compounds in the training set. In general, the drug-like compounds represent a unique chemical structure space, which can be considered as a subspace in the complete space of environmental chemicals. Hence, the established classifiers would make promising complements to the existing QSAR models of the PPARγ agonist activity.

IMPLICATION

In the 21st century, toxicology has embraced a paradigm shift that focuses on human cell-based *in vitro* tests rather than *in vivo* animal tests.[54] In 2010, the adverse outcome pathway framework was proposed to organize key toxicological events of varied spatial levels related to apical endpoints for risk assessment.[55] The transactivation of PPARγ is a molecular initiating event that is linked to multiple adverse outcomes including obesity, liver steatosis, etc.[26,56,57] Thus, *in silico* screening of potential environmental PPARγ agonist gains insight into the assessment of health risks.

In medicinal chemistry and drug discovery, a large amount of PPARγ agonist activity data of drug-like compounds have been generated, providing a potential data set for QSAR modeling. However, various experimental factors lead to data heterogeneity, which can affect the quality of the data set. This study demonstrated that despite the heterogeneity (reflected by various factors, e.g., cell lines), it is possible to make categorized data sets using an evidence based discretizing scheme. It is anticipated that this scheme can be extrapolated to heterogeneous data sets of other functional biomolecules.

The discrepancy in both efficacy and binding affinity between drug-like and environmental PPARγ agonists implies the importance of ADs in application of the classifiers trained by the data of drug-like compounds to the environmental chemicals. The usefulness of ADs could be determined by AD characterization methods (e.g., the fingerprint based ADs as proposed in this study and the Euclidean distance-based methods) and descriptor systems (e.g., ECFP4 fingerprints, MACCS keys, and non-3D Mordred descriptors). In fact, compared with the widely explored machine learning algorithms for classification, AD characterization methods for classifiers have been much less investigated and should be emphasized in the future.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.9b00498.

> Final ChEMBL data set; curated EHT set; simple inorganic and organic components/salts; categorized data set obtained with discretizing scheme; main parameters for initializing "RandomForestClassifier" instance; intersection-set compounds of EHT set and Tox21 set; Python codes for data manipulation, modeling, and applicability domain characterization can be requested by e-mailing the corresponding author (XLSX)

AUTHOR INFORMATION

Corresponding Author

Jingwen Chen − *Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China;* orcid.org/0000-0002-5756-3336; Phone: +86-411-84706269; Email: jwchen@dlut.edu.cn; Fax: +86-411-84706269

Authors

Zhongyu Wang − *Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Huixiao Hong − *National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.9b00498

Notes

The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was supported by the National Key R&D Program of China (2018YFE0110700, 2018YFC1801604) and the National Natural Science Foundation of China (21661142001).

REFERENCES

(1) Lehrke, M., and Lazar, M. A. (2005) The many faces of PPAR gamma. *Cell 123*, 993−999.

(2) Marciano, D. P., Kuruvilla, D. S., Boregowda, S. V., Asteian, A., Hughes, T. S., Garcia-Ordonez, R., Corzo, C. A., Khan, T. M., Novick, S. J., Park, H., Kojetin, D. J., Phinney, D. G., Bruning, J. B., Kamenecka, T. M., and Griffin, P. R. (2015) Pharmacological repression of PPAR gamma promotes osteogenesis. *Nat. Commun. 6*, 7443 DOI: 10.1038/ncomms8443.

(3) Al Sharif, M., Alov, P., Vitcheva, V., Pajeva, I., and Tsakovska, I. (2014) Modes-of-action related to repeated dose toxicity: Tissue-specific biological roles of PPAR gamma ligand-dependent dysregulation in nonalcoholic fatty liver disease. *PPAR Res. 2014*, 1.

(4) Pillai, H. K., Fang, M., Beglov, D., Kozakov, D., Vajda, S., Stapleton, H. M., Webster, T. F., and Schlezinger, J. J. (2014) Ligand Binding and Activation of PPAR gamma by Firemaster (R) 550: Effects on Adipogenesis and Osteogenesis in Vitro. *Environ. Health Perspect. 122*, 1225−1232.

(5) Holtcamp, W. (2012) Obesogens An Environmental Link to Obesity. *Environ. Health Perspect. 120*, A62−A68.

(6) Grün, F., Watanabe, H., Zamanian, Z., Maeda, L., Arima, K., Chubacha, R., Gardiner, D. M., Kanno, J., Iguchi, T., and Blumberg, B. (2006) Endocrine-disrupting organotin compounds are potent inducers of adipogenesis in vertebrates. *Mol. Endocrinol. 20*, 2141−2155.

(7) Chamorro-Garcia, R., Shoucri, B. M., Willner, S., Kach, H., Janesick, A., and Blumberg, B. (2018) Effects of perinatal exposure to dibutyltin chloride on fat and glucose metabolism in mice, and molecular mechanisms, in vitro. *Environ. Health Perspect. 126*, 057006.

