A cheminformatics workflow to select representative TSCA chemicals for New Approach Methodology (NAM) screening

Grace Patlewicz[a,*], Antony Williams[a], Matthew Adams[a,b], Imran Shah[a], Katie Paul-Friedman[a]

[a]Center for Computational Toxicology and Exposure (CCTE), US Environmental Protection Agency, 109 TW Alexander Dr, Durham, 27709 USA

[b]Oak Ridge Associated Universities (ORAU), Oak Ridge, 37831 USA

Abstract

The Toxic Substances Control Act (TSCA) requires the US EPA to evaluate the hazard and exposure of new and existing chemicals. New chemical notifications are typically data poor and EPA's Office of Pollution Prevention and Toxics (OPPT) has historically relied upon the use of tools including chemical categories and read-across approaches to fill data gaps. As part of a multi-year New Chemicals Collaborative Research Program between OPPT and EPA's Office of Research & Development (ORD), opportunities are being explored to leverage New Approach Methods (NAMs) in hazard and exposure assessments. In vitro NAMs will be used to generate mechanistic, hazard, and toxicokinetic data that will be evaluated for their applicability to new chemical assessments. Herein, we describe the cheminformatics workflow developed to identify a set of approximately 300 representative candidate case study chemicals for screening from the TSCA non-confidential active inventory. Chemicals were first categorised using EPA's New Chemicals Categories (NCC) to evaluate their scope, coverage, and potential gaps. Approximately half of the non-confidential active inventory with discrete structure representations could be assigned at least one NCC. Given this limitation, chemical classification information was instead generated from the freely available web application ClassyFire to categorise all discrete organic structures from the TSCA inventory into one of 68 primary categories. Large primary categories were subcategorised into smaller categories using hierarchical clustering. A total of 180 structural terminal categories were derived. The inventory was then profiled in a number of different ways to create filters that could aid in the selection of candidate substances from each terminal category. Informative physicochemical property predictions from OPERA QSAR models, such as melting point, boiling point, LogP, and vapour pressure were used to indicate likely physical form and volatility. PubChem was queried to identify the feasibility of finding a vendor from which a given substance could be procured. Analytical method detection amenability predictions for liquid-chromatography mass spectrometry were also generated to provide an indication of which chemicals lent themselves to aqueous based-screening. Structural rules based on expert judgement were used to identify potential explosive or highly reactive substances. Potential candidate substances were selected on the basis of being structurally representative of the terminal category and meeting the other screenability conditions. A final set of 318 chemicals were proposed which will undergo analytical quality control and screening in a range of broad and targeted biological technologies for human health relevant endpoints. As a second phase of this study, the candidate substances were also compared with the broader TSCA active landscape through the lens of different in silico tools to explore their predicted hazard profiles.

Keywords: TSCA, categorisation, structural categories, QSAR

1. Introduction

Under the Toxic Substances Control Act (TSCA), as amended by the Frank R. Lautenberg Chemical Safety for the 21st Century Act[1], the US Environmental Protection Agency (EPA) evaluates potential risks from new and existing chemicals and acts to address any unreasonable risks that chemicals may have on human health and the environment. The non-confidential TSCA Inventory contains 86,741 chemicals, of which 42,293 are active (at the time of writing February 2024) in commerce which renders the use of traditional approaches too resource and time intensive to generate relevant data to facilitate assessment. In 2022, EPA launched a new effort under TSCA to modernise the process and bring innovative science to the review of new chemicals before they are permitted to enter the marketplace. This included a multi-year collaborative research programme (known as the New Chemical Collaborative Research Program (NCCRP)) to refine existing approaches as well as develop and implement New Approach Methods (NAMs) to ensure the best available science could be applied in new chemical evaluations. Several key areas were proposed to: 1) refine read-across approaches; 2) digitise and consolidate existing information on chemicals and combine them with publicly available sources to broaden coverage and accessibility; 3) update Quantitative Structure Activity Relationship (QSAR) models for physical and (eco)toxicological properties; 4) explore ways of integrating and applying NAMs in new chemical assessments thereby reducing the reliance on animal testing and 5) develop decision support tools to integrate various data streams to facilitate new chemical risk assessment[2]. In terms of actualising how in vitro NAMs would be integrated and applied, a selection of different technologies were identified for evaluation as part of a proof-of-concept study. These comprise a battery of Tier 1 broad profiling and targeted Tier 2 NAMs as described in more detail in the 2019 Computational Toxicology bluePrint[3]. Broad profiling NAMs encompass high-throughput phenotypic profiling (HTPP)[4] technologies which measure a large number of cellular morphological features that can be used to inform point of departure, molecular initiating events and bioactivity. Targeted Tier 2 NAMs include high-throughput transcriptomic (HTTr)[5] and high-throughput screening (HTS) assays (e.g. such as ToxCast[6,7]). Collectively, these technologies have the potential to fill gaps related to molecular initiating events such as nuclear receptor targets, genotoxicity, oxidative stress. They can also be tailored to different routes of exposure or address specific toxicities that are more closely aligned with traditional adverse outcomes such as developmental toxicity. In addition to technologies to characterise biological activity, High Throughput Toxicokinetic (HTTK) approaches[8] to assess dosimetry and kinetics also form an important component of informing point of departure estimations. Herein, a cheminformatic workflow was developed to identify a set of approximately

*Corresponding author
Email address: patlewicz.grace@epa.gov (Grace Patlewicz)

300 representative candidate case study chemicals from the TSCA non-confidential active inventory, including some reference chemicals[9] in an effort to increase scientific confidence in the application of NAMs for informing chemical safety. These data will be used to evaluate performance of the selected NAMs for further application and may also inform evolving frameworks for using multiple data streams to perform bioactivity-based dose-response assessment and hazard identification.The aims of this study are therefore:

- Construct a structure-based version of the TSCA active non-confidential inventory (herein referred to as the 'TSCA landscape')
- Profile the TSCA landscape on the basis of the EPA New Chemical Categories and the ClassyFire ontology[10] to assign substances into broad structural categories
- Develop a categorisation scheme to assign substances into structural categories that maximise 'within structural similarity' whilst maintaining a total number of terminal categories that would facilitate candidate substance selection within a certain scope of resources
- Identify representative substances from the TSCA landscape for NAM screening taking into account structural considerations and technical constraints
- Evaluate the TSCA landscape in terms of selected predicted physicochemical and toxicity considerations to help prioritise structural categories based on their anticipated hazards of concern

2. Materials and Methods

2.1. Dataset

The list of ~32,000 TSCA active non-confidential inventory substances that had been registered on the EPA CompTox Chemicals Dashboard (August 2022)[11] was downloaded for subsequent processing and analysis. This was filtered to only retain substances for which a structure was available, mixtures and substances of unknown or variable composition, complex reaction products or of biological materials (UVCBs) were excluded from consideration. There were 14,247 substances with structural information. The dataset was processed using the Python package RDKit[12] to convert the simplified molecular-input line-entry system (SMILES) into molecular objects so that a structure-data-file (SDF) could be generated for input into the OPEn structure-activity Relationship App (OPERA) expert system[13] for physicochemical property prediction and QSAR-Ready SMILES derivation[14]. The latter are desalted, stereochemistry stripped versions of SMILES that are more amenable to QSAR predictions. Of the 14,247 substances processed, 51 substances could not be rendered into molecular objects by RDKit.

2.2. Categorisation Workflow

Using the dataset constructed, a categorisation workflow was then developed to inform the selection of approximately 300 substances that would be nominally representative of the TSCA landscape from a structural perspective but would also satisfy two practical constraints namely the substances could be readily procured from EPA's chemical supplier, Evotec and be testable in the Tier 1 and 2 NAM screening systems. The approach would rely on firstly assigning substances into predefined categories known as primary categories which were subsequently subset into smaller categories (secondary categories) using hierarchical clustering approaches. Secondary categories were expected to be more structurally similar on the basis of their chemical structural representation than the initial primary categories. Techniques to identify representative substances taking into account structural similarity as well as other practical/technical constraints were considered. Figure 1 shows the conceptual workflow developed that shaped the case study.

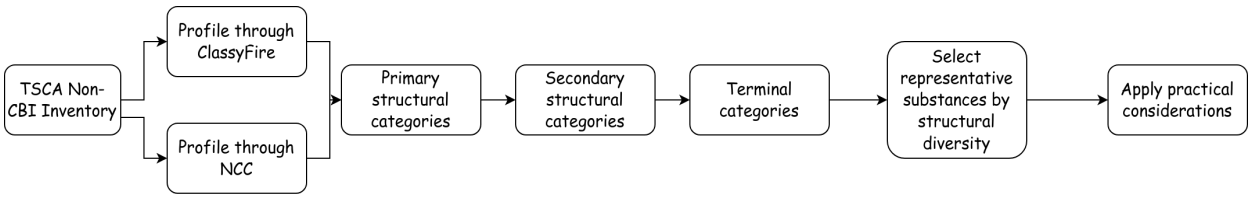


Figure 1: Categorisation workflow

2.3. Profiling Substances into Categories

A pragmatic number of final categories was needed to capture the breadth and diversity of the TSCA landscape, yet facilitate selection of ~300 representative substances. To that end, primary categories were first explored making use of 2 approaches - the EPA New Chemical categories (NCC) and the ClassyFire ontology[10].

2.3.1. Primary Categories

EPA New Chemical Categories (NCC)

Under section 5 of the Toxic Substances Control Act (TSCA), EPA's New Chemicals programme helps manage potential risk to human health and the environment from chemicals new to the marketplace. For the purposes of regulation under TSCA, a chemical that is on the TSCA Inventory is considered "existing" as in a substance is known to be in US commerce. Any substance that is not on the Inventory is considered "new". For a new chemical that will be imported or manufactured in the US, the manufacturer/importer is required to file a notice to EPA, known as a premanufacture notice (PMN). EPA then undertakes an assessment of that chemical to determine whether it poses a risk based on its hazard and exposure profile. One of the resources EPA makes use of during that assessment are the New

Chemical Categories (NCC). These are a set of categories whereby substances with shared chemical and toxicological properties have been grouped together. Many of the categories are defined based on their structural characteristics such as functional groups and physicochemical parameters. In certain cases, the categories may provide indications of what ecotoxicological or toxicological hazards might be associated with that category thereby providing recommendations of the types of data that would be informative for an assessment. There are currently 56 categories which were last updated in August 2010 (https://www.epa.gov/reviewing-new-chemicals-under-toxic-substances-control-act-tsca/chemical-categories-used-review-new) and are described in narrative form in a pdf document available for download. As part of this case study, the 56 categories were first re-implemented into a machine readable format making use of structure query syntax (as Simplified molecular-input-entry-system (SMILES) arbitrary target specification (SMARTS)) to facilitate profiling of substances into one or more of the categories. The xml file of the NCC encoded in version 4.5 of the OECD Toolbox[15] was exported and formed the foundation of creating an independent ruleset for the NCC which was augmented for certain categories and coupled with predictions of physicochemical parameters to provide a means of profiling the TSCA landscape into their respective NCCs.

ClassyFire

All substances were also assigned into classes using the structure-based chemical taxonomy developed by the Wishart laboratory[10]. ClassyFire assigns chemicals into a taxonomy consisting of >4800 different categories. The taxonomy comprises 11 different levels such as Kingdom, Super-Class, Class, SubClass etc. The webserver, accessible at(http://classyfire.wishartlab.com/), was used to query each substance by its hashed International Chemical Identifier (InChIKey)[16] and assign it into Kingdom-SubClass levels. For the 14,247 substances - 13,477 were assigned into Kingdom 'Organic compounds', 593 into Kingdom 'Inorganic' and the remaining 179 returned no information. These 179 were assigned into an arbitary class of "Other" to facilitate ongoing processing.

2.3.2. Secondary and Terminal categorisation

Since structural diversity was likely to be high within a primary category – an approach was needed to balance maximising 'within category' structural similarity relative to the total number of terminal categories. Hierarchical clustering (using Ward's criterion[17]) on the basis of Morgan chemical fingerprints[18] (using radius 3 and bitvector length 1024) was applied to those primary categories where membership exceeded 65 substances. The first generation of clusters was taken as the secondary category. Sixty-five was an arbitrary but pragmatic threshold based on the distribution of membership size. Thirty-nine primary categories had a membership size less than 65. Primary categories whose membership was less than 65 were termed terminal categories. Primary categories containing

more than 65 membership underwent further subcategorisation into secondary categories by applying hierarchical clustering. These secondary categories were hence termed terminal categories.

2.4. Representative Substance Identification

The nominally representative substance for a given terminal category was taken as the medoid. This was defined as the substance with the minimum pairwise distance from all other members of that category. This was used as an initial seed to then identify and additional set (up to 5) of structurally diverse substances within the category on the basis of their Morgan chemical fingerprints. The approach used in this case was the MaxMin procedure[19] within RDKit. As described in the RDKit documentation, the algorithm works as follows: 1. Generate chemical features (e.g. Morgan chemical fingerprints) for all the substances, both any initial seeds (in this case, the medoid computed) plus those to pick from (the candidate pool). 2. From the molecules in the candidate pool, find the one that has the maximum value for its minimum distance to substances in the picked set (hence the MaxMin name), calculating and recording the distances as required. This substance is the most distant one to those already picked so is transferred to the picked set. 3. Iterate back to step 2 until the number of substances required is met. The MaxMin procedure was performed for all terminal categories containing at least 5 or more members.

