HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2019 November 26.

Published in final edited form as:

J Chem Inf Model. 2018 November 26; 58(11): 2203–2213. doi:10.1021/acs.jcim.8b00450.

Chemistry-Wide Association Studies (CWAS): A Novel Framework for Identifying and Interpreting Structure-Activity Relationships

Yen S. Low[a,#], **Vinicius M. Alves**[a,b,#], **Denis Fourches**[c], **Alexander Sedykh**[d], **Carolina Horta Andrade**[b], **Eugene N. Muratov**[a,e], **Ivan Rusyn**[f], **Alexander Tropsha**[a,*]

[a]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.

[b]Laboratory for Molecular Modeling and Design, Department of Pharmacy, Federal University of Goias, Goiania, GO, 74605-170, Brazil.

[c]Department of Chemistry and Bioinformatics Research Center, North Carolina State University, Raleigh, NC, 27695, USA

[d]Sciome LLC, Research Triangle Park, NC 27709, USA

[e]Department of Chemical Technology, Odessa National Polytechnic University, Odessa, 65000, Ukraine.

[f]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, 77843, USA

Abstract

Quantitative Structure-Activity Relationships (QSAR) models are often seen as a "black box" because they are considered difficult to interpret. Meanwhile, qualitative approaches, *e.g.*, structural alerts (SA) or read-across, provide mechanistic insight, which is preferred for regulatory purposes, but predictive accuracy of such approaches is often low. Herein, we introduce the Chemistry-Wide Association Study (CWAS) approach, a novel framework that both addresses such deficiencies and combines advantages of statistical QSAR and alert-based approaches. The CWAS framework consists of the following steps: *(i)* QSAR model building for an endpoint of interest; *(ii)* identification of key chemical features; *(iii)* determination of communities of such features disproportionately co-occurring more frequently in the active than in the inactive class; and *(iv)* assembling these communities to form larger (and not necessarily chemically connected) novel structural alerts with high specificity. As a proof-of-concept, we have applied CWAS to model Ames mutagenicity and Stevens-Johnson Syndrome (SJS). For the well-studied Ames mutagenicity dataset, we have identified 76 important individual fragments and assembled co-occurring fragments into SA both replicative of known as well as representing novel mutagenicity

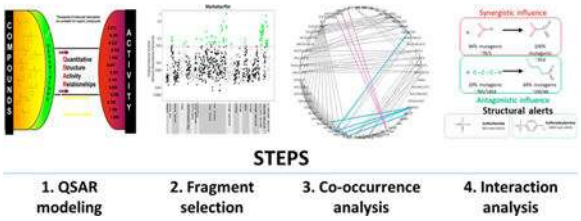
[*]**Corresponding Author:** Address for correspondence: 100K Beard Hall, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-2955; FAX: (919) 966-0204; alex_tropsha@unc.edu.
[#]These authors contributed equally

Associated content

Supporting information including curated datasets and calculated fragment descriptors used in this study is available via the Internet at http://pubs.acs.org.

alerts. For the SJS dataset, we identified 29 important fragments and assembled co-occurring communities into SA including both known and novel alerts. In summary, we demonstrate that CWAS provides a new framework to interpret predictive QSAR models and derive refined structural alerts for more effective design and safety assessment of drugs and drug candidates.

Graphical Abstract



Introduction

Quantitative structure-activity relationship (QSAR) modeling is a major computational approach used in drug discovery and chemical risk assessment.[1] Despite the broad application of QSAR models with well-defined metrics such as accuracy and external predictive power, they bear the stigma of being non-interpretable and therefore, for the most part, unsuitable for regulatory applications.[2] Alternatively, approaches such as structural alerts (SA),[3] chemical grouping,[4] and read-across (RA)[5] have earned regulatory acceptance due to their simplicity, transparency, and ease of interpretation[6] in spite of being often criticized for the lack of predictive accuracy.[7] For instance, we recently showed that the mere presence of structural alerts in a chemical is unreliable for predicting toxicants or PAINS and such exclusive use should be avoided.[8,9] However, we also suggested that when validated by predictive QSAR models, structural alerts may play a key role in the mechanistic understanding of the chemical activity and may help in designing novel compounds with the desired activity or those with lower toxicity.[8]

It has been shown that QSAR models can indeed offer mechanistic interpretation.[10] This realization led to the suggestions that QSAR and read-across approaches can be integrated to afford both statistical accuracy and interpretability of chemical toxicity prediction models.[11–14] For instance, we proposed a new framework that synergistically integrates structural alerts and rigorously validated QSAR models for a more transparent and accurate prediction of new chemicals.[8]

Herein, we introduce the Chemistry-Wide Association Study (CWAS), a novel framework to systematically assess and characterize the contribution of both individual chemical moieties and their combinations. Obviously, CWAS was inspired by GWAS (Genome-Wide Association Study), a well-established approach for simultaneously identifying and studying large numbers of genetic features potentially associated with a given phenotype (*e.g.,* disease).[15] Indeed, there is a striking similarity between approaches to describing correlations between genetic features and a phenotype (such as disease) and those between chemical features of compounds and their biological activity (Table 1).

By comparing the genotypes of patients with those of healthy individuals (control group), GWAS identifies the genetic determinants (single nucleotide polymorphisms, SNPs, or loci) associated with a phenotype on a genome-wide scale, while controlling for multiple hypotheses. The GWAS-like methodology has been implemented in several fields, leading to the dissemination of other subject-specific wide association studies (WAS), such as Environment-WAS (EWAS),[16,17] Phenome-WAS (PheWAS),[18] and Metabolome-WAS.[19] The same framework naturally lends itself to an implementation for QSAR analysis and interpretation. Two key steps in GWAS – to identify (i) SNPs statistically associated with the phenotype to serve as biomarkers and (ii) alleles that disproportionately co-occur in a phenotype using linkage disequilibrium – perfectly match the standard goals of QSAR, *i.e.*, to (i) build a predictive model of specific bioactivity; (ii) identify chemical features associated with such activity; and (iii) derive structurally meaningful alerts composed of disproportionately frequently co-occurring chemical features.

To the best of our knowledge, this popular GWAS-like concept has not been applied to problems in cheminformatics thus far. The goal of our study was to develop the CWAS approach as a novel framework for identifying structural moieties associated with a target property based on statistically validated QSAR models. Correct identification of such key structural fragments leads to improved interpretation of QSAR models. To demonstrate the utility of the proposed CWAS approach, we have applied it to both Ames mutagenicity and Stevens-Johnson Syndrome (SJS) datasets. Our objectives were to identify statistically significant SA (both known and new) and demonstrate how the combined effect of co-occurring substructures could be used to form novel SA with high specificity. Identification of such statistically significant SA could improve mechanistic understanding of structure-activity relationships and enable more effective design and safety assessment of industrial chemicals and drugs.

Methods

Datasets

Ames mutagenicity—The Ames mutagenicity dataset containing 5,864 compounds was retrieved from the CASE Ultra software (http://www.multicase.com/case-ultra). These data have been originally compiled from multiple sources.[20–22] After curation (see Data Curation Section), 5,439 compounds (2,121 actives and 3,318 inactives) were retained for this study.

Stevens-Johnson Syndrome (SJS)—For the purposes of this study, we employed a dataset studied in our recent work.[23] Briefly, the original SJS dataset originally consisted of 436 drugs extracted from VigiBase, the largest database of adverse drug reactions reports maintained by the World Health Organization Uppsala Monitoring Centre.[24] After curation (see Data Curation Section), 365 compounds (194 active and 170 inactive) remained.

Data curation

We have curated both datasets following the protocols we have developed earlier.[25–27] These protocols include structural normalization of specific chemotypes, such as aromatic and nitro groups; removal of inorganic salts, organometallic compounds, mixtures and large molecules

(MW > 2,000 DA); etc. Duplicates were identified, analyzed, and, if necessary, removed using ISIDA Duplicates[28] and HiT QSAR[29].

Generation of Substructural Molecular Fragments

The ISIDA Fragmentor software (freely available at http://infochim.u-strasbg.fr) was used to calculate 2D fragment descriptors.[30] Briefly, each molecular structure is split into subgraphs of two types: *sequences* and *augmented atoms*. *Sequences* represent linear sequences of atoms only (A), bonds only (B) and/or both atoms and bonds (AB). Only shortest paths from one atom to the other are used. We computed sequences between 2 to 8 atoms long. The second type (*augmented atom*) represents a selected atom with its immediate environment including atoms only (A), bonds only (B), and/or both atoms and bonds (AB). Atomic hybridization (Hy) is considered for augmented atoms with the atom (A) subtype. Fragments with low-variance (standard deviation < 10^{-6}) or highly correlated with each other (r^2 > 0.99) were removed. Thereafter, 967 and 1,091 fragments remained for modeling in Ames mutagenicity and SJS datasets respectively.

