


Spectrum of the past

Michael A. Skinnider

 Check for updates

Thirty-four years ago, Curry and Rumelhart described a neural network-based approach to annotate tandem mass spectra. Their ideas foreshadowed several important developments in computational mass spectrometry over the past decade, but many of the challenges they discuss remain relevant today.

REFERS TO Curry, B. & Rumelhart, D. E. MSnet: a neural network which classifies mass spectra. *Tetrahedron Comput. Methodol.* **3**, 213–237 (1990).

Identifying small molecules from their tandem mass spectra – also known as MS/MS or MS² – is a challenge that many undergraduates first encounter in second-year organic chemistry. But the simple examples presented in chemistry textbooks belie a profoundly challenging problem. Experts can often elucidate chemical structures from tandem mass spectra, but this process requires careful manual analysis by a highly trained chemist. This kind of laborious manual interpretation does not scale to the billions of mass spectra that are now collected by untargeted metabolomics¹ – large-scale studies of small molecules present in biological systems (or metabolites) – and, as a result, most of the spectra acquired in these experiments currently go unidentified².

By 1990, the possibility of automating structure elucidation from tandem mass spectra had already attracted a great deal of interest. Beginning in the 1960s, the Dendritic Algorithm (better known as DENDRAL) project sought to enable de novo structure elucidation by first inferring structural constraints from spectral data, and then computationally enumerating every possible structure consistent with these constraints³. A competing tool, named self-training interpretive and retrieval system (STIRS), sought to deduce the presence of key substructures based on the most similar spectra within a reference library⁴. Still other tools used simple statistical approaches to infer whether unidentified compounds are members of broad chemical classes (for example, polycyclic aromatic hydrocarbons) from their tandem mass spectra⁵.

Curry and Rumelhart took a different approach⁶. They sought to develop a machine learning model that, given a tandem mass spectrum as input, could predict the presence or absence of key chemical substructures in the unidentified molecule. To achieve this, they turned to artificial neural networks – an established class of models that in 1990 had “recently become the object of renewed interest” (ref. 6). Rumelhart was deeply familiar with these models, having published a seminal paper just a few years earlier on the backpropagation algorithm that underlies much of modern deep learning⁷.

To train their model, Curry and Rumelhart compiled a dataset that was remarkably large for its time, with 31,926 spectra in the training set alone, and a further 12,671 in the test set. These were low-resolution mass spectra in which intensities were recorded at integral mass-to-charge ratio (m/z) bins. Recognizing the importance of how these spectra were

represented as input to the neural network, the researchers devoted special effort to devising an appropriate set of mass spectral features. These included not just the intensities of fragment ions and neutral losses, but also a series of carefully handcrafted features, such as autocorrelation sums, modulo-14 sums and even sums of individually selected ion series (one feature, for instance, consisted of the sum of intensities at m/z ratios of 45, 57, 58, 59, 69, 70, 71 and 85) that were suggested to be particularly diagnostic for monofunctional aliphatic molecules (for example, see the [mass spectrum of caffeine](#) in Fig. 1). These handcrafted features, Curry and Rumelhart argued, would mitigate the noise inherent in raw ion intensities and instead “accentuate patterns which correlate with molecular structure”.

Curry and Rumelhart selected a set of 36 broad chemical classes – chosen for their subjective ‘chemical interest’ – for the model to use when evaluating the presence of, for instance, carbonyl, phenol or nitro groups in mass spectra used as input. They then trained a multi-task neural network, with a single hidden layer of 80 neurons, to classify the presence or absence of all 36 classes simultaneously from a given mass spectrum. This neural network was named MSnet and took about two weeks to train. However, Curry and Rumelhart noticed that many chemically interesting classes (for instance, phthalates) were so rare in their training set that, despite the presence of highly characteristic fragment ions, they could not be reliably identified by this initial model. To address this gap, they devised a hierarchical learning strategy, whereby a series of more specialized neural networks were trained to recognize sub-categories of each structural class. This strategy was found to enable the prediction of less common structural classes. For example, the “COO” subnetwork correctly learned to recognize more than 92% of the phthalates in the training set, whereas the top-level network never learned to recognize the presence of a phthalate group at all.

