METHODOLOGY

Open Access

APBIO: bioactive profiling of air pollutants through inferred bioactivity signatures and prediction of novel target interactions

Eva Viesi[1,2,5*], Ugo Perricone[4,5], Patrick Aloy[2,3] and Rosalba Giugno[1,5]

Abstract

More sophisticated representations of compounds attempt to incorporate not only information on the structure and physicochemical properties of molecules, but also knowledge about their biological traits, leading to the so-called bioactivity profile. The bioactive profiling of air pollutants is challenging and crucial, as their biological activity and toxicological effects have not been deeply investigated yet, and further exploration could shed light on the impact of air pollution on complex disorders. Therefore, a biological signature that simultaneously captures the chemistry and the biology of small molecules may be beneficial in predicting the behaviour of such ligands towards a protein target. Moreover, the interactivity between biological entities can be represented through combined feature vectors that can be given as input to a machine learning (ML) model to capture the underlying interaction. To this end, we propose a chemogenomic approach, called Air Pollutant Bioactivity (APBIO), which integrates compound bioactivity signatures and target sequence descriptors to train ML classifiers subsequently used to predict potential compound-target interactions (CTIs). We report the performances of the proposed methodology and, via external validation sets, demonstrate its outperformance compared to existing molecular representations in terms of model generalizability. We have also developed a publicly available Streamlit application for APBIO at ap-bio.streamlit.app, allowing users to predict associations between investigated compounds and protein targets.

Scientific contribution

We derived ex novo bioactivity signatures for air pollutant molecules to capture their biological behaviour and associations with protein targets. The proposed chemogenomic methodology enables the prediction of novel CTIs for known or similar compounds and targets through well-established and efficient ML models, deepening our insight into the molecular interactions and mechanisms that may have a deleterious impact on human biological systems.

Keywords Chemogenomic approach, Air pollutants' bioactivity, Compound-target interaction prediction

*Correspondence:
Eva Viesi
eva.viesi@univr.it
Full list of author information is available at the end of the article

Introduction

Several studies have already linked exposure to environmental air pollution to carcinogenesis [1, 2], inflammation and oxidative stress [3, 4], and other chronic or complex disorders [5, 6]. However, exploring the bioactivity of air-polluting compounds by discovering novel biological targets and binding activities may reveal still uninvestigated harmful and toxicological effects of such molecules on human health. Following this direction, we recently developed a Database of Air Pollutants (APDB) [7], that chemically characterizes molecules derived from air pollution sources and reports computed molecular descriptors and their similarities.

Computational or in silico prediction of new compound-target interactions (CTIs) can help overcome the limitations of experimental in vivo and in vitro methodologies, mainly related to time and resource cost [8]. In particular, chemogenomic or proteochemometric (PCM) approaches attempt to represent CTIs by a molecular feature vector able to capture the interaction between multiple ligands and proteins simultaneously [9, 10], without requiring an exhaustive list of active ligands for similarity search, as in the case of ligand-based approaches, or a solved three-dimensional (3D) structure, as in the case of molecular docking in virtual screening (VS) [11, 12]. A widely known ligand-based method is the Similarity Ensemble Approach (SEA) [13], which employs chemical similarity to identify targets for small molecules by comparing them to known ligands. However, predictions are challenging for targets with few active ligands and are particularly focused on well-characterized proteins. Proteochemometric (PCM) strategies, instead, integrate protein descriptors to provide additional knowledge for target prediction and combine them with chemical features to detect patterns of interactivity by machine learning models. This enables handling data imbalance and sparsity, facilitating the prediction of novel ligand-target pairs.

Various molecular features have been proposed and calculated from chemical structures and protein sequences to describe the interaction between small compounds and their biological targets [14]. At the same time, different methodologies have been developed to leverage these properties, such as kernel- and network-based approaches, and machine learning-based methods, which mainly exploit the chemical and genomic similarity or integrate molecule and target feature vectors to predict potential ligand-target interactions [15–18]. These approaches use common molecule representations, such as Morgan fingerprints or MACCS keys [19], without or partially including all the bioactivity knowledge that could be inferred for small compounds [20]. Moreover, they are mainly applied to drug-like molecules in the context of drug repurposing and discovery and do not consider environmental pollutants, such as volatile organic compounds (VOCs) or polycyclic aromatic hydrocarbons (PAHs), and their molecular interactions with human proteins in the view of better understanding their mechanisms of action, toxic and side effects.

In the endeavour to characterize molecules from a biological perspective, the Chemical Checker (CC) [21] resource supplies essential bioactivity data on a massive collection of small compounds. In particular, the CC is divided into five datasets (A-E) representing different contexts in which a small molecule can be studied, starting with its chemistry, moving on to its associated biological targets, pathways and networks, and ending with its effect on cells and diseases. Each dataset is in turn subdivided into five spaces (1–5) containing specific data for each category (such as two-dimensional (2D) fingerprints for chemistry or binding data for targets). Based on the information stored in the 25 CC spaces, a *Signaturizer* module has been implemented to predict a signature of biological activity for any molecule of interest [22].

Although the CC is a comprehensive repository which offers knowledge on about one million small molecules by describing their biological similarities to assist the drug discovery procedure, we observed that out of 1,830 air pollutant molecules in APDB only ~ 920 were common with the CC. Moreover, concerning the chemical feature space, APDB considers fingerprints, structural keys, and physicochemical properties common to the CC, but also complementary information by computing further molecular descriptors and quantum–mechanical properties. In addition, both resources provide different information on experimental bioassays and targets associated with the collected molecules.

Regarding targets, a vast number of open-source software are readily accessible to calculate sequence descriptors, such as the web servers PROFEAT [23, 24] and PseAAC [25], the R package and web application protr/ProtrWeb [26], and the iFeature Python-based package [27]. All these tools extract some of the most used and efficient features for characterizing biological target sequences, i.e., the amino acid composition, the dipeptide composition, the autocorrelation, the composition, transition and distribution, and the quasi-sequence-order descriptors [28].

Associations between molecules and targets can be extrapolated from several publicly available databases [15], including ChEMBL [29], which provides data on bioactivities, BindingDB [30], describing molecular binding affinity, DrugBank [31], which contains information mainly related to drugs, the Comparative Toxicogenomics Database (CTD) [32], storing manually curated toxicogenomics data, and PubChem BioAssay [33], a repository among the most comprehensive ones of small compound-associated biological experiments.

Bringing together all the information related to molecules, targets, and their interactions, the final CTI representation can be obtained by integrating compound and target features, such that the combined feature vector can be employed to fulfil a binary classification task via machine learning (ML) models as described above. Specifically, supervised machine learning classification methods require defining a set of positive and negative instances to perform prediction [15]. At this point, the identification of negative samples represents a fundamental step, since data is usually highly unbalanced due to the small number of validated interactions and negatives could also represent untested positive pairs [17, 18]. Many methods have been proposed for constructing a set of representative negative samples from a dataset of experimentally validated positive pairs to improve the prediction performances of the ML models. These approaches are mainly based on similarity/dissimilarity measures between molecule and target properties and/or the One-Class Support Vector Machine (OCSVM) classifier [34, 35]. In particular, the OCSVM model has been widely used in the field of anomaly and novelty detection to generate a decision boundary, starting from a set of positive training instances, capable of detecting outlier samples [36, 37]. A sample with a high negative distance from the boundary or hyperplane has less resemblance to positive data, therefore, based on this assumption, negative instances can be sampled according to the distances defined by the OCSVM.

In this paper, we propose a feature-based chemogenomic method which finds a proper representation of the molecular and biological properties of air pollutant molecules and their associated targets to capture and predict potential interactions.

Specifically, the first objective was to incorporate additional compounds and chemical properties into the CC tool to extend the original chemical knowledge, and, in turn, leverage all available biological information (chemistry, targets, networks, cells, clinics) to describe air pollutants according to their bioactivity. Furthermore, this integration was intended to place air pollutants within the broader context of known bioactive compounds. The consequent purpose was to provide a qualitative view of how specific categories of molecules, in particular air pollutants, interact at the molecular level with target proteins. Even though this does not directly address dose–response relationships or exposure levels, it could be the basis for further toxicological and chemical risk assessments.

Figure 1 shows the overall procedure of the proposed APBIO pipeline for CTI prediction. In particular, we derived new *Signaturizer* models from the raw data in APDB by leveraging the information stored in the CC to expand knowledge of air pollutants beyond their structural and physicochemical features and gain a more accurate understanding of their bioactivities and mechanisms of action in pathways and diseases.

We combined target sequence descriptors extracted via the iFeature package [27] with molecule bioactivity signatures inferred by the *Signaturizer* models according to the activity observed in the experimental bioassays from PubChem. We trained a One-Class SVM estimator with positive molecule-target interactions and calculated the distances to the hyperplane for the unlabelled interactions. This allowed the identification of a reliable set of negative instances. We constructed the final CTI dataset by merging the positive pairs with the selected negative pairs.

As a comparison, we applied the same procedure to build CTI datasets derived from the Morgan fingerprints and the *Signaturizer* models of the CC chemical spaces.

We used the generated datasets to evaluate the ability of four ML estimators in classifying compound-target interactions. We demonstrated model generalizability on unseen data consisting of active pairs from PubChem BioAssay not present in the training dataset, additional binding data provided by the CC, and specific protein targets from the CTD.

We first showed that negative samples selected by the OCSVM substantially increase the accuracy of the predictive models compared to random sampling. We then obtained model performances for all the datasets by applying a nested cross-validation (CV) procedure, observing comparable and outstanding results. Whereas, addressing generalization capabilities, we found that molecular features from APDB provided more stable results in terms of recall score compared to the other datasets in any type of tested scenario.

Finally, we supply an application, called APBIO, which, given as input a molecule structure and a target identifier, generates molecule signatures and target descriptors to predict potential interactions between the biological entities being investigated.

Methods

The following sections describe the collection of molecule and target data, the pipeline for the inference of bioactivity signatures and the extraction of target sequence descriptors. They also report the methodology for the prediction of CTIs for which we provide a Streamlit web application located at ap-bio.streamlit.app that enables the input of compound structures (SMILES) and target identifiers (UniProtKB) to obtain feature vectors used for estimating possible interactions.

