ApisTox: a new benchmark dataset for the classification of small molecules toxicity on honey bees

Jakub Adamczyk[1,†,*], **Jakub Poziemski**[2,†], **and Paweł Siedlecki**[2]

[1]AGH University of Krakow, Department of Computer Science, Cracow, Poland
[2]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[*]corresponding author: Jakub Adamczyk (jadamczy@agh.edu.pl)
[†]these authors contributed equally to this work

ABSTRACT

The global decline in bee populations poses significant risks to agriculture, biodiversity, and environmental stability. To bridge the gap in existing data, we introduce ApisTox, a comprehensive dataset focusing on the toxicity of pesticides to honey bees (Apis mellifera). This dataset combines and leverages data from existing sources such as ECOTOX and PPDB, providing an extensive, consistent, and curated collection that surpasses the previous datasets. ApisTox incorporates a wide array of data, including toxicity levels for chemicals, details such as time of their publication in literature, and identifiers linking them to external chemical databases. This dataset may serve as an important tool for environmental and agricultural research, but also can support the development of policies and practices aimed at minimizing harm to bee populations. Finally, ApisTox offers a unique resource for benchmarking molecular property prediction methods on agrochemical compounds, facilitating advancements in both environmental science and cheminformatics. This makes it a valuable tool for both academic research and practical applications in bee conservation.

Background & Summary

Global declines of bee population are a serious threat to agricultural production, environmental stability, and overall biodiversity. Because bees are essential to the pollination of plants, protecting bee populations is of crucial importance for environmental preservation and food security. To evaluate the effects of different stressors, such as pesticides, infections, and environmental changes, on bee health, thorough and trustworthy data are needed. This necessity motivated the creation of ApisTox, a new state-of-the-art dataset on pesticide toxicity for honey bees (*Apis mellifera*), useful for the research and development community.

Existing datasets on bee toxicity, such as BeeTox[1] and subsets of ECOTOX[2], offer valuable insights but are limited by their scope, consistency, and the comprehensiveness of data. Furthermore, data on the effects of various crop management systems, as captured in CropCSM[3] datasets, are often disconnected from bee toxicity information, hindering holistic analyses. To address these gaps, we have developed a comprehensive dataset that combines carefully filtered and manually curated data from the above-mentioned resources. This integration makes it the largest and most consistent dataset on pesticide bee toxicity currently available.

The motivation behind creating this dataset stems from the need to understand the interactions between bees and their environments. The global decline in bee populations is a complex issue influenced by many factors, including exposure to pesticides and other human introduced variables, and climate change. By consolidating and curating data from multiple sources, our dataset allows for an assessment of a broader chemical space and provides a more coherent and comprehensive basis for analysis.

Our dataset encompasses a wide range of data types, including but not limited to, toxicity levels of various pesticides, herbicides, insecticides and other chemicals, with respect of their time of registration and references to external chemical databases and source publications. The data have been carefully filtered to ensure accuracy, relevance, and consistency, addressing a critical need for high-quality, standardized data in bee research.

ApisTox is also one of the very few datasets outside medicinal chemistry that can be used for benchmarking molecular property prediction methods for graph classification. Predictive models utilizing molecular graphs are almost exclusively evaluated using data from de novo drug design, whereas agrochemical compounds possess quite different structural characteristics. Our dataset is also large enough for training and testing more data-demanding models, being larger than e.g. 17 datasets from Therapeutics Data Commons (TDC) benchmark[4]. ApisTox can serve as a part of challenging benchmarks for novel classification algorithms on molecular graphs, including out-of-distribution testing of models designed using medicinal

chemistry data.

The significance of this dataset extends beyond academic research. It can be useful for policy decisions and strategies, and for developing agricultural practices which minimize harm to bees. The dataset can help screen for bee-friendly chemicals and natural products, which help with the development of agricultural systems that promote bee health. Moreover, the dataset is designed with accessibility and interoperability in mind. This is intended to foster collaboration among researchers and encourage the development of novel, innovative solutions to protect bee populations. In conclusion, the creation of this new, state-of-the-art dataset for bee toxicity addresses a critical gap in existing data resources, offering a comprehensive, reliable, and accessible tool for studying the complex factors affecting bee health.

Methods

We gather data about the acute honey bee toxicity of pesticides, based on their median lethal dose (LD50). We enrich it with additional metadata, which allows deeper analyses. Dataset is assembled from various sources, with extensive preprocessing, cleaning and deduplication. Those steps are described below in this section. The resulting files are described in Data Records section, with their characteristics and quality analyzed in Technical Validation section.

Data sources

We base our dataset on three pre-cleaned, high quality data sources, widely utilized in environmental science[5–7]. They vary in their detail, format and provided metadata.

ECOTOX[2] database is maintained by US EPA (United States Environmental Protection Agency) and consists of relatively raw data and experimental measurements. For each substance, it typically has many entries from different sources, with varied measurement values, and is the most comprehensive data source for ecotoxicology data. However, it provides relatively less structured data than some other databases, and in particular does not contain SMILES strings, typically processed by computational methods, instead relying on CAS (Chemical Abstracts Service) registry numbers. It is updated quarterly, and we use the version updated on 14th December 2023, which covers almost 1.2 million measurements and over 13 thousand chemicals.

Pesticide Properties DataBase (PPDB)[8] is a database that catalogs defined pesticide chemical entities, i.e. active ingredients, along with their physicochemical properties, ecotoxicological data, environmental fate, and human health impacts. It was created and is maintained by the Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire. PPDB is a crucial source of curated, structured data for analyzing pesticide applications and facilitating risk management.

Bio-Pesticides DataBase (BPDB), also curated by the University of Hertfordshire, is our third source. It provides detailed information on the properties, efficacy, and application of pesticides originating from natural sources, such as microorganisms, plant extracts, and pheromones. BPDB entries typically comprise multiple components and lack a defined active ingredient.

Both the PPDB and BPDB are reviewed, managed, and updated through literature and legal resources, following established data curation protocols and guidelines. Those databases received endorsements from leading chemical and agrochemical organizations, including the International Union of Pure and Applied Chemistry (IUPAC) and the Food and Agriculture Organization (FAO). They are continuously updated, and we use data from 22nd February 2024.

All datasets undergo data processing and cleaning procedures, in order to unify their structure and fill missing fields. The entire workflow has been summarized in Figure 1. During the entire data cleaning and processing, we save all removed rows, along with removal reason, in a separate file, and include it along the rest of the raw and processed data (see Data Records section for details). Manual analysis of this data allowed us to verify the correctness of processing steps. In further sections, we describe them, and their rationale, in detail.

ECOTOX processing

ECOTOX database offers a rich set of filters, allowing flexible retrieval of data. We apply the following ones:

1. Effect “mortality” - effects are observed responses of different types when applying a given chemical to selected species. Mortality is the most common target for toxicological studies.

2. Endpoint “LD50” - endpoints are measurable numerical targets for given effects. We use median lethal dose (LD50), for consistency with PPDB and BPDB.

3. Species contains “apis mellifera” - we select *Apis mellifera* (honey bee) data. We use contains filter instead of exact match, because while the vast majority of measurements use the general species name, some also specify the subspecies, most commonly *Apis mellifera mellifera* (European dark bee).

4. Test locations “lab” - we exclude less reliable field measurements and measurements without stated test location. This filters out 10 rows.

