Harnessing high-dimensional data in environmental health sciences

Understanding the fate and transport of chemicals in the environment, and their impact on human health, are essential for evaluating their potential harms. Research by Dr Yike Shen and colleagues at Columbia University, New York City, USA, show an association between blood metal levels and changes in the gut microbiome of children. Shen and colleagues also describe improvements in environmental prediction models with the inclusion of extended connectivity fingerprints (ECFP) of pesticides and present a novel advanced autoencoder deep learning model which can learn from complex environmental and health science data.

Humans are exposed to hundreds of chemicals daily, including pharmaceuticals, personal care products, metals (such as lead, mercury, and cadmium), and pesticides. Impacts of those chemicals on human health could include changing the gut microbiome – the microbial composition, mainly bacteria, within the gut. These changes directly impact digestion and absorption of food, but may also indirectly affect mood, behaviour, cognition, and mental health via the endocrine and nervous systems. Establishing the potential harm of chemicals is important to protect humans and the environment, especially through novel methods in computation precision environmental health. Here we look at the wide-ranging research of Dr Yike Shen of Columbia University, New York, USA, and colleagues into assessing environmental exposures and their impact on health using statistical and machine learning methods.

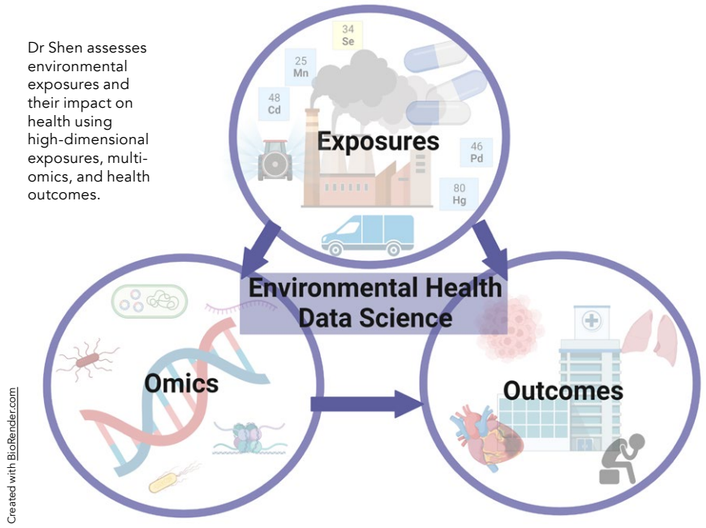
Kateryna Kon/Shutterstock.com

HOW DO CHEMICALS IMPACT THE GUT MICROBIOME?

Few studies have assessed the impact of chemicals on the gut microbiome. Shen and colleagues undertook one of the first epidemiological studies in healthy children to examine the influences of environmental metals on children's gut microbiome. They used genetic sequencing to identify the bacteria in the gut microbiome of children aged six to seven and looked at the association between this composition and the level of metals in the children's blood.

The study found that high levels of manganese (Mn) in the blood was associated with lower levels of bacteria from three families (a level of classification that includes several genera) and one phyla (a higher level of classification that includes several families) and increased abundance of a fourth family. High selenium (Se) concentration was associated with a higher relative abundance of two phyla. The study also investigated the impact of maternal exposure to Mn and Se during pregnancy. Both were found to impact the children's gut microbiome, although fewer phyla were affected.

When looking at the impact of blood metal levels on individual bacterial species, Shen and colleagues used shotgun metagenomic sequencing. This method allows sequencing of short DNA sequences from large numbers of microbes contained in biological samples, such as the gut microbiome, without having to culture individual microbes. The method identifies 'gene families' – groups of evolutionary related protein-coding sequences. The contribution of individual species in each 'gene family' was determined and termed 'gene family-inferred species'. This analysis showed that higher childhood blood cadmium (Cd) levels were associated with a higher relative abundance of 11 gene family-inferred species, including several species of the genera *Bacteroides* and *Bifidobacterium*. Three metals, Cd, Se, and lead (Pb), had positive associations with gene family-inferred species that can be either beneficial or harmful, depending on the context (for example *Bacteriodes vulgatus* and *Eubacterium rectale*). High childhood blood Cd and Pb had positive associations with potentially harmful species, such as *Flavonifractor plautii*. The findings suggest both long- and short-term associations between metal exposure and the childhood gut microbiome, with stronger associations observed with more recent exposure.



Dr Shen assesses environmental exposures and their impact on health using high-dimensional exposures, multi-omics, and health outcomes.

