Data Science and Cheminformatics Tools to Support Exposomics and Metabolomics

Icahn School of Medicine at Mount Sinai

Dinesh Kumar Barupal
Assistant Professor
Department of Environmental Medicine and Public Health
Icahn School of Medicine at Mt Sinai
New York, USA

Overview

- Opportunities in non-targeted analyses (NTA)
- Chemical to publication mapping
- Prioritizing chemicals for hazard assessments

Opportunities in non-targeted analyses

NTA for the disease prevention

Exposomics — Metabolomics

Risk factors	Diagnosis & screening	Treatment and prognosis
Substance registries EPA FDA	Electronic Health Records – blood chemistry MAYO CLINIC LABORATORIES	Electronic Health Records – blood chemistry MAYO CLINIC LABORATORIES
Risk assessment International Agency for Research on Cancer, World Health Organization; OEHHA	Blood chemical monitors and dried blood spot kits 	Drug targets in metabolic pathways DAILYMED
Biomonitoring CDC		

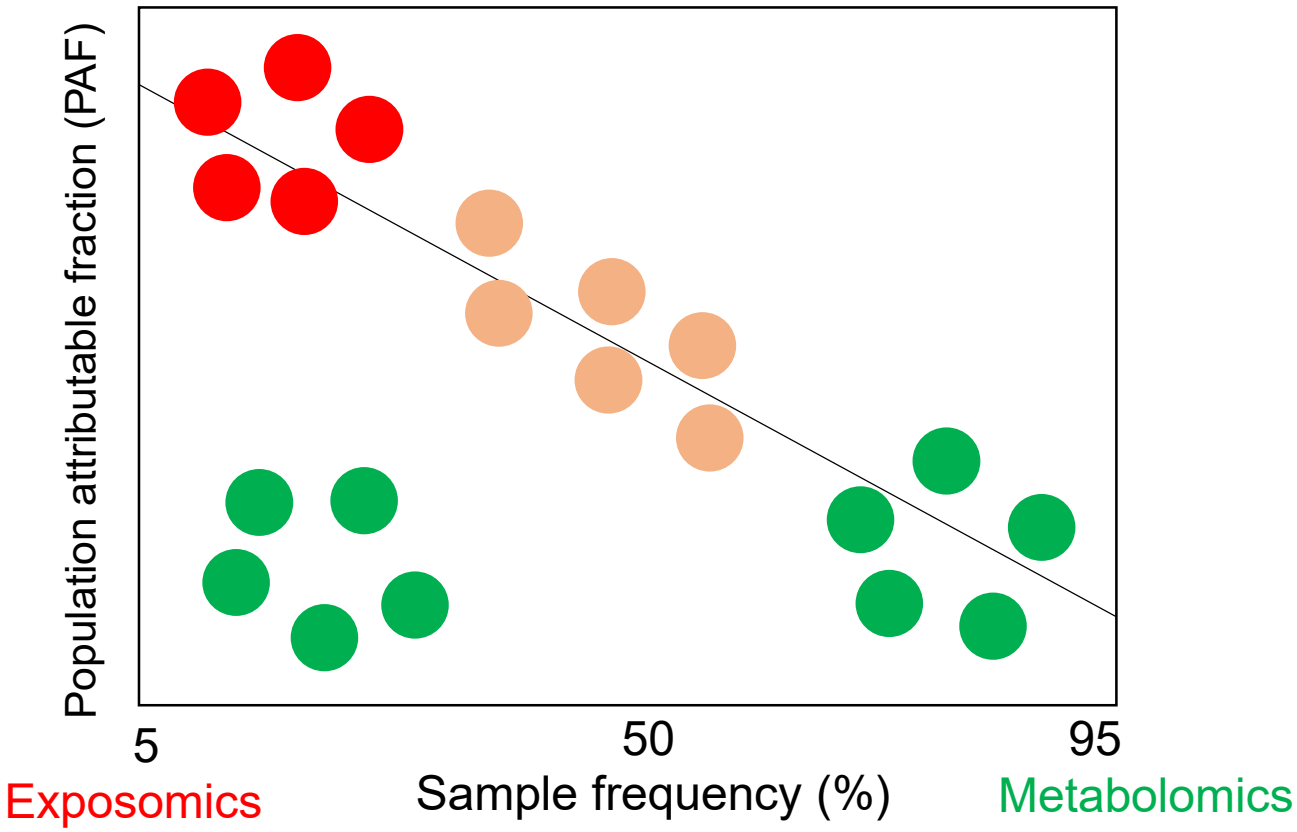
Prevention strategies

Primary — Secondary — Tertiary

Discussion point : How to prioritize NTA assays for identifying risk factors or discovering new metabolic reactions?

Rappaport, Stephen M., et al. "The blood exposome and its role in discovering causes of disease." *Environmental health perspectives* 122.8 (2014): 769-774.