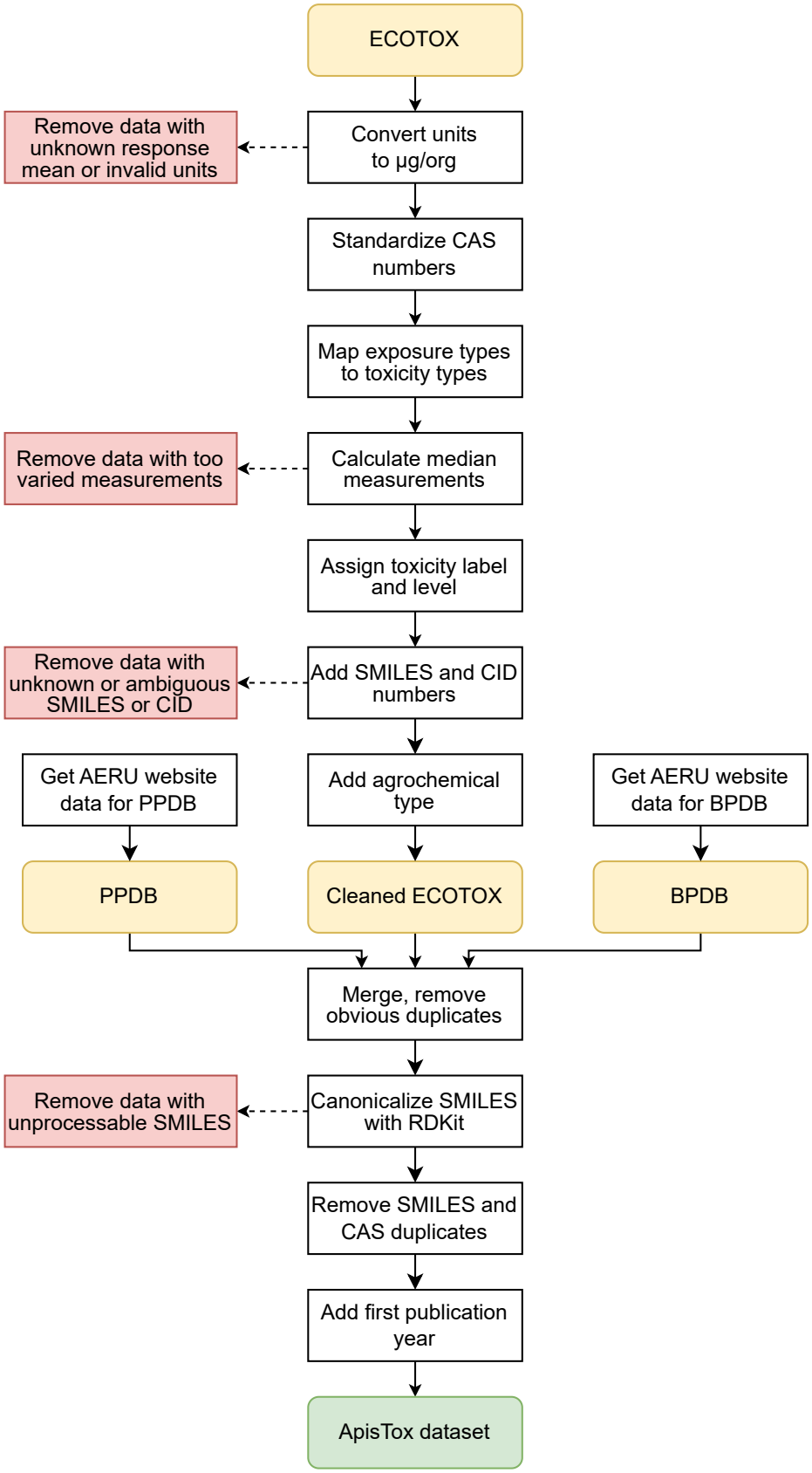

Figure 1. Data processing workflow

We do not filter data by observed duration, because it contains measurements with at most 8 days, which still corresponds to acute toxicity, and most measurements are for 72 hours or less. The resulting set contains 2674 raw measurements, which then undergo a series of processing steps to obtain data aggregated per pesticide. For each measurement, we select substance name, CAS number, exposure type, observed response mean and unit. They identify the pesticide used, how it was applied, and measured LD50.

Firstly, we remove all rows without mean response value, marked as "NR" (Not Reported). Then, we standardize measurements to unit μg / bee (also known as μg / org). This unit is used by PPDB, BPDB, and US EPA guidelines for measuring bee toxicity[9]. Overall, there are over 40 different units used. We convert those that can be unambiguously mapped to μg / bee (e.g. "pg/org", "ug/bee"), and remove all other samples. Removed units are, e.g. "%", "ae ug/org:x" or "mg/cm2".

ECOTOX uses non-standard notation for CAS numbers, without parentheses. We normalize such cases, e.g. changing 1194656 to 1194-65-6.

Toxicity is typically divided into oral, contact, or other way of application, and PPDB and BPDB also use this system. However, ECOTOX applies a more fine-grained classification, and as a first step of processing, we map the specific exposure types into three standardized toxicity types as follows:

- "Diet, unspecified", "Drinking water", "Food" - oral
- "Dermal", "Direct application", "Topical, general" - contact
- "Multiple routes between application groups", "Oral via capsule", "Spray, unspecified", "Environmental, unspecified" - other

To obtain a dataset with a single pesticide per row, raw measurements from ECOTOX need to be aggregated. EPA guidelines[9] suggest 11 μg / org as the threshold for acute bee toxicity, i.e. LD50 at or below this value marks a pesticide definitely toxic for honey bees. PPDB and BPDB[10], for consistency with EU and UK regulatory institutions, use 1 μg / org and 100 μg / org as thresholds, i.e. LD50 equal to or lower than 1 means highly toxic pesticide, between 1 and 100 moderately toxic, and higher than 100 means non-toxic or slightly toxic substance. We therefore calculate two such labels for each pesticide: binary non-toxic/toxic (EPA label), and ternary non-toxic/moderately toxic/highly toxic (PPDB level). They are numerically encoded as 0/1 and 0/1/2, respectively.

We identify each distinct substance by the CAS number, and for each one we calculate two labels for every toxicity type (oral, contact, other) available for that pesticide. Those measurements vary strongly for some pesticides, so we take a conservative approach here, to ensure high data quality. If the lowest and highest measurement agree for the EPA binary label, i.e. all measurements are below or above 11 μg / org, we assign positive or negative class, respectively. Otherwise, we mark the toxicity label as "Unspecified", and remove such rows. Then we assign PPDB ternary level based on median measurement. Lastly, for each pesticide, we take the strongest toxicity type (e.g. with the highest level), as representing the most toxic effect the pesticide can have on bees. We also save the information which toxicity type was the strongest for each pesticide.

Further, we add SMILES strings and PubChem CID (Compound ID) numbers. SMILES strings are required by computational libraries, and CID numbers enable easy and unambiguous lookup of molecules in PubChem[11], the largest openly available database of chemical information. We utilize PUG REST API[12], mapping CAS numbers to SMILES strings and CID numbers. We remove those molecules for which that operation was impossible or ambiguous.

Lastly, we add information about agrochemical type for each substance. In PubChem, compounds can have a separate section "Agrochemical information" with descriptions of agricultural applications. Using PUG REST API, we fetch descriptions from this section for each substance, and we search for keywords "herbicide", "fungicide" and "insecticide". If any of those words appear in the description, we note that information as agrochemical type. If none are found, but the compound has this section, it means that it has other agrochemical applications, e.g. as a growth agent or fertilizer, and we mark it as "other agrochemical". This creates a total of four boolean (binary) variables. In the last case, when the compound page does not have "Agrochemical information" section at all, it can have all zeros, meaning "unknown" pesticide type. We still include those substances, since we have bee toxicity measurements for them. They could have been researched as potential pesticides, but not yet registered or used.

PPDB and BPDB

Both PPDB and BPDB databases have the same structure organized in tables, available via web pages. We extract all relevant values using regular expressions on HTML responses. This approach always utilizes the latest available data. We downloaded the data on 22nd February 2024.

For each pesticide, LD50 values for oral, contact, and other honey bee toxicity can be provided. Often no measurement, or only one or two are available, and we ignore substances without any toxicity measurement for honey bees. If we have more than one measurement, we take the lowest value, i.e. the strongest toxicity. Additionally, some values are provided in a non-standard

format, especially for low toxicity, and marked as "Low", "Non-toxic", or in scientific notation, e.g. "10^3". We normalize those cases as non-toxic, i.e. label and level 0.

