Environmental Sciences Europe

RESEARCH

Open Access

The applicability domain of EPI Suite™ for screening phytotoxins for potential to contaminate source water for drinking

Inés Rodríguez-Leal* and Matthew MacLeod

Abstract

Background: Toxins produced by plants constitute a potential threat to water supplies in Europe, but have not been widely considered in systematic risk assessments. One way to begin to address potential risks of phytotoxins is to conduct screening-level assessments of known phytotoxins for their potential to contaminate source water for drinking due to persistence (P) and mobility (M). Chemical properties relevant for such an assessment (octanol–water partition coefficient K_{OW} and biodegradation half-life) can be estimated from the structure of phytotoxins with quantitative structure–property relationship (QSPR) models found in the United States Environmental Protection Agency's Estimation Program Interface (EPI Suite™) software, but predictions must be considered critically since these models have been developed using data for anthropogenic chemicals and many phytotoxins could lie outside their applicability domain.

Results: We analyzed two EPI Suite™ models—KOWWIN and BIOWIN5—by evaluating the quality of property predictions for their validation sets as a function of Euclidean distances d_E to the centroid of descriptor space of the models' training sets. We identified model-specific applicability domain boundaries as local maxima in plots of the difference between root mean square error (ΔRMSE) of modeled property values of validation set compounds within and outside applicability domain boundaries defined by a continuum of possible boundaries. And, we also evaluated four generic boundaries that have been suggested in literature. The ΔRMSE between validation set compounds outside and inside applicability domain boundaries had positive values for all but one of the possible boundaries we considered, indicating that properties of chemicals with d_E inside the boundaries were better predicted. With our proposed model-specific boundaries, 21% of 1586 phytotoxins produced by plants found in Switzerland were out of domain of KOWWIN, and 46% of were out of domain of BIOWIN5.

Conclusions: Estimates of Log K_{OW} and biodegradation half-life of phytotoxins that lie outside the domain of applicability of the QSPR models should be viewed as extrapolations that are subject to unquantified and potentially large errors. Phytotoxins outside the domain of applicability of QSPR models should be prioritized for property measurements as a basis to expand the training sets of QSPR models and to support hazard identification for better management of drinking water quality in Europe.

Keywords: Natural toxins, QSPR, Plant specialized compounds, Plant secondary metabolites

Background

Natural toxins are toxic chemicals produced by living organisms including plants, algae, fungi and cyanobacteria. They encompass several different chemical classes and constitute a potential risk to human and

*Correspondence: ines.rodriguezleal@aces.su.se

Department of Environmental Science, Stockholm University, Svante Arrhenius väg 8, 11418 Stockholm, Sweden

environmental health [1]. Natural toxins are environmental contaminants of emerging concern, particularly in the context of climate change, which could drive changes in species distribution patterns, and metabolic paths that induce increased toxin production [2].

Natural toxins produced by plants (phytotoxins) pose a hazard in Europe, but the quantity and severity is unknown [1]. An example of a problematic plant-produced natural toxin is ptaquiloside, which is synthesized by Bracken fern (*P. aquilinum*) and is known to leach into soil and reach surface and groundwater that is sourced for drinking [3]. Another example is aristolochic acids, which can leach from European birthwort (*A. clematitis*), contaminate groundwater, and are known to cause kidney damage [4]. Recently, there have been some studies that reported the presence and abundance of other groups of phytotoxins, such as quinolizidine alkaloids [5], pyrrolizidine alkaloids [6] and saponins [7], but field data of this kind are still scarce. Assessing phytotoxins based on their chemical properties to identify other potential persistent, mobile and toxic (PMT) chemicals that could impact source water for drinking is hence an important research priority [8].

Recently, Günthardt and co-workers published the Toxic Plants-PhytoToxins (TPPT) database [9], which includes 1586 toxins produced by plants in Switzerland. They scored the phytotoxins in the TPPT database for their potential to contaminate water based on their persistence and mobility using Log K_{OW} (i.e., the log of the octanol–water partition coefficient) and biodegradation half-lives predicted with quantitative structure–property relationship (QSPR) models in EPI Suite™ [10].

The QSPR models in EPI Suite™ have been widely applied for predicting physico-chemical properties and half-lives of chemicals for screening-level hazard assessment (for example, [11]). QSPR models, which are examples of QSAR (Quantitative Structure Activity Relationship) models use property measurements for a "training set" of compounds to build a model, and measurements for a "validation set" of compounds to quantify model performance. KOWWIN was developed to estimate measured Log K_{OW} values [12], and BIOWIN to estimate half-lives from expert opinions [13]. Both models use a fragment approach, in which structural fragments and other molecular descriptors are the building blocks of regression models. The EPI Suite™ models were developed using property data measured for training sets of mostly anthropogenic chemicals, such as persistent organic pollutants and other organochlorine pesticides, modern pesticides, personal care products and industrial chemicals. Relatively few compounds of natural origin were included, and only 2% and 0.6% of the phytotoxins in the TPPT database are found in the training sets for KOWWIN and BIOWIN5, respectively.

