Biology Methods and Protocols, 2020, 1–10

doi: 10.1093/biomethods/bpaa025
Advance Access Publication Date: 13 November 2020
Methods Manuscript

METHODS MANUSCRIPT

Concomitant prediction of environmental fate and toxicity of chemical compounds

Juan Antonio Garcia-Martin[1], Max Chavarría[2,3], Victor de Lorenzo[4,*], and Florencio Pazos [4,*]

[1]Bioinformatics for Genomics and Proteomics, National Centre for Biotechnology (CNB-CSIC), 28049 Madrid, Spain, [2]Escuela de Química/CIPRONA Universidad de Costa Rica, 11501-2060 San José, Costa Rica, [3]Centro Nacional de Innovaciones Biotecnológicas (CENIBiot), CeNAT-CONARE, 1174-1200 San José, Costa Rica and [4]Department of Systems Biology, National Centre for Biotechnology (CNB-CSIC), 28049 Madrid, Spain

*Correspondence address. (V.d.L.) Department of Systems Biology, National Centre for Biotechnology (CNB-CSIC), 28049 Madrid, Spain. E-mail: vlorenzo@cnb.csic.es; (F.P.) Department of Systems Biology, National Centre for Biotechnology (CNB-CSIC), 28049 Madrid, Spain. E-mail: fpazos@cnb.csic.es

Abstract

The environmental fate of many functional molecules that are produced on a large scale as precursors or as additives to specialty goods (plastics, fibers, construction materials, etc.), let alone those synthesized by the pharmaceutical industry, is generally unknown. Assessing their environmental fate is crucial when taking decisions on the manufacturing, handling, usage, and release of these substances, as is the evaluation of their toxicity in humans and other higher organisms. While this data are often hard to come by, the experimental data already available on the biodegradability and toxicity of many unusual compounds (including genuinely xenobiotic molecules) make it possible to develop machine learning systems to predict these features. As such, we have created a predictor of the "risk" associated with the use and release of any chemical. This new system merges computational methods to predict biodegradability with others that assess biological toxicity. The combined platform, named *BiodegPred* (https://sysbiol.cnb.csic.es/BiodegPred/), provides an informed prognosis of the chance a given molecule can eventually be catabolized in the biosphere, as well as of its eventual toxicity, all available through a simple web interface. While the platform described does not give much information about specific degradation kinetics or particular biodegradation pathways, *BiodegPred* has been instrumental in anticipating the probable behavior of a large number of new molecules (e.g. antiviral compounds) for which no biodegradation data previously existed.

Keywords: biodegradation; prediction; web server

Introduction

Although polluting emissions of anthropogenic origin have been a constant feature of human evolution [1], the onset of the industrial revolution and the more recent development of synthetic chemistry have had a more serious environmental impact. In some cases, the contaminants generated are intrinsically persistent (e.g. heavy metals) and the only way to deal with them is through their immobilization or re-speciation to less toxic forms [2]. In other cases, such as the plethora of chemicals found in petroleum (a mixture of natural compounds), these have been mobilized by modern industry to niches and ecosystems where they do not naturally belong,

Received: 3 September 2020; **Revised:** 9 November 2020; **Accepted:** 10 November 2020

producing a harmful shift in their biological balance. This latter scenario is epitomized, yet by no means limited to, oil tanker spills [3]. Less conspicuously, many components of fossil fuels are separated at their source and used as feedstock for the production of a large number of materials that form part of our daily life (polymers, solvents, etc.). Finally, modern synthetic chemistry further modifies organic molecules to generate thousands of functional molecules for the most diverse applications (from flame retardants to medications), generating an amazing range of structures that are often altogether new-to-nature [4].

Whether already part of the biosphere or created by synthetic chemists, time and time again these compounds end up in the environment, where they come face-to-face with the extraordinary catalytic armory of the microbial community residing in the niches in which they are deposited. In the best-case scenario, these contaminants are quickly degraded to less toxic intermediates or even completely metabolized (or co-metabolized) to carbon dioxide and water through the catabolic pathways available to the resident microbial populations. This is typical of the components of petroleum that, while potentially toxic or difficult to degrade, they have been in the biosphere for long enough to allow the evolution of efficient biochemical routes that can use them as carbon sources [5]. Things get more problematic when the pollutants involve molecules with chemical bonds that are either rare or only produced synthetically (e.g. halo-aromatic and nitro-organic compounds, ionic liquids, etc.). In these cases, it is possible that the corresponding biochemical degradation routes do not exist and thus, these compounds may remain intact for long periods of time until evolutionary mechanisms produce a solution to the challenge represented by their degradation [6].

How living catalysts successfully encounter and act on their target chemicals in a physicochemical landscape can be defined by at least six constraints: toxicity, abundance, concentration, biodegradability, bioavailability, and mobility [7]. While most of these boundaries are entirely dependent on the specific scenario, biodegradability and biological toxicity can be traced to the intrinsic chemical structure of the pollutant under consideration. Nevertheless, determining these parameters experimentally is expensive and slow, especially for biodegradability where the compound must be released into a controlled environment and its decay followed over a relatively long period of time. Therefore, it should come as no surprise that various computational efforts have been made over the past decade to predict the biodegradability and fate of specific molecules when released into the environment. Accordingly, the already classic University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) (now located at the Swiss Federal Institute of Aquatic Science and Technology: http://eawag-bbd.ethz.ch/: [8]) predicts plausible pathways for the microbial degradation of organic compounds. The system is based on a set of rules representing chemical transformations frequently found in reactions described in the scientific literature. Other approaches based on machine learning include the BDPServer system, that tries to correlate the triplets of atoms found in molecular structures with their biodegradative fate [9], and the ATLAS platform [10], a repository of both known and novel predicted biochemical transformations among biological compounds that contains ~150 000 possible reactions. By the same token, a number of *Quantitative structure-activity relationship* (QSAR) methods and databases are available that correlate structure with biological toxicity [11, 12]. A second important factor to consider when deciding whether to generate, use, and release a new chemical is its eventual toxicity for humans. To this end, a number of tools have been developed for the *in silico* prediction of the toxicity of chemical compounds adopting different approaches [see Raies and Bajic [13] for a review].

