The exposome and health: where chemistry meets biology

Authors:
Roel Vermeulen[1,2,*], Emma L. Schymanski[3], Albert-Laszlo Barabási[4], Gary W. Miller[5*]

Abstract 124, Body text 3123, Text Boxes 330, References count 40, Legends 395

*Corresponding authors: Roel Vermeulen and Gary W. Miller

Affiliations:

[1] Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, Utrecht University, Utrecht, the Netherlands

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands

[3] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 avenue du Swing, 4367 Belvaux, Luxembourg

[4] Network Science Institute, Northeastern University, Boston, MA, USA; Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA and Department of Network and Data Science, Central European University, Budapest, Hungary

[5] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA

Abstract

Despite extensive evidence that exposure to specific chemicals can lead to disease, current research approaches and regulatory policies fail to address the chemical complexity of our world. To safeguard current and future generations from the increasing number of chemicals polluting our environment, a systematic and agnostic approach is needed. The exposome concept strives to capture the diversity and range of exposures, including synthetic chemicals, dietary constituents, psychosocial stressors, and physical factors, as well as their corresponding biological responses. Technological advances, such as high-resolution mass spectrometry (HRMS) and network science, allow first steps towards comprehensive assessment of the exposome. Given the increased recognition of the dominant role non-genetic factors play in disease, an exposome effort at a scale comparable to the human genome is warranted.

One Sentence Summary:

The exposome concept, together with recent technological advances, has the potential to identify environmental contributors to health and disease in a manner complementary to the genome.

The Exposome

A basic tenet of biology is that the phenotype results from a combination of genes and environment. The field of genomics has provided an extraordinary level of genetic knowledge, chiefly capacitated by large-scale agnostic genome-wide association studies (GWAS). A similar level of analysis, however, is still lacking for our environment. The exposome concept was conceived in 2005 as a way to represent the environmental, i.e. non-genetic, drivers of health and disease (*1*). For these external forces to have an effect on health, they must alter our biology, suggesting that a detailed analysis of accessible biological samples at different molecular levels, coupled with information on environmental drivers, can provide snapshots of both the internal (biological perturbations) and external contributors to the exposome. As Rappaport and Smith described in 2010, "*toxic effects are mediated through chemicals that alter critical molecules, cells, and physiological processes inside the body…under this view, exposures are not restricted to chemicals (toxicants) entering the body from air, water, or food, for example, but also include chemicals produced by inflammation, oxidative stress, lipid peroxidation, infections, gut flora, and other natural processes" (2).* The exceptional variety and dynamic nature of non-genetic factors (Figure 1) confronts us with an array of sampling and analytical challenges. Fifteen years after the exposome concept was introduced, this review looks into progress in assessing the chemical component of the exposome and its implications on human health.

Insert Figure 1.

Environment >> Genes

Mapping the human genome revolutionized our ability to explore the genetic origins of disease, but also revealed the limited predictive power of individual genetic variation for many common diseases. For example, genetic factors contribute to less than half of the risk of heart disease, the leading source of mortality in the US and many other parts of the world (*3*). The health impact of environmental risk factors was highlighted by the Global Burden of Disease (GBD) project, which estimated the disease burden of 84 metabolic, environmental, occupational, and behavioral risk factors in 195 countries and territories, and found that these modifiable risks contribute to approximately 60% of deaths worldwide (*4*). Using established causal exposure-disease associations, nine million deaths a year (16% of all deaths worldwide) were attributed to air, water, and soil pollution alone (*5*). However, the true impact of the environment is likely to be grossly underestimated, as many of the known chemicals of concern were not considered and less than half of the non-genetic risk burden was explained, indicating the existence of missing exposome factors (*4*).These missing factors can be considered similar to the missing heritability challenge observed in genetic studies. Even with this incomplete inventory, the economic costs of chemical pollution are considerable, with healthcare and disability-related productivity loss estimated to amount to US $4.6 trillion per year, representing 6.2% of global economic output (*5*). Reducing or preventing chemical pollution is a multifaceted problem that involves medical, legal, and regulatory input (see Text Box 1).