We combine "Summary", "Description" and "Pesticide type" fields into a single text to extract the pesticide type. We search for keywords "herbicide", "fungicide" and "insecticide", similarly to processing PubChem data for ECOTOX. Since PPDB and BPDB databases contain only agrochemicals, if none of those words appear, we mark a given substance as "other agrochemical".

CAS numbers and SMILES strings are available for almost all pesticides, and we exclude ones without this information. CID numbers are also often available, but in case they are missing, we fill them using PubChem and PUG REST API, based on CAS and SMILES.

Combining datasets

After gathering three preprocessed datasets, we merge them. Firstly, we concatenate all rows and drop obvious duplicates, i.e. compounds with the same CAS, SMILES and label.

We ensure that the dataset consists of only distinct molecules. All SMILES are canonicalized using RDKit 2023.9.5[13], and this way we also make sure that all SMILES strings are valid and processable by this software. We then drop all duplicates, first using SMILES, and then using CAS. When removing duplicates, we keep rows from datasets in the order of preference: PPDB, BPDB, ECOTOX. This was motivated by the fact that PPDB and BPDB are additionally manually verified, and therefore can have slightly higher quality.

Lastly, we add the first publication year information to all molecules, using their CID numbers and PubChem literature records. While this is not exact information about first usage of a given substance as a pesticide, it still provides a good approximation of when a given compound was first created. We noticed that PubChem has three obvious errors in this regard, attributing an unreasonably large number of molecules to three publications[14–16]. We verified manually that this is a mistake, and we exclude those papers from this mapping. We sort the resulting dataset by this year. It contains a total number of 1035 molecules.

Data splitting

Statistical models, especially predictive machine learning (ML) models, require separate parts of data for fitting (training) models, and for their testing and verification of performance. For molecular graphs and chemical data especially, there are many non-obvious possible sources of data leakage, which lead to improper and overly optimistic estimation of models' performance. Having predetermined splits is beneficial for reproducible science in ML, therefore we compute train-test splits and distribute those files along the full dataset.

We split the dataset in three different ways: stratified random split, time split, and MaxMin split. Random and MaxMin are interpolative, validating predictive performance inside the domain of the dataset. Time split is extrapolative, meaning that it aims to verify the out-of-domain generalization of models to structurally novel compounds. In all cases, we use 80%-20% proportions, resulting in 828 training and 207 testing molecules.

We do not provide scaffold split[17], popular in medicinal chemistry, for the following reasons. It uses the Bemis-Murcko scaffolds to group molecules, and then puts the smallest groups in the test set. The idea is to put the most structurally different compounds into the test set, which requires generalization to new ares of chemical space. However, this assumes that those scaffolds can be calculated at all, and that the dataset will contain many small groups of scaffolds. Bemis-Murcko scaffolds are not defined for ring-free compounds, as well as those with disconnected components (e.g. salts), which constitute almost 20% of the data (see Technical Validation section for more details). This results in one large "no scaffold" group in the training set, therefore this approach does not differentiate them structurally. In the worst case, almost identical molecules can be both in training and test set, introducing data leakage. This is clearly a problem, and for this reason we recommend using MaxMin or time split instead.

Stratified random split puts data points randomly into training and testing parts, disregarding the features or internal structure of molecular graphs. Stratification ensures that the proportion of toxic and non-toxic classes is approximately the same in the full dataset and both splits. This is desirable, since ApisTox is imbalanced and toxic molecules constitute a minority class. We use binary toxicity labels here. This split is susceptible to clustering in chemical space, and therefore can result in structurally similar pesticides in both training and testing sets[17,18]. Depending on application, this may be seen as a form of data leakage and artificially increase test score.

Time split puts the newest molecules in the testing set, in order to simulate the actual discovery and adoption of pesticides. The underlying assumption is that intrinsically new substances are introduced over time. This is often a very realistic setting, especially for designing novel molecules, but this information is rarely available in molecular datasets[18,19]. Our literature-based year assignment, while not perfectly precise, allows using this kind of split.

MaxMin split utilizes the maximum diversity picking algorithm to select test molecules such that the sum of distances in the test set is maximized[20,21]. Typically, ECFP4 fingerprints with either Tanimoto or Dice distance are used for this purpose.

This way, the test set has very high coverage of the chemical space and tests the generalization of the algorithm to all kinds of compounds in the data. Due to maximization of test set distances, it selects molecules much more uniformly in the chemical space than random split, alleviating the problem of clustering.

We use Scikit-learn[22] for computing stratified random split, Pandas[23] for time split, and DeepChem[24] for MaxMin split (with Dice distance and 1024 bits ECFP4 fingerprints).

Data Records

ApisTox dataset is available on GitHub[25], as well as on Zenodo[26]. We distribute the files on all stages of processing: raw, cleaned, the final dataset, and data splits as described in Methods section. All files are in human-readable CSV format. Main dataset file is *dataset_final.csv*, in *outputs* directory, with structure described in Table 1.

Raw files are in *raw_data* directory. *ecotox.csv* contains raw outputs of ECOTOX database query, and uses pipe "|" as separator. All other files use commas as separators. *bpdb.csv* and *ppdb.csv* consist of data from BPDB and PPDB, respectively.

Processed data files are in *outputs* directory. *ecotox_cleaned_data.csv* is ECOTOX data after processing and cleaning, and *excluded_data.csv* contains rows removed at any point of processing, along with reason for exclusion. *dataset_final.csv* is the main dataset file.

Column	Type	Description
name	string	Chemical name
CID	integer	PubChem Compound ID number
CAS	string	Chemical Abstracts Service registry number
SMILES	string	Molecule structure in SMILES format
source	string	Compound source: ECOTOX, PPDB or BPDB
year	integer	First publication year in literature according to PubChem
toxicity_type	string	Strongest toxicity type: Contact, Oral or Other
herbicide	boolean	Is the chemical used as a herbicide?
fungicide	boolean	Is the chemical used as a fungicide?
insecticide	boolean	Is the chemical used as an insecticide?
other_agrochemical	boolean	Is the chemical used in other way as an agrochemical?
label	boolean	Binary toxicity label
ppdb_level	integer	Ternary toxicity level

Table 1. Features in the final ApisTox dataset.

For each split, described in Methods section, we provide training and testing subset of data after splitting in *outputs/splits* directory. All those files have exactly the same structure as the main dataset file.

Technical Validation

In this section, we present molecular characteristics of ApisTox dataset, and technical analyses with basic properties relevant to chemoinformatical applications. We use the data from *dataset_final.csv* file. Suggestions for further modelling and applications are in the Usage Notes section.

Dataset quality verification

Here, we perform the basic quality checks for molecular data. ApisTox, as a curated and unified collection of data, should cover all three source databases. We compare the number of molecules of the final dataset and source databases, all after the same data cleaning procedure outlined in the Methods section, in Table 2. ApisTox is indeed larger than all input datasets, validating our data combination process. In particular, it is also almost 25% larger than PPDB, the largest of the source datasets.

	ApisTox	PPDB	BPBD	ECOTOX
Total molecules	1035	831	115	521
Toxic molecules	296	228	18	189
Non-toxic molecules	739	603	97	332

Table 2. Comparison of the number of molecules between ApisTox and the source databases.

For more detailed analysis, we verify that all canonical SMILES from those cleaned source datasets are included in the final dataset. Results are summarized by a Venn diagram in Figure 2. We gain molecules from each data source, showing that creation of unified datasets using many input databases is beneficial for ecotoxicology data. Almost molecules are included in the final dataset, except for 11 SMILES strings from ECOTOX. However, this is not a mistake - manual check showed that those molecules are subtle duplicates, having the same CAS numbers as molecules from PPDB already included. This is the consequence of non-uniqueness of SMILES format, i.e. a given molecule can be written in many ways, and RDKit canonicalization is not always able to detect such cases. This shows that the two-step deduplication, described in the Methods section, which includes both canonical SMILES and CAS deduplication, is indeed required for proper merging of molecular datasets.