2.5. OPERA generated physicochemical parameters

The OPERA expert system[13] version 2.8 was used to generate a range of physicochemical properties predictions for all substances in the dataset. LogP (octanol water partition coefficient), melting point, boiling point and vapour pressure predictions were generated since these would provide a means of determining physical form at room temperature and volatility (both considerations for technical testing potential). Substances were standardised into QSAR READY format so that stereochemistry, salt forms etc. were removed from the substances under consideration. Accordingly some substances could not be converted into QSAR READY format. Merging OPERA predictions with the starting dataset found that 946 substances were not associated with any OPERA prediction either because substances were mixtures (6), were too large in size to be processed (12), could not be converted into a molecular objective by RDKit (51) or were inorganic (588) or organometallic (289) in nature. Predictions were possible for up to 13,299 substances from the starting dataset which were carried forward for processing and analysis.

2.6. Creation of a Constrained Landscape

The landscape was constrained to only consider substances that were potentially screenable in the NAM suite. Screenability was based on 3 main considerations: 1) (physico)chemical characteristics,

2) procurability and 3) out of scope due to existing NAM data. The physicochemical characteristics captured physical state at room temperature and pressure. These were inferred based on the OPERA predictions generated (see Section 2.5). Melting and boiling point thresholds as outlined in the US EPA Sustainable Futures Programme (https://www.epa.gov/sites/default/files/2015-05/documents/05-iad_discretes_june2013.pdf) was used to determine whether a substance was likely to be a solid, liquid or gas at room temperature and pressure. Volatility was modelled using vapour pressure (VP). If a substance had a logVP less than 2 mmHg, then it was considered screenable. The Lipinski rule of 5[20] uses simple heuristics to denote oral absorption namely LogP <5, MW <500 and Hydrogen bond donor or acceptors (< 10 or <5). If any of these criteria are met, this is referred to as a Lipinski failure. For the purposes of this study, if the number of Lipinski failures was less than 3, than a substance was considered screenable. Certain elements could also factor problematic for chemical procurement and/or testing. Only substances containing the following elements, C, H, N, O, P, S, Halogens and Si when not adjacent to an O were tagged for consideration. Procurability was estimated based on the number of vendors listed in [PubChem] (https://pubchem.ncbi.nlm.nih.gov/) for a given substance. The vendor count needed to be at greater than 1. Out of scope referred to the substance having already been tested in the ToxCast screening programme. Considering all these aspects, the inventory was filtered to only retain substances that were tagged as screenable. For this set of substances, medoids and MaxMin substances were also identified using the same approach as already described in Section 2.4.

2.7. Other considerations

Aside from selection of representative substances from the full and constrained inventory on the basis of the outlined considerations, several other aspects were also used to inform the final selection process. Analytical method detection amenability predictions for liquid-chromatography mass spectrometry (LM-MS) were generated using the QSAR model developed by Lowe et al.[21] to provide an indication of which chemicals lent themselves to aqueous-based screening. Structural considerations based on lists of substances on the Dashboard known to be potentially explosive or highly reactive were also tagged. Category membership size was another consideration - size bins were set at less than 20 members, between 20-70, between 70-150, between 150-300 and finally between 300-600. A final manual check was performed to review all proposed candidate substances for testing following application of all the aforementioned considerations.

2.8. Evaluation of Selected Candidate Substances

2.8.1. Physicochemical Comparison

t-distributed schocastic neighbourhood mapping[22] was performed using Morgan fingerprints computed for all the chemicals in the TSCA landscape. This was plotted as a scatterplot and colour coded by 1) screenability considerations and 2) the selection of candidate substances. The intention was to visually compare how representative the screenable and/or candidate selections were from a structural perspective relative to the TSCA landscape. In addition, 2D boxplots were created to compare the distribution of selected predicted physicochemical parameters of the TSCA landscape chemicals relative to that of the selected candidates.

2.8.2. Predicted Structural Alert Profile

The entire set of TSCA landscape chemicals were profiled through the Derek Nexus 2.5 system to generate structural alerts using default settings. This would captures the following endpoints including: carcinogenicity, genotoxicity, irritation, neurotoxicity, organ toxicity, reproductive toxicity, respiratory sensitisation, skin sensitisation and miscellaneous endpoints (such as chloracne, methaemogloneaemia, blood in urine). The output was processed and aggregated so that a structural alert fingerprint representation was constructed indicating presence or absence of that alert for a given substance. The fingerprint representations were explored through the lens of the terminal categories 1) to visually evaluate the consistency in the alert profiles across a category and 2) to identify what endpoints were overrepresentated in specific categories.

2.8.3. Predicted Toxicity Profile

Predictions of the entire landscape were made using the Toxicity Estimation Tool (TEST) version 5.1.2 (https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test) for the developmental toxicity and Ames mutagenicity endpoints. Projections onto 2D using t-SNE were used to explore the extent to which TEST predictions were consistent throughout a terminal category. Predictions of rat oral acute toxicity were also generated using OPERA specifically the CATMOS module[23]. Comparisons of the distributions of the potencies across the TSCA landscape relative to the candidate substances were made. An evaluation to identify which terminal categories were associated with most and least potent LD50 values was also undertaken.

2.8.4. Predicted ToxCast Profile

Using a deep learning graph attention mechanism network developed in Adams et al, in preparation, predictions were made for the TSCA landscape to profile the likely NAM profile with respect to selected ToxCast assays. The selected assays focused on one vendor, Attagene (ATG) which provides

insights into transcription factor activity in transformed HepG2 cells. A multi-task graph based approach as implemented within the DeepChem python library was used to develop the model that predicted the probability of a substance resulting in a positive hitcall.

2.8.5. Availability of in vivo toxicity data within terminal categories

In vivo oral data (non-cancer and reproductive/developmental) taken from Aurisano et al[24] were used to assess the availability of points of departure for the full landscape and the candidate set as well as the range of potencies. Potencies were also compared across different terminal categories.

3. Data Analysis and Code Availability

Data processing was conducted using the Anaconda distribution of Python 3.9 and associated libraries, NumPy[25], Pandas[26], Scikit-learn[27], Matplotlib[28] and Seaborn[29]. Jupyter Notebooks[30], scripts and datasets will be made available on github at XXXX and on Figshare at XXXX.

4. Results and Discussion

4.1. Primary and Secondary Structural categories

The TSCA landscape downloaded comprised 14,247 substances. Of this set, 51 substances could not be rendered into a molecular object by RDKit. OPERA predictions were possible for up to 13,299 substances from the starting dataset. 946 substances were not associated with any OPERA prediction either because substances were mixtures (6), were too large in size to be processed (greater than 100 carbon atoms which prevents descriptor generation within OPERA) (12), were not converted into a molecular objective by RDKit (51), had no SMILES (2), were inorganic (588) or organometallic (289) in nature.

Using the OPERA physicochemical predictions, substances were assigned into their most likely physical form. Based on the thresholds defined in the EPA Sustainable Futures guidance, 1029 substances (7%) could not be assigned a physical form (in the majority of cases, this was on account of no predicted melting point or boiling point prediction available, although for 68 cases this was because the criteria did not account for substances that might have a predicted value of exactly 25 deg C). For the substances that could be assigned, 54% (7690) were solids, 28% (5407) were liquids and 0.8% (121) were gases. Note if the thresholds were adjusted, those 68 substances would have been categorised as liquids. The distribution of the physical form suggested that only a very small number of substances within the TSCA landscape would merit consideration of an inhalation route of exposure.

EPA New Chemical Categories (NCC)

The dataset was processed through the NCC in conjunction with the physicochemical property information from OPERA to make category assignments. Over 46% of substances were not assigned to any NCC. Although some 948 could not be processed due to lack of QSAR ready SMILES, the primary reason for the non assignment was the lack of a relevant existing category. The next largest NCC assignment was neutral organics (18.7%), followed by esters, phenols and anilines. Substances could be assigned to more than 1 NCC e.g. Esters and Substituted Triazines. There were 141 unique combinations of NCC of which 99 comprised 2 or more NCC and the remaining 43 were single NCC. The first 50 NCC with their respective count are shown in Figure 2. The large number of substances that could not be assigned into a category represents a gap in coverage of the NCC for the TSCA landscape and presents an opportunity for new categories to be derived. This is the subject of an ongoing but related study.

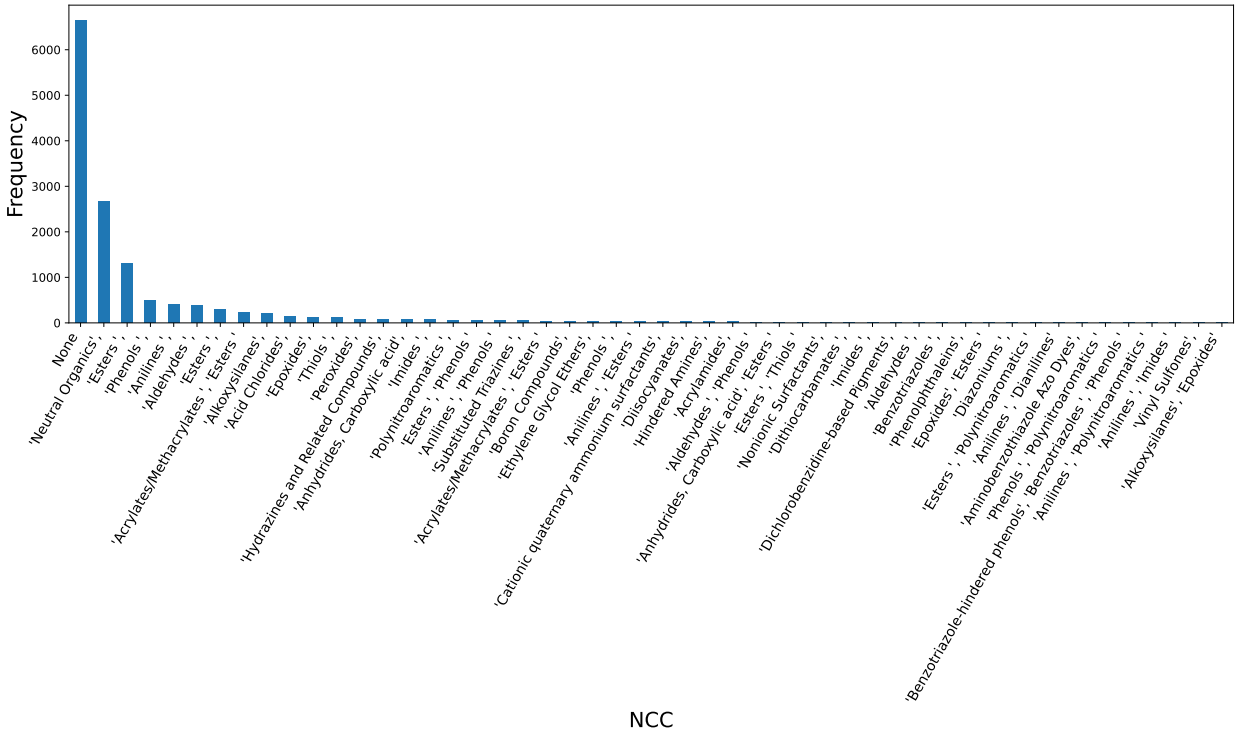


Figure 2: Frequency of NCC membership, first 50 categories shown

ClassyFire primary categories

Since almost 47% of substances could not be assigned to a NCC, the ClassyFire ontology was used to assign substances into primary categories based on its chemical taxonomy. Substances that returned no taxonomy information was assigned to an arbitrary "Other" class. The majority of substances (13477) were assigned to the Organic compounds Kingdom. At the next level of granularity,

Superclass was associated with 27 categories whereas the level down Class was associated with 270 unique categories. Given the objective of identifying approximately 300 candidates for testing, using "Class' as a basis for the primary category (per Figure 1) would rapidly exceed the number of practical categories from which to draw from. Exploring the frequency of the Superclass designation found that whilst there were 27 unique categories, 5 had memberships exceeding 1000. A hybrid approach of using Superclass as the primary category designation was taken for Superclasses that contained fewer than 1000 members whereas the 5 Superclasses within memberships exceeding 1000 (Benzenoids, Lipids and Lipid-like, Organoheterocyclic compounds, Organic acids and derivatives, Organic oxygen compounds) were expanded into their respective Class assignment. This resulted in 68 unique ClassyFire categories that formed the basis of the primary categories. The frequency of the TSCA landscape across these ClassyFire primary categories is shown in Figure 3.