Low signal prevalence is important



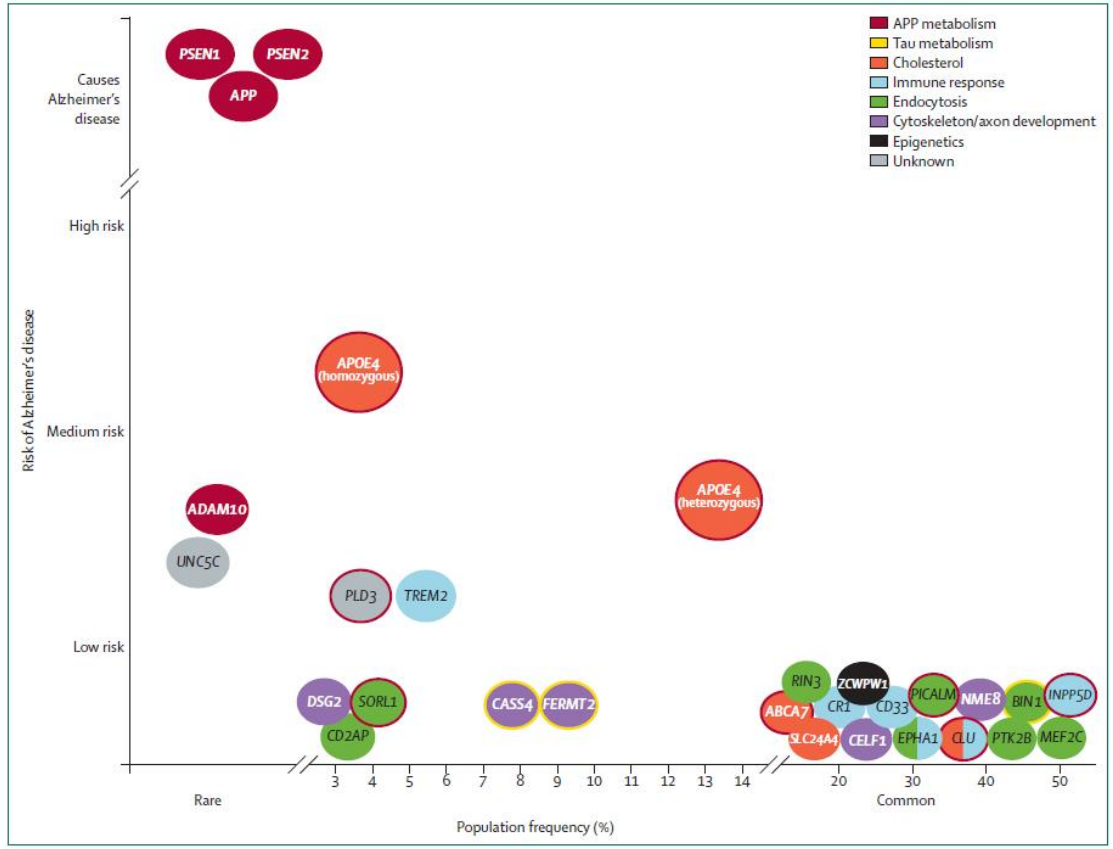


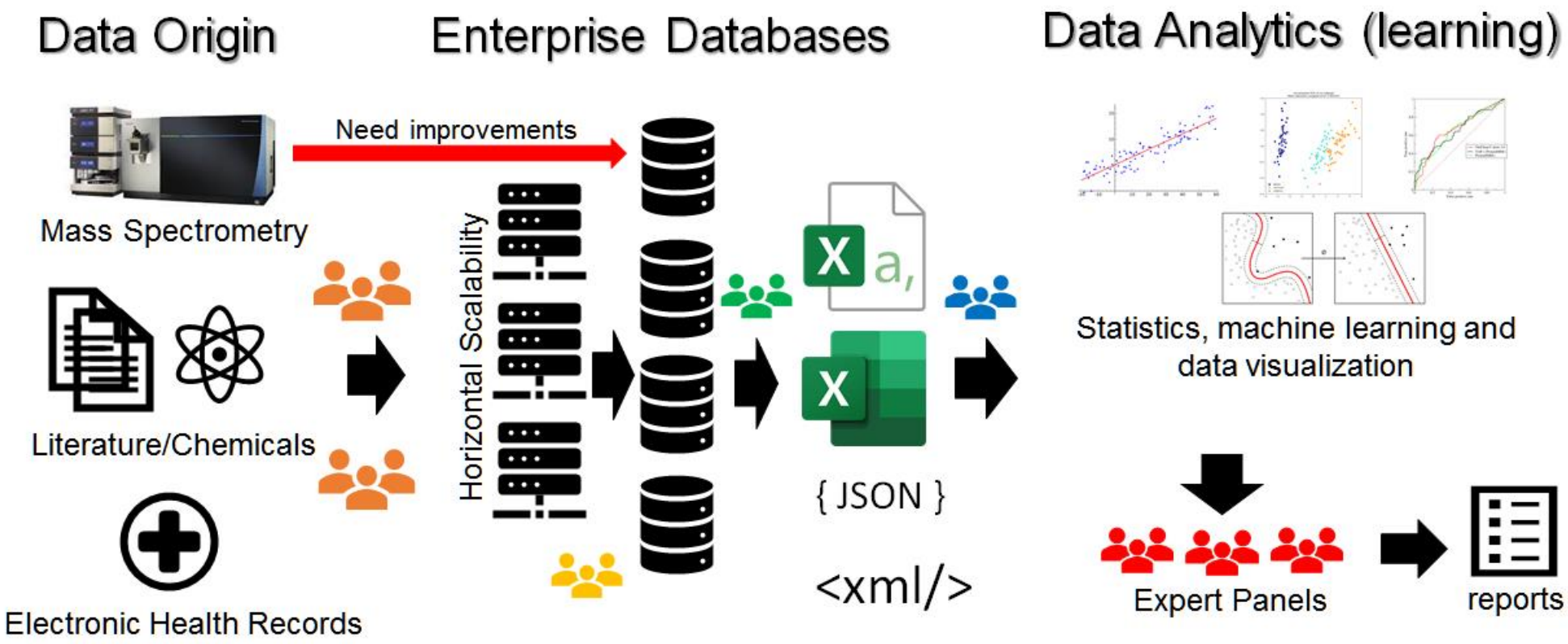
Figure: Schematic overview of genes linked to Alzheimer's disease

www.thelancet.com **Vol 388 July 30, 2016**

Raw variants carry more risk.

Discussion point : NTA studies should **avoid thresholding** signal prevalence so we don't miss rare signals with high PAFs.

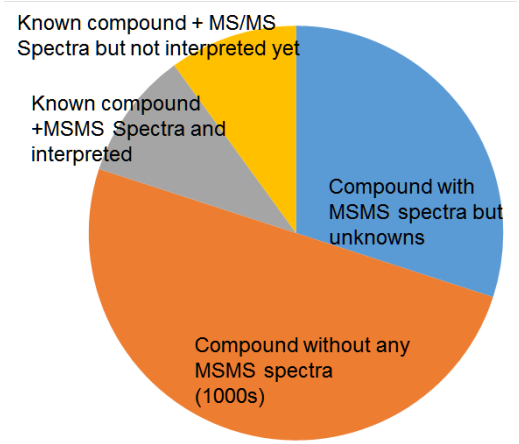
A basic data science environment



Discussion point : Raw LC/GC MS raw from NTA studies should be **indexed in enterprise databases** to support basic queries as well as advanced signal processing.

Lai, Zijuan, et al. "Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics." *Nature methods* 15.1 (2018): 53-56.

Annotation capacity building needs an integrated approach



Sample processing
compound enrichment
derivatization

Instruments
scan speed
precursor isolation width
sensitivity
spectra quality

Annotation workflows
scalable
automated
filtering criteria
multiple-evidences
context-dependent

