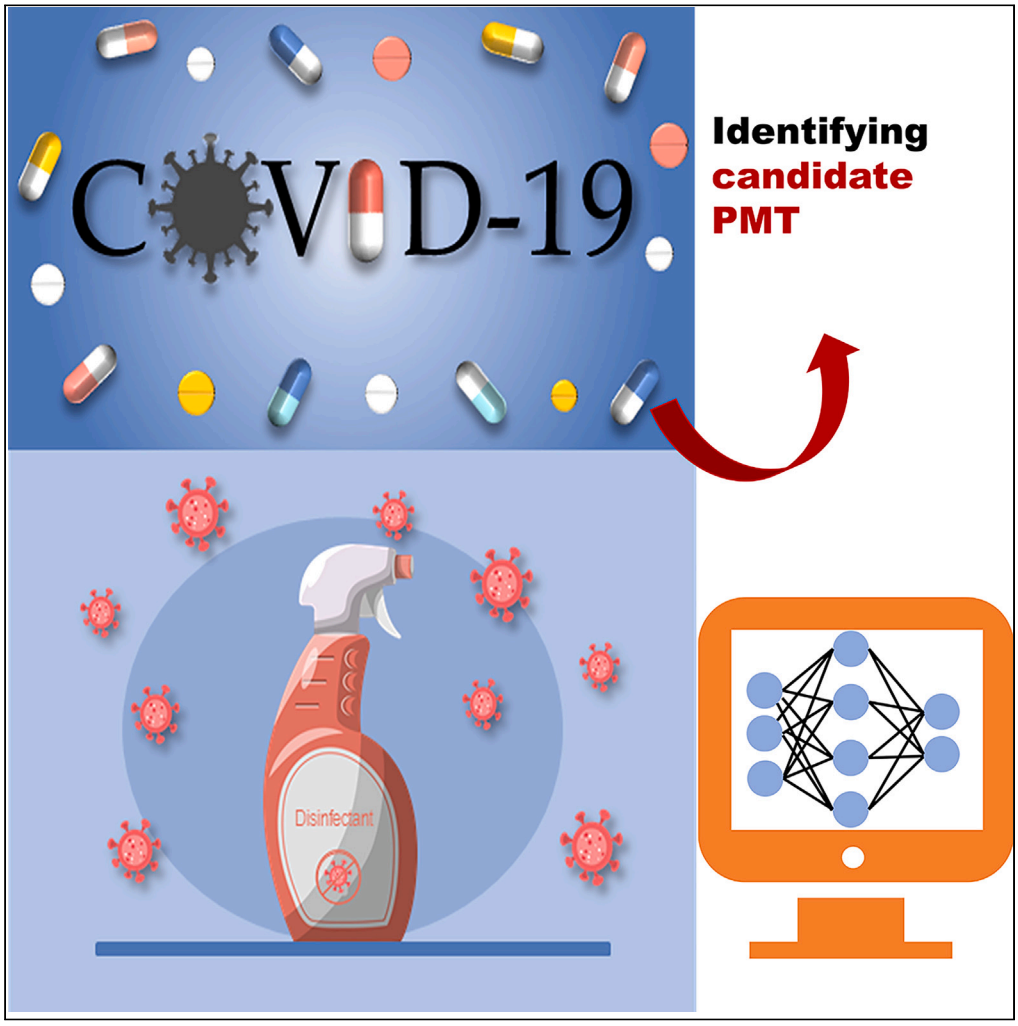
iScience

CellPress
OPEN ACCESS

Article

Machine learning coupled with causal inference to identify COVID-19 related chemicals that pose a high concern to drinking water



Min Han, Jun Liang, Biao Jin, Ziwei Wang, Wanlu Wu, Hans Peter H. Arp

jinbiao@gig.ac.cn

Highlights

PMT substances among COVID-19 related chemicals were identified by machine learning

Causations between PMT properties and molecular descriptors were understood

Over 60% of the COVID-19 chemicals considered are candidate PMT substances

Han et al., iScience 27, 109012
February 16, 2024
https://doi.org/10.1016/j.isci.2024.109012

iScience

CellPress
OPEN ACCESS

Article

Machine learning coupled with causal inference to identify COVID-19 related chemicals that pose a high concern to drinking water

Min Han,[1,2,3,4] Jun Liang,[5] Biao Jin,[1,2,3,4,8,*] Ziwei Wang,[1,2,3] Wanlu Wu,[1,2,3] and Hans Peter H. Arp[6,7]

SUMMARY

Various synthetic substances were utilized in large quantities during the recent coronavirus pandemic, COVID-19. Some of these chemicals could potentially enter drinking water sources. Persistent, mobile, and toxic (PMT) substances have been recognized as a threat to drinking water resources. It has not yet been assessed how many COVID-19 related substances could be considered PMT substances. One reason is the lack of high-quality experimental data for the identification of PMT substances. To solve this problem, we applied a machine learning model to identify the PMT substances among COVID-19 related chemicals. The optimal model achieved an accuracy of 90.6% based on external test data. The model interpretation and causal inference indicated that our approach understood causation between PMT properties and molecular descriptors. Notably, the screening results showed that over 60% of the COVID-19 chemicals considered are candidate PMT substances, which should be prioritized to prevent undue pollution of water resources.

INTRODUCTION

The severe acute-respiratory syndrome coronavirus 2 (SARS-CoV-2) caused the pandemic of coronavirus disease COVID-19 across the globe.[1,2] Though much attention has been placed on impacts to human health, there has been comparatively less attention on the environmental impacts. One environmental aspect is the consumption and emission of numerous chemicals, including disinfectants, antivirals, and other auxiliary drugs, that were used to control the outbreaks.[3,4] During the pandemic, a large quantity of disinfectants were consumed in indoor and outdoor settings and during wastewater treatment, and later entered the environment, including surface water and groundwater.[5–7] The residual chlorine concentrations in certain freshwater lakes in China were observed to have increased up to 0.4 mg/L during February and March 2020, likely due to the increased consumption of pharmaceuticals.[8] A significant proportion of these pharmaceuticals are excreted from the human body and remained unaltered in wastewater and surface water.[5] For instance, Zhang et al. found that the pandemic resulted in a significant increase in concentrations of lopinavir and ritonavir in wastewater effluents and surface water.[8] Wastewater discharge from municipal areas and hospitals is a major environmental emission pathway for COVID-19 related chemicals,[3,9–11] and some of these emissions may have further polluted drinking water sources by lowering water quality.[4]

When it comes to drinking water quality, persistent, mobile, and toxic (PMT) substances are becoming increasingly prioritized.[12–14] PMT substances are hard to degrade and persist in wastewater and water environments for a long period after emissions and, based on their mobility in soils, sludges, and sediments, can travel long distances through natural and artificial barriers, eventually entering groundwater and drinking water sources.[15] Very recently, the European Commission adopted PMT and very persistent, very mobile (vPvM) as two new hazard categories as part of the Classification, Labeling, and Packaging (CLP) regulation.[15,16] For health professionals, water resources managers and wastewater engineers, identifying and prioritizing PMT substances among COVID-19 chemicals are of primary importance to better manage their emissions and impact on water quality. Reliable determination of PMT/vPvM substances depends on high-quality experimental data such as experimentally determined half-life values in water, soil or sediment, as well as experimentally determined organic carbon-water partitioning coefficients, K_{OC}.[15,17]

Given the laboratory data are often unavailable for many chemicals, "*in-silico*" screening tools are utilized to identify potential PMT/vPvM substances as a pre-screening or weight-of-evidence approach. In recent years, machine learning (ML) models have become a powerful tool

[1]State Key Laboratory of Organic Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China
[2]CAS Center for Excellence in Deep Earth Science, Guangzhou 510640, China
[3]University of Chinese Academy of Sciences, Beijing 10069, China
[4]Guangdong Provincial Key Laboratory of Environmental Protection and Resources Utilization, Guangzhou 510640, China
[5]School of Software, South China Normal University, Foshan 528225, China
[6]Norwegian Geotechnical Institute (NGI), P.O. Box 3930 Ullevaal Stadion, N-0806 Oslo, Norway
[7]Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway
[8]Lead contact
*Correspondence: jinbiao@gig.ac.cn
https://doi.org/10.1016/j.isci.2024.109012

to address complex environmental problems due to their powerful fitting abilities.[18,19] One emerging application of ML in environmental research is predicting chemical properties.[20–25] Such ML models were developed to identify the correlation between chemical properties and molecular structures. For instance, Sun et al.[21] and Wang et al.[23] established ML models for screening persistent, bio-accumulative, and toxic (PBT) substances with a comprehensive consideration of attributes for persistence (P), bioaccumulation (B) and toxicity (T).

Although satisfactory predictive performance could be realized, the currently available ML models have "black box" features, leading to opaque decision-making process and the obscuring of mechanistic causality.[26–28] Recently, many studies explored using ML models to mechanisms using the SHapley Additive exPlanations (SHAP) method.[28–32] Very few studies have tested whether ML models combined with molecular features could understand causality between molecular features and chemical properties. The ability to use this approach for diverse molecular processes remains an open question. When an ML model makes predictions based on spurious correlation, the model is not trustworthy in spite of satisfactory predictive performances. For instance, McCloskey et al.[33] found that ML models still learn spurious correlations despite achieving perfect accuracy on test datasets. In order to improve the credibility of the model, it is of primary importance to verify whether the model makes accurate property predictions, such as between "PMT" and "Not PMT", based on correct and causal mechanistic reasons.[34] Furthermore, ML models can potentially identify complex patterns that experts have overlooked. Thus, understanding prediction mechanisms of ML models aids in providing new insights to human decision makers.[35]

Toward this goal, the specific aim of this study is to use ML models and causal inference approaches to better understand causal mechanism for how molecular descriptors influence PMT properties. This is herein utilized to identify PMT substances among the COVID-19 substances. The specific goals are: (1) to develop machine learning model to predict candidate PMT substances; (2) to validate our model with expert judgment; (3) to analyze prediction mechanisms by global interpretation, local interpretation and feature interaction analysis based on the SHAP method; and (4) to explore causality between molecular features and PMT properties by combing the SHAP method and causal inference. The results are discussed in the context of how to better identify COVID-19 substances that may pose a threat to water resources during an outbreak, as well as the wider context of new substances introduced to the market in high volumes.