It is crucial to define the applicability domain of QSPRs to evaluate the adequacy of the models to predict chemical properties [14]. The OECD principles for the validation of QSAR models highlight the need to have a defined domain of applicability which specifies the scope of the model [15]. Conceptually, a model's predictions for chemicals within the applicability domain can be viewed as interpolations and those for chemicals outside the applicability domain can be viewed as extrapolations [16]. When QSPR models are extrapolated to predict properties for chemicals outside the applicability domain, the predictions are subject to unknown and unquantifiable uncertainties. However, there is no single accepted method to define the applicability domain of a QSPR. One way of evaluating the applicability domain of a QSPR model is by determining the locations of chemicals in the training set in a chemical space defined by the molecular descriptors used in the model, and defining an applicability domain boundary in that descriptor space [17].

In this work, we define model-specific applicability domain thresholds for KOWWIN and BIOWIN5 and assess whether the 1586 plant toxins in the TPPT database are within the applicability domains of these models using both our model-specific thresholds and generic threshold values that have been proposed in the literature. Our goal is to identify phytotoxins in the database that are outside the applicability domain of the QSPR models, and thus have unquantifiable uncertainties in their modeled persistence and mobility profile. These substances should be assigned a high priority for experimental determination of Log K_{OW} and biodegradability to reduce uncertainties in their environmental fate profile, and to generate data that can be added to QSPR training sets to improve screening assessments of phytotoxins.

Materials and methods

The TPPT database details toxins produced by a variety of different plant species found in Switzerland and Central Europe [9]. The 1586 chemicals in the database are each assigned a "toxin number" (T1–T1586), and are characterized according to occurrence frequency in Switzerland, estimated toxicity and environmentally relevant properties, including half-lives and Log K_{OW} values. For most toxins, Günthardt et al. included physico-chemical properties and half-lives in the TPPT database that were estimated from structure using EPI Suite™ v4.11 [10].

EPI Suite™ is a compilation of QSPR models that predict environmentally relevant properties of organic chemicals, such as Log K_{OW} by KOWWIN and aerobic biodegradability by BIOWIN5. The QSPR models are based on molecular descriptors that represent the structure of molecules [18]. In EPI Suite™, structural fragment counts and correction factors are used as molecular descriptors [19]. Each chemical's structure is described as a numerical vector containing the number of times each descriptor used in the QSPR model is present. The models are formulated by multiple linear regression of the number of occurrences of each descriptor for the chemicals in the training set against the property of interest. KOWWIN has 2447 chemicals in the training set and 10,873 in the validation set, and BIOWIN5 has 589 chemicals in the training set and 295 in the validation set.

The entire list of molecular descriptors and coefficients for the QSPRs in EPI Suite™ is not directly accessible in the software, but the list of training and validation sets are available at http://esc.syrres.com/interkow/EpiSuiteData.htm We obtained the KOWWIN and BIOWIN5 training and validation sets by feeding the training and validation set compound structures into the models and parsing batch output files with *Python* v. 3.7.3 [20] and the *pandas* package [21]. This way, we re-assembled the dataset that was used to build the model, and we then constructed matrices that contained all information about the molecular descriptors of training and validation sets of KOWWIN and BIOWIN5. We used the same method to extract the descriptor counts for the phytotoxins and produce a matrix with the same format. The end results were three matrices of molecular descriptors for the training, validation and TPPT sets of chemicals, and the property data for the training and validation sets of chemicals.

To define the applicability domain of KOWWIN and BIOWIN5, we measured distance of chemicals to the centroid of the training set in multi-dimensional chemical descriptor space. To reduce the dimensionality of the chemical descriptor space, principal component analysis (PCA) was performed. In PCA, each component encompasses variability associated within several descriptors, reducing the number of dimensions. We therefore calculated the position of compounds in each of n PCA-reduced dimensions of molecular descriptor space. Then the unitless Euclidean distance ($d_{\mathrm{E},x}$) that each compound (x) lies from the centroid of the space occupied by compounds in the training set is calculated as:

$$d_{E,x} = \sqrt{\sum_{i=1}^{n} (p_{i,x} - \overline{p}_i)^2},$$

where n is the number of dimensions in the PCA-reduced space of the QSPR model, $p_{i,x}$ is the location of compound x in PCA-reduced dimension i, and $\overline{p}_i$ is the average location of training set compounds in PCA-reduced dimension i in the training set. Greater d_{E} values indicate higher structural dis-similarity to the compounds in the training set and hence d_{E} thresholds have been used to classify chemicals as being outside of domain [22]. We calculated d_{E} for every chemical in the TPPT set, the validation set and for the training set itself with R v. 4.0.2. [23], packages *Rmarkdown* [24] and *papaja* [25] as well as multiple helper functions [26–30].

Thresholds

We consider four threshold values of d_{E} from literature [14]; to distinguish between toxins inside and outside domain; (1) the maximum distance of a member of the training set (d_{Emax}), (2 and 3) three times ($3\overline{d}_{\mathrm{E}}$) and twice ($2\overline{d}_{\mathrm{E}}$), the average Euclidean distance of members of the training set ($\overline{d}_{\mathrm{E}}$), and 4) the 95th percentile distance (d_{E95}) of the members of the training set.