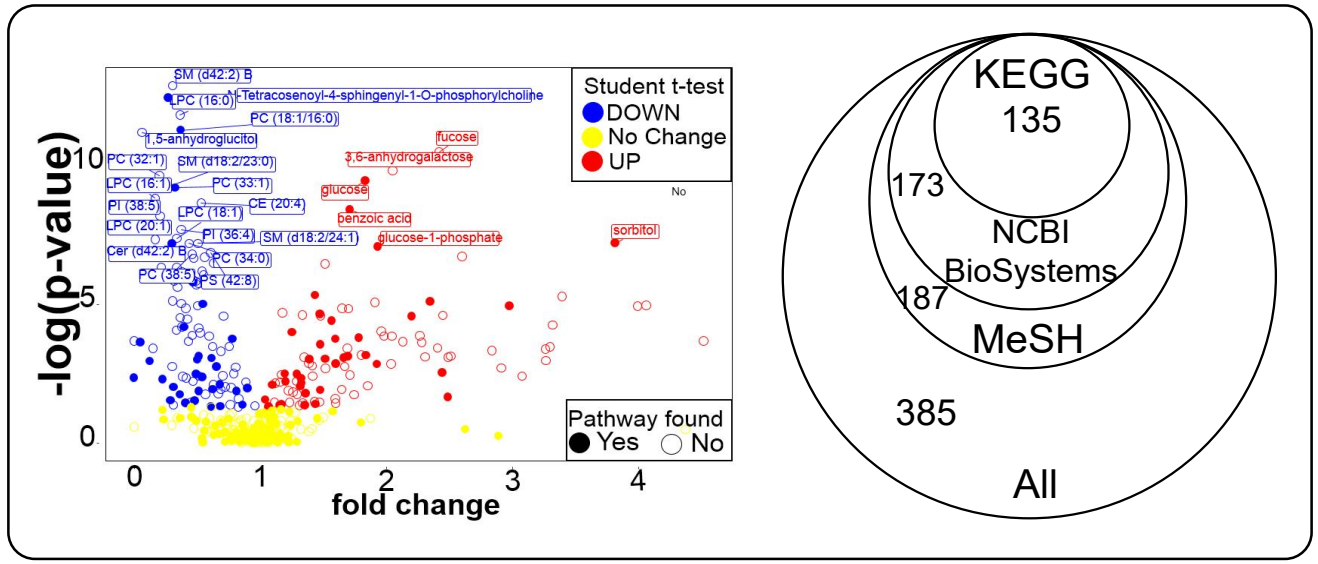
Mass spectral library
coverage and diversity
in-silico prediction
spectra quality

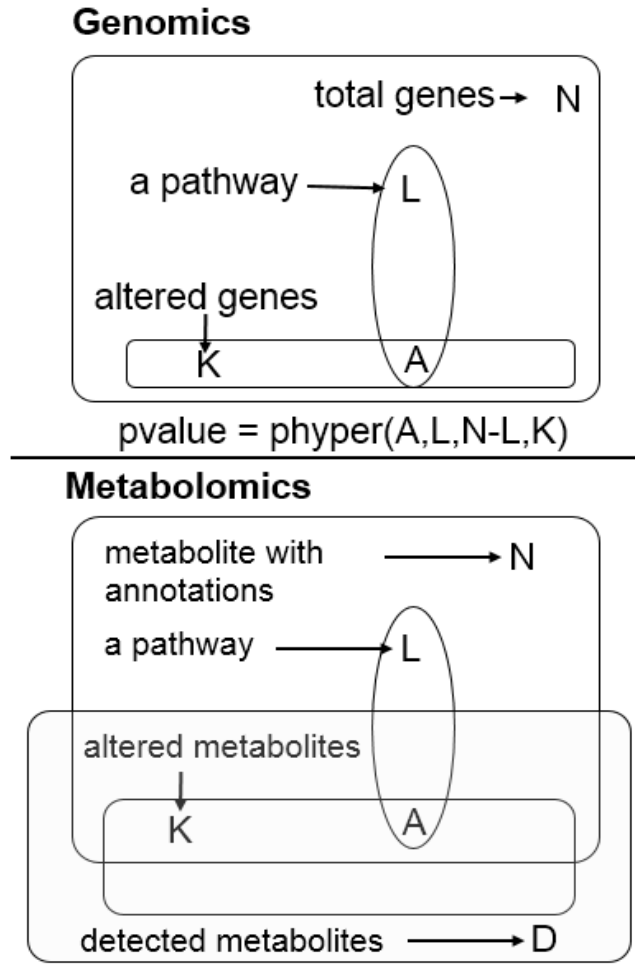
Continuous expansion

<u>Discussion point</u> : How to rank experimental and in-silico evidences for a peak annotation?

Barupal, Dinesh K., et al. "A comprehensive plasma metabolomics dataset for a cohort of mouse knockouts within the international mouse phenotyping consortium." *Metabolites* 9.5 (2019): 101.
Bonini, Paolo, et al. "Retip: retention time prediction for compound annotation in untargeted metabolomics." *Analytical Chemistry* (2020).
Lu, Wenyun, et al. "Improved annotation of untargeted metabolomics data through buffer modifications that shift adduct mass and intensity." (2020).

Poor coverage of NTA data in pathway DBs





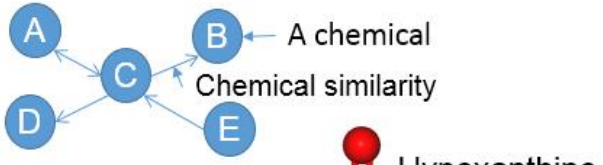
Discussion points :

1) A background database does not exist for NTA.
2) Assuming a statistical independence of chemicals is false.

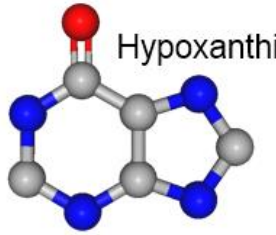
Barupal, Dinesh Kumar, and Oliver Fiehn. "Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets." *Scientific reports* 7.1 (2017): 1-11

Barupal, Dinesh K., et al. "MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity." *BMC bioinformatics* 13.1 (2012): 99.

Chemical similarity graph for NTA data



Tanimoto Chemical Similarity score
0.917



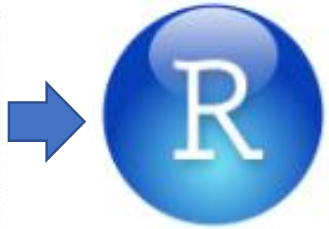
Tanimoto = AB / (A + B - AB)

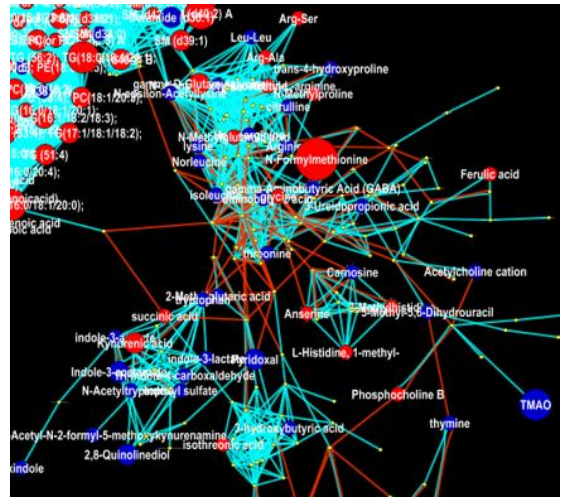
Substructure decomposition for calculations of chemical similarity