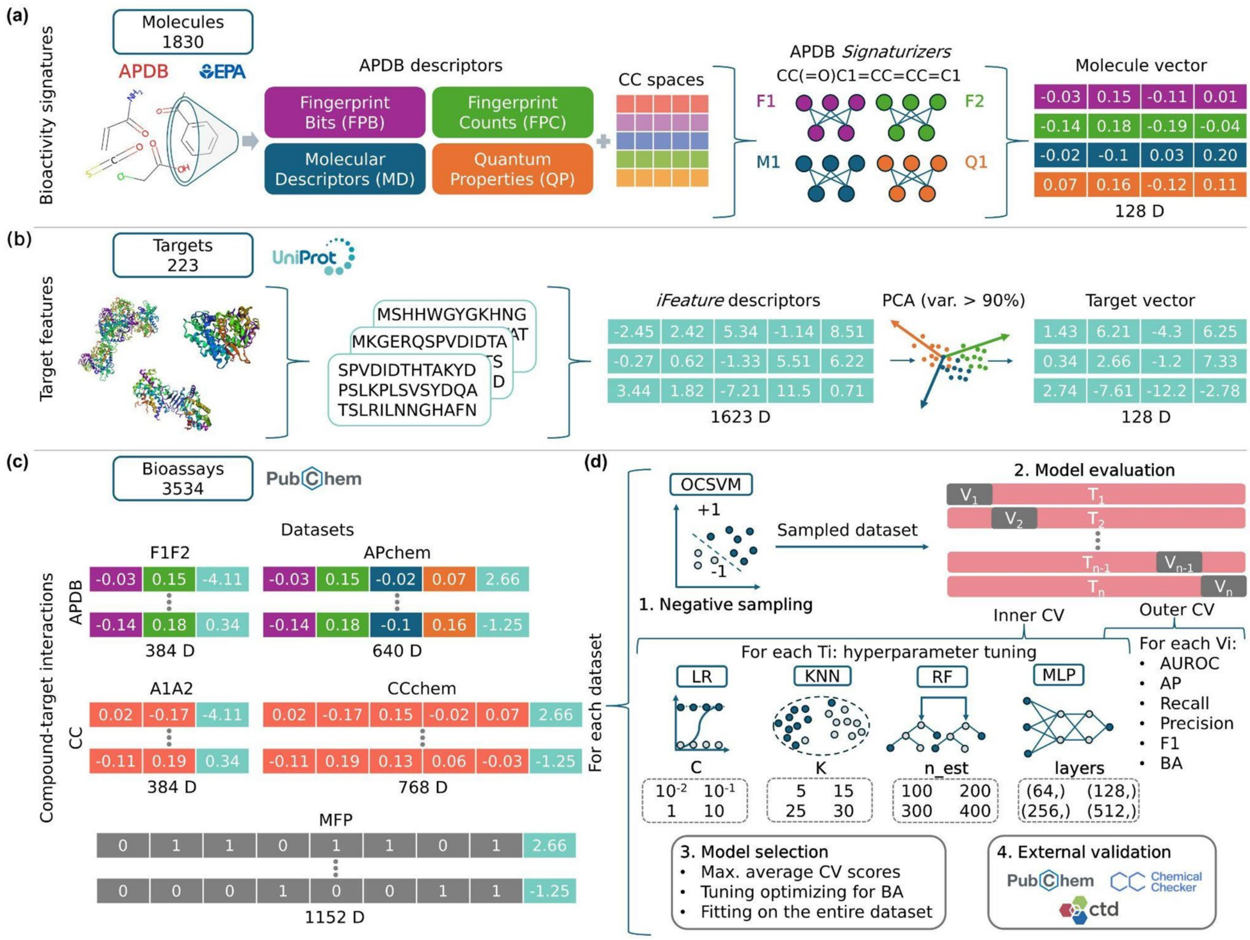

Fig. 1 APBIO pipeline. The **a** figure illustrates the procedure to derive bioactivity signatures (128 dimensional vectors) for the 1,830 molecules in the Database of Air Pollutants (APDB) collected from the Environmental Protection Agency (EPA). For each chemical space in APDB, namely, fingerprint bits (FPB), fingerprint counts (FPC), molecular descriptors (MD), and quantum properties (QP), the corresponding *Signaturizer* models, called F1, F2, M1, and Q1, respectively, are built by leveraging the information stored in the 25 Chemical Checker (CC) spaces. The **b** plot reports the steps to compute sequence descriptors of 223 protein targets. From the UniProtKB identifier, the FASTA file of the protein sequence is fetched and used to calculate descriptors by the iFeature module. The resulting feature vector of dimension 1623 is scaled and reduced by Principal Component Analysis (PCA) to a dimension of 128. In **c** are depicted the compound-target interaction (CTI) datasets built from the active pairs identified in 3,534 bioassays from PubChem by concatenating molecule and target feature vectors. F1F2 and APchem are the datasets derived from APDB, A1A2 and CCchem are the datasets derived from the CC, MFP is the dataset of Morgan fingerprints (1024 bits, radius 2). The **d** figure shows the CTI prediction workflow applied to each dataset independently. The first step involves the sampling of negative instances by the One-Class Support Vector Machine (OCSVM) classifier. Secondly, the sampled dataset is used as input for nested cross-validation (CV) of four machine learning (ML) models, i.e., Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Multi-layer Perceptron (MLP). For each training set of the outer stratified tenfold CV, an inner stratified fivefold CV is performed to tune model hyperparameters (C, K, n_est, and layers, respectively), while evaluation metrics are computed on each remaining validation fold. The final model is selected by averaging CV scores, tuning parameters, and fitting it on the entire dataset. Lastly, external validation is conducted on unknown CTIs from PubChem, the Chemical Checker, and the Comparative Toxicogenomics Database (CTD)

Data collection

We collected 1830 molecules from APDB [7], derived from the SPECIES_PROPERTIES table of the SPECIATE 5.1 database [38] provided by the Environmental Protection Agency (EPA) and from the EPA's Hazardous List of Air Pollutants [39].

We downloaded the molecule data stored in APDB related to four chemical spaces, specifically, the fingerprint bits (FPB), fingerprint counts (FPC), molecular descriptors (MD), and quantum properties (QP) tables later used to derive new molecular representations (Fig. 1a).

We retrieved bioassays from APDB and updated them with the version currently available in PubChem [33] collecting data via the PUG-REST service [40, 41] querying by PubChem's CID. We kept only active assays that were tagged as "Confirmatory" or "Summary" and had a specified activity value to minimize the number of false positives that could be found in primary screens [42], discarding those changed to "Unspecified". In cases of contradictory results for the same assay identifier (AID), we retained experiments reporting a majority or an equal number of active results over inactive ones. In the presence of concurrent inactive assays, the active assays were kept as considered more reliable. The targets associated with the queried bioassays were annotated by the reviewed UniProtKB-Swiss-Prot identifier, protein names, gene names, organism and sequence length using the UniProt ID Mapping REST service [43] (https://www.uniprot.org/help/id_mapping). Only human proteins were selected, corresponding to a total of 223 genes (Fig. 1b). The final number of selected assays is 3534 (Fig. 1c) corresponding to 2609 CTIs (499 compounds and 223 targets). The count of targets and the number of interactions with active ligands for each identified protein family are reported in Supplementary Figure S1.

Bioactivity signatures derivation

Bioactive profiling of small compounds goes beyond chemistry itself, as it seeks to explore the available bioactivity information of the molecules under study [21]. This is highly valuable since the bioactivity profile can be exploited to unravel hidden mechanisms and adverse effects of air pollution on human health.

Molecular representations in APDB are generated from structural, physico- and quantum-chemical properties by initially applying dimensionality reduction to the raw data, followed by node embedding on networks constructed from significant pairwise molecular similarities within each identified chemical space. Further details on their calculation can be found in the corresponding manuscript [7].

Although these embeddings capture a wide range of chemical characteristics, they do not integrate any bioactivity data; we therefore derived novel molecule signatures by incorporating additional information present in the CC, such as binding modes, cell effects, and clinical outcomes of bioactive molecules. The main advantage of this process is that bioactivity signatures can be efficiently predicted for any compound, providing a useful descriptor for downstream analyses, such as similarity search, clustering, interaction or activity prediction, replacing traditional molecular representations.

The pipeline for the creation of a new CC space and the inference of bioactivity signatures consists of four main steps and is described in detail at https://gitlabsbnb.irbbarcelona.org/packages/protocols.

Starting from the APDB raw data, which can be discrete, as in the case of fingerprint bits or counts, or continuous, as in the case of molecular descriptors and quantum properties, a cleaning and pre-processing step is applied, leading to *signature 0*. The output signatures are then compressed by using Latent Semantic Indexing (LSI) or Principal Component Analysis (PCA) reduction techniques, typically preserving 90% of the variance. The low-dimensional data, called *signature 1*, are used to derive similarity networks encoding statistically significant similarities between pairs of molecules. An embedding algorithm is applied to these networks to obtain fixed-length embedding vectors denoted as *signature 2*. The *signatures 2* for all CC and APDB compounds are fed into a Siamese Neural Network (SNN) to derive a data representation, named *signature 3*, explaining the initial similarity observed in the *signature 1* of each CC space and the newly created space. Finally, a Deep Neural Network (DNN) model is trained to predict the comprehensive *signature 3* from the Morgan fingerprints of molecules. The ultimate model is called *Signaturizer* and the resulting *signature 4* is the so-called bioactivity signature, i.e., a feature vector of dimension 128 [22]. We applied the entire pipeline to derive signatures from *0* to *4* for the four APDB chemical spaces. The final *Signaturizer* models are called F1 for fingerprint bits, F2 for fingerprint counts, M1 for molecular descriptors and Q1 for quantum properties (see Fig. 1a).

The pipeline also provides a diagnostic plot for each space to observe and evaluate the quality of the signatures produced. In Fig. 2 we report the diagnostic plot for the F1 space.

In Fig. 2a, the heatmap represents data values of 128 features (i.e., the signature dimension) for 100 selected molecules (or keys). The range of values for *signature 4* goes from – 0.2 to 0.2 (Fig. 2b, c) as the final result of the whole procedure. Instead, for example, if we had examined the raw data for the fingerprint bits space, we would have found 0 and 1 as the unique values.

