Published in final edited form as:
Comput Toxicol. 2019 November 01; 12: 1–13. doi:10.1016/j.comtox.2019.100100.

Comparing and contrasting the coverage of publicly available structural alerts for protein binding

Mark D. Nelms[a,b], **Ryan Lougee**[a,b], **David W. Roberts**[c], **Ann Richard**[b], **Grace Patlewicz**[b,*]

[a]Oak Ridge Institute for Science and Education (ORISE), 1299 Bethel Valley Road, Oak Ridge, TN 37830, USA.

[b]National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency (US EPA), 109 TW Alexander Dr, Research Triangle Park (RTP), NC 27711, USA

[c]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

Abstract

The molecular initiating event for many mechanisms of toxicological action comprise the reactive, covalent binding between an exogenous electrophile and an endogenous nucleophile. The target sites for electrophiles are typically peptides, proteins, enzymes or DNA. Of these, the formation of covalent adducts with proteins and DNA are perhaps the most established as they are most closely associated with skin sensitisation and genotoxicity endpoints. As such, being able to identify electrophilic features within a chemical structure provides a starting point to characterise its reactivity profile. There are a number of software tools that have been developed to help identify structural features indicative of electrophilic reactive potential to address various purposes, including: 1) to facilitate category formation for read-across of toxicity effects such as skin sensitisation potential, as well as 2) to profile substances to identify potential confounding factors to rationalise their activity in high-throughput screening (HTS) assays. Here, three such schemes that have been published in the literature as collections of SMARTS patterns and their associated chemical-biological reaction domains have been compared. The goals are 1) to better understand their scope and coverage, and 2) to assess their performance relative to a published skin sensitisation dataset where manual annotations to assign likely mechanistic domains based on expert judgement were already available. The 3 schemes were then applied to the Tox21 library and the consensus outcome was reported to highlight the proportion of chemicals likely to exhibit a reactivity response, specific to a mechanistic reaction domain, but non-specific with respect to target-tissue based activity. ToxPrint fingerprints were computed and activity enrichments computed to compare the structural features identified for the skin sensitisation dataset and Tox21

*Correspondence: Grace Patlewicz: Tel: +1 919 541 1540, patlewicz.grace@epa.gov.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Conflict of Interest
The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

chemicals for each ‘consensus’ reaction domain. Enriched ToxPrints were also used to identify ToxCast assays potentially informative for reactivity.

Keywords

Structural alerts; protein binding; reaction mechanistic domains; ToxPrint enrichments; Tox21; ToxCast

1. Introduction

As stated by Schultz et al. [1], ‘the identification of plausible molecular initiating events based on covalent reactions with nucleophiles such as peptides, proteins, and DNA provides the unifying concept for a framework for reactivity toxicity’. The formation of covalent adducts with proteins and DNA are perhaps the best-established examples as they are most closely associated with skin sensitisation and genotoxicity endpoints [2]. There are a number of software tools that have been developed to help identify structural features indicative of electrophilic reactive potential to assist in the assessment of skin sensitisation and mutagenicity endpoints. Examples include Derek Nexus (LHASA Ltd), which consists of a knowledgebase containing structural alerts for a range of different endpoints indicating potential toxicity. Other tools, such as the OECD (Q)SAR Toolbox, contain rulebases of alerts (known as profilers) that are intended to be used to group and profile chemicals based on their common structural/mechanistic/reactive potential to assist in the formation of categories for associated read-across [3, 4]. In this study, we considered three such rulebases/schemes that had been largely derived on the basis of skin sensitisation data, for which structural queries in the form of SMARTS (simplified molecular input line entry system (SMILES) arbitrary target specification) were published in the literature and/or implemented in publicly available open source tools. Given that the various schemes were derived from overlapping data sets and for common objectives, we were interested to assess the degree to which they overlap and to compare the breadth and coverage of the alerts relative to each other and in relation to manual expert assignments applied to a skin sensitisation dataset that had been compiled by Asturiol et al [5]. Two authors in this study had previously assigned the reaction mechanistic domains for the dataset in Asturiol et al [5] in a separate publication [6]. The expert assignments dataset from Asturiol et al [5] was used to pragmatically ‘ground truth’ the respective schemes in the absence of other experimental reactivity data.

The first scheme comprised a set of SMARTS published by Enoch et al [7] for each of the reaction domains that were first described by Aptula and Roberts [8] for skin sensitisation. These reaction mechanistic domains, namely Schiff Base formers, Michael addition, Acylating agents, S_N2 (substitution nucleophilic bimolecular) and S_NAr (aromatic nucleophilic substitution), are based on standard organic chemistry principles. The set of SMARTS described in [7] has also been implemented as a module in the open source tool, Toxtree (Ideaconsult Ltd).

The second scheme was a set of alerting groups reviewed by Enoch et al [9] and captured within the OECD Toolbox as a profiling scheme named ‘Protein binding alerts by OECD’.

This set of alerts were an expanded set based on a more extensive literature review rather than an attempt to encode the domains described by Aptula and Roberts [8].

The third scheme was a set of SMARTS published by researchers at The Dow chemical company [10] as part of a KNIME workflow for identifying reactive groups helpful in the assessment of acute inhalation toxicity.

The utility of these reactivity alert schemes is several-fold: 1) to identify potential electrophilic features implicated in skin sensitisation by virtue of proposed molecular initiating events (MIE),– e.g., the alerts by Enoch et al [7] and Enoch et al [9] were specifically intended to codify protein binding; 2) to profile and categorise substances for the purposes of deriving read-across predictions within analogue and category approaches -, e.g., Roberts et al [11] have long posited the utility of these reaction domains to facilitate mechanistic read-across for skin sensitisation or for the development of Quantitative Mechanistic Models (QMMs) for the prediction of skin sensitisation; and 3) to profile substances to identify potential confounding factors that might rationalise the activity (or lack of activity) outcomes of substances tested in high-throughput screening (HTS) assays such as ToxCast [12, 13]. Our motivation was to compare these protein binding reactivity schemes for 3 main purposes: 1) to understand the scope and coverage of the different schemes and the extent to which they are comparable to each other; 2) to assess the performance of the alerts relative to manual expert assignments of reaction mechanistic domains made for chemicals in the dataset published by Asturiol et al. [5]; and 3) to profile the Tox21 library [14] using reactivity alerts to identify which substances are more likely to be non-target-specific, reactive chemicals within a ‘consensus’ set of mechanistic reaction domains. Both datasets (Tox21 and the Asturiol et al. [5]) were also characterised in terms of their structural fingerprints using ToxPrint chemotypes (Chemotyper.org; [15]) to enable an enrichment analysis for each of the consensus reaction domains relative to the whole dataset. The enriched ToxPrints were then compared with ToxCast assay enrichments to identify assays potentially informative for, or impacted by, reactivity.

2. Materials and Methods

2.1 Construction of the reaction domain profiling schemes

Each of the three profiling reactivity schemes comprised SMARTS patterns associated with one of the five main reaction mechanistic domains that had been previously described by Aptula and Roberts [8], namely, bimolecular nucleophilic substitution (S_N2)-acting, Michael acceptors (MA), Acyl transfer agents, Schiff Base formers (SB) and nucleophilic aromatic substitution (S_NAr)-acting. These reaction domains are based on organic chemistry principles describing the reactions that occur between electrophiles and nucleophiles.

In each case, the file of SMARTS and their associated reaction mechanistic domain was converted to a Python dictionary where the mechanistic domain formed the key and the set of SMARTS patterns, the values. RDKit’s python library (RDKit.org) was then used to transform the SMARTS strings into chemical substructures.

2.2 Comparing and contrasting the performance of the reaction schemes using a dataset with known manually annotated reaction domains

The skin sensitisation dataset from Asturiol et al [5] that had been annotated with reaction domains by Patlewicz et al [6] was chosen as a 'benchmark' set to facilitate a comparison of how the 3 different schemes performed. This was a pragmatic approach in the absence of other experimental data that would objectively measure the reactivity and confirm the appropriate reaction domain, as well as in the absence of the complete training sets that were used to derive the original SMARTS in the respective schemes. The Asturiol et al [5] dataset will be referred to as the "JRC dataset" throughout this manuscript, where JRC refers to Asturiol's affiliation at the European Commission's Joint Research Centre. The JRC dataset included reported outcomes in *in chemico*, *in vitro* and *in vivo* tests for skin sensitisation. The local lymph node assay (LLNA) was the *in vivo* test result reported whereas the *in chemico* and *in vitro* tests make reference to 3 assays that have been mapped to key events (KEs) in the Adverse Outcome Pathway (AOP) for skin sensitisation [16]. The assays are namely the DRPA (direct reactivity peptide assay), an *in chemico* test for the molecular initiating event (MIE) [17], whereas the KeratinoSens test [18] and the h-CLAT (human cell line activation test) [19] are *in vitro* tests that characterise KEs 1 and 2, respectively. The benchmark dataset and the three reaction domain alert schemes are briefly summarised in Table 1.

DSSTox Substance Identifier (DTXSID) chemical IDs were identified and mapped to names and/or CAS registry numbers (CASRN) in the JRC dataset to facilitate the extraction of QSAR-ready SMILES from the EPA CompTox Chemicals Dashboard ([20]; https://comptox.epa.gov/dashboard, Accessed 29 January 2019). The DSSTox (Distributed Structure-Searchable Toxicity) database is underpinned by a chemical registration process to ensure the quality of structure mapping with chemical names, CASRN and INChIkeys. QSAR-ready SMILES are the result of a structure standardisation process where structures are normalised and desalted as described by Mansouri et al [21]. Mapping the JRC dataset identifiers to DTXSIDs would also aid subsequent comparisons with lists such as the Tox21 screening library, used in this study. A total of 222 chemicals with unique SMILES were identified. RDKit was then used to perform a substructure match for each SMILES string against any of the SMARTS in each alert scheme to identify any and all mechanistic domains associated for a given chemical.