Cpd/Substr.	1	2	3	4	5	6	7	8	9	10
A	1	1	0	1	0	1	1	0	0	0
B	1	1	0	1	0	1	1	0	0	0
C	1	0	1	0	1	1	0	0	0	0
D	1	0	1	0	1	1	0	0	0	0
E	1	0	1	0	1	1	0	0	0	0
F	1	0	1	0	1	1	0	1	1	0
G	1	0	0	0	1	1	0	0	0	0
H	0	1	0	0	1	1	0	0	0	0
I	0	1	0	0	1	1	0	0	0	0
J	0	1	0	0	1	1	0	0	0	0

	cpd a	cpd b	cpd c	cpd d	cpd e	cpd f	cpd g	cpd h	cpd i	cpd j
cpd a	1	0.231	0.333	0	0.111	0.167	0.143	0.026	0.03	0.03
cpd b	0.231	1	0.444	0.04	0.13	0.053	0.1	0.037	0.111	0.082
cpd c	0.333	0.444	1	0	0.154	0.143	0.13	0.035	0.075	0.072
cpd d	0	0.04	0	1	0.111	0	0.04	0.205	0.056	0.072
cpd e	0.111	0.13	0.154	0.111	1	0.6	0.625	0.086	0.036	0.057
cpd f	0.167	0.053	0.143	0	0.6	1	0.818	0.025	0.009	0.029
cpd g	0.143	0.1	0.13	0.04	0.625	0.818	1	0.037	0.019	0.036
cpd h	0.026	0.037	0.035	0.205	0.086	0.025	0.037	1	0.089	0.156
cpd i	0.03	0.111	0.075	0.056	0.036	0.009	0.019	0.089	1	0.037
cpd j	0.027	0.102	0.079	0.051	0.042	0.017	0.026	0.084	0.917	1

1) Compute similarity among chemicals



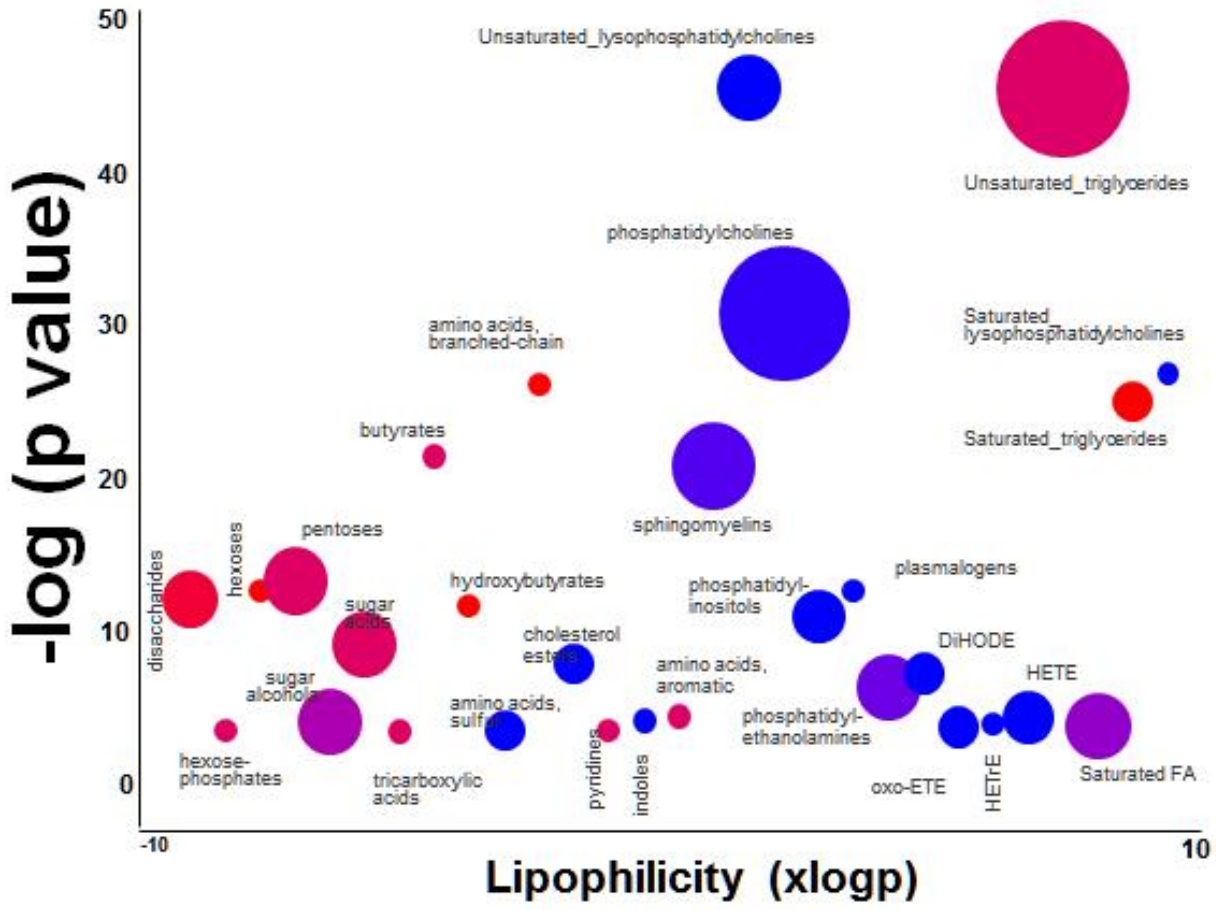


2) Network visualization

Discussion point : How to interpret large-scale network visualization for NTA data?

Barupal, Dinesh K., et al. "MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity." *BMC bioinformatics* 13.1 (2012): 99.

ChemRICH uses the MeSH ontology



- Node color indicate the proportion of node had a positive (red) or negative (blue) association with a phenotype.
- The Kolmogorov–Smirnov was used compute set level p-values (y-axis)

Perspectives | Brief Communication

Identifying Chemical Groups for Biomonitoring

http://dx.doi.org/10.1289/EHP537

Discussion points :

1) Prioritization of MeSH chemical ontology terms of biomonitoring
2) How to include unidentified metabolites into the set analysis ?

Barupal, Dinesh Kumar, and Oliver Fiehn. "Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets." *Scientific reports* 7.1 (2017): 1-11.

Well-known issues with the NTA data processing

1) A large number of signals (50-95%) remains **unknown**
2) **Slow** signal processing for a large batch of samples
3) **Errors** in peak grouping and deconvolution
4) Correction of retention time **drifts** for large sample sizes
5) Presence of **missing** values
6) **Low** frequency signals are often ignored
7) Presence of **artifacts** and background signals
8) **Issues** with data normalization
9) Challenging biological **interpretation**
10) Ethical issues in data sharing for **sensitive** analytes such as illicit drugs

Discussion point : How and when to address these issues in the NTA data processing?

Chemical to Literature Mapping

Chemical to literature mapping

Abstract

Am J Epidemiol. 2016 Feb 15;183(4):249-58. doi: 10.1093/aje/kwv242. Epub 2016 Jan 27.

Plasma Biomarkers of Inflammation, the Kynurenine Pathway, and Risks of All-Cause, Cancer, and Cardiovascular Disease Mortality: The Hordaland Health Study.

Zuo H, Ueland PM, Ulvik A, Eussen SJ, Vollset SE, Nygård O, Midttun Ø, Theofylaktopoulou D, Meyer K, Tell GS.