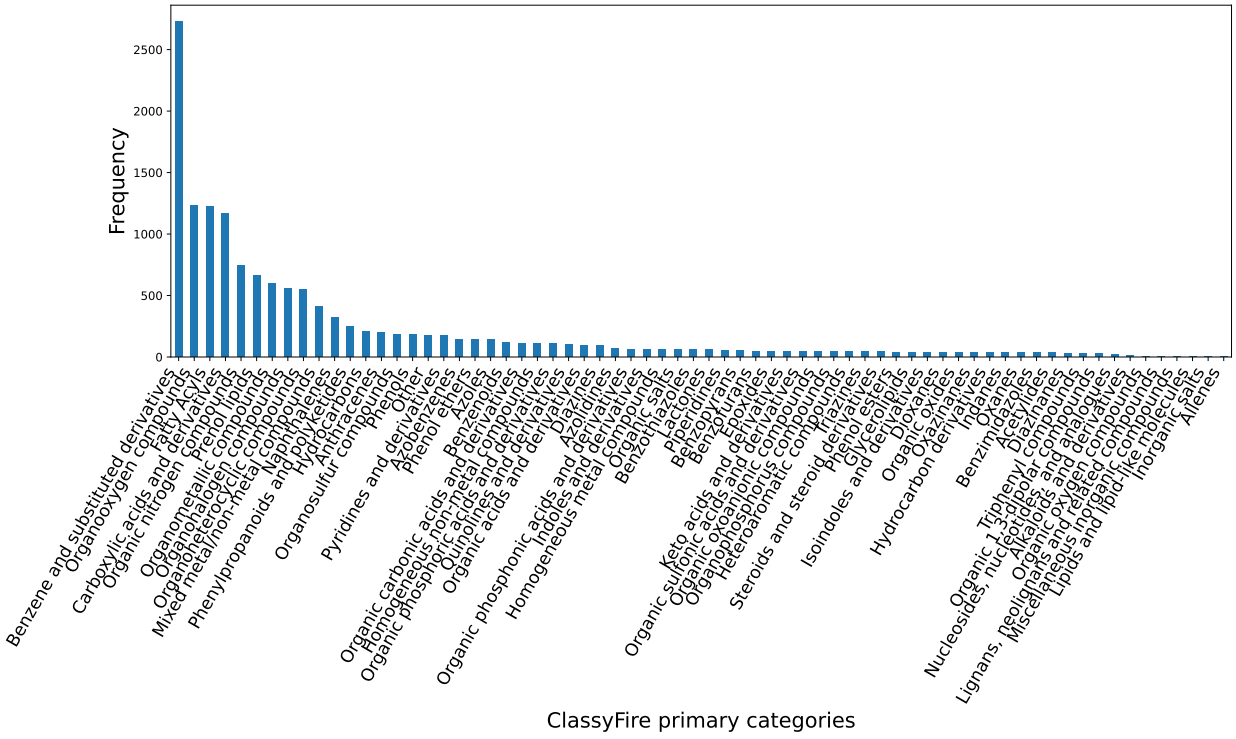


Figure 3: Frequency of ClassyFire primary categories for the TSCA landscape

Inspection of the substances without NCC assignment found that these were most associated with ClassyFire Classes including Benzene, Organonitrogen compounds, Fatty Acyls, Carboxylic acids, Organooxygen, organometalloid and naphthalenes.

4.2. Secondary-terminal categorisation

Medoids were first calculated for each of the 68 primary categories. The pairwise Jaccard distance between the medoid and the remaining members of the category were also computed. This was generated as a potential consideration for candidate chemical selection later - conceivably the medoid might not be feasible to test due to procurability issues but a close match on the basis of Jaccard distance might be a practical alternative. For primary categories containing less than 65 members, no further refinements were made - the primary category was effectively named as the terminal category. For primary categories with greater than 65 members of which there were 29 such categories, a hierarchical clustering was performed using Morgan fingerprints as inputs. The first generation of clusters were extracted for each primary category based on visual inspection of the dendrogram. These were arbitrarily tagged by 1,2,3 such that the secondary category was named using the primary category root + the cluster designation. For the 29 primary categories that contained more than 65 members, using the clustering approach resulted in an expansion to 141 secondary categories. Together with the 39 primary categories that contained less than 65 members, a total of 180 terminal categories were generated. Figure 4 shows the 37 terminal categories that have a membership exceeding 100 substances.

4.3. Diverse substance selection: TSCA landscape

Whilst medoids were selected as the most representative substance from each terminal category, a single chemical was unlikely to capture the breadth of diversity within a category. Additional substances to capture the breadth and structural diversity relied on the MaxMinPicker method[19] within RDKit. This method was used to select up to 5 further substances in addition to the centroid. For the terminal categories with fewer than 65 members, the MaxMin approach was feasible for 35 of the 39 relevant terminal cateories. No MaxMin approach was performed for terminal categories: Homogeneous metal compounds, Keto acids and derivatives, Organic salts and Organophosphorus compounds. A total of 189 diverse substances were selected from the 35 terminal categories. For the 141 terminal categories which originated from the primary categories with more than 65 members, the MaxMin was applicable to 139 categories. A total of 278 substances were identified following the MaxMin approach.

4.4. Diverse substance selection: constrained landscape

The TSCA landscape was filtered to only consider substances that could be potentially procured based on an inference that substances with greater than 1 vendor indexed in Pubchem were more likely to be procurable; was of the appropriate physical form (based on melting point and boiling point considerations); met the element conditions already outlined; was not likely to be volatile; had less

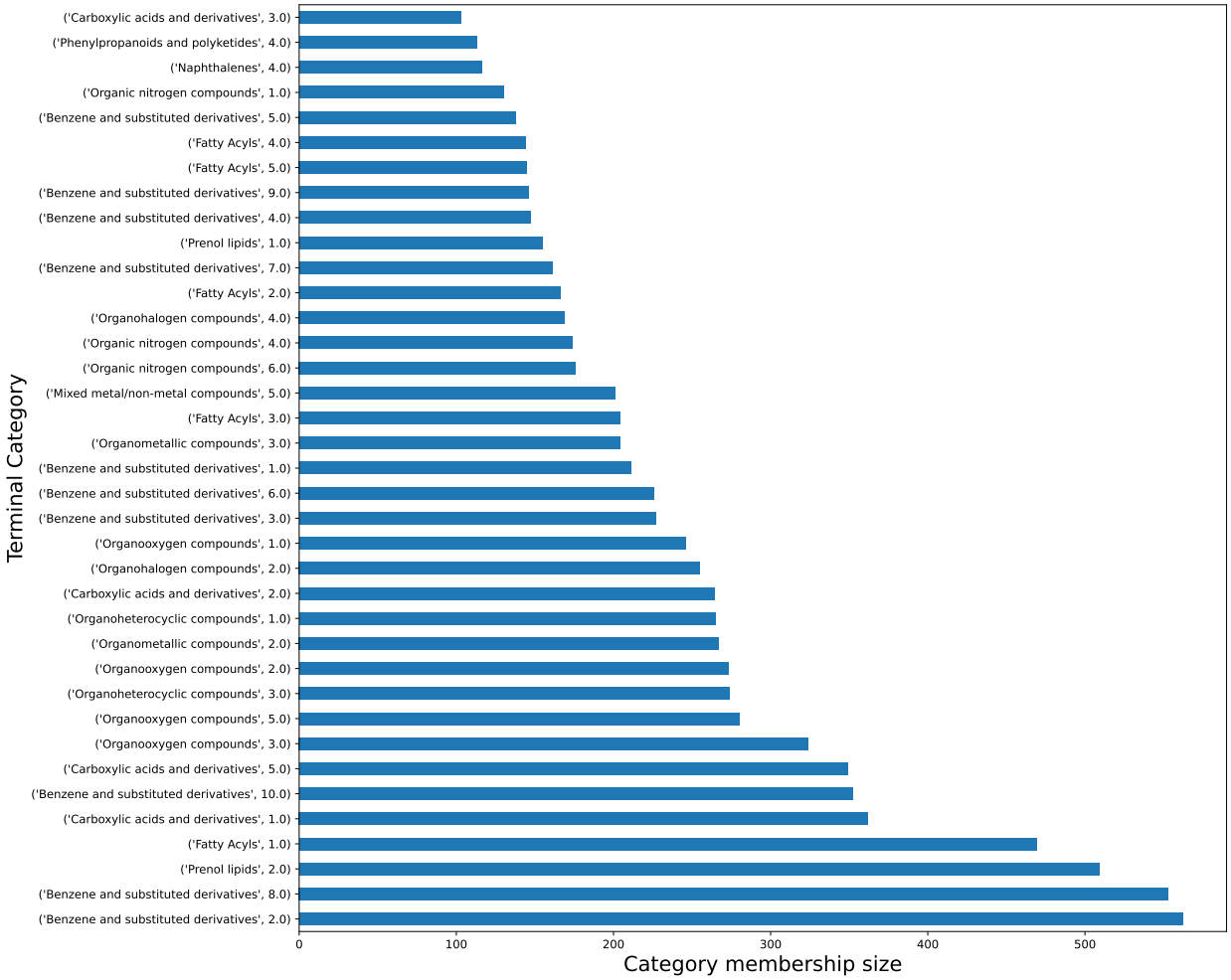


Figure 4: Terminal categories exceeding 100 members

than 3 Lipinski failures; and had not been tested in the ToxCast assays previously. Based on these considerations, 7565 substances were tagged as potentially amenable to screening for NAM assays spanning 160 (89%) of the 180 terminal categories.

The MaxMin approach could be applied to 146 of the 160 terminal categories. A total of 438 diverse substances were selected as a result of the MaxMin approach. The terminal categories where a MaxMin approach was not performed are listed in Table 1.

Table 1: List of the 14 terminal categories where a MaxMin approach was not applied

Terminal Category	Category Size
(Alkaloids and derivatives, nan)	3
(Azobenzenes, 1.0)	4
(Azoles, 3.0)	4
(Azolidines' 1.0)	3
(Azolidines, 5.0)	4
(Benzenoids, 4.0)	5
(Homogeneous non-metal compounds, 3.0)	1
(Hydrocarbon derivatives', nan)	3
(Lignans, neolignans and related compounds, nan)	4
(Lipids and lipid-like molecules, nan)	1
(Organic oxygen compounds, nan)	4
(Organic phosphoric acids and derivatives, 3.0)	2
(Other, 3.0)	4
(Quinolines and derivatives, 1.0)	3

4.5. Final proposed substances for testing

In terms of proposing a final set of substances for testing, substances were selected from the constrained landscape targeting terminal categories that had membership sizes greater than 20 but less than 300. Figure 5 shows the broad category membership sizes across the terminal categories. With this final constraint, 318 candidate substances that spanned 108 terminal categories were proposed for testing.

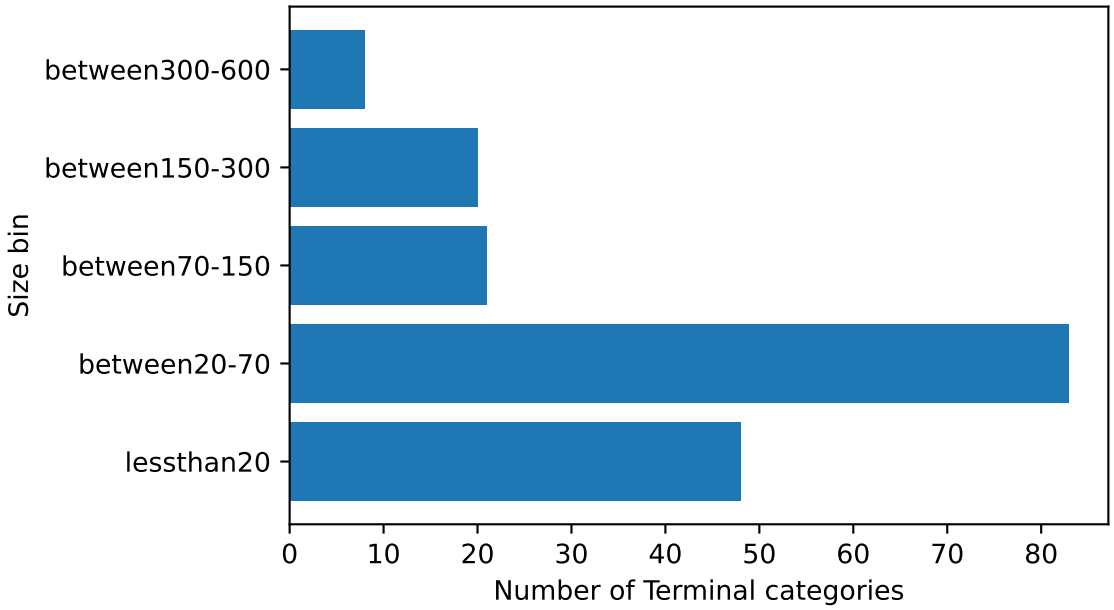


Figure 5: Terminal category membership size

4.6. Profiling of selected candidate substances

4.6.1. Structural diversity

The TSCA landscape was projected onto 2D using a t-SNE plot and colour coded in 2 ways; 1) using a tag of screenable to permit a comparison of the structural coverage of the screenable library relative to the full landscape and 2) using a tag of the proposed candidates for testing. Figure 6 (a) shows the screenable tag where 1 denotes substance that is member of the screenable inventory whereas Figure 6 (b) shows the structural coverage tagged by 1 to indicate proposed candidate substance for NAM testing. Figure 6 (a) shows how representative the constrained screenable library is of the larger landscape from a chemical structure perspective. In Figure 6 (b) the proposed candidate (denoted as final_picks) are evenly distributed throughout the landscape suggesting that from a structural perspective, they offer promise of being broadly representative of the TSCA landscape as a whole.

4.7. Profiling on the basis of predicted physicochemical properties

From the perspective of predicted physicochemical properties, the candidate test substances also represent the range of properties for the full landscape well. Figure 7 showcase the distribution of properties across the TSCA landscape and how that corresponds to the distribution of the 318 candidate substances.

