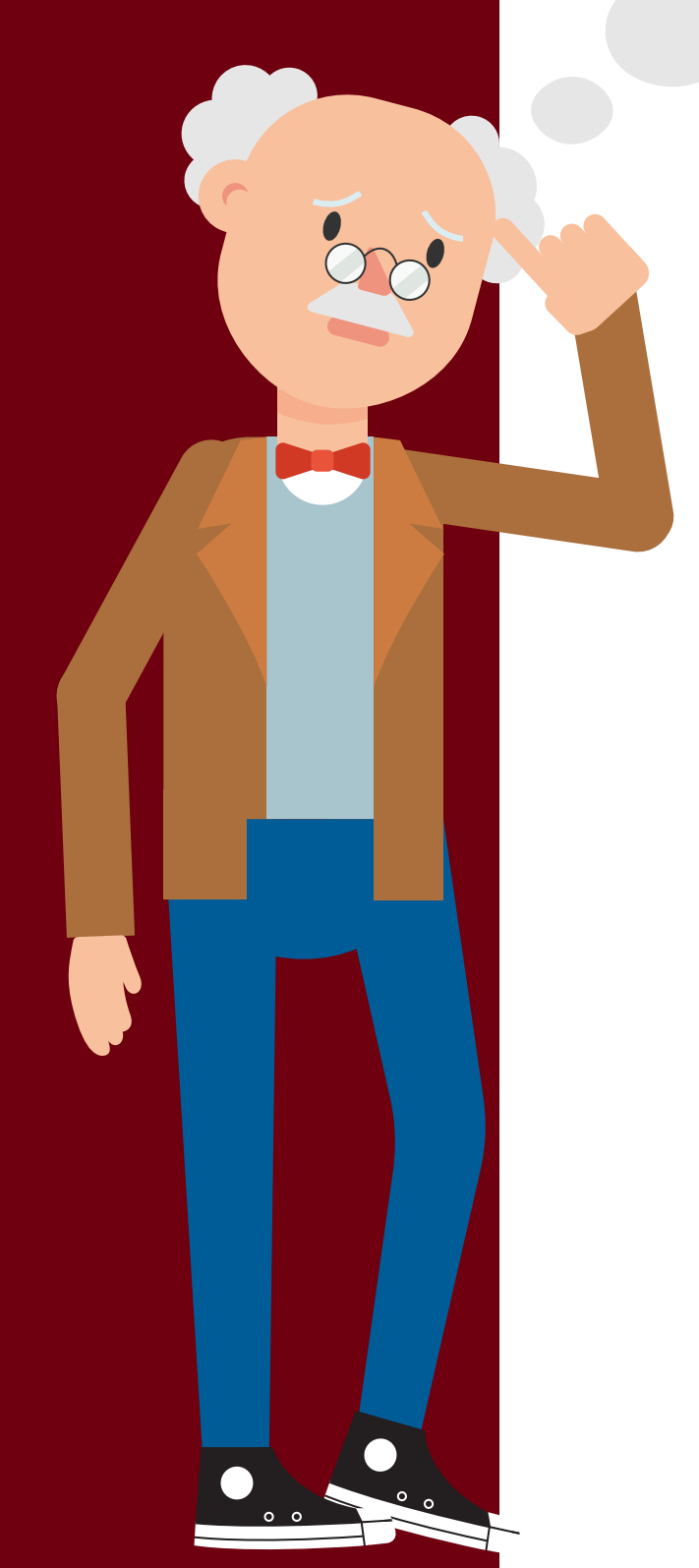
# Visualizing the History of LC-MS by Combining Text Mining and QSPR

Magnus Palmblad[1] and the
ASMS History Committee[2]

[1]Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, the Netherlands
[2]P. Jane Gale (Chair), Mariam ElNaggar, Michael Grayson, O. David Sparkman, Kenneth Tomer and Alfred Yergey

## OVERVIEW

- The applications of LC-MS and other separation techniques combined with several MS ionization methods were investigated using text mining and quantitative structure-property relationships (QSPR).
- The mass and polarity distributions of molecules found in the literature were visualized (excluding most peptides and proteins).
- The results reveal clear trends in LC-MS applications.