Abstract

We aimed to evaluate 10 biomarkers related to inflammation and the kynurenine pathway, including neopterin, kynurenine:tryptophan ratio, C-reactive protein, tryptophan, and 6 kynurenines, as potential predictors of all-cause and cause-specific mortality in a general population sample. The study cohort was participants involved in a community-based Norwegian study, the Hordaland Health Study (HUSK). We used Cox proportional hazards models to assess associations of the biomarkers with all-cause mortality and competing-risk models for cause-specific mortality. Of the 7,015 participants, 1,496 deaths were recorded after a median follow-up time of 14 years (1998-2012). Plasma levels of inflammatory markers (neopterin, kynurenine:tryptophan ratio, and C-reactive protein), anthranilic acid, and 3-hydroxykynurenine were positively associated with all-cause mortality, and tryptophan and xanthurenic acid were inversely associated. Multivariate-adjusted hazard ratios for the highest (versus lowest) quartiles of the biomarkers were 1.19-1.60 for positive associations and 0.73-0.87 for negative associations. All of the inflammatory markers and most kynurenines, except kynurenic acid and 3-hydroxyanthranilic acid, were associated with cardiovascular disease (CVD) mortality. In this general population, plasma biomarkers of inflammation and kynurenines were associated with risk of all-cause, cancer, and CVD mortality. Associations were stronger for CVD mortality than for mortality due to cancer or other causes.

Full-text

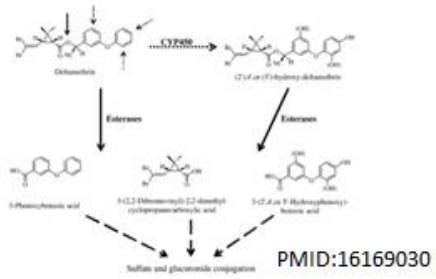
Table

PBDEs pg/ml ww	median	range	% detect
BDE-28/33	3.22	0.37–25.05	100
BDE-47	46.57	nd-463.94	97
BDE-99	9.19	nd-60.03	89
BDE-100	9.96	nd-93.94	97
BDE-153	59.64	20.31–180.91	100
BDE-209	18.39	nd-204.22	97

PMID:29396447

Table 1. Descriptive statistics for study participants

Figure



PMID:16169030

In -paragraph

"A halving in **serum folate** concentrations was moderately associated with increased risk of UCC (OR: 1.18; 95% CI: 0.98–1.43)" - PMC6899898

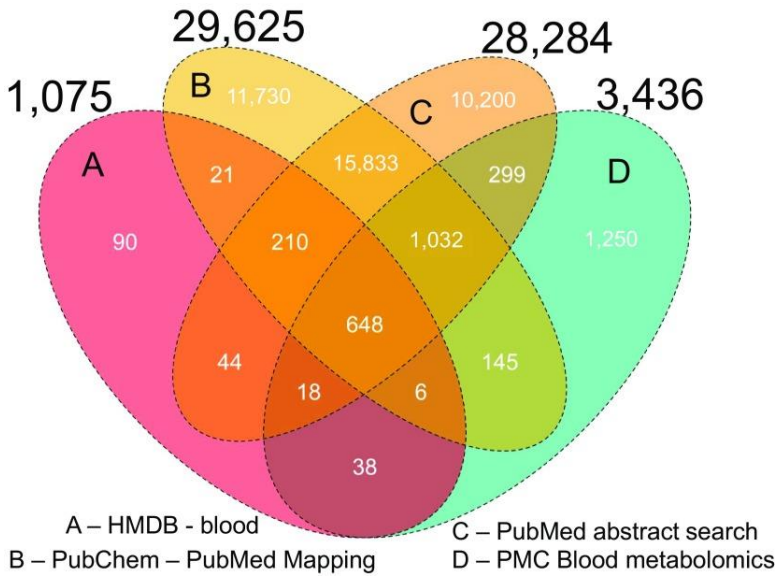
Supplementary data

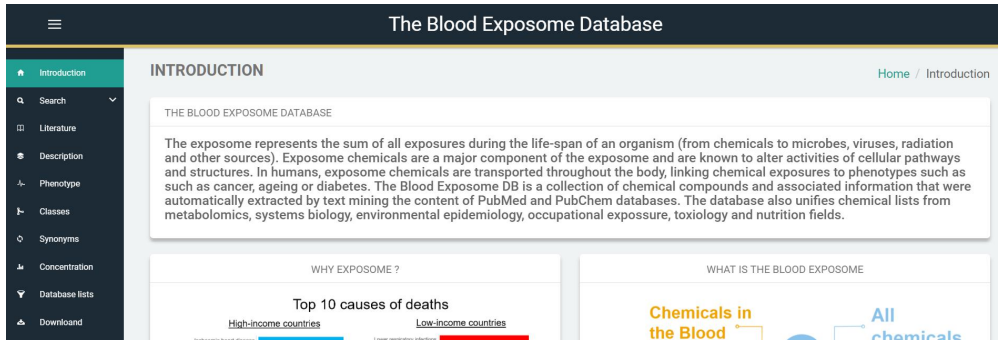
4-methylcatechol sulfate	Xenobiotics
4-methylguaiacol sulfate	Xenobiotics
4-vinylphenol sulfate	Xenobiotics
benzoate	Xenobiotics
catechol sulfate	Xenobiotics
guaiacol sulfate	Xenobiotics
hippurate	Xenobiotics
methyl-4-hydroxybenzoate sulfate	Xenobiotics
o-cresol sulfate	Xenobiotics
p-cresol sulfate	Xenobiotics
propyl 4-hydroxybenzoate sulfate	Xenobiotics
1,2,3-benzenetriol sulfate (2)	Xenobiotics
2,2'-Methylenebis(6-tert-butyl-p-cresol)	Xenobiotics
2-aminophenol sulfate	Xenobiotics
2-methoxyresorcinol sulfate	Xenobiotics
3-acetylphenol sulfate	Xenobiotics
3-hydroxypyridine sulfate	Xenobiotics
4-hydroxychlorothalonil	Xenobiotics
4-methylbenzenesulfonate	Xenobiotics
6-hydroxyindole sulfate	Xenobiotics
benzoylcarnitine*	Xenobiotics
bromine	Xenobiotics

