Published in final edited form as:

Environ Int. 2017 April ; 101: 19–26. doi:10.1016/j.envint.2016.10.006.

Incorporating ToxCast and Tox21 Datasets to Rank Biological Activity of Chemicals at Superfund Sites in North Carolina

Sloane K. Tilley¹, David M. Reif², and Rebecca C. Fry^{1,3,*}

¹Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina, Chapel Hill, NC, 27599, USA

²Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, NC, 27695, USA

³Curriculum in Toxicology, The University of North Carolina, Chapel Hill, NC, 27599, USA

Abstract

Background—The Superfund program of the Environmental Protection Agency (EPA) was established in 1980 to address public health concerns posed by toxic substances released into the environment in the United States. Forty-two of the 1328 hazardous waste sites that remain on the Superfund National Priority List are located in the state of North Carolina.

Methods—We set out to develop a database that contained information on both the prevalence and biological activity of chemicals present at Superfund sites in North Carolina. A chemical characterization tool, the Toxicological Priority Index (ToxPi), was used to rank the biological activity of these chemicals based on their predicted bioavailability, documented associations with biological pathways, and activity in *in vitro* assays of the ToxCast and Tox21 programs.

Results—The ten most prevalent chemicals found at North Carolina Superfund sites were chromium, trichloroethene, lead, tetrachloroethene, arsenic, benzene, manganese, 1,2-dichloroethane, nickel, and barium. For all chemicals found at North Carolina Superfund sites, ToxPi analysis was used to rank their biological activity. Through this data integration, residual pesticides and organic solvents were identified to be some of the most highly-ranking predicted bioactive chemicals. This study provides a novel methodology for creating state or regional databases of Superfund sites.

Conclusions—These data represent a novel integrated profile of the most prevalent chemicals at North Carolina Superfund sites. This information, and the associated methodology, is useful to toxicologists, risk assessors, and the communities living in close proximity to these sites.

Keywords

North Carolina; Risk Assessment; Superfund; Tox21; ToxCast; ToxPi

*Corresponding Author: Dr. Rebecca C. Fry, Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, CB 7431, University of North Carolina, Chapel Hill, NC, 27599, USA. rfry@unc.edu.

Competing interests

The authors declare that they have no competing interests.

1. Introduction

To conclude the “decade of the environment” of the 1970s, the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), also known as Superfund, was established to facilitate the remediation of abandoned hazardous waste sites around the United States (EPA, 2015a). Since its induction in 1980, over 1700 sites have been added to the National Priorities List (NPL), which is used by the Environmental Protection Agency (EPA) to prioritize the cleanup of sites that pose the greatest potential risk to public health. In 2006, the EPA celebrated the completion of construction at the one thousandth Superfund site (EPA, 2015a). The cleanup of these hazardous waste sites has had numerous public health and economical benefits on the surrounding communities (Arndt et al., 1999; EPA, 2011; Gamper-rabindran et al., 2011; Henry et al., 2015). However, many Superfund sites remain as hazardous waste sites that pose significant threats to human health through contaminant exposure via soil, surface water, or ground water (Carlin et al., 2015; Jung et al., 2011; Messier et al., 2012; U.S. EPA, 2013).

Environmental exposures to chemicals contribute to numerous adverse health effects in humans, including carcinogenesis, nephrotoxicity, hepatotoxicity, and immune system dysfunction (Prüss-Ustün et al., 2011). The World Health Organization (WHO) estimated that approximately 23% of deaths around the globe in 2012 were attributable to modifiable environmental risk factors (World Health Organization, 2016). Cleaning up areas known to be contaminated with toxicants harmful to human health, therefore, represents a significant way to reduce disease risk in a community (EPA, 2011). However, there are many difficulties associated with remediation of a hazardous waste site, including characterization of likely routes of exposure, the complications of exposures to chemical mixtures, consideration of health effects in susceptible populations, determining the extent of cleanup needed at a particular site, and prioritization of sites for cleanup (Anderson et al., 2015; Harris and Wrenn, 1988). The cleanup process is made more difficult by the paucity of data detailing the mechanism of action by which the more than 200 chemicals commonly found at Superfund sites could potentially harm human health and the dose at which they pose a threat (Committee on Developmental Toxicology et al., 2000; Cronin, 2003).

In the state of North Carolina, there are currently 42 Superfund sites on the NPL, spread throughout 22 different counties (EPA, 2016a). The University of North Carolina Superfund Research Program (UNC SRP) aspires to further identify, characterize, and communicate the human and environmental health risks associated with chemicals found at these sites (“UNC Superfund Research Program,” 2016). Although these hazardous waste sites have been added to the NPL, there is an insufficiency of data on which chemicals are present at North Carolina Superfund sites and what potential health risks they pose. While the EPA maintains information about Superfund site status and history, the relative biological activities of the contaminants located at each site are not publically available. Without these data, both the characterization and prioritization of potential risks that remain at Superfund sites remain difficult.

In this study, the ToxPi computational chemical prioritization tool was used to compare the biological activity of chemicals found at North Carolina Superfund sites (Reif et al., 2013,

2010). The advantage of using ToxPi is the ability to integrate different types of data, including experimental and predicted chemical properties of compounds, pathway analyses, and *in vitro* assays into a comprehensive predictive profile of a chemical's biological activity. In addition, our analysis provides a ranking system that could be used for chemical prioritization in remediation efforts at many hazardous waste sites throughout the United States. These results allowed us to aggregate data on each of North Carolina Superfund sites to create a database and methodology useful to both risk assessors and community members.