CAN TEXT MINING AND
COMPUTATIONAL CHEMISTRY REVEAL
THE HISTORY OF LC-MS?

## DATA AND METHODS

In this work we exclusively used open data and resources. Europe PMC[1] contains all of PubMed and PubMed Central and can be searched programmatically using REST or SOAP web services. Text-mined terms are available for 34 million abstracts and 1.8 million full-text articles. The online chemical database with modeling environment (OCHEM)[2] collects 2 million records for 545 properties from 12,776 sources and offers 152 public models for predicting a wide range of properties from molecular structures in the SMILES format. We combined these resources to match small molecule annotations in Europe PMC with their physico-chemical properties:

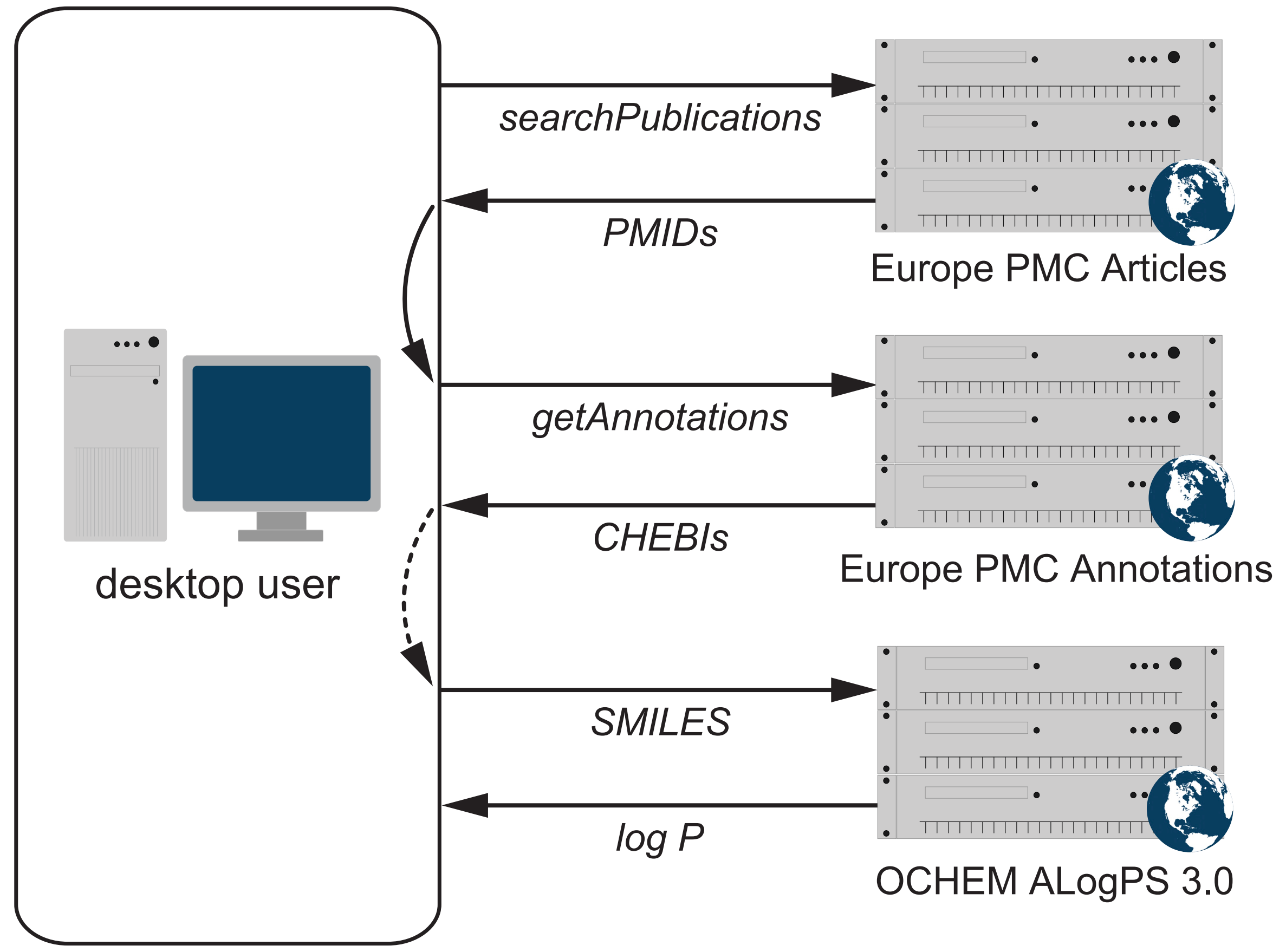


**Figure 1.** Literature searches were executed using the Articles API from Europe PMC, retrieving a list of PMIDs. The PMIDs for articles with text-mined items were submitted to the Annotations API to retrieve annotated chemicals (ChEBI identifiers). Physico-chemical properties such as log P (octanol-water partition coefficient) were predicted from the 90,698 unique chemical compounds in the ChEBI ontology[3] using the ALogPS 3.0 model. The results were then combined on the client side.

## RESULTS

The results from any literature search can be visualized as distributions over any properties that can be calculated, e.g. mass and polarity:

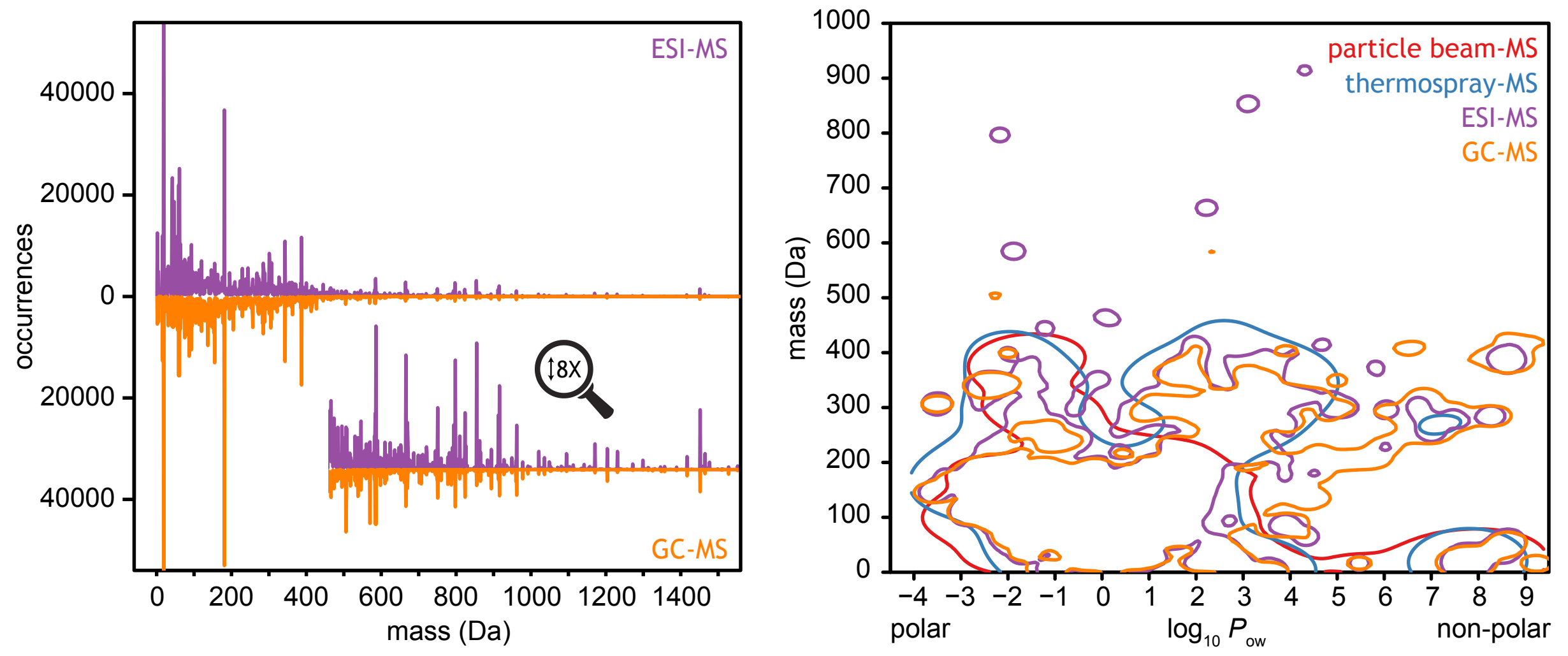


**Figure 2.** Chemical annotations in the literature displayed as “mass spectra” (left) or mass/polarity contours (right). Two frequent nominal masses are 18 (water) and 180 (e.g. glucose). The spectra in this example are based on 2,033,888 occurrences of 11,864 chemical compounds in 152,043 articles! All mass axes really are mass, not *m/z*!

To more easily identify ‘hot spots’, we can also visualize the literature search results as heatmaps (Figure 3) or RGB/CMYK plots (Figure 4):

