Interpretation of QSAR models: mining structural patterns taking into account molecular context

Mariia Matveieva[1], Mark T.D. Cronin[2], Pavel Polischuk[1,3,*]

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University and University Hospital in Olomouc, Hnevotinska 5, 77900 Olomouc, Czech Republic

[2] School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, United Kingdom

[3] A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kremlevskaya Str. 10, Kazan, Russia

Abstract

The study focused on QSAR model interpretation. The goal was to develop a workflow for the identification of molecular fragments in different contexts important for the property modelled. Using a previously established approach – Structural and physicochemical interpretation of QSAR models (SPCI) – fragment contributions were calculated and their relative influence on the compounds' properties characterised. Analysis of the distributions of these contributions using Gaussian mixture modelling was performed to identify groups of compounds (clusters) comprising the same fragment, where these fragments had substantially different contributions to the property studied. SMARTSminer was used to detect patterns discriminating groups of compounds from each other and visual inspection if the former did not help. The approach was applied to analyse the toxicity, in terms of 40 hour inhibition of growth, of 1984 compounds to *Tetrahymena pyriformis*. The results showed that the clustering technique correctly identified known toxicophoric patterns: it detected groups of compounds where fragments have specific molecular context making them contribute substantially more to toxicity. The results show the applicability of the interpretation of QSAR models to retrieve reasonable patterns, even from data sets consisting of compounds having different mechanisms of action, something which is difficult to achieve using conventional pattern/data mining approaches.

Introduction

Mechanistic interpretation of QSAR models is useful to understand the complex nature of biological or physicochemical processes. It can be applied to drug and product development to optimise the structures of studied compounds by increasing efficacy and reducing harmful effects. Model interpretation can also serve as a means to confirm the validity of a model, i.e. that the model has captured relevant and meaningful relationships between activity and structure [1]. It is also important for regulatory application, for example, the fifth of the OECD Principles for the Validation of QSARs requires, where possible, mechanistic interpretation on QSAR models. [2]. Whilst this principle is optional it is considered helpful to get round the long held belief that QSAR models were “black boxes” and interpretation was not always possible. Recently, several approaches have been proposed that have assisted in making QSAR models interpretable and have helped to establish a new paradigm for interpretation [1]. These methods are closely related to the matched molecular pairs (MMP) approach [3]. They allow for the calculation of the contributions of single atoms [4], arbitrary fragments [5], or predicted activity changes corresponding to given molecular transformations [6]. The contributions of identical fragments/transformations are usually averaged to reveal general trends in structure-property relationships [7-11]. Whilst useful, currently none of these approaches take into account the molecular context of the fragments considered which may significantly influence the fragment’s behavior, e.g. transforming a “safe” non-toxicophoric moiety into a reactive group. It has been demonstrated that consideration of the molecular environment may improve the outcome of the MMP analysis [12]. Therefore, we expect that capturing molecular context of fragments can improve interpretation of QSAR models.

The aim of this study was to develop and utilise an unsupervised approach to analyse fragment contributions and the influence of molecular context on biological activity. This influence was captured implicitly by analysing distributions of the contributions of fragments from different compounds. Three scenarios were considered as illustrated in **Error! Reference source not found.**. The distribution of fragment contributions may constitute a narrow range of values (**Error! Reference source not found.**A). Since each single contribution comes from a particular molecule, such a shape may indicate that the influence of molecular context is insignificant or all compounds comprise the fragment in a very similar context. Broad distributions indicate that the molecular context could substantially influence calculated contributions (**Error! Reference source not found.**B and C). Broad distributions can have a single, or several, peaks. Multiple peaks on a distribution (**Error! Reference source not found.**C) may indicate different contexts of the fragment. Therefore, it would be reasonable to identify subpopulations (clusters) of fragments having different contributions, and inspect compounds corresponding to these subpopulations for meaningful structural patterns. To achieve these aims Gaussian mixture modelling (GMM) was used. Compounds corresponding to different clusters were analyed using SMARTSminer [13] in order to find discriminative structural patterns (**Error! Reference source not found.**). We applied the approach developed to analyse QSAR models for the toxicity of a range of compounds to *Tetrahymena pyriformis* in order to identify toxicophoric patterns.

Materials and methods

Data set

The data set analysed was for the inhibition of growth of chemical substances to the ciliated protozoan *Tetrahymena pyriformis* represented as lg(1/IGC_{50}) (IGC_{50} in mol/l). Toxicity data were taken from the study of Ruusmann and Maran [14]. Standardizer was used for structure standardisation and tautomerisation, in addition structures were checked for errors [15] and duplicates were removed. The data set curation workflow is available from the public repository - https://bitbucket.imtm.cz/projects/STD/repos/std/browse. The curated data set comprised 1984 compounds whose structures and activity values are provided in Supplementary materials. All modelling steps including descriptor calculation, model development and validation, molecule fragmentation and calculation of fragment contributions were performed with the open-source *spci* software [9, 16].

Descriptors

Counts of fragments having 2-4 heavy atoms were used as the descriptor set. Fragments were determined with either all atoms connected or containing two disconnected parts. The atoms in the fragments were labelled according to their partial charge, lipophilicity, refractivity and ability to form H-bonds calculated using Chemaxon cxcalc utility [17]. Descriptors were calculated by the *sirms* tool included in the *spci* software [18]. More details about descriptor calculation are available in previous publications [8-9, 16].

Model building

Four regression QSAR models were built using Random Forest (RF), Partial Least Squares (PLS), Gradient Boosting Machine (GBM) and Support Vector machine (SVM) methods from *scikit-learn* Python package [19]. The predictive performance of the models was assessed using five-fold cross-validation. Q^2 and RMSE were used as statistical measures calclulated according to eq. (1) and (2) respectively. A consensus model (obtained by averaging predictions of individual models which had appropriate Q^2 and RMSE) was used for interpretation in this study as, in common with previous studies [9], individual models were in close agreement and using the consensus model helped compensate biases of these models [9].