The JRC dataset was profiled using each of the 3 reaction domain rulebases and the outcomes were compared to the expert-derived assignments. Performance metrics including precision, recall, F1 score, Matthews correlation coefficient (MCC), and Cohen's Kappa were calculated for each profiling scheme relative to the expert judgement calls. The discordant assignments relative to the expert judgement calls were also reviewed and rationalised to identify whether any potential refinements were merited, or gaps existed for the alerts themselves. The formulas for the different metrics are provided below:

Precision or positive predictive value (PPV) is defined as:

$$\frac{TP}{TP+FP}$$

Recall or True positive rate is also referred to for 2 class problems as the sensitivity. It is defined as the:

$$\frac{TP}{TP+FN}$$

F1 score is the harmonic mean of precision and recall and is defined as:

$$\frac{2TP}{(2TP+FP+FN)}$$

Mathews correlation coefficient (MCC) is defined as:

$$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)(FN+TP)(TN+FP)(TN+FN)}}$$

Cohen's kappa coefficient (κ) measures inter-rater agreement for categorical items that takes into account the possibility of agreement occurring by chance [22].

Performance metrics were calculated using available functions within the Scikit-learn (v0.19.1) (https://scikit-learn.org/stable/#), StatsModels (v0.9.0) (http://www.statsmodels.org/dev/index.html) and NumPy (v1.14.5) (www.numpy.org) python packages where version numbers are provided in parentheses.

2.3 Profiling of the ToxCast/Tox21 substances

The Tox21 Screening Library (Tox21SL) list of 8947 chemicals was downloaded from the EPA CompTox Chemicals Dashboard (available at: https://comptox.epa.gov/dashboard/chemical_lists/TOX21SL). This list was filtered to retain substances that were discrete organics with normalised structures, i.e. only substances that had an associated QSAR-ready SMILES record. There were 8360 chemicals with unique DTXSID identifiers and QSAR-ready SMILES after deduplication. Each of the 3 profiling schemes was applied to this set of chemicals and an overall 'consensus' reaction domain was predicted for each chemical based on the following logic. If the outcome from all 3 schemes was identical, this formed the final 'consensus' prediction. If 2 schemes were identical this consensus was reported as the final 'consensus' prediction. If only one scheme flagged one or more domains, this result was given a higher weight than a negative and served as the final prediction. In cases where the predicted reaction domains (or no domain) for all 3 schemes conflicted with each other, the final prediction was labeled as inconclusive.

Chemical structural fingerprints were generated for both the JRC dataset and the Tox21 database using the 729 Toxprint set of chemotypes [15], downloaded from the EPA CompTox Chemicals Dashboard as a tsv file (Chemotyper format) using the batch search capability. ToxPrint chemotypes were selected since these are a publicly available approach for representing molecules and are specifically tailored to provide good coverage of environmental, regulatory and commercial use chemical space. ToxPrints were used to explore whether particular structural features were enriched for each of the expert-assigned

or consensus predicted reaction domains relative to the whole dataset, for both the JRC dataset and the Tox21 dataset. A ToxPrint with an odds ratio equal or greater than 3, a p value less than or equal to 0.05 and the number of True Positives (TPs) greater than or equal to 3 (i.e., at least 3 chemicals within the reaction domain contain the ToxPrint) was defined as being significantly enriched in the reaction domain subset relative to the whole. The enriched ToxPrints within each 'consensus' reaction domain of the Tox21 screening library dataset were compared with the enriched ToxPrints within each expert-assigned reaction domains of the JRC dataset.

All of the above analyses were performed in Python 3.6 using NumPy (v1.14.5), Pandas (v0.23.3), RDKit (v2018.03.3.0), and Scikit-learn (v0.19.1). Plots were created in Seaborn (v0.9.0) and Matplotlib (v2.2.2). All code and datasets are available in associated Jupyter (v1.0.0) notebooks as part of the supplementary information available at https://gaftp.epa.gov/Comptox/NCCT_Publication_Data/PatlewiczGrace/CompTox-Protein_binding_alerts-comptoxicol/.

The enriched ToxPrints for each consensus reaction domain for both datasets were also compared with ToxCast assay enrichments to identify which HTS assays might be informative for reactivity. ToxPrint enrichment calculations in the ToxCast assay space were computed previously using the Chemotyper Enrichment Workflow (CTEW) code base developed by EPA researchers for application to ToxCast HTS datasets (see e.g. [23]). The CTEW employs the same odds ratio, p-value and TP statistical thresholds as applied in the current analysis to the JRC and Tox21 datasets. Enrichments were previously calculated and stored for all 1192 unique ToxCast assay endpoints, with the latter represented as baseline (binary) hit calls from level 5 (Mc%) within the ToxCast public data release (invitrodb_v2) (https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data).

3. Results and Discussion

3.1 Comparing and contrasting the reaction schemes

The 3 schemes were first compared by the number of alerts present in each of the reaction domains. The OECD alerts comprised 102 SMARTS patterns in contrast to the Enoch alerts which consisted of 66 SMARTS patterns and the Dow alerts that contained 71 SMARTS patterns. Based on visual inspection (Figure 1), each scheme appeared particularly enriched with alerts characterising the Michael acceptor (MA) domain whereas far fewer alerts captured the S_NAr domain. The Enoch scheme was unusual in that it comprised almost the same number of alerts for the S_NAr domain as the S_N2, but far fewer alerts in the S_N2 domain than the other schemes. The Dow scheme comprised only 4 alerts for Schiff base formers (SB) and the OECD alerts comprised only 1 alert for S_NAr.

The profile of the number of SMARTS patterns relative to each domain mirrors the expected profile for how chemicals tested for their skin sensitisation are categorised by domain. This is unsurprising given how many of these alerts were originally devised through evaluating the chemistry of skin sensitisation datasets. One related example, published by Roberts et al [24], for a dataset consisting of 210 chemicals, showed that Michael acceptors and S_N2 chemicals were the most abundant reaction domains based on expert derived manual

assignments (see Figure 2 that summarises the number of chemicals assigned to each domain).

However, it should be noted that the number of alerts per domain within the schemes does not necessarily reflect the breadth of chemical space covered; some alerts are more general than others, thus spanning more diverse chemicals, and some alerts are nested within others or may frequently co-occur with others, thus spanning fewer chemicals. Hence, a relevant reference dataset is needed to assess relative coverage of the three schemes. In this respect, comparing a dataset of pre-assigned mechanistic domains with the domains identified by each scheme would provide a better understanding of the coverage and applicability of the alerts. The reaction domain assignments applied to the JRC dataset were informed by mechanistically relevant chemical and biological knowledge and data, in addition to purely structural considerations, and thus serve as the baseline “ground truth” for the present analyses. The JRC dataset comprised 222 chemicals for which QSAR-ready SMILES were extracted from the EPA CompTox Chemicals Dashboard. The set of chemicals were profiled through each of the schemes using the 3 sets of SMARTS patterns to assign all and any reaction domain(s). The profiled domains together with the expert-assigned domains were formatted for consistency to enable comparisons to be made and performance metrics to be computed.

The comparisons were complicated by 2 factors: 1) although there were only 5 specific reaction domains identified by the respective profiling schemes, some chemicals flagged more than 1 domain; and 2) the expert assignments included additional pathways (denoted as ‘special case’) to account for chemicals undergoing other transformations such as autoxidation. For the JRC dataset, there were 15 unique expert judgement assignments (which captured both additional pathways as well as combinations of the 5 reaction domains) whereas the Enoch profiler gave rise to 11 different reaction domain combinations (i.e. only 1 or more of the 5 reaction domains). The OECD and Dow profilers produced 8 different assignments. Table 2 provides the macro average performance characteristics for the 3 schemes relative to the expert judgement calls, whereby the precision and recall for each reaction domain(s) is computed first and then averaged across the number of domain outcomes.

The MCC, defined earlier, is a useful measure to summarise the overall performance, where the minimum value is −1 and the best value is +1. On the basis of this metric, the Enoch alerts were overall ‘the best’ at classifying the JRC dataset into their respective domains relative to the expert judgement calls, whereas the OECD alerts were a close second. A factor that contributes to the performance characteristics could also be the coverage of the SMARTS patterns that were flagged for the chemicals in the JRC dataset relative to the total of SMARTS patterns that exist for each scheme. The number of unique SMARTS patterns for each of the schemes is 66 for the Enoch scheme, 71 for the Dow scheme and 102 for the OECD as referenced in Figure 1. The number of unique SMARTS patterns that were matched for the JRC chemicals was 47 for the Enoch scheme, 32 for the Dow scheme and 38 for the OECD scheme. Thus 47/66 (71%) of the Enoch SMARTS were triggered for the JRC scheme whereas fewer SMARTS patterns were flagged for the other two schemes, 32/71 (45%) for the Dow scheme and 38/102 (37%) for the OECD scheme. Interestingly,

even though the Enoch scheme contained the fewest total number of alerts (66), these were better aligned to, and represented within the corresponding JRC expert-assigned reaction domains than the much larger set of OECD scheme alerts. This can perhaps be attributed to the closer proximity and overlap of the JRC dataset studies to the earlier Enoch, 2008 study. The proportion of alerts that were flagged could be another factor in why the Enoch scheme appeared to perform slightly better in comparison to the other 2 schemes. Table 2 provides the micro per domain performance metrics for the 3 schemes relative to the expert judgement calls for the 5 principal reaction domains only. The intent was to evaluate whether one scheme was better at matching one specific domain relative to another scheme. Robust performance characteristics could not be computed for the multi-domain combinations and are not reflected in Table 3. Indeed, for inconclusive multi-domain cases, further evaluation would be needed to determine which reaction domain dominated either through analysis of the frontier molecular orbitals (using the shapes and energies of the Lowest Unoccupied Molecular Orbital) or through generating relevant *in chemico* data [7].