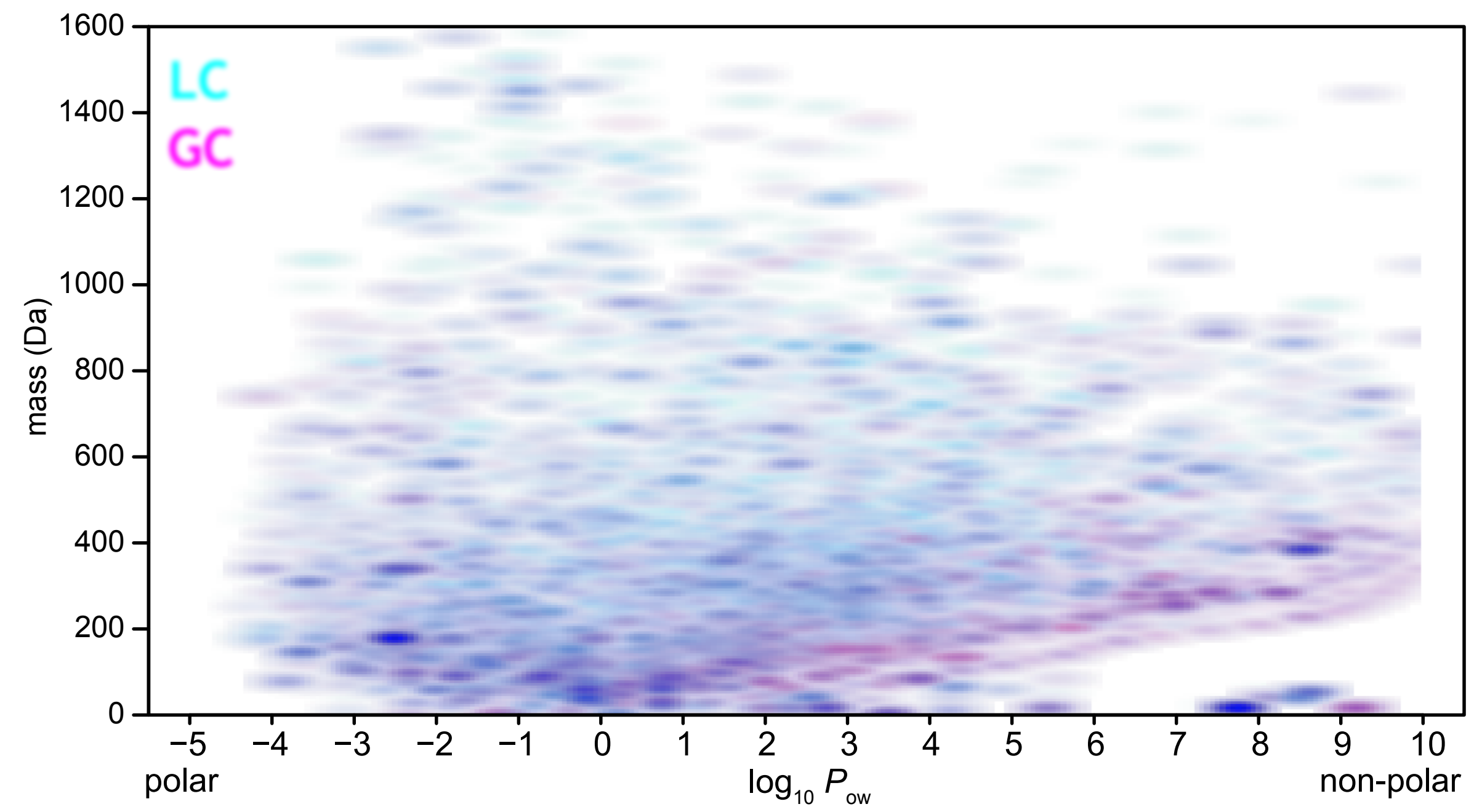


**Figure 3.** Application of liquid chromatography (cyan) and gas chromatography (magenta) as function of polarity and mass of chemical compounds co-appearing with these methods in the literature.