The 2D t-distributed stochastic neighbor embedding (t-SNE) projection of signatures (Fig. 2d) exhibits higher-density areas at the boundaries and the Euclidean (Fig. 2e) and Cosine (Fig. 2f) distance distributions reflect a fair degree of similarity of the molecule features considered in this space.

The remaining plots show the behaviour of F1 *signature 4* compared to the CC spaces for each dataset by taking as reference *signature 0* and *signature 1*. Each space in the CC is named with the dataset (A: chemistry (red), B: targets (purple), C: networks (blue), D: cells (green), E: clinics (yellow)) and a number from 1 to 5, such as A1.

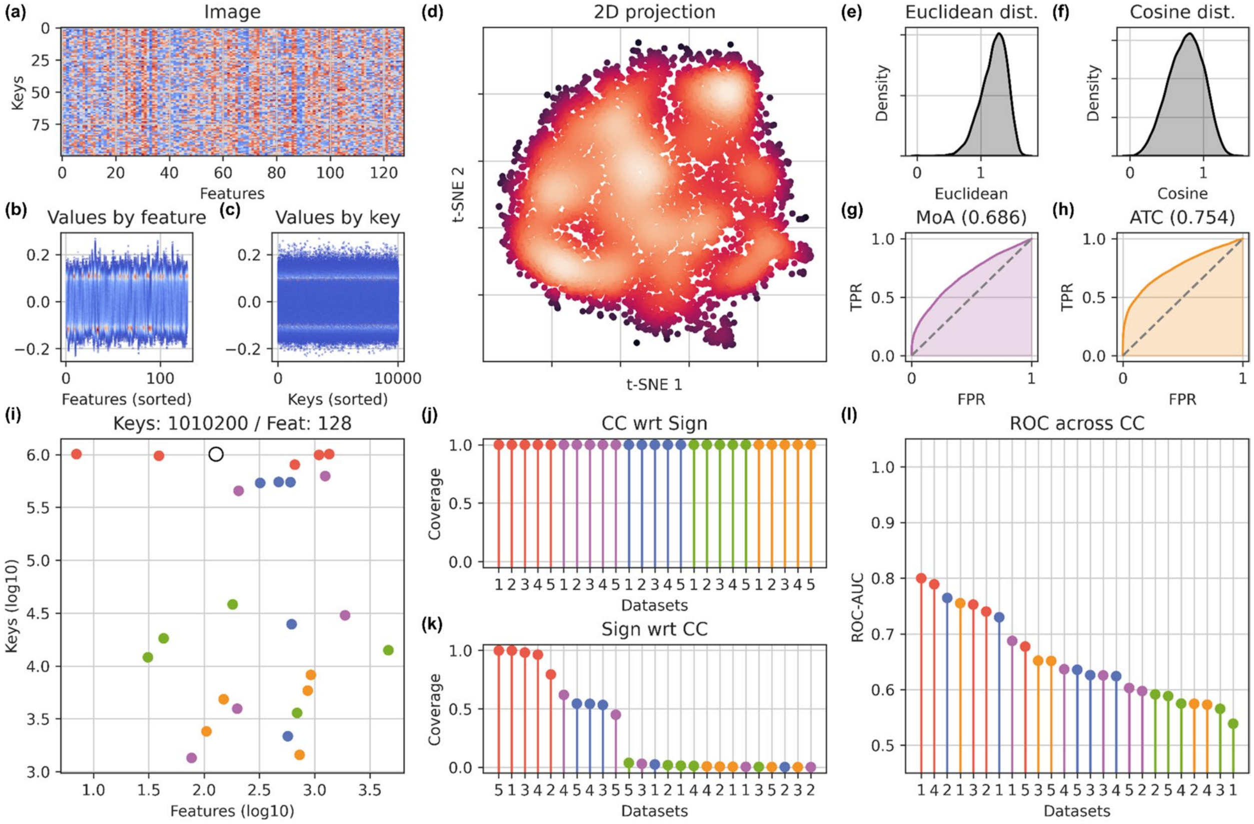

Fig. 2 A diagnostic plot to evaluate the quality of *signature 4* of the APDB F1 space and its comparison with the CC datasets. The **a** plot represents the distribution of feature values from minimum (blue) to maximum (red) per key (100 compounds are randomly selected by default). Similarly, the **b** and **c** plots depict the distribution of values present in the dataset per each feature and each key, respectively. The **d** figure is the two-dimensional t-distributed stochastic neighbor embedding (2D t-SNE) projection of the molecule signatures. The **e** and **f** density plots represent the pairwise Euclidean and Cosine distance distribution, respectively. The **g** and **h** are the receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC) value reflecting whether neighboring molecules for that signature tend to have similar mechanisms of action (MoA) or therapeutic code (ATC) (*signature 0* as reference). The **i** plot shows the log10 value of the number of keys and features of *signature 1* of each CC space and the *signature 4* of the created space (white dot). The **j** and **k** lollipop plots illustrate the proportion of *signature 4* keys covered by *signature 1* keys for each CC dataset and, conversely, the proportion of *signature 1* keys for each CC dataset covered by *signature 4* keys. The **l** figure displays the area under the receiver operating characteristic curve (ROC-AUC) values indicating whether neighboring molecules for that signature share similar characteristics in each CC space (*signature 0* as reference)

The bioactivity signature of F1 presents a slighter degree of similarity when looking at shared mechanisms of action (MoA), for which the area under the curve (AUC) is 0.686 (Fig. 2g). A closer resemblance is found when considering anatomical therapeutic chemical (ATC) classes, for which the AUC is 0.754 (Fig. 2h). Figure 2i illustrates the dimension of the *signature 4* of F1 (white dot) and of the *signature 1* of each CC space in terms of number of keys and features. Additionally, we can see that *signature 1* keys of each CC dataset cover all *signature 4* keys (Fig. 2j), but not vice versa (Fig. 2k), e.g., only ~ 60% of *signature 4* keys are present in B4 (binding data). The best recapitulation, defined as the capability to recover the KNNs identified in *signature 0* of the dataset under consideration, is obtained for the 2D fingerprints (A1) and the structural keys (A4), as expected (Fig. 2l).

Target feature extraction

We first retrieved the FASTA sequence for the 223 target proteins via UniProt RESTful APIs (https://www.uniprot.org/help/api_queries) querying by UniProtKB-Swiss-Prot identifier. Secondly, through the iFeature package [27], we computed the following sequence properties to define the target's biological space: the Amino Acid Composition (AAC), the Dipeptide Composition (DPC), the Pseudo-Amino Acid Composition (PAAC), the Moran, Geary, and Normalized Moreau-Broto autocorrelation descriptors (Moran, Geary, NMBroto), the Composition, Transition, and

Distribution features (CTDC, CTDT, CTDD), the Sequence-Order-Coupling Number (SOCNumber) and the Quasi-Sequence-Order (QSOrder) descriptors. In particular, from the software package provided in https://github.com/Superzchen/iFeature/, we used the *iFeature.py* program and the required files in the *codes* and *data* folders to obtain a.csv file for each descriptor. We then merged the calculated descriptors into a single dataset of 1623 features and applied feature scaling and principal component analysis (PCA) (StandardScaler and PCA implemented in the scikit-learn package [44]) to reduce the dimensionality to 128 features, explaining more than 90% of the variance (Fig. 1b). The fitted scaler and pca model can be used to derive feature vectors for novel targets to be predicted.

Compound-target interaction prediction

Datasets generation

Starting from the 1830 APDB molecules and the 223 associated targets, we generated five different CTI datasets by the concatenation of bioactivity signatures and target sequence descriptors (Fig. 1c). Compound signatures are derived from *Signaturizer* models by giving chemical structures as input in the form of SMILES strings, while target features are calculated from the biological sequence represented in FASTA format. Specifically, F1F2 is the dataset derived from the APDB fingerprint bits (F1) and fingerprint counts (F2) *Signaturizers* and APchem is the dataset derived from the APDB fingerprint bits (F1), fingerprint counts (F2), molecular descriptors (M1), and quantum properties (Q1) *Signaturizers*. Given that both molecule signatures and target descriptors have 128 features, the final datasets have dimensions 384 and 640, respectively. Concerning samples, we used all 1830 molecules and 223 targets to create CTI datasets, assigning the label 1 to the 2609 active pairs and the label 0 to the 405,481 remaining ones, for a total of 408,090 pairs.

For evaluation purposes, we created CTI datasets from the CC using the signaturizer Python module provided in [22]. In particular, A1A2 is the dataset derived from the 2D fingerprints (A1) and 3D fingerprints (A2) *Signaturizers* and CCchem is the dataset derived from the 2D fingerprints (A1), 3D fingerprints (A2), scaffolds (A3), structural keys (A4), and physicochemical parameters (A5) *Signaturizers*. Moreover, for further comparison, we employed the rdkit Python library (http://www.rdkit.org/) to generate a CTI dataset of Morgan fingerprints (1024 bits, radius 2) called MFP. Given that both molecule signatures and target descriptors have 128 features, the final datasets have dimensions 384, 768, and 1152, respectively.

Negative instances selection

For each generated CTI dataset, we performed a selection of negative samples (Fig. 1d). This is a crucial step to determine the performance of the final prediction model. Following the approaches in [35, 45], we trained a One-Class SVM (OCSVM) model using the 2,609 positive samples identified. In particular, we selected a linear kernel and tuned the *nu* parameter by taking [0.01, 0.03, 0.05, 0.1] as the range of values and maximizing the recall (> 0.9) to have a high proportion of correctly classified positives. We performed two fivefold cross-validation repetitions to select the best model and computed the distances to the estimated hyperplane for both positive and negative pairs. We ranked the negative signed distances in increasing order and picked the desired number of samples from the head of the list (Fig. 3a–e). Specifically, we set a positive-to-negative ratio of 0.1 to reproduce a realistic scenario in which non-active pairs are far more than active ones, resulting in a final number of 28,699 positive and negative samples. The number of selected APDB molecules and targets for each class and each dataset is reported in Supplementary Table S1.

To compare performances across different proportions of positives and negatives, we selected non-interacting pairs using a ratio of 0.2 and 0.01, resulting in a total number of 54,789 and 263,509 samples, respectively. Similarly to [34], we also generated corresponding datasets with randomly selected negative instances to evaluate the efficiency of the sampling strategy.