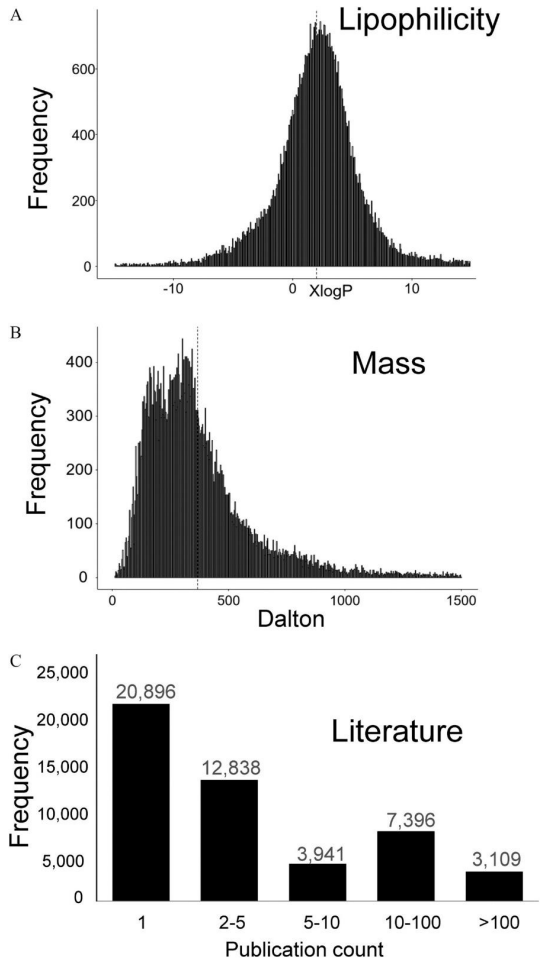
https://www.mdpi.com/2218-1989/10/1/34

Discussion point : How far we can go in developing a chemical to publication mapping resource?

The Blood Exposome Database







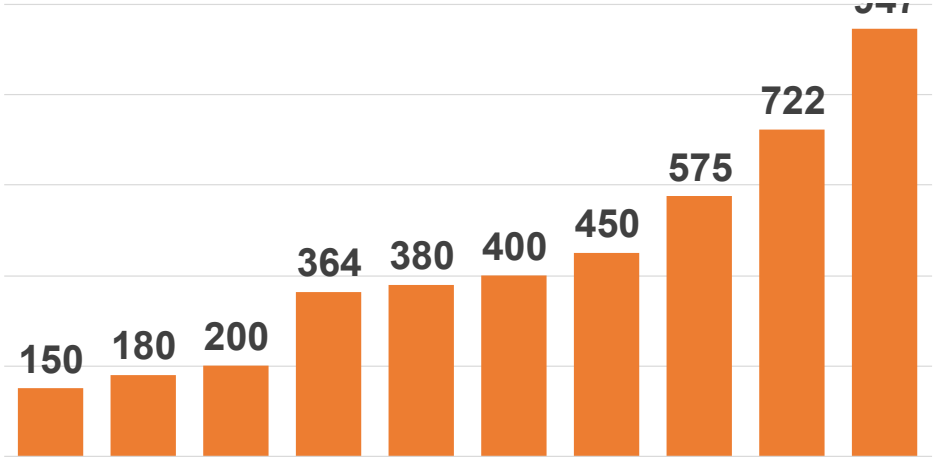
~ 42,000 unique 2D structures

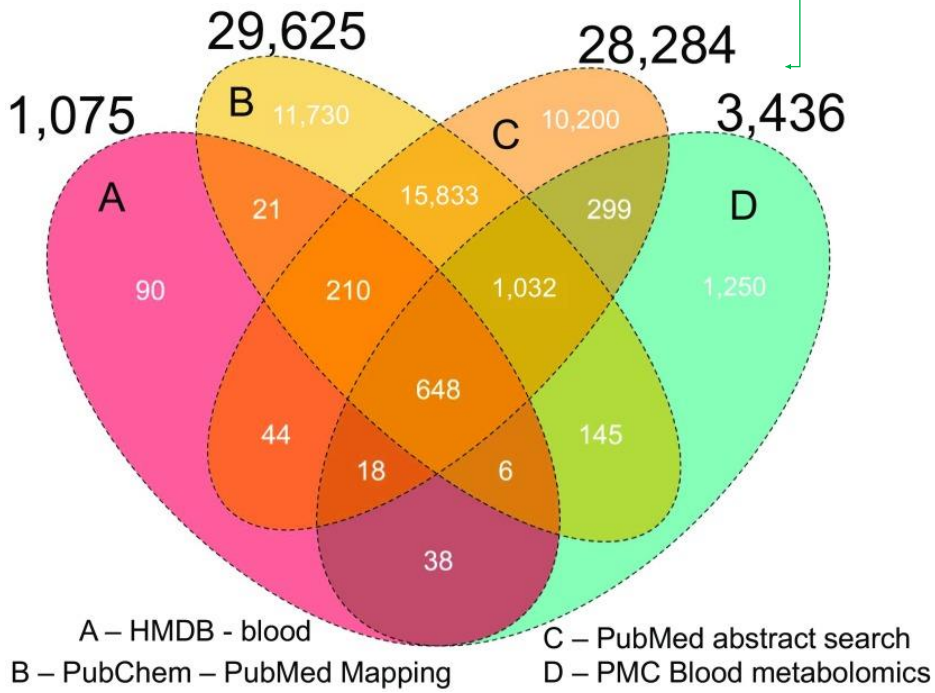
Discussion points : 1) How publication count for a chemical can improve peak annotation in NTA? 2) How to cover compounds that are not reported in an abstract text ?

Barupal, D. K., & Fiehn, O. (2019). Generating the Blood Exposome Database using a comprehensive text mining and database fusion approach. *Environmental health perspectives*, *127*(9), 097008.

Rise of the blood metabolome

Number of identified compounds in a blood metabolomics dataset by **metabolon**.