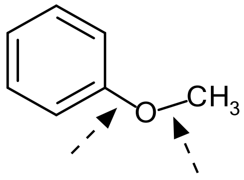
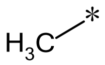
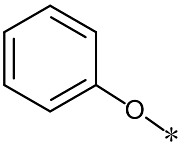
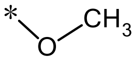
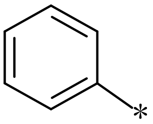
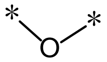
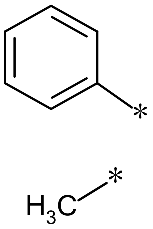
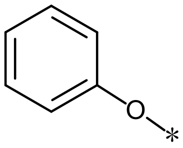
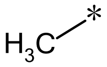
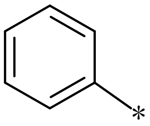
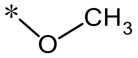
$$Q^2 = 1 - \frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{\sum_i (y_{i,pred} - \bar{y}_{obs})^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{N}} \quad (2)$$

Fragmentation of molecules

For the purpose of model interpretation exhaustive fragmentation was applied to the chemical structures from the data set to generate fragments. Fragments were enumerated by means of RDKit [20] using a SMARTS pattern [!#1]!@!=!#[!#1] matching bonds which can be cleaved during fragmentation. All possible fragments having at most three attachment points were generated from the training set compounds. An example of a molecule with all possible fragments generated is given in the first two columns of Table 1.

Table 1. Example of fragment generation by breaking of bonds matching a SMARTS pattern [!#1]!@!=!#[!#1]. Dashed arrows show the bonds cleaved.

Source molecule	Fragment for which the contribution is to be calculated	Counter-fragment	Fragment kept? (if a counter-fragment has at least 2 heavy atoms)
			Kept
			Kept
			Kept
			Discarded
			Kept

Model interpretation. Step 1: calculation of fragment contributions

In this study the structural and physicochemical interpretation (SPCI) approach for QSAR development [9] was applied; this utilises the concept of matched molecular pairs. The approach can be summarised as follows. For a compound A consisting of two fragments B and C the contribution of fragment C can be calculated as the difference between predicted activity values for the initial compound A and the counter-fragment B (obtained by removal of the fragment C from the molecule A) (**Error! Reference source not found.**). In this way the overall contribution of the fragment C in the units of a studied activity was calculated. The fragment of interest can also be “removed” in terms of a certain type of descriptors (e.g. descriptors encoding partial atomic charge) to calculate the contribution of the fragment from a corresponding physicochemical point of view (**Error! Reference source not found.**). This is *local* interpretation which gives information about the contribution of a fragment in individual compounds. Averaging of contributions of identical fragments allowed for the ranking of different fragments and revealed general trends in structure-activity relationships (*global* interpretation). In this study we extended the global interpretation by means of using GMM (see *„Step 2"* below).

In this study contributions were calculated only for those fragments whose counter-fragments had at least two atoms (Table 1) since only such structures can be properly encoded by the descriptors used.

Step 2: Analysis of fragment contributions using Gaussian mixture models and SMARTSminer

The contributions of fragments were calculated for all molecules where they occurred. If a particular fragment occurred in a molecule multiple times its contributions were calculated separately. Therefore the overall occurrence of the fragment and the number of its contributions calculated is greater than or equal to the number of compounds. Distributions of contributions were analysed for each fragment separately. The distribution can be represented by single or multiple Gaussians (**Error! Reference source not found.**). GMM utilizes the EM-algorithm for finding the optimal parameter values (mean and variance) by maximizing data log-likelihood function for a fixed number of Gaussian components. The number of components was chosen using integrated completed data likelihood criterion. Variance was set to be variable in this study. Cases where the distribution of fragment contributions is represented by multiple Gaussians can be due to the different molecular context of that fragment in different molecules. SMARTSminer was applied to find patterns discriminating compounds corresponding to different Gaussians (clusters). SMARTSminer takes as its input two sets of molecules and searches for discriminative patterns (SMARTS) which appear more often in one set ("positive") than in the other ("negative"). In the case of two clusters we submitted compounds corresponding to the cluster with lower contributions as "negative" and compounds corresponding to the cluster with higher contributions as "positive" and vice versa in order to find patterns discriminating both clusters from each other. In cases where more than two clusters were identified one cluster could be chosen as a "positive" set and the remaining ones could be combined into the "negative" set, or these could be considered separately one by one. Patterns determined were ranked according to the calculated σ-score [13]. Additionally the user can specify desired levels of positive and negative support. In this work minimum positive support was set to 0.7 (at least 70% of molecules in the "positive" set must contain a pattern) and maximum negative support 0.3 (at most 30% of molecules in the "negative" set must contain the same pattern). The top scored patterns outputted by SMARTSminer were analysed to find those that may influence or cause changes in toxicity.

Analysis of distributions of fragment contributions was performed using GMM from the *mclust* R package [21-22]. GMM model building and visualisation steps were implemented in the *rspi* R package to automate the analysis workflow [23].

Results and Discussion

This study utilised models based on the growth inhibition of a large data set to *T. pyriformis.* These data have been the subject of numerous previous QSAR analyses [24] and it is stressed that the purpose here was not the modelling of toxicity *per se* but the use of the proposed algorithm to extract useful and usable information from the data and, more importantly, to validate in this way the approach proposed.

Modeling of acute aquatic toxicological endpoints has a long history and is based on the premise that toxicity is governed by the ability of a chemical to reach the active site (e.g. the cellular membrane) and its ability to interact there [24-25]. In this context transport has usually been quantified by descriptors for hydrophobicity and interaction by descriptors for electrophilicity or more specific interaction such as receptor binding [24-26]. Many QSAR models for acute aquatic toxicity have been developed on a mechanistic basis [24-25, 27] currently more commonly referred to through Adverse Outcome Pathways [28].