Model evaluation and prediction setting

Feature vectors of each sampled CTI dataset with a positive-to-negative ratio of 0.1 (obtained above) were fed to four different machine learning classifiers, namely, Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Multi-layer Perceptron (MLP) [44]. To evaluate model performances, we applied nested cross-validation, in which the inner loop is executed to tune the hyperparameters of the models by optimizing for balanced accuracy while the outer loop serves for metrics computation, specifically, the area under the ROC curve, average precision, recall, precision, f1, and balanced accuracy. Since only the training set of the outer loop is passed to the inner one, it allows better generalization by avoiding overfitting; at the same time, the remaining test set is used to estimate performance [46]. We performed repeated (*n_repeats* = 2) stratified cross-validation by setting 5 folds and 10 folds, respectively. We defined a grid of parameters for each classifier, in particular, the *C* range for LR was [0.01, 0.1, 1, 10], the *n_neighbors* (K) range for KNN was [5, 15, 25, 35], the *n_estimators* (n_est) range for RF was [100, 200, 300, 400], and the *hidden_layer_sizes* (layers) range for MLP was

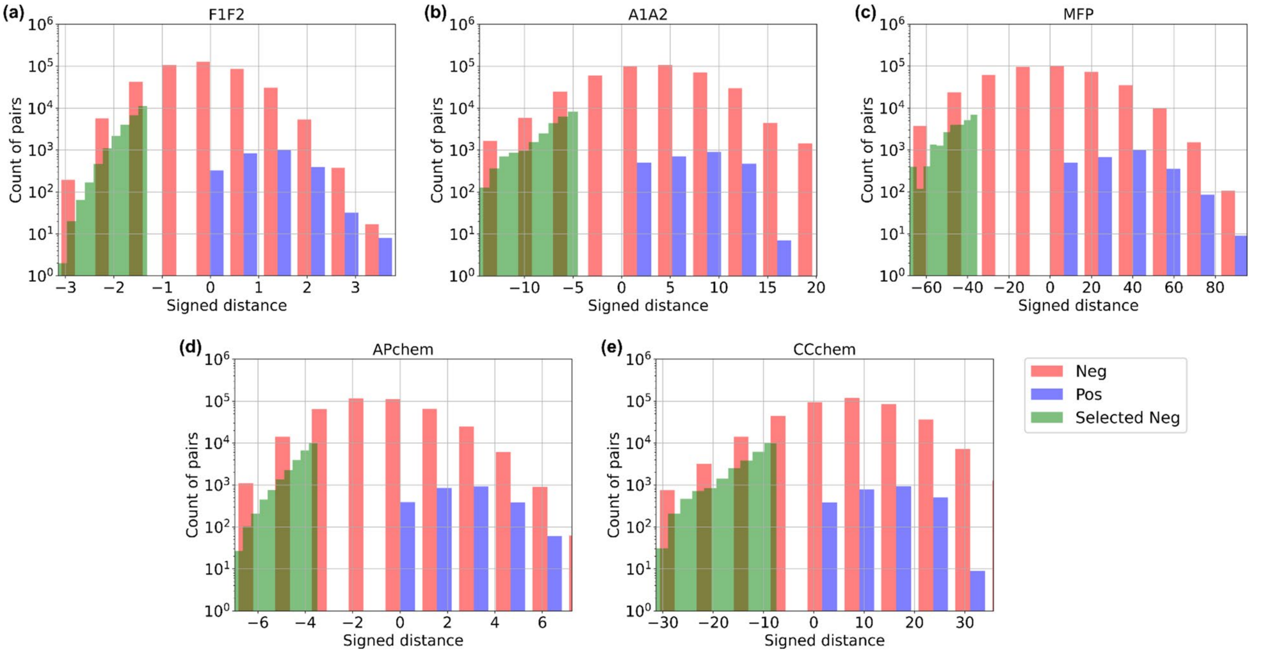

Fig. 3 Distribution of OCSVM signed distances of the compound-target pairs in (**a**) F1F2 dataset, (**b**) A1A2 dataset, (**c**) MFP dataset, (**d**) APchem dataset, and (**e**) CCchem dataset. Highlighted in green is the distribution of distances of the selected negative instances with a positive-to-negative ratio of 0.1

[64, 128, 256, 512]. To select the final model, we computed the average of all test scores and kept the estimator with the maximum value. Finally, we re-executed hyperparameter tuning via grid search fivefold cross-validation by maximizing for balanced accuracy (BA) and trained the best model on the entire dataset (Fig. 1d). We computed the final performances on a train-test split with a test set size of 0.3 reporting the precision-recall curve and the corresponding average precision value. We also calculated the same metrics on the different ratios of negatives and the randomly generated datasets for comparison purposes.

External validation dataset construction

To gain a more comprehensive understanding of pollutant molecule bioactivity and focus on specific protein families to evaluate the model's capabilities on external datasets, we conducted an enrichment analysis by exploiting biological annotations stored in the Chemical Checker (CC) [21].

Enrichment analysis is widely applied in the omics field to identify biological terms or pathways that are particularly enriched (or over-represented) in gene sets of interest [47–50].

In this framework, a common test used to determine statistical significance is the one-sided Fisher's exact test, which calculates an exact p-value based on a hypergeometric distribution. Following the formulation in [51, 52], for each term in a CC space, we constructed a contingency table by taking as background the CC molecules that had at least one annotation in the considered space and as a set of interest the APDB molecules found annotated with that particular term. For robustness, we considered sets of at least 5 molecules.

On this table, we executed the Fisher's test to find the terms whose associated molecules were significantly over-represented in the list of APDB molecules, compared to all the ones in the reference set [53]. To adjust apparently significant p-values due to multiple testing, we applied the Bonferroni correction, which controls the ratio of false positives by re-computing the significance threshold as $\alpha = 0.05/n$, where n is the number of tests executed [54]. We filtered out terms that were not statistically significant (adjusted p-value > 0.05) and we also inspected the strength of the observed association (or enrichment) by calculating the odds ratio value.

For the enrichment, we considered two CC datasets, namely, metabolic genes (B2 with dimension 1644 rows × 214 columns) and binding data (B4 with dimension 631,027 rows × 4635 columns) containing information on molecule targets derived from ChEMBL [29], BindingDB [30], and DrugBank [31]. Compounds were annotated through the CC Molset class implemented in the *molkit.py* module [21], which provides a dataset of annotation terms for each biological space of interest. In total, we have annotated 1,009,292 compounds from the

CC and 1830 from APDB, of which 921 are common to both datasets. APDB molecules with annotation are 44 in B2 and 146 in B4. The count of APDB molecules annotated in each CC space is shown in Supplementary Figure S2.

From the analysis, we found enriched terms in B2 related to the solute carrier (SLC) family of transporters. This protein family is known to be responsible for the transport of metabolites and nutrients that guarantee the correct functioning of vital organs, such as brain and heart [55, 56]. Moreover, recent studies have demonstrated that the expression of these transporters may be dysregulated by exposure to air pollution particles, potentially contributing to the onset of neurodegenerative diseases [57]. Concerning the B4 dataset, cytochrome P450 (CYP) enzymes, carbonic anhydrases (CA), and their related transcription factors emerged among the targets significantly associated with air pollutants. The first class of proteins is fundamental for the metabolism of xenobiotics (such as environmental pollutants and industrial chemicals) and it has been shown that the presence of different types of particulate matter (PM), like polycyclic aromatic hydrocarbons (PAHs), can lead to oxidative stress and subsequent inflammation by binding the aryl hydrocarbon receptor (AHR), an important transcription factor that modulates the expression of the cytochrome P450 CYP1A1 and CYP1A2 enzymes [58]. Relative to carbonic anhydrases, it has been found that they could be extremely useful in environmental biomonitoring due to their susceptibility to heavy metals and other contaminants [59].

Gathering all the information derived from the enrichment analysis described above, we constructed four validation datasets used to demonstrate the generalization capability of our classifier. Specifically, we selected 23 CTIs (13 compounds and 8 targets) considering the solute carrier (SLC) family of proteins, found to be enriched for APDB molecules in the B2 space, and additional binding data from the CC; 1059 CTIs (712 compounds and 38 targets) with the cytochrome P450 (CYP) enzymes and 152 CTIs (72 compounds and 11 targets) with the carbonic anhydrase (CA) proteins found in the Comparative Toxicogenomics Database, in particular, we downloaded the processed and organized CTD gene-chemical interactions from Harmonizome [60]. We also considered a validation set of CTIs from PubChem not present in APDB containing 34 new active pairs (21 compounds and 25 targets). The number of APDB and non-APDB molecules and targets for each external dataset is summarized in Supplementary Table S2.

We annotated targets with their reviewed UniProtKB-Swiss-Prot identifier, protein names, gene names, organism and sequence length and kept only human ones. Concerning CTD data, we converted each external substance ID to the corresponding PubChem Compound ID (note that some chemicals got lost during the conversion) and retrieved its SMILES and InChIkey (a 27-character string used as an identifier for molecules) through the PUG-REST web service.

For each validation dataset, we used the *Signaturizer* models to derive molecule signatures (see subsection *Datasets generation)*, and we applied the same procedure described in the section *Target feature extraction* to compute target descriptors.

Applicability domain study

The introduction of an applicability domain (AD) for a machine-learning classification model enables focusing on the most reliable predictions determined by the distribution of training data [61, 62]. Therefore, we defined an AD for each dataset using a distance-based method to represent the location in the chemical space and biological space of molecule signatures and target descriptors, respectively. In particular, we computed a distance cutoff to identify molecules and targets outside the applicability domain of our estimator to ensure robust predictions.

Following one strategy proposed in [63], we first calculated the average K-nearest neighbors (KNN) distance of each molecule in the training set by setting "cosine" as distance metric and $K=5$ by default (NearestNeighbors from scikit-learn [44]). We then used the computed distances to define an outlier threshold by considering the 95th percentile of the distribution. We applied the same procedure to targets. We also assigned a flag to molecules for each dataset, naming "positive" or "negative" those present only in the positive or negative pairs, respectively, "in" or "new" those present or not in the training set, respectively, and "out" the new molecules that do not fall within the applicability domain. We did the same for targets, excluding the "negative" flag. The number of molecules and targets for each flag category and each external dataset is reported in Supplementary Tables S3–S7.