2. Materials and Methods

2.1 Identifying North Carolina Superfund Sites

Superfund sites on the NPL and Superfund Alternative Approach sites in North Carolina were identified from the EPA Superfund website. Geographic locations were plotted by street address using CartoDB (<https://cartodb.com/>). Sites that were deleted from the NPL were excluded from further analysis, as remediation has been completed, and the EPA has determined that there is no longer a threat of human exposure. These were Martin-Marietta, Sodyeco, Inc., New Hanover County Airport Burn Pit, and the 210 Mile Roadside PCB Spill. Additionally, sites that are currently undergoing site investigation were excluded from further analysis since no information about contaminants present at the site was available at the time of analysis. These included Kerr-McGee Chemical Corporation and Cristex Drum.

2.2 Data Extraction and Compilation

2.2.1 Superfund Chemical Data Extraction and Compilation—Contaminant lists of concern were obtained for every site, where available. Several sites had incomplete chemical characterization, with only a few chemicals listed under the site description, but no official list of contaminants of concern. For these sites, which were CTS – Asheville, Ecusta Mill, Hemphill Road TCE, Holtra Chem, and Wright Chemical Corporation, the chemicals mentioned in the description of the site were used instead of a list of contaminants of concern. Chemical Abstracts Service Numbers (CASRN) were obtained by querying chemical names in the Aggregated Computational Toxicology Resource (ACToR) database. Data from chemicals with the same CASRN were combined so that each CASRN only corresponded to one chemical name.

2.2.2 Chemical and Physical Property Data Extraction and Compilation—From chemicalize.org by ChemAxon, the following data were gathered for each chemical by CASRN: average molecular weight (g/mol), polar surface area ($\text{\AA}^2$), 1-octanol-water coefficient (logP), bioavailability score (0/1), Lipinski's Rule of Five (0/1), Ghose Filter (0/1), Veber's Rule (0/1), Lead-like Rule (0/1), and Muegge Filter (0/1). Bioavailability is a measure of whether the compound is orally bioavailable upon ingestion. The remaining five filters are used in pharmacology to estimate the solubility and permeability of orally active compounds based on their physical and chemical properties (Bickerton et al., 2012; Lajiness et al., 2004; Lipinski, 2004; Lipinski and Hopkins, 2004; Muegge, 2002). In addition, the aqueous solubility, logS, of each compound was predicted in ALOGPS 2.1, if it was available in the database (Tetko et al., 2005; "VCCLAB, Virtual Computational Chemistry

Laboratory,” 2005). Where available, experimental logP and logS values were recorded instead of the predicted values.

2.2.3 KEGG Pathway Data Extraction and Compilation—Using the Comparative Toxicogenomics Database (CTD), each of the toxicants were queried in the “Chemicals” section of the CTD by CASRN, if available. In the cases where the substance did not have a CASRN (e.g. Pesticides), the name of the compound was queried instead. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that were significantly enriched ($p < 0.01$) among genes that interact with that chemical were downloaded and categorized by the functional annotations provided by the KEGG Pathway Database (<http://www.genome.jp/kegg/pathway.html>). The significance of enrichment was calculated by the hypergeometric distribution and adjusted for multiple testing using the Bonferroni method (CTD, 2016). Where pathway data were not available for a specific chemical, the parent compound pathway data were downloaded and used instead so all chemicals had KEGG pathway annotations.

2.2.4 ToxCast and Tox21 Assay Data Extraction and Compilation—Phase II ToxCast and Tox21 Summary Files were downloaded from the EPA website (https://www3.epa.gov/research/COMPTOX/toxcast_summary.html), with information from 353 *in vitro* assays, corresponding to 1,191 assay endpoints. A total of 9,076 chemicals had data available. All assays used in the Tox21 program are required to be compatible with the National Center for Advancing Translational Sciences assay criteria, which specify acceptable parameters of assay validation (Sittampalam et al., 2016). Before use in the Tox21 program, all assays are optimized and validated for reproducibility in triplicate against a reference library that includes positive controls (NCATS, 2016). In addition, all Standard Operating Procedures for each assay and the concentration-response curves for all assays for each chemical tested are made publically available (NCATS, 2016). Through this robust screening procedure, all assays used in the Tox21 program adhere to the overall validation framework developed by the Organization of Economic Co-operation and Development in collaboration with both the Interagency Coordinating Committee on the Validation of Alternative Methods and the European Centre for the Validation of Alternative Methods (Judson et al., 2013). Assay data were downloaded from the EPA website in the file ‘Assay_Summary_151020.csv’. The hit-call matrix identified chemicals by CASRN number, except for explicit mixtures and pharmaceutical compounds, where arbitrary identifiers were assigned beginning with “NOCAS”. If data were available, chemicals found at North Carolina Superfund sites were matched to chemicals assayed in the ToxCast and Tox21 programs. For the remaining chemicals, ToxCast and Tox21 data from isomers or compounds of these chemicals were used instead, where available. Where multiple compounds were available for one chemical, the following preference ranking was used: isomers, acetates, chlorides, sodium anhydrides, allyls/benzyls/oxides. A total of 127 chemicals present at North Carolina Superfund sites had *in vitro* assay data available for this analysis. All annotation information can be found in the Supplementary Information.

2.3 Data Analysis

2.3.1 Prevalence ranking of chemicals at N.C. Superfund Sites—Chemical presence at each N.C. Superfund site was given a binary score, as data on chemical concentrations were not available. Scores were analyzed across all sites to determine percent presence. Chemicals were ranked in order of presence at the highest percentage of N.C. Superfund sites.

