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

G Patlewicz[1*], K Paul Friedman[1], and AJ Williams[1],
[1] Center for Computational Toxicology & Exposure, U.S. Environmental Protection Agency, RTP, NC, USA

Grace Patlewicz | Patlewicz.grace@epa.gov | ORCID: 0000-0003-3863-9689 | 919-541-1540

www.epa.gov

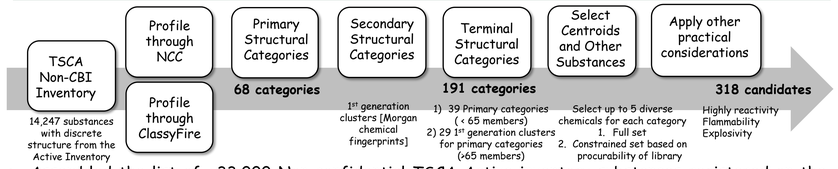
INTRODUCTION

The Toxic Substances Control Act (TSCA) requires the US EPA to evaluate the hazard and exposure to 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. The New Chemicals Collaborative Research Program (NCCRP) that has been established between OPPT and EPA's Office of Research & Development will explore opportunities to leverage New Approach Methods (NAMs).

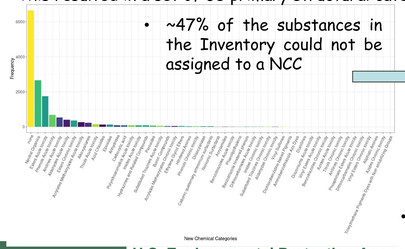
WHY NEW APPROACH METHODS (NAMs)?

- The non-confidential TSCA Inventory contains ~86,685 substances of which 42,170 are active in commerce.
- Use of traditional approaches will be too resource and time intensive to generate relevant data to facilitate assessment.
- Herein a cheminformatics workflow was developed to identify a set of ~300 representative candidate case study chemicals from the TSCA non-confidential active inventory that could be submitted screening in a range of different NAM approaches.

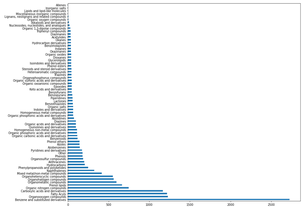
WORKFLOW



- Assembled the list of ~33,000 Non-confidential TSCA Active inventory substances registered on the CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard/chemical-lists/TSCA_ACTIVE_NCTI_0222).
- Filtered to identify those that were associated with a defined structure and could be processed by RDKit (rdkit.org) (14,247 substances).
- Profiled through the EPA New Chemicals Categories (NCC) to investigate the feasibility of creating initial primary structural categories.
- Profiled the Inventory using the Chemistry Ontology scheme called ClassyFire[1] to define pragmatic structural primary categories. This produces different levels of information: Kingdom, Superclass, Class. Based on the distribution of the substances across these levels – a hybrid of primary categories was established such that Superclass assignments were used unless the membership size exceeded 1000. This resulted in a set of 68 primary structural categories.



- ~47% of the substances in the Inventory could not be assigned to a NCC



- 13477 substances were organic, 593 were inorganics and 179 could not be assigned. The latter were assigned as "other".

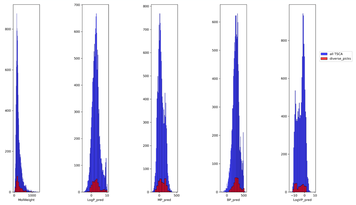
SECONDARY-TERMINAL CATEGORIES

- 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.
- Applied hierarchical clustering on the basis of Morgan fingerprints to those primary categories where membership exceeded 65. The first generation of clusters formed the secondary category. 65 was an arbitrary but pragmatic threshold based on the distribution of membership size. 39 primary categories had a membership size less than 65.
- 191 Terminal categories were derived which were either 'primary' or 'secondary' categories.