However, despite extensive studies and numerous approaches proposed to allocate compounds to mechanisms of action, e.g. the Verhaar scheme [29], that has been recently updated [30], the schemes available are still limited in their applicability. Thus, the mining of databases and datasets of toxicological information to determine relevant structural features is essential. This study has focused in particular on the capability to determine fragments and their molecular contexts associated with excess toxicity, i.e. toxicity caused by reactive species or specific mechanisms of action. Baseline toxicity (non-polar narcosis) was taken into consideration as well, though it is governed by hydrophobicity and no specific patterns are expected to be found.

Performance of the QSAR models to predict toxicity

Whilst not the overall aim of this investigation, it is important to assess the validity of the QSAR models created. The performance of the SVM, RF and GBM models was found to be reasonable, whereas the PLS model had low predictivity overall (Table 2). Therefore, a consensus prediction was developed by averaging of the predictions of the SVM, RF and GBM models. The contributions of fragments calculated from the individual models, as well as the consensus model, was analysed and there was close agreement between all models. In order not to bias the analysis by a specific statistical approach, the consensus model was therefore used for further analysis.

Table 2. Predictive performance of the QSAR models for the growth inhibition of compounds to *Tetrahymena pyriformis*, as estimated by 5-cold cross-validation.

Model	Q^2	RMSE
RF	0.76	0.51
SVM	0.73	0.55
GBM	0.77	0.50
PLS	0.35	0.85
Consensus (RF, SVM, GBM)	0.75	0.52

Analysis of fragment contributions using GMM and SMARTSMiner

All compounds were exhaustively fragmented and fragment contributions were calculated. The resulting number of distinct fragments obtained was 6742 (**Error! Reference source not found.**). 6431 fragments were excluded from consideration due to the lack of occurrence: they appeared less than in 10 compounds of the data set. GMM was applied to analyse contributions of the remaining 311 fragments. For 193 fragments only one Gaussian was detected. For the subsequent analysis 39 of these fragments having low variance of contributions (<= 0.25) were selected. They are relevant to analyse since we assume that contributions of such fragments would not substantially depend on their molecular context, hence, those of them with high average contribution can indicate toxicophore moieties *per se*. For 118 fragments two or more clusters were identified. Those fragments frequently represented the same structural motifs and were highly similar/homologous, e.g. differed by a methylene group, etc. Therefore, we focused our analysis to the most relevant and not overlapping patterns.

Fragments for which one Gaussian (cluster) was identified by GMM

The largest average contributions (around 1.0-2.0) amongst 39 fragments having narrow distributions corresponded to various aromatic fragments (**Error! Reference source not found.**). This can be explained to a large part by their high lipophilicity and hence implicit relationship to non-polar narcosis. However, some of them, such as benzaldehyde derivatives, can be reactive. Several structurally overlapping fragments were found comprising a methacrylate substructure and having close average contributions (around 0.7-0.9). Their toxic effect can be caused by the reactivity of the esters of acrylic acid which may participate as electrophilic substances in Michael addition.

Fragments for which multiple Gaussians (clusters) were identified by GMM

Halogens (single-atom fragments being halogens)

As expected, the halogens all have a positive contribution to toxicity i.e. the inclusion of a halogen atom on a molecule will make it more toxic in comparison to the parent. The median contribution of halogens (in rank order) was: F (0.25) < Cl (0.52) < Br (0.72) < I (0.91). This means that on average inclusion of a particular halogen atom on a molecule increases $pIGC_{50}$ on the specified value. This can be interpreted in two ways. The trend in the halogens is consistent with both hydrophobicity [31] and electrophilicity [32]. Depending on the configuration of the halogen the increase can be related to either. For instance, halogen substitution on an unsubstituted aromatic ring or alkyl chain will increase hydrophobicity and is hence related to non-polar narcosis [33]. A halogen that is adjacent to an activating group, however, will become unspecifically reactive through one of several electrophilic mechanisms of chemical reactions.

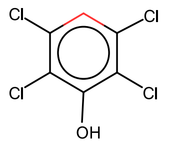
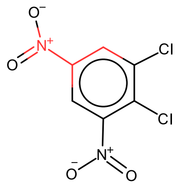
The distributions of the contributions of chlorine, bromine and iodine atoms to toxicity are broadly similar and illustrated in **Error! Reference source not found.**. The distributions had a large peak and a relatively long right tail, the tail being detected by GMM as a distinct cluster. Taking chlorine as an example, the majority of the contributions (96% of data) were in the first cluster with mean value of 0.47. The second cluster, with more significant contributions comprised only 4% of the data and had a mean contribution of 1.16. This finding is consistent with the hypothesis that if non-activated (halogens belonging to the first cluster), the addition of a halogen will have a contribution equivalent to its hydrophobicity [33]. However, an activated halogen (halogens belonging to the second cluster) will have a much greater contribution. For instance, one of the prevalent patterns corresponding to the second cluster was A[CD3H0](CCl)=[OX1-0]. It matches α-chloroketones, esters or amides present in the cluster. In the first cluster the prevalent highest scored patterns were simple aromatic carbon and other SMARTS matching aromatic compounds. Patterns with examples of compounds matching are shown in Table 3.

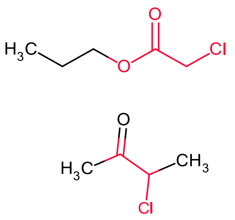
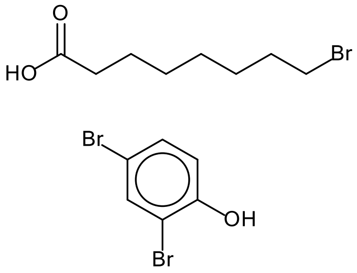
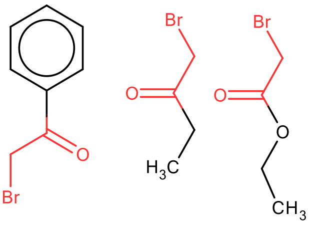
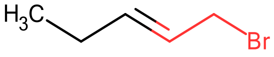
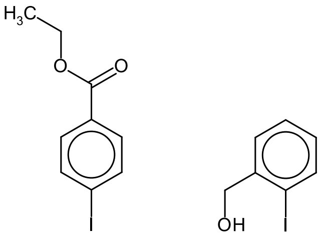
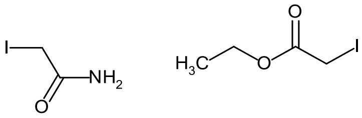
13% of bromine fragments fell into the second cluster (the right tail of the overall distribution) with mean contribution of 1.68 (**Error! Reference source not found.**). The highest scored SMARTS patterns were A[CX4][CX3]=[C,O], Br[CX4][CX3] and several others which matched the *α*-bromoketones and esters, analogously to chlorine, and additionally *α*-bromoalkenes, which are also reactive species (Table 3). No significant and consistent patterns were found for the first cluster which had a mean contribution to toxicity of 0.64.