Furthermore, we estimated model-specific d_{E} thresholds by exploring the performance of the models at calculating property values for members of the validation set as a function of d_{E}. To this end, we plotted the difference in root mean squared error (ΔRMSE) of chemicals outside a continuum of boundaries versus chemicals inside the boundaries by sorting the chemicals in the training set by increasing d_{E}, then iteratively calculating ΔRMSE for boundaries where ($n = 1,2,3,\ldots x$) of the training set chemicals with lowest d_{E} were assumed to be "in domain". We then plotted ΔRMSE of the validation set as a function of the percentile rank of the training set chemical in our ordered list that defined the d_{E} threshold. For reference, and to confirm that d_{E} is a meaningful metric for defining domain of applicability, we performed the same analysis 10 times with the training set ordered randomly instead of by increasing d_{E}. We then identified applicability domain thresholds unique to KOWWIN and BIOWIN5 at local maxima in the plot of ΔRMSE calculated for the training set ordered by increasing d_{E}.

Results

Euclidean distances for KOWWIN and BIOWIN5 for training, validation and phytotoxin sets

For KOWWIN, Euclidean distances of the compounds in the three datasets from the centroid of the training set (Fig. 1, left panel) range from 1.34 to 78.41 for the training set, from 1.31 to 98.03 for the validation set, and from 1.47 to 131.89 for the phytotoxin set. In BIOWIN5, they range from 0.75 to 31.51 for the training set, from 0.84 to 46.85 for the validation set, and for phytotoxins from 0.79 to 77.24 (Fig. 1, right panel).

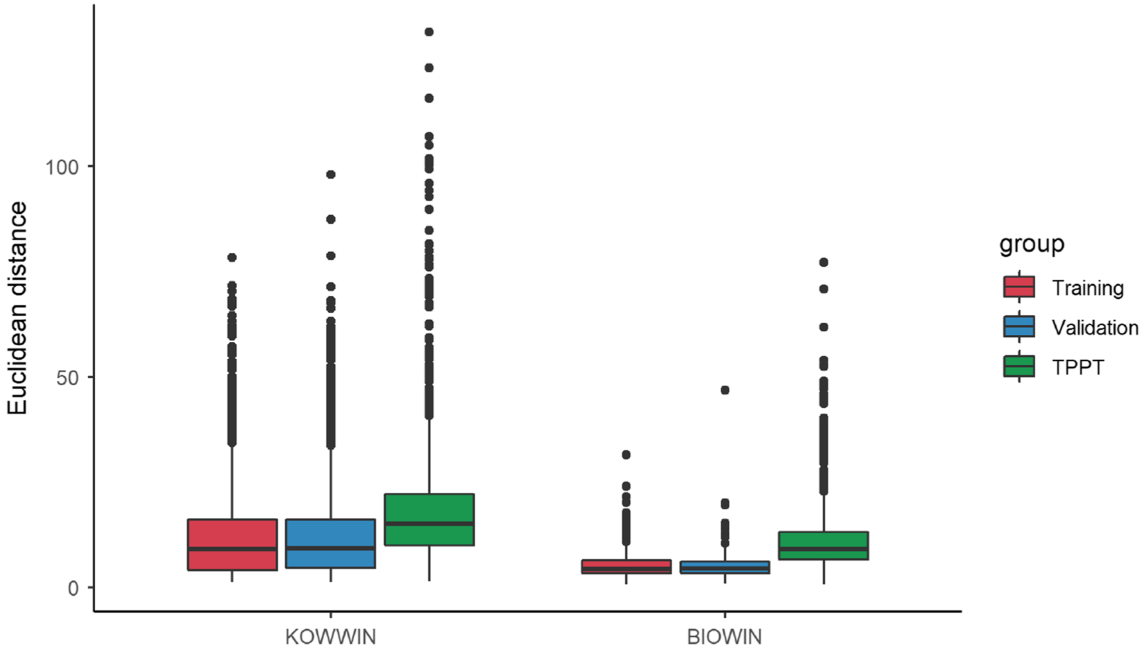

Fig. 1 Distribution of d_E values in the training, validation and phytotoxin compound sets of KOWWIN and BIOWIN5. Boxes encompass the interquartile range (IQR) and whiskers extend to 1.5 times IQR. Training and validation sets have similar d_E values, while the phytotoxin set has consistently greater values of d_E in the structural descriptor space of both models

Threshold estimation based on ΔRMSE

Results of our search for model-specific applicability domain boundaries at local maxima in prediction accuracy as a function of d_E for KOWWIN and BIOWIN5 are shown in Fig. 2. The applicability domain of QSPR models is a property of the model and its training set. Therefore, for each model the primary horizontal axis of Fig. 2 is percentile of Euclidean distance of members of the training set from its centroid.