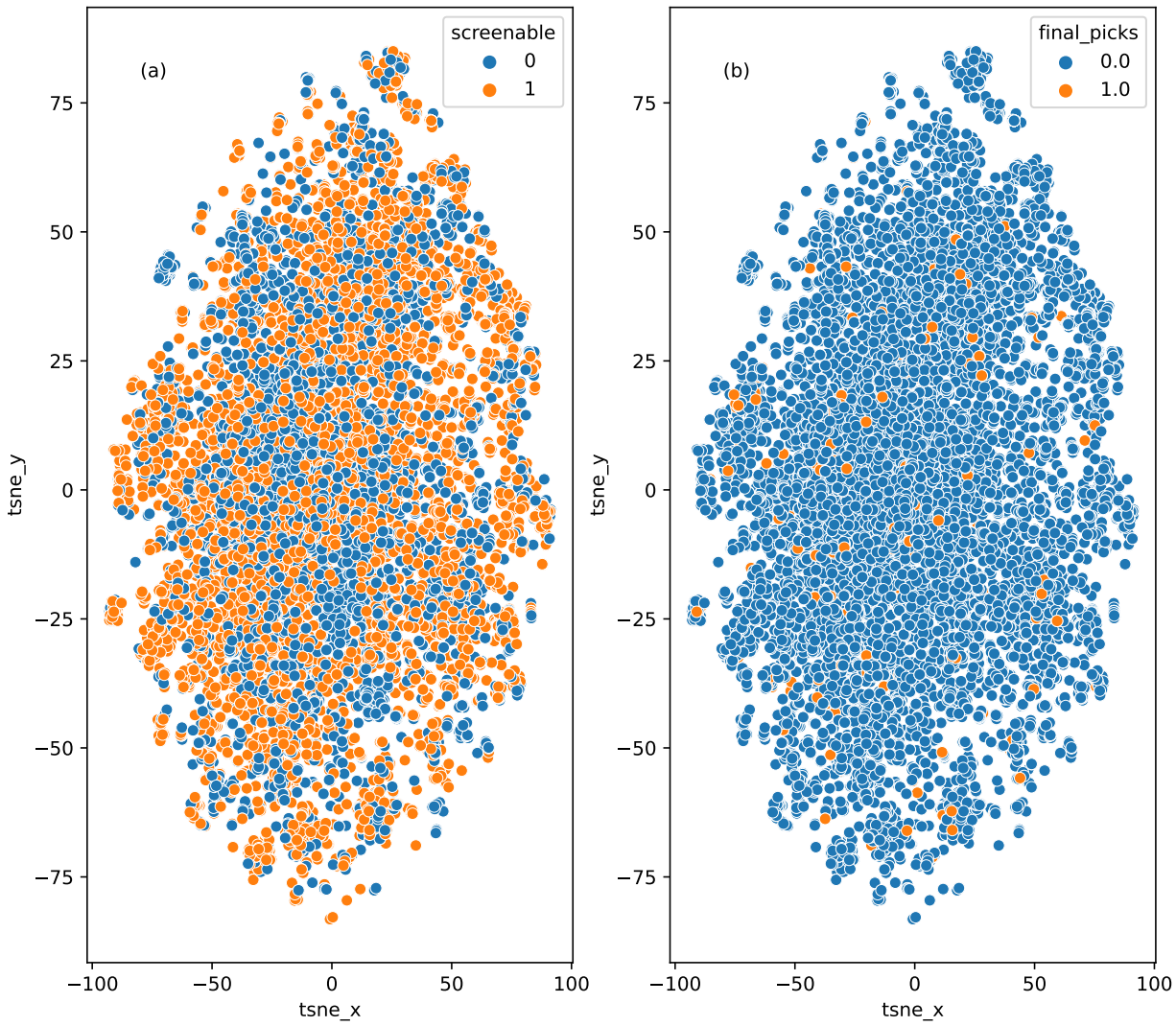


Figure 6: Selected substances within the constrained Landscape or whole TSCA Landscape

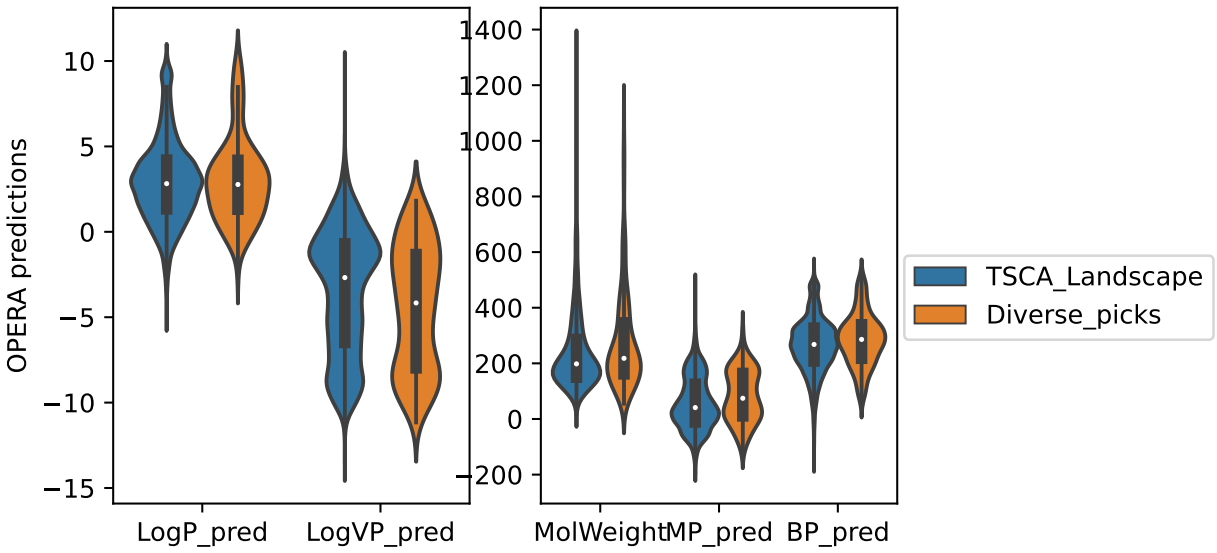


Figure 7: Physical property distributions between full inventory and candidate selections

4.8. Predicted structural alert profile

Derek Nexus predictions were made for all TSCA landscape substances and summarised in a bit vector format so that each substance was associated with a fingerprint made up of all possible endpoint-toxicophores present. There were 619 unique end-point-toxicophore combinations encompassing 53 different endpoints. The 318 substances identified flagged 217 of these end-point-toxicophore combinations encompassing 36 different endpoints (68% of the endpoints).

Pairwise distance matrices of the Derek fingerprints were computed to explore the extent to which there was consistency in alert profile across each terminal category. Four terminal categories are illustrated in Figure 8 to show the extent to which alert profiles varied.

Figure 8 (a - d) represent terminal categories ('Pyridines and derivatives', 1.0), ('Benzene and substituted derivatives', 9.0), ('Benzene and substituted derivatives', 9.0) and ('Lactones', nan) respectively.

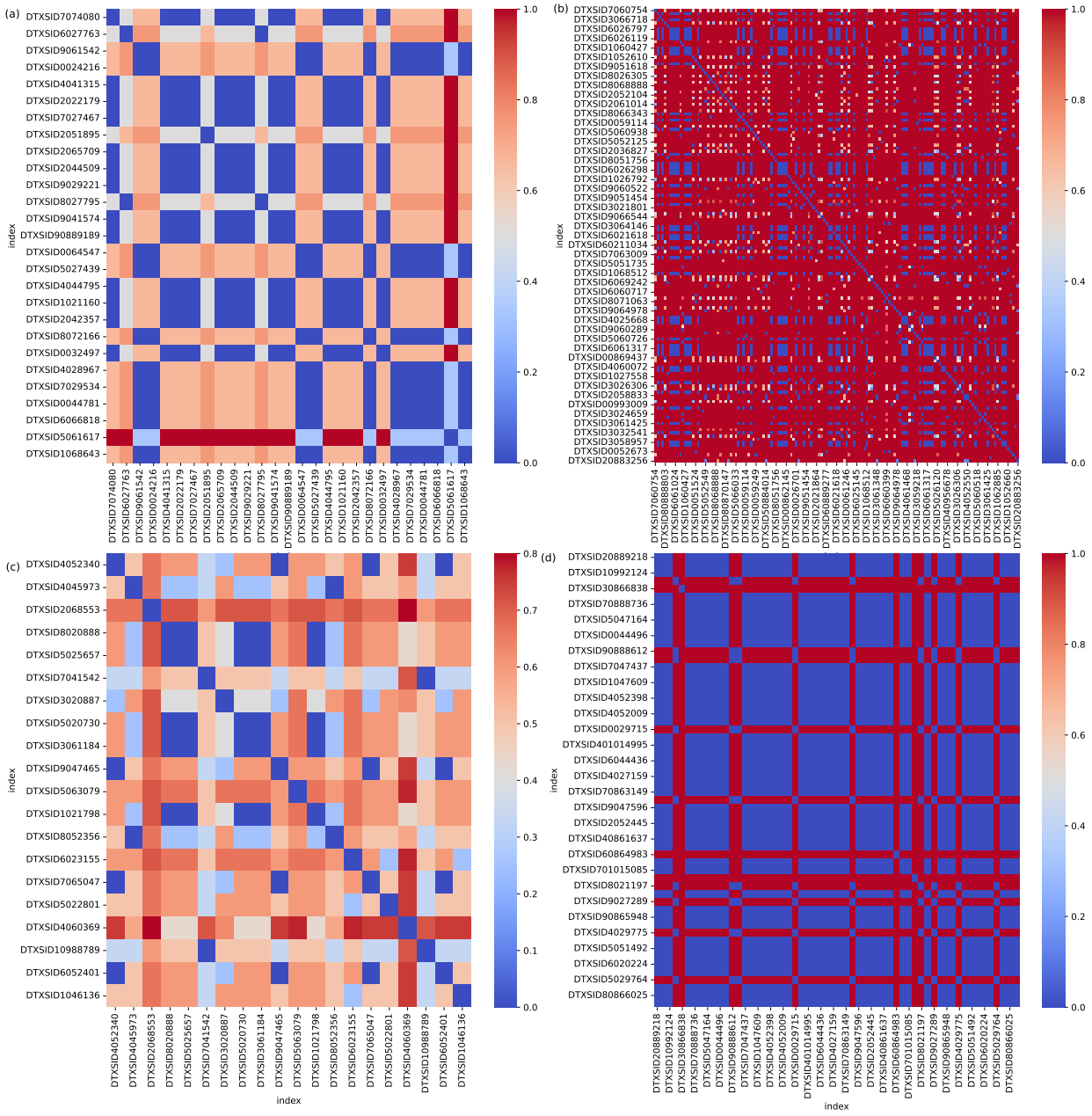
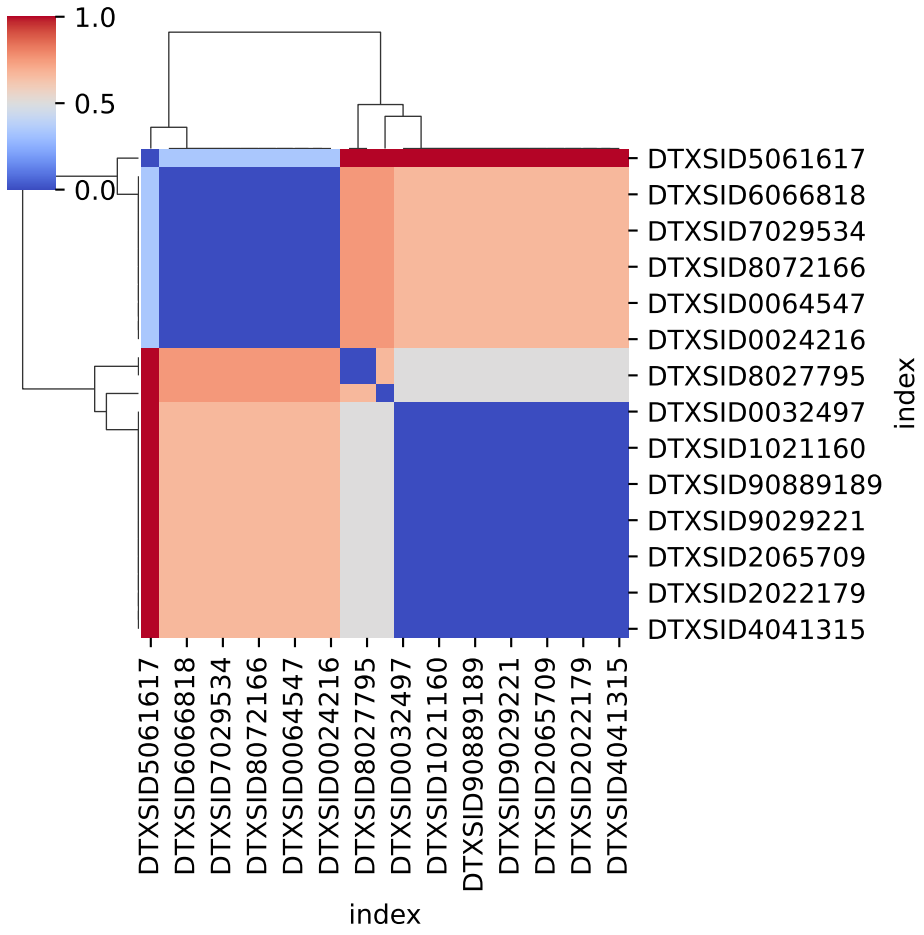


Figure 8: Heatmap of the pairwise distance matrices of selected terminal categories



For terminal category ('Pyridines and derivatives', 1.0), there are at least 3 clusters of substances which shared a common alert profile as shown in Table 2. On the otherhand, there were at least 4 clusters within terminal category ('Quinolines and derivatives', 3.0) which shared common alert endpoint profiles as shown in Table 3.