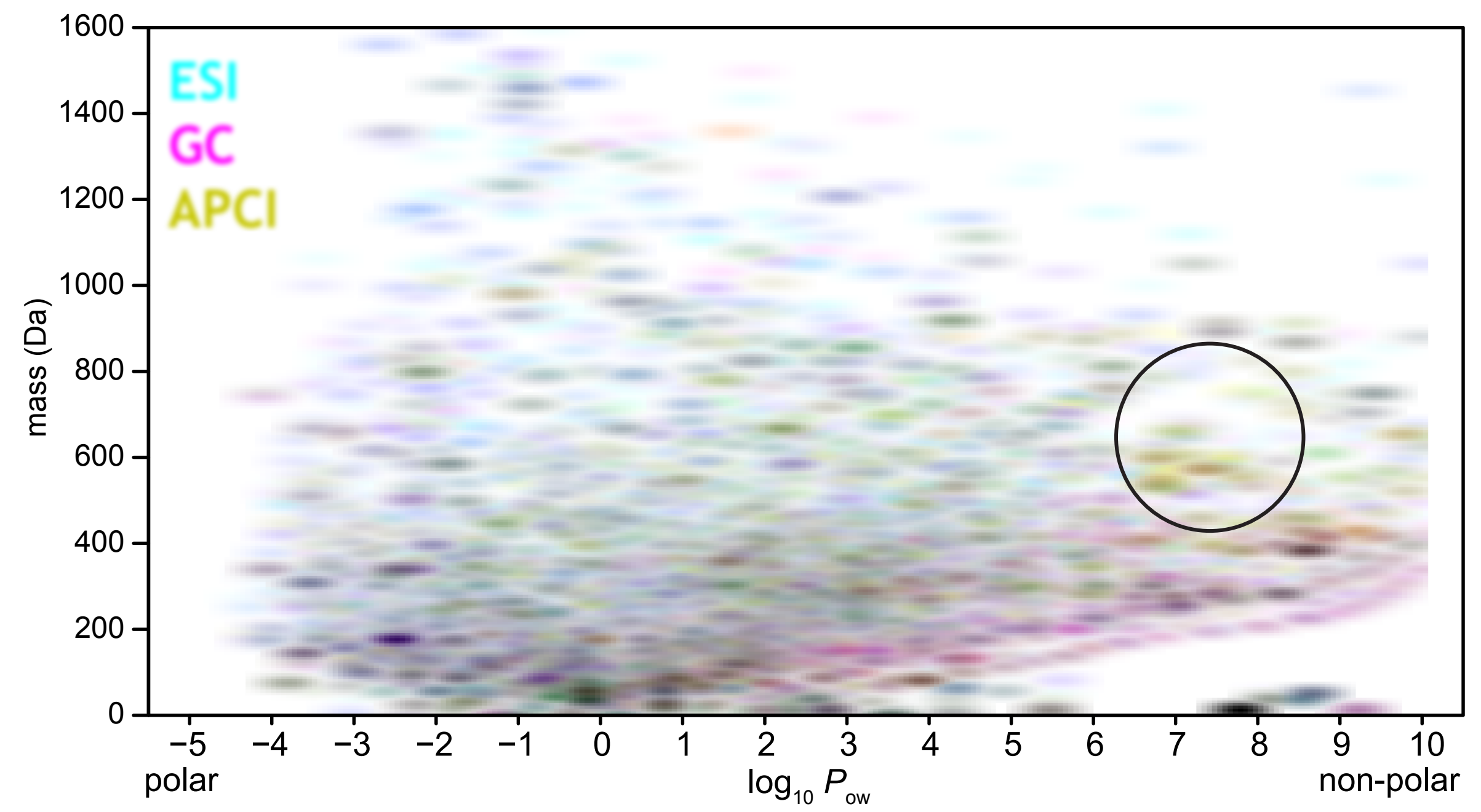


**Figure 4.** Comparison of ESI (cyan), GC (magenta) and APCI (yellow) reveals a mass/polarity ‘hot spot’ where APCI is most applicable (circle), as well as the region of small, non-polar analytes dominated by GC.

We can then restrict the searches to a defined time period and visualize how the application of a particular technology, e.g., LC-MS, has evolved:

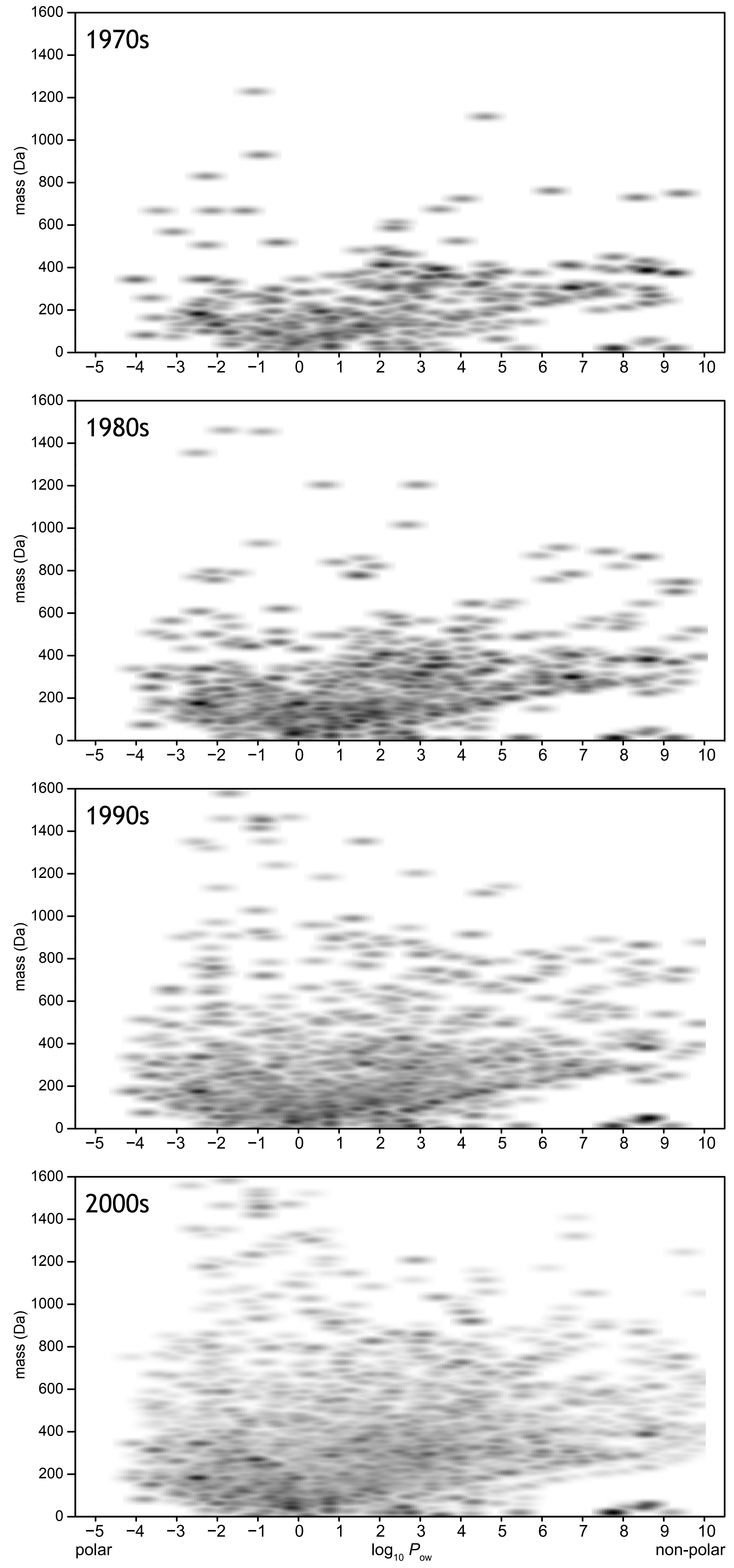


**Figure 5.** Evolution of LC-MS over four decades. Liquid chromatography in combination with softer ionization methods has granted access to a larger analyte space, including large and polar compounds.

Averaging over large numbers of publications on LC-MS reveals trends:

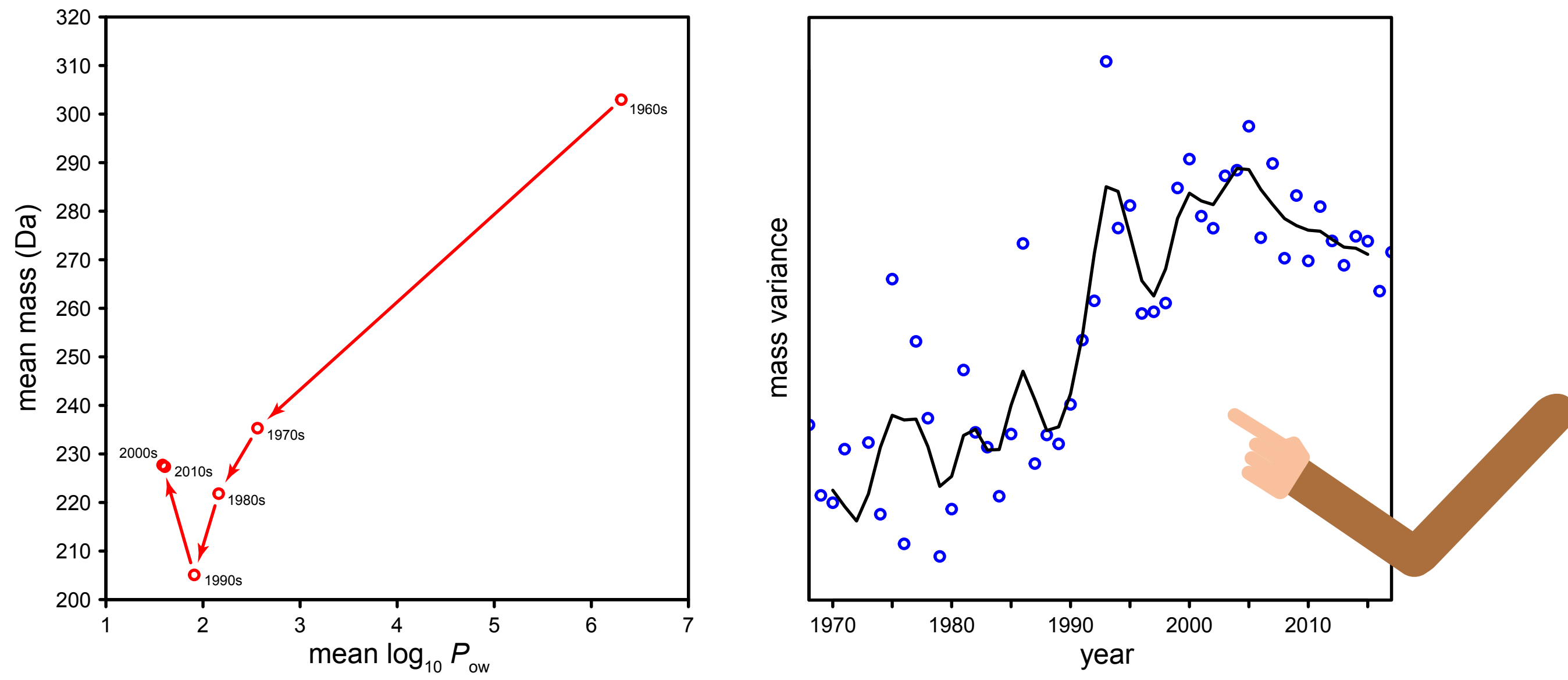


**Figure 6.** Changing applications of LC-MS over five decades. Before the 1990s, the main trend was toward incresing polarity (but smaller mass). Between the 1980s and the 2000s, the trend changed toward larger mass (left). The analyte mass range, as measured by variance, expanded most rapidly between 1989 and 1993 (right). The current rate of change is very slow, at least in these physico-chemical dimensions. The number of annotations increases over time, from 48 annotations in 8 papers on LC-MS in the 1960s (1 in 1968, 7 in 1969) to nearly 1.9 million annotations in 53,714 papers from the 2010s. The real effect is likely larger still, as few peptides and almost no proteins are covered by ChEBI.

## CONCLUSIONS

Can text mining combined with computational chemistry reveal trends in the applications of a particluar technology, such as LC-MS? Yes!

The analyses comfirm much of what is generally known to practicioners in the field. However, the bibliometric analyses are objective - i.e., data-driven and do not require any expertise in the investigated topics.

Research output has increased over time in most fields. For technical and licensing reasons, the coverage and number of annotations also increases with year of publication. This is a caveat in historical studies: absolute number of occurrences should not be compared over time, but properties and relative frequencies may be. Other caveats of text mining also apply, such as false and/or missing matches to search terms.

## Acknowledgements

The authors would like to thank Igor Tetko and Aswin Verhoeven for technical assistance. Professor Thinking Cartoon based on one by VideoPlasty (CC BY-SA 4.0) from Wikimedia Commons. The cartoon is not meant to represent contributers to the field in any way.

## References

1. Europe PMC: a full-text literature database for the life sciences and platform for innovation, The Europe PMC Consortium, *Nucleic Acids Res*. 2015 January 28; 43(Database issue): D1042–48.

2. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information., Sushko *et al.*, *J Comput Aided Mol Des*. 2011 Jun; 25(6): 533-54.

3. ChEBI in 2016: Improved services and an expanding collection of metabolites. Hastings *et al.*, *Nucleic Acids Res*. 2016 Jan 4; 44(D1): D1214-9.