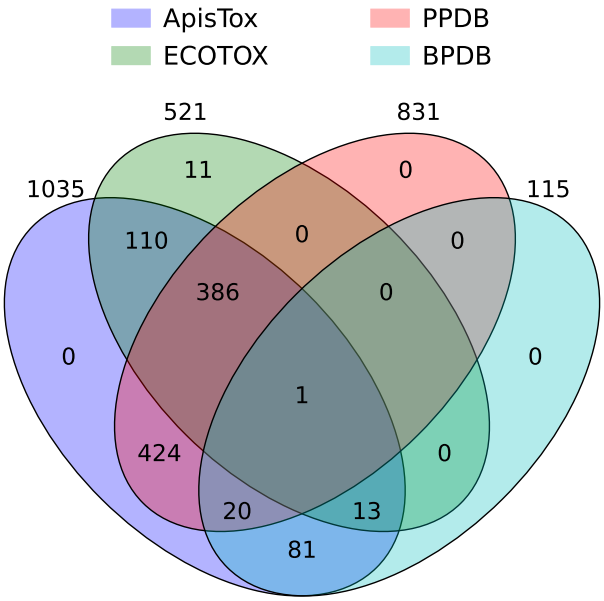

Figure 2. Number of common molecules between source datasets and ApisTox.

We also verify that ApisTox contains only valid entries that can be processed by RDKit, and contains no duplicated molecules (in terms of canonical SMILES and CAS numbers). To further validate our preprocessing workflow in this regard, we apply it to datasets previously proposed for honey bee toxicity: CropCSM[3], BeeTOX[1] and BeeToxAI[7]. As shown in Table 3, other datasets contain invalid SMILES and duplicates, even as much as 36% for BeeToxAI. Furthermore, ApisTox is considerably larger than all of them, containing many more toxic molecules in particular, which are crucial for understanding the underlying causes of pesticide toxicity for honey bees.

	ApisTox	**CropCSM**	**BeeTOX**	**BeeToxAI**
Initial number of molecules	1035	900	891	734
Invalid entries	0	1	3	0
Duplicated molecules	0	28	12	262
Cleaned dataset size	1035	871	876	472
Non-toxic molecules	739	638	645	282
Toxic molecules	296	233	231	121

Table 3. Comparison of the ApisTox and previous datasets on pesticide toxicity for honey bees. In case of BeetoxAI, only 403 molecules were reported with toxicity labels.

Molecular properties

We present distributions of six basic physico-chemical molecular properties in Figure 3: molecular weight (MW), logarithm of octanol-water partition coefficient (logP), topological polar surface area (TPSA), number of hydrogen bond acceptors and

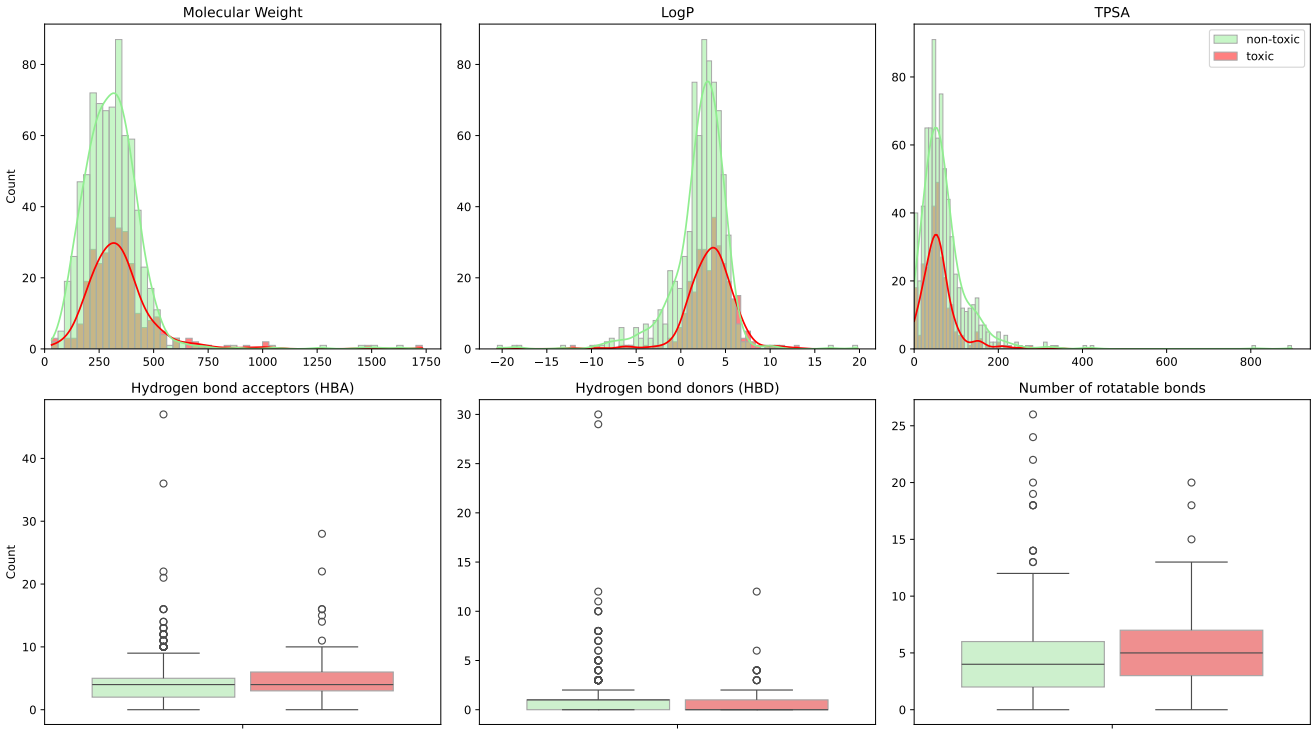

Figure 3. Comparison of basic physico-chemical properties of molecules in ApisTox for non-toxic and toxic pesticides.

donors (HBA, HBD) and number of rotatable bonds.

Distributions for non-toxic and toxic pesticides are similarly shaped. Positive logP and low TPSA (<100) are dominant, suggesting that most molecules are non-polar. This makes sense for pesticides, since non-polar molecules penetrate biological membranes much more effectively than polar ones[27]. Toxic molecules have slightly higher HBA and number of rotatable bonds, across all quartiles. ApisTox contains many large molecules, as measures by molecular weight, with some over 1000 daltons. This also explains outliers in terms of HBA and HBD.

Toxicity labels distributions

We present distributions of toxicity labels in Figure 4. In terms of binary toxicity label, ApisTox is moderately imbalanced, with pesticides toxic for honey bees constituting 29% of the data. Concretely, there are 739 non-toxic and 296 toxic compounds. The ternary level, following PPDB methodology, is more severely imbalanced, with 17% non-toxic, 66% moderately toxic, and 17% highly toxic molecules. This is due to the very wide definition of moderate toxicity in PPDB. Under this methodology, we have 177 non-toxic, 687 moderately toxic and 171 highly toxic molecules.

Such imbalance in class distributions influences metrics appropriate for validating predictive models, as discussed in the Usage Notes section.

Splits analysis

We analyze the effects of different proposed splits into training and testing sets. Preferably, both parts should have similar distribution of classes and pesticide types, to keep the basic characteristics of the dataset similar and have good representation of all segments of the data in the test set.