(8) Pereira-Fernandes, A., Vanparys, C., Hectors, T. L. M., Vergauwen, L., Knapen, D., Jorens, P. G., and Blust, R. (2013) Unraveling the mode of action of an obesogen: Mechanistic analysis of the model obesogen tributyltin in the 3T3-L1 cell line. *Mol. Cell. Endocrinol. 370*, 52−64.

(9) Feige, J. N., Gelman, L., Rossi, D., Zoete, V., Metivier, R., Tudor, C., Anghel, S. I., Grosdidier, A., Lathion, C., Engelborghs, Y., Michielin, O., Wahli, W., and Desvergne, B. (2007) The endocrine disruptor monoethyl-hexyl-phthalate is a selective peroxisome proliferator-activated receptor gamma modulator that promotes adipogenesis. *J. Biol. Chem. 282*, 19152−19166.

(10) Fang, M., Webster, T. F., and Stapleton, H. M. (2015) Activation of Human Peroxisome Proliferator-Activated Nuclear Receptors (PPAR gamma 1) by Semi-Volatile Compounds (SVOCs) and Chemical Mixtures in Indoor Dust. *Environ. Sci. Technol. 49*, 10057−10064.

(11) Riu, A., Grimaldi, M., le Maire, A., Bey, G., Phillips, K., Boulahtouf, A., Perdu, E., Zalko, D., Bourguet, W., and Balaguer, P. (2011) Peroxisome Proliferator-Activated Receptor gamma Is a Target for Halogenated Analogs of Bisphenol A. *Environ. Health Perspect. 119*, 1227−1232.

(12) Li, C.-H., Ren, X.-M., Ruan, T., Cao, L.-Y., Xin, Y., Guo, L.-H., and Jiang, G. (2018) Chlorinated polyfluorinated ether sulfonates exhibit higher activity toward peroxisome proliferator-activated receptors signaling pathways than perfluorooctanesulfonate. *Environ. Sci. Technol. 52*, 3232−3239.

(13) Fang, M., Webster, T. F., and Stapleton, H. M. (2015) Effect-Directed Analysis of Human Peroxisome Proliferator-Activated Nuclear Receptors (PPAR gamma 1) Ligands in Indoor Dust. *Environ. Sci. Technol. 49*, 10065−10073.

(14) Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014) QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem. 57*, 4977−5010.

(15) Ekins, S. (2014) Progress in computational toxicology. *J. Pharmacol. Toxicol. Methods 69*, 115−140.

(16) Todeschini, R., and Consonni, V. (2009) *Molecular Descriptors for Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA.

(17) Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W. D., Veith, G., and Yang, C. H. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. *ATLA, Altern. Lab. Anim. 33*, 155−173.

(18) Ciallella, H. L., and Zhu, H. (2019) Advancing computational toxicology in the big data era by artificial intelligence: Data-driven and mechanism-driven modeling for chemical toxicity. *Chem. Res. Toxicol. 32*, 536−547.

(19) Zhu, H., Zhang, J., Kim, M. T., Boison, A., Sedykh, A., and Moran, K. (2014) Big data in chemical toxicity research: The use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol. 27*, 1643−1651.

(20) Attene-Ramos, M. S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R. J., Austin, C. P., Shinn, P., Simeonov, A., Tice, R. R., and Xia, M. (2013) The Tox21 robotic platform for the assessment of environmental chemicals - From vision to reality. *Drug Discovery Today 18*, 716−723.

(21) Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013) Improving the human hazard characterization of chemicals: A Tox21 update. *Environ. Health Perspect. 121*, 756−765.

(22) Huang, R., and Xia, M. (2017) Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Front. Environ. Sci. 5*, 3 DOI: 10.3389/fenvs.2017.00003.

(23) Vo, A. H., Van Vleet, T. R., Gupta, R. R., Liguori, M., and Rao, M. S. (2020) An overview of machine learning and big data for drug toxicity evaluation. *Chem. Res. Toxicol. 33*, 20.

(24) Janesick, A. S., Dimastrogiovanni, G., Vanek, L., Boulos, C., Chamorro-Garcia, R., Tang, W. Y., and Blumberg, B. (2016) On the Utility of ToxCast (TM) and ToxPi as Methods for Identifying New Obesogens. *Environ. Health Perspect. 124*, 1214−1226.

(25) Banks, A. S., McAllister, F. E., Camporez, J. P. G., Zushin, P.-J. H., Jurczak, M. J., Laznik-Bogoslavski, D., Shulman, G. I., Gygi, S. P., and Spiegelman, B. M. (2015) An ERK/Cdk5 axis controls the diabetogenic actions of PPAR gamma. *Nature 517*, 391−U581.

(26) Ahmadian, M., Suh, J. M., Hah, N., Liddle, C., Atkins, A. R., Downes, M., and Evans, R. M. (2013) PPAR gamma signaling and metabolism: the good, the bad and the future. *Nat. Med. 19*, 557−566.

(27) Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res. 40*, D1100−D1107.

(28) Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krueger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. P. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res. 42*, D1083−D1090.

(29) Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2017) The ChEMBL database in 2017. *Nucleic Acids Res. 45*, D945−D954.

(30) Breiman, L. (2001) Random forests. *Machine Learning 45*, 5−32.