RESULTS AND DISCUSSION

Data description

The internal dataset for model training includes 132 PMT substances and 898 Not PMT substances which were reported in previous studies (see Table S1). Separately, 108 COVID-19 related substances were identified from the literature (Tables S2 and S3). The COVID-19 substances were categorized into three types based on their intended use: disinfectants,[36] antivirals,[37] and other auxiliary drugs (see Figure 1; section S1.1 for further description). The COVID-19 substances were each classified based on PMT assessment criteria (section S2.2), though only 32 of the 108 COVID-19 substances had available high-quality experimental data or weight-of-evidence data (see Table S4). Among these 32 substances, 22 were evaluated as "Not PMT" (including all 20 which were disinfectants), and 9 substances were evaluated "Potential PMT ++" and 1 "PMT" as shown in Table 1. Here, "Potential PMT++" represents potential PMT substances that have a high weight-of-evidence they likely fulfill the PMT criteria, though are lacking some experimental data for a more definitive classification.[13] An exemplary PMT substance used for COVID-19 is ibuprofen, a high-volume, anti-inflammatory drug with antipyretic functions, recommended by FDA as a medical treatment for management of COVID-19 symptoms.[38] Ibuprofen is hardly able to be metabolized completely by the human body, and later could enter wastewater and natural waters.[39,40] Therefore, the above-mentioned 10 PMT or Potential PMT ++ substances deserve special attention.

Model validation

The best model was selected among all the possible combinations of 4 molecular representations, 15 data balancing methods and 12 ML algorithms (see Method details, and Tables S5–S7). The models were evaluated based on two metrics (i.e., recall rate and balanced accuracy, see Method details) and only 4 models are simultaneously among the top 5% of the both metrics (see Table S8). The four models are: molecular descriptor (MD)-random undersampling (RU)-support vector machine with the radial basis function kernel (RSVM) (Model 1); MD-EasyEnsemble-linear support vector machine (LSVM) (Model 2); MACCS-RU-RSVM (Model 3); and MD-EasyEnsemble-XGBoost (Model 4). The data balancing methods of Model 1–4 are undersampling strategies, which balance the dataset by reducing the number of majority class samples (Not PMT) and thus greatly improve recall rate (the accuracy of minority class). Due to the priority of recall rate, we preferentially selected the model with the highest recall rate among the four models. Both Model 1 and Model 2 showed the highest recall of 86.4%, while Model 2 returned a higher balancing accuracy (78.2%) than Model 1 (see Figure 2A). Based on the results above, Model 2, combining MD, EasyEnsemble and LSVM, gave the best performance. EasyEnsemble could, in particular, effectively solve the problem of unbalanced data and reduce loss of information induced by undersampling (see Figure 1). In Figure 2B, Model 2 with 1804 MDs produced the highest recall rate (89.5%). All results of feature selections were summarized in Table S9. After hyperparameter optimization, Model 2 even achieved a better performance (recall of 90.2% and balancing accuracy of 79.6%). Furthermore, the area under curve (AUC) of receiver operating characteristic (ROC) curve reached a value of 0.87, indicating that Model 2 showed good overall predictive performances (see Figure 2C). It was therefore used for further analysis. Furthermore, the complexity of the best model was discussed in Supplemental information (Note S3).

A different strategy was also tested, which involved the construction of models based on sequential P, M, and T classification, as a three-step prediction process.[31] This was executed using the same training set. The data selection procedures of the three-step model are summarized in Supplemental Information (Section S1.4). The same model construction procedures and evaluation metrics (see Model establishment and evaluation in STAR Methods) were used for both the one-step and three-step models. The detailed information of model comparation and optimization for three-step model was presented in Supplemental information (Section S1.4; Figure S1; Tables S10–S12).

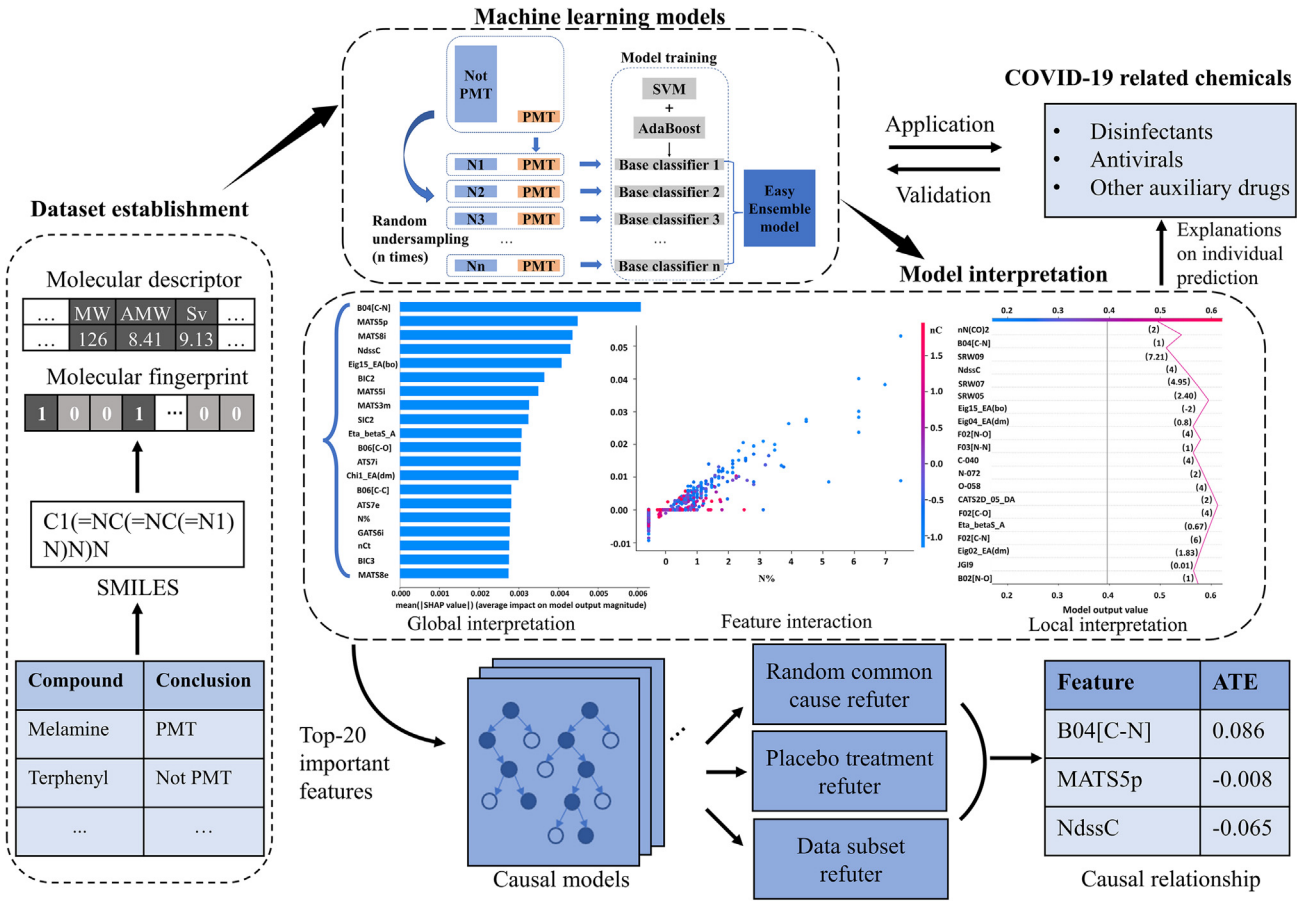


Figure 1. The workflow of dataset establishment, model construction using machine learning models, model application/validation, model interpretation and causal inference

Prior to applying our model to the target COVID-19 chemicals without experimental data, the prediction capability of the one-step model and the three-step model was tested by using the above-mentioned 32 chemicals. As shown in Table 1, the one-step model correctly identified 29 chemicals, while the three-step model correctly identified 27 chemicals. Specifically, the recall rate (90%) of the one-step model was higher than 70% of the three-step model. Besides, the one-step model resulted in higher balanced accuracy (90.5%), accuracy (90.6%), F-measure (85.7%) and precision (81.8%) in comparison with the three-step model (80.5%, 84.3%, 73.6%, and 77.8%). These results suggested that the one-step model achieved the better performance. Furthermore, previous models based on Quantitative Structure-Activity Relationship (QSAR) were calculated in parallel for the target chemicals for comparison (details of QSAR are provided in the STAR Methods and Note S2). Among the 32 chemicals, QSAR method correctly identified 26 chemicals, and the accuracy (81.3%) was lower than 90.6% of one-step model.

As false negatives for PMT substances are probably more problematic than false positives, we determined the best value for probabilistic decision cutoff such that it gives the highest recall rate (without affecting precision much) using a Precision-Recall curve. As shown in Figures 2E and 2F, the model resulted in the best performance when the cutoff was 0.495, achieving a recall rate of 100% and an accuracy of 93.5%. Furthermore, the highest value of the Youden index was achieved with a threshold value of 0.495.[41] Therefore, 0.495 was selected as the probabilistic decision cutoff. However, as a decrease in the cutoff value and the prioritization of recall rate would lead to an increase in false positives, the term "Potential PMT" was proposed to denote uncertain positive predictions. With this consideration, the model achieved the highest precision (100%) and accuracy (93.5%) when the cutoff was 0.548 (see Figures 2E and 2F). This suggests that there are no or few false positives in the validation set when the cutoff is 0.548. Therefore, this range (from 0.495 to 0.548) can be used to express the uncertainty of model predictions, and compounds that fall into this range could be concluded as "Potential PMT" based on this ML screening tool.

Global interpretation

As presented in Figure 1, the model interpretation involved three parts, including global interpretation, local interpretation and feature interaction analysis. Here the global explanations aim to summarize the relevance between input molecular features and predicted properties (i.e., PMT or Not PMT) based on SHAP values.[30] Figure 3A showed the 20 most important molecular features that affect model predictions. The

Table 1. Model predictions on Covid-19 related compounds (32 expert-verified compounds)