Chemistry-Wide Association Studies

The general methodological framework of CWAS (Table 2) consists of the following four steps: *(i)* QSAR modeling, *(ii)* fragment selection, *(iii)* co-occurrence analysis, and *(iv)* interaction analysis. In the first step, we build QSAR models according to the best practices for model development and validation.[31] Next, fragment selection (or feature selection) identifies the minimum subset of chemical fragments associated with chemical bioactivity (*i.e.*, chemical phenotype); this step also dramatically reduces the number of fragments for subsequent analysis. Then, co-occurrence analysis, which is one of the unique benefits of CWAS as compared to the conventional QSAR analysis, evaluates which combination of fragments co-occur in the active compounds more frequently than in the inactive ones. At last, such co-occurring fragments can be assembled into a larger connected or disconnected substructure, which, as a whole, may be more indicative of chemical activity than the sum of its parts.

Step 1. QSAR modeling—The QSAR modeling workflow used in this study includes three following major steps[31,32]: *(i)* data curation/preparation/analysis (selection of compounds and descriptors); *(ii)* model building; and *(iii)* model validation/selection. Here, we followed a *5*-fold external cross-validation procedure. The full set of compounds is randomly divided into five subsets of equal size of which one (20%) is set aside as a test set while the rest (80%) form the modeling set. This procedure is repeated five times allowing each of the five subsets to be used as a test set once. Each modeling set is divided into many internal training and validation sets; then models are built using compounds of each training set and applied to test set compounds to assess their properties.[1,31,33,34] It is important to emphasize that the test set compounds are never employed either to build or optimize the models.

In this study, we used the random forest (RF) algorithm[35]. We chose RF as a statistical modeling approach for its computational efficiency as well as because it enables straightforward mechanistic interpretation of the models in terms of relative variable

importance, which can be easily derived from randomized tree ensembles (cf. Step 2). Each forest was obtained as an ensemble of 500 trees (models), built from 500 bagged, *i.e.*, taken with replacement, subsamples with balanced stratified sampling such that the activity classes are balanced in each subsample used to construct a tree. All other settings were as default in the randomforest package for R. RF also reports an error estimated for out-of-bag set compounds, which is somewhat similar to the error of prediction estimated on the test set.[35]

Models were assessed by sensitivity, specificity, balanced accuracy (average of sensitivity and specificity) and area under curve (AUC) on the test set. Standard errors of all metrics were calculated by bootstrapping with 1000 trials. Additionally, errors for the out-of-modeling-bag sample (reported as OOB errors by RF) were used to determine a minimum subset of features.[35] Since the accuracy of each model is estimated for compounds in the test sets only, which were never used to derive, bias, or select models, this protocol ensures an objective estimation of the true external predictivity of the models.

Step 2. Fragment selection—The process of identifying significant fragments associated with the chemical phenotype is similar to that used in GWAS,[36] except that the compound's "genotype" is denoted by chemical fragments. In this study, the significance of the association was determined according to the variable importance as given by the RF models.[37]

To determine the minimal subset of the f most important chemical fragments, we ranked them according to their associated RF conditional importance.[38] Because the fragment ranking varied slightly across the five models generated from the 5-fold external cross-validation protocol, only the f fragments that were always present among the top 100 fragments in all five models were selected for rebuilding the reduced RF models. For each value of f, the reduced RF model's OOB_f error[35] was compared with that of the full RF model (OOB_{full}) incorporating all fragments. Optimal f (*i.e.*, $f_{\min}$) was defined as f with the minimum OOB_f error less than or equal to OOB_{full} error.

Step 3. Co-occurrence analysis—While the individual fragments identified from Step 2 demonstrate an association with activity and each may be a reasonable alert for the activity class, we posit that combinations of fragments may make better structural alerts. To identify such combinations, we used co-occurrence analysis to determine which pairs of fragments would occur in toxic compounds more frequently than in non-toxic compounds and then employed community detection to expand from pairs of fragments to their clusters.

For each possible pair of fragments i and j, we computed a two-tailed Fisher's exact test of association $t_{i,j}$. A pair of fragments was considered to co-occur more frequently than expected when its p-value was less than 0.1 after adjusting for multiple testing by permutation. To adjust the test value for the pairwise co-occurrence of fragments i and j, $t_{i,j}$ was compared against the null distributions, D_i and D_j, such that the larger of its quantiles along the null distributions, $\max(q_i, q_j)$, was taken as the adjusted p-value. The null distribution D_i is generated from the Fisher's test statistics of fragment i's pairwise co-occurrence with 1,000 noise fragments (randomly present or absent), $t_{i,\mathrm{noise1}}$, $t_{i,\mathrm{noise2}}$, …, $t_{i,\mathrm{noise1000}}$. This was repeated with fragment j to generate the other D_j distribution.

Co-occurring fragments were represented by an adjacency matrix of the respective co-occurrence *p*-values. Fragments that co-occurred with significant frequency (*p*-value < 0.1) were connected to form a network. Within this network, we looked out for densely connected subnetworks known as communities. These communities represent fragments that occurred as a cluster in compounds of one activity class disproportionately more frequently than in compounds of the other class. Such communities of fragments could serve as a collective indicator for an activity class. Further, these fragments could be spatially assembled into larger substructures (not all the fragments need to be chemically connected) as activity class indicators or structural alerts. The communities of fragments were detected by the walktrap algorithm proposed by Pons and Latapy[39]. Briefly, the algorithm agglomeratively merges nodes (individual fragments) into increasingly larger communities such that the probability of intracommunity connections is maximized than if they were connected at random. In the walktrap algorithm used here, the distance metric between two nodes is the probability distribution of random walks between the 2 nodes through 3 nodes instead of the shortest path. The intuition behind the random walks is that nodes frequently encountered in the indirect paths are likely to be near and thus, should belong to the same community.

For comparison, structural alerts obtained by the co-occurrence analysis were benchmarked against those derived from the maximum common substructure (MCS)[40] analysis, an established approach for determining structural alerts. MCS analysis extracted the largest substructures that were more frequently associated with the active compounds than with the inactive compounds using the following parameters: *(i)* size ≥8 atoms, *(ii)* frequency ratio ≥ 2, and *(iii)* present in ≥6 compounds. Ideally, co-occurrence analysis should reveal SA similar to those from MCS or include non-contiguous SA that MCS is not designed to uncover.

Step 4. Interaction analysis—Pairs of fragments with significant statistical interaction effects were identified. All possible pairwise interactions were inserted into the following lasso logistic regression equation (Eq. 1).

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \mu + \sum_{j=1}^{m} \beta_j x_{ij} \qquad \text{(Equation 1)}$$

Where $p_i = P(Y_i = 1)$ is the probability of activity in the i^{th} chemical compound, x_{ij} is the value of the j^{th} fragment for the i^{th} chemical, μ is the intercept, and $\beta = (\beta_j, \ldots, \beta_m)$ are the effects of the m fragments.

Important interaction effects were determined using LLARRMA,[41] a LASSO-based resample averaging model. Briefly, a random subsample was drawn without replacement and fitted to a lasso logistic regression. This process was repeated 100 times to obtain an ensemble lasso logistic regression. LLARRMA indicated the significance of each feature by a resample model inclusion probability (RMIP) metric,[41] which measured the proportion of times a feature is selected into the lasso logistic regression. Only interaction effects with RMIP > 0.85 were considered as significant interactions. Of the significant interaction

effects, those with positive beta coefficients were considered *synergistic* as their combination positively contributed to the chemical activity while those with negative coefficients were deemed *antagonistic* as their combination had negative contribution to the chemical bioactivity.

Results

As a proof-of-concept, we performed two case studies of CWAS application to previously studied Ames mutagenicity[42,43] and SJS[23] datasets (see Methods). For Step 1 (QSAR modeling), model accuracy and other evaluation metrics are shown in Table 3. For Step 2 (fragment selection), the 76 fragments associated with Ames mutagenicity with high mean variable importance scores are highlighted in green in the Manhattan plot (Figure 1). For Step 3 (co-occurrence analysis), fragments co-occurring disproportionately more frequently in one activity class than the other are described in Figure 2 and Figure 3 (Ames dataset) and Figure 4 (SJS dataset). For Step 4 (interaction analysis), the networks showing how the selected fragments statistically interact with one another to elicit a combined effect on mutagenicity are presented in Figure 5 for Ames and Figure 6 for SJS.