The findings for bromine and chlorine fragments are in accordance with the understanding that activated halogens (e.g. adjacent to an ester or other unsaturation) are electrophilic in nature and will have a strong influence on toxicity [34]. Specifically, the reactivity of α-haloactivated compounds occurs as a result of their reactivity in Phase II enzymes. It is mediated by a S_N2-type of transition state with the partially negative charged sulfur atom from the thiol groups of glutathione S-transferases. It was noted [32] that the halo-substituted compounds of this type were one of eight classes of S_N2 electrophiles.

There was a small number of compounds containing iodine atoms of which only a few compounds were detected as belonging to the second cluster with high contribution values (mean contribution is 2.72). For this reason, we did not apply SMARTSminer and manually found that these compounds were *α*-iodoketones and esters – as noted above, these are activated halogens and likely to act as S_N2 electrophiles.

Table 3. Examples of SMARTS patterns and molecules corresponding to each cluster detected by GMM for chlorine, bromine and iodine. Colours of Gaussians correspond to **Error! Reference source not found.**. SMARTS patterns matched in structures are coloured in blue.

Halogen	Gaussian				SMARTS found	σ-score	Molecule examples
	Cluster number	Mean	Standard deviation	Coverage, %			
Cl	First (Pink)	0.47	0.24	95	c	0.91	
					A[cH0]:[c,n]	0.91	
	Second (Orange)	1.16	1.02	5	A[CD3H0](CCl)=[OX1-0]	0.91	

					C(Cl)[CD3H0]	0.90	
Br	First (Pink)	0.64	0.24	77	-	-	
	Second (Orange)	1.68	0.79	23	A[CX4][CX3]=[C,O]	0.84	
					Br[CX4][CX3]	0.84	
I	First (Pink)	0.83	0.27	87	-	-	
	Second (Orange)	2.72	0.46	13	-	-	

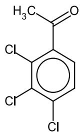
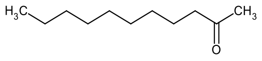
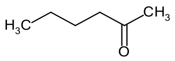
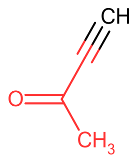
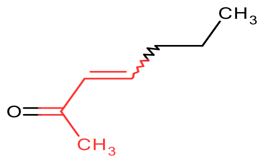
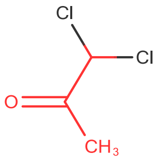
The results for fluorine were different from the rest of halogens (**Error! Reference source not found.**). The GMM algorithm did not identify clusters, however manual inspection of the data found that fluorine atoms had high contributions in molecules where multiple fluorine atoms were attached to benzene or pyridine ring. No other apparent discriminative patterns were found. However, there were a small total number of fluorine-containing compounds and thus any conclusions would have weak support.

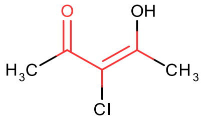
Methylcarbonyl (acetyl)

Whilst the methylcarbonyls were almost symmetrically distributed around zero (from -0.5 to 0.5), the right side tail covered by the second Gaussian contained fragments which demonstrated quite significant contributions to toxicity (**Error! Reference source not found.**). The polar nature of the unactivated fragment is likely to be the cause of its reduction in toxicity due to the reduction in hydrophobicity. However, some patterns e.g. C[CD3H0]([CD1H3])=[OX1-0] and C([CD1H3])[CX3]=[C,O] were found to be discriminative for the second cluster, one with increased toxicity (Table 4). The former matches acetyl itself connected to aliphatic carbon which appears to be not toxicophoric *per se*. The latter is also non-toxicophoric to our knowledge. These patterns thus can be considered artifacts. Visual inspection of compounds from the second cluster revealed that when the carbonyl group of methyl carbonyl moiety is conjugated with a double bond the corresponding compounds are potential Michael acceptors [32, 35-36]. The corresponding manually derived pattern [CD3H0]([CX3]=[CX3])=[OX1-0] was not found by the SMARTSminer with chosen settings due to its low positive support 58% (Table 4).

It should be noted that before modeling the most stable tautomers were generated for all compounds. That converted β-diketones to α,β-unsaturated ketones, see the last example in Table 4 which were identified as belonging to the second (orange) cluster. This shows the importance of choosing of appropriate tautomeric forms of molecules for modeling.

Table 4. Examples of SMARTS patterns and molecules corresponding to each cluster detected by GMM for methylcarbonyl. Colours of Gaussians correspond to **Error! Reference source not found.**. SMARTS patterns matched in structures are coloured in blue.