Chemical name	CAS number	Expert judgment	QSAR predicted	Three-step model	One-step model
Capric acid	334-48-5	Not PMT	Not PMT	Not PMT	Not PMT
Citric acid	77-92-9	Not PMT	Not PMT	Not PMT	Not PMT
Thymol	89-83-8	Not PMT	Not PMT	Not PMT	Not PMT
Ethanol (Ethyl alcohol)	64-17-5	Not PMT	Not PMT	Not PMT	Not PMT
Isopropanol (Isopropyl alcohol)	67-63-0	Not PMT	Not PMT	Not PMT	Not PMT
o-Phenylphenol	90-43-7	Not PMT	Potential PMT	PMT	PMT
Phenol	108-95-2	Not PMT	Not PMT	Not PMT	Not PMT
Glycolic acid	79-14-1	Not PMT	Not PMT	Not PMT	Not PMT
Hydrogen peroxide	7722-84-1	Not PMT	Not PMT	Not PMT	Not PMT
Octanoic acid	124-07-2	Not PMT	Not PMT	Not PMT	Not PMT
Peroxyacetic acid (Peracetic acid)	79-21-0	Not PMT	Not PMT	Not PMT	Not PMT
Glutaraldehyde	111-30-8	Not PMT	Not PMT	Not PMT	Not PMT
Tetraacetyl ethylenediamine	10543-57-4	Not PMT	Potential PMT	PMT	PMT
Triethylene glycol	112-27-6	Not PMT	Not PMT	Not PMT	Not PMT
C14 benzalkonium chloride	139-08-2	Not PMT	Not PMT	Not PMT	Not PMT
C16 benzalkonium chloride	122-18-9	Not PMT	Potential PMT	Not PMT	Not PMT
Dioctyl dimethyl ammonium chloride	5538-94-3	Not PMT	Not PMT	Not PMT	Not PMT
Didecyl dimethyl ammonium chloride	7173-51-5	Not PMT	Potential PMT	Not PMT	Not PMT
Hexadecyl trimethyl ammonium chloride	112-02-7	Not PMT	Not PMT	Not PMT	Not PMT
Tetradecyl trimethyl ammonium bromide	1119-97-7	Not PMT	Potential PMT	Not PMT	Not PMT
Ambroxol	18683-91-5	Potential PMT ++	Potential PMT ++	PMT	PMT
Amprenavir	161814-49-9	Potential PMT ++	Potential PMT ++	Not PMT	PMT
Baricitinib	1187594-09-7	Potential PMT ++	Potential PMT ++	PMT	PMT
Chlorpheniramine	113-92-8	Potential PMT ++	Potential PMT ++	Not PMT	PMT
Chlorpromazine	34468-21-8	Potential PMT ++	Potential PMT ++	PMT	PMT
Colchicine	64-86-8	Potential PMT ++	Potential PMT ++	PMT	PMT
Ibuprofen	15687-27-1	PMT	Not PMT	PMT	PMT
Remdesivir	1809249-37-3	Potential PMT ++	Potential PMT ++	Not PMT	PMT
Thalidomide	50-35-1	Potential PMT ++	Potential PMT	PMT	Not PMT
Tofacitinib	477600-75-2	Potential PMT ++	Potential PMT ++	PMT	PMT
Vitamin C (Ascorbic Acid)	50-81-7	Not PMT	Not PMT	Not PMT	Not PMT
Acetylcysteine	616-91-1	Not PMT	Not PMT	Not PMT	Not PMT

detailed information of these molecular features was summarized in supplemental information (Table S13). Specifically, the most important feature is B04[C-N] and the corresponding importance value is significantly higher than others (see Figure 3A). B04[C-N] is a 2D atom pair descriptor and denotes presence/absence of "C-N" at topological distance 4. When the fragment is present, the target compound is trending to be predicted as a PMT substance. This result could be due the occurrence of "C-N" contributes to impart the polarity in the molecules, and thus resulting in higher aqueous mobility,[42] and perhaps to some extent persistence.[43] Furthermore, the feature N% is a type of constitutional descriptor and denotes percentage of N atoms. In general, the molecular toxicity increases with the content of nitrogen atoms.[44] According to these results the N% is positively correlated with a PMT classification, which is consistent with the above-mentioned knowledge. In addition, the feature B06[C-C] denotes presence/absence of "C-C" at topological distance 6, which can be representative of hydrophobicity.[35] This is consistent with the knowledge learned by our model, where the presence of this fragment trends toward Not PMT predictions. The discussion of other molecular descriptors is presented in supplemental information (Note S5).

Given that the PMT classification is a combination of three properties, the three-step model was also interpreted by utilizing SHAP method to further explore the associations between PMT properties and molecular descriptors. Figure 3A summarized the effects of the above-mentioned top-20 molecular features on three properties (P, M, and T). Detailed information is presented in Figures S2–S21. In Figure 3A, the red arrows represent if the feature contributes positively on predicted property, and the blue arrows denote the negative influence. The absence of an arrow indicates that the feature exerts no influence on the property. The majority of the model interpretation results were consistent with

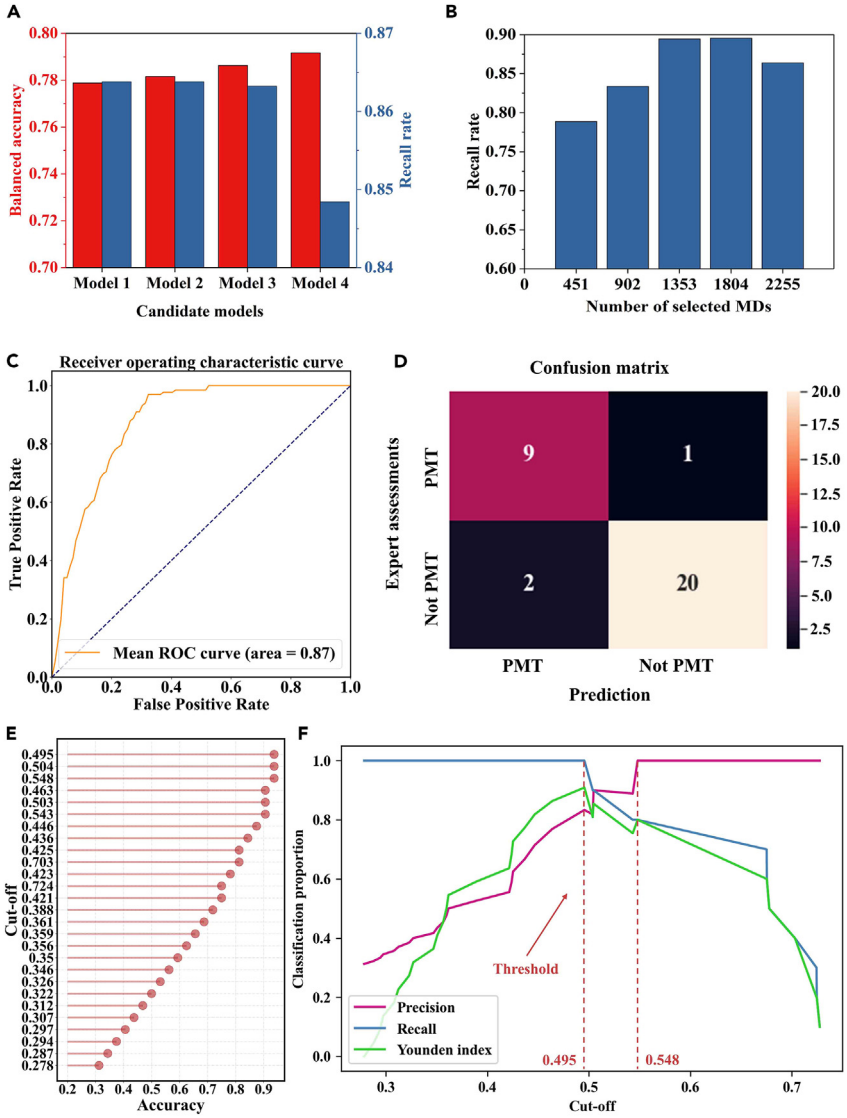


Figure 2. Model selection, feature selection, hyperparameter optimization and model calibration

(A) Recall rate and balanced accuracy of four candidate models (5-fold cross-validation).

(B) Recall rate of Model 2 with different numbers of selected molecular descriptors during feature selection.

(C) Receiver operating characteristic (ROC) curve and mean area under curve (AUC) of the final selected model (Model 2) with the optimal hyperparameter based on 5-fold cross-validation.

(D) Confusion matrix of Model 2 on external test set (COVID-19 related chemicals with available expert assessments).

(E) Accuracy plotted against cutoff.

(F) Youden index, precision and recall rate plotted against cutoff.

the domain knowledge. For instance, the feature B04[C-N] positively affected on P (persistence), and the feature N% exhibited positive contribution on T (toxicity). In addition, the model explanation offered novel insights that aid in comprehending the correlation between molecular descriptors and PMT characteristics. Specifically, the feature B04[C-N] also was positively correlated with T property, while the feature B06[C-C] had negative effects on P property. According to Figure 3A, we hypothesize that the relationship between the roles of molecular features in individual properties (P, M, and T) and their roles in the combined property (PMT) is that once a molecular feature has a negative contribution to any of the individual properties (P, M or T), the feature would have a negative contribution to the overall PMT classification.

Local interpretation

Global model interpretations cannot account for prediction mechanisms for individual compounds.[29] To better explore the influence of the molecular descriptors on model predictions at the local, individual compound level, we analyzed 32 compounds from the validation set.

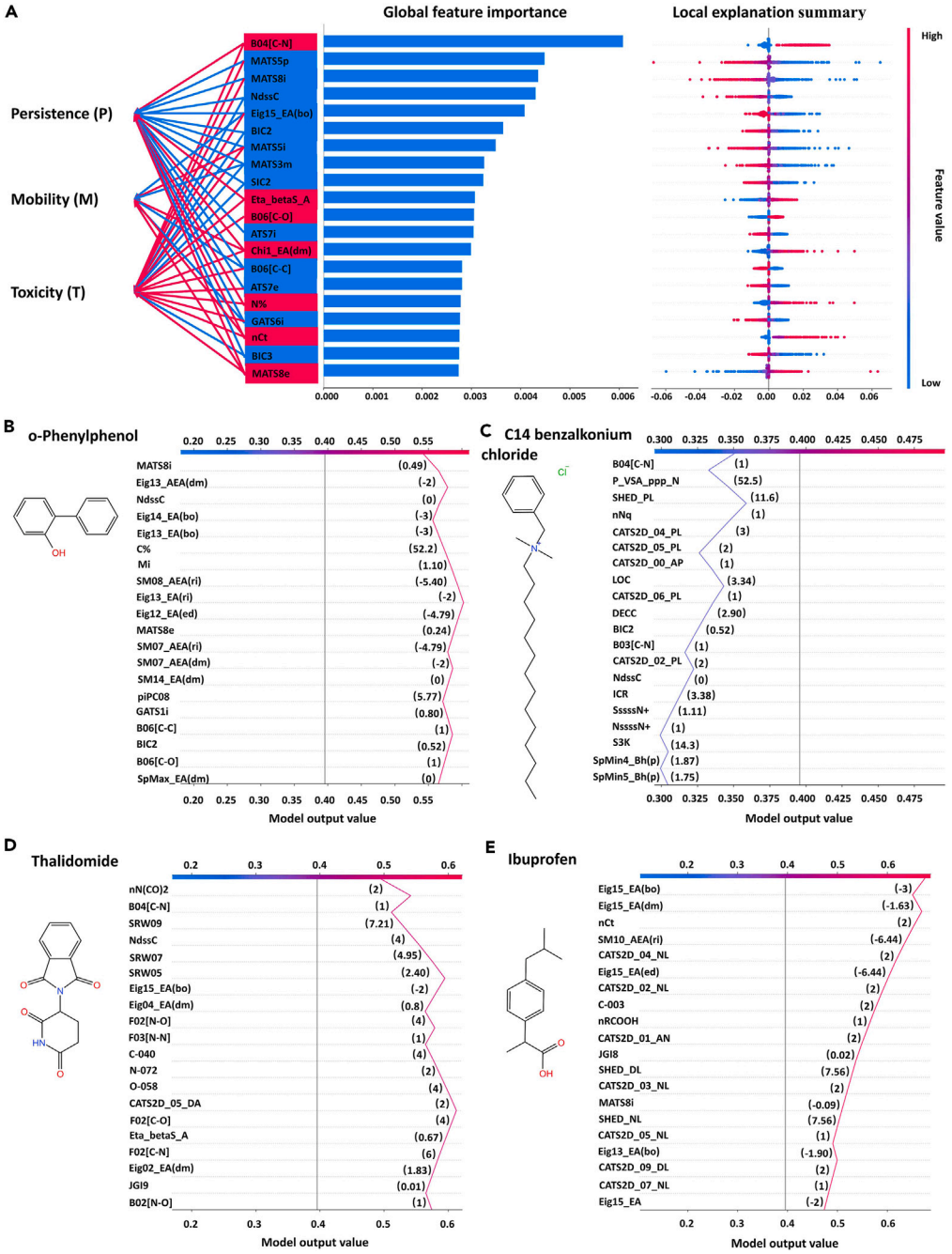


Figure 3. Model global and local interpretation

(A) Left: The effects of the top-20 molecular features on three properties (P, M, and T) individually. The red lines meant that the feature has a positive contribution on property and the blue lines represented the negative influence. The absence of an arrow indicates that the feature exerts no influence on the property. Median: Bar chart of the average impact on global model output magnitude, denoting feature importance values for our model. Right: Beeswarm plot of the model on the training dataset, presenting the magnitude, prevalence and direction of the above-mentioned molecular features' influences on model predictions by summarizing the multiple local explanations. The x axis represents SHAP values and the y axis denotes feature. SHAP >0 means that a specific feature has positive impact on the model predictions (i.e., tending to be PMT). Each dot refers to a compound and the color of the dots reflect the value of the feature. The redder color indicates a higher value of feature, and vice versa.