For KOWWIN (top panel in Fig. 2), the RMSE of property predictions for validation set compounds outside the applicability domain threshold is consistently higher than for compounds inside the domain threshold for nearly the entire range of possible thresholds (red line). ΔRMSE decreases with a shallow slope from a peak at about the 5th percentile d_E of training set compounds until a first inflection point at approximately the 85th percentile d_E of the training set compounds, where ΔRMSE begins to increase. A local maximum of ΔRMSE occurs at the 94th percentile d_E of the training set compounds, which is marked as our proposed model-specific applicability domain threshold (vertical dotted line in the top panel in Fig. 2). Above the 94th percentile ΔRMSE first decreases, then swings randomly as the number of compounds outside the domain threshold becomes small. For reference, ΔRMSE for the continuum of thresholds in ten randomly ordered lists of the KOWWIN validation set compounds (grey lines in top panel of Fig. 2) are close to zero across a wide range of thresholds, and deviate randomly only above the 94th percentile d_E of the training set compounds (grey lines in the top panel in Fig. 2).

For BIOWIN5 (bottom panel in Fig. 2), ΔRMSE for a continuum of thresholds of d_E (red line) is close to zero and within the range of variability of thresholds defined in randomly ordered lists of compounds (grey lines) up to about the 80th percentile d_E of the training set compounds. It rises to a local maximum at the 93rd percentile, which is marked in the bottom panel of Fig. 2 as our proposed applicability domain threshold (vertical dotted line). Above the 93rd percentile ΔRMSE plateaus, then increases with increasing d_E, in a region where the high variability in the 10 randomly ordered lists indicates that ΔRMSE is determined by model performance for a small number of compounds classified as outside of domain (compare red and grey lines in the bottom panel of Fig. 2).

Comparison of chemicals outside and inside domain using different thresholds

Our model-specific applicability domain threshold for KOWWIN is at the 94th percentile d_E of the training set, and is very close to the 95th percentile threshold suggested in previous literature [14] (Fig. 3 and Table 1). Our suggested model-specific threshold

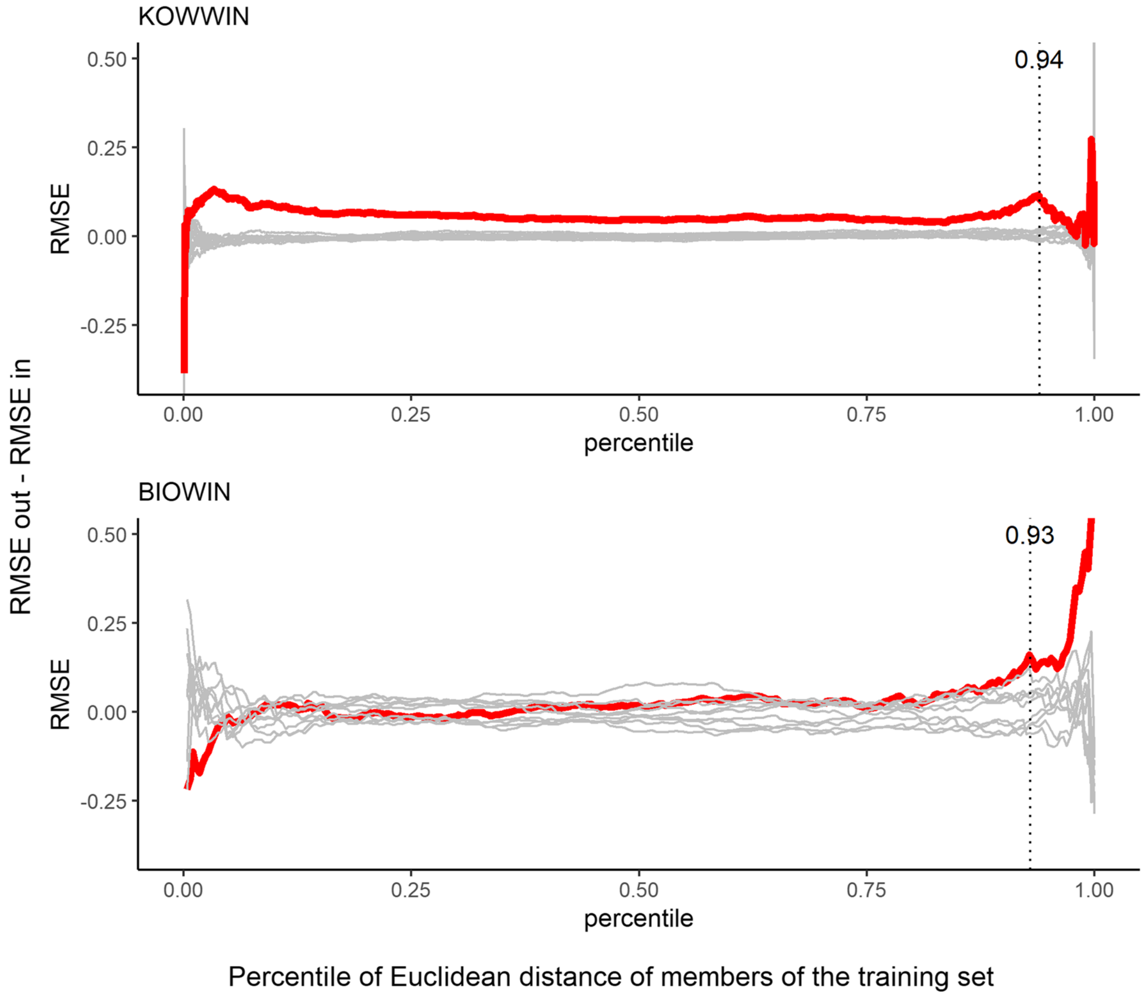

Fig. 2 Red: difference in root mean squared error of predictions (ΔRMSE) for validation set compounds with Euclidean distances below and above a continuum of thresholds set at percentiles of Euclidean distance of training set compounds from their centroid, in both models (top: KOWWIN, bottom: BIOWIN). Grey: ΔRMSE for 10 trials of randomly ordered lists of the validation set chemicals