Measuring chemicals *en masse*

Several research efforts have pioneered different approaches for the systematic mapping of the exposome, taking advantage of developments in mass spectrometry, sensors, wearables, study design, biostatistics, and bioinformatics (*6*); advances that now position us to pursue Dr. Wild's original vision of the exposome (*1*). A prime example is how high-resolution mass spectrometry (HRMS) has transformed our ability to measure multitudinous chemical species in a wide range of media, expanding our analytical window beyond targeted analysis of well-known metabolites and priority pollutants (*7*). HRMS provides the means to simultaneously

measure a vast number of exogenous and endogenous compounds, offering a description of the system and its changes in response to exposure to environmental factors (*6, 8*). As Figure 2 (top panel) indicates, HRMS is capable of measuring thousands to tens of thousands of chemical features in a single analytical run, although the majority of these features remain unannotated. While the systems biology approaches in metabolomics originally focused on detecting endogenous metabolites, HRMS methods also detect exogenously-derived small molecules, such as pharmaceuticals, pesticides, plasticizers, flame retardants, preservatives, and microbial metabolites, unavoidably intertwining biology and environmental science (*9*). Historically, these exogenous compounds were often viewed as noise and artifact, while in reality they carry direct evidence of the complex environments to which living organisms are exposed.

Insert Figure 2.

Data resources relevant for HRMS-based exposomics (i.e. the study of the exposome) range from specialized lists (e.g. (*10*)) to medium-sized databases containing tens to hundreds of thousands of chemicals, through to huge resources such as PubChem (*11*) with 96 million entries (see Figure 2). Of the >140,000 chemicals produced and used heavily since the 1950s, only about 5000 are estimated to be widely enough dispersed in the environment to pose a global threat to the human population, although many thousands more would be expected to impact individuals, local communities or specific occupational settings (*5*). Specialized lists compiled *e.g.,* by the US Environmental Protection Agency (EPA) (*12*) and environmental communities such as the NORMAN Network (*13*) often contain additional information (*e.g.,* exposure data, product information) to help annotate chemicals of interest in the study context. Medium-sized databases, such as Human Metabolome Database (HMDB) (*14*), are commonly used in approaches involving metabolic network analysis, offering typically one to a few possible chemicals per feature exact mass detected by HRMS. Databases that contain spectral information (*i.e.,* structural "fingerprints") can be used to increase the confidence of exact mass matching where experimental fragmentation information is available (*15, 16*). Comprehensive chemical resources such as PubChem are so large that they often offer several thousand possible chemical candidates per exact mass. Despite the exceptional size of the chemical space, the knowledge and computational tools required to interrogate this data are increasingly available (*15, 17*). For instance, incorporation of literature and patent information with *in silico* methods greatly improved annotation rates (from <22% to >70%) for >1,200 chemicals in HRMS experiments using PubChem (*17*).

Chemicals are not static entities – they react in our bodies and the environment to form metabolites or transformation products. Computational tools exist to predict such metabolic and environmental transformations (*15, 18*), but often produce many false positive and false negative candidates. Merely predicting first order reactions of PubChem chemicals would result in billions of possibilities (Figure 2, second from bottom). As a result, few studies have been able to successfully capitalize on this information in high-throughput identification efforts so far. The dispersed nature of the essential chemical, metabolite, and spectral information across a wide range of resources with various formats and forms of accessibility (fully open, academic use only, commercial, etc.) is a major impediment to progress in the field.

Integrating chemical knowledge

The interconnected nature of the available chemical information indicates the need for an interdisciplinary and integrative approach to further define the exposome and the associated data science challenges. Literature mining of PubMed and mapping to discrete chemicals can

be used to compile and synthesize the chemical information in scientific literature(*10, 19*). The expansion and automation of literature mining for more accurate chemical candidate retrieval during high throughput identification, *e.g*. with MetFrag (*17*) or other *in silico* approaches (*15*), will be crucial for faster, more efficient annotation of the complex and highly varying datasets that characterize studies of chemical exposures and health.

Many of the chemicals of interest in exposomics come from the same or related sources (*e.g.*, industrial processes, consumer goods, diet), meaning that such exposures exhibit a population structure (*i.e.*, complex correlations and dynamic patterns) akin to observed correlations in complex biological systems. Thus, reduction of dimensional complexity will be possible by grouping correlated exposures. Indeed, several reports have shown correlation patterns between different chemicals and chemical families within populations (*20, 21*). These relationships between chemicals can be presented as networks of chemicals (i.e exposure enrichment pathways) that unveil communities of exposures (*20, 21*), which in turn can be used to explore the impact that they have on the biological system (see network science below).