(B–E) Decision plots of our model prediction on *o*-Phenylphenol (false positive, B), and C14 benzalkonium chloride (true negative, C), and thalidomide (false negative, D), and ibuprofen (true positive, E). SHAP decision plots show how complex models arrive at their predictions.

Specifically, Figure 3 shows the local explanation of a true positive (ibuprofen), a false negative (thalidomide), a true negative (C14 benzalkonium chloride) and a false positive (*o*-Phenylphenol). The individual interpretations of the other 28 compounds are available in Supplemental information (Figures S22–S28). In the SHAP decision plots presented Figure 3, the x axis represents the model's output, where the larger the value is, the more possibly the compound is identified as a PMT substance. The compound is predicted as a PMT substance when the model output value is greater than 0.5, and vice versa for a Not PMT substance. The y axis of the SHAP decision plot (Figure 3) represents the 20 most important molecular features, which are sorted by descending importance. According to these SHAP decision plots, we counted the 20 most important features of each of the 32 compounds and sorted these features by number of occurrences in these compounds (Table S14). Notably, the top 5 features with the highest number of occurrences in compounds are B04[C-N], BIC2, NdssC, SIC2 and MATS8i (see Figure S29). These features belong to the top-20 important features of global interpretation, demonstrating the consistency between global and local explanations. Furthermore, these results suggested that the global interpretation prioritizes features that consistently influence most predictions.[27] However, it was also found that the top-20 molecular features of the individual compounds are not absolutely identical regarding the global importance ordering (see Figures 3 and S22–S28). This indicated that global interpretation derived from training data could not uncover feature patterns driving the prediction of individual compounds.

These results can be explained by the following reasons. First, the influence of molecular descriptors on the PMT classification differs for different values of the molecular descriptor. Specifically, B04[C-N] is ranked among the top 20 significant features for only 15 compounds, but not for all 32 compounds (see Figure S29). Significantly, the B04[C-N] values for all 15 compounds are consistently 1. Furthermore, of the remaining 17 compounds, 16 exhibit a B04[C-N] value of 0. The value of B04[C-N] of 1 represents the presence of "C-N" at topological distance 4 and vice versa. This suggests that the model might give priority to B04[C-N] when "C-N" at topological distance 4 is present. Similarly, the influence of BIC2 on model prediction is significant when the value of BIC2 deviates from a range of 0.6–0.8 (see Figures 3B, 3C, S22D, S23B, S23C, and S24B–S24D). In summary, we speculate that the mechanism of the model prediction is as follows. When making predictions for the PMT classification, the model would most likely prioritize the global important features. However, for certain compounds when the values of these features are approaching to the average of the training dataset or the feature fragments are absent, the global important features become less important.[27] Second, the impact of one feature on the PMT classification can be influenced by other features. Specifically, it was found that the effect of B04[C-N] on the PMT classification prediction for different compounds might vary despite these compounds possessing identical values for B04[C-N]. For instance, B04[C-N] significantly influences the PMT prediction of quaternary ammonium compounds (Figures 3C, S25B–S25D, S26A, and S26B), yet it does not rank among the top-20 important features in remdesivir (Figure S28B). In addition, the presence of multiple features with low global importance can markedly alter a single prediction. The molecular features of low global significance refer to those that do not rank within the top-20 most important features of global interpretation. For instance, although the most important feature has a positive effect on the PMT classification, the cumulative effects of the other features ultimately lead C14 benzalkonium chloride to have correctly received a Not PMT substance classification (see Figure 3C). These phenomena emphasize the importance of feature combination and might be caused by feature interaction (see the next section). In conclusion, global interpretations prioritize general feature correlations that account for many predictions, whereas local explanations allow for the examination of synergistic or compensatory effects that may not be globally apparent.[27]

Feature interaction analysis

The above-mentioned global and local model interpretation targeted the effect induced by a specific molecular feature when making model predictions; though did not address the issue of interactions between different molecular features. As shown in Figures 4A–4C, the SHAP values of the different compounds vary even when the feature values are the same. These results suggest that the SHAP value of a molecule descriptor could be affected by the values of other molecular descriptors within the substance. In other words, there are feature interactions between different molecular descriptors, which may also explain why feature importance orders vary by different compounds in Figures 3B–3E; from the physical chemistry literature, this is expected based on previous investigations on nonadditive effects on physicochemical properties due to intermolecular interactions of multifunctional molecules.[45] Figures 4D–4F depicts the interactions of the above-mentioned features (i.e., B04[C-N], B06[C-C] and N%) with other features with respect to the PMT/Not PMT prediction. Concerning the interaction between B04[C-N] and Chi0_AEA(ed) (Figure 4D), a higher Chi0_AEA(ed) value is always accompanied by smaller SHAP values when B04[C-N] is larger than 0. Chi0_AEA(ed) belongs to connectivity-like indices, molecular descriptors based on Kier-Hall Connectivity indices. These indices are calculated on the edge adjacency matrices by replacing the vertex degree with the edge degree. The edge degree is the number of edges adjacent to a given edge in the H-depleted molecular graph. Specifically, Chi0_AEA(ed) characterizes connectivity-like index of order 0 from augmented edge adjacency mat. weighted by edge degree. This suggests that lowering the Chi0_AEA(ed) values is moving model predictions away from a PMT classification when B04[C-N]>0. As for interaction of B06[C-C] and J_H2 (i.e., Balaban-like index from reciprocal squared distance matrix), a lower J_H2 leads to a positive impact on a Not PMT classification when the B06[C-C] > 0. According to Figure 4F, it is found that SHAP values are lowered along with higher nC (i.e., number of carbon atoms), which implies that increasing carbon atoms could weakens properties leading to a PMT classification when an N atom is present in a compound. In summary, there are interactions between different features and our model had already taken this influence into account.

Identification of the causation between PMT property and molecular descriptors

As discussed above, we aimed to interpret the relationship between molecular descriptors and PMT or Not PMT classifications by using the SHAP approach. Even though the SHAP method is efficient to explore the correlation between input features and model predictions, the

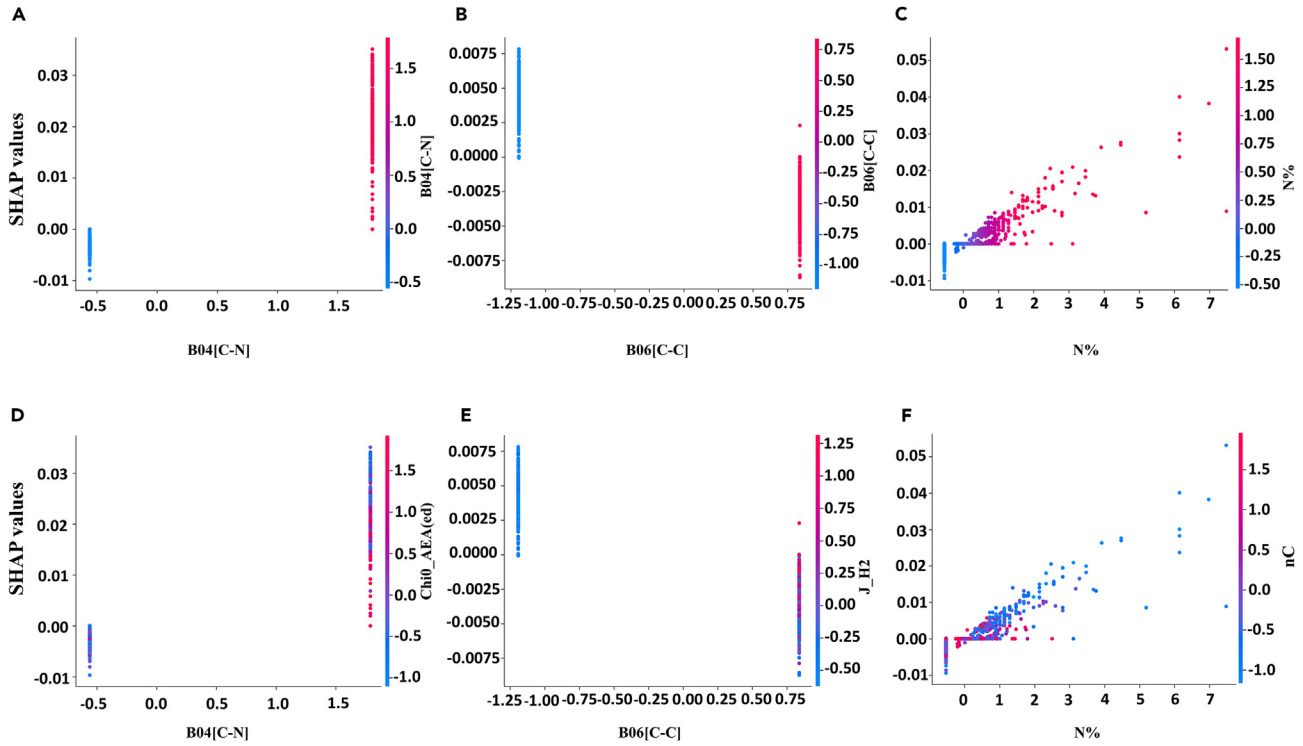


Figure 4. Feature interaction analysis

(A) SHAP dependence plot of B04[C-N] versus its SHAP value.

(B) SHAP dependence plot of B06[C-C] versus its SHAP value.

(C) SHAP dependence plot of N% versus its SHAP value.

(D) Plot of the SHAP interaction value of B04[C-N] with Chi0_AEA(ed).