Table 2: Alert profiles for terminal category ('Pyridines and derivatives', 1.0)

cluster number	alerts
1	'('Carcinogenicity mammal', 'Di- to poly-halogenated alkane')','('Hepatotoxicity mammal', '2-Halopyridine')','('Mutagenicity in vitro bacterium', 'Trichloromethyl aromatic compound')'
2	'('Hepatotoxicity mammal', '2-Halopyridine')','('Nephrotoxicity mammal', 'Aromatic nitrile')'

cluster number	alerts
3	'('Hepatotoxicity mammal', '2-Halopyridine')','('Skin irritation/corrosion mammal', 'Pyridine or analogue')'

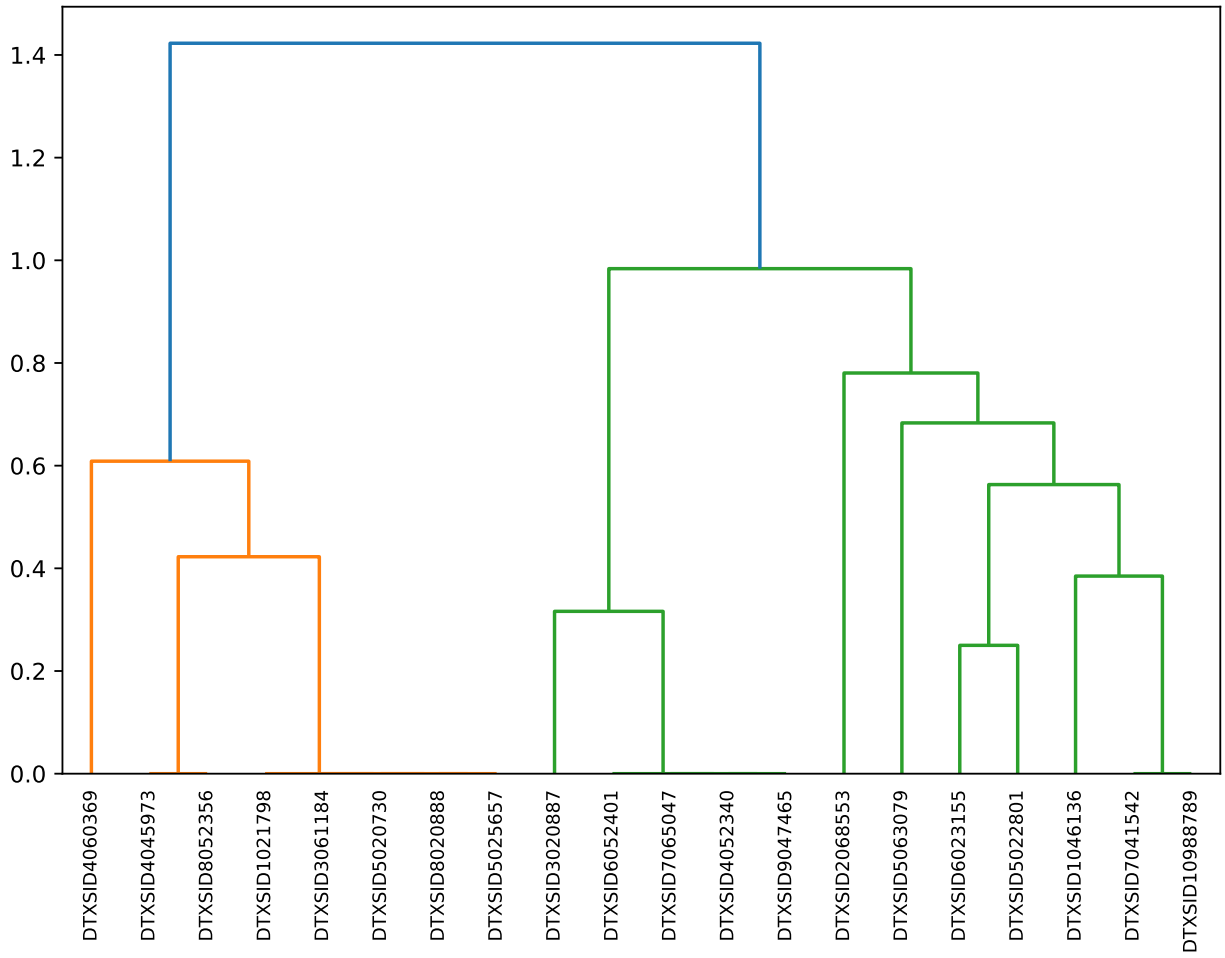


Table 3: Alert profiles for terminal category ('Quinolines and derivatives', 3.0)

cluster number	alerts
1	'('Carcinogenicity mammal', 'Aromatic amine or amide')','('Hepatotoxicity mammal', 'Quinoline')','('Mutagenicity in vitro bacterium', 'Aromatic amine or amide')','('Mutagenicity in vitro bacterium', 'Quinoline')','('Mutagenicity in vivo mammal', 'Aromatic amine or amide')','('Mutagenicity in vivo mammal', 'Quinoline')','('Skin irritation/corrosion mammal', 'Pyridine or analogue')'

cluster number	alerts
2	('Androgen receptor modulation mammal', 'Quinoline or analogue')','('Mutagenicity in vitro bacterium', 'Quinoline')','('Mutagenicity in vivo mammal', 'Quinoline')','('Skin irritation/corrosion mammal', 'Pyridine or analogue')'
3	'('Carcinogenicity mammal', 'Aromatic nitroso compound')','('Carcinogenicity mammal', 'Polyhalogenated aromatic')','('Mutagenicity in vitro bacterium', 'Aromatic nitroso compound')','('Mutagenicity in vitro bacterium', 'Quinoline')','('Mutagenicity in vivo mammal', 'Quinoline')','('Thyroid toxicity mammal', 'Aromatic iodo compound')'
4	'('Carcinogenicity mammal', '4-Aminobiphenyl, benzidine, naphthylamine or precursors')','('Carcinogenicity mammal', 'Aromatic azo compound')','('Mutagenicity in vitro bacterium', 'Aromatic azo compound')','('Mutagenicity in vitro bacterium', 'Quinoline')','('Mutagenicity in vivo mammal', 'Quinoline')'

The alert profiles within the terminal categories provide useful insights on the types of targets or toxicities to evaluate in the Tier 2 NAM assays.

No alert cluster profiles were produced for 10 terminal categories - namely ('Allenes', nan), ('Anthracenes', 5.0), ('Azobenzenes', 1.0), ('Azobenzenes', 5.0), ('Azobenzenes', 6.0), ('Azoles', 3.0), ('Hydrocarbons', 1.0), ('Mixed metal/non-metal compounds', 5.0), ('Organic oxygen compounds', nan), and ('Phenylpropanoids and polyketides', 1.0). Of these 4 terminal categories were associated with no alerts at all - ('Hydrocarbons', 1.0), ('Mixed metal/non-metal compounds', 5.0), ('Phenylpropanoids and polyketides', 1.0) and ('Allenes', nan). ('Allenes', nan) had 2 members and ('Mixed metal/non-metal compounds', 5.0) had 5 members but ('Phenylpropanoids and polyketides', 1.0) and ('Hydrocarbons', 1.0) had more members, 14 and 50 respectively. For the other 6 terminal categories, each was associated with a number of alerts as shown in Table 4. For example, terminal category ('Anthracenes', 5) was associated with 8 alerts that were present all 17 members; 7 of these were related to the anthraquinone moiety. The endpoints associated with these 7 alerts were carcinogenicity, mutagenicity, chromosomal damage, nepthrotoxicity, hepatotoxicity and mitochondrial dysfunction.

Table 4: Terminal categories that could not be clustered but were associated with many alerts

Terminal category	Number of alerts	Membership size
('Anthracenes', 5.0)	18	17
('Azobenzenes', 1.0)	9	7
('Azobenzenes', 5.0)	15	13
('Azobenzenes', 6.0)	21	23
('Azoles', 3.0)	13	4
('Organic oxygen compounds', nan)	9	6

For the remaining terminal categories, the maximum number of alerts was 101 for terminal category ('Benzene and substituted derivatives', 8.0) whereas the median number of alerts was 6 (see Figure 9).

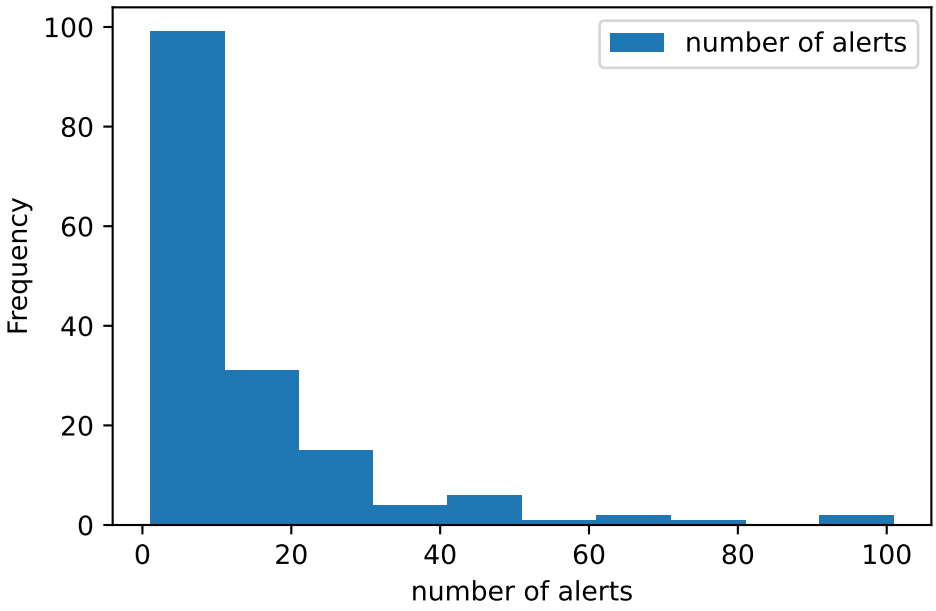


Figure 9: Distribution of Derek alert counts across terminal categories which were clustered

Profiling across all the substances in the landscape, the top 3 alerts that triggered the most substances were skin irritation/corrosion related with almost ~2000 substances presenting an alert. The next most populist alerts were for carcinogenicity, hepatoxicity and mutagenicity endpoints, approx 450-550 substances flagged each of these. The mean number of substances presenting an alert was 50 with the median number of substances presenting an alert being 11. Overall the alert profile across the TSCA landscape is quite sparse in terms of the number of alerts flagged. A number of the alerts were highly correlated with each other - at a threshold of 0.8, up to 215 alerts could dropped from

further consideration. The number of alerts across the 53 endpoints was summarised across the categories to gain a perspective of which categories were associated with what sorts of endpoints. Figure 10 shows a clustermap of the alert endpoints as a function of the terminal categories. For example, terminal categories containing benzene compounds and their derivatives or organohalogen compounds were typically associated with heptatoxicity, nephrotoxicity and carcinogenicity related endpoints.

4.9. Profiling toxicity predictions from TEST

Predictions were generated for the developmental and Ames mutagenicity endpoints using TEST. Consensus predictions were available for 14,194 substances. To explore the coverage of positive/negative predictions across the landscape, t-SNE plots were generated to aid visualisation. Figure 11 shows both endpoints. Positive developmental toxicity predictions were found to be quite evenly spread across the landscape whereas positive Ame results appear to cluster in some regions of the chemical space. To probe this a little more, the clustered region of positive predictions were overlaid relative to the terminal categories to start to gauge whether certain categories were visually overrepresented in positive Ames results. To simplify the representation, all terminal categories with fewer than 40 members were aggregated into one 'miscellaneous' category which were hidden in the resulting t-sne plot in Figure 12.

Many of the clustered positive predictions appear to be associated with substances in the Benzene and substituted derivatives but exploring the ratio of Ames positives to Ames negatives with the terminal categories showed that Azobenzenes and Anthracenes were substantially overrepresented (Figure 12). The top 10 terminal categories where a ratio between predicted Ames positive and negative outcomes exceeds 2 are shown in Table 5. In contrast ('Steroids and steroid derivatives', nan) and ('Phenylpropanoids and polyketides', 4.0) are examples of terminal categories where developmental toxicity predictions exceed no developmental toxicity (Table 6). t-SNE plots of 3 of the Ames and Developmental toxicity terminal categories where these ratios of positive:negative outcomes are high are shown in Figure 13.