First, we identify the intersection with the training set by using molecule InChIkey and target UniProtKB as identifiers; then, to establish whether a molecule or target is outside the AD, we inspect whether the average distance to its 5NN is higher than the defined cutoff. As a result, by defining the chemical and biological region in which the model performs effectively, the applicability domain enables determining the prediction confidence and reliability for an unknown compound-target interaction. An example of unseen molecules lying outside and inside the model's AD is illustrated in Supplementary Figure S3.

Results and discussion

APDB embeddings and CC signatures in comparison

Similarly to CC signatures, APDB provides embeddings for each chemical space, e.g., to calculate similarities between molecules, drawn from different types of molecular descriptors and properties. For this reason, we first investigated whether the information stored in the APDB and the CC is somehow analogous or complementary. The comparison between the APDB embeddings and the CC signatures is made in terms of recapitulation and follows the procedure reported in [21], specifically, we reimplemented the *compute_distance_pvalues()* function of the *signature_data.py* module and the *cross_roc()* function of the *diagnostics.py* module.

The recapitulation is given by the ability of the CC signatures to retrieve the nearest neighbors (NN) calculated from the APDB embeddings at different significance levels.

Thus, for each APDB space, namely, fingerprint bits (FPB), fingerprint counts (FPC), molecular descriptors (MD), and quantum properties (QP), we computed the pairwise cosine distances using the embedding vectors associated with each pollutant molecule and retrieved the NN at different p-value thresholds, namely 5×10^{-2}, 10^{-2}, 10^{-3}. We then predicted molecule signatures considering CC datasets representing different aspects of small-molecules, i.e., structural characteristics described by Morgan fingerprints (A1), target interactions from binding (B4) data, and association with diseases and toxicology (E4). We used the corresponding *Signaturizer* model [22] and calculated the pairwise cosine distances over signature vectors for positive and randomly chosen negative NN pairs from APDB data. Finally, we evaluated the results by displaying the receiver operating characteristic curve (ROC) and the precision-recall (PR) curve, of which we report the area under the curve (AUC) and the average precision (AP) values, respectively. As expected, for the A1 dataset representing the 2D or Morgan fingerprints of molecules (Fig. 4a–c), the best recapitulation is observed in the FPC embeddings at almost all p-value thresholds, given that the information encoded by these descriptors is similar (Fig. 4a). For binding data, represented by the B4 dataset (Fig. 4d–f), the AUC and AP values of FPC and MD are nearly comparable (Fig. 4d, e), reflecting the fact that the binding behaviour accounts for both structural and physicochemical characteristics. The E4 dataset (Fig. 4g-i), reporting the associations found between molecules and diseases, shows a worse recapitulation in general, due to the difficulty of capturing such relationships just from chemical information. Finally, we can observe that QP represents some complementary information concerning the signatures found in the CC; therefore, in this case, the recapitulation in terms of neighbors appears distinct (Fig. 4c, f, i).

In summary, we demonstrated that the knowledge captured by the CC and the APDB is fairly similar when looking at the structural and physicochemical characteristics of molecules, but not identical. Furthermore, the quantum properties from APDB provide supplementary chemistry-related information. This motivated the derivation of specific bioactivity signatures for pollutant molecules from APDB data.

APDB and CC signatures in comparison

We likewise compared the created APDB signatures and the CC signatures of 2D fingerprints (A1) and physicochemical parameters (A5), being the chemical spaces most similar to those found in APDB. The highest area under the curve (AUC) values are observed for the F1 and F2 spaces, indicating that the final *signature 4* recapitulates well the characteristics of the initial dataset represented by *signature 0* of fingerprint bits and counts (Fig. 5a, b). Even if for the M1 and Q1 spaces some information on physicochemical and quantum properties got lost during the signaturization procedure, the performances are still acceptable (Fig. 5c, d), and the bioactivity signature is capable of capturing both the original raw data and the knowledge derived from the CC. Instead, we can observe that the A1 and A5 CC *Signaturizer* models recapitulate less information found in the original APDB datasets (Fig. 5e–h) compared to the generated APDB *Signaturizers* (Fig. 5a–d). Therefore, the methodology described in the section *Bioactivity signatures derivation* allowed the incorporation of air-polluting molecules not present in the CC resource and the derivation of a bioactivity signature well reflecting their original similarities. This confirms that the developed *Signaturizer* models can be used to generate ad hoc representations for investigated pollutants, explore similar behaviours, and predict interactions with biological targets.

Model performances and novel CTIs prediction

We first analysed the performances of the different classifiers, namely, LR, KNN, RF, and MLP, to select the best model for each dataset. From the nested cross-validation, we averaged the average scores on the validation set for each metric (Fig. 6) and selected the estimator with the maximum value. Logistic regression was found to be the model with the most successful or comparable performance for all the datasets. We then re-execute hyperparameter tuning for the LR model and re-fit it on the whole dataset. We also fit the final classifier on the datasets with a positive-to-negative ratio of 0.2 and 0.01 selected by both strategies, OCSVM and randomly. Results are evaluated in terms of precision-recall curve and show

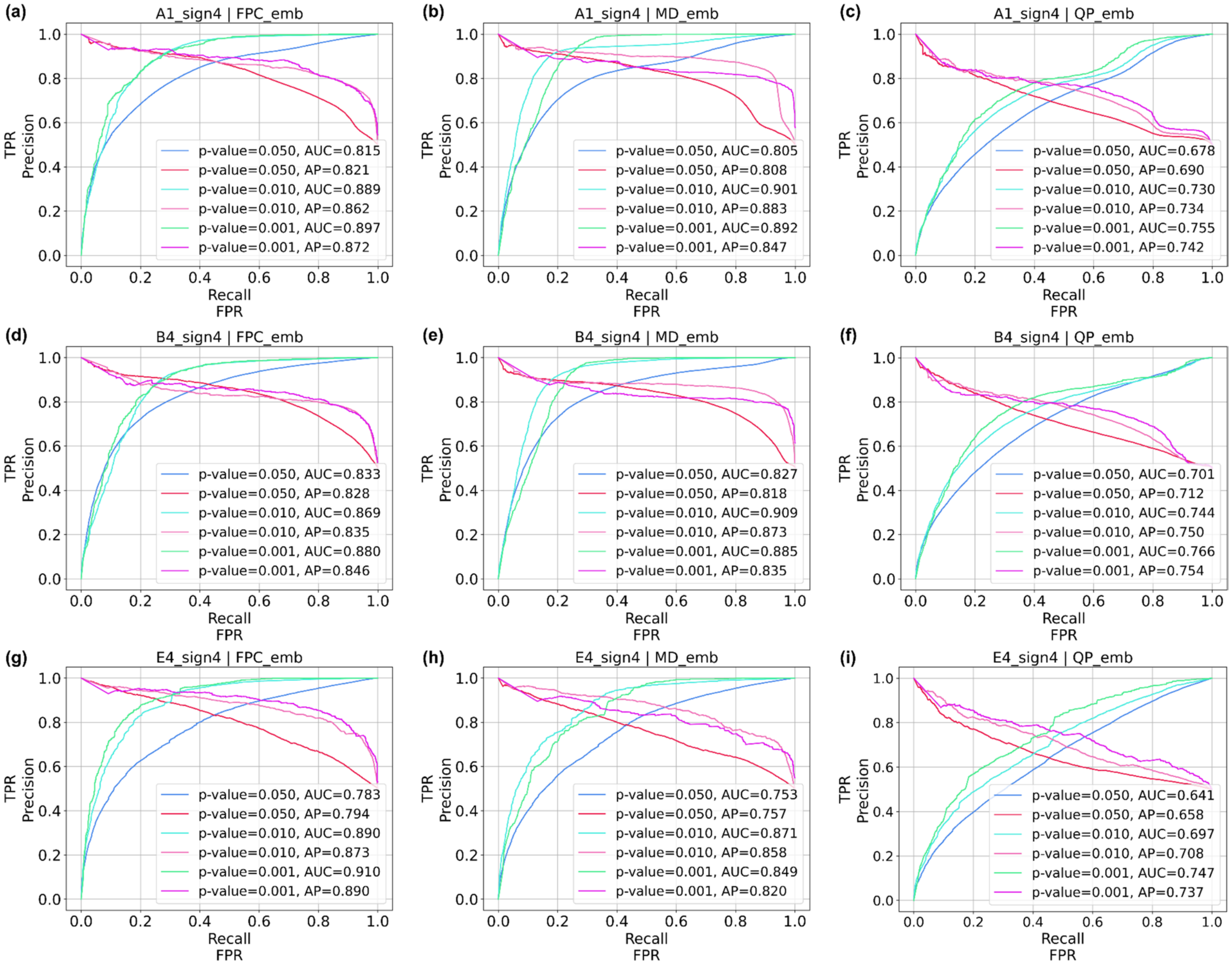

Fig. 4 The receiver operating characteristic (ROC) and precision-recall (PR) curves and the corresponding area under the curve (AUC) and average precision (AP) values comparing APDB embeddings of **a**, **d**, **g** fingerprint counts (FPC), **b**, **d**, **e** molecular descriptors (MD), and **c**, **f**, **i** quantum properties (QP) spaces with CC bioactivity signatures of **a–c** 2D fingerprints (A1), **d–f** binding (B4), and **g–i** diseases and toxicology (E4) datasets

that average precision scores are high in all datasets when considering a sampling ratio of 0.1 and 0.2, while are substantially decreasing when the number of negatives is 100 times more than the number of positives, particularly for the A1A2 and CCchem datasets (Fig. 7a–e). Furthermore, we can conclude that by adopting an appropriate sampling strategy to select negative instances, the model performances are significantly improved compared to a random approach, as evidenced by the decrease in average precision values when considering randomly sampled negatives (AP1 versus AP2 values in Fig. 7a–e).