Much of our current knowledge about the health effects of chemicals comes from epidemiological and toxicological studies, in which a limited number of pollutants are analyzed in relation to a specific phenotype, representing a hypothesis-driven path towards understanding exposure-disease relationships. Yet, our exposures are not a simple sum of a handful of chemicals. To overcome the limitations of traditional epidemiological studies, environment-wide association studies (EWAS) were proposed to identify new environmental factors in disease and disease-related phenotypes at scale. EWAS was inspired by the analytical procedures developed in GWAS (*22*) in which a panel of "exposures" (genotype variants) is studied in relation to a phenotype of interest. For example, using the National Health and Nutrition Examination Survey dataset, an EWAS study explored the associations of 543 environmental attributes with type 2 diabetes, identifying five statistically significant associations (including persistent organic pollutants and pesticides) validated across independent cohorts (*22*). By focusing on a predetermined list of chemicals, these initial EWAS studies could suffer from the same limitations of candidate gene searches. Further, current EWAS approaches do not test for interactions and/or combinations of factors (mixtures). Recent efforts have been undertaken to develop statistical methods to screen for interactions and to test the effect of mixtures or to apply frameworks such as aggregated exposure and adverse outcome pathways to study combinatorial effects (*9*).

As exposomics approaches moves forward to elucidate the impact of our chemical constellation on health in a systematic manner, they must integrate increasingly rich and high-dimensional data that capture the continuum of exposure to health (Figure 3), while adequately defining appropriate frameworks for establishing controls, background and negative responses enabling causal inference. To aid such inference, more insights into the boundaries of 'normal' responses is required and would necessitate definitions of a reference exposome.

Network science to tackle exposome complexity

The challenge in understanding the role of the exposome on our health lies not only in the large number of chemical exposures in our daily lives, but also in the complex ways they interact with cells, ultimately affecting our health. A reductionist framework aims to isolate the role of a single variable, inadequately capturing the complexity of the exposome. Network science (*23*), whose applications are well developed in network medicine and system biology (*24*), offers a platform to achieve this, helping us systematically explore the mechanistic role of the chemical compounds in our exposome. Each chemical exerts its effect on health through

interactions with various cellular components, supplying or perturbing the networks within our cells. To capture the diversity in these interactions, we must first catalogue the sum of all physical interactions as a multilayer network (*25*) consisting of several distinct biological layers (Figure 3). While each of these networks rely on different biological mechanisms, they are not independent: protein production is governed by the regulatory network, and the catalysis of the metabolic reactions is governed by enzymes and protein complexes, members of the protein interaction network (*26*).

To fully understand the role of the exposome, we must similarly develop a multilayer network-based framework, capable of unveiling the role of chemicals, their combinations, and biological perturbations on our health. However, there are several data and methodological challenges. The first is the paucity of systematic data on the various dimensions of exposure, from bioavailability to binding information of the hundreds of thousands of exposome molecules to human proteins. The US National Toxicology Program, US EPA, and the European Molecular Biology Laboratory are developing platforms to generate, collate, and organize data on chemical-biological interactions, but there is a need for high throughput approaches that offer greater coverage (*12, 27, 28*). Second, the current statistical toolset, including EWAS, assumes that we are faced with a collection of random variables that are independent and identically distributed and measured with equal precision. In a network environment, these assumptions are inherently false, mechanistic interactions couple the probability distribution of most network-based variables. Furthermore, most of the chemicals we are exposed to represent communities of exposures, hence the effect of a chemical can be rarely observed in isolation. Therefore, identifying meaningful associations from high-dimensional exposomic data poses significant statistical and computational challenges that need to be addressed in parallel with the experimental developments. Third, beyond cataloging interactions, we must also understand the dynamics (*29*) of the biochemical pathways through which the different exposome factors affect our health. Indeed, the human interactome, representing the sum of all physical interactions within a cell (Figure 3), often depicted as a static graph, is in reality a temporal network (*30*), whose nodes and links disappear and re-emerge based on factors ranging from the cell cycle to variability in environmental exposures across the life course. Modeling the fully temporal nature of these networks remains a challenge, as the kinetic constants underlying metabolic processes are not known and we currently lack systematic tools to identify them (*31*).

Insert Figure 3.

Informative exposome study designs

The systematic and unbiased assessment of the exposome (*i.e.,* not focusing on a selected set of readily measured or priority chemicals), requires access to biological samples that reflect exposures, biological effects, and preferably the health phenotype of interest. This is challenging, as it will be rare that the variability of exposures (E) aligns perfectly with the kinetics of the biological effect (B) and the etiological time-window of the health phenotype, including developmental and transgenerational effects (P). Optimizing each step (E-B and B-P) in separate studies, however, has the disadvantage that overlapping patterns in each step restrict us from unveiling the true association between exposure and the health phenotype (E-P). The meet-in-the-middle (MITM) design, in part, addresses this challenge (*32*). In MITM, exposures can be assessed in individuals using HRMS or upstream estimates of external factors (Figure 1) and are compared to downstream biological changes in persons who develop a specific health phenotype and those that do not.