In addition, we want the test set to be diverse, and to be reasonably structurally different from the training data in order to avoid data leakage. For measuring those qualities, we utilize commonly used ECFP4 (Morgan) fingerprints with 1024 bits and Tanimoto distance (one minus Tanimoto similarity), which allow us to represent the molecules in a vector space. We measure diversity by calculating the average distance between test molecules, as high values mean that we avoid "clustering" the test samples, which would measure generalization only in small subsets of a chemical space. To calculate structural separation, we compute average distance between each test sample and closest training sample. A high value of such metric means that test molecules are structurally different from training ones, and we avoid data leakage from too similar compounds.

We summarize the class distributions and distances in Table 4. An additional table with pesticide types distributions is available in the supplementary information.

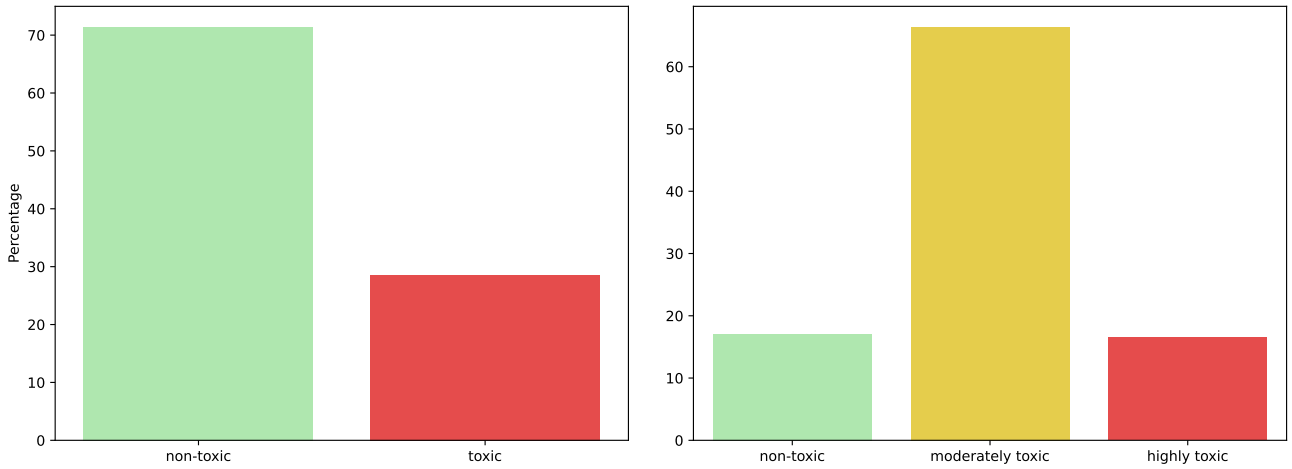

Figure 4. Toxicity binary labels and ternary levels distributions.

Class distributions for both binary labels and ternary PPDB levels are similar in all cases, which indicates that no splitting method introduces unwanted additional imbalance. The diversity is highest for MaxMin split and the lowest for stratified random split, which follows their motivation outlined in the Methods section. Train-test separation is also the highest for MaxMin split, meaning that not only its test set covers the chemical space of the dataset well, but also at the same time is not too similar to training compounds.

	Binary label		**Ternary level**		**Test set diversity**	**Train-test separation**
Split type	Train	Test	Train	Test	-	-
Stratified random	71% / 29%	71% / 29%	65% / 18% / 17%	70% / 16% / 14%	0.471	0.900
Time	69% / 31%	80% / 20%	66% / 18% / 16%	69% / 19% / 12%	0.515	0.882
MaxMin	69% / 31%	80% / 20%	66% / 18% / 16%	67% / 22% / 11%	0.611	0.944

Table 4. Dataset splits statistics.

Pesticides timeline

Using literature publication dates from PubChem, we present a timeline plot with the total number of available pesticides per year in Figure 5. The results align with agrochemical literature, with the oldest pesticides like benzoic acid or calcium carbonate known and used in the 19th century, and the majority of older generation pesticides (often toxic and outdated by contemporary standards) developed in the 1970s (carbamates), 1980s (pyrethroids) and 1990s (neonicotinoids)[28,29]. The decrease in new developments in the 21st century is also supported by both literature and industry trends, e.g. no herbicides with new mode of action have been introduced commercially in the last 30 years[30].

Molecular filter rules

A common approach to drug design in medicinal chemistry is the application of molecular filters[31]. They consist of conditions (rules) that have to be satisfied by new drug candidates, ensuring that they are bioavailable, have high absorption and permeation, or have other desirable properties. Filter rules are typically based on statistics of molecule properties derived from large collections of compounds of a particular type, e.g. drug-like molecules. While the filter-based approach is conservative and may limit the diversity of novel compounds for drug design, they can also be used to verify the quality of the data. For high-quality datasets, a reasonable percentage of compounds should meet the requirements of typical filters.

In the context of pesticides, multiple filters have been designed and used. The most widely used Lipinski's rule of five has been designed for bioavailable, drug-like molecules with high absorption and permeation[32]. It has been shown that it also works for pesticides and their subtypes (e.g. insecticides)[33], which also have to possess similar bioavailability properties. Specialized filters for agrochemistry have been designed, most prominently Hao's pesticide filter[34], and Tice's filters for herbicides and insecticides[33]. Since filters are quite conservative and can often reject specific groups of molecules (e.g. macrolides for Lipinski's rule), a common variant allows violating one of the conditions.

We present results of Lipinski, Hao and two Tice filters in Table 5, i.e. what percentage of the data fulfill the given filter conditions. We analyze both the whole ApisTox dataset and subsets with particular pesticide types. Results are provided for all

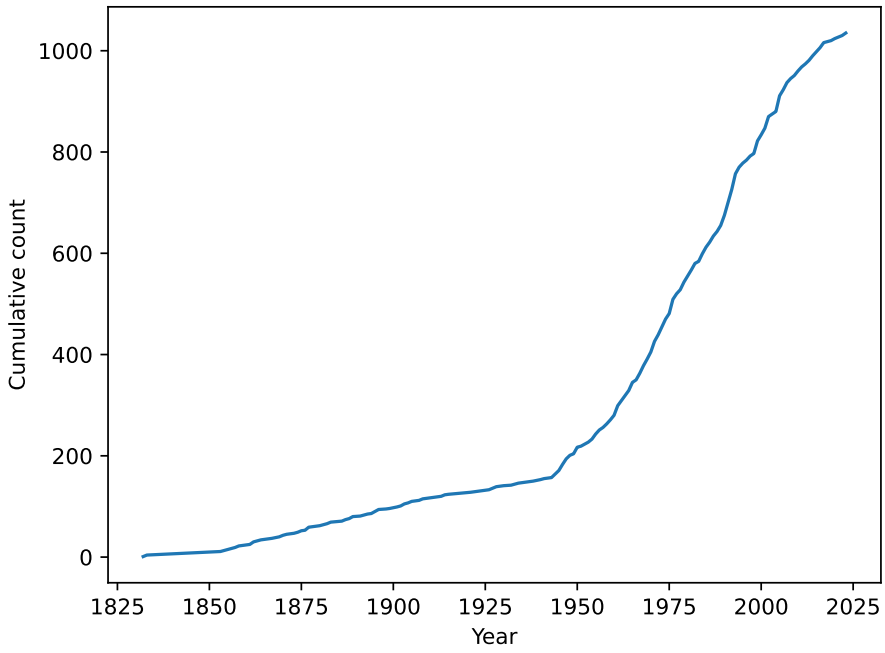

Figure 5. Cumulative count of pesticides by year.

rules satisfied, and for the more relaxed variant with one violation allowed.

The vast majority of both the full dataset and pesticide type subsets fulfill those filters when one violation is allowed. This means that our data follows established rules for bioavailable drugs, pesticides, and their types. Even when no violation is allowed, the majority of ApisTox molecules satisfy the filters' conditions. At the same time, lower percentages for Hao and Tice filters with all rules indicate that data is varied and represents a rich collection of pesticides, not only following the most common trends.