To the best of our knowledge, a single tool to predict environmental fate that combines intrinsic biodegradability with toxicity, trained on experimental data of these features, has yet to be developed. In this context, we present here a merged predictive platform (BiodegPred) that exploits otherwise scattered data from a variety of biodegradation- and toxicity-related resources. Trained on the experimental data available at these resources, this tool can predict biodegradability and toxicity using only the chemical structure described in SMILES code as the input. BiodegPred not only facilitates such predictive exercises with an extremely simple web-based interface but also it enables the destiny of collections of structurally unrelated chemicals (e.g. new pharmaceuticals) to be appraised when they are released into the environment. We show that the tool has a good predictive performance in cross-validation tests. It also produces good results in a more demanding scenario involving chemicals never seen during the training/testing cycles, and can generate blind predictions for a set of antiviral compounds of medical interest.

Materials and methods

An overview of the methodology used here is shown in Fig. 1. In short, different datasets of biodegradable and recalcitrant compounds were obtained from a variety of resources that use different criteria to quantify and define biodegradability. All compounds available at these resources at the time of retrieval were taken. For each resource, a support vector machine (SVM) predictor was trained to discriminate biodegradable from recalcitrant compounds using a vector representation of these chemicals. The performance of the individual predictors was assessed using a leave-one-out cross-validation and, for all the resources, an additional set of compounds was retrieved at a later stage for validation. The final predictors have been made accessible through a web interface.

Datasets

- UM-BBD. Now EAWAG Biocatalysis/Biodegradation Database (http://eawag-bbd.ethz.ch/; [8]). For this resource, biodegradability was defined as in Gomez *et al.* [9] and Pazos *et al.* [14], whereby a compound is defined as "biodegradable" if it is possible to find a route from that compound to the Central Metabolism within the network of chemical reactions contained in the database, otherwise it is considered "non-biodegradable." Hence, this dataset does not reflect experimental criteria of biodegradability but rather, this is computationally inferred from the biochemical information available.
- University of Hertfordshire Pesticide Properties DataBase (PPDB, http://sitem.herts.ac.uk/aeru/ppdb/en/index.htm; [15]). The biodegradability criteria here are based on the DT_{50} for soil degradation in laboratory at 20°C: "non-persistent" ($DT_{50} \leq 30$) and "persistent" ($DT_{50} > 30$).
  - In addition, PPDB compounds with mammalian oral toxicity information are used to generate the toxicity predictor. Although toxicity experiments are performed in different mammals (rat, mice, rabbit, etc.), the results are usually taken as a proxy for human toxicity. Toxicity classification is based on the acute oral LD_{50} in mammals: "low-toxicity", $LD_{50} > 2000$ mg/kg and "high-toxicity" $LD_{50} \leq 2000$ mg/kg. These

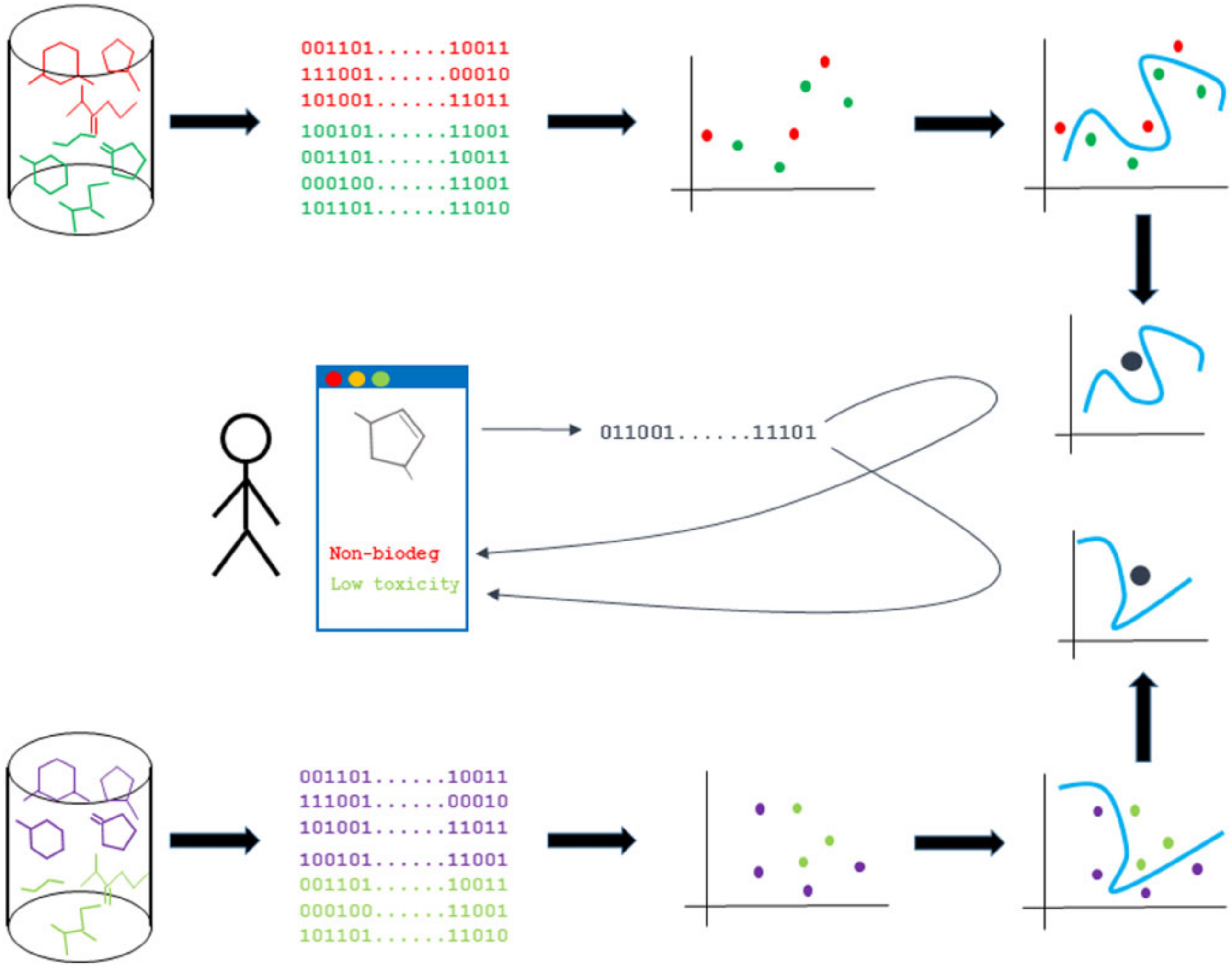


Figure 1: Scheme of the methodology. To construct the biodegradability predictor (top), a given database with experimental annotations on biodegradable (green) and recalcitrant (red) compounds is used. The chemical structures of the compounds are coded as binary fingerprint vectors, which can be represented in a multidimensional space (shown in the figure as a two-dimensional space). A SVM is used to locate the optimal hyperplane maximizing the separation between the two sets of compounds (blue curve in this 2-D representation). After training, the system can be used to classify a compound submitted by a user (middle panel): the vector representing the query compound is classified according to the hyperplane generated during training. A similar procedure is performed to generate the predictor of toxicity (bottom), starting with a database with annotations on toxic (purple) and non-toxic (light green) compounds.

compounds are further divided into the two categories of annotation confidence used in this resource: Q4 ("verified data") and Q3 ("unverified data of an unknown source") and two predictors were independently trained for these two categories.