Table 5: Top 10 terminal categories where the ratio between positive:negative exceeds 2

	Mutagenicity		
Terminal Category	Negative	Mutagenicity Positive	Ratio
('Anthracenes', 3.0)	1.0	21.0	21.0
('Anthracenes', 7.0)	1.0	15.0	15.0
('Azobenzenes', 7.0)	1.0	13.0	13.0

Terminal Category	Mutagenicity Negative	Mutagenicity Positive	Ratio
('Anthracenes', 9.0)	5.0	46.0	9.2
('Quinolines and derivatives', 2.0)	1.0	9.0	9.0
('Anthracenes', 6.0)	2.0	18.0	9.0
('Anthracenes', 4.0)	2.0	17.0	8.5
('Anthracenes', 5.0)	2.0	14.0	7.0
('Azobenzenes', 2.0)	3.0	11.0	3.67
('Acetylides', nan)	8.0	19.0	2.38

Table 6: Top 10 terminal categories where the ratio between positive:negative exceeds 10

Terminal Category	Developmental NON-toxicant	Developmental toxicant	Ratio
('Steroids and steroid derivatives', nan)	1.0	31.0	31.0
('Phenylpropanoids and polyketides', 4.0)	6.0	96.0	16.0
('Organic salts', nan)	1.0	16.0	16.0
('Phenol ethers', 3.0)	1.0	15.0	15.0
('Azolidines', 5.0)	1.0	15.0	15.0
('Azobenzenes', 3.0)	2.0	28.0	14.0
('Benzenoids', 2.0)	1.0	12.0	12.0
('Triphenyl compounds', nan)	2.0	23.0	11.5
('Diazines', 1.0)	3.0	32.0	10.67
('Organic phosphoric acids and derivatives', 1.0)	3.0	31.0	10.33

In terms of the profile of the selected candidate substances relative to the full landscape, the number of substances predicted positive and negative is shown in Figure 14. The profile trend between the full landscape and the selected candidate substances are consistent but the number of substances predicted to be developmental toxicants is surprisingly high. This is at least consistent with the test set performance metrics reported in the user manual (https://www.epa.gov/sites/default/files/2016-

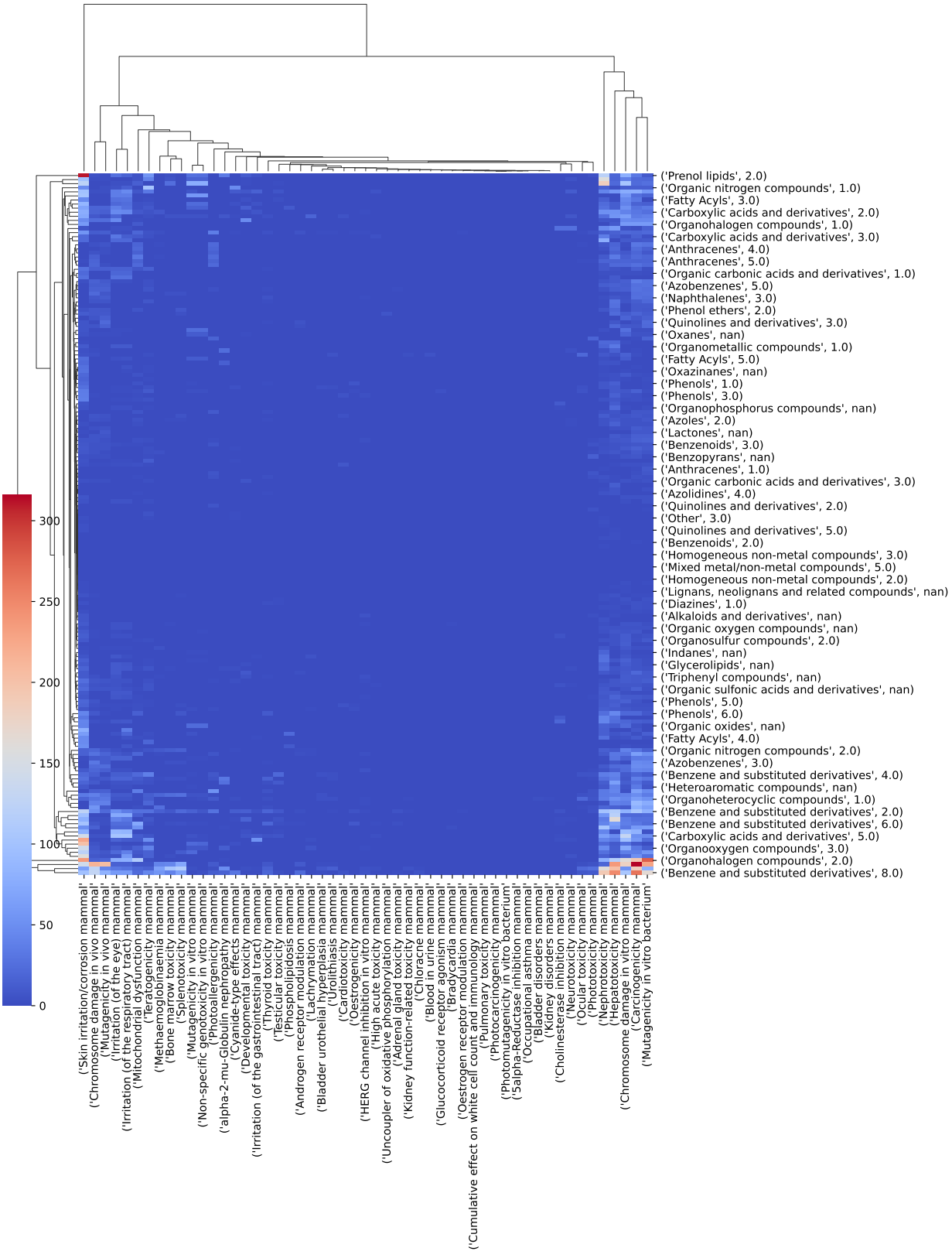


Figure 10: Distribution of Derek alert counts across terminal categories which were clustered

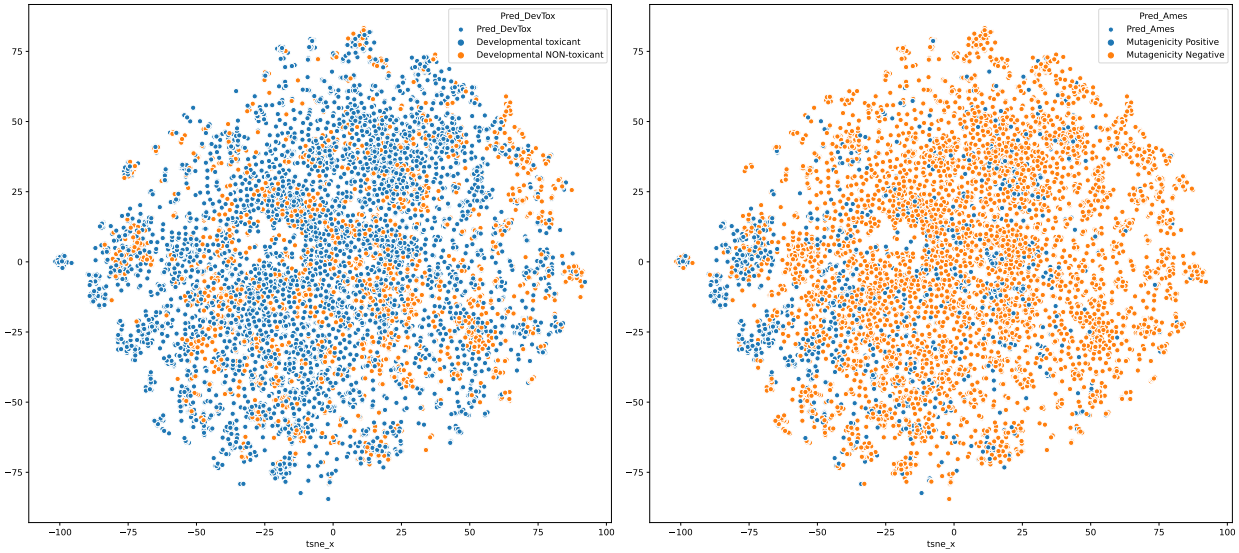


Figure 11: t-SNE plots of TEST mutagenicity and developmental toxicity predictions

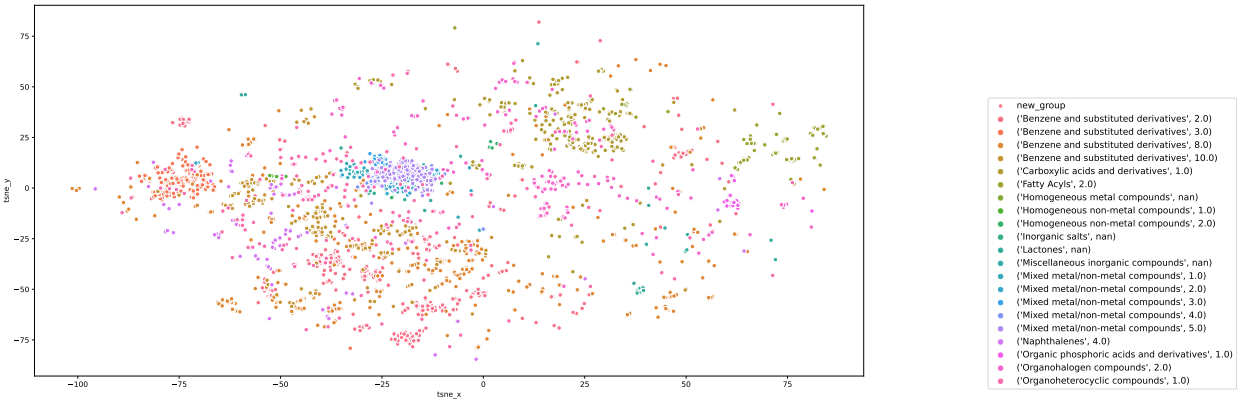


Figure 12: t-SNE plots of TEST mutagenicity predictions for categories with members exceeding 40

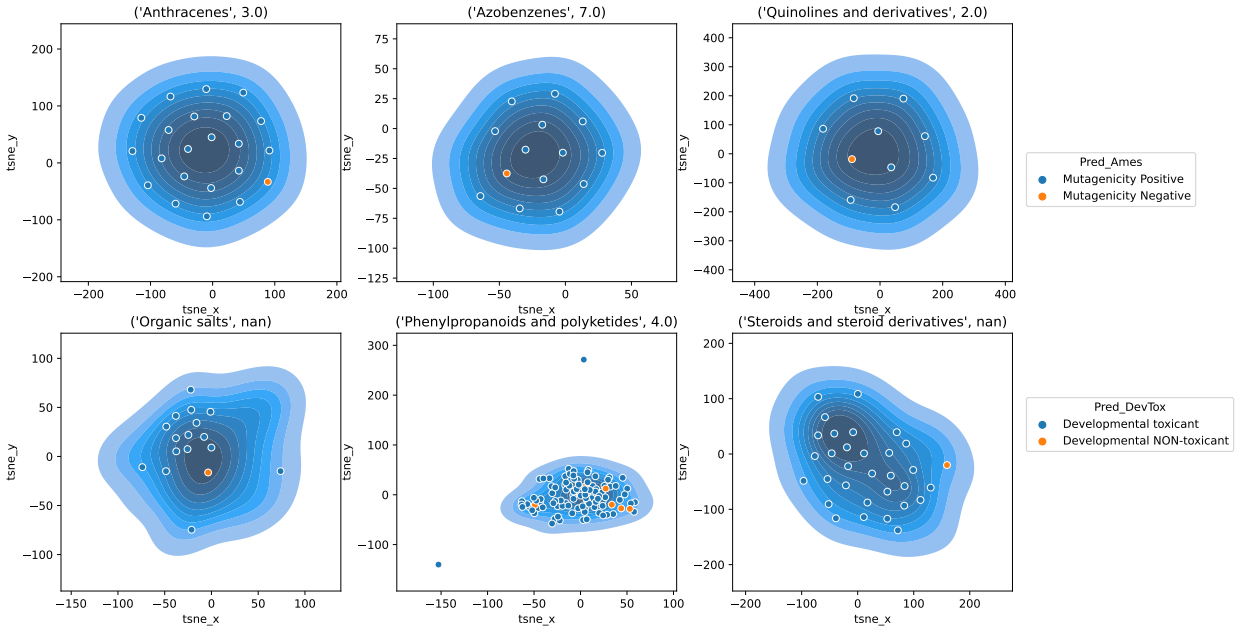


Figure 13: t-SNE plots of selected categories that are overrepresented by Ames positive predictions

05/documents/600r16058.pdf) namely that the sensitivity of the developmental toxicity consensus model was 0.9 in contrast to the specificity 0f 0.471. The test set performance metrics were similar for sensitivity and specificity, 0.79 and 0.756 respectively.

4.10. Profiling toxicity predictions from OPERA

Predictions were generated using OPERA v2.8 for all TSCA landscape substances. Figure 15 presents the distribution of CATMOS predictions to show the range of LD50 acute oral values for the landscape as a whole and how this is represented for the candidate selections. The medians of the TSCA landscape and candidate selections were very similar, with values ~2300 mg/kg and ~2100 mg/kg which are borderline low toxicity. The median LD50 values differed across the terminal categories with ('Organosulfur compounds', 2.0), ('Allenes', nan) and ('Allenes', nan) having a median LD ~245 mg/kg whereas ('Benzenoids', 5.0) and ('Fatty Acyls', 3.0) had a median ~7000 mg/kg. Figure 16 showcase the ECDFs for these selected terminal categories to highlight the difference in potencies. Those terminal categories with lower LD50 values (more potent) have ECDFs that are more right shifted.