Gaussian				SMARTS found	σ-score	Molecule examples
Cluster number	Mean	Standard deviation	Coverage, %			
First (Pink)	0.12	0.25	86	-	-	  
Second (Orange)	1.24	0.72	14	C[CD3H0]([CD1H3])=[OX1-0]	0.86	
				C([CD1H3])[CX3]=[C,O]	0.86	
				C[CD3H0]([CX4])=[OX1-0]	0.86	

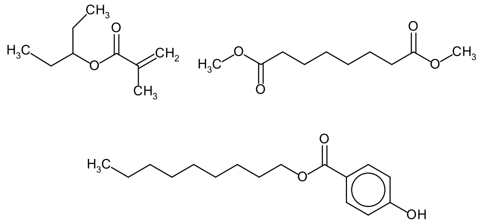
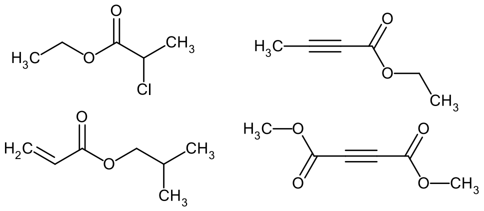
				[CD3H0]([CX3]=[CX3])=[OX1-0] *	0.75	

* the pattern was derived manually

Ester group

Two clusters were found for ester fragments. The cluster with higher contributions (in orange in **Error! Reference source not found.**) corresponds to esters of α,β-unsaturated acids (mainly acrylic and 2-butynoic acid) and α-halogen carboxylic acids that were found by visual inspection (Table 5). These compounds can participate in Michael addition or nucleophilic substitution reactions [32] and hence are associated with excess toxicity. SMARTSminer did not identify them, since each of these two patterns has positive support about 50% which is lower than the chosen threshold (70%) and because the algorithm implemented in SMARTSminer does not use generalised bond patterns, e.g. “double or triple bond”. Ester groups with lower contributions (the pink cluster in **Error! Reference source not found.**) correspond to different esters of aliphatic and aromatic acids and no patterns were found by SMARTSminer. Simple esters are known to act as non-polar narcotics to *T. pyriformis* but are metabolised to reactive toxicants in other species e.g. fish [37].

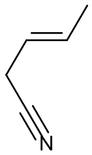
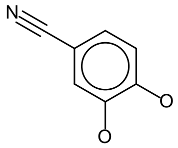
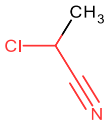
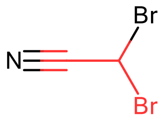
Table 5. Examples of SMARTS patterns and molecules corresponding to each cluster detected by GMM for ester group. Colours of Gaussians correspond to **Error! Reference source not found.**.

Gaussian				SMARTS found	σ-score	Molecules examples
Cluster number	Mean	Standard deviation	Coverage, %			
First (Pink)	0.22	0.22	75	-	-	
Second (Orange)	1.04	0.7	25	-	-	

Cyano group

Two clusters were detected by GMM for cyano group contributions The second cluster, with higher contribution values, was made up predominantly of compounds with a halogen in α-position to the cyano group: C(#N)[CX4][F,Cl,Br,I] (Table 6, **Error! Reference source not found.**). Such polarized cyano compounds can add via Michael addition, causing significant increased toxicity. Surprisingly, many cyano-containing compounds are considered to act by a narcotic-type mechanism (probably polar narcosis) to *T. pyriformis* [38], this is consistent with the observation in the first cluster not being associated to significant increase in toxicity and that no clear patterns were detected.

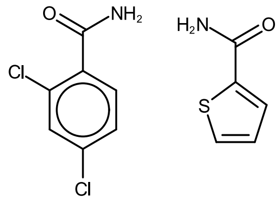
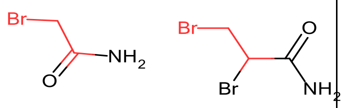
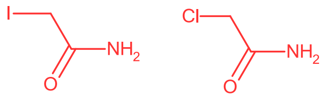
Table 6. Examples of SMARTS patterns and molecules corresponding to each cluster detected by GMM for cyano group. Colours of Gaussians correspond to **Error! Reference source not found.**. SMARTS patterns matched in structures are coloured in blue.

Gaussian				SMARTS found	σ-score	Molecules examples
Cluster number	Mean	Standard deviation	Coverage, %			
First (Pink)	0.22	0.24	90	-	-	 
Second (Orange)	1.04	1.01	10	C(#N)[CX4][F,Cl,Br,I]	0.89	
				A[CX4][F,Cl,Br,I]	0.85	

Carbamoyl group (-C(=O)NH_2)

In most cases, carbamoyl group appeared in aliphatic and aromatic hydrocarbons and appeared to decrease the toxicity of compounds (**Error! Reference source not found.**). This is consistent with the polar and unreactive nature of these fragments, thus their addition to a molecule will reduce hydrophobicity and hence baseline narcotic potency. No patterns were found by SMARTSminer for first cluster. Only four carbamoyl containing compounds had significantly elevated toxicity, which formed the second cluster. Those compounds feature halogens in α-position to the carbamoyl moiety Table 7 and are expected to be electrophilic through nucleophilic substitution reactions [32].

Table 7. Examples of SMARTS patterns and molecules corresponding to each cluster (with the second consisting of four outliers) detected by GMM for carbamoyl. Colours of Gaussians correspond to **Error! Reference source not found.**. SMARTS patterns matched in structures are coloured in blue.

Gaussian				SMARTS found	σ-score	Molecules examples
Cluster number	Mean	Standard deviation	Coverage, %			
First (Pink)	-0.25	0.33	91	-	-	
Second (Orange)	1.87	0.12	9	C[CD2H2][F,Cl,Br,I]	0.98	
				[CD3H0]([CX4][F,Cl, Br,I])([ND1H2])=[OX 1-0]	0.92	

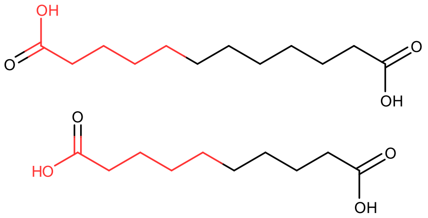
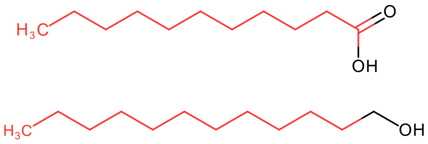
Saturated linear C_8 and C_9 alkylene moieties

Two clusters with large difference between average contribution values were detected by GMM for the longer chain alkyl groups (**Error! Reference source not found.**). According to the SMARTS patterns found, compounds containing aliphatic carboxylic group appeared predominantly in the first cluster. Whereas C([C,O][C,O][CD1H3])[CD2H2][CD2H2][CD2H2][CD2H2][CD2H2][CD2H2][C,O] pattern encoding long linear alkyl chain was found discriminative for the second cluster (Table 8). However, visual inspection showed that dicarboxylic acids were mainly present in the first cluster while the second one is more populated with monocarboxylic acids (Table 8). This subtlety could not be captured by SMARTSminer due to inability to search for disjoint or variable length patterns.