The MITM approach using HRMS data has successfully identified single and combinatorial effects of chemicals. (*e.g.,*(*33-36*)). For example, the HELIX study explored the early-life exposome of population-based birth cohorts and identified several environmental exposures, the majority of which were chemicals, associated with lung function in children (*35*). The EXPOsOMICs study showed how air pollution alters biological pathways, in particular linoleate metabolism, which predicted for the occurrence of adult onset asthma and cardiovascular disease (*36*).

Scaling-up

By pooling studies, sample sizes for GWAS have increased from a few thousand to tens and hundreds of thousands of individuals over the last decade (*37*). However, enrollment in studies of non-genetic environmental exposures remains relatively low. The large-scale genomic consortia efforts allowed GWAS to detect many common genetic traits related to health phenotypes and, although the combined effects of the identified traits are still modest, they provide insights into the underlying biological pathways of disease. It is estimated that sample sizes of 500,000 to 2,000,000 are needed to explain a substantial portion of the projected genomic heritability of common chronic diseases (*38*). For the multitude of factors within the exposome, most of which likely exert small effects, similar or even greater sample sizes would be required for future environment/exposome-wide association studies (EWAS) (*22*). Scaling exposome research to these numbers will require a joint effort across multiple cohort consortia and research programs. Recently funded programs to work towards a Human Exposome Project are a first step towards reaching tens of thousands of people with detailed environmental and biological analysis of exposures. While these numbers are likely large enough to identify the most prevalent and strongest chemical risk factors, a progressive increment in sample size will be needed for a systematic understanding of the impact of combinatorial exposome factors on specific and rare phenotypes. The systematic identification of the impact of non-genetic factors and chemical exposures would enable the establishment of Exposome Risk Scores (ERS) akin to Polygenetic Risk Scores (PRS) (see Text Box 2).

Next steps for the exposome

The rate, volume, and variety of chemicals being introduced into our environment continues to expand. The importance of these chemical exposures on human health is exemplified by the still large, unexplained proportion of disease that is known to be related to yet elusive factors(*3*). Indeed, the non-genetic component exceeds the known and missing heritability. We therefore need innovative ways to study these factors and translate our findings into policy. Currently, many of the regulatory agencies are expanding their computational and high-throughput approaches to account for the ever-increasing number of chemicals, but there are still major challenges regarding prioritization and new approaches are urgently needed (see Text Box 1). Exposome research can contribute to this process through large scale nontargeted screenings of chemicals and by systematically unveiling their associations to health outcomes. Open science efforts such as the Global Natural Product Social Molecular Networking (GNPS), which allows users to archive huge amounts of raw data, offering computational mass spectrometry workflows coupled with open mass spectral libraries and continuous updates of new identifications in return, are beginning to be leveraged for large-scale studies (*20*). However, as discussed above, we must address several challenges to exploit the full potential of exposome research as it relates to improving our understanding of exposure to complex chemical mixtures: 1) we must improve our technology to screen for exogenous chemicals and their biological consequences at higher throughput rates and lower costs; 2) continue to develop the current chemical and spectral data resources needed to identify these signals in samples; 3)

increase the scale and scope of studies to a level that provides the necessary statistical power to precisely characterize the effects of the chemicals and their combinations; 4) further develop and support cheminformatic and bioinformatic tools, including network theory and network medicine, to enable elucidation of the constellation of the chemical environment and its biological consequences; and 5) ensure adequate protection for the generation of false positive results by insisting on replication in independent studies and the use of methods to establish causation such as Mendelian Randomization, within-sibling comparisons, and exposure negative, and outcome negative controls.

Conclusion

A concerted and systematic effort to profile the non-genetic factors associated with disease and health outcomes is urgently needed as we lack important insights to curtail the ever-growing burden of chronic disease on society. Emerging exposome research frameworks are poised to enable the systematic analysis of non-genetic factors involved in disease. Technology has enabled the first generation of studies evolving to the comprehensive study of combinatorial chemical exposures. Future efforts must ensure that the analytical approaches and study designs are rigorous and validated. A coordinated and international effort to characterize the exposome, akin to the Human Genome Project, would provide rigorous data to allow exposome-based EWAS to be conducted at the scale of GWAS. By taking advantage of the nontargeted nature of HRMS, EWAS provides a true complement to GWAS. Consolidating knowledge garnered from GWAS and EWAS would allow us to map the gene and environment interface, which is where nature meets nurture and chemistry meets biology.

Text Box 1. The exposome and regulation.