Step 1. QSAR modeling

QSAR models for both Ames mutagenicity and SJS were developed using RF on ISIDA substructural fragments. Models validated using 5-fold external cross validation successfully passed the Y-randomization test[44], and hence, they were unlikely to be obtained by chance. The QSAR models for Ames mutagenicity using the full set of 967 fragments afforded balanced accuracy of 85% and AUC of 91% (Table 3), which is comparable with experimental reproducibility of the Ames test.[20] Although the Ames dataset is indeed well-studied, honest comparison could be made only using exactly the same test sets. Among these, two studies Sushko et al.[42] and Alves et al.[43] have used the same dataset, except that a subset of Ames dataset was used as a single test set in these two studies, whilst in our current study we have used 5-fold external cross validation procedure[31]. Thus, although direct comparison will not reflect the real picture, CWAS demonstrated similar or slightly better performance as compared to other studies[42,43]. For the full coverage of Ames dataset, balanced accuracies were reported as high as 50–90% in the study Sushko et al.[42] and 79% in the study of Alves et al.[43]. In our study, full and reduced (in terms of descriptors) models presented balanced accuracies of 85% and 87%, respectively. The full model of SJS using all 1,091 fragments presented a balanced accuracy of 71% and AUC of 77%.

Step 2. Fragment selection

We progressively rebuilt RF models using the f most important fragments yielding the model with OOB error less than or equal to that developed with all the fragments. The optimal number of important fragments to achieve this goal was f≒76 for the Ames dataset. The prediction performances of the reduced model with 76 fragments and the full models with all 967 fragments were similar (Table 3). The corresponding Manhattan plot (Figure 1) shows that the 76 selected fragments (in green) have high mean variable importance score in the full RF model. These selected chemical fragments represent various chemical functional groups such as amine, sulfide, phenol, etc. Similarly, for the SJS dataset we have identified

29 important fragments used for building a reduced model with comparable OOB error as the full model containing 1,091 fragments (Table 3).

Step 3. Co-occurrence analysis

The co-occurrence analysis for the Ames mutagenicity dataset revealed which of the 76 fragments frequently occurred together in the active class. These fragments could be fused to generate larger meaningful substructures of potential value as structural alerts of mutagenicity.

The heatmap (Figure 2A) shows the presence of 76 fragments in the active and inactive classes. The triangle map (Figure 2B) shows the pairwise co-occurrence significance determined by Fisher's exact test: low adjusted *p*-values (< 0.1) were shaded for significance while insignificant values were unshaded. This triangle map formed the basis of an adjacency matrix for constructing a network graph (Figure 2C): the 76 fragment nodes were connected if they co-occurred with significant frequency (*p*-value < 0.1). From this network graph, the walktrap community detection algorithm identified seven distinct subnetworks (or communities) of frequently co-occurring fragments (represented by different colors). For example, one of the smallest communities (purple, sulfonate) is made up of two fragments O-S=O and O(-C'-S') both of which are more likely to be present among mutagenic compounds than non-mutagens in the heatmap. Note that "*" denotes aromatic bonds, "-" denotes single bonds, and "=" denotes double bonds. Another example is the aromatic nitro (blue community) constructed from five co-occurring fragments almost exclusively present among mutagens.

As expected, the precision of structural alerts indicating mutagenicity increased with the number of co-occurring fragments used (Figure 3). The number of co-occurring fragments required to achieve high precision depends on the size of the fragment, its strength of association with mutagenicity and the extent of overlap among the co-occurring fragments. For example, the aromatic nitro SA (blue community), took only three out of five fragments to achieve the same precision as the polyaromatic hydrocarbon (PAH) SA (pink community) made up of over 10 fragments. This is because the two chemical classes have different modes of action involving chemical functional groups of very different sizes.[45] Aromatic nitro compounds (upon metabolic activation) easily form reactive adducts and can be readily identified from at least an aryl fragment (e.g., C*C-N) and a nitro fragment (e.g., C*C-N=O). In contrast, PAH are mutagenic through intercalation, slipping between DNA bases and distorting the DNA structure. They are much larger, comprising of multiple adjacent benzenes, and require more fragments for identification as a PAH SA, especially when its constituent aromatic fragments are highly ubiquitous and non-specific (e.g., C*C, C*(C'*C')*C).

For the SJS dataset, the triangle map shows the pairwise co-occurrence between the 29 important fragments (Figure 4A). Significantly co-occurring pairs of fragments (*i.e.*, p-value < 0.1) were connected in the network graph (Figure 4B). Community detection partitioned the network graph into five densely connected subnetworks representing communities (C1-C5, distinctly colored) of frequently co-occurring fragments (Figure 4B). To illustrate, Figure 4C shows two SA reconstructed from the fragments in communities C1 (pink,

sulfonarylamine) and C2 (blue, beta lactam with adjacent sulfur). Indeed, as shown in the heatmap (Figure 4C), the fragments creating these 2 SA were simultaneously present among SJS inducers much more frequently than among non-inducers. The densely populated upper left corner (Figure 4C) suggests that SJS inducers often contain sulfones (O=S=O), sulfonylarenes (C*C*C*C-S=O), and thia-aza groups (S-C-N), while non-inducers (bottom right of the heat map) often contain arenes with aliphatic chains (H-C-C-C*C*C) and secondary amines (C-C-N-C-C). The two SA discussed, sulfonylarylamine and beta-lactam, were consistent with those obtained from MCS analysis.

Step 4. Interaction analysis

Statistically significant co-occurring fragments that affect mutagenicity are shown as an interaction network (Figure 5A). The fragment O=N=O is linked to mutagenicity in 84% of the cases (620 out of 738 mutagens); it forms a synergistic interaction with C(*C'-N'*O'), but also an antagonistic interaction with C-C-C-H. The synergistic interaction confers an additional mutagenic effect beyond the sum of its component fragments (O=N=O and C(*C'-N'*O')). In contrast, the antagonistic interaction reduces mutagenicity effect beyond the sum of its component fragments (O=N=O and C-C-C-H). Figure 5B illustrates how the synergistically interacting pair of fragments O=N=O and C(*C'-N'*O') are more prevalent (100%) among mutagens and, conversely, how the antagonistically interacting pair of fragments are much less prevalent (69%) in mutagens.

Four following structural alerts (see Figure 6) derived from smaller co-occurring fragments were identified for SJS: *(i)* sulfonylarylamine; *(ii)* β-lactam ring with adjacent sulfur; *(iii)* fluorquinolones; and *(iv)* tetracyclines. Sulfonylarylamine is reconstituted from five co-occurring fragments and is the most common structural alert present among SJS-active drugs.

Discussion

This study demonstrates that GWAS-like approaches can be successfully adapted for cheminformatics both to build statistically significant models of bioactivity and identify key substructures associated with a chemical bioactivity. Herein, we show that such analysis can elucidate and analyze chemical substructures that jointly influence biological activity (or toxicity) of compounds, in which these substructures occur.

As a proof of concept, we have developed robust and predictive QSAR models with the entire set of fragments for Ames mutagenicity (CCR=85%, AUC = 91%) and SJS (CCR=71%, AUC = 77%) data sets. Following the model building step, the application of CWAS identified a minimal number of key fragments required for obtaining reliable QSAR models with slightly improved characteristics for Ames mutagenicity (76 fragments, CCR=87%, AUC = 94%) and SJS (29 fragments, CCR=74%, AUC = 81%). Then, fragment co-occurrence analysis determined the communities of fragments disproportionately frequent in the active class such that they could be fused together to form more meaningful and reliable structural alerts compared to individual constitutional fragments. These alerts were comparable to those derived from the MCS analysis, lending support to the notion that

co-occurrence analysis is a valid method for determining structural alerts. Finally, we characterized pairs of fragments that statistically interact to potentiate or negate activity.

Four steps described above are combined into the CWAS workflow, which can be employed for the analysis of any chemical bioactivity dataset both for predicting chemical bioactivity and interpretation and understanding of structure-activity relationships.

Application of CWAS to identify novel toxicity alerts

A few structural alerts known to be associated with mutagenicity and SJS were in concordance with CWAS results. For instance, nitro groups in aromatic rings and sulfonate esters[45] alerts were validated for Ames mutagenicity. As observed in the results (Figure 5), the fragment O=N=O statistically interacts with C(*C'-N'*O') synergistically and with C-C-C-H antagonistically. The chemical rationale behind the synergistic interaction (*i.e.* more mutagenic than individual fragments) is that the combined aromatic nitro is more readily activated into a resonance stabilized nitro anion radical which in turn form genotoxic metabolites. [47] If the nitro group were instead paired with an aliphatic chain C-C-C-H into an antagonistic pair, their combined mutagenic effect is reduced as the aliphatic fragment does not provide resonance stability like the arene fragment. Thus, metabolic activation of an aliphatic nitro into a mutagenic form is less likely than an aryl nitro.

In another example, synergistic interactions between O-S=O and S-O-C fragments potentiated mutagenicity because taken together, they form a sulfonate ester, a potential alkylating agent related to mutagenicity.[46] Conversely, the antagonistic interaction between O-S=O and O-H reduces mutagenicity because the resultant sulfonic acid O=S-O-H can be readily cleared given that the addition of sulfo group is an endogenous mechanism.[47] The above pairwise interactions provided plausible explanations for mutagenicity variation consistent with already established mechanisms, which serve as a validation of our methodology.