The Enoch scheme appears to be overall the best performing for 4 of the 5 domains listed based on the F1 score. The Dow and OECD schemes appear to be better at correctly assigning chemicals within the S_N2 domain, albeit requiring many more alerts, i.e., 23 and 39, respectively, vs. only 9 alerts for the Enoch scheme. By the same token, the 23 patterns characterising the S_N2 domain in the Dow scheme appear to be sufficient to cover the scope of S_N2 reacting chemicals in this dataset. In other cases, such as the Enoch scheme for predicting in the Acyl domain, fewer alerts (9) appeared to perform as well or better than the 11 or 17 alerts for the Dow and OECD schemes, respectively. Clearly, however, relative performance of these schemes in assigning reaction domains on a reference dataset, such as the JRC dataset in this case, would be expected to depend on the reaction domain knowledge base, or training set, used in their development. Hence, the better performance of the Enoch scheme on the JRC dataset may be due to closer alignment of those studies. The JRC dataset was also profiled to showcase which SMARTS pattern was triggered for each substance to result in its associated reaction domain score (Figure 3). Actual SMARTS patterns were coded to designate source origin for ease of plotting. Although this provided an overall visual perspective of the consistency in reaction domain assignment across the schemes and allowed some inspection of which SMARTS pattern(s) were responsible in case, it also highlighted how difficult it was to reconcile one SMARTS pattern relative to another from a different scheme. This difficulty in part prompted the subsequent exploration of enrichments using ToxPrints that are objective and well-defined features.

The dataset was then filtered to consider the set of substances for which all 3 schemes were in agreement with each other, but in conflict with the expert judgement assignments. There were 25 substances identified that met these conditions, with 21 of these being cases where no alerts were triggered by any of the 3 schemes. These are listed in Table 4 together with their 2D chemical structural representation.

Of the 25 cases, 3 of the original expert assignments appeared to be erroneous as discussed in Table 4; correcting these would result in only 22 cases where all 3 schemes were in agreement with each other but in conflict with the expert assignments. These included 1-Benzoylacetone (changed from Acyl to SB), and Hexyl salicylate and 5-Dodecanolide,

both changed to likely non-reactive. The conflicts were typically observed for substances that were expected to act by an alternative pathway, either directly via S_N1 (unimolecular nucleophilic substitution) or indirectly by a free radical mechanism as a result of oxidation. This is not a specific limitation in this study as our primary interest was to have a means of identifying inherently electrophilic reactivity.

3.2 Pairwise comparison of the 3 schemes

Cohen's kappa coefficient was used to measure the inter-rater agreement for the results of the 3 schemes pairwise across the reaction domain assignments. The kappa value for each of the pairs of reaction domain schemes was >0.6 (values were 0.69 for the OECD and Enoch pair, 0.67 for the OECD and Dow pair, and 0.61 for the Enoch and Dow pair), indicating that whilst the schemes have good agreement with one another, no scheme was completely replicated by another scheme. Accordingly, rather than excluding any particular scheme for any subsequent profiling activity or application, a 'consensus' from all three schemes was used.

3.3 Profiling of the Tox21 substances

The Tox21 list was identified from the EPA CompTox Chemicals Dashboard and all substances in the list were 'sent' to the batch search to download additional chemical information, including QSAR-ready SMILES where available and the ToxPrint fingerprints. The list was filtered to remove all inorganics and mixtures, resulting in a set of 8360 unique substances with QSAR-ready SMILES. Six SMILES could not be resolved by RDKit, such that the final Tox21 dataset used in the remainder of the analysis comprised 8354 substances with QSAR-ready SMILES. Each of the Tox21 SMILES was processed through the 3 protein binding alert schemes to predict their reaction mechanism domain(s). These were then aggregated into a 'consensus' outcome as described in Materials & Methods. Figure 4 shows the profile of the Tox21 chemicals across the consensus reaction domains, which include the 5 main domains, the inconclusive category, as well as 13 combination domains (in absolute terms and on a log-scale).

Over 55% of the Tox21 substances lacked an alerting feature suggesting either that they were either not inherently reactive or the knowledgebase was insufficient to categorise them as reactive. However, the 45% that remained are indicated to have the potential to act as electrophiles on the basis of the alerts: 12% of the library was categorised as acyl transfer agents (Acyl), 8% as S_N2 acting and 7% as Michael acceptors and Schiff base formers. For the 7% that were categorised as inconclusive, some degree of reactivity is indicated, but there is insufficient consensus to assign the chemical to a particular reaction mechanism domain. In these cases, alerts from 2 or more schemes triggered an alert in a substance, but they were associated with different reaction domains. For example, a chemical could be assigned as MA based on the Enoch scheme, SB based on the OECD scheme, and both [MA, SB] based on the Dow scheme. Given the large proportion of substances highlighting an alert, ongoing work should be focused on exploring approaches to evaluate the robustness of these alerts (given many of them are strongly influenced by human experts and bounded by the chemicals assessed for skin sensitisation potential), and investigate the feasibility of deriving new alerts by comparison to actual reactivity data. A new cross-partner Tox21

project has just commenced to generate experimental reactivity data to serve such a purpose (https://tox21.gov/projects/).

3.4 ToxPrint chemotype characterisation and enrichment of the Tox21 library

ToxPrint chemotype fingerprints (ChemoTyper format) were extracted from the EPA CompTox Chemicals dashboard for the Tox21SL chemical list and merged with the consensus reaction domain prediction based on presence or absence of alerting features. There are 729 ToxPrints in total, though many of these are structured in a hierarchy of different levels representing more-to-less generalised chemical features. For profiling of the library, a reduced number of higher level (more general) ToxPrints were used, so-called level 2 ToxPrints of which there are 70 in total. These were defined on the basis of taking the first and second name elements from the original ToxPrint name; e.g., if the full original ToxPrint name was 'bond:C#N_cyano_acylcyanide' the corresponding Level 2 name would take the first and second name components, hence 'bond:C#N'. The ToxPrints computed for the Tox21 library were mapped to the level 2 ToxPrints for profiling purposes. Figure 5 shows the profile of the Tox21 library using ToxPrints projected onto the 5 main consensus-alert reaction domain assignments.

Upon inspection of the barplots in Figure 5, the types of ToxPrints represented in each domain appear similar, though their relative frequencies differ. To better parse out ToxPrints that might be more specific to each domain, an enrichment analysis was performed to identify which ToxPrints were significantly enriched in each consensus-alert defined reaction domain relative to the full set of Tox21 chemicals. Table 5 shows the top 5 enriched chemotypes (by odds ratios) for each of the 5 main consensus reaction domains for the Tox21 chemical set.

Noteworthy is that there are no overlaps of ToxPrints across the 5 reaction mechanism domains, i.e., no ToxPrint in the top 5 enriched set is in more than one reaction mechanism domain. Hence, these particular ToxPrints are highly specific to each distinct reaction mechanism domain.

The enrichment was also compared with that for the JRC dataset used earlier in the study to compare and contrast which ToxPrints were enriched in each domain and how this differed for the 2 datasets. For comparative reasons, the consensus domain was predicted for the JRC set and used for the enrichment analysis. The top 5 odds ratios and chemotype comparisons for the JRC dataset is reflected in Table 6.

Once again, there is no overlap in the sets of top 5 (or less) chemotypes within the JRC set, indicating distinct chemical feature signatures for each of these reaction mechanism domains. In addition, and perhaps more surprisingly, there was minimal overlap in the top 5 chemotypes for 4 of the 5 reaction mechanism domains when comparing the consensus reaction domain results for Tox21 (Table 5) to those for the JRC dataset (Table 6). The MA and S_NAr/S_N2 domains all shared only 1 ToxPrint in common. The S_N2 domain shared no ToxPrints whereas 4/5 ToxPrints were common in both datasets for the SB domain.

The JRC dataset is largely made up of chemicals used in the cosmetics sector which have prompted the study of skin sensitisation. In addition, 181 of the chemicals in the JRC dataset overlapped with the Tox21 dataset. Many of these chemicals are fragrances, whereas the Tox21 library covers a much broader spectrum of environmental chemicals and use-categories in the Industrial sector, including drugs and pesticides [13]. Given the much larger size of Tox21 in relation to the JRC dataset (40:1), and the intentional structure diversity of the former versus a narrowed bias of the latter based on testing and concern for skin sensitisation, it is also possible, and perhaps likely, that Tox21 covers a larger domain of reactive chemistry than is captured by the JRC dataset. If this were the case, the 55% of Tox21 chemicals lacking an alerting feature might contain “false negatives” in reaction chemistry space. Likewise, the much larger structural diversity of Tox21 in relation to the small JRC dataset in the “active” region, i.e., in the nearly 4000 Tox21 chemicals containing a reactive alerting feature, opens up possibilities for refining or expanding the expert-based or consensus alerts if additional confirmational HTS data could be generated.

A comparison of the ToxPrints significantly enriched within each consensus reaction domain for the 2 datasets was performed and is available in the supplementary information. In total for the MA domain, there were 33 enriched ToxPrints in Tox21 and 14 in the JRC dataset. Of these, 13 enriched ToxPrints were in common to both datasets, 1 ToxPrint was unique to the JRC dataset, and 20 chemotypes were unique to the Tox21 library, the latter indicating a richer capture of structural diversity in relation to the alert-based reaction groups. Table 7 lists the number of unique and commonly enriched ToxPrints for each of the 5 consensus reaction domains.

As an illustration, the enriched chemotypes for the MA domain were mapped back to the original chemical structures in the respective datasets to showcase the types of chemicals that alerted for this domain. Tables 8 and 9 show illustrative chemicals for the common enriched chemotypes and those unique to the Tox21 set, respectively.

3.5 Enriched ToxPrints projected in the ToxCast assay space

The enriched ToxPrints derived from the JRC and Tox21 datasets were projected in the ToxCast assay space to identify assays sharing these particular enrichments that would be potentially informative for, or impacted by, reactivity. Full results are provided in the supplementary information. Here, we only consider the assays that are enriched with the same ToxPrints as are also enriched for the consensus MA domain within the JRC dataset.