	Lipinski[32]		**Hao**[34]		**Tice**[33]	
Dataset	All rules	1 violation	All rules	1 violation	All rules	1 violation
ApisTox	82.6%	95.1%	70.9%	87.7%	-	-
Herbicides	91.8%	99.4%	68.6%	90.1%	61.3%	97.2%
Fungicides	88.3%	94.9%	79.7%	90.9%	-	-
Insecticides	67.0%	91.6%	66.1%	85.0%	60.8%	84.1%

Table 5. Percentage of molecules satisfying given chemical rules

Molecule structures analysis

Here we present analyses concerning internal structure of the molecules, i.e. their Bemis-Murcko scaffolds, functional groups and frequent subgraphs. High quality molecular datasets should, in general, be structurally valid and not be dominated by a few substructures. At the same time, we expect the existence of discriminative structures common only among pesticides toxic or non-toxic for honey bees, which would enable data analysis and interpretability of predictive algorithms.

Firstly, we validate the distribution of molecular scaffolds, which motivated our omission of scaffold split in the Methods section. Among 1035 molecules, there are 424 Bemis-Murcko scaffolds, and 324 occur in only one molecule, indicating a very diverse dataset. This is also quite high number compared to typical medicinal chemistry datasets. Within the group of toxic molecules, scaffolds composed of 6-membered aromatic rings containing carbons and nitrogens (e.g."c1ccccc1", "c1cncncnc1", "c1ccncc1", "c1ncncn1") are common, constituting 22% (67 out of 296) of these molecules.

However, Bemis-Murcko scaffolds are defined only for connected compounds with ring systems, and 186 molecules (almost 20% of the ApisTox dataset) have no scaffold at all (141 have multiple fragments, 45 do not have any rings). In particular, it

means that the ones with disconnected components could share an entire subgraph with those with scaffolds available, and the scaffold split method would not be able to detect that.

Next, we check the distribution of the functional groups, also called fragments in RDKit. In Figure 6, we visualize ten most discriminative functional groups, i.e. those with the largest frequency difference between pesticides toxic and non-toxic to honey bees.

We see that there is a high proportion of insecticide-specific functional groups in toxic molecules. Phosphorus is very commonly found in insteciticides, most often in a form of organophosphates[35,36] (*phos_ester* and *phos_acid* fragments). Rich presence of *sulfide* fragment is explained by common usage of sulphur in pesticides, in particular fungicides[37,38]. Pyrethrins and pyrethroids constitute a large group of insecticides[39], to which bees are very sensitive[40], containing an ester group and also very often a cyano group (*ester* and *nitrile* fragments). Neonicotinoids are yet another large group of insecticides, very rich in nitrogen, within which a hydrazine group is often detected, as well as a guanidinium group[41,42] (*hdrzine* and *guanido* fragments). All of those results show that ApisTox data aligns with literature on pesticide toxicity of honey bees.

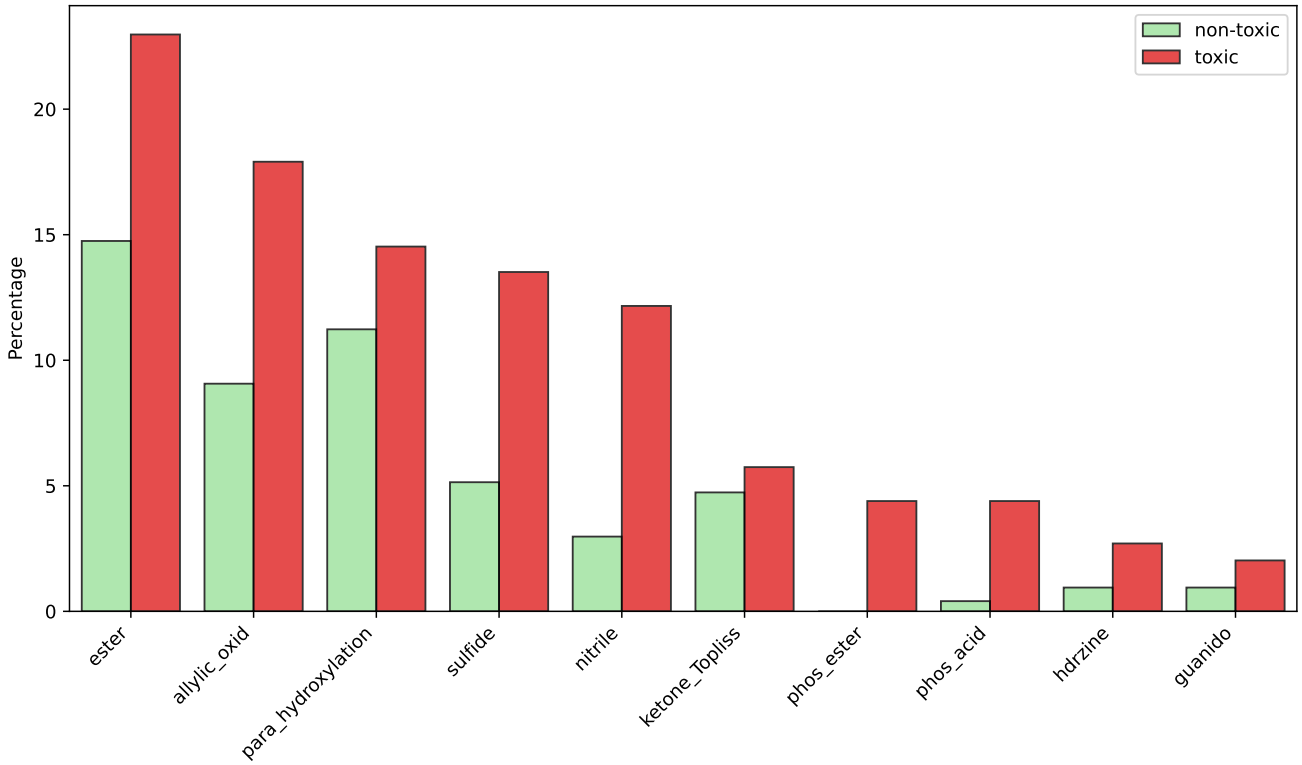

Figure 6. Functional groups with the largest frequency difference between non-toxic and toxic molecules. They are sorted by total frequency in the dataset. Groups are named following their corresponding function names in RDKit.

Lastly, we apply molecular frequent subgraph mining, in order to identify discriminative substructures, i.e. both frequent overall and at the same time much more frequent among one toxic molecules than non-toxic ones (or vice versa). This is a more data-driven solution compared to functional groups, since we derive subgraphs from the data itself. It can also detect smaller subgraphs compared to typical functional groups. Again, we expect the data to reflect the usage of common chemical elements and their influence on pesticide toxicity to honey bees.

We use the MoSS tool[43] with default settings, to mine the most common subgraphs. We then identify their frequency in toxic and non-toxic classes, and select ten with the largest frequency difference between classes. Results are presented in Figure 7. The results again align with literature, with phosphate-containing subgraphs much more common among toxic compounds, and sulfur often found in both types of compounds.