for KOWWIN is set at local maximum of ΔRMSE, where Log K_{OW} of validation set compounds inside the threshold are predicted 0.17 log units more accurately on average than Log K_{OW} of compounds outside the threshold (Table 1). However, ΔRMSE is also positive and similar in size for three of the other thresholds that have been suggested in literature, $3\overline{d}_E$, $2\overline{d}_E$ and d_{E95q}. The exception is d_{Emax}, which has a ΔRMSE value close to zero that likely reflects random variability in prediction errors for the few validation set compounds that lie above this threshold (see Fig. 3).

Our model-specific threshold for BIOWIN5 is at the 93rd percentile d_E of the training set, and is again close to the 95th percentile threshold suggested in previous literature (Fig. 4 and Table 1). Because ΔRMSE for the BIOWIN validation set increases with increasing d_E above our model-specific threshold (Fig. 2), the ΔRMSE values for our model-specific threshold are not the highest of the group of possible thresholds (Table 1).

Distribution of in and out-of-domain phytotoxins across chemical classes

Among chemical classes, phytotoxins, steroids and saponins are most likely to be out of domain in both KOWWIN and BIOWIN when using our model-specific thresholds (Table 2). All saponins and 95% of steroids were classified as out of domain for BIOWIN5. On the other hand, 97% of the alkaloids and phenylpropanoids in the TPPT database are classified as within the applicability domain of KOWWIN, but higher percentages (46% and 19%) of these chemicals are out-of-domain for BIOWIN5 (Table 2).