In this case, there were 14 enriched ToxPrints, which if projected on the ToxCast assay space identified 359 different assays that were, likewise, significantly enriched for those ToxPrints within the assay active “hit” space. The assays covered a spectrum of different vendors (9) with a wide range of assays per vendor; from 6 Tanquay assays to 238 Bioseek assays (Table 10). For more specific ToxPrints such as ‘bond: C=0_carbonyl_ab-unsaturated_aliphatic_(michael_acceptors)’, a more targeted set of assays (22) were found which covered 6 different vendors (ATG, BSK, CEETOX, TOX21, OT and NVS).

4. Summary & Conclusions

In this study, three schemes which encode protein binding structural alerts for 5 major reaction mechanism domains as captured in SMARTS queries, were compared. Their relative information content and coverage was assessed by comparing the number of SMARTS queries in each of the 5 reaction domains and how their domain assignments for each chemical, i.e. based on presence/absence of alerting features, contrasted relative to manual expert assignments that had been annotated in a published skin sensitisation dataset. The overall performance of the 3 alert schemes relative to the manual assignments were reasonable; differences were observed on a per reaction domain basis, with the scheme by Enoch [8] generally performing better perhaps due to a closer alignment with the benchmark study that provided expert reaction domain assignments. Inspecting the cases where all 3 of the schemes were the same, but conflicted with the manual expert assignments, revealed that these largely included cases where an alternative reaction pathway could be postulated that fell outside of the 5 reaction domains and, typically, where some transformation was required.

The Tox21 screening library was then profiled with the 3 schemes to derive a consensus alert outcome. In the case of Tox21, we hypothesised that the presence of an alerting feature for reactivity might provide a marker for a non-specific reactivity response, i.e. one in which a specific enzyme target interaction is likely not taking place. A large fraction of the Tox21 library (45% of the total) alerted for 1 or more reaction domains. The profile of these predicted reaction domains was explored using the publicly available ToxPrint chemotypes, which offer a more transparent, interpretable, and standardized means for representing chemical features than SMARTS. ToxPrints provided a means to characterise the reactivity subsets relative to the whole Tox21 set, as well as to compare different sets of chemicals within each reaction domain through a common chemical interface. The types of chemotypes appeared to be broadly similar for chemicals in each of the predicted consensus reaction domains but their frequencies within the domains differed.

A chemotype enrichment analysis was performed for both the JRC dataset and the Tox21 library to identify which chemotypes were particularly enriched for a specific reaction domain, to what extent they differed, and to what extent the JRC and Tox21 results overlapped. In both cases, the top 5 enriched chemotypes in each consensus reaction domain were highly specific to that domain, with very few enriched chemotypes spanning in multiple domains. In addition, and somewhat surprisingly, there were very few overlapping chemotypes in this top 5 enriched set when comparing the results for the Tox21 library with those of the JRC dataset. This is likely due to the very different chemical constituents of the two datasets. Expanding the comparison to the more complete set of ToxPrint enrichments in each case reveals a much larger degree of overlapping enrichments. This and the greater size and diversity of the Tox21 library in relation to the JRC dataset was illustrated for the MA domain, where there was 1 ToxPrint specific to the JRC dataset, 20 unique to the Tox21 library, and 13 common to both datasets. Example chemicals that presented these enriched ToxPrints are highlighted for the MA domain. The enriched ToxPrints were projected on to the ToxCast assay space to identify assays sharing these particular enrichments that would be potentially informative for, or impacted by, reactivity. For the consensus MA domain,

within the JRC dataset, there were 359 potentially informative assays for the 14 enriched ToxPrints, covering a number of different vendors. A broader look at assays across all the consensus domains might be more informative to readily identify a subset of assays that are more indicative of general reactivity. The profiled consensus alert-predicted reaction domains and the ToxPrints that were identified as enriched within these domains will be investigated in greater detail as part of a cross partner Tox21 project where *in chemico* reactivity data will be generated to evaluate the utility and relevance of the alerting schemes and ToxPrints.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors wish to acknowledge Dr Steve Enoch, Liverpool John Moores University, for kindly providing the SMARTS patterns for the Protein binding schemes in an electronic form.

References

[1]. Schultz TW, Carlson RE, Cronin MT, Hermens JL, Johnson R, O'Brien PJ, Roberts DW, Siraki A, Wallace KB, Veith GD. A conceptual framework for predicting the toxicity of reactive chemicals: moderlling soft electrophilicity. SAR QSAR Environ Res. 17 (2006) 413–428. [PubMed: 16920662]

[2]. Schwöbel JA, Koleva YK, Enoch SJ, Bajot F, Hewitt M, Madden JC, Roberts DW, Schultz TW, Cronin MT. Measurement and estimation of electrophilic reactivity for predictive toxicology. Chem Rev. 111 (2011) 2562–2596. doi: 10.1021/cr100098n. [PubMed: 21401043]

[3]. Dimitrov SD, Diderich R, Sobanski T, Pavlov TS, Chankov GV, Chapkanov AS, Karakolev YH, Temelkov SG, Vasilev RA, Gerova KD, Kuseva CD, Todorova ND, Mehmed AM, Rasenberg M, Mekenyan OG. QSAR Toolbox - workflow and major functionalities. SAR QSAR Environ Res. 19 (2016) 1–17.

[4]. Patlewicz G, Helman G, Pradeep P, Shah I. Navigating through the minefield of read-across tools: A review of in silico tools for grouping. Computational Toxicology. 3 (2017) 1–18. doi: 10.1016/j.comtox.2017.05.003 [PubMed: 30221211]

[5]. Asturiol D, Casati S, Worth A. Consensus of classification trees for skin sensitisation hazard prediction. Toxicol In Vitro. 36 (2016) 197–209. doi: 10.1016/j.tiv.2016.07.014. [PubMed: 27458072]

[6]. Patlewicz G, Casati S, Basketter DA, Asturiol D, Roberts DW, Lepoittevin JP, Worth AP, Aschberger K. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? Regul. Toxicol. Pharmacol. (2016). pii: S0273–2300(16)30228–8. doi: 10.1016/j.yrtph.2016.08.007.

[7]. Enoch SJ, Madden JC, Cronin MT. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. SAR QSAR Environ Res. 19 (2008) 555–578. doi: 10.1080/10629360802348985. [PubMed: 18853302]

[8]. Aptula AO, Roberts DW. Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. Chem Res Toxicol. 19 (2006) 1097–1105. [PubMed: 16918251]

[9]. Enoch SJ, Ellison CM, Schultz TW, Cronin MT. A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. Crit Rev Toxicol. 41 (2011): 783–802. doi: 10.3109/10408444.2011.598141. [PubMed: 21809939]

[10]. Wijeyesakere SJ, Wilson DM, Settivari R, Auernhammer TR, Parks AK, Marty S. Development of a profiler for facile chemical reactivity using the open-source Konstanz information miner. Appl In Vitro Toxicol. 4 (2018) 202–213.

[11]. Roberts DW, Aptula AO, Patlewicz G, Pease C. Chemical Reactivity Indices and Mechanism-based read across for non-animal based assessment of skin sensitization potential. J. Appl. Toxicol. 28 (2008) 443–454. [PubMed: 17703503]

[12]. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstrueur N, Knudsen T, Martin M, Padilla S, Reif D, Richard A, Rotroff D, Sipes N, Dix D. Updated on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem Res Toxicol. 2012. 25(7): 1287–1302. Doi: 10.1021/tx3000939. [PubMed: 22519603]

[13]. Richard A, Judson R, Houck K, Grulke C, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin M, Wambaugh J, Knudsen T, Kancherla J, Mansouri K, Patlewicz G, Williams A, Little S, Crofton K, Thomas R. The ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem. Res. Toxicol. 2016. 29(8): 1225–1251. doi: 10.1021/acs.chemrestox.6b00135. [PubMed: 27367298]

[14]. Tice RR, Austin CP, Kavlock RJ, Bucher JR. Improving the human hazard characterization of chemicals: a Tox21 update. Environ Health Perspect. 121 (2013) 756–765. [PubMed: 23603828]

[15]. Yang C, Tarkhov A, Marusczyk J, Bienfait B, Gasteiger J, Kleinoeder T, Magdziarz T, Sacher O, Schwab CH, Schwoebel J, Terfloth L, Arvidson K, Richard A, Worth A, Rathman J. New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. J Chem Inf Model. 55 (2015) 510–528. doi: 10.1021/ci500667v. [PubMed: 25647539]

[16]. OECD. 2012. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Part 1: Scientific Evidence. Series on Testing and Assessment No. 168 ENV/JM/MONO(2012)10/PART1

[17]. OECD. 2015. Organisation for Economic Cooperation and Development. Test Guideline 442C: In Chemico, Skin Sensitisation, Direct Peptide Reactivity Assay (DPRA). http://www.oecd-ilibrary.org/environment/test-no-442c-in-chemicoskin-sensitisation_9789264229709-en

[18]. OECD. 2016. Organisation for Economic Cooperation and Development. Test Guideline 442D: In vitro Skin Sensitisation, ARE-Nrf2 Luciferase Test Method. DOI: 10.1787/9789264264359-en

[19]. OECD. 2015. Organisation for Economic Cooperation and Development. Test Guideline 442E: In vitro Skin Sensitisation, Human Cell Line Activation Test (h-CLAT). http://www.oecd-ilibrary.org/content/book/9789264229822-en

[20]. Williams A, Grulke C, Edwards J, McEachran A, Mansouri K, Baker N, Patlewicz G, Shah I, Wambaugh J, Judson R. The CompTox Chemistry Dashboard – A Community Data Resource for Environmental Chemistry. J. Cheminformatics. 9 (2017) 61 doi: 10.1186/s13321-017-0247-6

[21]. Mansouri K, Grulke CM, Richard AM, Judson RS, Williams AJ. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ Res. 27 (2016) 939–965. [PubMed: 27885862]

[22]. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 22 (2012) 276–282.