Many of the influential regulatory bodies currently residing in Europe and North America have been expanding their computational and high-throughput approaches to address the increasing number of chemicals, but there are still major challenges regarding prioritization. Networks such as NORMAN (*13*) that bridge scientists, regulators, and practitioners, are becoming increasingly valuable avenues of knowledge exchange. Large-scale exposome studies provide a systematic approach to prioritization, allowing regulatory bodies to focus on those chemicals that have the most significant adverse effects on health. If systematic analysis would reveal major adverse effects on human health from exposure to currently approved or potential replacement chemicals, then those compounds should be removed from the marketplace. While thousands of compounds are classified as "generally recognized as safe", they were never subjected to the scientifically rigorous testing systems currently in place. A data-driven exposome approach ignores historical decision-making and can help evaluate the effects of classes of chemicals on specific biological pathways known to be perturbed and help design new compounds with minimal impact on human health and the environment.

Text Box 2. Towards an Exposome Risk Score (ERS)

There has been significant progress in the identification of genetic risk factors of chronic diseases. Analysis of high-risk mutations and estimation of polygenic risk score (PRS) for these diseases are now becoming routine and can be included when developing individual-based (i.e. precision) prevention and treatment strategies. Similarly, the establishment of Exposome Risk Scores (ERS) would help summarize relevant non-genetic risk factors, enabling identification of hotspots of concern where multiple environmental factors come together, and would aid in the prioritization of risk factors based on their population and individual impact. For example, an ERS could provide data on exposure to chemical toxicants based on the biological processes or organ systems that are most vulnerable and couple them with indices of associated biological injury or response. Such ERS scores, in contrast to PRS, would be time-varying and dynamic through age-related exposures and susceptibilities.