(E) Plot of the SHAP interaction value of B06[C-C] with J_H2.

(F) Plot of the SHAP interaction value of N% with nC. The x axis is the value of the feature and the y axis is the SHAP value for that feature. The color corresponds to a second feature that may have an interaction effect with the feature.

correlations are not equivalent with causality due to mediation or confounding effects (see the SI Section S1.6).[46,47] In order to eliminate pseudo-correlation, causal models are built for each of the top 20 important features selected by SHAP to quantify the causal effects (average treatment effect or ATE). As presented in Table 2, the causal effects of 19 features were nonzero, indicating that these features causally associated with a PMT classification, albeit with large uncertainties.

To ensure the reliability of causal inference results, we tested the robustness of each of the above-mentioned causal relationships following the procedures presented in the Method Details (see Figure 1). Although the causal effects of some features were high, they failed to pass all the robustness tests. For instance, ATS7i holds the highest negative causal effect (−0.589) but only passes 1 test, indicating that the robustness of the causal relationship is insufficient. In the end, only 6 features (i.e., B04[C-N], MATS5p, NdssC, MATS5i, MATS3m, and MATS8e) went through all robustness tests. In particular, B04[C-N] and MATS5p, the top two features identified via SHAP, passed all robustness tests, resulting in causal effects of 0.086 and −0.008, respectively. This partly suggests that our model captured causation between the PMT classification and molecular descriptors. However, the twentieth-ranked feature MATS8e and the eighth-ranked MATS3m resulted the higher positive causal effect (0.118) and negative causal effect (−0.103) than the ones of the top two features. This could be explained by that MATS8e has positive contributions on three properties (P, M, and T) while B04[C-N] only positively contributes to P and T property (see Figure 3A). Similarly, MATS3m has negative effects on three properties (P, M, and T) but MATS5p only negatively correlated with P property (see Figure 3A). Notably, the contributions of MATS5i on PMT predictions based on SHAP and causal inference are opposite. This requires additional validation in future research. Based on the interpretations above, six features (B04[C-N], MATS5p, NdssC, MATS5i, MATS3m, and MATS8e) are likely to be causally related to PMT properties. The detailed information of the causation between Not PMT properties and molecular descriptors was presented in Note S7 and Table S15.

PMT substances among COVID-19 related chemicals

After conducting the model validation, model interpretation and causal inference, our approach was applied to the COVID-19 chemicals without expert judgments. An analysis of the applicability domain (AD) was conducted to evaluate if the model could be applied to the target

Table 2. Causal effect and refutation results of top 20 important molecular features

Feature	Definitions	[a]ATE	[b]Threshold of RCC 5%	Threshold of RCC 1%	Threshold of RCC 5‰	[c]PT	[d]DS
[e]**B04[C-N]**	Presence/absence of C - N at topological distance 4	0.086	[f]P	P	P	P	P
MATS5p	Moran autocorrelation of lag 5 weighted by polarizability	−0.008	P	P	P	P	P
MATS8i	Moran autocorrelation of lag 8 weighted by ionization potential	−0.002	P	P	P	P	[g]F
NdssC	Number of atoms of type dssC	−0.065	P	P	P	P	P
Eig15_EA(bo)	eigenvalue n. 15 from edge adjacency mat. weighted by bond order	−0.531	F	F	F	P	F
BIC2	Bond Information Content index (neighborhood symmetry of 2-order)	−0.096	F	F	F	P	F
MATS5i	Moran autocorrelation of lag 5 weighted by ionization potential	0.015	P	P	P	P	P
MATS3m	Moran autocorrelation of lag 3 weighted by mass	−0.103	P	P	P	P	P
SIC2	Structural Information Content index (neighborhood symmetry of 2-order)	−0.293	P	P	P	P	F
Eta_betaS_A	eta sigma average VEM count	0.151	P	P	F	P	P
B06[C-O]	Presence/absence of C - O at topological distance 6	−0.016	P	P	P	P	F
ATS7i	Broto-Moreau autocorrelation of lag 7 (log function) weighted by ionization potential	−0.589	F	F	F	P	F
Chi1_EA(dm)	connectivity-like index of order 1 from edge adjacency mat. weighted by dipole moment	−0.003	F	F	F	F	P
B06[C-C]	Presence/absence of C - C at topological distance 6	0.000	P	P	P	F	P
ATS7e	Broto-Moreau autocorrelation of lag 7 (log function) weighted by Sanderson electronegativity	0.240	P	F	F	P	F
N%	percentage of N atoms	0.027	F	F	F	P	P
GATS6i	Geary autocorrelation of lag 6 weighted by ionization potential	0.058	P	P	P	P	F
nCt	number of total tertiary C(sp3)	1.111	P	P	P	P	F
BIC3	Bond Information Content index (neighborhood symmetry of 3-order)	0.042	F	F	F	P	F
MATS8e	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity	0.118	P	P	P	P	P

[a]ATE: average treatment effect (quantification of causal effects). Specially, ATE >0 means that a specific feature has positive causal effect on PMT property and vice versa. When the absolute value of ATE for one feature is less than 0.01, there is little causal relationship between that feature and the PMT classification.
[b]Threshold of RCC: RCC refers to adding random common cause test and threshold represents the maximum allowed variation of an estimate.
[c]PT: placebo treatment test.
[d]DS: data subset refuter.
[e]features marked in bold represent passing all tests.
[f]P: pass.
[g]F: fail.

compounds. The AD was determined by evaluating the similarity between the target compound and those within the training dataset. Here, Euclidean distance was utilized to calculate the similarity between two compounds. Specifically, the smaller the Euclidean distance, the higher the similarity. The Euclidean distance between all target compounds and the compounds of training set were less than the threshold value of AD. Thus, all the target compounds were within applicability domain of our model. The detailed information of applicability domain is summarized in the SI (Section S1.8). In total 46 chemicals including 7 disinfecting chemicals, 25 antivirals and 14 other auxiliary drugs were predicted as "PMT" substances, accounting for 60.5% of the target COVID-19 chemicals (see Figure 5). Moreover, 13 "Potential PMT" substances were predicted. A recent study indicated that detection frequency and concentrations of specific chemicals such as azithromycin dramatically increased during the COVID-19 pandemic.[5] Validation of these PMT assessments using high-quality experimental data is recommended to confirm this assessment.

Conclusions

Clean drinking water is essential to human health.[48] Intensive applications of COVID-19 substances may pose risks to drinking water resources, as many of them could meet the PMT/vPvM hazard classification.[4] Limited experimental data and inefficient screening tools have hindered identification and prioritization of PMT substances.[49] Here, we developed a machine learning model to screen for PMT substances, and validated our model with expert judgment, fulfilling the first two aims of the study. The optimal model was validated and achieved an accuracy of 90.6% based on external test data, and later was applied on 76 different COVID-19 substances.

The third aim of the study was to analyze prediction mechanisms of our model by using global interpretation, local interpretation and feature interaction analysis based on SHAP methods. The results indicate that our model can implicitly acquire chemical knowledge. By combining the model interpretation of one-step and three-step models, we found that molecular features would have a negative contribution to a PMT classification once a molecular feature has a negative contribution to any of the properties P, M or T. Furthermore, the molecular features of low global importance may be influential due to the observation that combinations of different features together could lead to accumulative effects on model predictions, due to intermolecular interactions. Some potential causal relationships between chemical properties (PMT or Not PMT) and molecular descriptors were identified. The results indicated that B04 [C-N] (Presence/absence of C-N at topological distance 4) may be causally associated with both properties leading to PMT classification and Not PMT classification, which indicates that our model captured some causation. Identifying this causality between molecular features and PMT properties by combing the SHAP method and causal inference thereby addressed our fourth aim of the study. The prioritization of PMT substances identified in this study will facilitate risk-based management of substances produced in high-volume related to the COVID-19 pandemic. The identified COVID-19 substances that can be considered PMT substances are of relevance for further follow-up, including determination of the experimental persistence and mobility, as well as prioritization for monitoring and testing remediation technology, since their occurrence in natural water cycle could lead to negative impact on drinking water resources. It is therefore recommended that follow-up chemical property analysis, and monitoring studies should be conducted to investigate the relation between outbreaks and the occurrence of the identified COVID-19 PMT/vPvM substances in water resources. This would help ensure proper environmental management or appropriate policy development toward public and environmental health. The model developed here, along with our previous study[43] could further facilitate regulatory compliance to the new PMT/vPvM hazard classes in Europe,[16] as a way to rapidly screen new and existing chemicals being introduced to the market.

Limitations of the study

There are a few limitations of this study that should be highlighted. First, the underlying datasets to calibrate the model were limited, especially in terms of positive samples, resulting in insufficient training for the current model. In future work, the performance of our model will be enhanced with an increasing number of high-quality experimental data. In addition, further external validation through experimental testing could be employed to assess the model's generalizability in future studies. Second, while many types of molecular description methods were used to describe the structural information of compounds, molecular images and molecular graphs have not yet been applied. This could be done in future work. Third, some of the relevant confounders might remain uncaptured due to limited domain knowledge. In subsequent research, the integration of additional domain knowledge and sophisticated causal inference methodologies would be advantageous for a more comprehensive exploration of causality between P, M, and T properties with chemical structures.[43] PMT hazard assessments based solely on models have uncertainties and should be confirmed with high-quality experimental data. In the context of this later point, it important to mention this study was conducted prior to the official, European criteria of the PMT/vPvM criteria and hazard classes were established.[16] Therefore, the PMT substances herein should be considered candidate PMT/vPvM substances, pending a formal classification, as the model developed did not make differentiations between candidate PMT substances and those that are also candidate vPvM substances.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