[23]. Wang J, Hallinger D, Murr A, Buckalew A, Lougee R, Richard AM, Laws S, Stoker T. High-Throughput Screening and Chemotype-Enrichment Analysis of ToxCast Phase II Chemicals Evaluated for Human Sodium-Iodide Symporter (NIS) Inhibition. Environ. Int. 126 (2019) 377–386. 10.1016/j.envint.2019.02.024 [PubMed: 30826616]

[24]. Roberts DW, Patlewicz GY, Kern PS, Gerberick GF, Kimber I, Dearman RJ, Ryan CA, Basketter DA, Aptula AO. Mechanistic Applicability Domain Classification of a Local Lymph Node Assay Dataset for Skin Sensitization. Chem. Res. Toxicol. 20(2007) 1019–1030. [PubMed: 17555332]

Highlights

- Three alert schemes for protein binding were evaluated
- Predicted reaction domains were compared with expert manual assignments for a published skin sensitisation dataset.
- Tox21 screening library was profiled by the 3 alert schemes to determine a ‘consensus’ reaction domain.
- Highly enriched ToxPrints were extracted for both datasets and example chemicals are highlighted.
- Enriched ToxPrints were compared with the ToxCast assay space to identify relevant assays informative for reactivity.

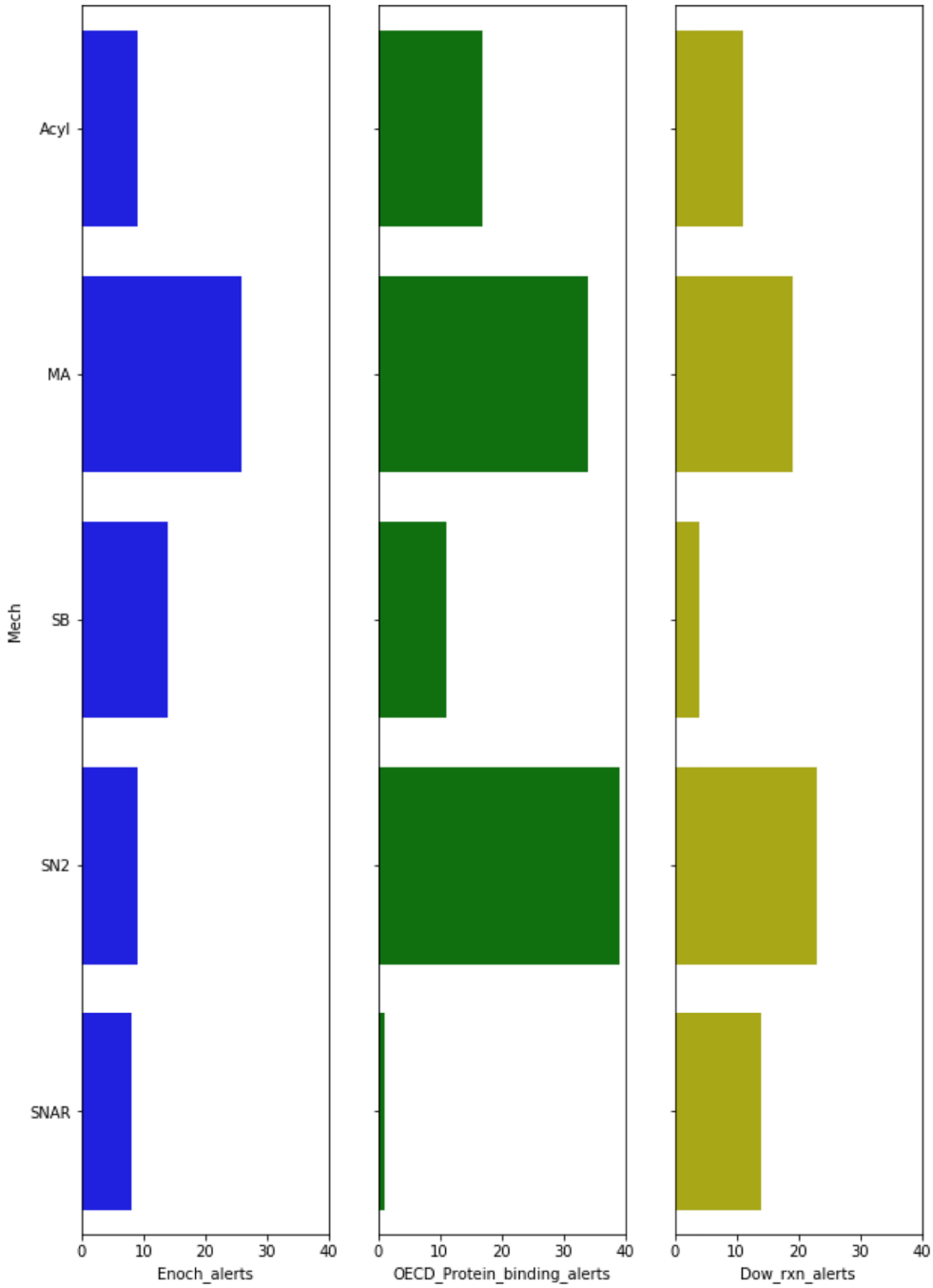


Figure 1.

Countplots of the different reaction schemes to summarise the number of alerts per reaction domain

Key: S_NAr = Substitution Nucleophilic Aromatic, S_N2 = Biomolecular nucleophilic substitution, SB = Schiff base formers, MA = Michael acceptors, Acyl = Acyl transfer agents

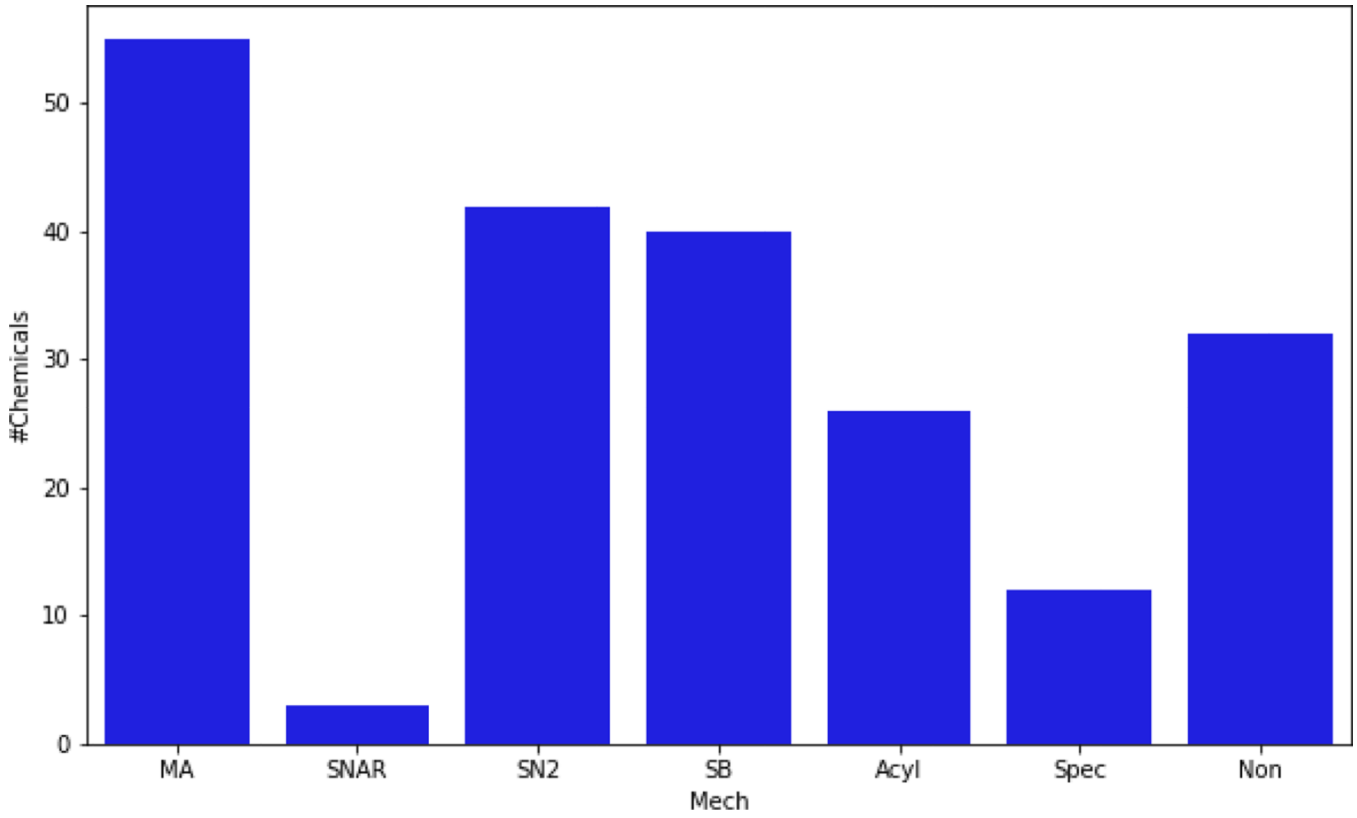


Figure 2.