The co-occurrence analysis on SJS dataset identified larger, *i.e.*, more specific substructures, that were expected to yield true positives (*i.e.*, higher precision) compared to known expert-based structural alerts inferred from epidemiological studies of drug classes associated with SJS (Figure 6).[48,49] For example, in the present dataset, all compounds containing a sulfonylarylamine were associated with SJS. However, another structural alert, sulfonamide, was found in both actives and inactives. Consequently, blind use of only sulfonamide as an alert predicting SJS could wrongly discard plenty of widely used sulfonamide-containing antibacterials (*e.g.*, sulfafurazole, sulfadiazine, sulfadimethoxine, etc.). As we have recently shown, structural alerts should be used carefully: usually the alerts fail to work as an actual predictor, but, after proper statistical validation and, if needed, refinement, they could be used as a mechanistic hypothesis.[8]

Indeed, the additional chemical structures condensed in the larger structural alerts identified by CWAS offered important mechanistic clues. For instance, the sulfonamide-containing antibacterials have been implicated with SJS,[48] although sulfonamides alone do not induce SJS.[50] Immunogenic reactions related to SJS has been attributed to the arylamine group within the sulfonylarylamine structural alert.[51] The supposed mechanism involves the

metabolic transformation of the arylamine group into a reactive nitroso metabolite, which covalently binds to cellular macromolecules to initiate an immune response consistent with the hapten hypothesis.[50–52] Arylamines are generally rare among drugs due to their reactivity at the nitrogen site. Exceptions often contain electron-withdrawing groups in the *para*-position to help stabilize the arylamine, such as sulfonamide-containing antibacterials, which involve a stabilizing *para*-sulfone group (SO_2).[53]

Another structural alert, β-lactam ring with adjacent sulfur, suggests that the additional sulfur atom may be necessary for SJS activity. By specifying the adjacent sulfur atom, precision increased to 100%: all β-lactam antibacterials containing it are SJS inducers in our dataset. On the contrary, analogs without the adjacent sulfur atom such as latamoxef were found to be non-inducers.

Fluoroquinolone was also identified as a structural alert since all the quinolones in the present dataset were fluoroquinolones; it is not possible to extrapolate that quinolone alone is associated with SJS. Nevertheless, this distinction seems irrelevant since non-fluorinated quinolones have been discontinued in favor of the more efficacious fluoroquinolones.[54] The analysis of tetracycline antibacterials revealed that only tetracyclines were inducers of SJS, while all three anthracyclines were non-inducers. This demonstrates that the four-ring system present in both tetracyclines and anthracyclines is not a statistically significant structural alert. By using a more refined structural alert that can differentiate the SJS-inducing tetracycline antibacterials from the non-inducing anthracyclines, we were able to improve significantly the accuracy of prediction.

Other substructures such as aromatic rings in anticonvulsants have been suggested as a structural alert for SJS in a previous study.[55] However, we did not find this structural pattern in our study using our expanded set of drugs including non-anticonvulsants. One reason may be the ubiquity of aromatic rings in both SJS inducers and non-inducers.

On the importance of using statistically significant structural alerts

Structural alerts are molecular moieties associated with a particular adverse outcome pathway and they are widely used by toxicologists and regulatory agencies to flag potential chemical hazards.[4,6] We recently showed through several case studies that the mere presence of structural alerts in a chemical is most likely an unreliable method to discriminate toxicants and thus should be avoided.[8] Structural alerts act within the whole chemical structure, even though their actual effect on chemical toxicity critically depends on their structural environment.

Nevertheless, alerts play an important role in understanding the mechanism underlying the chemical activity. They provide hypotheses of possible toxicological effects to guide further investigation and safer drug design. Although structural alerts may be derived by expert rules and/or (Q)SAR models, they must be validated by statistical analysis. More importantly, structural alerts should be used with caution for any dataset and molecular context in which they have been derived. To that end, we recently proposed a new framework that systematically integrates structural alerts and rigorously validated QSAR models for both transparent and accurate toxicity prediction of new chemicals.[8]

Conclusions

We have introduced the CWAS framework for assessing and interpreting the combined effects of molecular fragments towards the overall chemical activity. CWAS further extends the concept of quantitatively validated structural alerts by merging the advantages of well-established methods in GWAS and systems biology and QSAR modeling. Proposed CWAS framework consists of the following steps: *(i)* development of predictive QSAR model for an endpoint of interest; *(ii)* identification of important chemical fragments for this endpoint; *(iii)* determination of communities of co-occurring fragments which can be assembled to form larger and not necessarily connected structural alerts; and *(iv)* establishing the combined contribution of the co-occurring fragments into the activity. Steps (ii)-(iv) of CWAS contribute to enhance the interpretability of QSAR models.

We have applied CWAS for modeling Ames mutagenicity and SJS datasets. While for well-studied Ames mutagenicity dataset we have identified combined SA consistent with established knowledge of mutagenicity, for less-studied SJS dataset, in addition to existing SA, we identified new SA with increased precision. These alerts revealed mechanistic clues, contributing to better understanding of SJS. Overall, our results demonstrate that CWAS represents a new approach that improves the interpretation of QSAR models while preserving their predictive power. Combined structural alerts derived by CWAS consider the mutual influence of the fragments in the molecule and are useful for both effective design and safety assessment of drugs as well as for mechanistic interpretation of their action.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported in part by NIH (grants 1U01CA207160 and GM5105946), FAPEG (grant 201310267001095), and CNPq (grant 400760/2014-2). VA thanks CAPES for graduate scholarship. DF also thanks the NC State Chancellor's Faculty Excellence Program.

References

(1). Cherkasov A; Muratov EN; Fourches D; Varnek A; Baskin II; Cronin M; Dearden J; Gramatica P; Martin YC; Todeschini R; Consonni V; Kuz'min VE; Cramer R; Benigni R; Yang C; Rathman J; Terfloth L; Gasteiger J; Richard A; Tropsha A QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. 2014, 57, 4977–5010. [PubMed: 24351051]

(2). Raunio H. In Silico Toxicology-Non-Testing Methods. Front. Pharmacol. 2011, 2, 33. [PubMed: 21772821]

(3). OECD. Report of the workshop on structural alerts for the OECD (Q)SAR application toolbox http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2009)4&doclanguage=en (accessed Sep 4, 2018).

(4). Enoch SJ; Roberts DW Approaches for Grouping Chemicals into Categories In Chemical Toxicity Prediction: Category Formation and Read-Across; Cronin M, Madden J, Enoch S, Roberts D, Eds.; Royal Society of Chemistry, 2013; pp 30–43.

(5). Cronin MTD An Introduction to Chemical Grouping, Categories and Read-Across to Predict Toxicity In Chemical Toxicity Prediction: Category Formation and Read-Across; Cronin M, Madden J, Enoch S, Roberts D, Eds.; RCS, 2013; pp 1–29.

(6). Cronin MTD Evaluation of Categories and Read-Across for Toxicity Prediction Allowing for Regulatory Acceptance In Chemical Toxicity Prediction: Category Formation and Read-Across; Cronin M, Madden J, Enoch S, Roberts D, Eds.; Royal Society of Chemistry, 2013; pp 155–167.

(7). Stepan AF; Walker DP; Bauman J; Price DA; Baillie TA; Kalgutkar AS; Aleo MD Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. Chem. Res. Toxicol. 2011, 24, 1345–1410. [PubMed: 21702456]

(8). Alves V; Muratov E; Capuzzi S; Politi R; Low Y; Braga R; Zakharov AV; Sedykh A; Mokshyna E; Farag S; Andrade C; Kuz'min V; Fourches D; Tropsha A. Alarms about Structural Alerts. Green Chem. 2016, 18, 4348–4360. [PubMed: 28503093]

(9). Capuzzi SJ; Muratov EN; Tropsha A. Phantom PAINS: Problems with the Utility of Alerts for P an-A Ssay IN Terference Compound S. J. Chem. Inf. Model. 2017, 57, 417–427. [PubMed: 28165734]

(10). Polishchuk P; Kuz'min V; Artemenko A; Muratov E. Universal Approach for Structural Interpretation of QSAR/QSPR Models. Mol. Inform. 2013, 32, 843–853. [PubMed: 27480236]

(11). Low Y; Sedykh A; Fourches D; Golbraikh A; Whelan M; Rusyn I; Tropsha A. Integrative Chemical-Biological Read-across Approach for Chemical Hazard Classification. Chem. Res. Toxicol. 2013, 26, 1199–1208. [PubMed: 23848138]

(12). Benfenati E; Roncaglioni A; Petoumenou MII; Cappelli CII; Gini G. Integrating QSAR and Read-across for Environmental Assessment. SAR QSAR Environ. Res. 2015, 26, 605–618. [PubMed: 26535447]

(13). Price N; Chaudhry Q. Application of in Silico Modelling to Estimate Toxicity of Migrating Substances from Food Packaging. Food Chem. Toxicol. 2014, 71, 136–141. [PubMed: 24923263]

(14). Lozano S; Poezevara G; Halm-Lemeille MP; Lescot-Fontaine E; Lepailleur A; Bissell-Siders R; Crémilleux B; Rault S; Cuissart B; Bureau R. Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology. J. Chem. Inf. Model. 2010, 50, 1330–1339. [PubMed: 20726596]

(15). Klein RJ Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 2005, 308, 385–389. [PubMed: 15761122]

(16). Patel CJ; Bhattacharya J; Butte AJ An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One 2010, 5, e10746.