Created with BioRender.com

However, directly altering the gut microbiome is not the only way environmental chemicals can impact human health. Antibiotics released into the environment may lead to the development of antibiotic resistance in bacteria, which in turn can alter the gut microbiome when ingested. Shen and colleagues investigated the impact of antibiotic-containing irrigation water at soil level on the bacterial community of lettuce crops (in the shoots, roots, soil around the roots [rhizosphere], and the bulk soil [soil beyond the rhizosphere]). Bacterial diversity in the bulk soil and lettuce shoots decreased with antibiotic-containing water but remained unchanged in the rhizosphere and lettuce roots. These unchanging rhizosphere and root microbiomes may be due to substances exuded by the roots protecting and stabilising the bacterial community, making them more resilient to external stresses. When considering the number and relative abundance of antibiotic resistance genes and mobile genetic elements (MGEs; genes that can transfer between species), these were higher in the rhizosphere and bulk soil than in the lettuce root and shoot samples. With antibiotic exposure, multidrug-resistant genes decreased to undetectable levels in the rhizosphere but MGEs were consistently increased. Another study by Shen and colleagues highlighted that the risk of lettuce crops being enriched with antibiotic-resistant genes was greater with overhead irrigation than with soil-surface irrigation.

Humans are exposed to hundreds of chemicals daily, including pharmaceuticals, personal care products, metals, and pesticides.

Stokkete/Shutterstock.com

We are exposed daily to many chemicals, including pesticides that we consume through vegetables and fruit.

EXPLORING THE FATE AND TRANSPORT OF CONTAMINANTS IN THE ENVIRONMENT

Understanding the fate and transport of chemicals is essential when evaluating their potential harm to human health. However, the large number of chemicals makes it impossible to evaluate each using traditional laboratory and animal testing methods. Machine learning ('in silico') models provide a novel approach in predicting the impact of chemicals, such as their bioaccumulation and dissipation in the environment. Machine learning involves inputting established data into a computer model, then testing its accuracy to classify a 'verification' data set – a new set of data where the 'correct' classifications are known. As the model's inaccurate classifications are corrected, it learns and becomes increasingly accurate. Shen and colleagues took four machine

Jarun Ontakrai/Shutterstock.com

High childhood blood Cd, Se, and Pb had positive associations with bacteria that can be either beneficial or harmful, depending on the context.

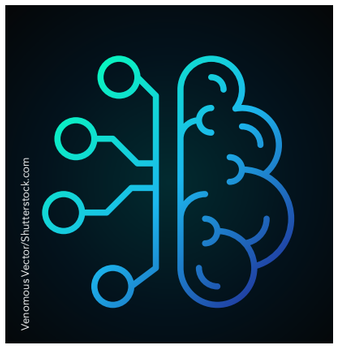
learning models that predict the pesticide dissipation half-lives in plants and developed them to use extended connectivity fingerprints (ECFP) of pesticides (identifiers or 'fingerprints' that are common to pesticides), temperature, plant type, and four plant component classes (plant surface, interior, root, unclassified) as model inputs. Incorporating ECFP into a model has two advantages: 1) directly using the molecular structure will enhance the applicability of the model as many physiochemical properties are not readily available, and 2) ECFP-based input features overcome the difficulty of directly relating pesticide dissipation half-lives/chemical bioaccumulation factors with chemical structures. Results of the study suggest ECFP-based input features in combination with gradient boosted regression tree model (GBRT-ECFP) successfully predicted pesticide dissipation half-life intervals in plants could help establish preharvest interval ranges for pesticides (the time required between last application and harvest). A further application reported by Shen and colleagues is the prediction of root concentration factors (the root:soil ratio of contaminant concentration) of organic (carbon-containing) contaminants in crops (including wheat, carrot, radish, turnips, spinach, celery, Chinese cabbage, maize, pumpkin, and barley).

CAN COMPUTATIONAL MODELS ESTIMATE THE DANGERS POSED BY CHEMICALS?

One measure used to indicate the danger posed by a chemical to humans and the wider environment is the 'hazard concentration for 50', or HC50. HC50 is calculated using the effect concentration 50% (EC50; the concentration of a chemical required to achieve 50% of its maximal effect, such as 50% reduction in algal growth) and lethal concentration 50% (LC50; the concentration of a chemical required to kill 50% of a group during an observation period).

Advanced machine learning models could become invaluable resources in environmental health data science research.

Evaluating a chemical's HC50 requires labour-intensive laboratory-based investigations and often involves animal testing. Traditional quantitative structure–activity relationship (QSAR) models (predictive mathematical models that are usually linear) have their limitations and have difficulty capturing complex, non-linear relationships. Machine learning models that can capture complex, non-linear relationships using multiple decision trees, such as one known as 'the random forest', are available. However, the advanced machine learning models developed by Shen and colleagues increase the prediction accuracy, thus reducing the need for animal testing.



Venomous Vector/Shutterstock.com

The machine learning models Shen and colleagues developed assist toxicity testing.