Summary of mechanistic domain assignments from Roberts et al [24]

Key: S_NAr = Substitution Nucleophilic Aromatic, S_N2 = Biomolecular nucleophilic substitution, SB = Schiff base formers, MA = Michael acceptors, Acyl = Acyl transfer agents, Spec = Special cases, Non = Non-reactive

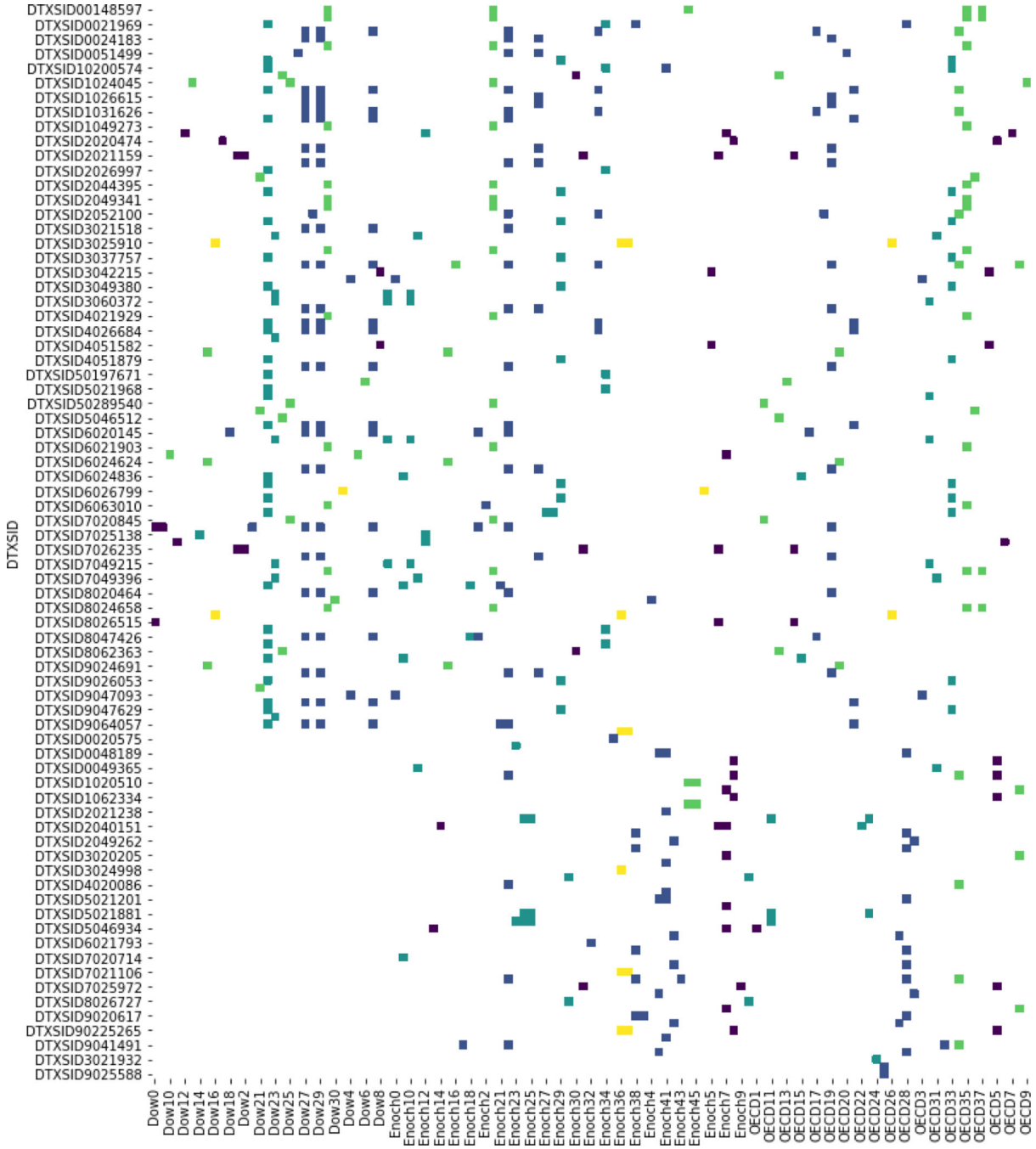


Figure 3:

Heatmap to illustrate which SMARTS patterns from each of the schemes were flagged for each substance and reflect this by the associated reaction domain, where purple = Acyl, dark blue = MA, dark green/blue = SB, green = SN2 and yellow = S_NAr. Substances that did not trigger any SMARTS or were associated with more than 1 domain are not shown.

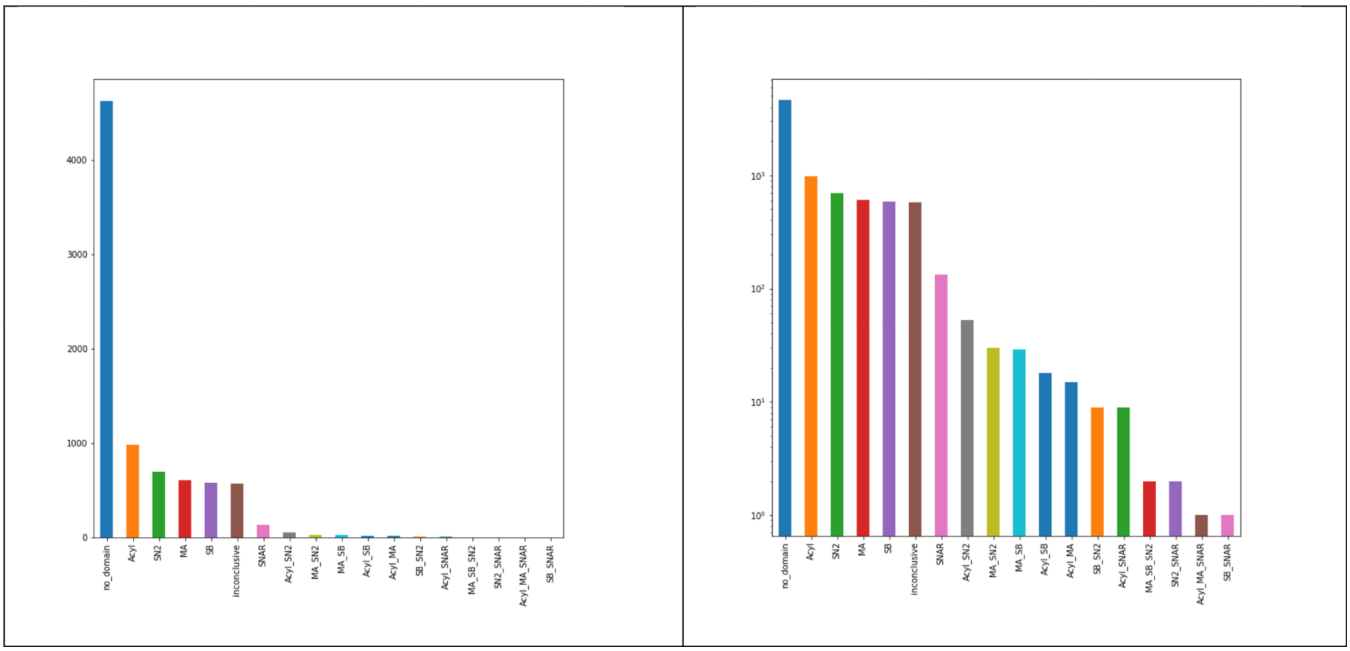


Figure 4.
Barplot of the frequencies across the reaction domains for the Tox21 library

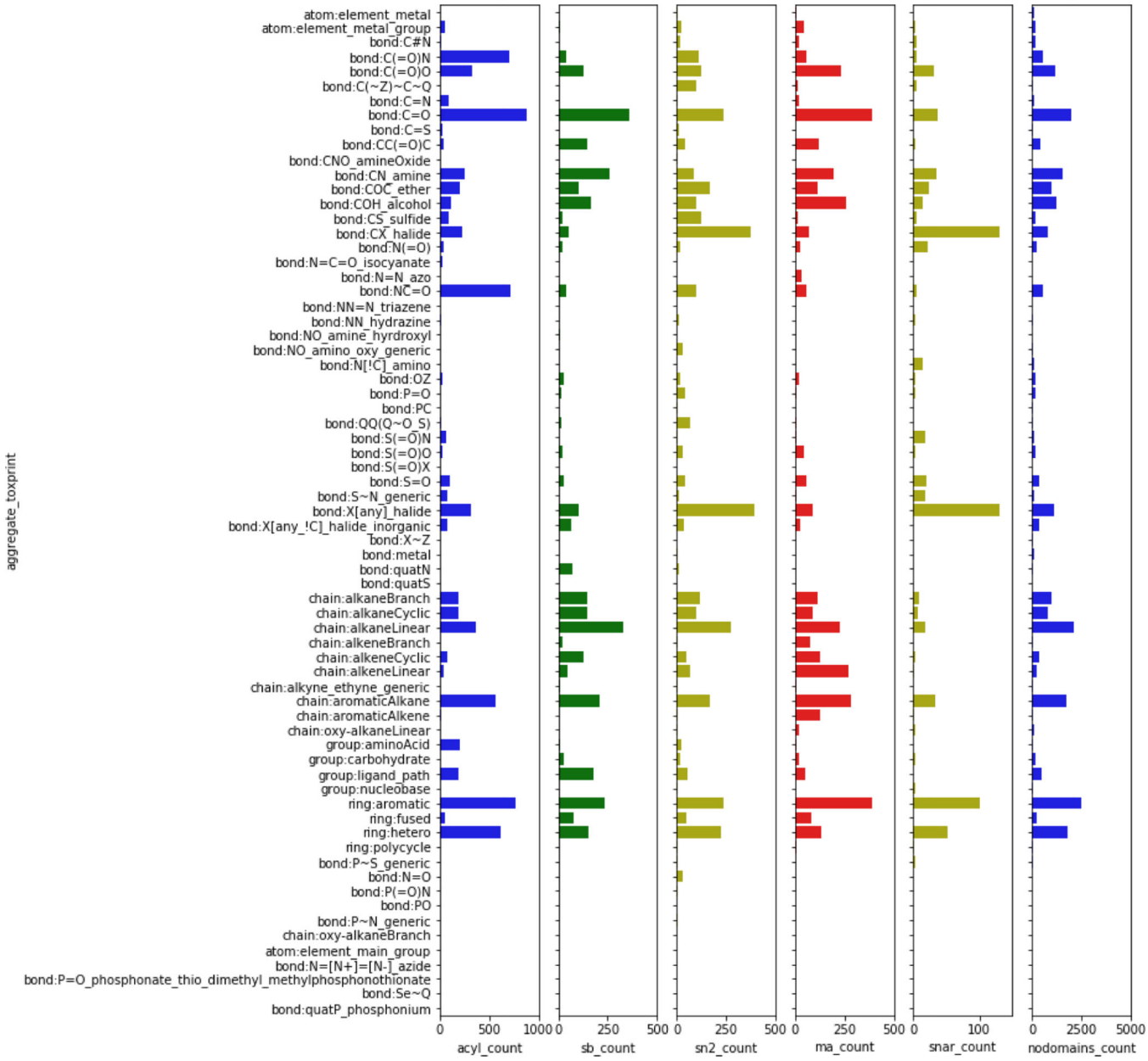


Figure 5.