- Ministry of International Trade and Industry (Japan) MITI-I test data (NITE, https://www.nite.go.jp/en/chem/qsar/evaluation.html; [16]). In this case, the criterion is based on the "MITI-I test," in which the substance is inoculated and incubated with 30 mg/l sludge. The biological oxygen demand (BOD) is then measured continuously over a 28-day test period. If the BOD amounts to ≥60% of the theoretical oxygen demand (ThOD), the substance is considered "ready-biodegradable," otherwise, it is regarded as "non-ready-biodegradable."

The distribution of the cases in all these datasets is shown in Table 1. All the datasets are slightly unbalanced. We decided not to balance the data as for relatively large datasets such as these SVMs have been shown to be quite robust to any imbalances [17]. Moreover, this imbalance in the databases probably reflects the true biodegradable/non-biodegradable distribution in the current chemical space and consequently, the compounds introduced by the potential users of the system would have the same distribution.

The datasets used for training/testing were retrieved from the corresponding resources on January 2012. In order to assess the performance of the method on an independent validation set not used during the training/testing cycle we retrieved from the same resources the new compounds deposited since that date up to April 2020 (Table 1), except for UM-BBD (see the "Results" section).

In addition to these datasets with compounds of a known biodegradation fate, we generated "blind" predictions for a set of 148 antiviral compounds extracted from Drugbank [18]. We selected only "antiviral agents" and "small molecules" to avoid peptides, nucleic acids, etc. The final list, which contained antiviral agents at all clinical stages (approved, investigational, and experimental), was checked to ensure that none of these compounds were in the training/testing datasets used.

Table 1: Distribution of the cases in the datasets

Dataset	Class 1	Class 2
UM-BBD (**710**/–)	Biodeg. (**567**/–)	No Biodeg. (**143**/–)
PPDB (**620**/180)	Non-persistent (**348**/76)	Persistent (**271**/104)
NITE (**1433**/2024)	Ready Biodeg. (**517**/847)	Non-ready Biodeg. (**916**/1177)
PPDB-Tox Q3 (**1219**/389)	High toxicity (**703**/200)	Low toxicity (**516**/189)
PPDB-Tox Q4 (**596**/184)	High toxicity (**407**/65)	Low toxicity (**189**/119)

The number of cases used to train/test the predictor is shown in bold, while the cases retrieved later and used for validation are shown in normal font. Note that "PPDB-Tox Q4" is contained in "PPDB-Tox Q3" as the latter represents a more relaxed reliability criterion.

Table 2: Optimal classification parameters

Parametertype	UM-BDD C-classification	PPDB nu-classification	NITE C-classification	PPDB-Tox (Q3) nu-classification	PPDB-Tox (Q4) C-classification
Gamma	0.015625	0.001953125	0.0625	0.03125	0.0078125
Cost	32	NA	2	NA	4
Nu	NA	0.5	NA	0.5	NA

C-classification doesn't require a value for nu and a cost value doesn't apply to nu-classification.

Compound vector encoding

The chemical structures of these compounds must be represented as numeric vectors given that machine learning in general, and SVMs in particular, can only handle this type of data. We used OpenBabel [19] to generate the FP2 fingerprints of the SMILES representation of the compounds obtained from the original resources. The FP2 fingerprint of a chemical compound is a binary vector of length 1024 where each component encodes the presence of a particular chemical feature in the molecule.

SVM parameters and training

SVMs are a class of machine learning methods intended for the binary classification of a set of vectors. These approaches look for the hyperplane defined by a kernel function that best separates the two classes of vectors in the training set. After training, they predict the class of an unseen case (vector) by just looking at which side of the hyperplane it falls (Fig. 1). In addition, the distance to the hyperplane can be used as a prediction's score, as the further a vector is from the hyperplane separating the two classes the clearer its classification.

The R package "e1071" serves as an interface for the "SVM" library [20], and it was used to model and train the SVM models with the chemical compound data described previously. This package allows the appropriate kernel to be selected and to fit the parameters to their optimal values. By evaluating the dataset complexity, and looking for a balance between effectiveness and time restrictions, a radial function basis kernel type was selected, defined as $k(u, v) = e^{-\gamma \cdot |u-v|^2}$ [21].

Using the "tune" function of the "e1071" package, we performed a grid search with two classification methods (C-classification and nu-classification), evaluating a range of values for gamma (2^{-10} to 2), nu (2^{-8} to 0.5), and cost (2 to 2^{20}) in the search for the optimal values of these parameters. For each set of parameter values, a 10-fold cross-validation was used to assess the SVM performance: 10% of the cases were removed from the dataset and used for testing, while the rest were used for training. Table 2 shows the optimal set of parameters for each dataset.

Performance evaluation

Once the optimal parameters were set, performance was measured by applying a leave-one-out strategy: a new model was constructed with the parameters specified using the full dataset as training set, except for one compound that was taken as the test set. The prediction performance for that compound is evaluated and the process is repeated for every other compound in the dataset. Furthermore, we evaluated the relationship between performance and SVM output score based on the percentage of correct predictions with a score equal to or higher than a given value. This allows a value of reliability to be associated with a future prediction score.

An "area under the ROC curve" (AUC) analysis [22] was used to evaluate performance. This analysis quantifies how well a classifier separates two sets of examples (positives and negatives) based on the score it associates to these. In our case, the two sets were the corresponding biodegradable and non-biodegradable compounds for each predictor, and the score is that generated by the predictor for each compound. A ROC analysis generates a plot of the true-positive rate (TPR) versus the false-positive rate (FPR) as the list of examples sorted by the score is traversed from top to bottom. A perfect classifier, whose scores situate all positives at the top of the list and all negatives at the bottom (or vice versa), would lead to an AUC value of 1.0. By contrast, a random classifier that distributes positives and negatives uniformly across the range of scores would produce a value of 0.5.