Shen and colleagues developed a novel technique to predict a chemical's HC50 using more advanced computation modelling. The new technique is known as an autoencoder (or an artificial neural network) deep learning model. HC50 data for 1,815 chemicals contained in the USEtox database and their 691 physiochemical features were used to develop the model. The model compresses these 691 features without losing essential information and 'embeds' them into low-dimensional (only the important features needed to learn/represent the data) latent space (a space where meaningful representations of inputted features can be encoded). Once embedded, the 'chemical embeddings' are learnt (a process called encoding) and can be used to reconstruct (or decode) the chemical features. This encoding–decoding technique allows the model to perform non-linear dimensionality reduction – ie, it can reduce the number of variables inputted from a complex data source, thereby learning meaningful representations of chemical features that other models are unable to learn.

Advanced machine learning models could become invaluable resources in environmental health data science research, given the size and complexity of the data sources involved. The research of Shen and her team is continuing to yield highly valuable findings in environmental health data science.

Behind the Research

Dr Yike Shen

E: ys3419@cumc.columbia.edu **T:** +1 517 488 9282 **W:** yikeshen.github.io @shen_yike

Research Objectives

Dr Shen integrates exposures, multi-omics, and machine learning to assess the effect of environmental exposure on human health.

Detail

Address
630 West 168th Street
P&S Building, Room 16-416
New York, NY 10032, USA

Bio
Dr Yike Shen is a trained environmental scientist with focus on environmental health data science. She is currently a postdoctoral research scientist in the Department of Environmental Health Sciences at Columbia University in the City of New York. She obtained her PhD in environmental toxicology at Michigan State University. She received her BS in environmental science from the University of Alberta.

Funding
NIEHS R01ES027845, R35ES031688, and P30ES009089 (Grants awarded to Andrea A Baccarelli)

Major collaborators
- Dr Andrea A Baccarelli
- Dr Wei Zhang
- Dr Feng Gao

References

Shen, Y, Lane, HE, Shrubsole, MJ, et al, (2022) Associations of childhood and perinatal blood metals with children's gut microbiomes in a Canadian gestation cohort. *Environmental Health Perspectives*, 130(1), 017707-1-10. doi.org/10.1289/EHP9674

Gao, F, Zhang, W, Baccarelli, AA, Shen, Y, (2022) Predicting chemical ecotoxicity by learning latent space chemical representations. *Environmental International*, 163, 107224. doi.org/10.1016/j.envint.2022.107224

Shen, Y, Zhao, E, Zhang, W, et al, (2022) Predicting pesticide dissipation half-life intervals in plants with machine learning models. *Journal of Hazardous Materials*, 436, 129177. doi.org/10.1016/j.jhazmat.2022.129177

Gao, F, Shen, Y, Sallach, JB, et al, (2022) Predicting crop root concentration factors of organic contaminants with machine learning models. *Journal of Hazardous Materials*, 424, 127437. doi.org/10.1016/j.jhazmat.2021.127437

Shen, Y, Ryser, ET, Li, H, Zhang, W, (2021) Bacterial community assembly and antibiotic resistance genes in the lettuce-soil system upon antibiotic exposure. *Science of the Total Environment*, 778, 146255. doi.org/10.1016/j.scitotenv.2021.146255

Gao, F, Shen, Y, Sallach, JB, et al, (2021) Direct prediction of bioaccumulation of organic contaminants in plant roots from soils with machine learning models based on molecular structures. *Environmental Science & Technology*, 55, 16358–16368. doi.org/10.1021/acs.est.1c02376

Shen, Y, Stedtfeld, RD, Guo, X, et al, (2019) Pharmaceutical exposure changed antibiotic resistance genes and bacterial communities in soil-surface- and overhead-irrigated greenhouse lettuce. *Environmental International*, 131, 105031. doi.org/10.1016/j.envint.2019.105031

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Personal Response

How widely do you think your new autoencoder model could be applied?

"The autoencoder model could be used to learn representation from high-dimensional data. Representation learning is actively being used in other fields, but not yet in environmental health sciences. One of the most active topics in the health sciences field now is omics. Omics is a broad scientific field that measures biological molecules in a high-throughput way. Examples of omics include metagenomics, metatranscriptomics, metabolomics, epigenomics, etc. Omics data can serve as biomarkers in response to environmental exposures or health outcomes, and the data are high dimensional. For example, high-resolution mass spectrometry measures thousands of chemical exposures simultaneously, and shotgun metagenomics can sequence hundreds of species, potential pathways, and thousands of strains. Our autoencoder model can be used in omics science. I am currently developing research titled 'representation learning in high dimensional exposure, multi-omics, and health outcomes in environmental health'."

What are the next steps for your research?

"In the next five to ten years, my mission is to protect human and environmental health by using novel computational tools to untangle the complex relationships between high-dimensional exposure and omics data and health outcomes. One of the future directions is in the machine learning direction of the aforementioned representation learning in high-dimensional omics data in predicting health outcomes and learning their patterns in relation to exposures. Another integral part is microbiome epidemiology direction integrating microbiome multi-omics data in association with environmental chemical exposures, and evaluating if microbiome can mediate the environmental exposures and adverse health outcomes."