As for external validation, we report the prediction results for all tested scenarios (see subsection *External validation dataset construction*), i.e., the novel pairs present in current PubChem bioassays, the interactions with the targets found enriched in the CC B2 (metabolic genes) space and the unseen targets in the CC B4 (binding data) space, and the CTIs with the CYP and the CA enzymes from the CTD. Performances are given in terms of recognition rate, or recall, since we are seeking a model primarily capable of recognizing positive instances. Concerning PubChem pairs, predictions are outstanding for both MFP and APchem datasets (Fig. 8a, c, d), which allow retrieval of almost all positive pairs, while CC signatures provide smaller recall values (Fig. 8b, e). For B2 and B4 targets, the maximum recall score is observed when using CC features (Fig. 8b, e) and comparable values are obtained with APchem features (Fig. 8a, d), whereas the MFP dataset leaks a large portion of hits (Fig. 8c). For CYP enzymes, the APchem dataset gives the best performances (Fig. 8d), followed by the F1F2 dataset (Fig. 8a) and the MFP dataset (Fig. 8c), while CC signatures show again a lower true positive rate (Fig. 8b, e); the same applies to CA targets, for which MFP and APchem scores are almost comparable (Fig. 8c, d), whereas the CC datasets report less successful hits (Fig. 8b, e). However,

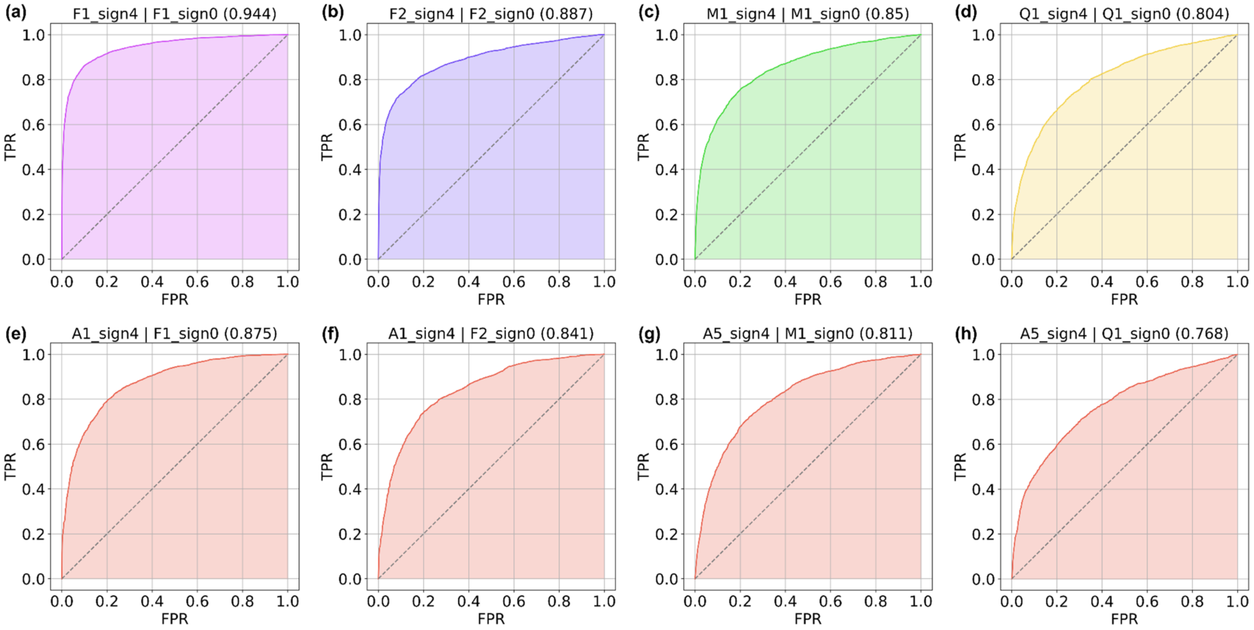

Fig. 5 The recapitulation of *signature 0* (original data in APDB) starting from *signature 4* (bioactivity signature) in terms of neighboring molecules in the **a** fingerprint bits (F1), **b** fingerprint counts (F2), **c** molecular descriptors (M1), and **d** quantum properties (Q1) spaces, respectively. The recapitulation of *signature 0* of **e**, **f** fingerprint bits (F1) and counts (F2) starting from *signature 4* of 2D fingerprints (A1). The recapitulation of *signature 0* of **g**, **h** molecular descriptors (M1) and quantum properties (Q1) starting from *signature 4* of physicochemical parameters (A5)

considering the applicability domain of the model, the scores improve considerably for the A1A2 and CCchem datasets, especially for interactions with CA targets, while remaining consistent for the other datasets.

Overall, we can deduce that the APchem dataset demonstrates a more stable behaviour compared to the others in finding positive associations. Therefore, new CTIs can be robustly predicted by our pre-trained estimators.

Conclusions

In this work, we proposed an in silico methodological pipeline to identify novel compound-target interactions by computing and integrating the bioactivity signature of air-polluting molecules with target sequence descriptors into a single vector representation. Indeed, representing their properties in the form of numerical vectors allowed us to apply established machine learning models to identify unknown pairs of interacting molecules with similar bioactivity behaviour. The outlined methodology, through the inclusion and inference of biological information related to small compounds, demonstrated its generalization capabilities on external data focused on specific targets for air pollution, providing support for future predictions. Moreover, the bioactivity signatures derived specifically for air pollutants demonstrated an improved consistency in reporting positive associations compared to other datasets' features.

As a future step, we will evaluate different negative selection strategies, e.g. sampling around the mean [45], which we demonstrated to be valuable compared to a random approach.

We also supply a web application at ap-bio.streamlit.app to compute molecule and target feature vectors, visualize the 2D t-SNE representation of molecule signatures focusing on the query molecule and its nearest neighbors, and predict compound-target interactions from molecule SMILES and target UniProt.

In this way, unknown ligands can be prioritized for a given target (or vice versa) and can subsequently be tested through molecular docking simulations or validated via in vitro experiments for further studies on gene expression, chemical toxicity, and hazard evaluation.

A final aspect to be emphasised is the field of application of this methodology. Air pollutants are today one of the most important threats for human health [64–66] and there are still few free resources for the study of the mechanisms of action behind their toxicological effect. In this scenario, APBIO stands as a noteworthy free tool for

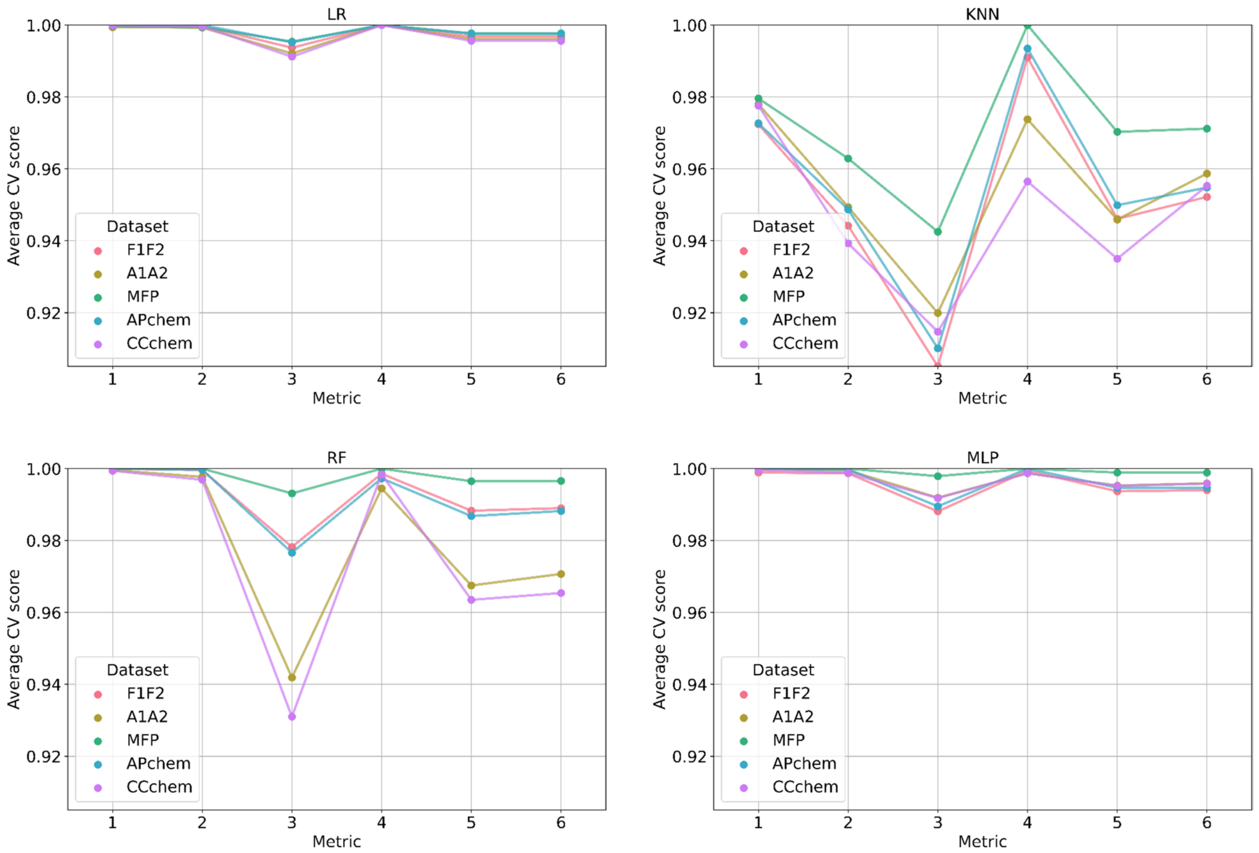

Fig. 6 Nested cross-validation performances of Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Multilayer Perceptron (MLP) models in terms of area under the receiver operating characteristic curve (ROC-AUC) (1), average precision (AP) (2), recall (3), precision (4), f1 (5), and balanced accuracy (6) for the F1F2, A1A2, MFP, APchem, and CCchem datasets