2.3.2. ToxPi Analysis—The dataset was separated into three main categories of data: chemical properties, pathways, and assays. Within each of these larger annotations, data were broken down into ToxPi components based on functional relevance. All chemicals were compared and ranked in ToxPi. Chemicals that were not evaluated in ToxCast/Tox21 were input into ToxPi as having missing values for all assay endpoints. Each of the three data types (chemical properties, pathways, and assays) was assigned total equal weights of the ToxPi (33.3% per type). In order to ensure that the lack of assay data for untested chemicals did not bias the results, three additional supplementary analyses were performed. First, only chemicals for which there were both pathway and assay data were evaluated in ToxPi. Second, all chemicals were evaluated in ToxPi using only the KEGG pathway data. Third, only chemicals for which there were assay data available were evaluated in ToxPi only using the assay data. Chemical properties data were used in all of these supplementary analyses. The purpose of the supplementary analyses was to elucidate if one dataset was driving the results and to compare bioactivity-ranking profiles using pathway and *in vitro* data. In the two supplementary analyses that utilized just one of either the pathways or assays data types, the chemical properties were assigned 33.3% of the ToxPi and the other component was assigned the remaining 66.7% of the ToxPi. The weight of each individual component for all analyses was then calculated by the following formula:

$$\text{Slice Weight} = \frac{\text{Number of Components in Slice}}{\text{Number of Components in Data Type}} \times \text{Total Data Type Percentage of ToxPi}$$

Component weights were rounded to the nearest whole number. Data normalization schemes were applied so that higher values would be indicative of a more toxic chemical, and thus correspond to a larger ToxPi radius. All missing data values were entered as “NA”.

The chemical properties annotation consisted of four components – (1) LogP, (2) LogS, (3) Polar Surface Area and Molecular Weight, and (4) Bioavailability. These four components were combined in the overall ToxPi score to represent a relative quantitative ranking of biological availability of each compound. Only one value comprised each of the LogP and LogS components. These were the (LogP value + absolute value(minimum (LogP values))) and the (LogS value + absolute value(minimum (LogS values))) for each chemical. The absolute value of the minimum value of each component was added to all values so that negative values would not be interpreted as missing values by ToxPi. LogP values were scored in ToxPi using the built-in $-\ln(x) + \max(x)$ normalization scheme, as higher LogP values are indicative of poor absorption and thus would be assigned lower values within ToxPi, as they are less readily absorbed by the body and thus assumed to be less biological active. LogS values were scored in ToxPi using the built-in linear(x) normalization scheme,

as higher LogS values indicate higher water solubility, which is a marker for good bodily absorption. Thus, higher LogS values were scored as more biological active and assigned larger ToxPi radii. Two values made up the Polar Surface Area and Molecular Weight component. These were, polar surface area (square angstroms ($\text{\AA}^2$)) and molecular weight (g/mol), both of which were obtained from the Human Metabolome Database (HMDB) recorded ChemAxon predictions. Low values of both of these characteristics are indicative of high absorption, and thus this component was scored using the $-\ln(x) + \max(x)$ scheme in ToxPi, where compounds with lower polar surface areas and molecular weights would be assigned higher values in ToxPi. The last chemical property component, Bioavailability, was comprised of the chemical's score in Bioavailability, Lipinski's Rule of Five, Ghose Filter, Veber's Rule, Lead-like Rule, and Muegge Filter. The Bioavailability component was scored by the hit-count method using ToxPi. The more drug-like filters the compound satisfied, the more biologically available, and thus potentially biological active, it was scored. The pathway data consisted of 387 KEGG pathways that were assigned into seven ToxPi components based on the KEGG pathway annotation. Corrected p-values were entered for each chemical into ToxPi. For chemicals that did not reach statistical significance for enrichment in a KEGG pathway, a "1" was entered in order to record that the data were not missing, but not significant. All pathway components were scored using the $-\ln(x) + \max(x)$ scheme in ToxPi so that smaller, more significant p-values corresponded to larger ToxPi radii. The ToxCast and Tox21 assay data were grouped into sixteen ToxPi components by their intended target family, as suggested in the Assay Annotation User Guide. Assay endpoints that could not be determined were entered as missing data. All assay data were scored using the hit-count operation in ToxPi.

2.3.3 Hierarchical Clustering of Chemicals at North Carolina Superfund Sites Based on Biological Activity Profiles—

In order to discover how different chemicals found at North Carolina Superfund sites grouped according to their biological activity profiles, unsupervised hierarchical clustering by Euclidean distance by Ward's Method on mean standardized ToxPi statistics was used. Analyses were carried out using Partek Genomics Suite (Partek Genomics Suite, St Louis, Missouri).

2.3.4 Summary Characterization of North Carolina Superfund Sites—

In order to provide a more accurate prioritization ranking of North Carolina Superfund sites, we summed the overall ToxPi scores for each chemical found at every site into a cohesive score for each site. In order to account for potential artificial lowering of summed ToxPi scores at sites which contained chemicals missing ToxCast/Tox21 data, we adjusted the summed ToxPi score for each site using the following formula,

$$\text{Adjusted Summed ToxPi Score} = \text{ToxPi Summed Score} + (n \cdot t),$$

where n is the number of chemicals at each site for which ToxCast/Tox21 data is not available, and t is the mean of summed ToxPi scores from only the assay components of the algorithm for the n=127 chemicals for which ToxCast/Tox21 data was available. In addition, for each site, it was recorded whether site construction was complete, human exposure was under control, ground water contamination was under control, and if the site was ready for

anticipated use. There were some sites where the label of “Insufficient Data” was listed as the current status for one or more of these outcomes. Estimates of population size living within 1 mile of each Superfund site were obtained for all sites included in our analysis, where available. All of these data are publically available on every sites’ Site Facts page. Superfund Alternative Approach Sites, Battery Tech/Duracell and Flander’s Filter’s Inc., did not have population and site status data available.