To evaluate MSnet, Curry and Rumelhart compared their model to STIRS – the strongest baseline they could identify in the literature. However, this comparison did not prove to be straightforward. STIRS had been trained to predict a different set of substructures than MSnet, using a different dataset of reference spectra as the training set. Moreover, the substructures that overlapped between MSnet and STIRS were sometimes defined differently by the two models: for instance, MSnet used a definition of ‘phenol’ that encompassed all aromatic alcohols, whereas STIRS included only true phenols. Even the metrics used to evaluate performance were defined differently by the developers of STIRS and MSnet. Curry and Rumelhart reported that MSnet demonstrated superior performance to STIRS on 12 of 16 overlapping structural classes, but it is difficult to interpret this result in context. Notwithstanding this, the researchers argued that MSnet also possessed a number of conceptual advantages over STIRS, perhaps most notably the fact that it was capable of estimating well-calibrated probabilities of class membership for any given spectrum.

Although the experimental and computational techniques for acquiring and analysing MS/MS data have evolved dramatically since 1990, it is remarkable how much of Curry and Rumelhart’s manuscript still rings true today. The researchers emphasized the need for automated methods for mass spectral interpretation, the huge amount of effort that had already been brought to bear on this challenge and the

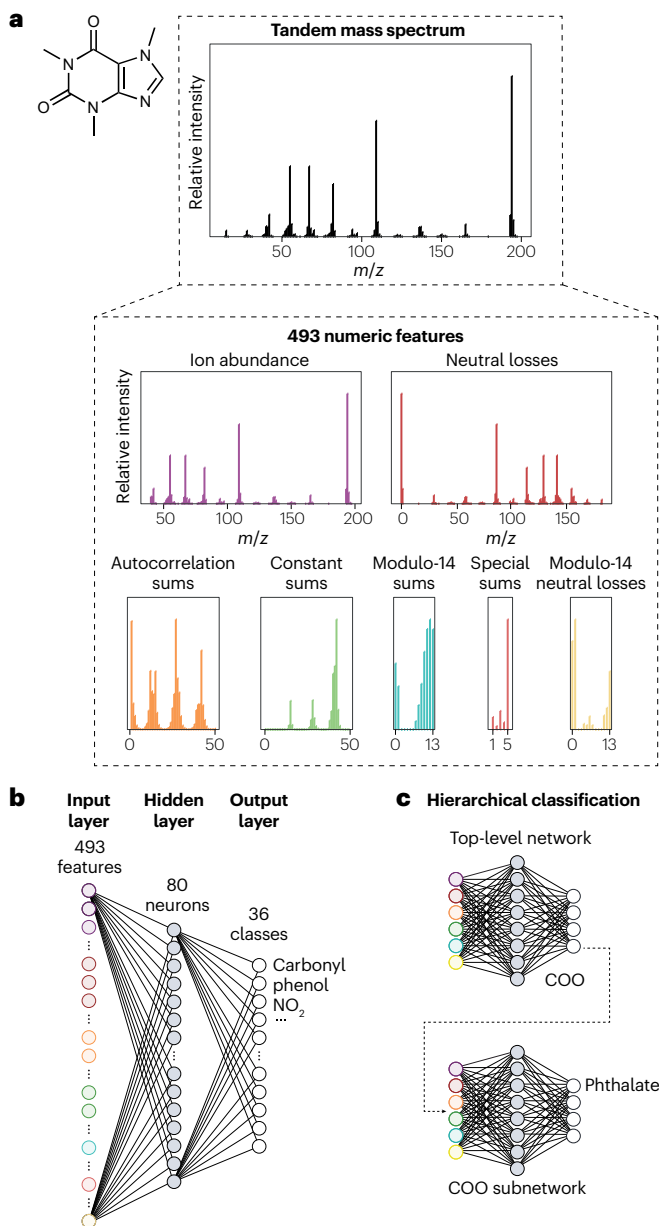


Fig. 1 | Overview of the design of MSnet. **a**, Top, structure and tandem mass spectrum of caffeine. Bottom, overview of the 493 features calculated from an input mass spectrum in MSnet. **b**, Architecture of the top-level neural network in MSnet with 493 input features, a hidden layer of 80 neurons, and an output layer of 36 classes. Only a subset of inputs and connections are illustrated. **c**, Schematic of the hierarchical classification strategy implemented in MSnet. *m/z*, mass-to-charge ratio.