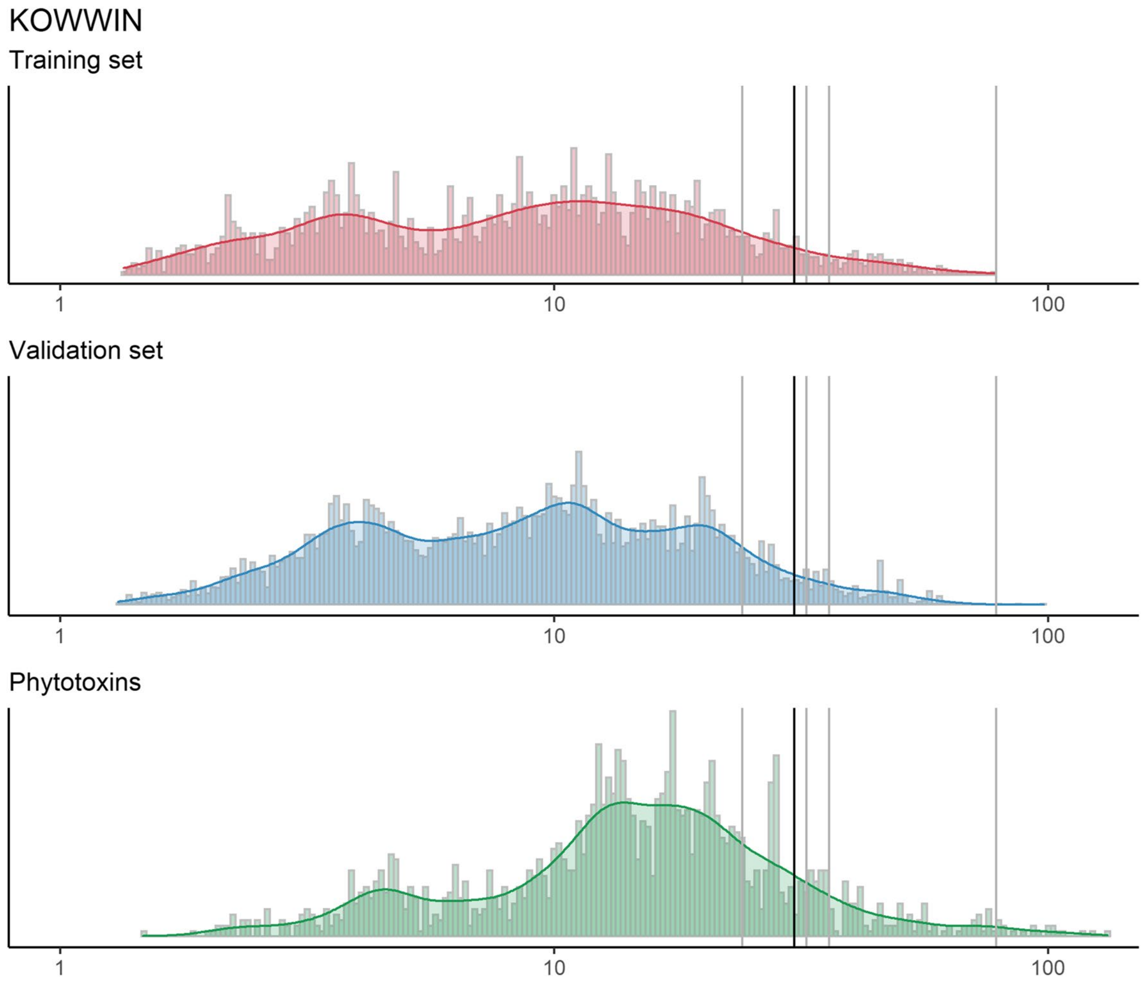

Fig. 3 KOWWIN: Euclidean distance to the centroid of the training set (d_E, plotted on a log scale on the horizontal axis) versus number density of chemicals in the training set (red), validation set (blue), and for the phytotoxins (green). Vertical grey lines correspond to thresholds from literature [14] (d_{Emax}, $3\bar{d}_E$, $2\bar{d}_E$ and d_{E95q}) and our estimated threshold from Fig. 2 is represented as a black line, d_{Emodel}

Discussion

There is no universal method to establish the applicability domain of a QSPR model [14, 31, 32]. In this work, we define applicability domain using different Euclidean distance thresholds. Structural similarity is measured exclusively using the chemical space defined by structural descriptors that are part of the models. All Euclidean distance threshold approaches to define applicability domain are thus specific to the selected model and its descriptors since a different chemical space defined by different molecular descriptors would be distorted [31].

Our results show that up to 46% of the TPPT database predictions for half-lives reported by Günthardt and co-workers [9] are outside the applicability domain of BIOWIN and thus subject to unquantifiable errors in the predictions. Moreover, multiple stereoisomers and a high number of ionized forms are expected to be present

Table 1 Comparison of five different thresholds to define the domain of applicability of KOWWIN and BIOWIN5

Model	Threshold	Value	ΔRMSE	Toxins
KOWWIN	d_{Emax}	78.41	− 0.03	1.20
	$3\bar{d}_E$	36.04	0.09	7.57
	$2\bar{d}_E$	24.02	0.13	20.81
	d_{E95q}	32.44	0.14	10.84
	d_{Emodel}	30.65	0.17	11.92
BIOWIN	d_{Emax}	31.51	0.98	3.96
	$3\bar{d}_E$	15.66	0.40	16.79
	$2\bar{d}_E$	10.44	0.14	40.25
	d_{E95q}	10.53	0.14	38.93
	d_{Emodel}	9.72	0.08	45.60

d_{Emax}—maximum Euclidean distance of a member of the training set from the centroid of the training set, $3\bar{d}_E$ and $2\bar{d}_E$—3 and 2 three times the average Euclidean distance of members of the training set from the centroid, d_{E95q}—the 95th percentile distance of the members of the training set from the centroid, and d_{Emodel}—the model-specific threshold identified at an inflection point of ΔRMSE (Fig. 2)