Web server

A web interface was created for interested users to utilize the predictors, mainly coded in PHP and JavaScript. It uses the JSME molecular editor [23] to convert user drawn molecular structures into SMILES and the BKChem library to generate graphical representations of molecules.

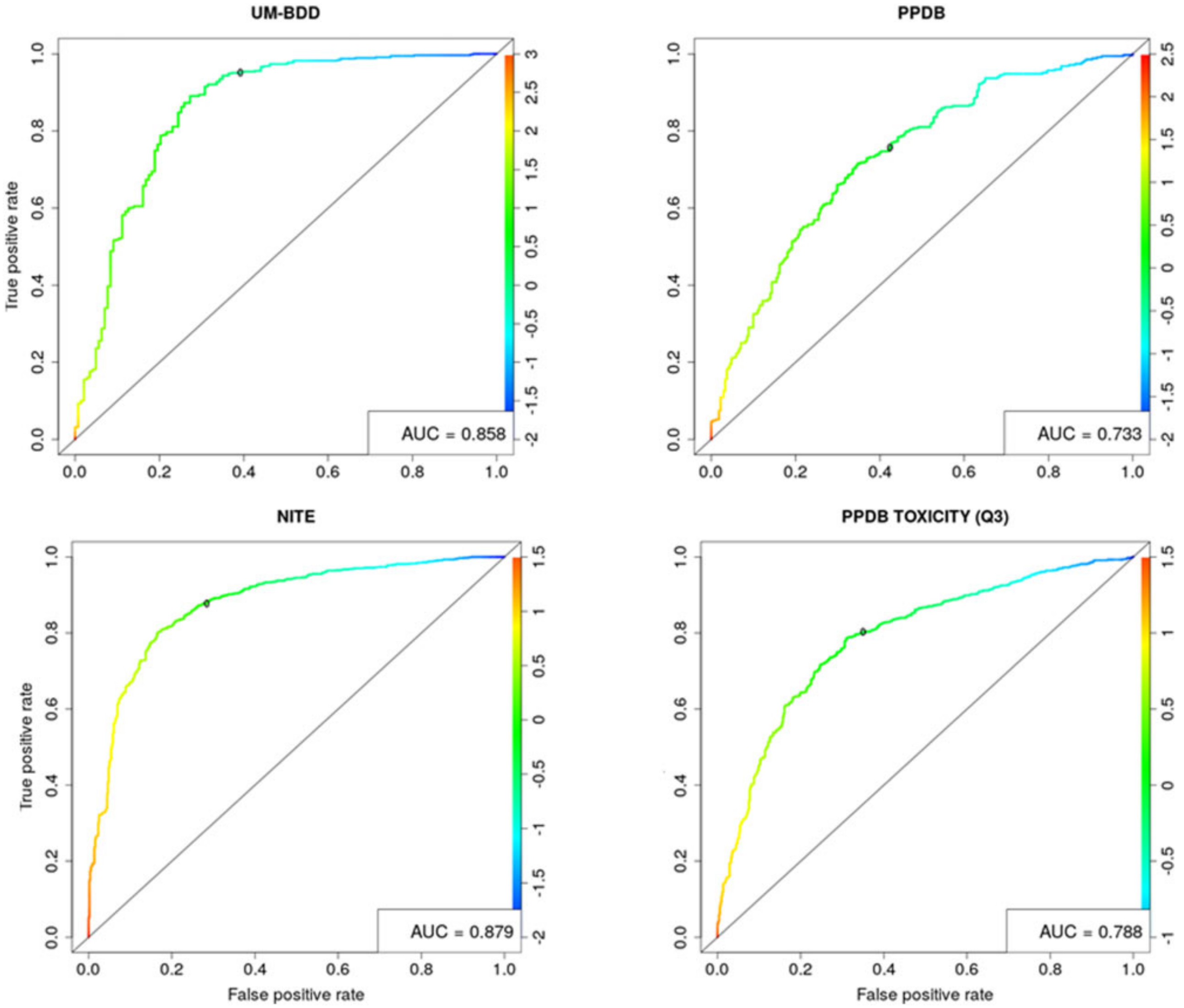


Figure 2: Performance of the predictors. ROC curves represent the performance of the four predictors on the training/testing datasets. The SVM score associated to each point of the curves is represented in a color code (see right). The "area under the curve" (AUC) quantification of performance is also shown.

Results

Predictor performance

We obtained the ROC plots for the three biodegradability predictors and the PPDB toxicity predictor trained with the Q3 compounds (see the "Materials and methods" section). As can be seen, all predictors present ROC curves clearly far from the diagonal that represents a random prediction (Fig. 2). Indeed, these curves lay relatively close to the top/left corner of the plot that would represent the perfect classifier, and the corresponding AUC values were well situated within the 0.5–1.0 range of a random to a perfect predictor: 0.86 for UM-BBD, 0.73 for PPDB, 0.88 for NITE, and 0.79 for PPDB toxicity. The ROC plots also represent the predictor's score associated to each region of the curve on a color scale. Consequently, it is possible to extract the expected TPR and FPR yielded by the predictor for any given score. Accordingly, the predictor based on the NITI data and biodegradability criteria is that which performed best, followed by UM-BBD. It should be remembered that UM-BBD is not trained on experimental biodegradability data but rather, on "assumed" biodegradability based on the information from chemical reactions.

Figure 3 shows the same results in terms of precision versus score plots, where precision is the ratio of the correct predictions out of all the predictions. Thus, for any given predictor's score (*x*-axis), it is possible to extract the fraction of correct predictions that would be obtained (*y*-axis). For example, if we take all the compounds predicted to be "ready biodegradable," with a score of -1.0 or better (lower), for NITE we will be right in 90% of the cases, whereas if we restrict the score to -1.5 or better this proportion increased to 95%. The corresponding values for "non-ready biodegradable" compounds were around 92% for a score $\geq+1.0$ and 97% for a score $\geq+1.5$. Thus, the prediction of recalcitrant compounds is slightly better in this particular case. Note that the "precision" performance metric provides information on the ratio of true positives relative to all those regarded as positives for a given score (i.e. all the predictions), yet no information is given about how many positives are lost (positives below the score threshold). However, this latter value could be extracted from other metrics, as well as from the TPR of the