(31) Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal Of Chemical Information And Computer Sciences 43*, 1947−1958.

(32) Capuzzi, S. J., Politi, R., Isayev, O., Farag, S., and Tropsha, A. (2016) QSAR modeling of Tox21 challenge stress response and nuclear receptor signaling toxicity assays. *Front. Environ. Sci. 4*, 3 DOI: 10.3389/fenvs.2016.00003.

(33) OECD (2007) *Guidance document on the validation of (quantitative) structure activity relationships (Q)SAR models, Technical Report for OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 69*, Organization for Economic Co-operation and Development, Paris.

(34) Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., Zhao, T., Austin, C. P., and Simeonov, A. (2016) Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat. Commun. 7*, 10425 DOI: 10.1038/ncomms10425.

(35) Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J., Xiao, J., Zhang, J., and Bryant, S. H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res. 38*, D255−D266.

(36) Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018) Mordred: a molecular descriptor calculator. *J. Cheminf. 10*, 4 DOI: 10.1186/s13321-018-0258-y.

(37) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742−754.

(38) (2005) *MACCS Structural Keys*, Symyx Software, San Ramon, CA.

(39) Jaccard, P. (1901) Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vandoise des Sciences Naturelles 37*, 547−579.

(40) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal Of Machine Learning Research 12*, 2825−2830.

(41) Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters 27*, 861−874.

(42) Bradley, A. P. (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition 30*, 1145−1159.

(43) Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal Of Chemical Information And Computer Sciences 44*, 1912−1928.

(44) Casey, W. M., Chang, X., Allen, D. G., Ceger, P. C., Choksi, N. Y., Hsieh, J.-H., Wetmore, B. A., Ferguson, S. S., DeVito, M. J., Sprankle, C. S., and Kleinstreuer, N. C. (2018) Evaluation and Optimization of Pharmacokinetic Models for in Vitro to in Vivo Extrapolation of Estrogenic Activity for Environmental Chemicals. *Environ. Health Perspect. 126*, 097001.

(45) Jian, Y., He, Y., Yang, J., Han, W., Zhai, X., Zhao, Y., and Li, Y. (2018) Molecular modeling study for the design of novel peroxisome proliferator-activated receptor gamma agonists using 3D-QSAR and molecular docking. *Int. J. Mol. Sci. 19*, 630.

(46) Vedani, A., Descloux, A.-V., Spreatico, M., and Ernst, B. (2007) Predicting the toxic potential of drugs and chemicals in silico: A model for the peroxisome proliferator-activated receptor-gamma (PPAR gamma). *Toxicol. Lett. 173*, 17−23.

(47) Stanton, D. T. (1999) Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *Journal Of Chemical Information And Computer Sciences 39*, 11−20.

(48) Moreau, G., and Broto, P. (1980) The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim. 4*, 359−360.

(49) Stefaniak, F. (2015) Prediction of compounds activity in nuclear receptor signaling and stress pathway assays using machine learning algorithms and low-dimensional molecular descriptors. *Front. Environ. Sci. 3*, 77 DOI: 10.3389/fenvs.2015.00077.

(50) Wu, L., Liu, Z., Auerbach, S., Huang, R., Chen, M., McEuen, K., Xu, J., Fang, H., and Tong, W. (2017) Integrating Drug's Mode of Action into Quantitative Structure-Activity Relationships for Improved Prediction of Drug-Induced Liver Injury. *J. Chem. Inf. Model. 57*, 1000−1006.

(51) Ribay, K., Kim, M. T., Wang, W., Pinolini, D., and Zhu, H. (2016) Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data. *Front. Environ. Sci. 4*, 12 DOI: 10.3389/fenvs.2016.00012.

(52) Riu, A., le Maire, A., Grimaldi, M., Audebert, M., Hillenweck, A., Bourguet, W., Balaguer, P., and Zalko, D. (2011) Characterization of Novel Ligands of ER alpha, Er beta, and PPAR gamma: The Case of Halogenated Bisphenol A and Their Conjugated Metabolites. *Toxicol. Sci. 122*, 372−382.

(53) Stumpfe, D., and Bajorath, J. (2012) Exploring activity cliffs in medicinal chemistry miniperspective. *J. Med. Chem. 55*, 2932−2942.

(54) Collins, F. S., Gray, G. M., and Bucher, J. R. (2008) Toxicology - Transforming environmental health protection. *Science 319*, 906−907.

(55) Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrrano, J. A., Tietge, J. E., and Villeneuve, D. L. (2010) Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem. 29*, 730−741.

(56) Allen, T. E. H., Goodman, J. M., Gutsell, S., and Russell, P. J. (2014) Defining molecular initiating events in the adverse outcome pathway framework for risk assessment. *Chem. Res. Toxicol. 27*, 2100−2112.

(57) Tsakovska, I., Al Sharif, M., Alov, P., Diukendjieva, A., Fioravanzo, E., Cronin, M. T. D., and Pajeva, I. (2014) Molecular modelling study of the PPAR gamma receptor in relation to the mode of action/adverse outcome pathway framework for liver steatosis. *Int. J. Mol. Sci. 15*, 7651−7666.