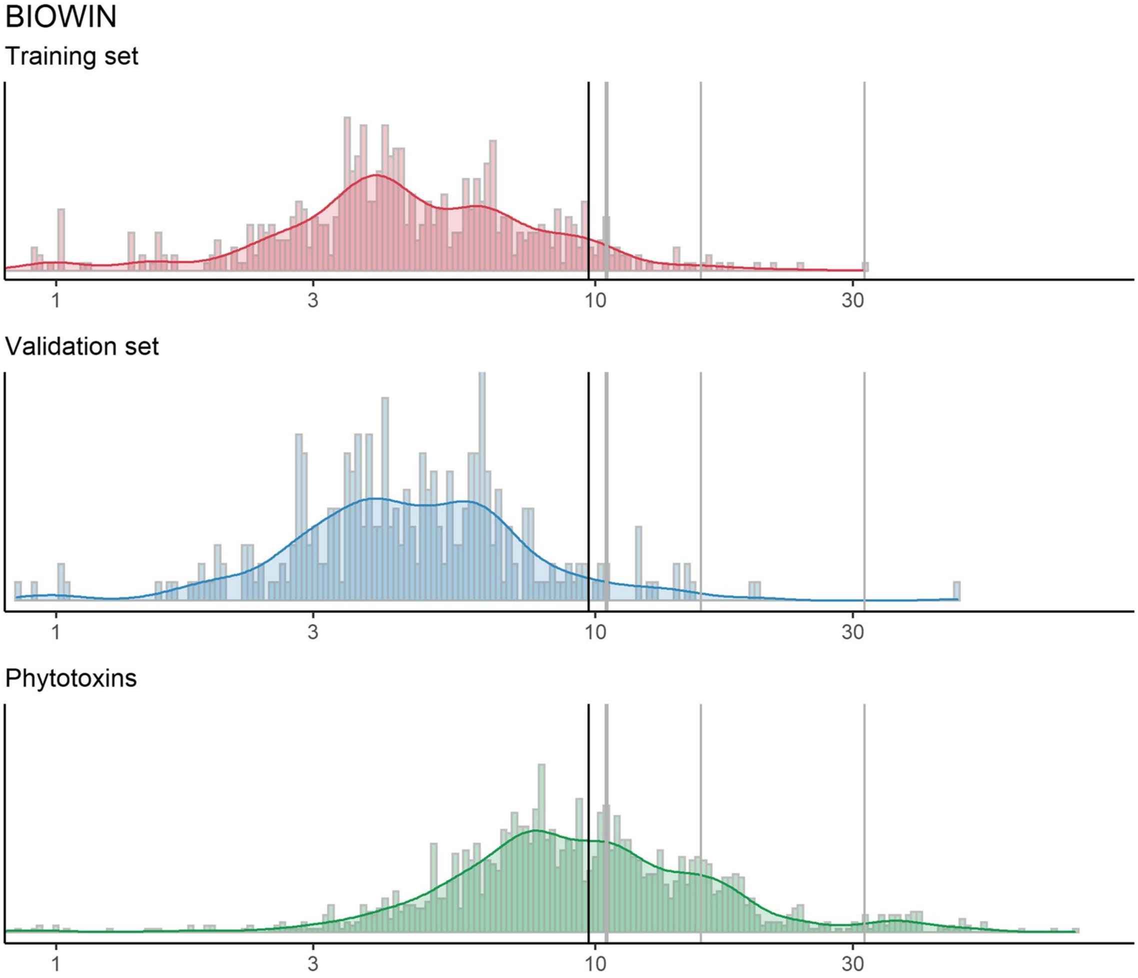

Fig. 4 BIOWIN: Euclidean distance to the centroid of the training set (d_E, plotted on a log scale on the horizontal axis) versus number density of chemicals in the training set (red), validation set (blue), and for the phytotoxins (green). Vertical grey lines correspond to thresholds from literature [14] (d_{Emax}, 3 $\overline{d}_E$, 2 $\overline{d}_E$ and d_{E95q}) and our estimated threshold from Fig. 2 is represented as a black line, d_{Emodel}

among many members of this group, which are not represented in the BIOWIN5 descriptor set.

The applicability domain of KOWWIN encompasses more of the natural toxins in the TPPT database. Only up to 21% of toxins in the database are outside the applicability domain of KOWWIN and thus subject to unquantifiable errors. Research to improve exposure hazard screening of phytotoxins should prioritize chemicals outside the applicability domain of KOWWIN and BIOWIN for experimental determination of properties, as these are the chemicals where new measurements will constrain unquantified uncertainties, and also could be used to expand the training sets of QSPRs in a way that also expands their domain of applicability.

The TPPT database of phytotoxins is focused on toxins produced by plants that grow in Switzerland, and thus is not fully representative of plant biodiversity in Europe. However, the database from Günthardt et al. is the largest compendium of European plant toxins to date, and a good starting point for screening assessment. Nevertheless, more specific knowledge of biodiversity and occurrence of toxin-producing plants is needed to estimate potential phytotoxin hazard for a specific region.

Conclusions

Our distance approach for determining applicability domain of QSPR models shed light on the adequacy of using KOWWIN and BIOWIN5 models for physicochemical property estimation in the TPPT database. The model-specific approach produced applicability domain thresholds close to the 95th percentile for both

Table 2 Distribution of in and out-of-domain phytotoxins across chemical classes using the model-specific thresholds (d_{Emodel})

	Out-of-domain (n)	In-domain (n)	% of toxins out-of-domain
KOWWIN			
Alkaloids	18	599	3
Terpenoids	36	348	9
Steroids	48	63	43
Polyketides	22	89	20
Saponins	46	28	62
Phenylpropanoids	2	72	3
Others	17	198	8
BIOWIN			
Alkaloids	285	332	46
Terpenoids	153	231	40
Steroids	105	6	95
Polyketides	39	72	35
Saponins	74	0	100
Phenylpropanoids	14	60	19
Others	50	165	23

models. The 95th percentile threshold [14] could thus be a preferred choice for defining the applicability domain of other similar QSPR models.

Using the model-specific threshold, almost half of the TPPT phytotoxins were out of domain of BIOWIN, and steroids and saponins were nearly all out of domain. Saponins are the biggest out-of-domain group in both models, suggesting their structure is underrepresented in the training sets. Therefore, measuring property data for these substances should be prioritized. Model predictions for out-of-domain phytotoxins are extrapolations, and thus their predicted hazard properties carry unknown uncertainties and should be interpreted with caution.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-022-00676-2.

Additional file 1. Euclidean distances and classification as in- or out-of-domain of BIOWIN5 for toxins in the TPPT database.

Additional file 2. Euclidean distances and classification as in- or out-of-domain of KOWWIN for toxins in the TPPT database.

Acknowledgements
The NaToxAq ITN is supported by the Horizon 2020 research and innovation programme of the European Union under Marie Sklodowska-Curie grant agreement No. 722493. We would like to thank all partners and colleagues involved.

Author contributions
IRL has been responsible for the concept of the manuscript and drafted the manuscript. MM helped to further elaborate the manuscript and contributed to specific aspects. All authors read and approved the final manuscript.