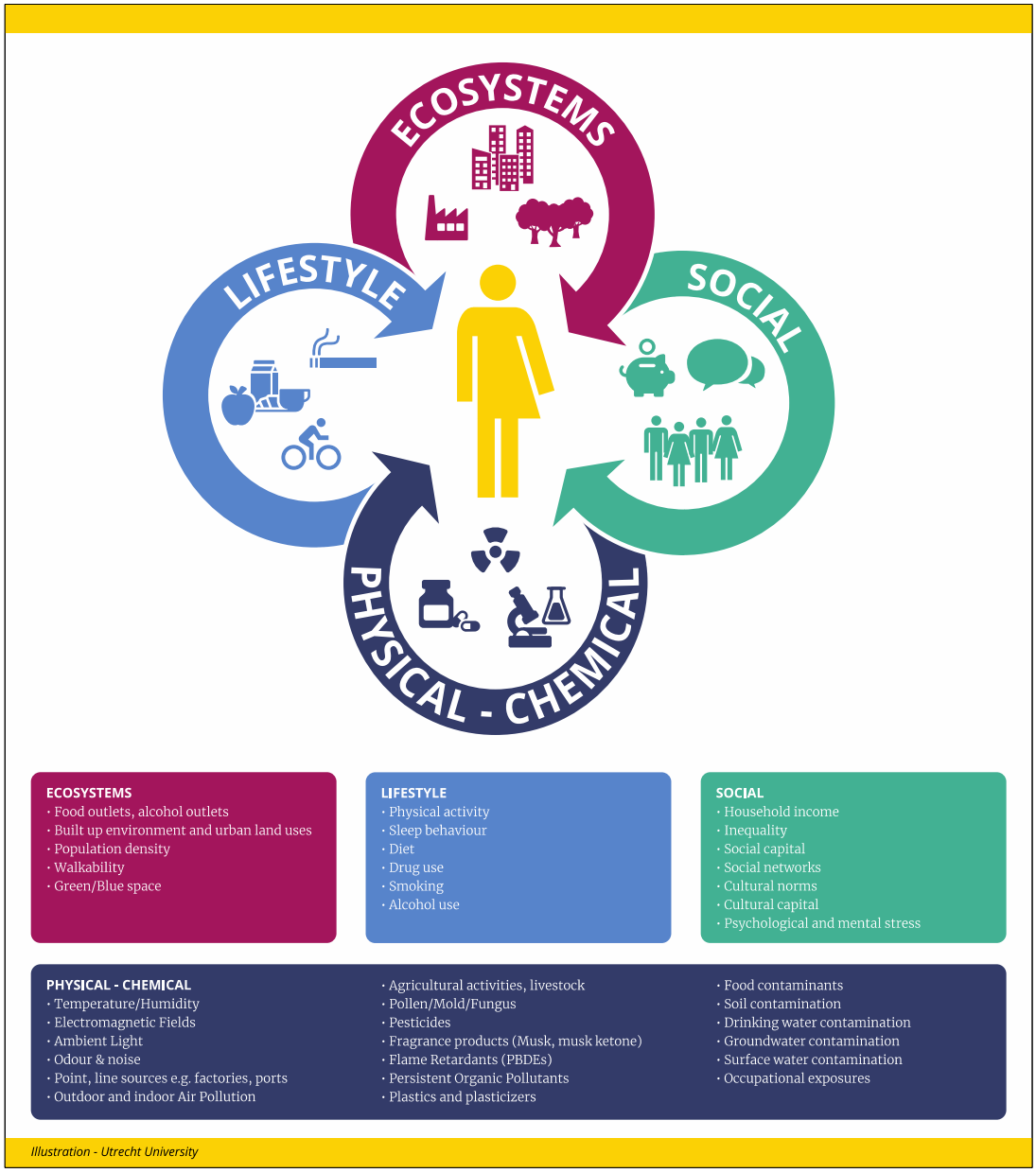

Figure 1. **The exposome concept**. The exposome is an integrated function of exposure on our body including what we eat and do, our experiences, and where we live and work. The chemical exposome is an important and integral part of the exposome concept. Examples of external stressors are adapted from (*39*). These stressors are reflected in internal biological perturbations (Figure 3); thus, exposures are not restricted to chemicals (toxicants) entering the body, but also include chemicals produced by biological and other natural processes.

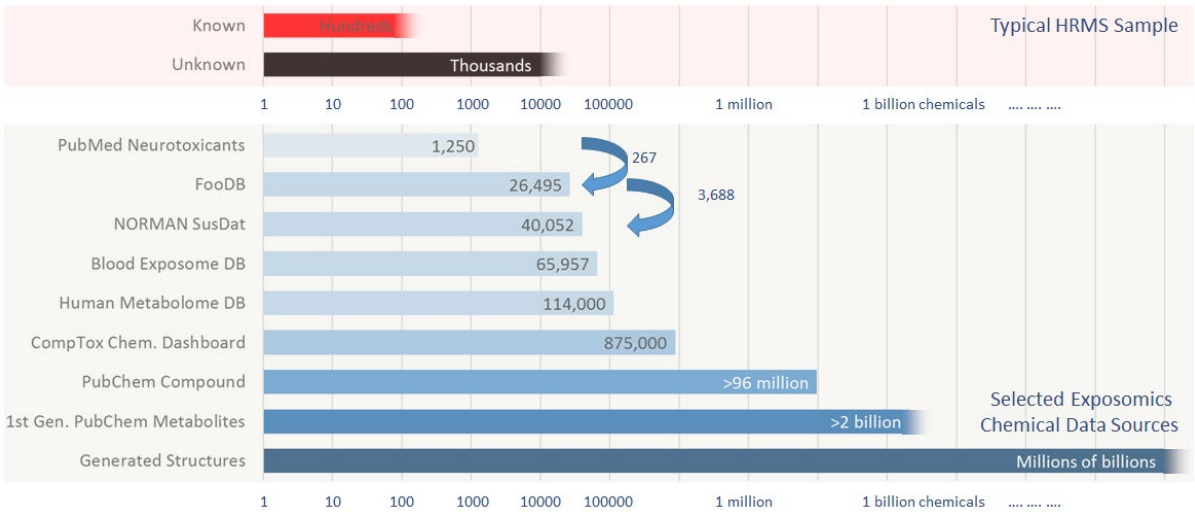

Figure 2. **Chemical complexity of HRMS and the exposome.** Top: Known versus unknown features in a typical HRMS measurement (data from (*7*)). Bottom: Selected data sources relevant to the chemical exposome (*10-14, 19*). Arrows show the overlap of potential neurotoxicants in FooDB (foodb.ca), and FooDB components in NORMAN SusDat (prioritized chemicals of environmental interest).