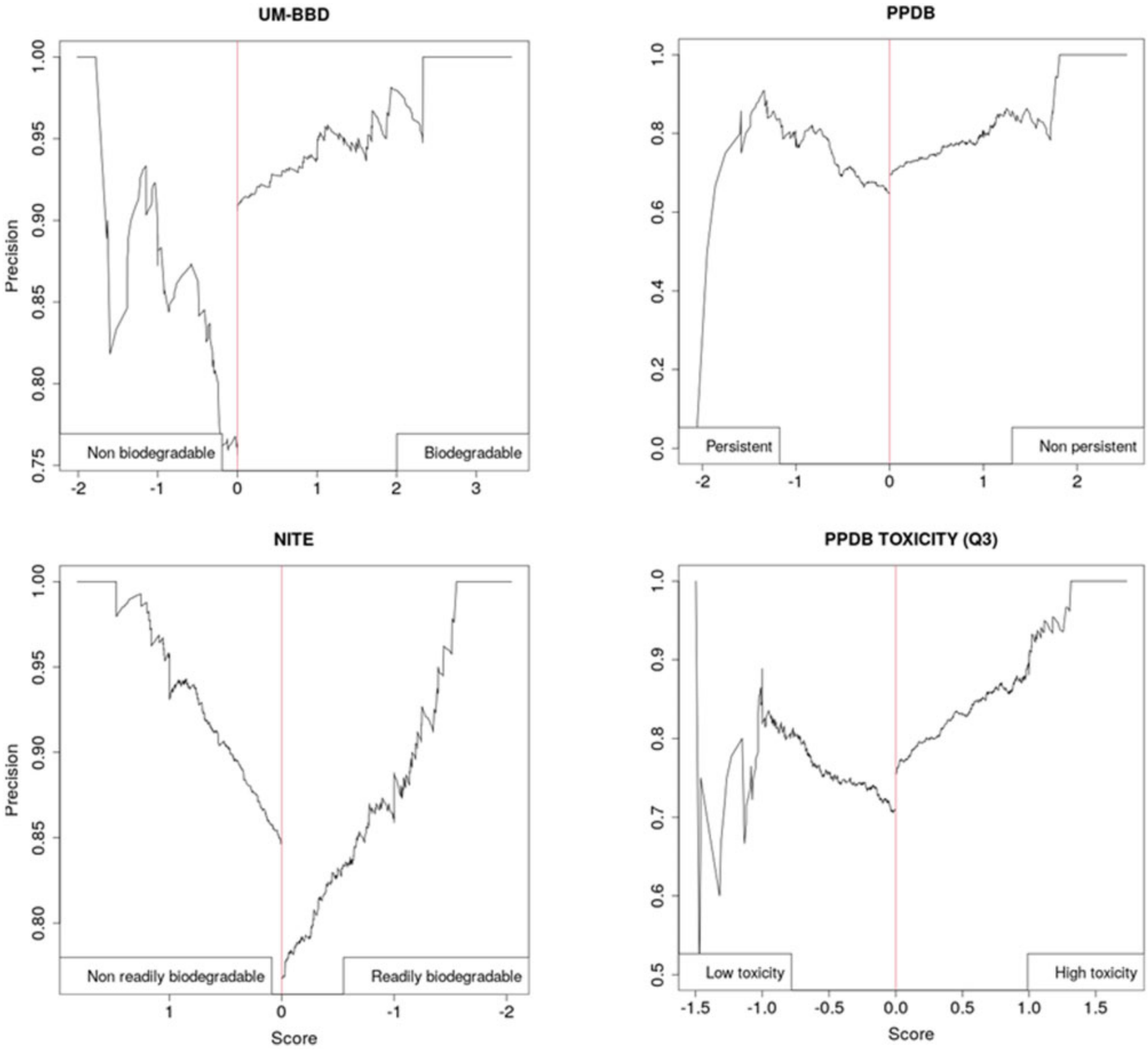


Figure 3: Relationship between precision and the SVM score. The precision (fraction of correct predictions) obtained for a given SVM score threshold is shown for the four predictors. A SVM score of 0.0 represents the boundary between the two classes (biodegradable and non-biodegradable, or high and low toxicity).

ROC plots (Fig. 2). The counterintuitive decrease in performance of PPDB in predicting recalcitrant compounds with a very high score (i.e. precision drops from 0.8 at score ≤ -1.5 to almost 0.0 for the more restrictive score ≤ -2.0) is due to the fact that there are only 10 compounds in that range and unfortunately, the compound with the lowest score is incorrectly predicted. Such variation due to a small number of instances is also evident in PPDB when predicting compounds with low toxicity.

The corresponding results for PPDB toxicity Q4 are shown in Supplementary Fig. S1. The performances of the Q3 and Q4 toxicity predictors are quite similar, indicating that restricting the training to a set of compounds with highest toxicity annotation confidence does not help for this particular dataset.

Although the compounds used for testing were never included in the corresponding training due to the cross-validation procedure (see the “Materials and methods” section), it is always good practice to include an additional set of examples not used in the training/testing cycles in order to assess how well the final predictor can be generalized to other cases. For this validation, we used the compounds deposited in the same databases after those that were retrieved in 2012 for the training/testing, i.e. the compounds deposited between 2012 and 2020 (Table 1). This was not done for UM-BBD as reproducing the network-based criteria of biodegradability for these new compounds (see the “Materials and methods” section) would require re-generating the complete biodegradation network (see [9, 14]). For PPDB, the results from this validation with new compounds were worse than those of the training/testing set, both for the prediction of biodegradability and toxicity (Fig. 4), although they were still highly significant. By contrast, these new compounds were even better predicted than those used to train/test the predictor for NITE.