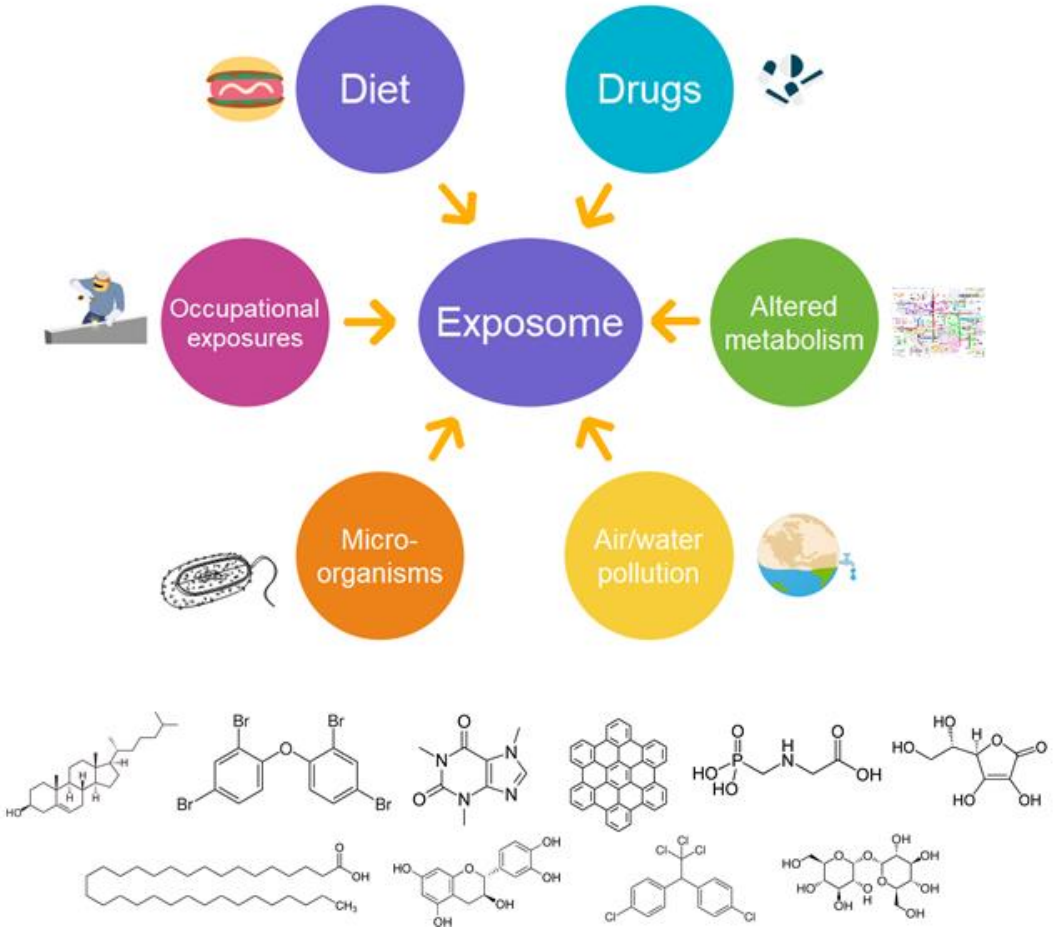


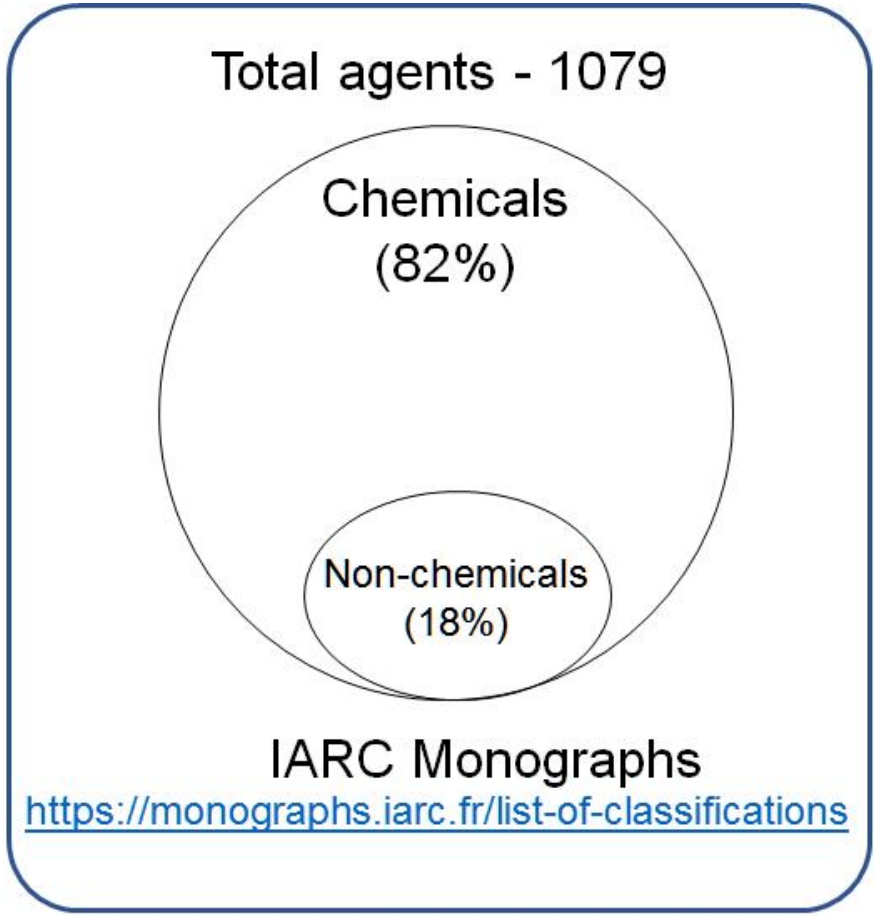


Discussion point : We should ensure that existing mass spectral libraries have EI/ESI spectra for these compounds.