(17). Lind PM; Risérus U; Salihovic S; Bavel , B. van; Lind L. An Environmental Wide Association Study (EWAS) Approach to the Metabolic Syndrome. Environ. Int. 2013, 55, 1–8. [PubMed: 23454278]

(18). Denny JC; Ritchie MD; Basford MA; Pulley JM; Bastarache L; Brown-Gentry K.; Wang D; Masys DR; Roden DM; Crawford DC PheWAS: Demonstrating the Feasibility of a Phenome-Wide Scan to Discover Gene-Disease Associations. Bioinformatics 2010, 26, 1205–1210. [PubMed: 20335276]

(19). Holmes E; Nicholson JK Human Metabolic Phenotyping and Metabolome Wide Association Studies. Ernst Schering Found. Symp. Proc. 2007, No. 4, 227–249.

(20). Hansen K; Mika S; Schroeter T; Sutter A; ter Laak A; Steger-Hartmann T; Heinrich N; Müller K-R Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. J. Chem. Inf. Model. 2009, 49, 2077–2081. [PubMed: 19702240]

(21). Kazius J; McGuire R; Bursi R Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005, 48, 312–320. [PubMed: 15634026]

(22). CCRIS: A TOXNET DATABASE https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS (accessed Aug 16, 2018).

(23). Low YS; Caster O; Bergvall T; Fourches D; Zang X; Norén GN; Rusyn I; Edwards R; Tropsha A Cheminformatics-Aided Pharmacovigilance: Application to Stevens-Johnson Syndrome. J. Am. Med. Inform. Assoc. 2015, No. January 2016, ocv127.

(24). Lindquist M. VigiBase, the WHO Global ICSR Database System: Basic Facts. Drug Inf. J. 2008, 42, 409–419.

(25). Fourches D; Muratov E; Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J. Chem. Inf. Model. 2010, 50, 1189–1204. [PubMed: 20572635]

(26). Fourches D; Muratov E; Tropsha A. Curation of Chemogenomics Data. Nat. Chem. Biol. 2015, 11, 535–535. [PubMed: 26196763]

(27). Fourches D; Muratov E; Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. J. Chem. Inf. Model. 2016, 56, 1243–1252. [PubMed: 27280890]

(28). Varnek A; Fourches D; Horvath D; Klimchuk O; Gaudin C; Vayer P; Solov'ev V; Hoonakker F; Tetko I; Marcou G. ISIDA-Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. Curr. Comput. Aided-Drug Des. 2008, 4, 191–198.

(29). Kuz'min VE; Artemenko AG; Muratov EN Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. J. Comput. Aided. Mol. Des. 2008, 22, 403–421. [PubMed: 18253701]

(30). Varnek A; Fourches D; Hoonakker F; Solov'ev VP Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. J. Comput. Aided. Mol. Des. 2005, 19, 693–703. [PubMed: 16292611]

(31). Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform. 2010, 29, 476–488. [PubMed: 27463326]

(32). Tropsha A; Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. Curr. Pharm. Des. 2007, 13, 3494–3504. [PubMed: 18220786]

(33). Oprisiu I; Varlamova E; Muratov E; Artemenko A; Marcou G; Polishchuk P; Kuz'Min V; Varnek A. QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids. Mol. Inform. 2012, 31, 491–502. [PubMed: 27477467]

(34). Tetko IV; Sushko I; Pandey AK; Zhu H; Tropsha A; Papa E; Oberg T; Todeschini R; Fourches D; Varnek A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. J. Chem. Inf. Model. 2008, 48, 1733–1746. [PubMed: 18729318]

(35). Breiman LEO. Random Forests. Mach. Learn. 2001, 45, 5–32.

(36). Schwarz DF; König IR; Ziegler A. On Safari to Random Jungle: A Fast Implementation of Random Forests for High-Dimensional Data. Bioinformatics 2010, 26, 1752–1758. [PubMed: 20505004]

(37). Bureau A; Dupuis J; Falls K; Lunetta KL; Hayward B; Keith TP; Van Eerdewegh P. Identifying SNPs Predictive of Phenotype Using Random Forests. Genet. Epidemiol. 2005, 28, 171–182. [PubMed: 15593090]

(38). Strobl C; Boulesteix A-L; Kneib T; Augustin T; Zeileis A. Conditional Variable Importance for Random Forests. BMC Bioinformatics 2008, 9, 307. [PubMed: 18620558]

(39). Pons P; Latapy M. Computing communities in large networks using random walks http://arxiv.org/pdf/physics/0512106v1.pdf (accessed Sep 18, 2018).

(40). Chakravarti SK; Saiakhov RD; Klopman G. Optimizing Predictive Performance of CASE Ultra Expert System Models Using the Applicability Domains of Individual Toxicity Alerts. J. Chem. Inf. Model. 2012, 52, 2609–2618. [PubMed: 22947043]

(41). Valdar W; Sabourin J; Nobel A; Holmes CC Reprioritizing Genetic Associations in Hit Regions Using LASSO-Based Resample Model Averaging. Genet. Epidemiol. 2012, 36, 451–462. [PubMed: 22549815]

(42). Sushko I; Novotarskyi S; Körner R; Pandey AK; Cherkasov A; Li J; Gramatica P; Hansen K; Schroeter T; Müller K-R; Xi L; Liu H; Yao X; Öberg T; Hormozdiari F; Dao P; Sahinalp C; Todeschini R; Polishchuk P; Artemenko A; Kuz'min V; Martin TM; Young DM; Fourches D; Muratov E; Tropsha A; Baskin I; Horvath D; Marcou G; Muller C; Varnek A; Prokopenko VV; Tetko IV Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. J. Chem. Inf. Model. 2010, 50, 2094–2111. [PubMed: 21033656]

(43). Alves VM; Golbraikh A; Capuzzi SJ; Liu K; Lam WI; Korn DR; Pozefsky D; Andrade CH; Muratov EN; Tropsha A. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent

Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. J. Chem. Inf. Model. 2018, 58, 1214–1223. [PubMed: 29809005]

(44). Rücker C; Rücker G; Meringer M. Y-Randomization and Its Variants in QSPR/QSAR. J. Chem. Inf. Model. 2007, 47, 2345–2357. [PubMed: 17880194]

(45). Benigni R; Bossa C. Structural Alerts of Mutagens and Carcinogens. Curr. Comput. Aided-Drug Des. 2006, 2, 169–176.

(46). Glowienke S; Frieauff W; Allmendinger T; Martus H-J; Suter W; Mueller L. Structure-Activity Considerations and in Vitro Approaches to Assess the Genotoxicity of 19 Methane-, Benzene- and Toluenesulfonic Acid Esters. Mutat. Res. 2005, 581, 23–34. [PubMed: 15725602]

(47). Gamage N. Human Sulfotransferases and Their Role in Chemical Metabolism. Toxicol. Sci. 2005, 90, 5–22. [PubMed: 16322073]

(48). Roujeau JC; Kelly JP; Naldi L; Rzany B; Stern RS; Anderson T; Auquier A; Bastuji-Garin S; Correia O; Locati F. Medication Use and the Risk of Stevens-Johnson Syndrome or Toxic Epidermal Necrolysis. N. Engl. J. Med. 1995, 333, 1600–1607. [PubMed: 7477195]

(49). Roujeau J-C; Bricard G; Nicolas J-F Drug-Induced Epidermal Necrolysis: Important New Piece to End the Puzzle. J. Allergy Clin. Immunol. 2011, 128, 1277–1278. [PubMed: 22133320]

(50). Toler SM; Rodriguez I. Not All Sulfa Drugs Are Created Equal. Ann. Pharmacother. 2004, 38, 2166–2167. [PubMed: 15494382]

(51). Brackett CC; Singh H; Block JH Likelihood and Mechanisms of Cross-Allergenicity between Sulfonamide Antibiotics and Other Drugs Containing a Sulfonamide Functional Group. Pharmacotherapy 2004, 24, 856–870. [PubMed: 15303450]

(52). Naisbitt DJ; Hough SJ; Gill HJ; Pirmohamed M; Kitteringham NR; Park BK Cellular Disposition of Sulphamethoxazole and Its Metabolites: Implications for Hypersensitivity. Br. J. Pharmacol. 1999, 126, 1393–1407. [PubMed: 10217534]

(53). Uetrecht J. N-Oxidation of Drugs Associated with Idiosyncratic Drug Reactions. Drug Metab. Rev. 2002, 34, 651–665. [PubMed: 12214672]