Usage Notes

The important application of ApisTox dataset is understanding and predicting agrochemical compounds' toxicity for honey bees, using data mining and machine learning (ML) methods. Those tasks can be broken into unsupervised and supervised

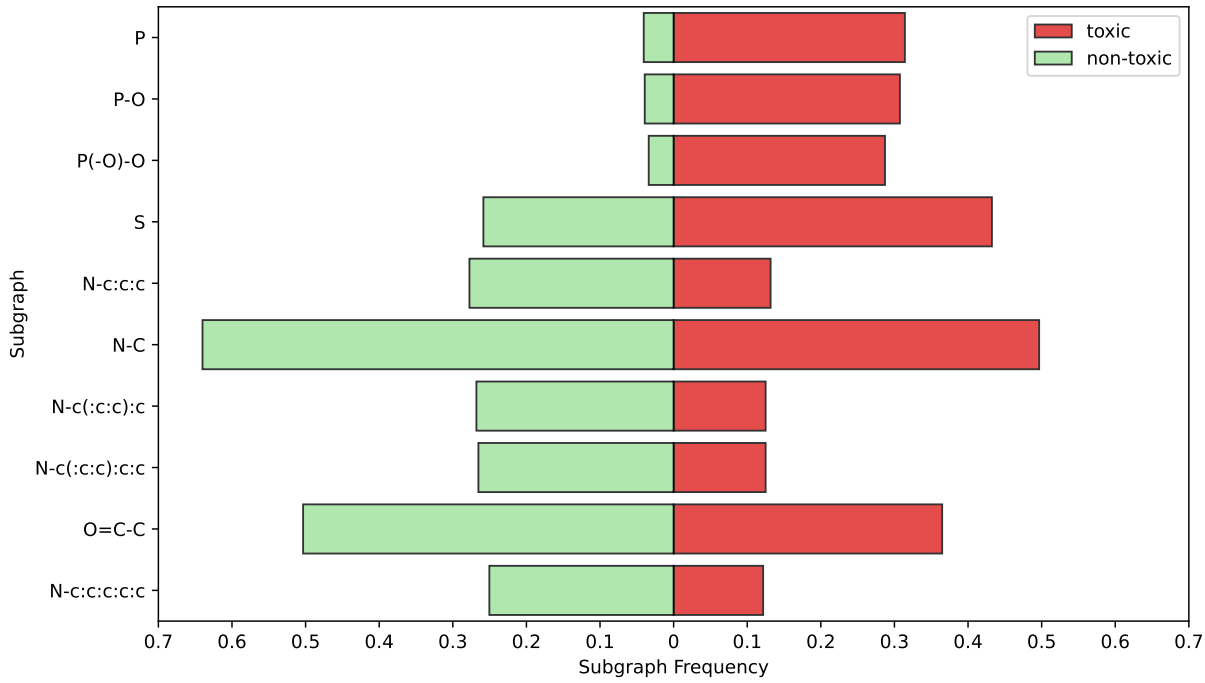

Figure 7. Frequency of occurrence of subgraphs identified by MoSS within toxic and non-toxic molecules. We plot groups with the highest frequency difference between classes.

applications, depending on whether they explicitly model the dependence between molecule features and selected targets or not.

Analytical tasks include detecting functional groups, scaffolds, binding sites, and other molecular fragments that influence the toxicity of chemicals for honey bees. Those analyses can be carried out either on the entirety of the data, or on its subsets, e.g. for herbicides, fungicides, or insecticides only. Understanding those factors can lead to design of safer pesticides in the future. Techniques like graph clustering, dimensionality reduction, and frequent subgraph mining can be utilized[44,45]. We stress that those methods, even when used as unsupervised learning techniques, often have hyperparameters and settings to tune, and they should also be validated on external data not used during the initial analysis. Therefore, provided train-test splits still have to be used, even if explicit labels are not utilized. In this regard, time split would be especially useful, as a direct approximation of the process of developing new agrochemicals.

Predictive models for modelling toxicity of pesticides will most likely use *label* or *ppdb_level* columns as targets. They should utilize only the training data for parameter estimation. For hyperparameter tuning, cross-validation with stratified sampling is recommended, due to class imbalance and relatively small dataset size. Test data must not be used at any point until the final validation of generalization performance. We recommend using MaxMin split for this purpose, since it covers the chemical space of the dataset more uniformly than other splits, requiring good generalization across the entire domain of the dataset[20,21].

For measuring the performance of predictive models, metrics that take into consideration class imbalance should be used, and at least two or three metrics should be reported to take into consideration different aspects of model performance. In particular, we recommend Area Under Receiver Operating Characteristic curve (AUROC), since it works well with imbalanced data and also takes into consideration probabilistic outputs of classifiers, as well as Matthews Correlation Coefficient (MCC), which can sometimes detect model failures despite high AUROC value[46,47]. Among other popular metrics, F1-score, precision, and recall can be used.

ApisTox can also be used as a benchmark dataset for molecular graph classification. Since datasets and benchmarks in this area come almost exclusively from medicinal chemistry, performance of many models, like molecular fingerprints, graph kernels, and graph neural networks (GNNs), has been evaluated exclusively on tasks directly related to pharmacological de novo drug design. Our dataset allows validation of the generalization performance of such models on new domains of agrochemistry and ecotoxicology. In this context, usage of unified and predefined split and metrics is of paramount importance, to allow

comparison of different models. We recommend usage of MaxMin split, as well as reporting both AUROC and MCC.

When using *ppdb_level* as target variable, we stress that this is not a three class classification problem, but rather an ordinal regression problem, also known as ordinal classification[48]. Toxicity levels are ordered integers, and classes 0 and 2 (non-toxic and highly toxic) are more distant than 1 and 2 (moderately toxic and highly toxic). Therefore, appropriate models should be used, e.g. ordinal logit model instead of logistic regression. In this case, additionally reporting regression metrics like MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) is recommended, with additional corrections for class imbalance[49].

Code availability

Code is available on GitHub at `https://github.com/j-adamczyk/apis_tox_dataset`. Code uses Python 3.10. To ensure full reproducibility, we pinned all external library dependencies (including transitive dependencies) using Poetry dependency manager[50]. We also include *poetry.lock* file with all dependency versions, as well as *requirements.txt* file exported from it.

The entire dataset can be recreated from scratch using *create_dataset.py* script. By default, it uses PPDB and BPDB files from *raw_data* directory, downloaded at 22nd February 2024, to ensure reproducible results.

References

1. Wang, F. *et al.* Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction. *Sci. Bull.* **65**, 1184–1191, https://doi.org/10.1016/j.scib.2020.04.006 (2020).
2. Olker, J. H. *et al.* The ECOTOXicology knowledgebase: A curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. *Environ. Toxicol. Chem.* **41**, 1520–1539, https://doi.org/10.1002/etc.5324 (2022).
3. Pires, D. E. V., Stubbs, K. A., Mylne, J. S. & Ascher, D. B. cropCSM: designing safe and potent herbicides with graph-based signatures. *Briefings Bioinforma.* **23**, bbac042, https://doi.org/10.1093/bib/bbac042 (2022).
4. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In Vanschoren, J. & Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1 (2021).
5. Kramer, L. *et al.* Curated mode-of-action data and effect concentrations for chemicals relevant for the aquatic environment. *Sci. Data* **11**, 60, https://doi.org/10.1038/s41597-023-02904-7 (2024).
6. Schür, C., Gasser, L., Perez-Cruz, F., Schirmer, K. & Baity-Jesi, M. A benchmark dataset for machine learning in ecotoxicology. *Sci. Data* **10**, 718, https://doi.org/10.1038/s41597-023-02612-2 (2023).
7. Moreira-Filho, J. T. *et al.* BeeToxAI: An artificial intelligence-based web app to assess acute toxicity of chemicals to honey bees. *Artif. Intell. Life Sci.* **1**, 100013, https://doi.org/10.1016/j.ailsci.2021.100013 (2021).
8. Lewis, K. A., Tzilivakis, J., Warner, D. J. & Green, A. An international database for pesticide risk assessments and management. *Hum. Ecol. Risk Assessment: An Int. J.* **22**, 1050–1064, https://doi.org/10.1080/10807039.2015.1133242 (2016).
9. EPA Pollinator Risk Assessment Guidance. https://www.epa.gov/pollinator-protection/pollinator-risk-assessment-guidance. Accessed: 2024-03-01.
10. The University of Hertfordshire Agricultural Substances Databases - Background and Support Information. https://sitem.herts.ac.uk/aeru/ppdb/en/docs/Background_and_Support.pdf. Accessed: 2024-03-01.
11. Kim, S. *et al.* PubChem 2023 update. *Nucleic acids research* **51**, D1373–D1380, https://doi.org/10.1093/nar/gkac956 (2023).
12. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. 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 (2018).
13. RDKit: Open-source cheminformatics. https://www.rdkit.org, 10.5281/zenodo.10633624. Accessed: 2024-03-01.
14. Swann, R., Laskowski, D., McCall, P., Vander Kuy, K. & Dishburger, H. A rapid method for the estimation of the environmental parameters octanol/water partition coefficient, soil sorption constant, water to air ratio, and water solubility. In *Residue Reviews: Residues of Pesticides and Other Contaminants in the Total Environment*, 17–28, https://doi.org/10.1007/978-1-4612-5462-1_3 (Springer, 1983).