“Blind” predictions for antivirals

The proportion of the 148 antiviral agents assigned to each class by the different predictors is shown in Fig. 5. According to the

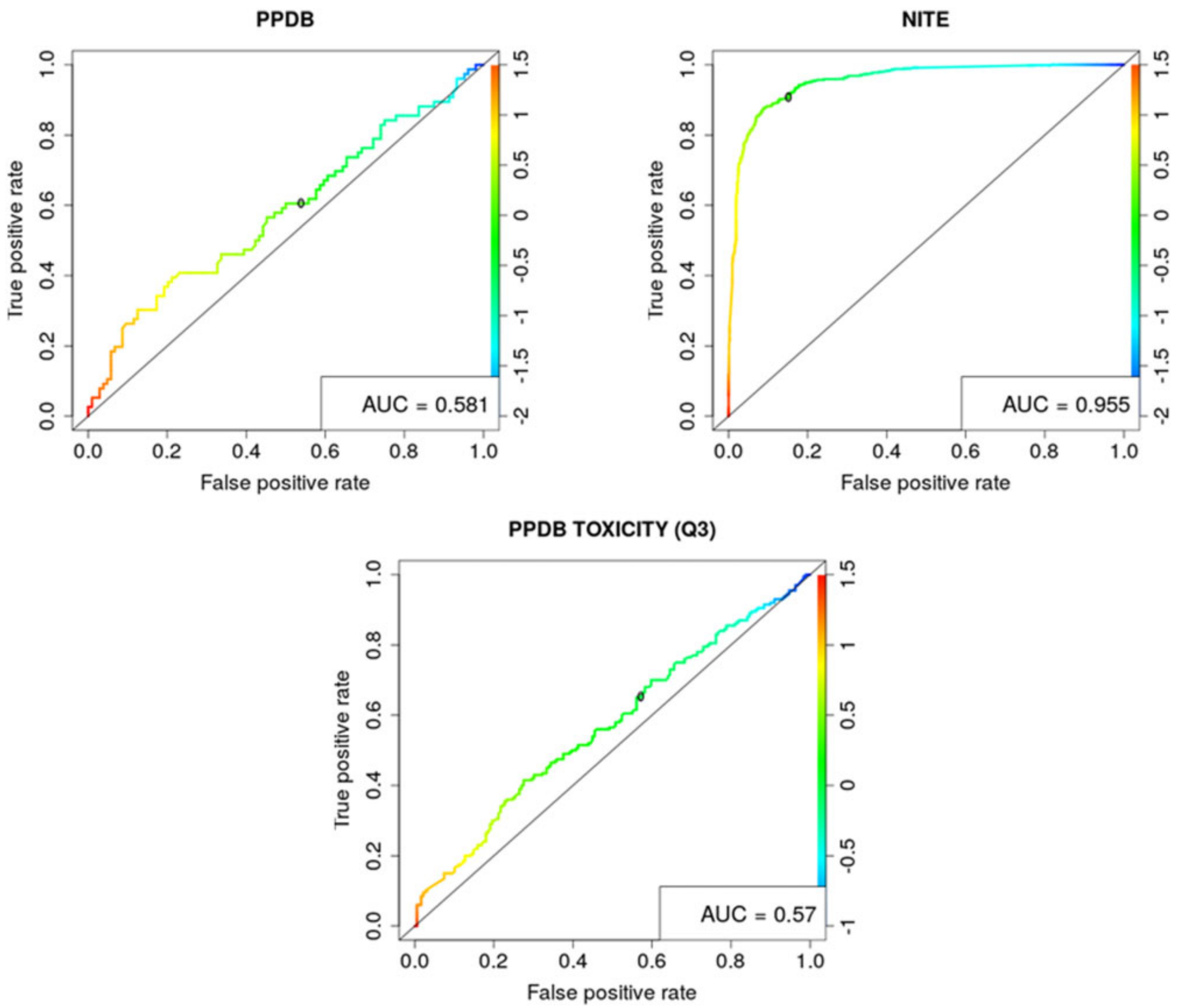


Figure 4: Performance of the predictors in the validation dataset. As in Fig. 2, yet for the compounds in the validation datasets that were not used in the training/testing cycles.

UM-BBD and PPDB criteria, about half of the compounds were predicted to be biodegradable while the other half were predicted to persist in the environment. Interestingly, according to the NITE criteria, only 3 of the 148 antiviral agents would belong to the "ready biodegradable" category. Regarding the predicted toxicity, about three-fourth of the compounds were predicted to be non-toxic in mammals according with the NITE Toxicity Q3 predictor. These antiviral agents are listed in Supplementary Table S1, including the antiviral's name, SMILES string, and molecular structure. The first 19 (from Remdesivir to Vazegepan) of these antiviral agents are currently being investigated or used for SARS-2 treatment.

Web server

A web interface was created for interested users to access the predictors, the main input for the interface being the SMILES codification of the molecule of interest. If this SMILES string is not known, it is possible to draw the molecule in a built-in JSME molecular editor [23] that will generate the corresponding SMILES code. After introducing the input molecule, the user selects one or more of the predictors to be applied to it (Fig. 6). The output of the server is the classification of the molecule according to the predictors selected, as well as the associated score and reliability values. The predicted classes are highlighted in color: red, recalcitrant; green, biodegradable; purple, toxic; and light green, non-toxic. An image of the input molecule generated with the BKChem library (http://bkchem.zirael.org/) is also shown. The web server is freely available at https://sysbiol.cnb.csic.es/BiodegPred/.

Discussion

Human activities lead to the release into the environment of a large number of chemicals not previously encountered by nature. Two very important factors that should be considered when designing new chemicals are their safety, insomuch as their toxicity to humans, and their environmental fate, i.e. whether they will be degraded by biotic and abiotic components in the environment or remain recalcitrant for a considerable time. The experimental determination of these two features is expensive and time consuming, especially in the case of