3. Results and Discussion

3.1 North Carolina Superfund Sites

The first step in our analysis was to compile and analyze all North Carolina Superfund sites. A final list of 43 Superfund sites was analyzed and included 37 final NPL sites and 6 non-NPL sites (Figure 1). The non-NPL sites are Superfund Alternative Approach (SAA) sites that are eligible for inclusion on the NPL and require remedial action cleanup (EPA, 2015b). As such, it was determined that these sites met the criteria for inclusion.

3.2 Prevalence of Chemicals at North Carolina Superfund Sites

Next, we aimed to determine the relative prevalence of each toxicant present at North Carolina Superfund sites. In total, 158 unique contaminants were present at one or more North Carolina Superfund site (Supplementary Table 1). The most common toxicant found at North Carolina Superfund sites was chromium, followed by trichloroethene, lead, tetrachloroethene, and arsenic. Interestingly, among the 29 contaminants present at >25% of North Carolina Superfund sites, 13 (45%) were metals or metalloids (Figure 2). Many metals are naturally occurring in North Carolina (Kim et al., 2011; Sanders et al., 2014, 2011; Spangler and Spangler, 2009), but are also used for various commercial purposes, including wood preservation, ingredients of dyes and pigments, and in production of electronics (Miller et al., 2016; Zeng et al., 2016). Also common among the chemicals frequently found at North Carolina Superfund sites were industrial organic solvents, such as trichloroethene and tetrachloroethene. These solvents are used in a wide variety of operations, including production of plastics, furniture, and fuel production (Falck et al., 2015; Huang et al., 2014). Forty-six of 158 (29%) contaminants were found at only one North Carolina Superfund site (Supplementary Table 1).

3.3 Top Ranking Chemicals Enriched for Bioactivity at North Carolina Superfund Sites

To rank chemicals according to biological activity, we integrated chemical property data, pathway data, and in vitro assay data in ToxPi. The top 10 ranking toxic chemicals from the ToxPi analysis were methylcyclohexane, 1,3,5-trimethylbenzene, endosulfan, metals/ unspecified mixture, bis(2-chloroethyl) ether, dibutyl phthalate, guthion, methoxychlor, 1,3-dichlorobenzene, and chloride (Figure 3, Supplementary Table 2). Interestingly, these top-ranking bioactive chemicals were some of the rarest chemicals at North Carolina Superfund sites, with none appearing at more than four sites (Supplementary Table 1). These top-ranked chemicals are used primarily as chemical reagents and pesticides or insecticides, with the exception of dibutyl phthalate, which is a plasticizer, and metals, which have a variety of uses in industry in addition to being naturally occurring in North Carolina (Sanders et al., 2011). Nine out of these top 10 compounds showed significant KEGG pathway enrichment

for cancer. Interestingly, none displayed enrichment for DNA binding assays *in vitro*, suggesting that the carcinogenic effects of these chemicals may be potentiated through non-genotoxic mechanisms such as altered cellular signaling (Figure 3, Supplementary Table 3).

Interestingly, the top two ranking chemicals, namely methylcyclohexane and 1,3,5-trimethylbenzene, are anthropogenic benzene derivatives. Specifically, methylcyclohexane is used as an organic solvent and 1,3,5-trimethylbenzene is a precursor in the processes of making pure chemical substances (Bennett, 2002). These two compounds were highly-ranked among all 158 chemicals due to enrichment in all KEGG pathway annotations, including metabolism and genetic information processing (Figure 3, Supplementary Table 3).

3.4 Clusters of Chemicals at North Carolina Superfund Sites

In order to visualize the functional enrichment among the most biological active chemicals as determined using the ToxPi ranking system, unsupervised hierarchical clustering on mean standardized ToxPi statistics was used. This revealed 10 clusters of chemicals, including four clusters that were enriched for biological activity in ToxCast and Tox21 assays, two clusters that were highly enriched among KEGG pathway annotations, and four clusters that displayed little or no enrichment for bioactivity markers in either the pathway or assay data (Figure 4, Table 1). Interestingly, almost all chemicals found at North Carolina Superfund sites that were enriched in the ToxCast and Tox21 assays were pesticides. Given a well-known mechanism of pesticide toxicity is via binding to nuclear receptors (Tebourbi et al., 2006), it was not surprising that all of these top-ranking insecticides showed enrichment among assays testing for dysregulation of nuclear receptors. In addition, all of these compounds were scored as being extremely bioavailable, as they are designed to be lipophilic in order to permeate cell membranes and to bioaccumulate (Brooks and Roberts, 1999).

3.5 Summary of ToxPi Scores, Populations, and Site Status of North Carolina Superfund Sites

In order to provide a comprehensive profile of each North Carolina Superfund site, an adjusted summed ToxPi score accounting for missing data, the 2010 Census estimation of population living within a 1-mile radius of the site, site location, and four parameters of site status were integrated into one database (Supplementary Table 4). The summed, adjusted ToxPi scores ranged from the lowest (22.58) at Holcomb Creosote Company to the highest (1804.41) at Camp Lejeune (Supplementary Table 4). Gaston County contained the largest number of Superfund sites of any one county, with four sites located within its boundaries. Fourteen of the 42 sites (33%) had upper bounds of populations living within a 1-mile radius of the site of 100 people, while only four of the 42 sites (9.5%) had an upper estimate greater than or equal to 10,000.