The overriding effect on toxicity is due to the relative hydrophobicity of these fragments. An eight carbon chain will have an intrinsic logarithm of the octanol-water partition coefficient (log P) of about 4 units, increasing considerably the toxicity of the molecule. This will be tempered by the presence of the carboxylic acid groups. Monocarboxylic acids have higher lipophilicity relatively to dicarboxylic acids and this can explain the substantially different contributions of alkylene chains in these compounds [39].

Table 8. Examples of SMARTS patterns and molecules corresponding to each cluster detected by GMM for C_8 and C_9 linear alkylene groups. Colours of Gaussians correspond to **Error! Reference source not found.**. SMARTS patterns matched in structures are coloured in blue.

Fragment	Gaussian				SMARTS found	σ-score	Molecules examples
	Cluster number	Mean	Standard deviation	Coverage, %			

C_8 and C_9	First (Pink)	1.22-1.24	0.37-0.41	10-11	[CD2H2][CD2H2][CD2H2][CD2H2][CD3H0][OX2]	0.84	
C_8 and C_9	Second (Orange)	2.67-2.93	0.18-0.28	89-90	C([C,O][C,O][CD1H3])[CD2H2][CD2H2][CD2H2][CD2H2][CD2H2][CD2H2][C,O]	0.89	

Hydroxyl and carboxyl groups

Only a single cluster was detected by GMM for hydroxyl and carboxyl fragments (**Error! Reference source not found.**). Therefore, SMARTSminer could not be applied. However, the left shoulder on the distribution of carboxyl group contributions was observed and two peaks could be visually detected on the distribution of hydroxyl group contributions as well. We checked whether these observations were related to context-dependence or were artifacts. Fragmentation SMARTS patterns were applied which match explicitly aliphatic and aromatic carboxyl and hydroxyl groups. Distributions of contributions of both these groups in the case of aliphatic and aromatic derivatives were significantly different according to the Kolmogorov-Smirnov two-sided test. Aliphatic hydroxyl groups (e.g. in aliphatic alcohols) have lower contributions to the toxicity in comparison to aromatic OH groups (in phenols). On **Error! Reference source not found.** smoothed densities of aromatic and aliphatic hydroxyl and carboxyl group contributions are shown in orange and pink. A carboxylic group showed lower toxicity in aromatic compounds than in aliphatic ones. This example demonstrates that GMM models cannot always separate contributions of fragments when distributions substantially intersect. Therefore, visual inspection of contribution distributions would be required to detect such cases. Both functional groups will reduce toxicity due to their polar nature and through the reduction of hydrophobicity. However, higher toxicity of aromatic hydroxyl group compared to aliphatic can be explained as follows. Certain combinations of di-hydroxy aromatic compounds (e.g. the in the –ortho or –para configuration) are responsible for increased toxicity via their oxidation to the corresponding quinone which, in turn, is electrophilic [40-41]. Besides that, phenols themselves are also associated with a number of different modes of action including non-polar narcosis and respiratory uncoupling [42].

Physicochemical interpretation of fragment contributions

Since the descriptors used for modelling encoded different physicochemical properties, the contribution of different physicochemical terms to the studied toxicity could be estimated (**Error! Reference source not found.**). The polarisability of halogen atoms, cyano and carbamoyl groups had high contribution (the second clusters on **Error! Reference source not found.**) to toxicity. This is consistent with the reactivity of the patterns detected as described above. Also as discussed above, the major contribution factor to the high toxicity of alkylene chains was their hydrophobicity which is supported by experimental findings [39]. Thus, physicochemical interpretation can provide more detailed knowledge about the contributions of fragments and help shed light on mechanisms of action.

Applying SMARTSminer directly to the whole dataset

SMARTSminer was also applied directly to the whole set of compounds modelled in order to find possible toxicophoric patterns and make a comparison of such a straightforward approach to that applied here. Since the approach is only suitable for working with classification tasks, two subsets of compounds were selected based on thresholds: the “negative” set with 500 compounds having pIGC$_{50}$ <= 2.5 and the “positive” set with 406 compounds having pIGC$_{50}$ >= 5. No patterns were found by running SMARTSminer with the chosen settings for positive and negatives (0.7 and 0.3, respectively). Decreasing positive and negative thresholds to 0.6 and 0.2, respectively, helped retrieve about 100 patterns. They mostly matched aromatic and some heteroaromatic substructures which are abundant in the “positive” set of compounds (**Error! Reference source not found.**) and less frequent in the “negative” set.

Further decreasing support values did not help to any great extent; numerous general patterns matching mainly aromatic substructures were found. The patterns identified by our approach could not be found because all of them had low positive support values (<0.1). Poor performance of SMARTSminer might be explained by the high structural diversity of the compounds in the data set and different, or mixed, mechanisms of toxic action.

Comparison to analysis of fragments of a greater size

In principle, analysis of structural context can be replaced by using fragments of a greater size, which will include both a fragment and its context. This can be done in some particular cases but is not applicable in general. One of main reasons is that structurally identical fragments can have different number and positions of attachment points making them not identically represented (e.g. as SMILES). This results in several variants of one fragment which are less frequent, and if their occurrence will be below the threshold, they would be ignored (see Figure 3). It is not trivial to unite such fragments in order to analyse their contributions together. Ignoring attachment points in fragments will make various different fragments indistinguishable, e.g. methoxy (CO[*]) and hydroxymethyl ([*]CO) groups. Our approach by design is more flexible in this respect, since the context can be expressed as more abstract pattern, compared to “fixed” large fragment.