15. Bidleman, T. F. Atmospheric processes. *Environ. science & technology* **22**, 361–367, https://doi.org/10.1021/es00169a002 (1988).

16. Franke, C. *et al.* The assessment of bioaccumulation. *Chemosphere* **29**, 1501–1514, https://doi.org/10.1016/0045-6535(94)90281-X (1994).

17. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. science* **9**, 513–530, https://doi.org/10.1039/C7SC02664A (2018).

18. Deng, J. *et al.* A systematic study of key elements underlying molecular property prediction. *Nat. Commun.* **14**, 6395, https://doi.org/10.1038/s41467-023-41948-6 (2023).

19. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. chemical information modeling* **53**, 783–790, https://doi.org/10.1021/ci400084k (2013).

20. Ashton, M. *et al.* Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct. Relationships* **21**, 598–604, https://doi.org/10.1002/qsar.200290002 (2002).

21. Kpanou, R., Dallaire, P., Rousseau, E. & Corbeil, J. Learning self-supervised molecular representations for drug–drug interaction prediction. *BMC bioinformatics* **25**, 47 (2024).

22. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830, https://doi.org/10.5555/1953048.2078195 (2011).

23. The Pandas development team. pandas-dev/pandas: Pandas, 10.5281/zenodo.10107975 (2023).

24. Ramsundar, B. *et al.* *Deep Learning for the Life Sciences* (O'Reilly Media, 2019). https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

25. ApisTox GitHub repository. https://github.com/j-adamczyk/apis_tox_dataset.

26. ApisTox Zenodo repository. https://zenodo.org/records/11062077.

27. Cooper, G. *The Cell: A Molecular Approach. 2nd edition.* (Sunderland (MA): Sinauer Associates, 2000).

28. Aktar, W., Sengupta, D. & Chowdhury, A. Impact of pesticides use in agriculture: their benefits and hazards. *Interdiscip. toxicology* **2**, 1–12, https://doi.org/10.2478%2Fv10102-009-0001-7 (2009).

29. Wood, T. J. & Goulson, D. The environmental risks of neonicotinoid pesticides: a review of the evidence post 2013. *Environ. Sci. Pollut. Res.* **24**, 17285–17325, https://doi.org/10.1007/s11356-017-9240-x (2017).

30. Umetsu, N. & Shirai, Y. Development of novel pesticides in the 21st century. *J. Pesticide Sci.* **45**, 54–74, https://doi.org/10.1584%2Fjpestics.D20-201 (2020).

31. Kralj, S., Jukič, M. & Bren, U. Molecular Filters in Medicinal Chemistry. *Encyclopedia* **3**, 501–511, https://doi.org/10.3390/encyclopedia3020035 (2023).

32. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. drug delivery reviews* **64**, 4–17, https://doi.org/10.1016/s0169-409x(00)00129-0 (2012).

33. Tice, C. M. Selecting the right compounds for screening: does Lipinski's Rule of 5 for pharmaceuticals apply to agrochemicals? *Pest Manag. Sci. formerly Pesticide Sci.* **57**, 3–16, https://doi.org/10.1002/1526-4998(200101)57:1%3C3::aid-ps269%3E3.0.co;2-6 (2001).

34. Hao, G., Dong, Q. & Yang, G. A Comparative Study on the Constitutive Properties of Marketed Pesticides. *Mol. informatics* **30**, 614–622, https://doi.org/10.1002/minf.201100020 (2011).

35. Christen, V., Joho, Y., Vogel, M. & Fent, K. Transcriptional and physiological effects of the pyrethroid deltamethrin and the organophosphate dimethoate in the brain of honey bees (Apis mellifera). *Environ. pollution* **244**, 247–256, https://doi.org/10.1016/j.envpol.2018.10.030 (2019).

36. Chaimanee, V., Evans, J. D., Chen, Y., Jackson, C. & Pettis, J. S. Sperm viability and gene expression in honey bee queens (Apis mellifera) following exposure to the neonicotinoid insecticide imidacloprid and the organophosphate acaricide coumaphos. *J. insect physiology* **89**, 1–8, https://doi.org/10.1016/j.jinsphys.2016.03.004 (2016).

37. EPA R.E.D. FACTS - Sulphur. https://www3.epa.gov/pesticides/chem_search/reg_actions/reregistration/fs_PC-077501_1-May-91.pdf. Accessed: 2024-03-01.

38. Hassan, A. Inorganic-Based Pesticides: A Review Article. *Egypt Sci J Pestic* **5**, 39–52 (2019).

39. Schleier III, J. J. & Peterson, R. K. Pyrethrins and pyrethroid insecticides. *Green trends insect control* **11**, 94–131 (2011).

40. Zhou, T. *et al.* Effects of pyrethroids on neuronal excitability of adult honeybees apis mellifera. *Pesticide biochemistry physiology* **100**, 35–40 (2011).

41. Araújo, M. F., Castanheira, E. M. S. & Sousa, S. F. The buzz on insecticides: A review of uses, molecular structures, targets, adverse effects, and alternatives. *Molecules* **28**, 10.3390/molecules28083641 (2023).

42. Gupta, R. C. & Milatovic, D. Chapter 23 - insecticides. In Gupta, R. C. (ed.) *Biomarkers in Toxicology*, 389–407, https://doi.org/10.1016/B978-0-12-404630-6.00023-3 (Academic Press, Boston, 2014).

43. Borgelt, C., Meinl, T. & Berthold, M. MoSS: a program for molecular substructure mining. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, 6–15, https://doi.org/10.1145/1133905.1133908 (Association for Computing Machinery, New York, NY, USA, 2005).

44. Traoré, H. *et al.* Clustering pesticides according to their molecular properties, fate, and effects by considering additional ecotoxicological parameters in the TyPol method. *Environ. Sci. Pollut. Res.* **25**, 4728–4738, https://doi.org/10.1007/s11356-017-0758-8 (2018).

45. Borgelt, C. & Berthold, M. R. Mining molecular fragments: finding relevant substructures of molecules. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 51–58, https://doi.org/10.1109/ICDM.2002.1183885 (IEEE, 2002).

46. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData mining* **10**, 35, https://doi.org/10.1186/s13040-017-0155-3 (2017).

47. Chicco, D. & Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **16**, 4, https://doi.org/10.1186/s13040-023-00322-4 (2023).

48. Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F. & Hervas-Martinez, C. Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowl. Data Eng.* **28**, 127–146, https://doi.org/10.1109/TKDE.2015.2457911 (2015).

49. Baccianella, S., Esuli, A. & Sebastiani, F. Evaluation Measures for Ordinal Regression. In *2009 Ninth international conference on intelligent systems design and applications*, 283–287, http://dx.doi.org/10.1109/ISDA.2009.230 (IEEE, 2009).

50. Poetry: Python packaging and dependency management made easy. https://python-poetry.org. Accessed: 2024-03-01.

Acknowledgements

Research was supported by the grant from "Excellence Initiative - Research University" (IDUB) for the AGH University of Krakow.

Author contributions statement

J.A. and J.P.: Conceptualization, Investigation, Methodology, Data Curation, Software, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing. P.S.: Supervision, Methodology, Writing - Original Draft, Writing - Review & Editing. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.