# Generative AI in Chromatographic Analysis

## The Speed of Development of AI Applications Has Increased Exponentially Across all Areas of Research

*Generative AI refers to the subset of artificial intelligence focused on creating new content and analyzing existing content, whether that be text, images, or in the case of chromatography, generating hypothetical data sets, methods, and models. Generative AI can impact the field of chromatography in various innovative ways, however, great care needs to be taken to carefully validate any use of this emerging technology from a technical, quality and regulatory perspective.*

It is important to draw proper distinctions with the field of artificial intelligence and this article will concentrate on deep learning generative machine and large language models which, when trained with a suitable amount of raw data, can produce statistically probably outputs.

Existing machine learning models develop artificial intelligence through learning gathered from input data, developing and refining this learning from the patterns which emerge without human intervention. Perhaps the turning point in the exploration and interest in more generative models, was the development and introduction of variational autoencoders (VAE's) which added a critical ability to produce variations on the training data, are easier to scale and can handle volumes of data which were previously unmanageable (in human terms). In the prevailing years, a huge number and variety of artificial intelligence models have been introduced, including the generative adversarial networks (GANs), which produce original data based on transformer tools, working with unlabeled data in a parallel fashion, significantly speeding up the training process.

### Innovation in Chromatographic Method Development

Chromatography relies heavily on the optimization of numerous variables to achieve the desired separation. There have been a growing number of studies which cite the use of generative AI in predicting chromatographic retention times, optimizing separations and helping to discover previously unresolved analytes.

DeepLC, a deep learning peptide retention time predictor using peptide encoding based on atomic composition has been used to predict the retention time of (previously unseen) modified peptides. Several other studies are cited in this area and include the use of deep learning for retention time prediction in reversed-phase liquid chromatography and predictive retention index modelling of organic compounds in one and two-dimensional gas chromatography.

© Element Materials Technology

Tony Taylor,
Element Materials
Technology

In his 2021 paper 'Perspective on the Future Approaches to Predict Retention in Liquid Chromatography', Fabrice Gritti postulates that in order to add further rigor to the simulated method development process, account needs to be taken of the fundamental solid-liquid adsorption process which, when using Monte Carlo or Molecular Dynamics simulations, can explain the complexi-

©Anchalee - stock.adobe.com

ties of retention data in reversed phase HPLC systems which are beyond empirical or statistical models.

## AI in Data Analysis

Perhaps the most active area of development for generative AI is the prediction, deconvolution and interpretation of mass spectral data. Whilst statistical tools for these operations have been in use for several years, generative AI approaches are gathering pace in the literature.

IDSL_MINT is a customizable deep-learning framework to train and utilize new models to predict molecular fingerprints from spectra for the compound annotation workflows in LC-MS/MS of untargeted metabolomics and exposomics datasets.

General adversarial models have been used to generate predictions for the rose oil characteristics of the Taif Rose, based on a training set of GC-MS spectra. This could lead to a reduction in laboratory testing and higher throughput screening to select genotypes for cultivation which retain a controlled profile of volatiles.

> *"The most active area of development for generative AI is the prediction, deconvolution and interpretation of mass spectral data."*

Deep learning models have been used to interpret the LC-MS spectra of novel psychoactive substances and transformer models have been used to elucidate the structures of small molecules within the metabolomics sphere, with the aim of maximum coverage of biologically feasible small molecules in this space.

## Generative AI for the People

For those in the laboratory who wait for the VAE and GAN based models to be successfully commercialized, what is possible using the popular text (and now image) based AI engines such as ChatGPT (GPT4) or Perplexity AI.

We have recently used GPT4 to undertake image recognition of problematic chromatograms or the mass spectra of compounds which are not contained in popular or in-house libraries.

For troubleshooting chromatographic data, we have been very pleasantly surprised by the LLMs ability to discern problematic features and offer suggestions to overcome the issues. Generally, some fundamental understanding of chromatography theory and experience are necessary for positive identification of issues, but as a diagnostic aide, LLMs offer a useful contribution.

We have also successfully used ChemCrow as an API for GPT4 to provide more domain expertise. This model contains enables the generation of molecular structures from SMILES strings and vice versa, can cite patents including the structure and deliver physicochemical information such as Tanimoto similarity indices, pKa and LogP(D) data etc. The inclusion of this domain expertise within a generative context can be very useful.

## Future Perspectives

Instrument vendors are currently exploring the use of the huge amounts of instrument telemetry data which are gathered from each chromatographic analysis to build models which could be used to predict failure rates and modes and to monitor equipment and columns to assess the likelihood for a system being fit for purpose for a particular analysis (will the column last for these 100 injections!).

We also believe that the production of synthetic data sets from limited training data represents an exciting possibility for future development of generative tools.

## Conclusions

Following the introduction of VAE and GAN generative models, the speed of development of AI applications has increased exponentially across all areas of research. The speed, efficiency and data handing capabilities of these models lend themselves well to further development of tools to assist the chromatographer. However, the data produced needs to be carefully validated and we need to consider the use of these generative tools in line with regulatory and ethical frameworks.

***Tony Taylor, Chief Scientific Officer, Element Materials Technology, Manchester, UK***

- tony.taylor@element.com
- www.element.com

References can be requested from the author.