At the majority ($25/42 = 60\%$) of these sites, site construction has been completed, meaning that necessary construction measures have been completed or that they were not deemed necessary. Only one site, namely, CTS-Mills Gap located in Asheville, NC, was listed as not having human exposure under control, although for an additional 5 sites this condition

classification was listed as “Insufficient Data.” Thus, at the vast majority of sites (35/42 = 83%), human exposure is considered to be under control. Similarly, ground water contamination was documented to be under control at 27 of 42 sites (64%), with 2 (4.8%) and 12 (29%) of the 42 sites documented as not under control or as having insufficient data, respectively. Interestingly, only 6 of the 42 (14%) sites were documented as “Site-Wide Ready for Anticipated Use,” meaning that 35 out of 42 sites (83%) have yet to “(i) meet all the cleanup goals of media may affect current and reasonably anticipated future land uses of the site, so that there are no unacceptable risks; and (ii) all institutional or other controls required in the Record(s) of Decision or other remedy decision document(s) have been put in place (EPA, 2016b).” At many of the sites that were not documented as “Site-Wide Ready for Anticipated Use,” the characterization of the risks at the site presently contained a statement such that indicated some documentation of potential continued exposure.

3.5 Study Limitations

While this study represents a novel integrated approach to characterizing prevalence and biological activity of chemicals present at Superfund sites, it is not without limitations. First, as this is the first study to relatively characterize the biological activities of chemicals found at North Carolina Superfund sites, there is no current way to validate the ranking. Second, there is a missingness of data for some of the most prevalent North Carolina Superfund chemicals within the ToxCast/Tox21 databases. This was specifically true for toxic metals. We addressed this limitation in our study by adjusting the overall ToxPi site scores according to how many chemicals with missing data were present at each site. In addition, this missing data did not seem to affect the relative ranking of bioactivity of the chemicals themselves, as shown by the similar results obtained when excluding the assay data for all chemicals (Supplementary Table 2). Third, we were limited in our assessment of North Carolina Superfund sites due to the data that were publically available. Specifically, there was a lack of any quantitative measures of chemical concentration at North Carolina Superfund sites. As chemicals’ interactions have often been shown to be dose-dependent (Sarigiannis and Hansen, 2012), this lack of quantitative data prohibited any attempts of accurate chemical mixture or co-exposure assessment. Due to this limitation, the adjusted summed ToxPi score may be over-simplistic in the assumption that the bioactivity of all of these compounds is necessarily additive. In addition, insufficient data were available on the environmental conditions in which each contaminant existed. In combination with a lack of quantitative data, the paucity of data on the type of media that contained each contaminant at sites and inability to account for variable weather patterns, soil composition, and other natural processes affecting biodegradation, precluded an assessment of biopersistence of contaminants at each site. Finally, the ability of *in vitro* data to predict *in vivo* data is contested. As toxicological models aim to limit the use of animals in favor of *in vitro* systems, the question of whether high-throughput screening (HTS) systems reliably predict *in vivo* health endpoints becomes increasingly important. ToxCast data has been used to demonstrate significant associations between *in vitro* assay results and *in vivo* endpoints (Judson et al., 2010; Kleinstreuer et al., 2013). Yet, in a comprehensive analysis that assessed the predictive performance of the ToxCast Phase I *in vitro* assays for more than 60 *in vivo* endpoints, it was found that HTS did not accurately predict *in vivo* health effects (Thomas et al., 2012). However, the authors do emphasize the positive performance of the

assays in predicting potential molecular initiating events or key events along adverse outcome pathways of toxicity, specifically noting that the *in vitro* assays represent useful tools for chemical prioritization (Thomas et al., 2012). To address this limitation, Phase III of the ToxCast program aims to use more biologically relevant cell lines and some three-dimensional model organ systems to improve *in vitro* to *in vivo* predictability (NTP, 2016).

3. CONCLUSIONS

These results provide a comprehensive description of the prevalence and predicted biological activity of chemicals found at Superfund sites in North Carolina. Metals remain highly prevalent at North Carolina Superfund sites, along with organic solvents and residual pesticides. These findings also highlight the advantage and importance of integrating *in vitro* and pathway data for ranking chemicals based on enriched biological activity profiles. In addition, these analyses represent a novel methodology that could be applied at a larger scale to establish databases of Superfund sites across the United States. We believe that these data could be used as Tier III toxicity information in the EPA's assessment of risks present at Superfund sites, as they represent novel, peer-reviewed, publically available toxicity values for contaminants of concern (Cook, 2003). Alternatively, this prioritization could be used as a starting point for a ToxPi model that integrates these data with expert knowledge (additional data sources, reapportionment of assays, etc.) into formal ranking system. By incorporating assessments of both chemical prevalence at sites and individual chemicals' enrichment bioactivity profiles, this methodology could allow for policy makers and the Integrated Risk Information System (IRIS) at the EPA to better assess the relative risk posed by Superfund sites in different regions to help with the prioritization of sites for clean-up and to more accurately forecast the extent and complexity of the remediation effort.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This research was supported by grants from the National Institutes of Health (<http://www.nih.gov>): R01 ES019315, P42 ES005948, and P30ES025128. The funders of this study had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations

ATSDR	Agency of Toxic Substances and Disease Registry
CASRN	Chemical Abstracts Service Numbers
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CTD	Comparative Toxicogenomics Database
EPA	Environmental Protection Agency

HMDB	Human Metabolome Database
IRIS	Integrated Risk Information Systems
KEGG	Kyoto Encyclopedia of Genes and Genomes
N.C.	North Carolina
NPL	National Priorities List
SAA	Superfund Alternative Approach
ToxPi	Toxicological Priority Index
UNC SRP	University of North Carolina Superfund Research Program
WHO	World Health Organization