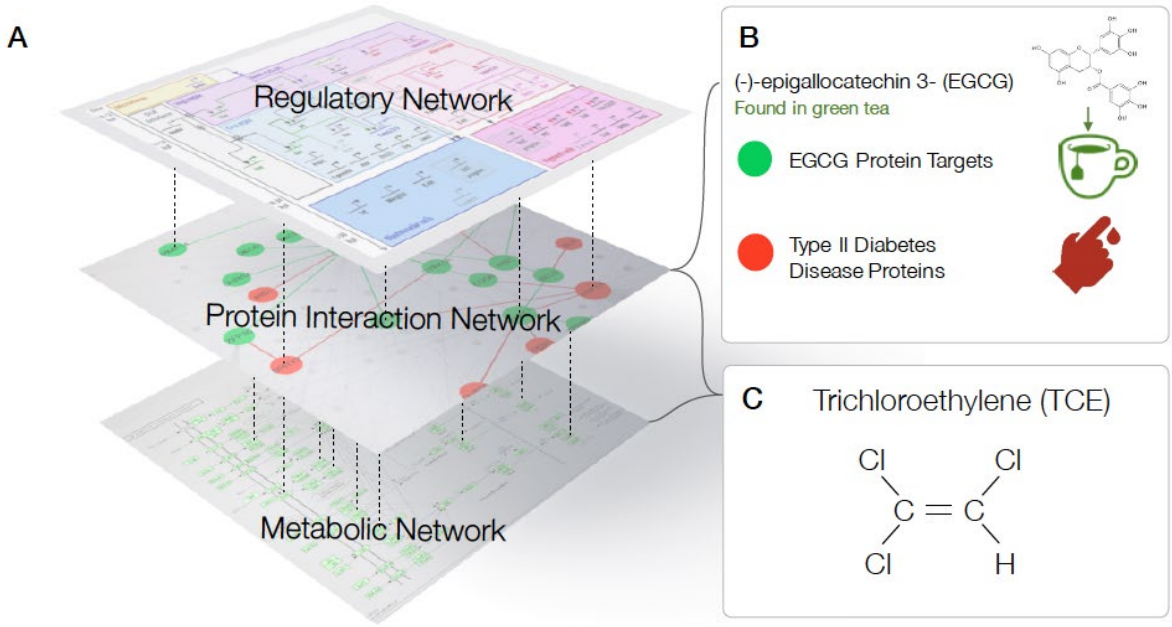

Figure 3. **The impact of the exposome on subcellular networks**. **A.** Network medicine views the cell as a multilayer network, with three principal, interdependent layers: (i) regulatory network, capturing all interactions affecting RNA and protein expression; (ii) protein interaction network, that captures all binding interactions responsible for the formation of protein complexes and signaling; (iii) metabolic network, representing all metabolic reactions, including those derived from the microbiome, a network of interacting bacteria linked through the exchange of metabolites. Exposome-related factors can affect each layer of this multilayer network. **B.** The polyphenol EGCG, a biochemical compound in green tea, with potential therapeutic effects on type 2 diabetes mellitus (T2D), binds to at least 52 proteins (*40*). Network-based metrics reveal a proximity between these targets and 83 proteins associated with T2D, suggesting multiple mechanistic pathways to potentially account for the relationship between green tea consumption and reduced risk of T2D. **C.** Trichloroethylene (TCE) is a volatile organic compound that was widely used in industrial settings and is now a widespread environmental contaminant present in drinking water, indoor environments, ambient air, groundwater, and soil. Multiple lines of evidence support a link between TCE exposure and kidney cancer and probably non-Hodgkin lymphoma (*33*). TCE perturbs at least two different layers of the cellular network: it covalently binds to proteins from the protein interaction network, altering their function, and affects cellular metabolic network, eventually leading to ATP depletion. Network-based tools could be used to explore the mechanistic role of many other exposome chemicals on our health, and to build experimentally testable hypotheses.

References

1. C. P. Wild, Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **14**, 1847-1850 (2005).
2. S. M. Rappaport, M. T. Smith, Epidemiology. environment and disease risks. *Science* **330**, 460-461 (2010).
3. J. P. Ioannidis, E. Y. Loy, R. Poulton, K. S. Chia, Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci Transl Med* **1**, 7ps8 (2009).
4. Global Burden of Disease Study Group, Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1345-1422 (2017).
5. P. J. Landrigan *et al.*, The Lancet Commission on pollution and health. *Lancet* **391**, 462-512 (2018).
6. M. M. Niedzwiecki *et al.*, The exposome: molecules to populations. *Annu Rev Pharmacol Toxicol* **59**, 107-127 (2019).
7. E. L. Schymanski *et al.*, Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol* **48**, 1811-1818 (2014).
8. D. C. Sevin, A. Kuehne, N. Zamboni, U. Sauer, Biological insights through nontargeted metabolomics. *Curr Opin Biotechnol* **34**, 1-8 (2015).
9. B. I. Escher *et al.*, From the exposome to mechanistic understanding of chemical-induced adverse effects. *Environ Int* **99**, 97-106 (2017).
10. E. L. Schymanski *et al.*, Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challenges. *Environ Sci Process Impacts*, (2019).
11. S. Kim *et al.*, PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **47**, D1102-D1109 (2019).
12. A. J. Williams *et al.*, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **9**, 61 (2017).
13. V. Dulio *et al.*, Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. *Environ Sci Eur* **30**, 5 (2018).
14. D. S. Wishart *et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46**, D608-D617 (2018).
15. I. Blazenovic, T. Kind, J. Ji, O. Fiehn, Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8**, (2018).
16. E. L. Schymanski *et al.*, Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **48**, 2097-2098 (2014).
17. C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **8**, 3 (2016).
18. Y. Djoumbou-Feunang *et al.*, BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform* **11**, 2 (2019).