Conclusions

The results of the study demonstrated that interpretation of QSAR models can retrieve reasonable and rational structural patterns within molecules. Using Gaussian mixture modelling in combination with SMARTSminer allowed for the detection of the influence of the different molecular contexts of the fragments having high contributions to the studied property. The developed approach was applied to study the toxicity of various classes of organic compounds to *Tetrahymena pyriformis*. Patterns indicating different mechanisms of action were identified. For example, halogens were associated with substantially higher contributions to the toxicity when being a part of α-haloketones, esters or amides than in other compounds. In general, the results obtained from structural and physicochemical interpretation of QSAR models in this study were consistent and corresponded to expert knowledge about environment toxicophores and their mechanisms of action. This confirmed the validity of the approach developed. However, the proposed workflow to determine molecular context of important fragments has some limitations. If contributions of fragments in different contexts were numerically similar, GMM could not separate them into clusters to perform further analysis, e.g. as observed for aliphatic and aromatic hydroxyl and carboxyl groups.

Overall, SMARTSminer helped automate the search for the molecular context of fragments. However, due to its limitations, e.g. the absence of generalised bond patterns or disjoint patterns that were not captured, SMARTSminer could not retrieve results in some cases and manual inspection was required to retrieve reasonable patterns. Moreover, our observations suggest that there is difficulty in applying SMARTSminer directly to data sets when the compounds studied have a variety of different patterns or different mechanisms of action. This issue was due to the low supports of the discriminative patterns. Interpretation of QSAR models is more suitable in that case because the only prerequisite is the possibility to build a predictive model. This makes interpretation of QSAR models a more versatile approach to retrieve structure-property relationships from data sets of chemical compounds. The proposed workflow of implicit detection of molecular context can be also used for MMP analysis.

Acknowledgement:

This work was supported by the grant of the Ministry of education, youth and sport of Czech republic (MSMT-5727/2018-2) and the grant of the Ministry of Education and Science of the Russian Federation (Agreement No14.587.21.0049)

[1] P. Polishchuk. *J. Chem. Inf. Model.* **2017,** *57*, 2618-2639.
[2] *OECD Papers* **2006,** *6*, 79-157.
[3] A. G. Leach; H. D. Jones; D. A. Cosgrove; P. W. Kenny; L. Ruston; P. MacFaul; J. M. Wood; N. Colclough; B. Law. *J. Med. Chem.* **2006,** *49*, 6672-82.
[4] S. Riniker; G. A. Landrum. *J. Cheminf.* **2013,** *5*, 43.
[5] P. G. Polishchuk; V. E. Kuz'min; A. G. Artemenko; E. N. Muratov. *Mol. Inf.* **2013,** *32*, 843-853.
[6] Y. Sushko; S. Novotarskyi; R. Korner; J. Vogt; A. Abdelaziz; I. V. Tetko. *J. Cheminf.* **2014,** *6*, 48.
[7] V. E. Kuz'min; A. G. Artemenko; E. N. Muratov. *J. Comput.-Aided Mol. Des.* **2008,** *22*, 403-421.
[8] V. E. Kuz'min; A. G. Artemenko; P. G. Polischuk; E. N. Muratov; A. I. Khromov; A. V. Liahovskiy; S. A. Andronati; S. Y. Makan. *J. Mol. Model.* **2005,** *11*, 457-467.
[9] P. Polishchuk; O. Tinkov; T. Khristova; L. Ognichenko; A. Kosinskaya; A. Varnek; V. Kuz'min. *Chem. Inf. Model.* **2016,** *56*, 1455-1469.
[10] Y. Sushko; S. Novotarskyi; R. Korner; J. Vogt; A. Abdelaziz; I. Tetko. *J. Cheminf.* **2014,** *6*, 48.
[11] Y.-Y. Zhang; H. Liu; S. G. Summerfield; C. N. Luscombe; J. Sahi. Mol. *Pharmaceutics* **2016**.
[12] G. Papadatos; M. Alkarouri; V. J. Gillet; P. Willett; V. Kadirkamanathan; C. N. Luscombe; G. Bravi; N. J. Richmond; S. D. Pickett; J. Hussain; J. M. Pritchard; A. W. J. Cooper; S. J. F. Macdonald *J. Chem. Inf. Model.* **2010,** *50*, 1872-1886.
[13] S. Bietz; K. T. Schomburg; M. Hilbig; M. Rarey. *J. Chem. Inf. Model.***2015,** *55*, 1535-46.
[14] V. Ruusmann; U. Maran *J. Comput.-Aided Mol. Des.* ***2013,*** *27*, 583-603.
[15] *Standardizer 16.9.12.*, 16.9.12.; ChemAxon (http://www.chemaxon.com): 2016.
[16] P. G. Polishchuk *SPCI: Structural and physico-chemical interpretation tool, https://github.com/DrrDom/spci*.
[17] *cxcalc 16.9.12*, 16.9.12; ChemAxon (http://www.chemaxon.com): 2016.
[18] P. G. Polishchuk Simplex representation of molecular structure - a chemoinformatic tool for calculation of simplex descriptors v. 1.1.1. https://github.com/DrrDom/sirms. https://github.com/DrrDom/sirms
[19] *Scikit-learn 0.18*, 0.18; Pedregosa et al.: 2016.
[20] *RDKit, Open-Source Cheminformatics 2017.09.1 http://www.rdkit.org*.
[21] C. Fraley; A. E. Raftery. *J. Am. Stat. Assoc.* **2002,** *97*, 611-631.
[22] T. B. M. Adrian E. Raftery, and Luca Scrucca *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597*; 597; University of Washington: 2012.
[23] P. G. Polishchuk *Analysis of fragments contributions calculated by SPCI software. https://github.com/DrrDom/rspci.*
[24] M. T. Cronin. *Environ. Sci. Process. Impacts* **2017,** *19*, 213-220.
[25] T. W. Schultz; M. T. Cronin. *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 304-9.
[26] M. T. Cronin; J. C. Dearden; J. C. Duffy; R. Edwards; N. Manga; A. P. Worth; A. D. Worgan. *SAR QSAR Environ. Res.* **2002,** *13*, 167-76.
[27] M. T. D. Cronin. Quantitative Structure–Activity Relationships (QSARs) – Applications and Methodology. In *Recent Advances in QSAR Studies,* T. Puzyn, J. L., M. Cronin, Ed. Springer: London, 2010; pp 3-11.
[28] C. M. Ellison; P. Piechota; J. C. Madden; S. J. Enoch; M. T. Cronin. *Environ. Sci. Technol.* **2016,** *50*, 3995-4007.
[29] H. J. M. Verhaar, van Leeuwen, C.J. Hermens, J.L.M. . *Chemosphere* **1992,** *25*, 471-491.
[30] C. M. Ellison; J. C. Madden; M. T. Cronin; S. J. Enoch. *Chemosphere* **2015,** *139*, 146-54.
[31] T. Fujita, Isawa J.., Hansch C. *J. Am. Chem. Soc.* **1964,** *86*, 5175-5180.
[32] S. J. Enoch; C. M. Ellison; T. W. Schultz; M. T. Cronin. *Crit. Rev. Toxicol.* **2011,** *41*, 783-802.