References

- Anderson BE, Naujokas MF, Suk WA. Interweaving Knowledge Resources to Address Complex Environmental Health Challenges. *Environ Health Perspect.* 2015; 123doi: 10.1289/ehp.1409525
- Arndt V, Vine MF, Weigle K. Environmental chemical exposures and risk of herpes zoster. *Environ Health Perspect.* 1999; 107:835–41. [PubMed: 10504152]
- Bennett, GF. *J Hazard Mater.* 4th. 2002. Sittig's Handbook of Toxic and Hazardous Chemicals and Carcinogens.
- Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat Chem.* 2012; 4:90–98. DOI: 10.1038/nchem.1243 [PubMed: 22270643]
- Brooks, GT., Roberts, T. *Pesticide Chemistry and Bioscience: The Food-Environment Challenge.* Elsevier Science; 1999. Woodhead Publishing Series in Food Science, Technology and Nutrition
- Carlin DJ, Naujokas MF, Bradham KD, Cowden J, Heacock M, Henry HF, Lee JS, Thomas DJ, Thompson C, Tokar EJ, Waalkes MP, Birnbaum LS, Suk WA. Arsenic and Environmental Health: State of the Science and Future Research Opportunities. *Environ Health Perspect.* 2015; doi: 10.1289/ehp.1510209
- Committee on Developmental Toxicology, Board on Environmental Studies and Toxicology, Commission on Life Sciences. , editor. *Scientific Frontiers in Developmental Toxicology and Risk Assessment.* National Academies Press; Washington, DC: 2000.
- Cook, MB. *Human Health Toxicity Values in Superfund Risk Assessments,* Environmental Protection. Washington, DC; 2003.
- Cronin, MTD. Quantitative Structure-Activity Relationships for Acute Aquatic Toxicity: The Role of Mechanism of Toxic Action in Successful Modeling. In: Benigni, R., editor. *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens.* CRC Press; Boca Raton: 2003. p. 237
- CTD. Chemical Enriched Pathways Help [WWW Document]. 2016. URL <http://ctdbase.org/help/chemPathwayDetailHelp.jsp?popup=true#fn1> (accessed 4.15.16)
- EPA. Search for Superfund Sites Where You Live [WWW Document]. 2016a. URL <http://www.epa.gov/superfund/search-superfund-sites-where-you-live> (accessed 10.14.15)
- EPA. Glossary for the Superfund Site Profile [WWW Document]. 2016b. URL <https://cumulis.epa.gov/supercpad/SiteProfiles/index.cfm?fuseaction=modules.glossary&id=0#GPRA> (accessed 5.2.16)
- EPA. Superfund History [WWW Document]. 2015a. URL <http://www2.epa.gov/superfund/superfund-history> (accessed 10.14.15)
- EPA. Superfund Alternative Approach [WWW Document]. 2015b. URL <http://www.epa.gov/enforcement/superfund-alternative-approach> (accessed 2.15.16)

- EPA. Handbook on the Benefits, Costs and Impacts of Land Cleanup and Reuse 1. Washington, DC: 2011.
- Falck AJ, Mooney S, Kapoor SS, White KMR, Bearer C, El Metwally D. Developmental Exposure to Environmental Toxicants. *Pediatr Clin North Am*. 2015; 62:1173–1197. DOI: 10.1016/j.pcl.2015.05.005 [PubMed: 26318946]
- Gamper-rabindran S, Greenstone M, Grider B, Hamilton J, Harris J, Herndon J, Jenkins R, Kiel K, Kuminoff N, Marten A, Nagavarapu S, Newbold S, Phaneuf D, Sheriff G, Sigman H, Smith K, Thomas M, Taylor L, Walsh P, Walsh R. Valuing the Benefits of Superfund Site Remediation: NBER Work. Pap. 2011
- Harris RH, Wrenn GC. Making Superfund work. *Issues Sci Technol*. 1988; 4:54–58.
- Henry H, Naujokas MF, Attanayake C, Basta NT, Cheng Z, Hettiarachchi GM, Maddaloni M, Schadt C, Scheckel KG. Bioavailability-Based In Situ Remediation To Meet Future Lead (Pb) Standards in Urban Soils and Gardens. *Environ Sci Technol*. 2015; 49:8948–8958. DOI: 10.1021/acs.est.5b01693 [PubMed: 26140328]
- Huang B, Lei C, Wei C, Zeng G. Chlorinated volatile organic compounds (Cl-VOCs) in environment - sources, potential human health impacts, and current remediation technologies. *Environ Int*. 2014; 71:118–138. DOI: 10.1016/j.envint.2014.06.013 [PubMed: 25016450]
- Judson R, Kavlock R, Martin M, Reif D, Houck K, Knudsen T, Richard A, Tice RR, Whelan M, Xia M, Huang R, Austin C, Daston G, Hartung T, Fowle JR, Wooge W, Tong W, Dix D. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *ALTEX*. 2013; 30:51–66. DOI: 10.14573/altex.2013.1.051 [PubMed: 23338806]
- Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, Reif DM, Rotroff DM, Shah I, Richard AM, Dix DJ. In vitro screening of environmental chemicals for targeted testing prioritization: The toxcast project. *Environ Health Perspect*. 2010; 118:485–492. DOI: 10.1289/ehp.0901392 [PubMed: 20368123]
- Jung D, Matson CW, Collins LB, Laban G, Stapleton HM, Bickham JW, Swenberg JA, Di Giulio RT. Genotoxicity in Atlantic killifish (*Fundulus heteroclitus*) from a PAH-contaminated Superfund site on the Elizabeth River, Virginia. *Ecotoxicology*. 2011; 20:1890–9. DOI: 10.1007/s10646-011-0727-9 [PubMed: 21706406]
- Kim D, Miranda M, Tootoo J. Spatial Modeling for Groundwater Arsenic Levels in North Carolina. *Environ Sci Technol*. 2011; 45:1–15. DOI: 10.1021/es103336s.Spatial [PubMed: 21186836]
- Kleinstreuer NC, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Paul KB, Reif DM, Crofton KM, Hamilton K, Hunter R, Shah I, Judson RS. In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. *Toxicol Sci*. 2013; 131:40–55. DOI: 10.1093/toxsci/kfs285 [PubMed: 23024176]
- Lajiness MS, Vieth M, Erickson J. Molecular properties that influence oral drug-like behavior. *Curr Opin Drug Discov Devel*. 2004; 7:470–7.
- Lipinski CA. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov Today Technol*. 2004; 1:337–341. DOI: 10.1016/j.ddtec.2004.11.007 [PubMed: 24981612]
- Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. *Nature*. 2004; 432:855–861. DOI: 10.1038/nature03193 [PubMed: 15602551]
- Messier KP, Akita Y, Serre ML. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ Sci Technol*. 2012; 46:2772–80. DOI: 10.1021/es203152a [PubMed: 22264162]
- Miller MD, Marty MA, Landrigan PJ. Children's Environmental Health: Beyond National Boundaries. *Pediatr Clin North Am*. 2016; 63:149–165. doi:<http://dx.doi.org/10.1016/j.pcl.2015.08.008>. [PubMed: 26613694]
- Muegge I. Pharmacophore features of potential drugs. *Chem - A Eur J*. 2002; 8:1976–1981. doi: 10.1002/1521-3765(20020503)8:9<1976::AID-CHEM1976>3.0.CO;2-K.
- NCATS. Work with Tox21 [WWW Document]. 2016. URL <http://www.ncats.nih.gov/tox21/work> (accessed 8.23.16)
- NTP. Tox21 Research Phases [WWW Document]. 2016. URL <https://ntp.niehs.nih.gov/results/tox21/researchphases-index.html> (accessed 8.16.16)