Funding
Open access funding provided by Stockholm University. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 722493.

Availability of data and materials
A list of toxins in the TPPT database, their Euclidean distance, and whether they are in- or out-of-domain according to the different boundaries discussed in this paper is available in Additional file 1and 2.

Declarations

Ethics approval and consent to participate
There are no issues in the reported work that required ethic approval in any of the contributing institutions. Consent on participation was obtained from all co-authors.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 May 2022 Accepted: 10 September 2022
Published online: 27 September 2022

References

1. Bucheli TD (2014) Phytotoxins: environmental micropollutants of concern? Environ Sci Technol 48:13027–13033. https://doi.org/10.1021/es504342w
2. Visser PM, Verspagen JMH, Sandrini G et al (2016) How rising CO_2 and global warming may stimulate harmful cyanobacterial blooms. Harmful Algae 54:145–159. https://doi.org/10.1016/j.hal.2015.12.006
3. Rasmussen LH, Kroghsbo S, Frisvad JC, Hansen HCB (2003) Occurrence of the carcinogenic Bracken constituent ptaquiloside in fronds, topsoils and organic soil layers in Denmark. Chemosphere 51:117–127. https://doi.org/10.1016/S0045-6535(02)00694-X
4. Tung K-K, Chan C-K, Zhao Y et al (2019) Occurrence and environmental stability of aristolochic acids in groundwater collected from Serbia: links to human exposure and Balkan endemic nephropathy. Environ Sci Technol 54:1554–1561
5. Hama JR, Jorgensen DBG, Diamantopoulos E et al (2022) Indole and quinolizidine alkaloids from blue lupin leach to agricultural drainage water. Sci Total Environ 834:155283
6. Hama JR, Strobel BW (2021) Occurrence of pyrrolizidine alkaloids in ragwort plants, soils and surface waters at the field scale in grassland. Sci Total Environ 755:142822
7. Jiang X, Strobel BW, Cedergreen N et al (2019) Stability of saponin biopesticides: hydrolysis in aqueous solutions and lake waters. Environ Sci Process Impacts 21:1204–1214
8. Reemtsma T, Berger U, Arp HPH et al (2016) Mind the gap: persistent and mobile organic compounds. Environ Sci Technol 50:10308–10315. https://doi.org/10.1021/acs.est.6b03338
9. Günthardt BF, Hollender J, Hungerbühler K et al (2018) Comprehensive toxic plants and its application in assessing aquatic micropollution potential. J Agric Food Chem 66:7577–7588. https://doi.org/10.1021/acs.jafc.8b01639
10. EPA U (2009) Exposure assessment tools and models, estimation program interface (EPI) suite
11. Reppas-Chrysovitsinos E, Sobek A, MacLeod M (2018) In silico screening-level prioritization of 8468 chemicals produced in OECD countries

to identify potential planetary boundary threats. Bull Environ Contam Toxicol 100:134–146. https://doi.org/10.1007/s00128-017-2253-9

12. Meylan WM, Howard PH (1995) Atom/fragment contribution method for estimating octanol–water partition coefficients. J Pharm Sci 84:83–92
13. Tunkel J, Howard PH, Boethling RS et al (2000) Predicting ready biodegradability in the Japanese ministry of international trade and industry test. Environ Toxicol Chem: Int J 19:2478–2485
14. Sahigara F, Mansouri K, Ballabio D et al (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17:4791–4810. https://doi.org/10.3390/molecules17054791
15. Economic Co-operation O for (OECD) D (2004) The report from the expert group on (quantitative) structure-activity relationships [(q) SARs] on the principles for the validation of (q) SARs. OECD Environment Health and Safety Publications 206
16. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701. https://doi.org/10.1002/qsar.200610151
17. Sahigara F, Ballabio D, Todeschini R, Consonni V (2014) Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective. Curr Comput Aided Drug Des 10:137–147
18. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley-Interscience, New York
19. Hansch C, Rockwell SD, Jow PYC et al (1977) Substituent constants for correlation analysis. J Med Chem 20:304–306. https://doi.org/10.1021/jm00212a024
20. Van Rossum G, Drake FL (2009) Python 3 reference manual. CreateSpace, Scotts Valley
21. McKinney W et al (2010) Data structures for statistical computing in python. In: Proceedings of the 9th python in science conference. Austin TX, p 51–56
22. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Altern Lab Anim 33:445–459
23. R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
24. Xie Y, Allaire JJ, Grolemund G (2018) R markdown: the definitive guide. Chapman; Hall/CRC, Boca Raton
25. Aust F, Barth M (2020) papaja: Create APA manuscripts with R Markdown
26. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York
27. Kassambara A (2020) Ggpubr: 'ggplot2' based publication ready plots
28. Wickham H, Henry L (2020) Tidyr: Tidy messy data
29. Bache SM, Wickham H (2014) Magrittr: A forward-pipe operator for r
30. Dowle M, Srinivasan A (2019) Data.table: Extension of 'data.frame'
31. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim 33:445–459
32. Eriksson L, Jaworska J, Worth AP et al (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.