[33] C. M. Ellison; M. T. Cronin; J. C. Madden; T. W. Schultz. *SAR QSAR Environ. Res.* **2008,** *19*, 751-83.

[34] T. W. Schultz; M. T. Cronin; T. I. Netzeva; A. O. Aptula. *Chem. Res. Toxicol.* **2002,** *15*, 1602-9.

[35] Y. K. Koleva; J. C. Madden; M. T. Cronin. *Chem. Res. Toxicol.* **2008,** *21*, 2300-12.

[36] T. W. Schultz; T. I. Netzeva; D. W. Roberts; M. T. Cronin. *Chem. Res. Toxicol.* **2005,** *18*, 330-41.

[37] J. S. Jaworska; R. S. Hunter; T. W. Schultz. *Arch. Environ. Contam. Toxicol.* **1995,** *29*, 86-93.

[38] M. T. Cronin; S. E. Bryant; J. C. Dearden; T. W. Schultz. *SAR QSAR Environ. Res.* **1995,** *3*, 1-13.

[39] R. Jaffe. *Environ. Pollut.* **1991,** *69*, 237-57.

[40] A. O. Aptula; D. W. Roberts; M. T. Cronin; T. W. Schultz. *Chem. Res. Toxicol.* **2005,** *18*, 844-54.

[41] F. Bajot; M. T. Cronin; D. W. Roberts; T. W. Schultz. *SAR QSAR Environ. Res.* **2011,** *22*, 51-65.

[42] S. J. Enoch; M. T. Cronin; T. W. Schultz; J. C. Madden. *Chemosphere* **2008,** *71*, 1225-32.

Figure 1. Workflow for the analysis of the context-dependence of fragment contributions. A: a narrow range of values indicates either “stable” contribution of a fragment regardless of molecular environment it appears in or a very similar context of a fragment in all compounds of a studied data set. B, C: a broader range suggests context-dependence of fragment influence on a modelled property. Distributions can be analysed with GMM to detect clusters (in the case C they will appear). Clusters can have distinct molecular contexts indicating important patterns, e.g. “fragment + context = toxicophore” (combination of fragment and context in which the fragment has high contributions). Search for patterns was performed using SMARTSminer or manually.

Figure 2. Schemes for the structural and physicochemical interpretation of QSAR models. In the case of physicochemical interpretation the contribution of H-bonding is calculated for the fragment C. Subscript E, H, D and HB denote descriptors representing electrostatic, hydrophobic, dispersive interactions and H-bonding, respectively.

Figure 3. Decision tree illustrating the workflow for the analysis of fragments. Green boxes contain fragments to be analysed. The upper green box contains the fragments of main interest to this study since clusters were found in their distributions. The lower green box contains fragments having narrow distributions with no clusters (sd <= 0.25, sd – standard deviation).

Figure 4. Distributions of contributions of fragments having the highest average contributions to the toxicity and narrow distributions of values.

Figure 5. Distributions of contributions of halogens (Cl, Br, I) with regard to their toxicity to T. pyriformis. Histograms and dashed lines represent actual fragment distribution. Coloured lines represent Gaussians detected by GMM.

Figure 6. The distribution of fluorine atom contributions. Coloured line represents Gaussian detected by GMM.

Figure 7. The distribution of methylcarbonyl fragment contributions. Coloured lines represent Gaussians detected by GMM.

Figure 8. The distribution of contributions of ester group. Coloured lines represent Gaussians detected by GMM.

Figure 9. The distribution of contributions of cyano group. Coloured lines represent Gaussians detected by GMM.

Figure 10. The distribution of contributions of carbamoyl group. Coloured lines represent Gaussians detected by GMM.

Figure 11. Distributions of contributions of C8 and C9 linear alkylene groups. Coloured lines represent Gaussians detected by GMM.

Figure 12. Distributions of the contributions of carboxyl (right) and hydroxyl (left) groups with smoothed densities (black dashed line) and subpopulations of fragments in aliphatic and aromatic context matched explicitly (solid colored lines).

Figure 13. Median physicochemical contributions of fragments to their toxicity to Tetrahymena pyriformis (M denotes the number of compounds having a particular fragment and N – the overall fragment occurrence).

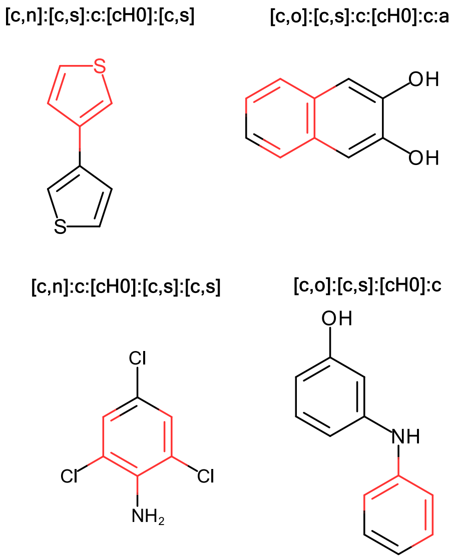


Figure 14. Top-ranked discriminative patterns found by SMARTSminer to discriminate high from low toxicity compounds and examples of matched compounds from the „positive“ set.