4.11. Predicted ToxCast Profile

Figure 17 shows an overview of regions of the chemical landscape as well as different types of assays that give rise to positive predictions. 10% of substances were positive in 48 of the 81 assays whereas only 2 assays 'ATG_NRF2_ARE_CIS', 'ATG_PXRE_CIS' resulted in more than 50% of substances predicted to be positive. Pregnane X receptor (PXR) is a nuclear receptor that regulates the

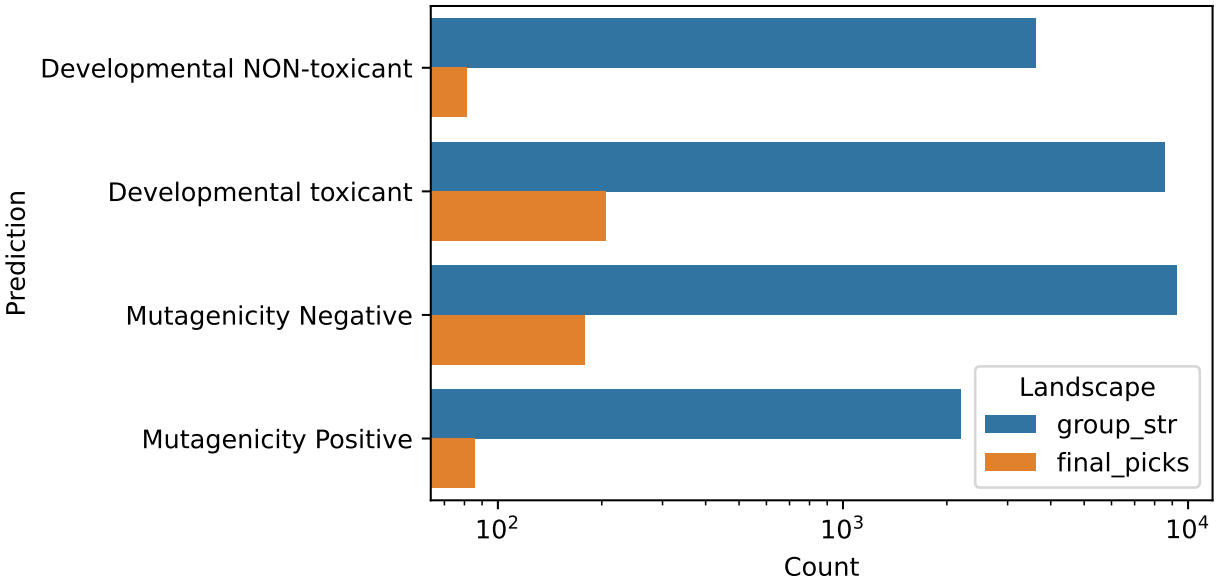


Figure 14: t-SNE plots of selected categories that are overrepresented by Ames positive predictions

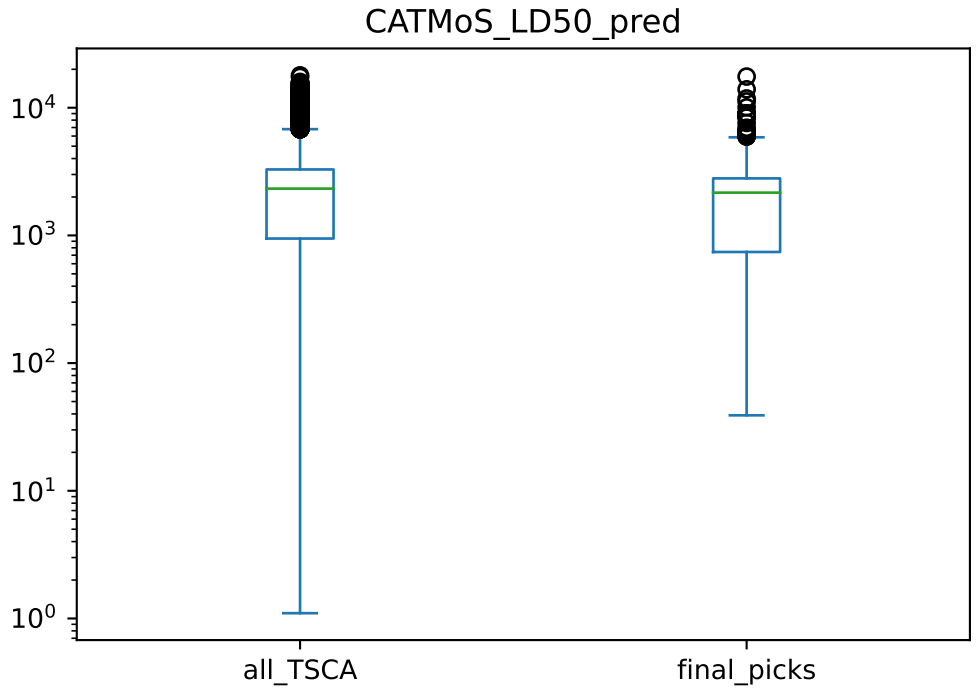


Figure 15: Distribution of acute oral toxicity predictions for the TSCA Landscape relative to the candidate substances

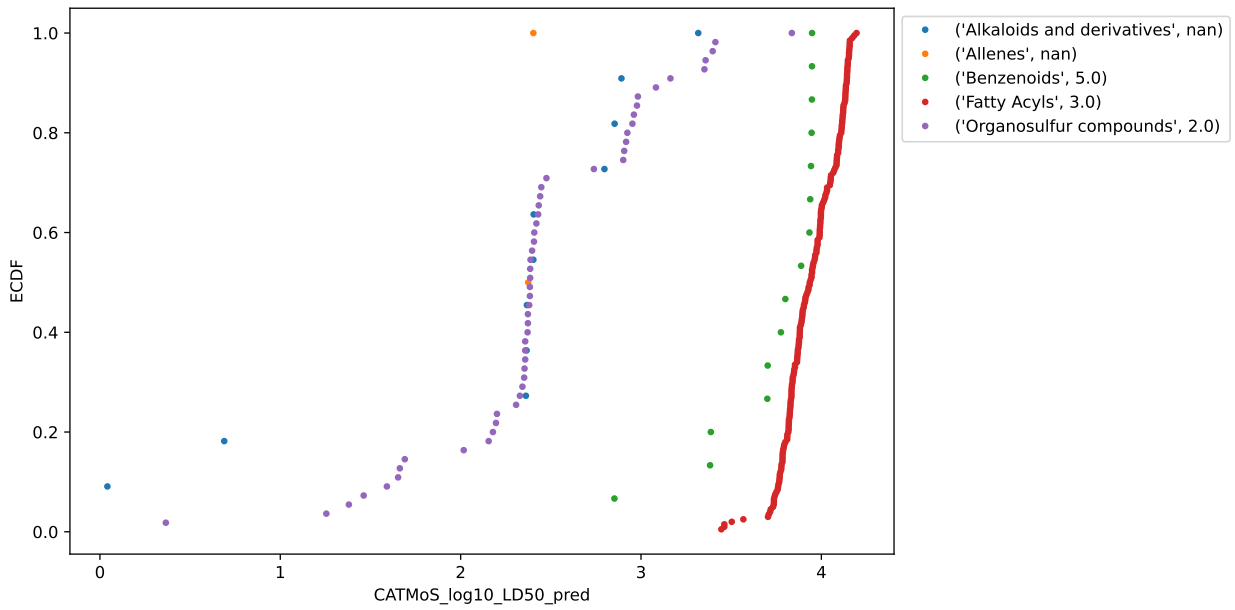


Figure 16: ECDFs of acute oral toxicity predictions for selected terminal categories

metabolism and disposition of various xenobiotics and endobioitics. ATG_NRF2_ARE_CIS is an assay that provides a measure oxidative stress. 30% of substances were predicted to be active in the following 8 assays 'ATG_DR4_LXR_CIS', 'ATG_ERE_CIS', 'ATG_ERa_TRANS', 'ATG_NRF2_ARE_CIS', 'ATG_PPARg_TRANS', 'ATG_PXRE_CIS', 'ATG_PXR_TRANS', 'ATG_VDRE_CIS' - some of which are associated with endocrine activity or oxidative stress. Across the chemicals the terminal categories that were associated to these positive predictions were ('Benzene and substituted derivatives', 8.0), ('Benzene and substituted derivatives', 2.0), ('Prenol lipids', 2.0), ('Fatty Acyls', 1.0), ('Benzene and substituted derivatives', 10.0), ('Organooxygen compounds', 3.0).

4.12. Availability of in vivo toxicity data within terminal categories

There were only 24 non-cancer and 32 reproductive/developmental PODs and for the candidate set compared with 2355 and 2353 respectively for the full landscape. The variation in potency differed quite considerably. For reproductive/developmental PODs, the min, median and max values were 10.93, 114 and 838 mg/kg bw respectively for the candidate set. In contrast for the full landscape, these values were 4E-08, 70.6 and 12,890 mg/kg bw highlightly a much broader variation in potencies. For the non-cancer endpoints, the min, median and max values were 3E-05, 21 and 4587 mg/bw for the full landscape but 0.346, 26 and 949 mg/bw for the candidate set. Across the terminal categories, ('Isoindoles and derivatives', nan) and ('Organosulfur compounds', 2.0) has the lowest median non-cancer PODs whereas ('Azolidines', 4.0) and ('Triazines', nan) had the lowest median reproductive/developmental PODs demonstrating that the potencies varied considerably across both

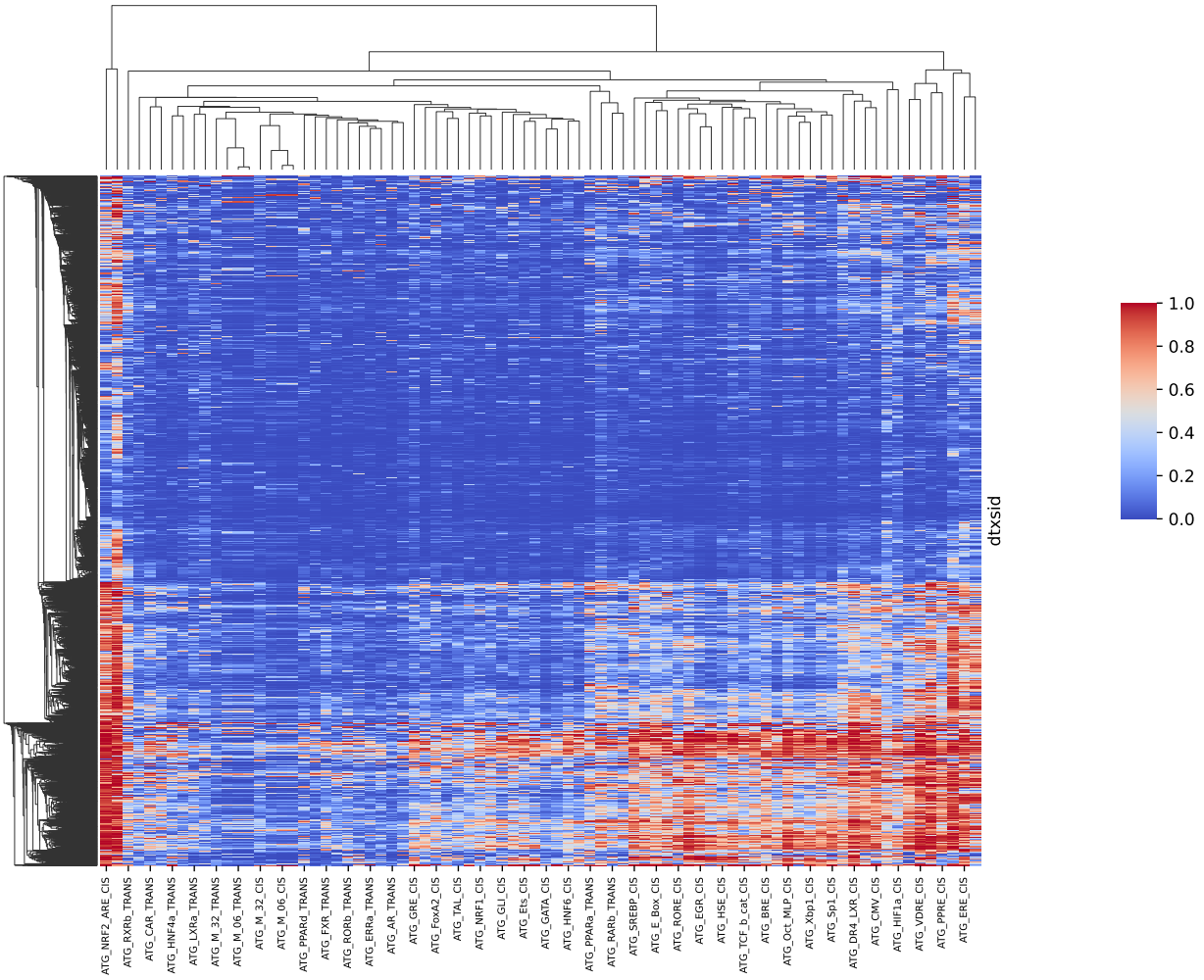


Figure 17: Clustermap of the TSCA landscape with respect to predicted probabilites of Attagene assay hitcalls

endpoints and chemical categories.

5. Conclusions

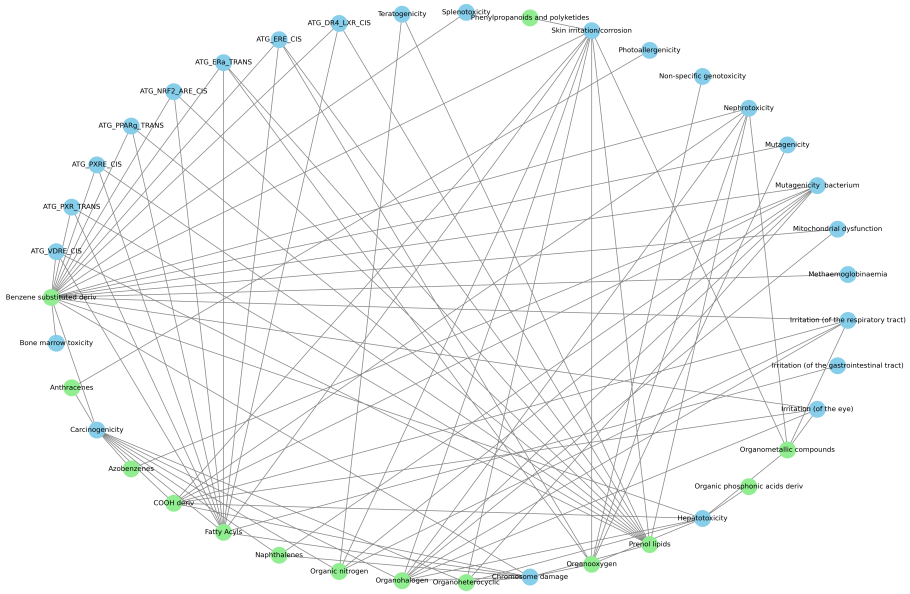
A cheminformatics workflow was developed in an effort to identify a set of approximately 300 representative candidate case study chemicals from the TSCA landscape that could be submitted for screening in a range of different NAM approaches. A categorisation scheme making use of the Classy-Fire ontology formed the basis of initial primary categories which were then subset into smaller more structurally similar categorises following hierarchical clustering. Since NAM testing was envisaged, the landscape was bounded by technical testing constraints. Physicochemical properties, reactivity flags, procurabillity, category size and analytical method detection amenability considerations were used to subset a portion of the landscape from which substances could be identified for testing. A set of 318 substances were proposed that spanned 106 (~59%) of the terminal categories. From a coarse grain perspective, the proposed 318 substances spanned the overall structural diversity of both the constrained and full landscape well. The physicochemical properties profile was similar for both the candidate substances and the landscape and the median potencies of the acute oral toxicity predictions were also very similar. All these factors led credence that the candidate set of substances were a reasonable and representative set of the broader TSCA landscape to the extent that it could be procured and tested. Attention was then turned to exploring the TSCA landscape through the lens of the terminal categories constructed to gauge the bioactivities expected, the sorts of toxicity endpoints that were likely to be exhibited as well as their potencies. To summarise such insights network graphs were constructed to capture the associations between categories and their active responses in ToxCast assays, specific endpoints as flagged by Derek alerts or predicted acute oral toxicity hazard classifications (Figure 18).