- Prüss-Ustün A, Vickers C, Haefliger P, Bertollini R. Knowns and unknowns on burden of disease due to chemicals: a systematic review. *Environ Health*. 2011; 10:9.doi: 10.1186/1476-069X-10-9 [PubMed: 21255392]
- Reif DM, Martin MT, Tan SW, Houck KA, Judson RS, Richard AM, Knudsen TB, Dix DJ, Kavlock RJ. Endocrine profiling and prioritization of environmental chemicals using toxcast data. *Environ Health Perspect*. 2010; 118:1714–1720. DOI: 10.1289/ehp.1002180 [PubMed: 20826373]
- Reif DM, Syta M, Lock EF, Wright FA, Wilson A, Cathey T, Judson RR, Rusyn I. ToxPi GUI: An interactive visualization tool for transparent integration of data from diverse sources of evidence. *Bioinformatics*. 2013; 29:402–403. DOI: 10.1093/bioinformatics/bts686 [PubMed: 23202747]
- Sanders AP, Desrosiers TA, Warren JL, Herring AH, Enright D, Olshan AF, Meyer RE, Fry RC. Association between arsenic, cadmium, manganese, and lead levels in private wells and birth defects prevalence in North Carolina: a semi-ecologic study. *BMC Public Health*. 2014; 14:955.doi: 10.1186/1471-2458-14-955 [PubMed: 25224535]
- Sanders AP, Messier KP, Shehee M, Rudo K, Serre ML, Fry RC. Arsenic in North Carolina: Public Health Implications. *Environ Int*. 2011; 38:10–16. DOI: 10.1016/j.envint.2011.08.005.Arsenic [PubMed: 21982028]
- Sarigiannis DA, Hansen U. Considering the cumulative risk of mixtures of chemicals - A challenge for policy makers. *Environ Heal*. 2012; 11:S18.doi: 10.1186/1476-069X-11-S1-S18
- Sittampalam GS, Coussens NP, Editor AS, Arkin M, Auld D, Austin C, Bejcek B, Glicksman M, Inglese J, Iversen PW, Mcgee J, Mcmanus O, Minor L, Napper A, Peltier JM, Riss T, Trask OJ, Weidner J. *Assay Guidance Manual*. 2016:261–263.
- Spangler AH, Spangler JG. Groundwater manganese and infant mortality rate by county in north carolina: An ecological analysis. *Ecohealth*. 2009; 6:596–600. DOI: 10.1007/s10393-010-0291-4 [PubMed: 20232227]
- Tebourbi O, Sakly M, Rhouma K, Ben. *Molecular Mechanisms of Pesticide Toxicity*. 2006
- Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV. Virtual computational chemistry laboratory - Design and description. *J Comput Aided Mol Des*. 2005; 19:453–463. DOI: 10.1007/s10822-005-8694-y [PubMed: 16231203]
- Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, Andersen ME, Wolfinger RD. A Comprehensive Statistical Analysis of Predicting In Vivo Hazard Using High-Throughput In Vitro Screening. *Toxicol Sci*. 2012; 128:398–417. DOI: 10.1093/toxsci/kfs159 [PubMed: 22543276]
- U.S. EPA. Use of Amendments for In Situ Remediation at Superfund Sediment Sites. *Off Superfund Remediat. Technol Inf*. 2013 doi:OSWER Directive 9200.2-128FS.
- UNC. Superfund Research Program [WWW Document]. 2016. URL <http://sph.unc.edu/superfund-pages/srp/> (accessed 2.14.16)
- VCCLAB. Virtual Computational Chemistry Laboratory [WWW Document]. 2005. URL <http://www.vcclab.org> (accessed 11.25.15)
- World Health Organization. Preventing Disease Through Healthy Environments A Global Assessment of the Burden of Disease from Environmental Risks. 2016
- Zeng X, Xu X, Boezen HM, Huo X. Children with health impairments by heavy metals in an e-waste recycling area. *Chemosphere*. 2016; 148:408–415. DOI: 10.1016/j.chemosphere.2015.10.078 [PubMed: 26829309]

Figure 1. Forty-eight Superfund NPL or Superfund Alternative Approach sites plotted by street address. Sites are color coded by type.