(54). King DE; Malone R; Lilley SH New Classification and Update on the Quinolone Antibiotics. Am. Fam. Physician 2000, 61, 2741–2748. [PubMed: 10821154]

(55). Handoko KB; van Puijenbroek EP; Bijl AH; Hermens W. a J. J.; Zwart-van Rijkom JEF; Hekster Y. a; Egberts TCG Influence of Chemical Structure on Hypersensitivity Reactions Induced by Antiepileptic Drugs: The Role of the Aromatic Ring. Drug Saf. 2008, 31, 695–702. [PubMed: 18636788]

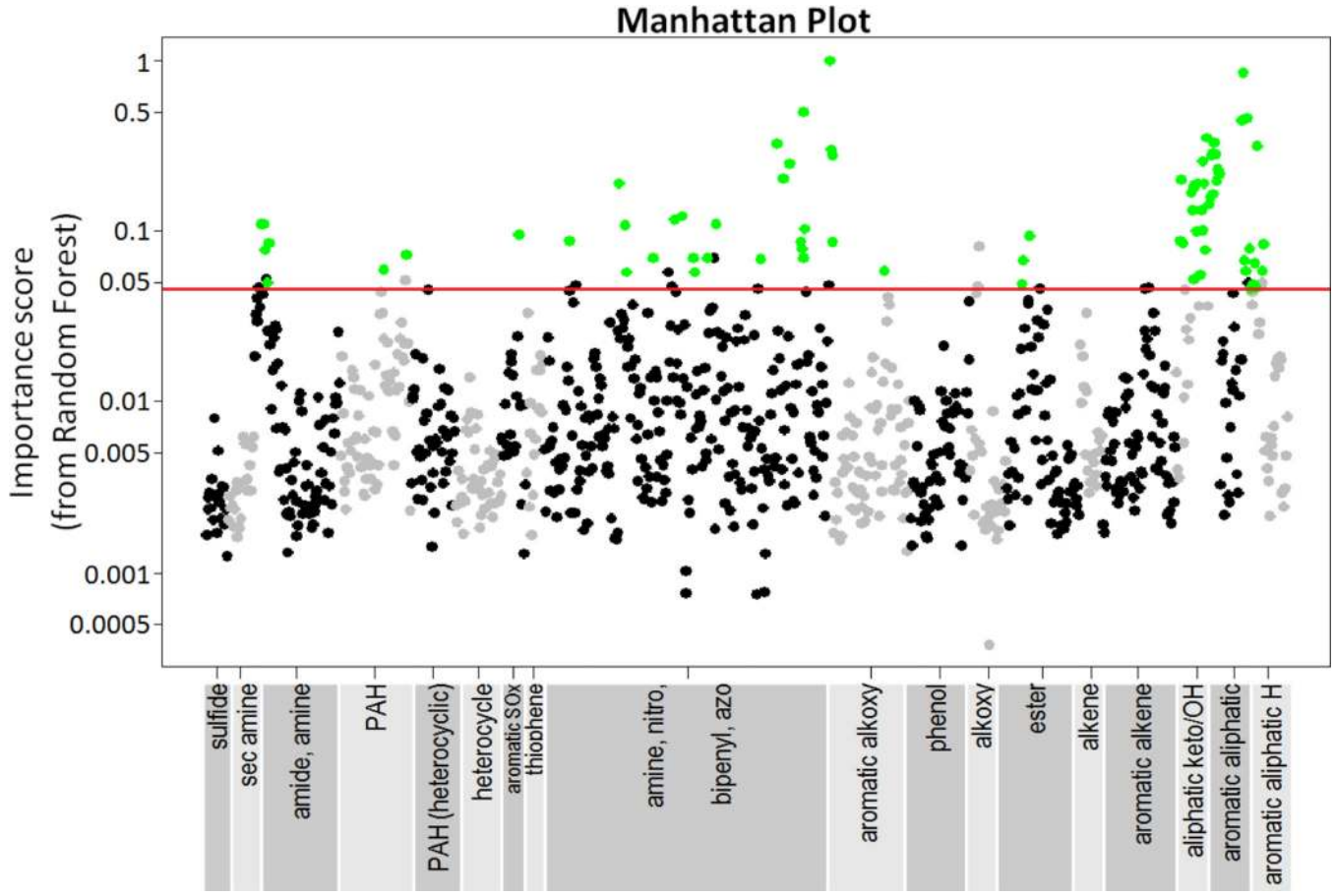


Figure 1.

Manhattan plot for Ames mutagenicity showing minimal set of 76 chemical fragments (in green) predictive of mutagenicity in all folds. Other fragments are shown in either black or gray (alternately colored for visual clarity).

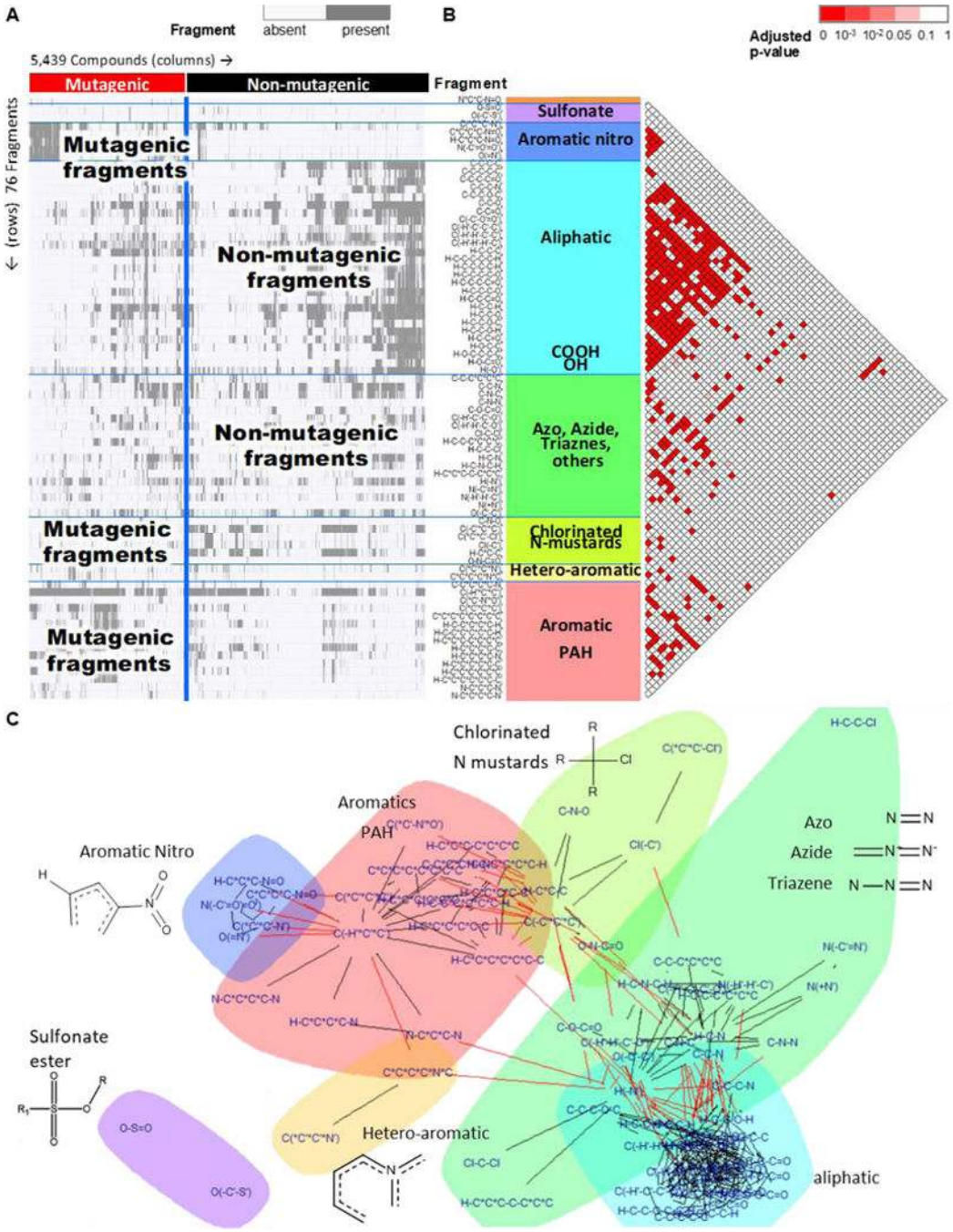


Figure 2.

Results of the co-occurrence analysis of chemical fragments for the Ames data set. (A) Heatmap shows the joint presence of 76 fragments (rows) in the mutagenic and non-mutagenic compounds (columns). (B) Triangle map shows the pairwise co-occurrence determined by Fisher's exact test: low adjusted *p*-values (< 0.1) were shaded for significance while insignificant values were unshaded. (C) The 76 fragment nodes were connected if they co-occurred with significant frequency (*p*-value < 0.1). From this network graph, the

walktrap community detection algorithm identified seven distinct subnetworks (or communities) of frequently co-occurring fragments (represented by different colors).