REPRESENTATIVE SUBSTANCE SELECTION

- The nominally representative substance for a given category was taken as the medoid. This is 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 additional structurally diverse substances within the category on the basis of their Morgan chemical fingerprints.
- Next the landscape of substances was constrained by various considerations including procurability; number of Lipinski failures as an indicator of *in vitro* screening amenability; properties including physical form, volatility; only containing elements C,H,O,N,P,S; Halogens and Si when not adjacent to an O; and not on the existing ToxCast inventory (more likely to be data-poor).
  - Availability of vendors was based on the vendor information accessible on the Pubchem website on a per chemical basis. Physical form was inferred on the basis of predicted melting and boiling points using OPERA[2] - a substance with a melting point greater than 25 deg C would be a solid, whereas a melting point less than 25 deg C would be a liquid and a boiling point greater than 25 deg C would be a gas.
  - Vapour pressure using a threshold of 100 mmHg was used as a surrogate for volatility.
  - Lipinski failures and vapour pressure were intended to characterise potential technical constraints for NAM testing and assessment.
  - Analytical method detection amenability predictions[3] for liquid-chromatography mass spectrometry were also generated to provide an indication of which chemicals lent themselves to screening. Internal Dashboard lists were also used to identify potential explosive or highly reactive substances.
  - Category size also played a role in prioritising selection using thresholds of less than 20, between 20-70, between 70-150, between 150-300 and final between 300-600.
  - A final manual check reviewed all 318 proposed candidate substances for testing.

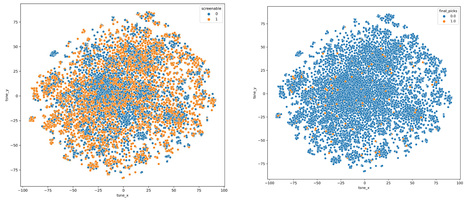
EVALUATION OF THE SELECTED SUBSET



t-SNE plots based on Morgan fingerprints to show :

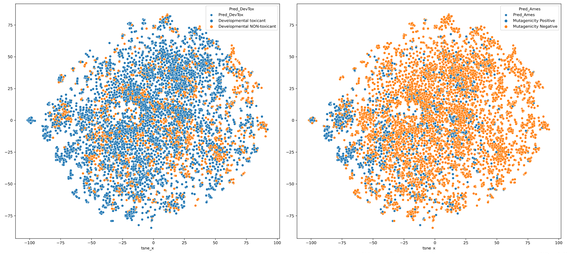
- (a) how much of the TSCA Active inventory is potentially screenable by NAMs
- (b) where the selected substances lie relative to the full TSCA inventory

- Histograms of the predicted physicochemical properties using OPERA to compare the selected subset relative to the full inventory.
- Generally across these properties, the diverse pick of substances represent the range of properties well.



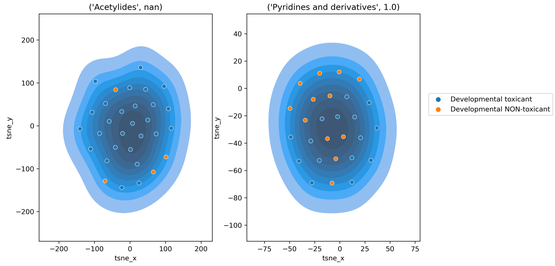
(a) (b)

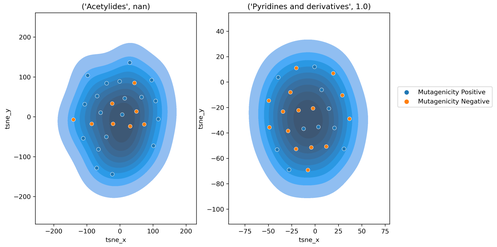
SELECTED TOXICITY PROFILES OF TSCA SPACE



- t-SNE plots of the full inventory colour coded with predicted developmental toxicity and Ames mutagenicity outcomes using the expert system TEST.

- t-SNE plots of two selected terminal categories colour coded with predicted developmental toxicity and Ames mutagenicity outcomes using the expert system TEST[4].





- More consistency in Developmental toxicity is observed across the Acetylides category than the Pyridines & derivatives category. Further work will explore these categories in more detail and profile relative to known structural alerts.

CONCLUSIONS

- Potential candidate substances were selected on the basis of being structurally similar to the centroid and meeting the other screenability conditions. The final set of ~300 chemicals will undergo analytical quality control and screening in a range of broad and targeted biological technologies for human health relevant endpoints.
- A cheminformatics analysis of the TSCA inventory with respect to these structural categories is in progress to refine the categories based on other information such as predicted toxicity profiles including structural alerts.

References:
1) doi: 10.1186/s13321-016-0174-y 2) doi: 10.1186/s13321-018-0263-1
3) doi: 10.1007/s00216-021-03713-w. 4) https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

Disclaimer: The views expressed are those of the authors and do not necessarily reflect the views or policies of the US EPA.