From these representations, such insights as to which categories are typically associated with a specific outcome or an array of outcomes can be more readily identified which are informative for prioritisation efforts.

REFERENCES

[1] U. EPA, The frank r. lautenberg chemical safety for the 21st century act (2016).
URL https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act-law

[2] U. EPA, The new chemicals collaborative research program: Modernizing the process and bringing innovative science to evaluate new chemicals under tsca. office of chemical safety and pollution prevention and office of research and development. (2022).



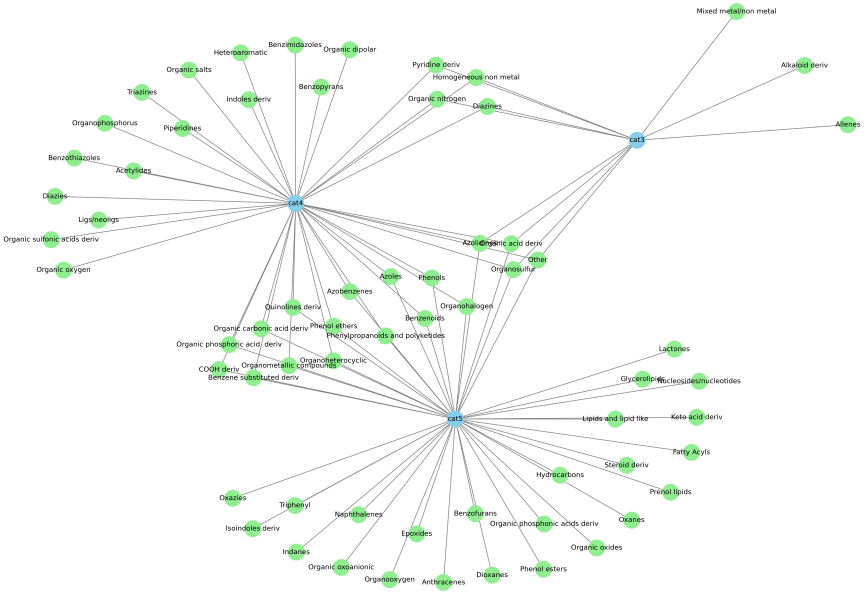


Figure 18: Network graphs of primary categories and their predicted Derek endpoint, ToxCast assay outcome and acute oral hazard classification.

[3] R. S. Thomas, T. Bahadori, T. J. Buckley, J. Cowden, C. Deisenroth, K. L. Dionisio, J. B. Frithsen, C. M. Grulke, M. R. Gwinn, J. A. Harrill, M. Higuchi, K. A. Houck, M. F. Hughes, E. S. Hunter, K. K. Isaacs, R. S. Judson, T. B. Knudsen, J. C. Lambert, M. Linnenbrink, T. M. Martin, S. R. Newton, S. Padilla, G. Patlewicz, K. Paul-Friedman, K. A. Phillips, A. M. Richard, R. Sams, T. J. Shafer, R. W. Setzer, I. Shah, J. E. Simmons, S. O. Simmons, A. Singh, J. R. Sobus, M. Strynar, A. Swank, R. Tornero-Valez, E. M. Ulrich, D. L. Villeneuve, J. F. Wambaugh, B. A. Wetmore, A. J. Williams, The next generation blueprint of computational toxicology at the u.s. environmental protection agency 169 (2) 317–332. doi:10.1093/toxsci/kfz058.

[4] J. Nyffeler, C. Willis, R. Lougee, A. Richard, K. Paul-Friedman, J. A. Harrill, Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling 389 114876. doi:10.1016/j.taap.2019.114876.

[5] J. Harrill, I. Shah, R. W. Setzer, D. Haggard, S. Auerbach, R. Judson, R. S. Thomas, Considerations for strategic use of high-throughput transcriptomics chemical screening data in regulatory decisions 15 64–75. doi:10.1016/j.cotox.2019.05.004.

[6] R. S. Judson, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah, A. M. Richard, D. J. Dix, In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project 118 (4) 485–492. doi:10.1289/ehp.0901392.

[7] A. M. Richard, R. S. Judson, K. A. Houck, C. M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M. T. Martin, J. F. Wambaugh, T. B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A. J. Williams, S. B. Little, K. M. Crofton, R. S. Thomas, ToxCast chemical landscape: Paving the road to 21st century toxicology 29 (8) 1225–1251. doi:10.1021/acs.chemrestox.6b00135.

[8] B. A. Wetmore, J. F. Wambaugh, B. Allen, S. S. Ferguson, M. A. Sochaski, R. W. Setzer, K. A. Houck, C. L. Strope, K. Cantwell, R. S. Judson, E. LeCluyse, H. J. Clewell, R. S. Thomas, M. E. Andersen, Incorporating High-Throughput Exposure Predictions With Dosimetry-Adjusted In Vitro Bioactivity to Inform Chemical Toxicity Testing, Toxicological Sciences 148 (1) (2015) 121–136. arXiv:https://academic.oup.com/toxsci/article-pdf/148/1/121/16693373/kfv171.pdf, doi:10.1093/toxsci/kfv171.
URL https://doi.org/10.1093/toxsci/kfv171

[9] R. S. Judson, R. S. Thomas, N. Baker, A. Simha, X. M. Howey, C. Marable, N. C. Kleinstreuer, K. A. Houck, Workflow for defining reference chemicals for assessing performance of in vitro assays 36 (2) 261–276. doi:10.14573/altex.1809281.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784312/

[10] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D. S. Wishart, ClassyFire: automated chemical classification with a comprehensive, computable taxonomy 8 61. doi:10.1186/s13321-016-0174-y.

[11] A. J. Williams, C. M. Grulke, J. Edwards, A. D. McEachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson, A. M. Richard, The comptox chemistry dashboard: a community data resource for environmental chemistry, Journal of Cheminformatics 9 (1) (2017) 61, number: 1 PMID: 29185060 PMCID: PMC5705535. doi:10.1186/s13321-017-0247-6.

[12] G. L. Landrum, RDKit: Open-source cheminformatics;.
URL http://www.rdkit.org

[13] K. Mansouri, C. M. Grulke, R. S. Judson, A. J. Williams, Opera models for predicting physicochemical properties and environmental fate endpoints, Journal of Cheminformatics 10 (1) (2018) 10. doi:10.1186/s13321-018-0263-1.
URL https://doi.org/10.1186/s13321-018-0263-1

[14] K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson, A. J. Williams, An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling 27 (11) 939–965. doi:10.1080/1062936X.2016.1253611.

[15] T. W. Schultz, R. Diderich, C. D. Kuseva, O. G. Mekenyan, The OECD QSAR toolbox starts its second decade 1800 55–77.

doi:10.1007/978-1-4939-7899-1_2.

[16] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC international chemical identifier 7 (1) 23. doi:10.1186/s13321-015-0068-4.
URL https://doi.org/10.1186/s13321-015-0068-4

[17] J. H. Ward, Hierarchical grouping to optimize an objective function 58 (301) 236–244, publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845. doi:10.1080/01621459.1963.10500845.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845

[18] D. Rogers, M. Hahn, Extended-connectivity fingerprints 50 (5) 742–754, publisher: American Chemical Society. doi:10.1021/ci100050t.
URL https://doi.org/10.1021/ci100050t

[19] M. Ashton, J. Barnard, F. Casset, M. Charlton, G. Downs, D. Gorse, J. Holliday, R. Lahana, P. Willett, Identification of diverse database subsets using property-based and fragment-based molecular descriptions 21 (6) 598–604, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qsar.200290002. doi:10.1002/qsar.200290002.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qsar.200290002

[20] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 46 (1) 3–26. doi:10.1016/s0169-409x(00)00129-0.

[21] C. N. Lowe, K. K. Isaacs, A. McEachran, C. M. Grulke, J. R. Sobus, E. M. Ulrich, A. Richard, A. Chao, J. Wambaugh, A. J. Williams, Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis 413 (30) 7495–7508. doi:10.1007/s00216-021-03713-w.

[22] L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE 9 (86) 2579–2605.
URL http://jmlr.org/papers/v9/vandermaaten08a.html

[23] K. Mansouri, A. L. Karmaus, J. Fitzpatrick, G. Patlewicz, P. Pradeep, D. Alberga, N. Alepee, T. E. H. Allen, D. Allen, V. M. Alves, C. H. Andrade, T. R. Auernhammer, D. Ballabio, S. Bell, E. Benfenati, S. Bhattacharya, J. V. Bastos, S. Boyd, J. B. Brown, S. J. Capuzzi, Y. Chushak, H. Ciallella, A. M. Clark, V. Consonni, P. R. Daga, S. Ekins, S. Farag, M. Fedorov, D. Fourches, D. Gadaleta, F. Gao, J. M. Gearhart, G. Goh, J. M. Goodman, F. Grisoni, C. M. Grulke, T. Hartung, M. Hirn, P. Karpov, A. Korotcov, G. J. Lavado, M. Lawless, X. Li, T. Luechtefeld, F. Lunghini, G. F. Mangiatordi, G. Marcou, D. Marsh, T. Martin, A. Mauri, E. N. Muratov, G. J. Myatt, D.-T. Nguyen, O. Nicolotti, R. Note, P. Pande, A. K. Parks, T. Peryea, A. H. Polash, R. Rallo, A. Roncaglioni, C. Rowlands, P. Ruiz, D. P. Russo, A. Sayed, R. Sayre, T. Sheils, C. Siegel, A. C. Silva, A. Simeonov, S. Sosnin, N. Southall, J. Strickland, Y. Tang, B. Teppen, I. V. Tetko, D. Thomas, V. Tkachenko, R. Todeschini, C. Toma, I. Tripodi, D. Trisciuzzi, A. Tropsha, A. Varnek, K. Vukovic, Z. Wang, L. Wang, K. M. Waters, A. J. Wedlake, S. J. Wijeyesakere, D. Wilson, Z. Xiao, H. Yang, K. G. Zahoranszky, A. V. Zakharov, F. F. Zhang, Z. Zhang, T. Zhao, H. Zhu, K. M. Zorn, W. Casey, N. C. Kleinstreuer, CATMoS: Collaborative acute toxicity modeling suite 129 (4) 047013, publisher: Environmental Health Perspectives. doi:10.1289/EHP8495.
URL https://ehp.niehs.nih.gov/doi/full/10.1289/EHP8495

[24] N. Aurisano, O. Jolliet, W. A. Chiu, R. Judson, S. Jang, A. Unnikrishnan, M. B. Kosnik, P. Fantke, Probabilistic Points of Departure and Reference Doses for Characterizing Human Noncancer and Developmental/Reproductive Effects for 10,145 Chemicals, Environmental Health Perspectives 131 (3) (2023) 037016, publisher: Environmental Health Perspectives. doi:10.1289/EHP11524.
URL https://ehp.niehs.nih.gov/doi/10.1289/EHP11524

[25] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy 585 (7825) 357–362, number: 7825 Publisher: Nature Publishing Group. doi:10.1038/s41586-020-2649-2.

URL https://www.nature.com/articles/s41586-020-2649-2

[26] T. p. d. team, pandas-dev/pandas: Pandas. doi:10.5281/zenodo.8092754.
URL https://zenodo.org/record/8092754

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, . Duchesnay, Scikit-learn: Machine learning in python 12 (85) 2825–2830.
URL http://jmlr.org/papers/v12/pedregosa11a.html

[28] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in Science & Engineering 9 (3) (2007) 90–95. doi:10.1109/MCSE.2007.55.

[29] M. Waskom, M. Gelbart, O. Botvinnik, J. Ostblom, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. Warmenhoven, J. B. Cole, E. t. Hoeven, J. d. Ruiter, J. Vanderplas, S. Hoyer, C. Pye, A. Miles, C. Swain, K. Meyer, M. Martin, P. Bachant, S. Molin, E. Quintero, G. Kunter, S. Villalba, Brian, C. Fitzgerald, C. Evans, M. L. Williams, D. O'Kane, mwaskom/seaborn: v0.12.2 (december 2022). doi:10.5281/zenodo.7495530.
URL https://zenodo.org/record/7495530

[30] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, Jupyter notebooks – a publishing format for reproducible computational workflows, in: F. Loizides, B. Schmidt (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas, IOS Press, 2016, pp. 87 – 90.