surge of interest in neural networks as a means to solve this problem – all of which are equally true 30 years later. On the other hand, Curry and Rumelhart noted with some dismay that their training dataset “contain[s] a shockingly high proportion (estimated at about 6%) of completely erroneous spectra” (ref. 6), and the quality and integrity of annotated MS/MS data remain an issue today. Whereas much energy is invested in developing new model architectures, comparatively less is devoted to assembling and curating the data required to train these models. Contemporary efforts to evaluate the performance of methods for MS/MS interpretation also face many of the same challenges as Curry and Rumelhart encountered. New machine learning tools are trained and evaluated on different datasets, with different metrics and different evaluation set-ups, and benchmarked selectively against relevant baselines, all of which can make it difficult to discern whether a particular model presents a real advantage over others.

There are also noteworthy conceptual similarities between MSnet and modern methods for MS/MS interpretation. A direct connection can be drawn between the prediction of chemical substructures in MSnet and the prediction of chemical fingerprints in Compound Structure Identification (CSI):FingerID⁸ – the current state-of-the-art method for identifying small molecules from their MS/MS spectra. Unfortunately, Curry, Rumelhart and their contemporaries did not seem to recognize that, even when individual substructures cannot be predicted with perfect accuracy, a large number of weakly accurate predictions can enable unambiguous molecule identification – a key conceptual leap that allowed CSI:FingerID to substantially advance the state of the art.

A notable difference between MSnet and contemporary machine learning approaches is the former’s reliance on handcrafted features. Today, deep neural networks are noted for their ability to automatically extract representations relevant to the task at hand from raw input data. However, tandem mass spectra are often provided as input to neural networks by summing ion intensities within small *m/z* bins (for example, 0.01 Da) to produce very long, sparse vectors⁹. In this setting, Curry and Rumelhart’s creative ideas about how to create features from MS/MS input spectra for machine-learning models might be worth revisiting in the present day.

Curry and Rumelhart’s paper coincided with a surge of interest in developing machine learning approaches to interpret MS/MS spectra. However, by the end of the 1990s, this enthusiasm had largely faded, leading one commentator to conclude that “the ‘heyday’ of the development of techniques based on artificial intelligence (AI) for automated mass spectral interpretation is past” (ref. 10). Today, with renewed interest in the possibility of deciphering mass spectra with chemical AI, it is exciting to reflect on how far the state of the art has advanced since Curry and Rumelhart’s seminal work. However, the fact that many of the challenges that Curry and Rumelhart identified remain as salient today as they were in 1990 points to clear opportunities for continued progress.

Michael A. Skinnider

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. ²Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, USA.

✉ e-mail: skinnider@princeton.edu

Published online: 05 January 2024

References

- Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 0054 (2017).
- da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
- Smith, D. H., Gray, N. A. B., Nourse, J. G. & Crandell, C. W. The dendral project: recent advances in computer-assisted structure elucidation. *Anal. Chim. Acta* **133**, 471–497 (1981).
- Dayringer, H. E., Pesyna, G. M., Venkataraghavan, R. & McLafferty, F. W. Computer-aided interpretation of mass spectra. Information on substructural probabilities from stirs. *Org. Mass Spectrom.* **11**, 529–542 (1976).
- Lohninger, H. & Varmuza, K. Selective detection of classes of chemical compounds by gas chromatography/mass spectrometry/pattern recognition: polycyclic aromatic hydrocarbons and alkanes. *Anal. Chem.* **59**, 236–244 (1987).
- Curry, B. & Rumelhart, D. E. MSnet: A neural network which classifies mass spectra. *Tetrahedron Comput. Methodol.* **3**, 213–237 (1990).
- Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
- Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
- Voronov, G. et al. Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data. Preprint at <https://arxiv.org/abs/2207.02980> (2023).
- Palmer, P. T. Gas chromatography/mass spectrometry. In *Encyclopedia of Analytical Chemistry: Applications, Theory, and Instrumentation* (ed. Meyers, R. A.) (Wiley, 2007).

Competing interests

The authors declare no competing interests.

Related links

Caffeine on the National Institute of Standards and Technology (NIST) chemistry webbook: <https://webbook.nist.gov/cgi/cbook.cgi?ID=C58082&Mask=2>