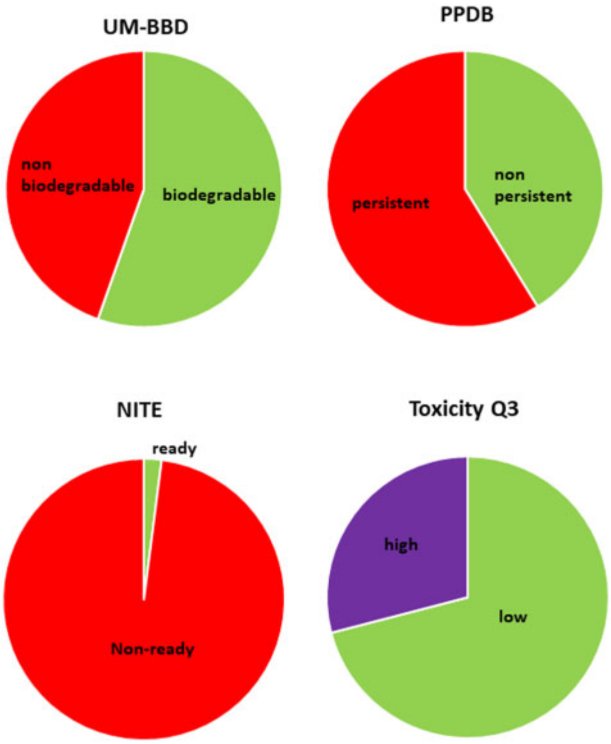
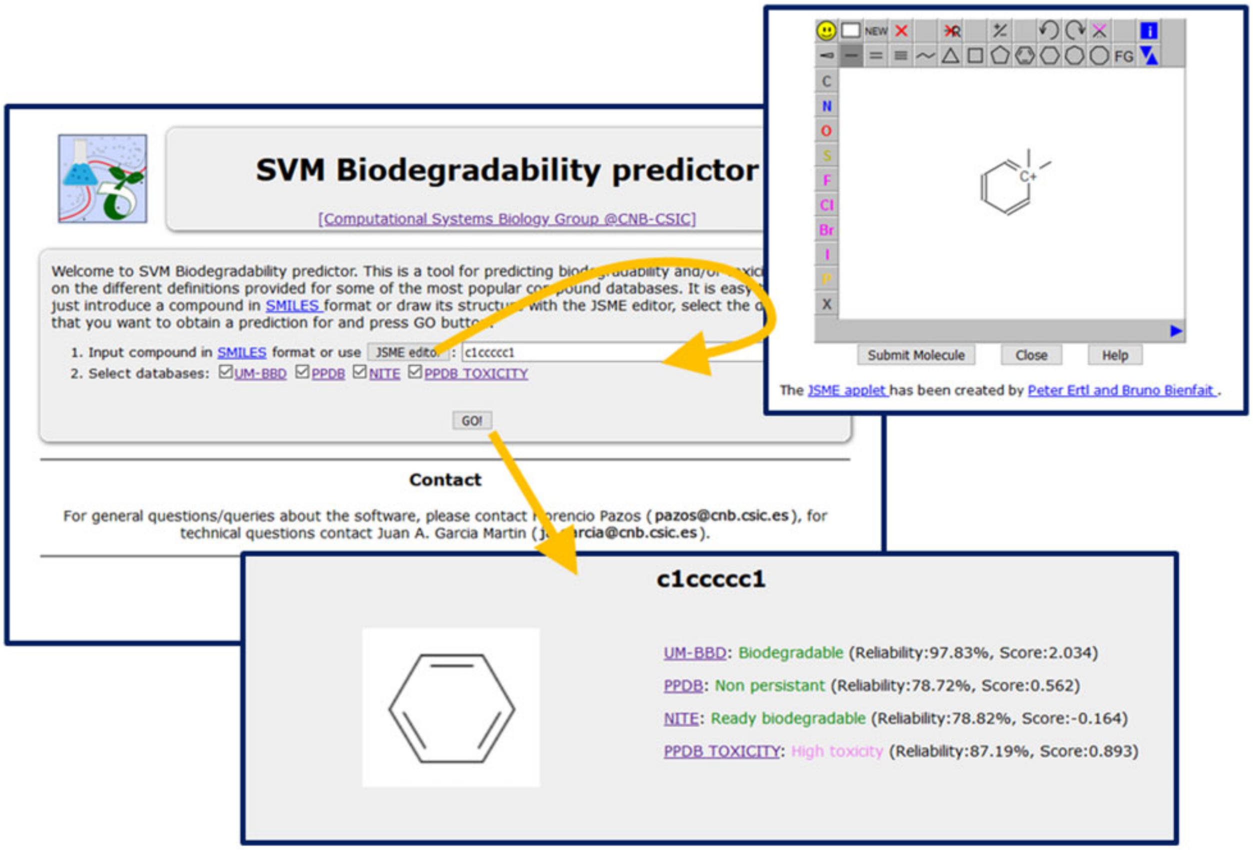


Figure 5: Blind predictions for the antiviral datasets. Fractions of the Drugbank antiviral agents predicted in each class by the four predictors.

biodegradability, as it involves tracking different concentrations of the compound in a controlled environment over relatively long periods of time. Consequently, these experiments cannot cope with the ever increasing number of new compounds that are being generated. Hence, *in silico* systems that can predict these characteristics from the chemical structure of the compounds alone are particularly useful. Even if not perfectly accurate, these systems can give a good initial idea of the potential toxicity and biodegradability of the compounds, thereby guiding future experimental work.

Accordingly, we have generated a multi-predictor of compound biodegradability that concomitantly predicts mammalian toxicity. As there is currently no accepted classification of biodegradable/non-biodegradable compounds (i.e. different agencies use distinct criteria, mostly based on quantitative measures), we developed different predictors trained on some of these diverse criteria. This lack of a definite criterion for classifying a compound as biodegradable/recalcitrant justifies the development of a tool which offers to the user an "overview" of the predictions based on different criteria, so that he/she can interpret the consensus/differences.

These predictors were trained and tested on large datasets involving hundreds of compounds and they showed good classification performance. In addition, they also performed well when validated in datasets built with the newer compounds deposited in the same databases, structures that had not been used during the training/testing process. For the NITE classifier, the performance in this validation was even superior to that observed during training/testing. However, this performance in



The JSME applet has been created by Peter Ertl and Bruno Bienfait .

SVM Biodegradability predictor

[Computational Systems Biology Group @CNB-CSIC]

Welcome to SVM Biodegradability predictor. This is a tool for predicting biodegradability and/or toxic on the different definitions provided for some of the most popular compound databases. It is easy just introduce a compound in SMILES format or draw its structure with the JSME editor, select the d that you want to obtain a prediction for and press GO button.

1. Input compound in SMILES format or use JSME editor: c1ccccc1
2. Select databases: ☑UM-BBD ☑PPDB ☑NITE ☑PPDB TOXICITY

Contact

For general questions/queries about the software, please contact Florencio Pazos (**pazos@cnb.csic.es**), for technical questions contact Juan A. Garcia Martin (**ja.garcia@cnb.csic.es**).

c1ccccc1

UM-BBD: Biodegradable (Reliability:97.83%, Score:2.034)

PPDB: Non persistant (Reliability:78.72%, Score:0.562)

NITE: Ready biodegradable (Reliability:78.82%, Score:-0.164)

PPDB TOXICITY: High toxicity (Reliability:87.19%, Score:0.893)

Figure 6: Screenshots of the web interface. In the main interface (left), the user can either enter the SMILES coding of the compound or, alternatively, the JME molecular editor (right) can be used to interactively draw its molecular structure, that is then converted into smiles. The results window (bottom) shows the classification of the compound according to the selected predictors.

validation was worse in the case of both biodegradability and toxicity for PPDB, which could be due to the newer compounds deposited in this period having certain peculiarities or a bias toward certain types of chemicals.