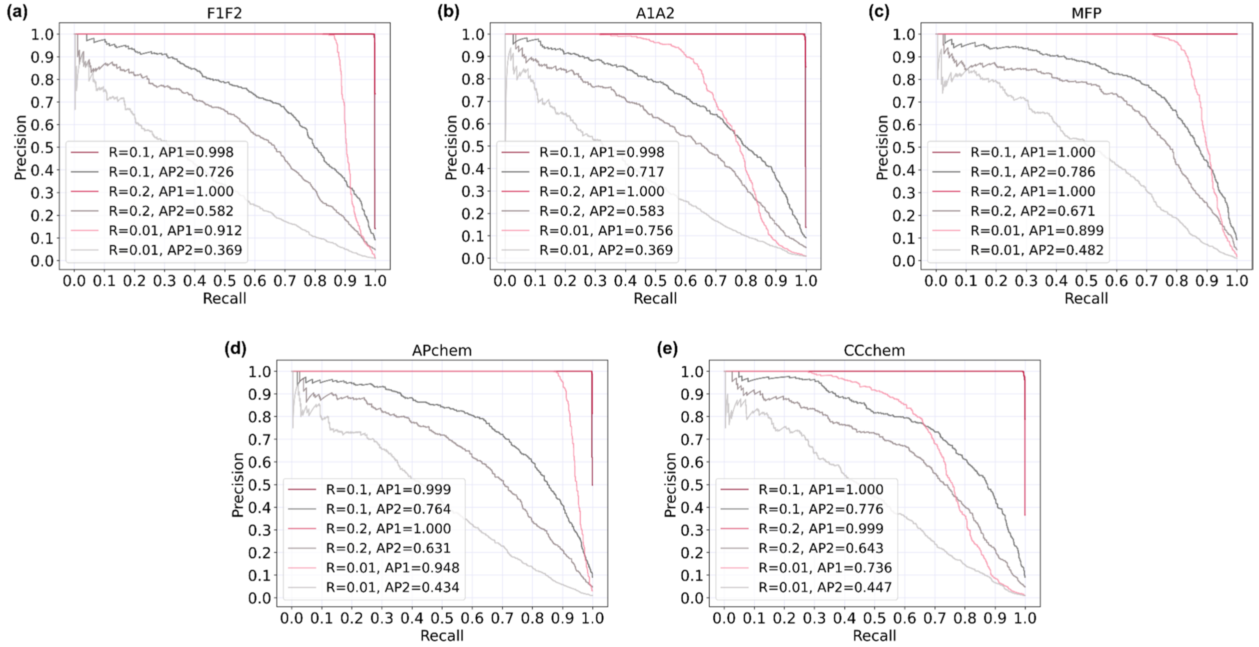

Fig. 7 Performances of OCSVM and random negative selection strategies in terms of precision-recall curve and average precision value of the final classifier (AP1 and AP2, respectively) on the train-test split with test size 0.3 for each (**a–e**) dataset with positive-to-negative ratio of 0.1, 0.2, and 0.01

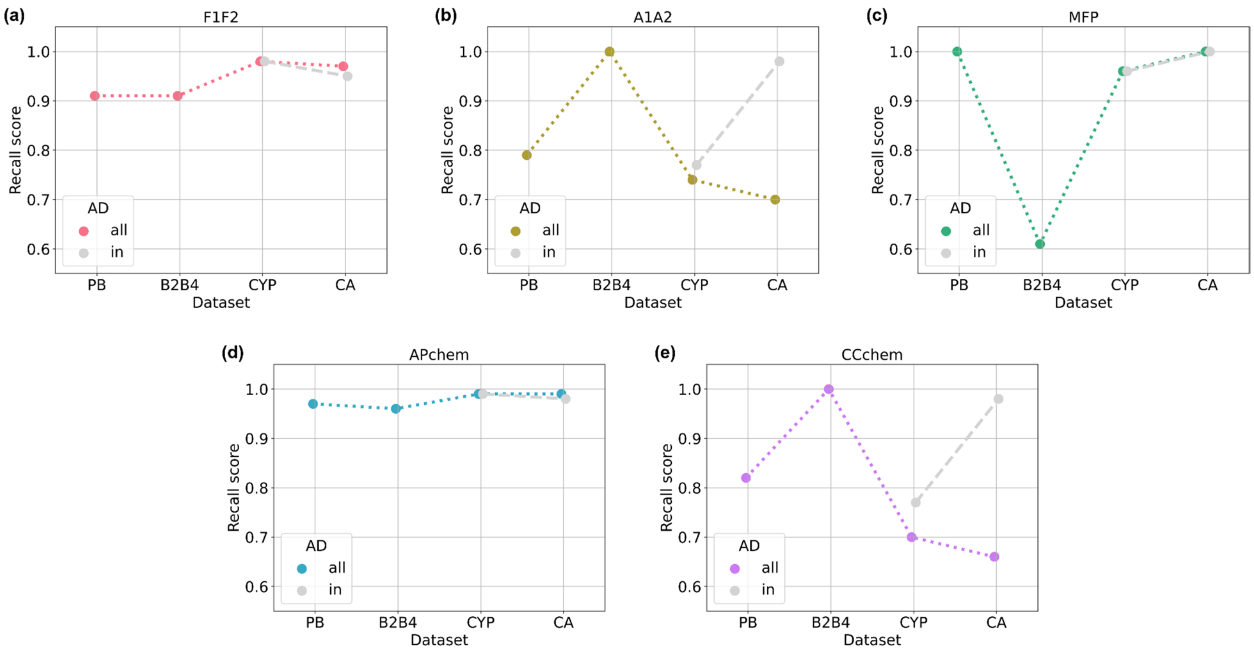

Fig. 8 Recall scores for each tested scenario on external validation datasets, namely, PubChem (PB) data, CC metabolic genes and binding data (B2B4), cytochrome P450 enzymes (CYP) and carbonic anhydrases (CA) from CTD for the **a** F1F2, **b** A1A2, **c** MFP, **d** APchem, and **e** CCchem datasets. The coloured dots are the scores for all interaction pairs (all), while the grey dots are the scores for the interaction pairs within the applicability domain (in)

the prediction of air pollutants targets, allowing scientists to deepen possible biological processes in pollutants toxicology. For this reason, our group is working to deepen the application of chemoinformatics and bioinformatics to unveil the role of pollutants from different matrices in different diseases.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AAC	Amino acid composition
AD	Applicability domain
AHR	Aryl hydrocarbon receptor
AP	Average precision
APBIO	Air pollutant bioactivity
APDB	Database of air pollutants
ATC	Anatomical therapeutic chemical
AUC	Area under the curve
CA	Carbonic anhydrase
CC	Chemical checker
CYP	Cytochrome P450
CTD	Comparative toxicogenomics database
CTDC	Composition-transition-distribution composition
CTDD	Composition-transition-distribution distribution
CTIs	Compound-target interactions
CTDT	Composition-transition-distribution transition
CV	Cross-validation
DNN	Deep neural network
DPC	Dipeptide composition
EPA	Environmental Protection Agency
FPB	Fingerprint bits
FPC	Fingerprint counts
KNN	K-nearest neighbors
LR	Logistic regression
LSI	Latent semantic indexing
MD	Molecular descriptors
ML	Machine learning
MLP	Multi-layer perceptron
MoA	Mechanism of action
NMBroto	Normalized Moreau-Broto
NN	Nearest neighbors
OCSVM	One-class support vector machine
PAAC	Pseudo-amino acid composition
PAHs	Polycyclic aromatic hydrocarbons
PCA	Principal component analysis
PM	Particulate matter
PR	Precision-recall
QP	Quantum properties
QSOrder	Quasi-sequence-order
RF	Random forest
ROC	Receiver operating characteristic curve
SEA	Similarity ensemble approach
SLC	Solute carrier
SNN	Siamese neural network
SOCNumber	Sequence-order-coupling number
t-SNE	T-distributed stochastic neighbor embedding
VOCs	Volatile organic compounds
VS	Virtual screening

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00961-1.

Supplementary Material 1.

Acknowledgements

The authors would like to thank the 'National Biodiversity Future Center' (identification code CN00000033, CUP B73C21001300006) on 'Biodiversity', financed under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 1.4 'Strengthening of research structures and creation of R&D "national champions" on some key enabling technologies'—Call for Tender No. 3138 of 16 December 2021, rectified by Decree No. 3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union-NextGenerationEU; Award No.: Project Code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B73C21001300006, Project title 'National Biodiversity Future Center—NBFC', the #NEXTGENERATIONEU and the Italian Ministry of University and Research, National Recovery and Resilience Plan (PNRR), projects MNESYS (PE0000006), HEAL ITALIA (PE0000019), and the "PREPARE" project (n. F/310130/05/X56—CUP: B39J23001730005)—D.M. MiSE 31/12/2021. We gratefully acknowledge the help and support provided by the members of the SBNB laboratory at the Institute for Research in Biomedicine (IRB) in Barcelona. In particular, we would like to sincerely thank Arnau Comajuncosa-Creus, Dylan Dalton, Gema Rojas Granado, and Elena Pareja Lorente. We are also deeply grateful to Giada De Simone for her assistance and contribution to this work.

Author contributions

All authors conceived the idea and designed the study. E.V. coded the software, performed all data analyses, and prepared the data repository. All authors discussed the findings. E.V. and R.G. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work has been partially supported by #NEXTGENERATIONEU and the Italian Ministry of University and Research, National Recovery and Resilience Plan (PNRR), projects MNESYS (PE0000006), HEAL ITALIA (PE0000019), National Biodiversity Future Center – NBFC (CN_00000033), and the "PREPARE" project (n. F/310130/05/X56—CUP: B39J23001730005)—D.M. MiSE 31/12/2021.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Computer Science, University of Verona, Verona, Italy. [2]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. [4]Molecular Informatics Unit, Ri.MED Foundation, Palermo, Italy. [5]NBFC, National Biodiversity Future Center, Palermo, Italy.