Prioritizing chemicals for hazard assessments

Most exposures are chemicals

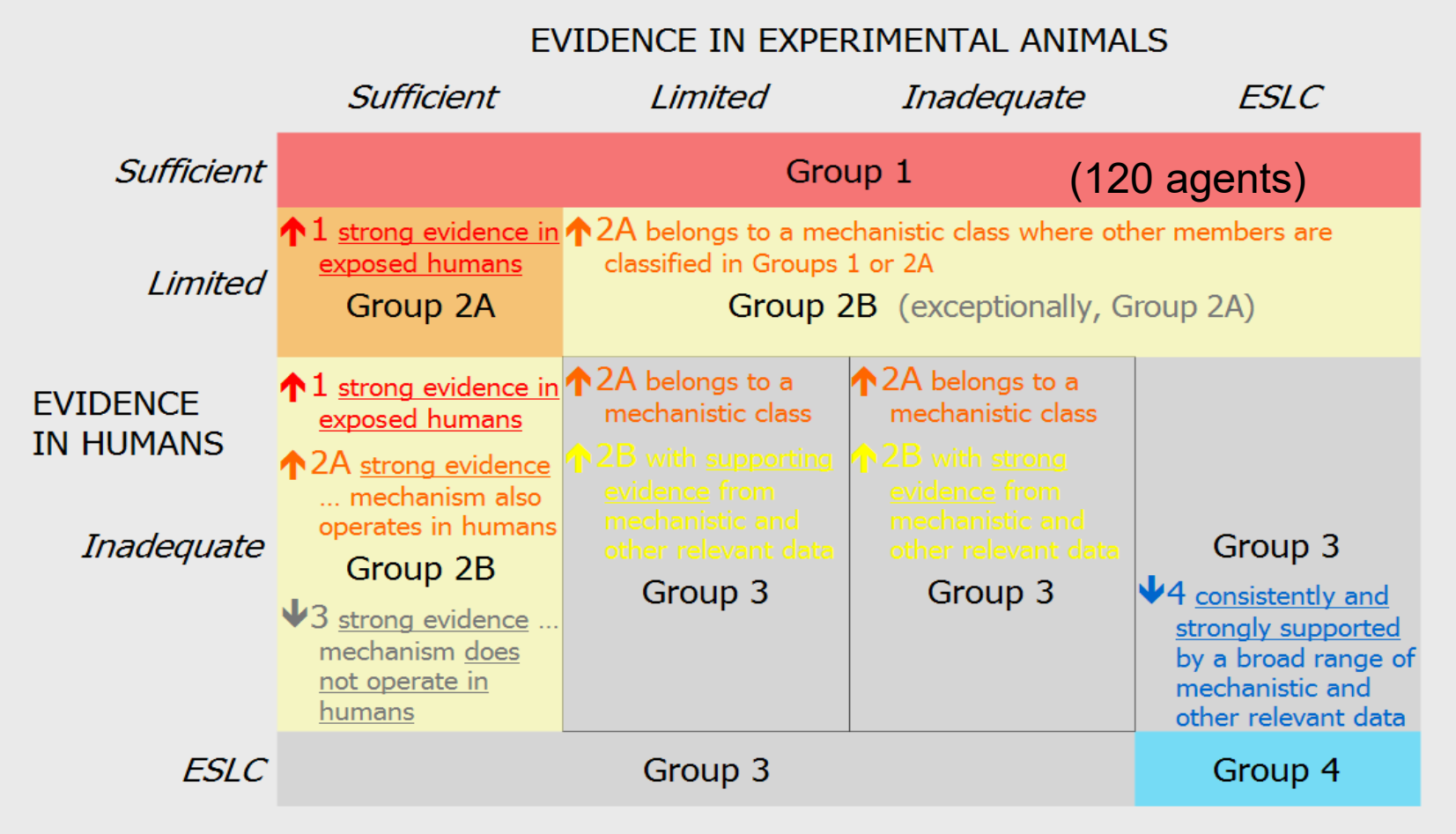




Mechanisms are in place to identify, monitor and regulate exposure to a specific chemical.

Vermeulen, Roel, et al. "The exposome and health: Where chemistry meets biology." *Science* 367.6476 (2020): 392-396.

Evidence based hazard assessments



ESLC : **Evidence** suggesting lack of **carcinogenicity**

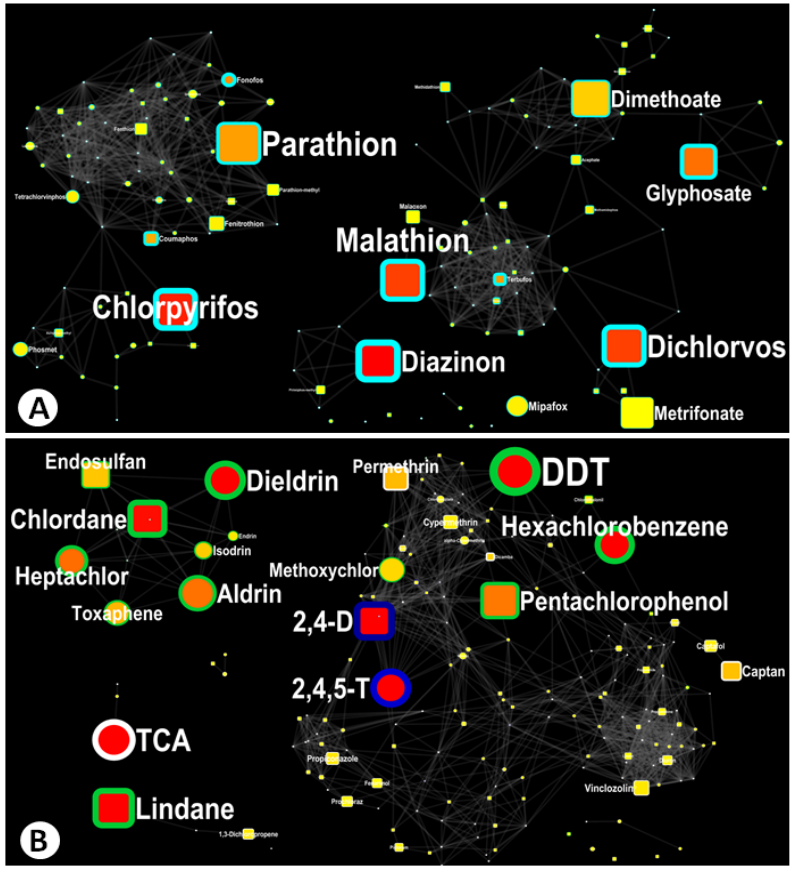
https://monographs.iarc.fr/wp-content/uploads/2018/06/Evaluations.pdf

IARC Monographs

International Agency for Research on Cancer

World Health Organization

Text mining for prioritizing chemicals



- Individual pesticides are represented as nodes on the chemical similarity maps. The node size is proportional to the number of publications overall on a pesticide and cancer: larger nodes represent more publications.
- The node border width represents the number of publications on epidemiology, cancer and the pesticide: a thicker border represents more papers. The node color, ranging from yellow to red, also represents the number of publications on epidemiology, cancer and the pesticide: red represents the highest count of publications.
- The node shape indicates whether results for a particular pesticide were available in the ToxRefDB database (circle = absent; square = present).
- The node border color represents the KEGG pesticide classification: green = Organochlorine, navy blue = Phenoxy, light blue = Organophosphorus, white= Others.

Environ Health Perspect. 2016 Dec;124(12):1823-1829. Epub 2016 May 10.

Prioritizing Chemicals for Risk Assessment Using Chemoinformatics: Examples from the IARC Monographs on Pesticides.

Guha N[1], Guyton KZ, Loomis D, Barupal DK.

IARC Monographs on the Evaluation of Carcinogenic Risks to Humans

Meeting 112: Some Organophosphate Insecticides and Herbicides: Diazinon, Glyphosate, Malathion, Parathion, and Tetrachlorvinphos (3-10 March 2015)

Call for Data (closing date 3 February 2015)
Call for Experts (closing date 30 July 2014)
Request for Observer Status (closing date 3 November 2014)
WHO Declaration of Interests for this volume

Meeting 113: Some Organochlorine Insecticides and Some Chlorphenoxy Herbicides (2-9 June 2015)

Call for Data (closing date 2 May 2015)
Call for Experts (closing date 10 October 2014)
Request for Observer Status (closing date 2 February 2015)
WHO Declaration of Interests for this volume

Discussion points : 1) Chemically similar agents can be evaluated together as they might have similar toxicological profile. 2) We can develop a similar approach for the California Biomonitoring program chemical list ?

IARC, Monographs Priorities Group. "Advisory Group recommendations on priorities for the IARC Monographs." *The Lancet. Oncology* 20.6 (2019): 763.

Conclusions

- Non-targeted analysis has a great potential for detecting high-priority chemicals for exposome research in biospecimens.
- However, a proper combination of analytical chemistry and data science needs to be planned ahead.
- Indexing raw data into enterprise databases and avoiding a signal prevalence threshold are needed for exposomics.
- Computational text mining can improve the prioritization process by linking chemicals to publications.
- Interpretational bias remains a major challenges in mining NTA.

Acknowledgment

Thanks to current and former collaborators at :

Icahn School of Medicine at Mount Sinai

Institute for Exposomic Research

International Agency for Research on Cancer

World Health Organization

UCDAVIS

UNIVERSITY OF CALIFORNIA

Special thanks to NIH for funding these initiatives

NIH Common Fund Metabolomics Program

HHEAR Human Health Exposure Analysis Resource

Thanks