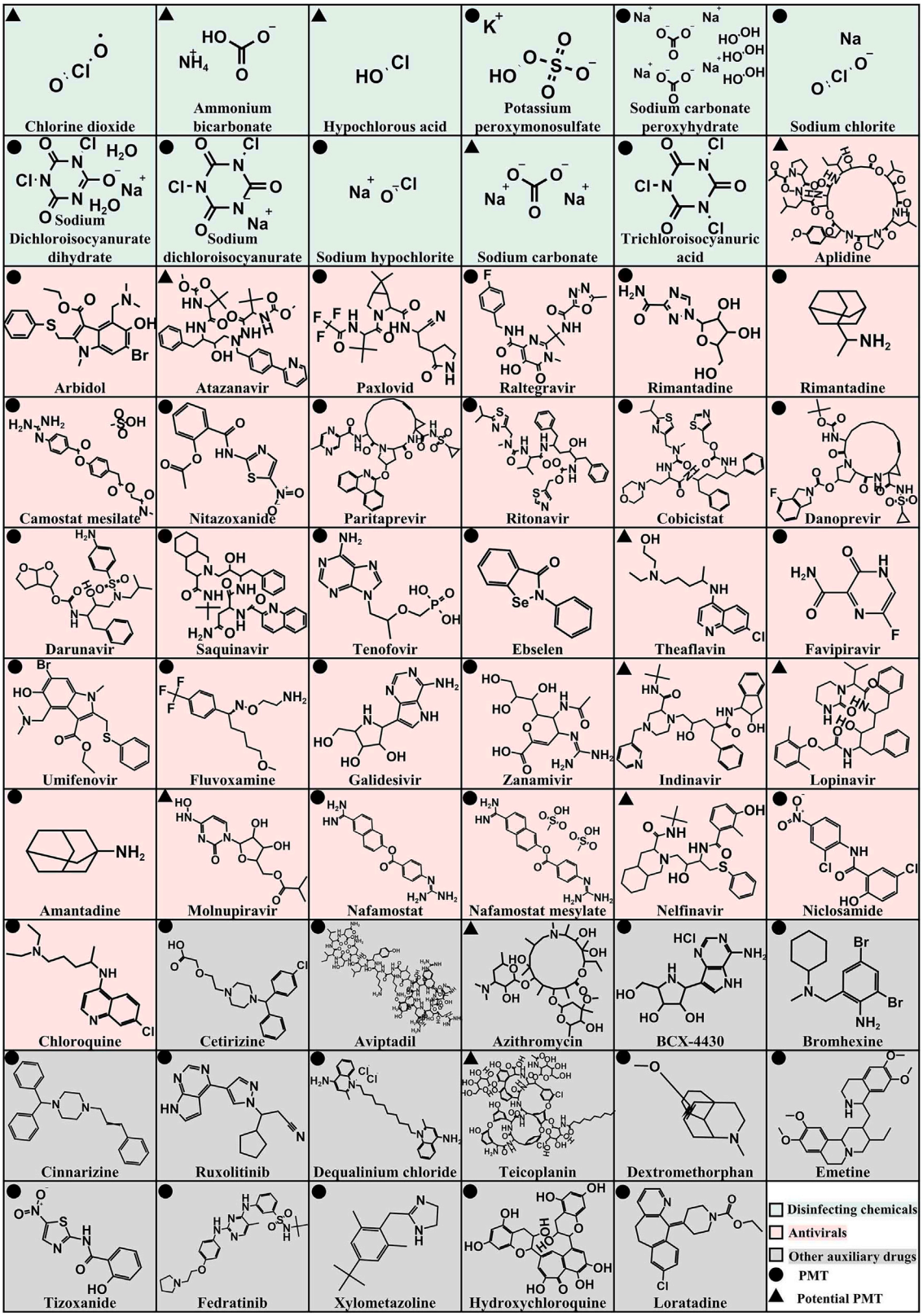


Figure 5. The candidate "PMT" substances and "Potential PMT" substances among COVID-19 related chemicals based on model prediction

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109012.

ACKNOWLEDGMENTS

This work was supported by Guangdong Major Project of Basic and Applied Basic Research (2023B0303000007). B.J. acknowledges funding from Foundation for Science and Technology Research (2023B1212060049); H.P.H.A. acknowledges funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101036756, ZeroPM.

AUTHOR CONTRIBUTIONS

Conceptualization, B.J. and H.P.H.A.; Methodology, M.H., B.J., and J.L.; Investigation, M.H., W.L.W., and Z.W.W.; Writing – Original Draft, M.H. and B.J.; Writing – Review and Editing, H.P.H.A. and B.J.; Funding Acquisition, B.J. and H.P.H.A.; Resources, B.J.; Supervision, B.J., J.L., and H.P.H.A.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 15, 2023
Revised: January 7, 2024
Accepted: January 22, 2024
Published: January 24, 2024

REFERENCES

1. Zheng, G., Filippelli, G.M., and Salamova, A. (2020). Increased Indoor Exposure to Commonly Used Disinfectants during the COVID-19 Pandemic. Environ. Sci. Technol. Lett. *7*, 760–765. https://doi.org/10.1021/acs.estlett.0c00587.
2. Kupferschmidt, K., and Cohen, J. (2020). Will novel virus go pandemic or be contained? Science *367*, 610–611. https://doi.org/10.1126/science.367.6478.610.
3. Li, D., Sangion, A., and Li, L. (2020). Evaluating consumer exposure to disinfecting chemicals against coronavirus disease 2019 (COVID-19) and associated health risks. Environ. Int. *145*, 106108. https://doi.org/10.1016/j.envint.2020.106108.
4. Zhang, H., Tang, W., Chen, Y., and Yin, W. (2020). Disinfection threatens aquatic ecosystems. Science *368*, 146–147. https://doi.org/10.1126/science.abb8905.
5. Chen, X., Lei, L., Liu, S., Han, J., Li, R., Men, J., Li, L., Wei, L., Sheng, Y., Yang, L., et al. (2021). Occurrence and risk assessment of pharmaceuticals and personal care products (PPCPs) against COVID-19 in lakes and WWTP-river-estuary system in Wuhan, China. Sci. Total Environ. *792*, 148352. https://doi.org/10.1016/j.scitotenv.2021.148352.
6. Kuroda, K., Li, C., Dhangar, K., and Kumar, M. (2021). Predicted occurrence, ecotoxicological risk and environmentally acquired resistance of antiviral drugs associated with COVID-19 in environmental waters. Sci. Total Environ. *776*, 145740. https://doi.org/10.1016/j.scitotenv.2021.145740.
7. Bandala, E.R., Kruger, B.R., Cesarino, I., Leao, A.L., Wijesiri, B., and Goonetilleke, A. (2021). Impacts of COVID-19 pandemic on the wastewater pathway into surface water: A review. Sci. Total Environ. *774*, 145586. https://doi.org/10.1016/j.scitotenv.2021.145586.
8. Zhang, Z., Zhou, Y., Han, L., Guo, X., Wu, Z., Fang, J., Hou, B., Cai, Y., Jiang, J., and Yang, Z. (2022). Impacts of COVID-19 pandemic on the aquatic environment associated with disinfection byproducts and pharmaceuticals. Sci. Total Environ. *811*, 151409. https://doi.org/10.1016/j.scitotenv.2021.151409.
9. Weinmann, T., Gerlich, J., Heinrich, S., Nowak, D., Mutius, E.v., Vogelberg, C., Genuneit, J., Lanzinger, S., Al-Khadra, S., Lohse, T., et al. (2017). Association of household cleaning agents and disinfectants with asthma in young German adults. Occup. Environ. Med. *74*, 684–690. https://doi.org/10.1136/oemed-2016-104086.
10. Dumas, O., Varraso, R., Boggs, K.M., Quinot, C., Zock, J.P., Henneberger, P.K., Speizer, F.E., Le Moual, N., and Camargo, C.A. (2019). Association of Occupational Exposure to Disinfectants With Incidence of Chronic Obstructive Pulmonary Disease Among US Female Nurses. JAMA Netw. Open *2*, e1913563. https://doi.org/10.1001/jamanetworkopen.2019.13563.
11. Bhat, S.A., Sher, F., Kumar, R., Karahmet, E., Haq, S.A.U., Zafar, A., and Lima, E.C. (2022). Environmental and health impacts of spraying COVID-19 disinfectants with associated challenges. Environ. Sci. Pollut. Res. Int. *29*, 85648–85657. https://doi.org/10.1007/s11356-021-16575-7.
12. Jin, B., Huang, C., Yu, Y., Zhang, G., and Arp, H.P.H. (2020). The Need to Adopt an International PMT Strategy to Protect Drinking Water Resources. Environ. Sci. Technol. *54*, 11651–11653. https://doi.org/10.1021/acs.est.0c04281.
13. Huang, C., Jin, B., Han, M., Yu, Y., Zhang, G., and Arp, H.P.H. (2021). The distribution of persistent, mobile and toxic (PMT) pharmaceuticals and personal care products monitored across Chinese water resources. Journal of Hazardous Materials Letters *2*, 100026.
14. Arp, H.P.H., Brown, T.N., Berger, U., and Hale, S.E. (2017). Ranking REACH registered neutral, ionizable and ionic organic chemicals based on their aquatic persistency and mobility. Environ. Sci. Process. Impacts *19*, 939–955. https://doi.org/10.1039/c7em00158d.
15. Arp, H.P.H., and Hale, S.E. (2022). Assessing the Persistence and Mobility of Organic Substances to Protect Freshwater Resources. ACS Environ. Au *2*, 482–509. https://doi.org/10.1021/acsenvironau.2c00024.
16. EU (2023). Delegated Regulation Amending Regulation 1272/2008 as Regards Hazard Classes and Criteria for the Classification, Labelling and Packaging of Substances and Mixtures.
17. Li, L., Zhang, Z., Men, Y., Baskaran, S., Sangion, A., Wang, S., Arnot, J.A., and Wania, F. (2022). Retrieval, Selection, and Evaluation of Chemical Property Data for Assessments of Chemical Emissions, Fate, Hazard, Exposure,

and Risks. ACS Environ. Au *2*, 376–395. https://doi.org/10.1021/acsenvironau.2c00010.

18. Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., et al. (2021). Machine learning: new ideas and tools in environmental science and engineering. Environ. Sci. Technol. *55*, 12741–12754. https://doi.org/10.1021/acs.est.1c01339.
19. Liu, X., Lu, D., Zhang, A., Liu, Q., and Jiang, G. (2022). Data-Driven Machine Learning in Environmental Pollution: Gains and Problems. Environ. Sci. Technol. *56*, 2124–2133. https://doi.org/10.1021/acs.est.1c06157.
20. Wang, Z., Chen, J., and Hong, H. (2021). Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. Environ. Sci. Technol. *55*, 6857–6866.
21. Xiangfei, S., Xianming, Z., Muir, D.C.G., and Zeng, E.Y. (2020). Identification of Potential PBT/POP-Like Chemicals by a Deep Learning Approach Based on 2D Structural Features. Environmental science & technology *54*.
22. Wang, L., Zhao, L., Liu, X., Fu, J., and Zhang, A. (2021). SepPCNET: Deeping Learning on a 3D Surface Electrostatic Potential Point Cloud for Enhanced Toxicity Classification and Its Application to Suspected Environmental Estrogens. Environ. Sci. Technol. *55*, 9958–9967. https://doi.org/10.1021/acs.est.1c01228.
23. Wang, H., Wang, Z., Chen, J., and Liu, W. (2022). Graph Attention Network Model with Defined Applicability Domains for Screening PBT Chemicals. Environ. Sci. Technol. *56*, 6774–6785. https://doi.org/10.1021/acs.est.2c00765.
24. Yang, Z., Luo, S., Wei, Z., Ye, T., Spinney, R., Chen, D., and Xiao, R. (2016). Rate constants of hydroxyl radical oxidation of polychlorinated biphenyls in the gas phase: A single−descriptor based QSAR and DFT study. Environ. Pollut. *211*, 157–164. https://doi.org/10.1016/j.envpol.2015.12.044.
25. Ye, T., Wei, Z., Spinney, R., Dionysiou, D.D., Luo, S., Chai, L., Yang, Z., and Xiao, R. (2017). Quantitative structure–activity relationship for the apparent rate constants of aromatic contaminants oxidized by ferrate (VI). Chem. Eng. J. *317*, 258–266. https://doi.org/10.1016/j.cej.2017.02.061.
26. Yu, F., Wei, C., Deng, P., Peng, T., and Hu, X. (2021). Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. Sci. Adv. *7*, eabf4130. https://doi.org/10.1126/sciadv.abf4130.
27. Rodríguez-Pérez, R., and Bajorath, J. (2021). Explainable Machine Learning for Property Predictions in Compound Optimization. J. Med. Chem. *64*, 17744–17752. https://doi.org/10.1021/acs.jmedchem.1c01789.
28. Zhong, S., Zhang, K., Wang, D., and Zhang, H. (2021). Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. Chem. Eng. J. *405*, 126627. https://doi.org/10.1016/j.cej.2020.126627.
29. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
30. Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. *2*, 573–584. https://doi.org/10.1038/s42256-020-00236-4.
31. Zhao, Q., Yu, Y., Gao, Y., Shen, L., Cui, S., Gou, Y., Zhang, C., Zhuang, S., and Jiang, G. (2022). Machine Learning-Based Models with High Accuracy and Broad Applicability Domains for Screening PMT/vPvM Substances. Environ. Sci. Technol. *56*, 17880–17889. https://doi.org/10.1021/acs.est.2c06155.
32. Wu, Z., Lei, T., Shen, C., Wang, Z., Cao, D., and Hou, T. (2019). ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches. J. Chem. Inf. Model. *59*, 4587–4601. https://doi.org/10.1021/acs.jcim.9b00801.
33. McCloskey, K., Taly, A., Monti, F., Brenner, M.P., and Colwell, L.J. (2019). Using attribution to decode binding mechanism in neural network models for chemistry. Proc. Natl. Acad. Sci. USA *116*, 11624–11629. https://doi.org/10.1073/pnas.1820657116.
34. Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.L. (2020). Evaluation of Methods for Causal Discovery in Hydrometeorological Systems. Water Resour. Res. *56*, e2020WR027251. https://doi.org/10.1029/2020WR027251.
35. Khan, K., Benfenati, E., and Roy, K. (2019). Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds. Ecotoxicol. Environ. Saf. *168*, 287–297. https://doi.org/10.1016/j.ecoenv.2018.10.060.
36. Li, Z., Song, G., Bi, Y., Gao, W., He, A., Lu, Y., Wang, Y., and Jiang, G. (2021). Occurrence and Distribution of Disinfection Byproducts in Domestic Wastewater Effluent, Tap Water, and Surface Water during the SARS-CoV-2 Pandemic in China. Environ. Sci. Technol. *55*, 4103–4114. https://doi.org/10.1021/acs.est.0c06856.
37. Ashour, N.A., Abo Elmaaty, A., Sarhan, A.A., Elkaeed, E.B., Moussa, A.M., Erfan, I.A., and Al-Karmalawy, A.A. (2022). A Systematic Review of the Global Intervention for SARS-CoV-2 Combating: From Drugs Repurposing to Molnupiravir Approval. Drug Des. Dev. Ther. *16*, 685–715. https://doi.org/10.2147/dddt.S354841.
38. Wu, Q., Coumoul, X., Grandjean, P., Barouki, R., and Audouze, K. (2021). Endocrine disrupting chemicals and COVID-19 relationships: A computational systems biology approach. Environ. Int. *157*, 106232. https://doi.org/10.1016/j.envint.2020.106232.
39. Tiwari, B., Sellamuthu, B., Piché-Choquette, S., Drogui, P., Tyagi, R.D., Vaudreuil, M.A., Sauvé, S., Buelna, G., and Dubé, R. (2019). The bacterial community structure of submerged membrane bioreactor treating synthetic hospital wastewater. Bioresour. Technol. *286*, 121362. https://doi.org/10.1016/j.biortech.2019.121362.
40. Goswami, P., Guruge, K.S., Tanoue, R., Tamamura, Y.A., Jinadasa, K.B.S.N., Nomiyama, K., Kunisue, T., and Tanabe, S. (2022). Occurrence of Pharmaceutically Active Compounds and Potential Ecological Risks in Wastewater from Hospitals and Receiving Waters in Sri Lanka. Environ. Toxicol. Chem. *41*, 298–311. https://doi.org/10.1002/etc.5212.
41. Li, K., Sun, R., and Guo, G. (2023). The rapid increase of urban contaminated sites along China's urbanization during the last 30 years. iScience *26*, 108124. https://doi.org/10.1016/j.isci.2023.108124.
42. Pandey, S.K., Ojha, P.K., and Roy, K. (2020). Exploring QSAR models for assessment of acute fish toxicity of environmental transformation products of pesticides (ETPPs). Chemosphere *252*, 126508. https://doi.org/10.1016/j.chemosphere.2020.126508.
43. Han, M., Jin, B., Liang, J., Huang, C., and Arp, H.P.H. (2023). Developing machine learning approaches to identify candidate persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances based on molecular structure. Water Res. *244*, 120470. https://doi.org/10.1016/j.watres.2023.120470.
44. Hossain, K.A., and Roy, K. (2018). Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in Dugesia japonica and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. Ecotoxicol. Environ. Saf. *166*, 92–101. https://doi.org/10.1016/j.ecoenv.2018.09.068.
45. Goss, K.-U., Arp, H.P.H., Bronner, G., and Niederer, C. (2009). Nonadditive effects in the partitioning behavior of various aliphatic and aromatic molecules. Environ. Toxicol. Chem. *28*, 52–60. https://doi.org/10.1897/08-189.1.
46. Kumar, R., Le, N., Oviedo, F., Brown, M.E., and Reineke, T.M. (2022). Combinatorial Polycation Synthesis and Causal Machine Learning Reveal Divergent Polymer Design Rules for Effective pDNA and Ribonucleoprotein Delivery. JACS Au *2*, 428–442. https://doi.org/10.1021/jacsau.1c00467.
47. Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.L. (2020). Evaluation of Methods for Causal Discovery in Hydrometeorological Systems. Water Resour. Res. *56*, e2020WR027251. https://doi.org/10.1029/2020WR027251.
48. Tortajada, C., and van Rensburg, P. (2020). Drink more recycled wastewater. Nature *577*, 26–28. https://doi.org/10.1038/d41586-019-03913-6.
49. Hale, S.E., Neumann, M., Schliebner, I., Schulze, J., Averbeck, F.S., Castell-Exner, C., Collard, M., Drmač, D., Hartmann, J., Hofman-Caris, R., et al. (2022). Getting in control of persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances to protect water resources: strategies from diverse perspectives. Environ. Sci. Eur. *34*, 22. https://doi.org/10.1186/s12302-022-00604-4.
50. Arp, H.P.H., and Hale, S.E. (2019). REACH: Improvement of Guidance Methods for the Identification and Evaluation of PM/PMT Substances (German Environment Agency (UBA). ISBN: 1862–4804), p. 130. UBA TEXTE 126/2019.
51. Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G. (2002). Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci. *42*, 1273–1280. https://doi.org/10.1021/ci010132r.
52. Rogers, D., and Hahn, M. (2010). Extended-Connectivity Fingerprints. J. Chem. Inf.

Model. *50*, 742–754. https://doi.org/10.1021/ci100050t.

53. Wassermann, A.M., Wawer, M., and Bajorath, J. (2010). Activity Landscape Representations for Structure–Activity Relationship Analysis. J. Med. Chem. *53*, 8209–8223. https://doi.org/10.1021/jm100933w.
54. Kar, S., Roy, K., and Leszczynski, J. (2018). Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. Methods Mol. Biol. *1800*, 141–169. https://doi.org/10.1007/978-1-4939-7899-1_6.
55. Garcia, V., Mollineda, R.A., and Sanchez, J.S. (2009). Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. held in Povoa de Varzim, PORTUGAL, Jun 10-12. pp. 441.
56. United States Environmental Protection Agency (2022). List N:Disinfectants for Coronavirus (COVID-19). https://www.epa.gov/pesticide-registration/list-n-disinfectants-coronavirus-covid-19.
57. Purohit, A., Kopferschmitt-Kubler, M.C., Moreau, C., Popin, E., Blaumeiser, M., and Pauli, G. (2000). Quaternary ammonium compounds and occupational asthma. Int. Arch. Occup. Environ. Health *73*, 423–427. https://doi.org/10.1007/s004200000162.
58. Dewey, H.M., Jones, J.M., Keating, M.R., and Budhathoki-Uprety, J. (2021). Increased Use of Disinfectants During the COVID-19 Pandemic and Its Potential Impacts on Health and Safety. ACS Chem. Health Saf. *29*, 27–38. https://doi.org/10.1021/acs.chas.1c00026.
59. Lundberg, S.M., and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. held in Long Beach, CA, Dec 04-09.
60. Chen, H., Covert, I.C., Lundberg, S.M., and Lee, S.-I. (2023). Algorithms to estimate Shapley value feature attributions. Nat. Mach. Intell. *5*, 590–601. https://doi.org/10.1038/s42256-023-00657-x.
61. Kang, Q., Song, X., Xin, X., Chen, B., Chen, Y., Ye, X., and Zhang, B. (2021). Machine Learning-Aided Causal Inference Framework for Environmental Data Analysis: A COVID-19 Case Study. Environ. Sci. Technol. *55*, 13400–13410. https://doi.org/10.1021/acs.est.1c02204.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Internal dataset	This paper	See Table S1
COVID-19 related compounds	This paper	See Tables S2–S4
Software and algorithms		
Python programming language, version 3.8.8	Python Software Foundation	https://www.python.org/
alvaDesc (software for molecular descriptors calculation) version 1.0.22, 2021	Alvascience	https://www.alvascience.com
Anaconda	Continuum Analytics	https://www.anaconda.com/
SHAP (SHapley Additive exPlanations)	Open-Source	https://shap-lrjball.readthedocs.io/en/latest/api.html
imbalanced-learn	Open-Source	https://imbalanced-learn.org/
EconML	Open-Source	https://econml.azurewebsites.net/
Dowhy	Open-Source	https://www.pywhy.org/dowhy/v0.8/getting_started/intro.html
Machine learning model	Han et al., 2023[43]	https://doi.org/10.1016/j.watres.2023.120470
Causal model	GitHub	https://github.com/6yunq6/hm_first/tree/master/causal

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Biao Jin (jinbiao@gig.ac.cn).

Materials availability

This study did not generate new unique materials.

Data and code availability

- The datasets generated during this study are available at Supplemental information and the key resources table.
- All original code has been deposited at our GitHub repository (https://github.com/6yunq6/hm_first/tree/master/causal) and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact on request.

METHOD DETAILS

The methods used here are largely based on our previous publication,[43] where a summary is presented here with an emphasis on unique differences to this study. Specifically, this study exclusively constructed a machine learning model for PMT substances rather than PMT and vPvM substances, and thus the training datasets are different. Secondly, the model prediction mechanism was interpreted by using both SHapley Additive exPlanations (SHAP) method and causal inference. Lastly, for the first time COVID-19 related chemicals were screened for PMT substances.