19. D. K. Barupal, O. Fiehn, Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. *Environ Health Perspect* **127**, 97008 (2019).
20. J. M. Gauglitz *et al.*, Untargeted mass spectrometry-based metabolomics approach unveils molecular changes in raw and processed foods and beverages. *Food Chem* **302**, 125290 (2019).
21. S. Li *et al.*, Understanding mixed environmental exposures using metabolomics via a hierarchical community network model in a cohort of California women in 1960's. *Reprod Toxicol*, (2019).
22. C. J. Patel, J. Bhattacharya, A. J. Butte, An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
23. A.-L. Barabási, *Network Science*. (Cambridge University Press, Cambridge, United Kingdom, 2016), pp. xviii, 456 pages.
24. J. Loscalzo, A.-L. Barabási, E. K. Silverman, *Network Medicine : Complex Systems in Human Disease and Therapeutics*. (Harvard University Press, Cambridge, Massachusetts, 2017), pp. xi, 436 pages.
25. G. Bianconi, *Multilayer Networks : Structure and Function*. (ed. First edition., 2018), pp. xiv, 402 pages.
26. A.-L. Barabási, Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113 (2004).
27. D. Mendez *et al.*, ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **47**, D930-D940 (2019).
28. R. R. Tice, C. P. Austin, R. J. Kavlock, J. R. Bucher, Improving the human hazard characterization of chemicals: a Tox21 update. *Environ Health Perspect* **121**, 756-765 (2013).
29. A. Barrat, M. Barthelemy, A. Vespignani, *Dynamical Processes on Complex Networks*. (Cambridge University Press, Cambridge, UK ; New York, 2008), pp. xvii, 347 pages.
30. P. Holme, J. Saramaki, in *Understanding Complex Systems,*. (Springer Berlin Heidelberg : Imprint: Springer,, Berlin, Heidelberg, 2013), pp. VIII, 352 p. 181 illus., 386 illus. in color.
31. M. Santolini, A. L. Barabasi, Predicting perturbation patterns from the topology of biological networks. *Proc Natl Acad Sci U S A* **115**, E6375-E6383 (2018).
32. M. Chadeau-Hyam *et al.*, Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen* **54**, 542-557 (2013).
33. D. I. Walker *et al.*, High-resolution metabolomics of occupational exposure to trichloroethylene. *Int J Epidemiol* **45**, 1517-1527 (2016).
34. B. Warth *et al.*, Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing. *Anal Chem* **89**, 11505-11513 (2017).
35. L. Agier *et al.*, Early-life exposome and lung function in children in Europe: an analysis of data from the longitudinal, population-based HELIX cohort. *Lancet Planet Health* **3**, e81-e92 (2019).
36. A. Jeong *et al.*, Perturbation of metabolic pathways mediates the association of air pollutants with asthma and cardiovascular diseases. *Environ Int* **119**, 334-345 (2018).
37. T. Beck, R. K. Hastings, S. Gollapudi, R. C. Free, A. J. Brookes, GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet* **22**, 949-952 (2014).

38. Y. Zhang, G. Qi, J. H. Park, N. Chatterjee, Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318-1326 (2018).
39. M. C. Turner *et al.*, Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations. *Annu Rev Public Health* **38**, 215-239 (2017).
40. H. Iso *et al.*, The relationship between green tea and total caffeine intake and risk for self-reported type 2 diabetes among Japanese adults. *Ann Intern Med* **144**, 554-562 (2006).

Acknowledgements: We thank the following colleagues for critical review of this manuscript: Rudi Balling, Marc Chadeau-Hyam, George Downward, Linda P. Fried, Dean P. Jones, Vrinda Kalia, Virissa Lenters, Giulia Menichetti, Randolph Singh, Ítalo Valle, Bob van de Water, and Jelle Vlaanderen. **Funding:** RV is supported by the EU H2020-EXPANSE grant, NWO Gravitation Program, and intramural funding from Utrecht University. ELS is supported by the Luxembourg National Research Fund (FNR, Grant A18/BM/12341006). GWM is funded by NIH U2C030163. LB is funded by NIH P01HL132825 and American Heart Association (#151708). **Author contributions:** All authors conceived, wrote, and edited the manuscript. **Competing interests:** Authors declare competing interests. **Data and materials availability:** All data is available in the main text and supplemental material.