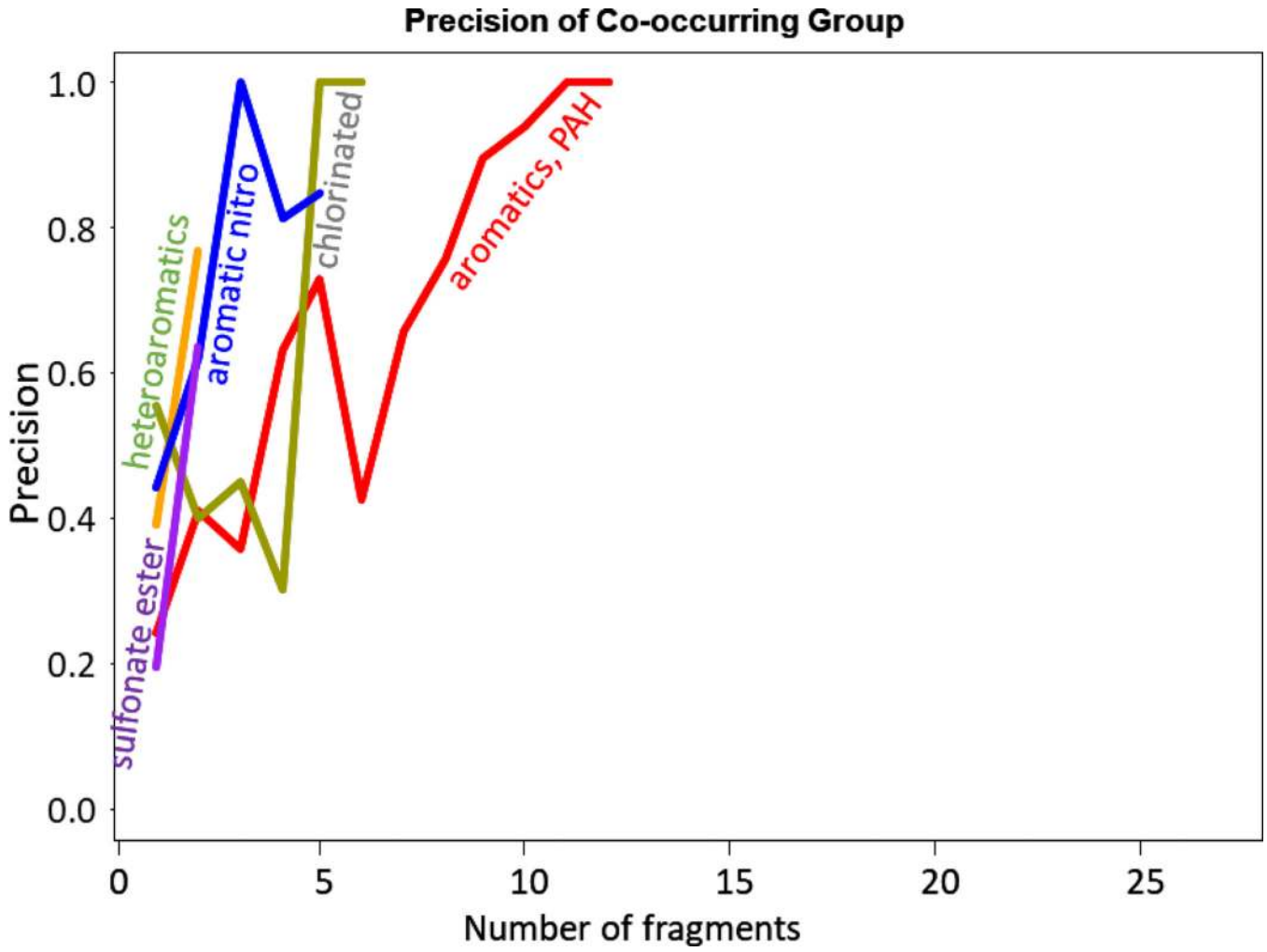


Figure 3.

Precision of five structural alerts. Precision of each structural alert increases with the number of co-occurring fragments used to construct the structural alert.

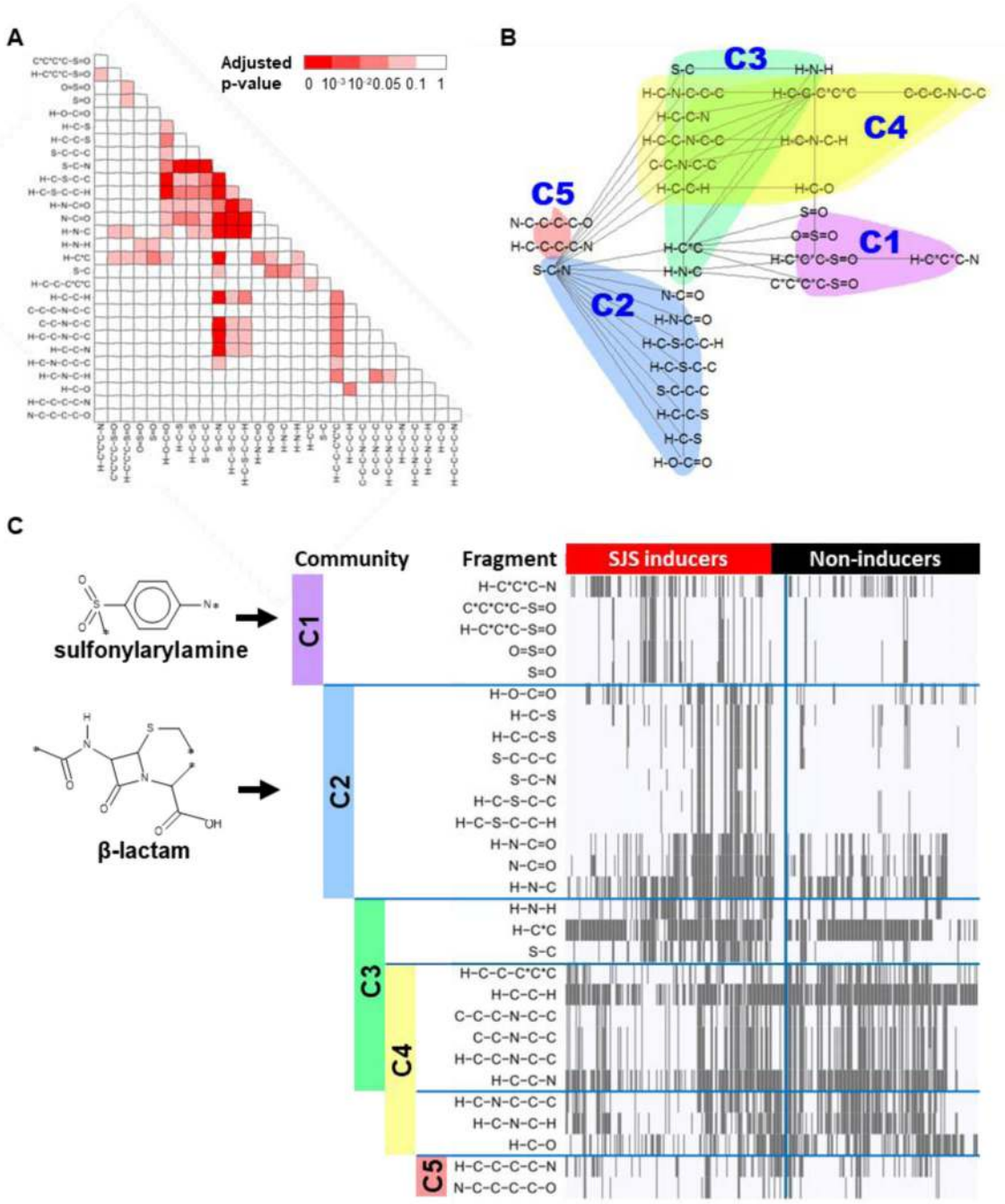


Figure 4.

Results of co-occurrence analysis of chemical fragments in the SJS data set. (A) Adjusted *p*-values show the association between pairwise co-occurrence of the 29 fragments and SJS activity. Rows show the first 28 fragments while columns show the 2nd to 29th fragment to avoid showing the self-identity diagonal (*i.e.,* pairwise co-occurrence with itself). (B) Fragment nodes are connected if significantly co-occurring (adjusted *p*-value < 0.1). The network graph is partitioned into densely connected subnetworks or communities C1–5 of co-occurring fragments using walktrap community detection. (C) As an example, the SA

reconstructed from the fragments in communities C1 (sulfonylarylamine) and C2 (beta-lactam with adjacent sulfur) are shown. The heatmap demonstrates the joint presence of the co-occurring fragments within a community (*e.g.*, C1, sulfonylarylamine), is more likely among SJS-inducing drugs than among non-inducers (fragments are mostly absent).