Chemical fingerprint profile on a consensus reaction domain perspective

Notes: Aggregate_ToxPrint represents the Level 2 ToxPrints. The figure shows the counts of each of these Level 2 ToxPrints against each consensus reaction domain. No domains_counts highlights the Level 2 ToxPrints for chemicals that did not trigger any domain.

Table 1.

Data and Schemes considered in this analysis

Publication	Rulebase/Alerts	No of SMART patterns	Tag used in this publication	Source	Comments
Asturiol et al (2016) [5] Patlewicz et al (2016) [6]	Expert-assigned reaction domains		JRC dataset	Skin sensitisation dataset with assigned reaction domains as described in [6]	
Enoch et al (2008) [7]	SMARTS patterns for each of the 5 reaction domains plus 3 'pro' domains to denote alerts requiring activation either metabolically or chemically to produce an electrophilic species.	66	Enoch alerts	SMARTS patterns were taken from Table 2 of the original publication. Alerts derived from skin sensitisation datasets	Each of the pro domains was aggregated into their respective parent domain for practical reasons.
Wijeyesakere et al (2018) [10]	SMARTS patterns for each of the 5 reaction domains	71	Dow alerts	SMARTS extracted from Table S2 in the supplementary information of the original article. Alerts derived principally using *in chemico* data relevant for skin sensitisation	
Enoch et al (2011) [9]	The SMARTS patterns that are incorporated into the protein binding alerts by the OECD profiler comprise a designation of mechanistic domain, alert class, alert name.	102	OECD alerts	SMARTS provided by Dr Enoch. Alerts derived from literature review covering different endpoints including skin sensitisation, aquatic toxicity, skin irritation as well as data from the chromosomal aberration test.	

Table 2.

Summary Performance metrics

Relative to the Expert assignments	Precision (TP/(TP+FP))	Recall (TP rate)	F1 score	MCC
Enoch alerts	0.71	0.73	0.71	0.66
Dow alerts	0.68	0.62	0.59	0.547
OECD alerts	0.68	0.68	0.66	0.612

Table 3.

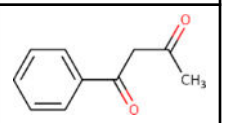
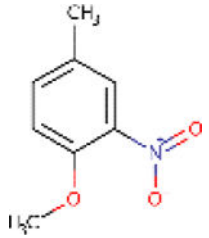
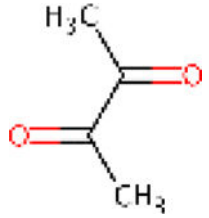
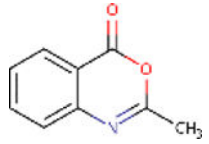
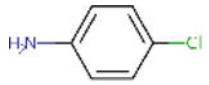
Micro performance characteristics (on a per reaction domain basis)

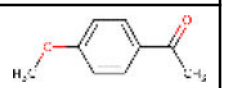
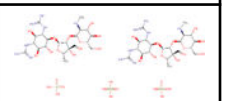
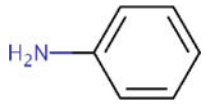
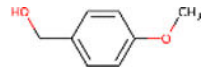
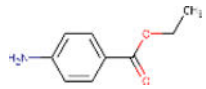
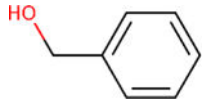
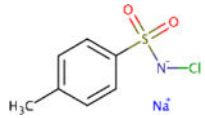
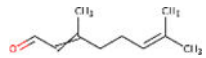
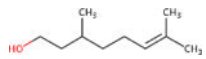
	#Chemicals per domain by Expert assignment in JRC dataset	Dow alerts				Enoch alerts				OECD alerts			
		#Alerts	Precision	Recall	F1 score	#Alerts	Precision	Recall	F1 score	# Alerts	Precision	Recall	F1 score
Acyl	14	11	75	43	55	9	61	79	69	17	69	64	67
MA	52	19	95	38	55	26	88	83	85	34	89	65	76
SB	34	4	74	68	71	14	76	76	76	11	71	59	65
S_N2	32	23	93	81	87	9	100	62	77	39	93	84	89
S_NAr	6	14	100	50	67	8	100	100	100	1	100	33	50

Notes: The remaining 84 chemicals were either not categorised by any domain (62 chemicals) or were assigned to more than 1 domain/special cases (22 chemicals)

Table 4.

List of substances where the profiling outcomes from all three schemes were “incorrect” relative to the expert assignment.

DTXSID	Chemical Name	Expert Mechanistic assignment	OECD	Enoch	Dow	Comments	Structure
DTXSID3021803	1-Benzoylacetone	[‘Acyl’]	[‘SB’]	[‘SB’]	[‘SB’]	Possible that acyl transfer could happen: Nu attacks at the aliphatic CO, PhCOCH2 anion leaves (and picks up H+). This is what happens when Nu is a hydroxide ion. Schiff base formation is expected. Misclassified in the original analysis. The profiling schemes are correct.	
DTXSID3029152	1-Methoxy-4-methyl-2-nitrobenzene	[‘MA’]	[]	[]	[]	Assigned as pro-MA. The OMe could be demethylated then oxidised to a quinone methide, or it could even be an SN2 Me transfer agent (phenoxide leaving group activated by ortho NO2) or it could be insufficiently (pro)reactive to cause a sensitisation response. In the LLNA, this was non-sensitising though 2 of the *in vitro* assays categorised this as positive and hence reactive.	
DTXSID6021583	2,3-Butanedione	[‘MA’, ‘SB’]	[‘SB’]	[‘SB’]	[‘SB’]	The expert MA categorisation is considered erroneous, Michael addition is not possible. The schemes have correctly assigned this substance.	
DTXSID0049313	2-Methyl-4H,3,1-benzoxazin-4-one	[‘Acyl’]	[]	[]	[]	Acyl -Ominus is enolic, so pKa is more like a phenol than an aliphatic alcohol, sufficiently good leaving group for acyl transfer.	
DTXSID9020295	4-Chloroaniline	[‘MA’, ‘SB’]	[]	[]	[]	Assigned as either pro or pre MA or pseudo SB pseudo-SB via NH2 -> N=O, or pre/pro-MA via ring oxidation. Sensitising in the LLNA.	

DTXSID	Chemical Name	Expert Mechanistic assignment	OECD	Enoch	Dow	Comments	Structure
DTXSID2044347	4’-Methoxyacetophenone	[‘SB’]	[]	[]	[]	Originally categorised as non-reactive SB, though SN2 Me transfer looks plausible. Non sensitising in the LLNA, active in the KeratinoSens assay.	
DTXSID7022047	Abietic acid	[‘pre’]	[]	[]	[]	Sensitising in the LLNA and other in vitro assays. Pre – activation by oxidation and likely by a free radical route rather than an electrophilic-nucleophilic scheme	
DTXSID8020090	Aniline	[‘MA’, ‘SB’]	[]	[]	[]	Annotated as Pre/pro-MA, pseudoSB. Sensitising in the LLNA and the h-CLAT but due to impurities	
DTXSID6044357	(4-Methoxyphenyl)methanol	[‘MA’]	[]	[]	[]	Originally annotated as pro/preMA Potential to be pro-SN2 – OH is biosulphated and becomes an SN2 LG. However, was only weakly sensitising in the LLNA.	
DTXSID8021804	Benzocaine	[‘SB’]	[]	[]	[]	Annotated as pro/pre-SB. Sensitising in the LLNA and positive in other KE assays	
DTXSID5020152	Benzyl alcohol	[‘SN2’]	[]	[]	[]	pro-SN2 via biosulphation non reactive and non sensitising in the LLNA	
DTXSID6040321	Chloramine-T	[‘Acyl’]	[]	[]	[]	Previously categorised as Acyl Could be SN2 attack at Cl	
DTXSID6024836	3,7-Dimethyl-2,6-octadienal	[‘MA’]	[‘SB’]	[‘SB’]	[‘SB’]	Nominally MA but hindered by Me to preferentially allow for SB formation.	
DTXSID3026726	Citronellol	[‘pre’]	[]	[]	[]	pre – autoxidation	

DTXSID	Chemical Name	Expert Mechanistic assignment	OECD	Enoch	Dow	Comments	Structure
DTXSID7029871	Clotrimazole	['Special_case']	☐	☐	☐	Special case – originally thought to be a SN1 reaction but could be a nucleophilic sensitiser	
DTXSID2046545	Diphenylcyclopropenone	['SB']	☐	☐	☐	Originally assigned as SB though MA is likely to be the initial step in a sequence involving ring opening. Needs revisiting especially given positive results in LLNA and DRPA	
DTXSID4038924	Hexyl salicylate	['Special_case']	☐	☐	☐	Probably a false positive assignment by the experts	
DTXSID5052414	2,4,6-Trimethylcyclohex-3-ene-1-methanol	['pre']	☐	☐	☐	pre	

DTXSID	Chemical Name	Expert Mechanistic assignment	OECD	Enoch	Dow	Comments	Structure
DTXSID2029612	Limonene	['pre']	☐	☐	☐	pre	
DTXSID7025502	Linalool	['pre']	☐	☐	☐	pre	
DTXSID8060618	N,N-Dibutylaniline	['SB']	☐	☐	☐	pro/pre-SB	
DTXSID9047596	5-Dodecanolide	['MA']	☐	☐	☐	Previously assigned as MA. Should have been annotated as unreactive	
DTXSID30200334	Squaric acid diethyl ester	['MA', 'SB']	☐	☐	☐	SB or MA but highly reactive as SB former	

DTXSID	Chemical Name	Expert Mechanistic assignment	OECD	Enoch	Dow	Comments	Structure
DTXSID9026053	Streptomycin sulfate (2:3)	[]	[‘SB’]	[‘SB’]	[‘SB’]	Non reactive in LLNA and DRPA	
DTXSID8049416	Tropolone	[‘MA’, ‘SN1’]	[]	[]	[]	MA (could be followed by loss of H20 to effectively give replacement of OH by RS) or SN1	

Notes: MA = Michael acceptor, SN2 = substitution nucleophilic bimolecular, Acyl = acyl transfer agent, SB = Schiff base former, SN1 = substitution nucleophilic unimolecular, SNAr = substitution nucleophilic aromatic, pre = pre-electrophilic (indirect acting, requiring chemical transformation), pro = pro-electrophilic (indirect acting, requiring enzymic transformation), empty brackets “[]” = either no expert reaction mechanism domain assigned (JRC) or no alert triggered in the scheme.