Received: 19 September 2024 Accepted: 20 January 2025
Published online: 31 January 2025

References

1. Vineis P, Husgafvel-Pursiainen K (2005) Air pollution and cancer: biomarker studies in human populations. Carcinogenesis 26:1846–1855. https://doi.org/10.1093/carcin/bgi216
2. Turner MC, Andersen ZJ, Baccarelli A et al (2020) Outdoor air pollution and cancer: an overview of the current evidence and public health recommendations. CA Cancer J Clin 70:460–479. https://doi.org/10.3322/caac.21632
3. Chuang K-J, Chan C-C, Su T-C et al (2007) The effect of urban air pollution on inflammation, oxidative stress, coagulation, and autonomic dysfunction in young adults. Am J Respir Crit Care Med 176:370–376. https://doi.org/10.1164/rccm.200611-1627OC
4. Arias-Pérez RD, Taborda NA, Gómez DM et al (2020) Inflammatory effects of particulate matter air pollution. Environ Sci Pollut Res 27:42390–42404. https://doi.org/10.1007/s11356-020-10574-w
5. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and health impacts of air pollution: a review. Front Public Health. https://doi.org/10.3389/fpubh.2020.00014
6. Costa LG, Cole TB, Dao K et al (2020) Effects of air pollution on the nervous system and its possible role in neurodevelopmental and neurodegenerative disorders. Pharmacol Ther 210:107523. https://doi.org/10.1016/j.pharmthera.2020.107523
7. Viesi E, Sardina DS, Perricone U, Giugno R (2023) APDB: a database on air pollutant characterization and similarity prediction. Database. https://doi.org/10.1093/database/baad046
8. Agamah FE, Mazandu GK, Hassan R et al (2020) Computational/in silico methods in drug target and lead prediction. Brief Bioinform 21:1663–1675. https://doi.org/10.1093/bib/bbz103
9. Ezzat A, Wu M, Li X-L, Kwoh C-K (2019) Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 20:1337–1357. https://doi.org/10.1093/bib/bby002
10. Jacob L, Vert J-P (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24:2149–2156. https://doi.org/10.1093/bioinformatics/btn409
11. Vázquez J, López M, Gibert E et al (2020) Merging ligand-based and structure-based methods in drug discovery: an overview of combined virtual screening approaches. Molecules 25:4723. https://doi.org/10.3390/molecules25204723
12. Ferreira L, Dos Santos R, Oliva G, Andricopulo A (2015) Molecular docking and structure-based drug design strategies. Molecules 20:13384–13421. https://doi.org/10.3390/molecules200713384
13. Keiser MJ, Roth BL, Armbruster BN et al (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25:197–206. https://doi.org/10.1038/nbt1284
14. Yu H, Chen J, Xu X et al (2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLoS ONE 7:e37608. https://doi.org/10.1371/journal.pone.0037608
15. Chen X, Yan CC, Zhang X et al (2016) Drug–target interaction prediction: databases, web servers and computational models. Brief Bioinform 17:696–712. https://doi.org/10.1093/bib/bbv066
16. Zhao L, Zhu Y, Wang J et al (2022) A brief review of protein–ligand interaction prediction. Comput Struct Biotechnol J 20:2831–2838. https://doi.org/10.1016/j.csbj.2022.06.004
17. Ding H, Takigawa I, Mamitsuka H, Zhu S (2014) Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Brief Bioinform 15:734–747. https://doi.org/10.1093/bib/bbt056
18. Sachdev K, Gupta MK (2019) A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform 93:103159. https://doi.org/10.1016/j.jbi.2019.103159
19. Cereto-Massagué A, Ojeda MJ, Valls C et al (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005
20. Sydow D, Burggraaff L, Szengel A et al (2019) Advances and challenges in computational target prediction. J Chem Inf Model 59:1728–1742. https://doi.org/10.1021/acs.jcim.8b00832
21. Duran-Frigola M, Pauls E, Guitart-Pla O et al (2020) Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol 38:1087–1096. https://doi.org/10.1038/s41587-020-0502-7
22. Bertoni M, Duran-Frigola M, Badia-i-Mompel P et al (2021) Bioactivity descriptors for uncharacterized chemical compounds. Nat Commun 12:3932. https://doi.org/10.1038/s41467-021-24150-4
23. Li ZR, Lin HH, Han LY et al (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 34:W32–W37. https://doi.org/10.1093/nar/gkl305
24. Rao HB, Zhu F, Yang GB et al (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 39:W385–W390. https://doi.org/10.1093/nar/gkr284
25. Shen H-B, Chou K-C (2008) PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388. https://doi.org/10.1016/j.ab.2007.10.012
26. Xiao N, Cao D-S, Zhu M-F, Xu Q-S (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of

protein sequences. Bioinformatics 31:1857–1859. https://doi.org/10.1093/bioinformatics/btv042
27. Chen Z, Zhao P, Li F et al (2018) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34:2499–2502. https://doi.org/10.1093/bioinformatics/bty140
28. Ong SA, Lin HH, Chen YZ et al (2007) Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinform 8:300. https://doi.org/10.1186/1471-2105-8-300
29. Zdrazil B, Felix E, Hunter F et al (2024) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192. https://doi.org/10.1093/nar/gkad1004
30. Gilson MK, Liu T, Baitaluk M et al (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053. https://doi.org/10.1093/nar/gkv1072
31. Knox C, Wilson M, Klinger CM et al (2024) DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res 52:D1265–D1275. https://doi.org/10.1093/nar/gkad976
32. Davis AP, Wiegers TC, Johnson RJ et al (2023) Comparative toxicogenomics database (CTD): update 2023. Nucleic Acids Res 51:D1257–D1262. https://doi.org/10.1093/nar/gkac833
33. Wang Y, Bryant SH, Cheng T et al (2017) PubChem BioAssay: 2017 update. Nucleic Acids Res 45:D955–D963. https://doi.org/10.1093/nar/gkw1118
34. Liu H, Sun J, Guan J et al (2015) Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics 31:i221–i229. https://doi.org/10.1093/bioinformatics/btv256
35. Zheng Y, Peng H, Zhang X et al (2019) Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces. BMC Bioinform 20:605. https://doi.org/10.1186/s12859-019-3238-y
36. Schölkopf B, Williamson RC, Smola A, et al (1999) Support vector method for novelty detection. In: Solla S, Leen T, Müller K (eds) Advances in Neural Information Processing Systems. MIT Press
37. Pimentel MAF, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Signal Process 99:215–249. https://doi.org/10.1016/j.sigpro.2013.12.026
38. U.S. Environmental Protection Agency SPECIATE. https://www.epa.gov/air-emissions-modeling/speciate. Accessed 26 Oct 2023
39. U.S. Environmental Protection Agency Initial List of Hazardous Air Pollutants with Modifications. https://www.epa.gov/haps/initial-list-hazardous-air-pollutants-modifications. Accessed 26 Oct 2023
40. Kim S, Thiessen PA, Bolton EE, Bryant SH (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. Nucleic Acids Res 43:W605–W611. https://doi.org/10.1093/nar/gkv396
41. Kim S, Thiessen PA, Cheng T et al (2018) An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res 46:W563–W570. https://doi.org/10.1093/nar/gky294
42. Han L, Wang Y, Bryant SH (2009) A survey of across-target bioactivity results of small molecules in PubChem. Bioinformatics 25:2251–2255. https://doi.org/10.1093/bioinformatics/btp380
43. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169. https://doi.org/10.1093/nar/gkw1099
44. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
45. Mei S, Zhu H (2015) A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. Sci Rep 5:8034. https://doi.org/10.1038/srep08034
46. Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6:10. https://doi.org/10.1186/1758-2946-6-10
47. Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97. https://doi.org/10.1093/nar/gkw377
48. Raudvere U, Kolberg L, Kuzmin I et al (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47:W191–W198. https://doi.org/10.1093/nar/gkz369
49. Eden E, Navon R, Steinfeld I et al (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinform 10:48. https://doi.org/10.1186/1471-2105-10-48
50. Sherman BT, Hao M, Qiu J et al (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 50:W216–W221. https://doi.org/10.1093/nar/gkac194
51. Boyle EI, Weng S, Gollub J et al (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–3715. https://doi.org/10.1093/bioinformatics/bth456
52. Rivals I, Personnaz L, Taing L, Potier M-C (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23:401–407. https://doi.org/10.1093/bioinformatics/btl633
53. Reimand J, Isserlin R, Voisin V et al (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 14:482–517. https://doi.org/10.1038/s41596-018-0103-9
54. Lee S, Lee DK (2018) What is the proper way to apply the multiple comparison test? Korean J Anesthesiol 71:353–360. https://doi.org/10.4097/kja.d.18.00242
55. Zhang Y, Zhang Y, Sun K et al (2019) The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. J Mol Cell Biol 11:1–13. https://doi.org/10.1093/jmcb/mjy052
56. Pérez-Escuredo J, Van Hée VF, Sboarina M et al (2016) Monocarboxylate transporters in the brain and in cancer. Biochim Biophys Acta - Mol Cell Res 1863:2481–2497. https://doi.org/10.1016/j.bbamcr.2016.03.013
57. Puris E, Saveleva L, Górová V et al (2022) Air pollution exposure increases ABCB1 and ASCT1 transporter levels in mouse cortex. Environ Toxicol Pharmacol 96:104003. https://doi.org/10.1016/j.etap.2022.104003
58. Vogel CFA, Van Winkle LS, Esser C, Haarmann-Stemmann T (2020) The aryl hydrocarbon receptor as a target of environmental stressors—implications for pollution mediated stress and inflammatory responses. Redox Biol 34:101530. https://doi.org/10.1016/j.redox.2020.101530
59. Lionetto M, Caricato R, Giordano M et al (2012) Carbonic anhydrase as pollution biomarker: an ancient enzyme with a new use. Int J Environ Res Public Health 9:3965–3977. https://doi.org/10.3390/ijerph9113965
60. Rouillard AD, Gundersen GW, Fernandez NF et al (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016:baw100. https://doi.org/10.1093/database/baw100
61. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic classification methods and their applicability domain. Mol Inform 35:160–180. https://doi.org/10.1002/minf.201501019
62. Klingspohn W, Mathea M, ter Laak A et al (2017) Efficiency of different measures for defining the applicability domain of classification models. J Cheminform 9:44. https://doi.org/10.1186/s13321-017-0230-2
63. Sahigara F, Mansouri K, Ballabio D et al (2012) Comparison of different approaches to define the applicability domain of qsar models. Molecules 17:4791–4810. https://doi.org/10.3390/molecules17054791
64. Apte JS, Manchanda C (2024) High-resolution urban air pollution mapping. Science 385:380–385. https://doi.org/10.1126/science.adq3678
65. Lewis A, Misonne D, Scotford E (2024) Harnessing science, policy, and law to deliver clean air. Science 385:362–366. https://doi.org/10.1126/science.adq4721
66. Huang W, Xu H, Wu J et al (2024) Toward cleaner air and better health: current state, challenges, and priorities. Science (80-) 385:386–390. https://doi.org/10.1126/science.adp7832

Publisher's Note