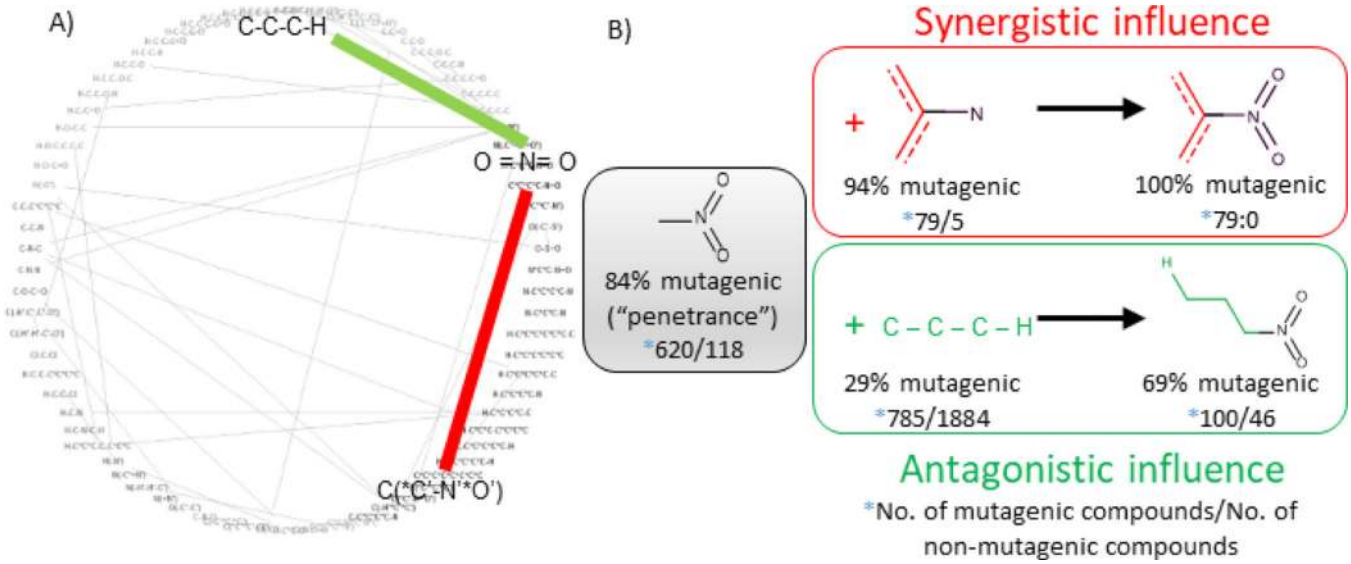


Figure 5.

(A) Interaction network plot showing synergistic (red) and antagonistic (green) interactions among selected chemical fragments in the Ames dataset; (B) Example of synergistic and antagonistic interactions affecting mutagenicity.

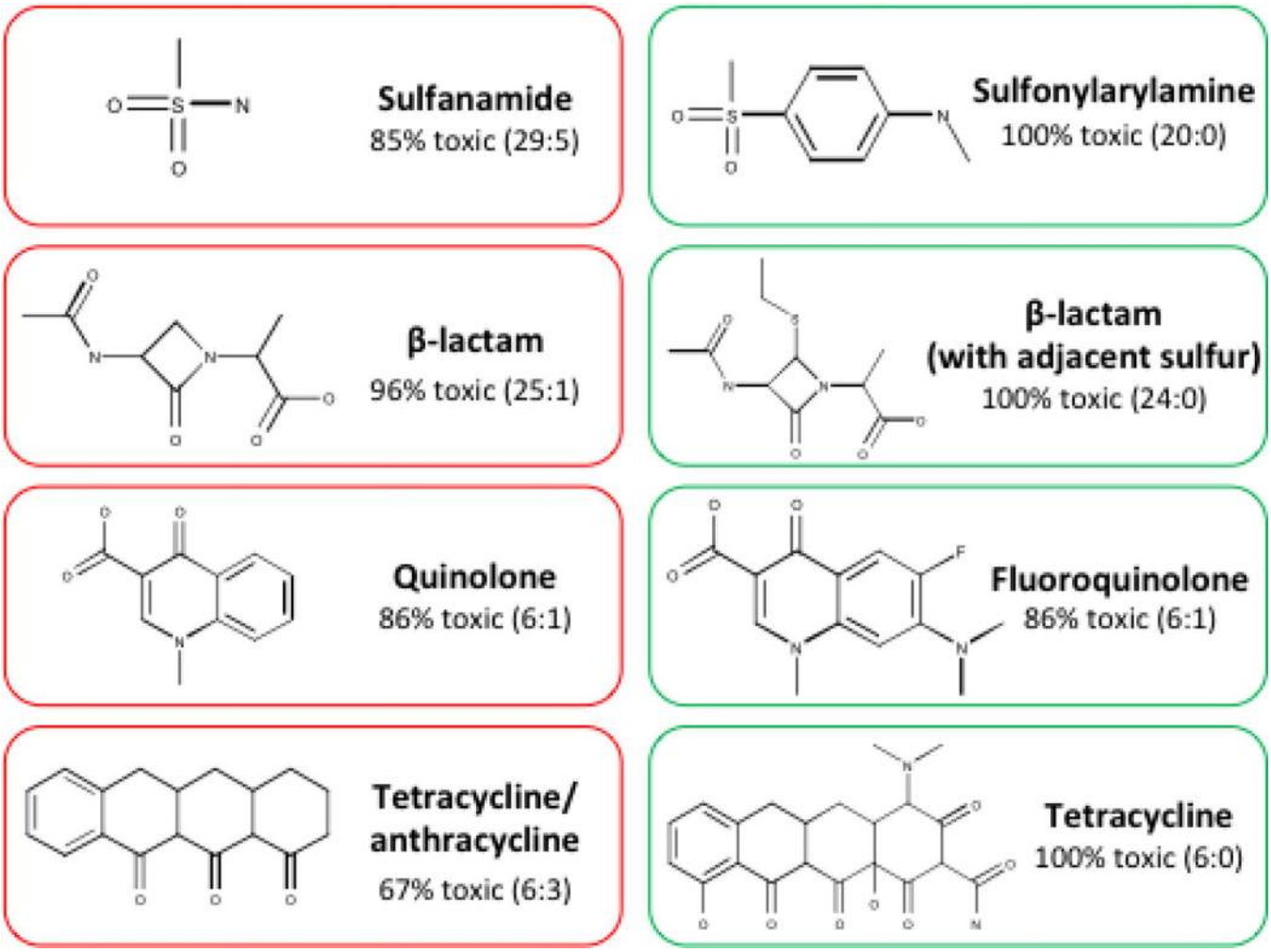


Figure 6.
Structural alerts for SJS. Left column shows expert-based structural alerts and right column shows structural alerts revealed by QSAR analysis.

Table 1.

Parallels between GWAS and QSAR.

	GWAS	QSAR
Samples	Patients	Chemical compounds
Features	Single Nucleotide Polymorphisms (SNPs)/loci	Chemical descriptors *e.g.*, fragments)
Response	Phenotype (*e.g.*, disease/no disease)	Activity (*e.g.*, active/inactive)
Objectives	Identify SNPs/loci associated with phenotype Predict phenotype from SNPs/loci	Identify structures associated with activity Predict activity from structures

Table 2.

Principal steps of CWAS: (i) QSAR modeling; (ii) fragment selection; (iii) co-occurrence analysis; and (iv) interaction analysis.

Step	Method	Output
QSAR modeling	Random forest (RF), following best practices for model development and validation	Predictive QSAR models with external CCR, sensitivity, specificity, and AUC above 0.6
Fragment selection	Variable importance on RF	Minimum set of fragments predictive of activity
Co-occurrence analysis	Fisher's exact test and network clustering by walktrap community	Structural alerts
Interaction analysis	Lasso regression with 2-way interaction	Combined effect of interacting fragments on activity

Table 3.

Prediction performance (estimated by 5-fold external CV) of QSAR models developed with all the fragments (step 0) and with reduced number of fragments (step 1).

	Mutagenicity dataset (*n=5,439*)		**SJS dataset (*n=364*)**	
	Full model (967 fragments)	**Reduced model (76 fragments)**	**Full model (1,091 fragments)**	**Reduced model (29 fragments)**
Balanced accuracy	0.85 ± 0.005	0.87 ± 0.005	0.71 ± 0.02	0.74 ± 0.02
Sensitivity	0.78 ± 0.005	0.81 ± 0.005	0.74 ± 0.04	0.77 ± 0.04
Specificity	0.92 ± 0.009	0.92 ± 0.009	0.69 ± 0.03	0.71 ± 0.03
AUC	0.91 ± 0.004	0.94 ± 0.003	0.77 ± 0.02	0.81 ± 0.02

n = number of compounds.