Table 5:

Top 5 enriched chemotypes per consensus-alert defined reaction domain for the Tox21 dataset

Reaction_domain	ToxPrint	TP	Odds Ratio	P-val
MA	bond:C#N_nitrile_ab-unsaturated	16	20.95	1.33643E-12
	ring:fused_[5_6]_indene	3	19.23	0.003416
	bond:N=N_azo_aromatic	30	17.46	4.42522E-21
	bond:C(=O)O_carboxylicEster_alkenyl	118	15.74	1.12978E-73
	chain:aromaticAlkene_Ph-C2_acyclic_generic	98	15.34	3.70467E-61
Acyl	bond:N=C=O_isocyanate_[O_S]	27	209.63	1.23981E-24
	bond:N=C=O_isocyanate_generic	18	138.44	2.46705E-16
	bond:C=O_acyl_halide	28	72.52	1.79837E-23
	bond:C(=O)O_acidAnhydride	26	50.39	7.49293E-21
	ring:hetero_[5_6]_N_isoindole_1-one	19	29.24	3.99817E-14
S_N2	bond:CX_halide_alkyl-X_ethyl	140	323.62	2.4999E-148
	bond:CX_halide_alkyl-X_aromatic_alkane	27	311.05	9.98123E-29
	bond:CX_halide_alkyl-X_aromatic_generic	27	311.05	9.98123E-29
	bond:CX_halide_alkyl-Cl_ethyl	95	304.65	9.3452E-100
	bond:CX_halide_alkyl-X_benzyl_generic	21	239.76	2.93727E-22
SB	bond:C=O_aldehyde_aromatic	6	41.78	4.45712E-05
	bond:C=O_aldehyde_generic	20	30.77	2.23394E-13
	chain:alkeneBranch_mono-ene_2-butene_2-propyl_(tiglate)	4	25.93	0.001856254
	chain:alkeneBranch_diene_2_6-octadiene	4	25.93	0.001856254
	bond:CC(=O)C_ketone_aromatic_aliphatic	3	9.35	0.0244
S_NAr	bond:CX_halide_aromatic-X_generic	131	856.24	1.0054E-110
	bond:CX_halide_aromatic-Cl_dichloro_pyridine_(1_2-)	12	822.1	1.8706E-21
	bond:CX_halide_aromatic-Cl_dichloro_pyridine_(1_4-)	12	411	1.29203E-20
	bond:CX_halide_aromatic-X_trihalo_benzene_(1_2_3-)	29	385.54	3.29585E-48
	bond:X[any]_halide	131	331.07	2.4174E-69

Table 6.

Top 5 chemotypes (or less if total is less than 5) per reaction domain from the JRC dataset

Reaction_domain	ToxPrint	TP	Odds Ratio	P-val
MA	bond:C(=O)O_carboxylicEster_alkenyl	11	24.58	2.39298E-07
	bond:C=O_carbonyl_ab-unsaturated_aliphatic_(michael_acceptors)	9	18.72	7.90222E-06
	chain:alkeneLinear_mono-ene_ethylene_generic	23	12.57	2.80398E-10
	chain:alkeneLinear_mono-ene_ethylene_terminal	7	10.16	0.000505946
	bond:C(=O)O_carboxylicEster_acyclic	13	9.05	5.23066E-06
	chain:alkeneLinear_diene_1_2-butene	9	7.85	0.000241917
Acyl	ring:hetero_[5]_Z_1-Z	5	49.29	9.51501E-06
	ring:hetero_[5]_O_oxolane	5	49.29	9.51501E-06
	ring:hetero_[5_6]_Z_generic	5	36.79	2.08278E-05
	chain:aromaticAlkane_Ph-C1_cyclic	5	24.29	7.22608E-05
S_N2	chain:alkaneLinear_hexadecyl_C16	3	26.73	4.72881E-03
	bond:CS_sulfide	6	20.42	7.56699E-05
	chain:alkaneLinear_dodedyl_C12	6	15.24	0.000174814
	bond:X[any]_halide	13	9.59	2.66263E-06
	chain:alkaneLinear_decyl_C10	6	8.57	0.00112602
SB	bond:C=O_aldehyde_aromatic	6	41.78	4.45712E-05
	bond:C=O_aldehyde_generic	20	30.77	2.23394E-13
	chain:alkeneBranch_mono-ene_2-butene_2-propyl_(tiglate)	4	25.93	0.001856254
	chain:alkeneBranch_diene_2_6-octadiene	4	25.93	0.001856254
	bond:C=O_carbonyl_1_2-di	5	8.26	0.004416916

Notes: The S_NAr domain did not have any ToxPrints that were enriched based on the criteria of Odds Ratio (OR)=> 3, TP>=3 and p =< 0.05

Table 7.

Number of enriched ToxPrints per reaction domain

Reaction domain	# Enriched ToxPrints unique to the JRC dataset	# Enriched ToxPrints unique to the Tox21 dataset	#Enriched ToxPrints common to both datasets
MA	1	20	13
Acyl	5	57	1
S_N2	8	58	2
S_NAr	0	30	0
SB	0	0	13

Table 8.

Illustrative chemicals with common MA enriched chemotypes

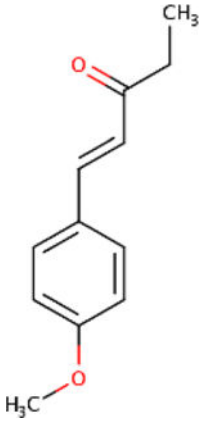
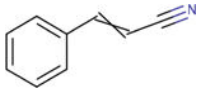
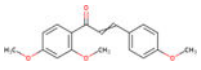
DTXSID	Structure	Enriched_ToxPrints
DTXSID00231670		bond:C=O_carbonyl_ab-unsaturated_aliphatic_(michael_acceptors) bond:C=O_carbonyl_ab-unsaturated_generic bond:CC(=O)C_ketone_alkene_generic chain:alkeneLinear_diene_1_2-butene chain:alkeneLinear_mono-ene_allyl chain:alkeneLinear_mono-ene_ethylene chain:alkeneLinear_mono-ene_ethylene_generic chain:aromaticAlkene_Ph-C2_acyclic_generic chain:aromaticAlkene_Ph-C2
DTXSID2052100		chain:alkeneLinear_diene_1_2-butene chain:alkeneLinear_mono-ene_allyl chain:alkeneLinear_mono-ene_ethylene chain:alkeneLinear_mono-ene_ethylene_generic chain:aromaticAlkene_Ph-C2_acyclic_generic chain:aromaticAlkene_Ph-C2
DTXSID9046152		bond:C=O_carbonyl_ab-unsaturated_aliphatic_(michael_acceptors) bond:C=O_carbonyl_ab-unsaturated_generic bond:CC(=O)C_ketone_alkene_generic chain:alkeneLinear_diene_1_2-butene chain:alkeneLinear_mono-ene_allyl chain:alkeneLinear_mono-ene_ethylene chain:alkeneLinear_mono-ene_ethylene_generic chain:aromaticAlkene_Ph-C2_acyclic_generic chain:aromaticAlkene_Ph-C2

Table 9.

Prototypic chemicals with enriched MA chemotypes specific to the Tox21 library

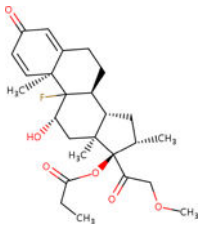
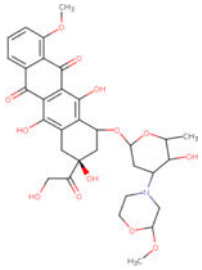
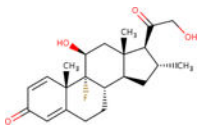
DTXSID	Structure	Enriched_ToxPrints
DTXSID4057631		bond:CC(=O)C_ketone_alkene_cyclic_2-en-1-one chain:alkeneCyclic_ethene_C_(connect_noZ)
DTXSID6057619		bond:CC(=O)C_ketone_alkene_cyclic_2-en-1-one bond:CC(=O)C_quinone_1_4-benzo bond:CC(=O)C_quinone_1_4-naphtho bond:COH_alcohol_aromatic
DTXSID3045647		bond:CC(=O)C_ketone_alkene_cyclic_2-en-1-one chain:alkeneCyclic_ethene_C_(connect_noZ)

Table 10.

Enriched ToxPrints projected on the ToxCast assay space

Vendor	# Assays	#ToxPrints represented out of 14
APR	37	6
ATG	155	13
TOX21	98	11
BSK	238	14
NVS	82	11
OT	41	9
CEETOX	12	6
Tanquay	6	3
CLD	12	6