After demonstrating that the predictors render good performance with compounds of known biodegradation fates and toxicity, we generated "blind predictions" by using them to predict the behavior of compounds of interest for which such information is not readily available (at least it is not standardized in databases). Due to the current COVID19 pandemic, antiviral agents are a particularly interesting set of compounds, some of which are currently being used to treat this disease. The different tools render different fractions of antiviral agents predicted to be non-biodegradable, which might be expected given that the criteria to define biodegradability differ in each, as indicated above. A potential user facing such apparent inconsistences should check these criteria carefully and interpret the predictions accordingly. Interestingly, according to the NITE criteria, that rendered the best performance in our tests, most of these antivirals belong to the "non-ready biodegradable" class. Most of these compounds are predicted to be non-toxic in mammals, as would be expected for such drugs. Nevertheless, it is not strange to find that some of them are predicted to be "toxic" since, as explained in the Materials and methods" section, this list includes experimental antiviral agents and others that are not yet approved.

As a result, we anticipate that the tool presented here will be useful both for guiding experimental biodegradation studies and for informing regulatory bodies when considering the approval of new active molecules for widespread use. To facilitate its use, we have developed a freely available web server where any interested user can generate predictions for his/her molecule of interest.

Data availability

All datasets used in this work are available upon request.

Supplementary data

Supplementary data are available at *Biology Methods and Protocols* online.

Acknowledgements

The authors thank the developers and curators of the biodegradation-related resources used in this work.

Funding

The work of F.P. was partially supported by the Spanish Ministry of Economy and Competitiveness with European Regional Development Fund [SAF2016-78041-C2-2-R]. The work of V.d.L. was partially funded by the SETH [RTI2018-095584-B-C42, MINECO/FEDER] and SyCoLiM [ERA-COBIOTECH 2018—PCI2019-111859-2] Projects of the Spanish Ministry of Science and Innovation, the MADONNA [H2020-FET-OPEN-RIA-2017-1-766975], BioRoboost [H2020-NMBP-BIO-CSA-2018-820699], SynBio4Flav [H2020-NMBP-TR-IND/H2020-NMBP-BIO-2018-814650], and MIX-UP [MIX-UP H2020-BIO-CN-2019-870294] Contracts of the European Union, as well as the InGEMICS-CM [S2017/BMD-3691] Project of the Comunidad de Madrid—European Structural and Investment Funds (FSE, FECER).

References

1. De Vleeschouwer F, Gérard L, Goormaghtigh C *et al.* Atmospheric lead and heavy metal pollution records from a Belgian peat bog spanning the last two millennia: human impact on a regional to global scale. *Sci Tot Environ* 2007;**377**:1–10.
2. Dixit RWMD, Pandiyan K, Singh UB *et al.* Bioremediation of heavy metals from soil and aquatic environment: an overview of principles and criteria of fundamental processes. *Sustainability* 2015;**7**:2189–212.
3. Sharma R. Bioremediation of oil-spills from shoreline environment. In: Oves M, Ansari MO, Zain Khan M, Shahadat M, M.I. Ismail I (eds.), *Modern Age Waste Water Problems*. New York: Springer, 2020, 275–91.
4. Rogowska J, Cieszynska-Semenowicz M, Ratajczyk W *et al.* Micropollutants in treated wastewater. *Ambio* 2020;**49**:487–503.
5. Varjani SJ. Microbial degradation of petroleum hydrocarbons. *Bioresour Technol* 2017;**223**:277–86.
6. Fewson CA. Biodegradation of xenobiotic and other persistent compounds: the causes of recalcitrance. *Trends Biotechnol* 1988;**6**:148–53.
7. de Lorenzo V, Prather KL, Chen G-Q *et al.* The power of synthetic biology for bioproduction, remediation and pollution control: the UN's Sustainable Development Goals will inevitably require the application of molecular biology and biotechnology on a global scale. *EMBO Rep* 2018;**19**:e45658.
8. Gao J, Ellis LBM, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Res* 2010;**38**:D488–D491.
9. Gomez MJ, Pazos F, Guijarro FJ *et al.* The environmental fate of organic pollutants through the global microbial metabolism. *Mol Syst Biol* 2007;**3**:114.
10. Hafner J, MohammadiPeyhani H, Sveshnikova A *et al.* Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *ACS Synth Biol* 2020;**9**:1479–82.
11. Dimitrov S, Breton R, MacDonald D *et al.* Quantitative prediction of biodegradability, metabolite distribution and toxicity of stable metabolites. *SAR QSAR Environ Res* 2002;**13**:445–55.
12. Dimitrov S, Kamenska V, Walker JD *et al.* Predicting the biodegradation products of perfluorinated chemicals using CATABOL. *SAR QSAR Environ Res* 2004;**15**:69–82.
13. Raies AB, Bajic VB. *In silico* toxicology: computational methods for the prediction of chemical toxicity. *Wires Comput Mol Sci* 2016;**6**:147–72.
14. Pazos F, Valencia A, De Lorenzo V. The organization of the microbial biodegradation network from a systems-biology perspective. *EMBO Rep* 2003;**4**:994–9.
15. Lewis KA, Tzilivakis J, Warner DJ *et al.* An international database for pesticide risk assessments and management. *Hum Ecol Risk Assess Int J* 2016;**22**:1050–64.
16. Takatsuki M, Takayanagi Y, Kitano M. In: Peijnenburg WJGM, Karcher W (eds). An attempt to SAR of biodegradation. In: Peijnenburg WJGM, Karcher W (eds.), *Proceedings of the Workshop "Quantitative Structure Activity Relationships for Biodegradation"*. Bilthoven, The Netherlands: National Institute of Public Health and Environmental Protection (RIVM), 1995, 67–103.

17. Tang Y, Zhang Y-Q, Chawla NV, *et al.* SVMs modeling for highly imbalanced classification. *IEEE Trans Cybernet* 2009;**39**: 281–8.
18. Wishart DS, Feunang YD, Guo AC *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
19. O'Boyle NM, Banck M, James CA *et al*. Open babel: an open chemical toolbox. *J Cheminform* 2011;**3**:33.
20. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:1–27.
21. Buhmann MD. *Radial Basis Functions: Theory and Implementations*. Cambridge, UK: Cambridge University Press, 2003.
22. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006;**27**:861–74.
23. Bienfait B, Ertl P. JSME: a free molecule editor in JavaScript. *J Cheminform* 2013;**5**:24.