Data collection and preprocessing

The internal dataset with high-quality data is collected from the previous studies.[13,14,50] As presented in Figure 1, the canonical SMILES (i.e., simplified molecular input line-entry system) codes of these chemicals are obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov/). In order to better exact the chemical information of compounds, 4 molecular representations including two-dimensional (2D) molecular descriptors (MDs) and three molecular fingerprints (MFs) are compared. Specifically, the MDs with missing values and constant values are removed and finally 2255 MDs are selected. Given irrelevant and redundant features might cause overfitting problems, feature selection is applied on MDs. In order to unify the units and scales of different MDs, the descriptor values are standardized (Equation 1):

$$x^{*} = \frac{x - \mu}{\sigma} \quad \text{(Equation 1)}$$

where x* is the standardized value of the original molecular descriptor x; μ is the mean value of all molecular descriptors; and σ indicates the standard deviation of all molecular descriptors. Furthermore, MFs includes Molecular ACCess System Fingerprints[51] (MACCS), Extended Connectivity Fingerprints[52] (ECFP) and Path Fingerprints (PFP). The details of MFs are summarized in supplemental information (Table S5). These MDs and MFs are calculated by alvaDesc (version 1.0.22) based on SMILES.

Considering that our data are class-imbalanced (i.e. the number of Not PMT substances is much greater than the PMT substances), 15 common data balancing methods (Table S6, Note S9) are used to improve model performance. The above mentioned data balancing methods are realized using Python 3.8.8 with an imbalanced-learn 0.8.0 package.[53]

Model establishment and evaluation

Machine learning models are developed through the internal dataset. The internal dataset is divided into training data (80%) and validation data (20%), which contain same fraction of the positive (i.e., "PMT") and negative samples (i.e., "Not PMT"). During model establishment, a total of 12 ML algorithms (Table S7 and Note S10) are compared. The 12 ML algorithms are combined with the above-mentioned 4 molecular representations and 15 data balancing methods to develop machine learning models. Finally, the optimal combination of molecular representation, data balancing method and ML algorithm are selected based on matrices of performance evaluation (see below). The establishment of the applicability domain (AD) aims to evaluate if the model could be applied to a certain target compound.[54] The further description of feature selection, hyperparameter optimization and AD are summarized in Supplemental information (Section S1.8; Figure S30). The model training and test was accomplished on a Tianhe-2 Supercomputer.

After model training, the model performance was evaluated by common metrics such as accuracy, recall rate, precision, F-measure and balanced accuracy (Equations 2, 3, 4, 5, and 6):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad \text{(Equation 2)}$$

$$\text{Recall rate} = TP/(TP + FN) \quad \text{(Equation 3)}$$

$$\text{Precision} = TP/(TP + FP) \quad \text{(Equation 4)}$$

$$F\_measure = (2 \times \text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision}) \quad \text{(Equation 5)}$$

$$\text{Balanced accuracy} = 0.5 \times (TP/(TP + FN) + TN/(TN + FP)) \quad \text{(Equation 6)}$$

where TN is true negative, TP is true positive, FN is false negative, and FP is false positive. Accuracy indicates the proportion of correctly predicted samples in the data sets. Recall rate represents the percentage of correctly classified positive samples in all positive samples. Precision denotes the proportion of correctly classified positive samples in all positive predictions. F-measure is the harmonic mean of recall rate and precision. Balanced accuracy is computed as the average of the accuracy of positives and the accuracy of negatives. On imbalanced data, accuracy cannot be primary performance metric. In addition, balanced accuracy is more suitable for evaluating binary classification on unbalanced datasets than F-measure.[55] Recall rate represents the accuracy of the PMT substance assessment according to the dataset. Due to more attention given to prediction of PMT substances (positive), recall rate and balanced accuracy are selected as primary performance metrics and recall rate is given priority.

To better evaluate the performance of the above-mentioned models, we used a 5-fold cross-validation method to eliminate the impact of dataset partitioning. Firstly, the dataset D is divided into five mutually exclusive subsets of similar size, that is, $D = D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_5$, $D_i \cap D_j = \varnothing$ $(i \neq j)$ (see Figure S31). Then, the union of 4 subsets is used as the training set each time, and the remaining subset is used as the test set. In this way, 5 sets of corresponding training and test sets can be obtained, so that 5 training and testing times can be performed. The mean of the 5 test results is used as the final result.

Model validation and application

In Figure 1, the optimal machine learning model is validated and applied on COVID-19 related chemicals. Here, we divide COVID-19-related compounds into disinfectants, antivirals, and other treatments. 605 disinfectants products are obtained from the U.S. Environmental Protection Agency's (EPA) List N: Disinfectants for Use Against SARS-CoV-2.[56] About half of products contain quaternary ammonium compounds (QACs) which are active components in hospital and household cleaners.[57] 8 common QACs are selected as our target compounds according to a previous study.[58] Finally, a total of 39 disinfectants compounds are summarized in our list for PMT screening. Furthermore, the antivirals and other auxiliary drugs are obtained from the previous studies in the ISI Web of Science. In total 69 compounds are collected from 27 literatures for PMT screening (see Tables S2 and S3). From this literature, a total of 108 COVID-19-related compounds are included (see Table S4).

The above-mentioned 108 compounds in our list were screened manually for their PMT properties based on available experimental data and weight-of-evidence data using the approach demonstrated in the previous studies[13] (see Table S4). The screening results for PMT substances are divided into two categories based on the quality of data. One kind is the results based on high quality data (expert judgement). When the expert-verified results of compounds were unavailable, the QSAR method is utilized to evaluate the P, M and T properties of these substances. The method applies existing modeling tools to evaluate the persistency (e.g., using the BIOwin and P-estimator tools in QSAR

Toolbox (https://qsar-toolbox.org/), mobility (e.g., calculating pH dependant octanol-water distribution coefficients, log D_{ow}, ins ChemAxon, https://www.chemaxon.com/) and toxicity (e.g., QSAR Toolbox) and the final identification conclusion is given by combining the above-mentioned results (see Note S2). Finally, the 32 chemicals with available expert judgments are retained to evaluate the performance of model. The optimal machine learning model is applied to the remaining 76 chemicals.

SHAP method and causal inference

In order to break up the "black box" of machine learning model, a locally interpretable explanatory method termed SHAP[59] is utilized to explore the mechanism of our model. SHAP is a model interpretation method based on game theory,[60] which could interpret the output of any machine learning model. For the predicted compound x_i which has n molecular features, SHAP method calculates the Shapley value ($\varphi_j(f, x_i)$) of each feature to measure the impact of the features on the final prediction value ($f(x_i)$) as shown in Equation 7:

$$f(x_i) = \varphi_0(f, x) + \sum_{j=1}^{n} \varphi_j(f, x_i) \quad \text{(Equation 7)}$$

where $\varphi_0(f, x)$ is the base value and the $\varphi_j(f, x_i)$ is SHAP value which denotes contribution of the jth molecular feature in the compound x_i to the final predicted value. The mathematical definition of SHAP value ($\varphi_j(f, x_i)$) is shown in Equation 8:

$$\varphi_j(f, x_i) = \sum_{z' \subseteq x'} \frac{|z'|!(n - |z'| - 1)!}{n!} [f_x(z') - f_x(z'\ j)] \quad \text{(Equation 8)}$$

where x' is the set of all possible feature combinations containing feature j, n is the number of all molecular features, $|z'|$ is the number of non-zero entries in feature combination z', $z'\ j$ denotes feature combination which removes feature j from z', $f_x(z')$ and $f_x(z'\ j)$ represent model predictions for feature combinations z' and $z'\ j$, respectively. In summary, the SHAP value of feature j is obtained by weighting and averaging the difference with and without feature j in all possible feature combinations. A more detailed explanation of the SHAP method was summarized in Supplemental information (Note S11). Model interpretation was carried out with Kernel Explainer module within shap 0.40.0 package.

Although the SHAP method is useful to extract correlation information, we cannot determine the causal impact.[46] Thus, we trained causal models and estimated the causal effects for each of top 20 important molecular features selected by SHAP method. As a promising approach for discovering causal relationships,[34] structural causal model (SCM) is selected to explore the causation between molecular descriptors and PMT/Not PMT classifications. For causal inference, the first step is defining the causal model in form of causal graph based on the domain knowledge and assumptions. For instance, to estimate the causal effect of feature B04[C-N] to PMT properties, we selected the molecular feature B04[C-N] as the treatment feature (T) and select PMT properties as outcome (Y). The other 19 top important features were selected as co-variates (X) and the rest 1784 features were selected as confounders (W). Here, X refer to the variables which are used to estimate heterogeneous treatment effect. Thus, a causal graph is defined (see Figure S32). The second step is building estimators to estimate the causal effects between T and Y according to causal graph. Recently, many estimators have been proposed for causal inference, mainly including machine learning based methods. Specially, the DMLOrthoForest method performs particularly well in the presence of high-dimensional confounding factors due to the orthogonalization aspect of the method. Therefore, we select DMLOrthoForest method to build estimators for each feature. A more detailed introduction to the DMLOrthoForest method can be found in the Supplemental information (Section S1.12). Finally, checking the robustness of the estimates is the most important step in the causal analysis. We obtained an estimate using steps 1-2, but each step might have made certain assumptions that could lead to wrong results. This step relies on refutation tests by using various robustness checks to verify the correctness of the estimate. Here, we used three refutation test methods, adding random common cause (RCC), placebo treatment (PT) and data subset refuter (DS) to test the robustness of causality. The RCC checks the following: Does the estimation method change its estimate after we add an independent random variable as a common cause to the dataset? Here, the PT replaces the true intervening variable with an independent random variable to determine whether the causal effect will go to zero. DS replaces the given dataset with a randomly selected subset to test whether the causal effect will change significantly. For a robust causal relationship, the new causal effects of RCC and DS refutation test should be similar with the original causal effects but the new causal effects under PT test should be zero. To quantify the robust check criterion (similarity between new causal effects under refutation test and the original causal effects), we use variance between the new causal effect and the original value as evaluation thresholds for RCC and DS tests. For this, three evaluation thresholds were set (5%, 1% and 5‰), suggesting an increasingly strict criterion.[61] The potential causal relationship passed the initial refutation when the variance was less than 5%. Furthermore, the variance of the estimates under RCC test with respect to the original values within 1% represents the second level. In order to make sure the robustness of causal relationships, the causality is considered robust only when passing the strictest criterion (5‰).[61] Furthermore, the causal relationship between molecular descriptors and "Not PMT" classifications was also explored by labeling "PMT" as 0 and "Not PMT" as 1. The causal inference was implemented by using Python 3.8.8 with EconML 0.13.1 package and DoWhy 0.7.1 package. The causal inference was accomplished on Tianhe-2 Supercomputer.