Only through machine learning and AI can we come to grips with all chemicals around us, say researchers
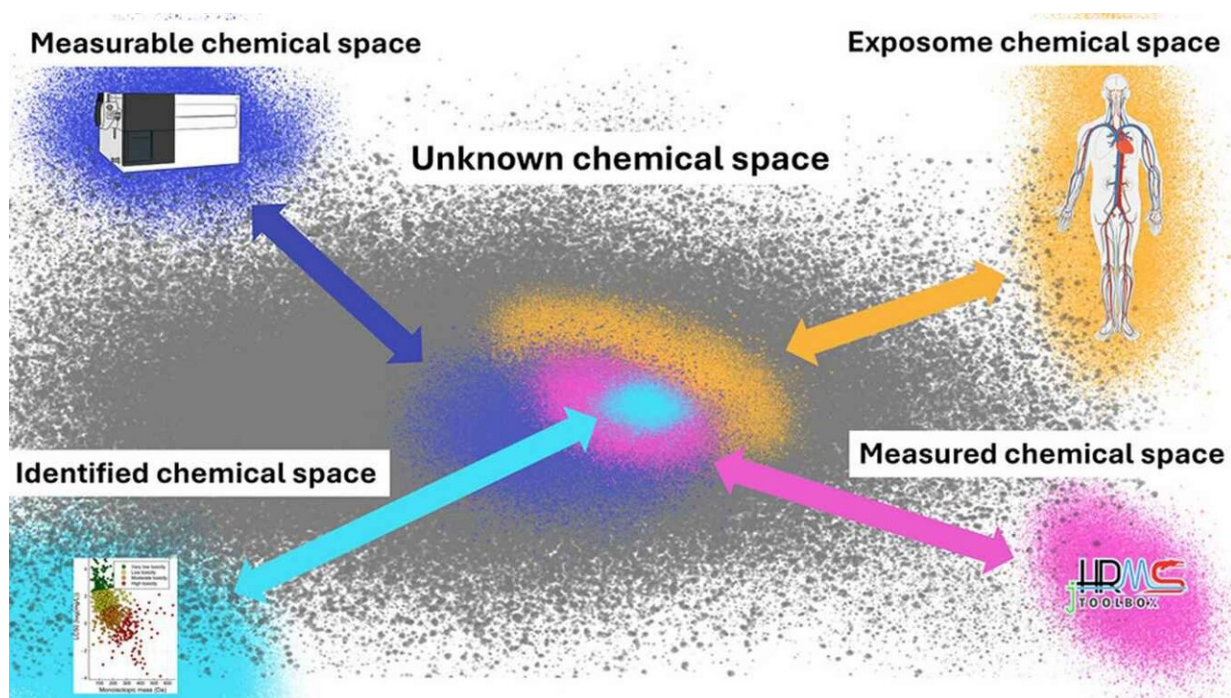
July 1 2024

Illustration of the problem at hand. Of the vast amount of molecules in the chemical space, current technology can detect only a limited amount. The fraction of molecules that have indeed been identified is even smaller. The exposome chemical space—the molecules we are exposed to—lies far beyond the realm of these measurable, measured and identified molecules. Credit: HIMS / JACS.

The *JACS Au* has just [published](#) an invited perspective by Dr. Saer Samanipour and his team on the daunting challenge of mapping all the chemicals around us. Samanipour, an Assistant Professor at the Van 't Hoff Institute for Molecular Sciences of the University of Amsterdam (UvA) takes inventory of the available science and concludes that currently a real pro-active chemical management is not feasible.

To really get a grip on the vast and expanding chemical universe, Samanipour advocates the use of machine learning and AI, complementing existing strategies for detecting and identifying all molecules we are exposed to.

In science lingo the aggregate of all the molecules we are exposed to is called the "exposome chemical space" and it is central to Samanipour's scientific endeavors. It is his mission to explore this vast molecular space and map it to the most "remote" corners. He is driven by curiosity, but even more so by necessity.

Direct and indirect exposure to a myriad of chemicals, mostly unknown, poses a significant threat to the human health. For instance, estimates are that 16% of global premature deaths are linked to pollution.

The environment suffers as well, which can be seen, for example, in the loss of biodiversity. The case can be made, according to Samanipour, that humankind has surpassed the safe operating space for introducing human-made chemicals into the system of planet Earth.

Current approach is inherently passive

"It is rather unsatisfactory that we know so little about this," he says. "We know little about the chemicals already in use, let alone that we can keep up with new chemicals that are currently being manufactured at an unprecedented rate."

In a previous study, he estimated that less than 2% of all chemicals that we are exposed to have been identified.

"The way society approaches this issue is inherently passive and at best reactive. Only after we observe some sort of effects of exposure to chemicals do we feel the urge to analyze them. We attempt to determine their presence, their effect on the environment and [human health](#), and we try to determine the mechanisms by which they cause any harm.

"This has led to many problems, the latest being the crisis with PFAS chemicals. But we have also seen major issues with flame retardants, PCBs, CFCs and so on," he adds.

Moreover, regulatory measures are predominantly aimed at chemicals with a very specific molecular structure that are produced in large quantities.

Samanipour says, "There are countless numbers of other chemicals out there where we don't know much about. And these are not only man-made; nature also produces chemicals that can harm us. Through purely natural synthetic routes, or through the transformation of man-made chemicals."

In particular the latter category has been systematically overlooked according to Samanipour. "Conventional methods have catalogued only a fraction of the exposome, overlooking transformation products and often yielding uncertain results."

We need a data-driven approach

The paper in JACS Au thoroughly reviews the latest efforts in mapping the exposome chemical space and discusses their results. A main bottleneck is that conventional chemical analysis is biased towards

known or proposed structures, since this is key to interpreting data obtained with [analytical methods](#) such as chromatography and mass spectrometry (GC/LC-HRMS). Thus the more "unexpected" chemicals are overlooked. This bias is avoided in so-called non-targeted analysis (NTA), but even then results are limited.

Over the past 5 years, 1600 chemicals have been identified while every year around 700 new chemicals are introduced into the US market alone.

Samanipour says, "When you take the potential transformation products of these novel chemicals into account, you have to conclude that the speed of NTA studies is far too slow to be able to catch up. At this rate, our chemical exposome will continue to remain unknown."

The paper lists these and many more bottlenecks in current analytical science and suggest ways to improve results. In particular the use of machine learning and [artificial intelligence](#) will really push the field ahead, Samanipour argues.

He says, "We need a data-driven approach along several lines. Firstly, we should intensify the datamining efforts to distill information from existing chemical databases. Already recorded relations between structure, exposure and effect of identified chemicals will lead us to new insights. They could for instance help predict the health effects of related chemicals that are yet unidentified.

"Secondly, we have to perform retrospective analysis on already available analytical data obtained with established methods, expanding the identified chemical space. We will for sure find molecules there that have been overlooked until now. And thirdly, we can use AI to work on understanding the structure and scope of the exposome chemical space."

Of course all this is a very complex, daunting matter, Samanipour

realizes. But as a sort of astronaut in molecular space—just like the explorers of the factual universe—he won't let that complexity put him off. "We have to work hard to tackle this. I have no illusion that during my scientific career we will be able to fully chart the exposome [chemical space](#). But it is imperative that we face its complexity, discuss it and take the first steps towards getting to grips with it," he adds.

Provided by University of Amsterdam