Figure 2.
Twenty-nine chemicals are present at >25% of North Carolina Superfund sites.

Figure 3.
ToxPis of the top 10 predicted chemicals most enriched for bioactivity found at North Carolina Superfund sites. Overall ToxPi scores are included for each chemical.

Figure 4.

Heatmap of 158 chemicals found at N.C. Superfund sites displayed with relative toxicity rankings from ToxPi. Light blue color indicates that chemical displayed moderately enriched predicted biological activity (one standard deviation above the mean), and dark blue color indicates that chemical displayed robustly enriched predicted biological activity (two standard deviations above the mean). Gray color indicates that they chemical was not enriched for bioactivity compared to the other chemicals found at N.C. Superfund sites.

Table 1

Characterization of chemicals from 10 clusters derived from unsupervised hierarchical clustering of the component scores from ToxPi for each of the 158 chemicals found at North Carolina Superfund sites. The color panels match the colored clusters shown in Figure 4.

		Characterization	Chemicals
A		Non-Enriched Chemicals	Xylene, Dibromochloropropane, Bromoform, Dibromochloromethane, Bromodichloromethane, Volatile Organic Compounds, Tetrachloroethene, Trichloroethane, Chlorobenzene, Ethylbenzene, 1,2,4-Trimethylbenzene, Picric acid, 1,2-Dichloropropane, Trans, 1,3-Dichloropropene, 1,1-Dichloroethene, 1,2-Dichloroethene, 1,2-Dichloroethane, 1,1-Dichloroethane, Chloroform, Ethyl chloride, Vinyl chloride, Dichloromethane, Methyl chloride, Acetone, Benzoic Acid, Cyclonite, Petroleum (Decene), Chrysene, Benzo[k]fluoranthene, Acenaphthylene, 1,4-Dichlorobenzene, Naphthalene, Phenanthrene, Trichloroacetic acid, Diethyl phthalate, 1,3,5-Trinitrobenzene, 2,4,6-Trinitrotoluene, Benzoic acid, Pesticides (Metalaxyl-M), Fluorene, 3,3'-Dichlorobenzidine, Base Neutral Acids, Trichloroethene, Benzene, Methyl bromide, Carbon tetrachloride, Toluene, Creosote, Carbon disulfide, Acetophenone, Methyl isobutyl ketone, 2-Hexanone, Phenol, Methyl tert-butyl ether, Methyl ethyl ketone, 3-Methylphenol, 2-Chlorophenol, Dibenzofuran, 2-Methylphenol, Cyanide, Formaldehyde, Dinitrosodimethylamine, Antimony, Barium, Strontium, Thallium, Molybdenum, Alumina, Silver, Magnesium, Perchloric acid, Potassium, Sodium, Aluminum, Vanadium, Iron, Beryllium, Ammonia, Benzo[ghi]perylene, PCB-1260, PCB-1242, PCB-1248, Hexachlorobenzene, PCB-1254, PCB-1268, Heptachlorodibenzo-p-dioxin, Hexachlorodibenzo-p-dioxin, 2,3,7,8,- Tetrachlorodibenzofuran, Tritium, Dioctyl phthalate, n-, Indeno[1,2,3-cd]pyrene, Dibenzo[a,h]anthracene, Bis(2-ethylhexyl)phthalate
B			
C			
D			
E		Chemicals Enriched in KEGG Pathway Annotations	Mercury, Lead, Manganese, Chromium, Selenium, Calcium, Acid Compounds, Arsenic, Zinc, Cadmium, Copper, Cobalt, Sulfate, Nickel, Nitrobenzene, Isophorone, Malononitrile, 2-Methylnaphthalene, Carbazole, Benzo[a]pyrene, 2-Chloronaphthalene, Tetrachloroethane, 1,3-Dichlorobenzene, Polycyclic Aromatic Hydrocarbons, Tetrachlorodibenzo-p-dioxins/Toxicity equivalents, Dioxins/furans, Unspecified mixture, Chloride, Metals, Unspecified Mixture, 1,3,5-Trimethylbenzene, Methylcyclohexane
F			
G		Chemicals Enriched in the ToxCast and Tox21 Assays	2,4-Dinitrophenol, Dinitrosodiphenylamine, 4-Methylphenol, 3,4-Dimethylphenol, Toxaphene, Lindane, Benzylbutylphthalate, Disyston, Dieldrin, 4-Nitrophenol, 2,4-Dichlorophenol, Anthracene, 2,4-Dimethylphenol, Benzo[a]anthracene, Pyrene, Fluoranthene, Heptachlor epoxide, Acenaphthene, Benzo[b]fluoranthene, Bis(2-chloroethyl) ether, Dibutyl phthalate, Endrin, Malathion, Carbaryl, Guthion, Endosulfan Sulfate, Hexachlorobutadiene, Pentachlorophenol, Methoxychlor, Endosulfan, Dichlorodiphenyldichloroethane, Aldrin, Heptachlor, Dichlorodiphenyltrichloroethane, Chlordane
H			
I			
J			