ARTIFICIAL INTELLIGENCE IN CHEMISTRY: CURRENT LANDSCAPE AND FUTURE OPPORTUNITIES

CAS

A division of the
American Chemical Society

Introduction

Artificial intelligence (AI) is the ability of machines to simulate human intelligence. Where regular computers are programmed to act based on pre-programmed rules, such as true/false or if/else statements, AI is designed to understand the relationships between data and develop novel solutions to problems. Since the 1950s, many types of AI have emerged, including machine learning (Figure 1). Each specialty trains computers to learn from data in unique ways.

AI techniques have been widely employed across a range of disciplines, particularly in scientific research where AI has been used to understand molecular properties, design molecules, and predict reaction outcomes. One area of science that has seen a huge increase in research related to AI is chemistry.

Since 2015, publications and patents using AI methodologies have increased drastically. Through AI, researchers have been able to make leaps in data processing that would otherwise have taken several decades if undertaken manually. Some examples include:

- Predicting the bioactivity of new drugs
- Optimizing reaction conditions
- Suggesting synthetic routes to complex target molecules

Despite the advancements in AI-related problem solving, there may still be significant opportunities for disciplines within chemistry that have low AI adoption. To understand where these opportunities are, we examined the landscape of AI in chemistry and what the barriers to adoption may be.

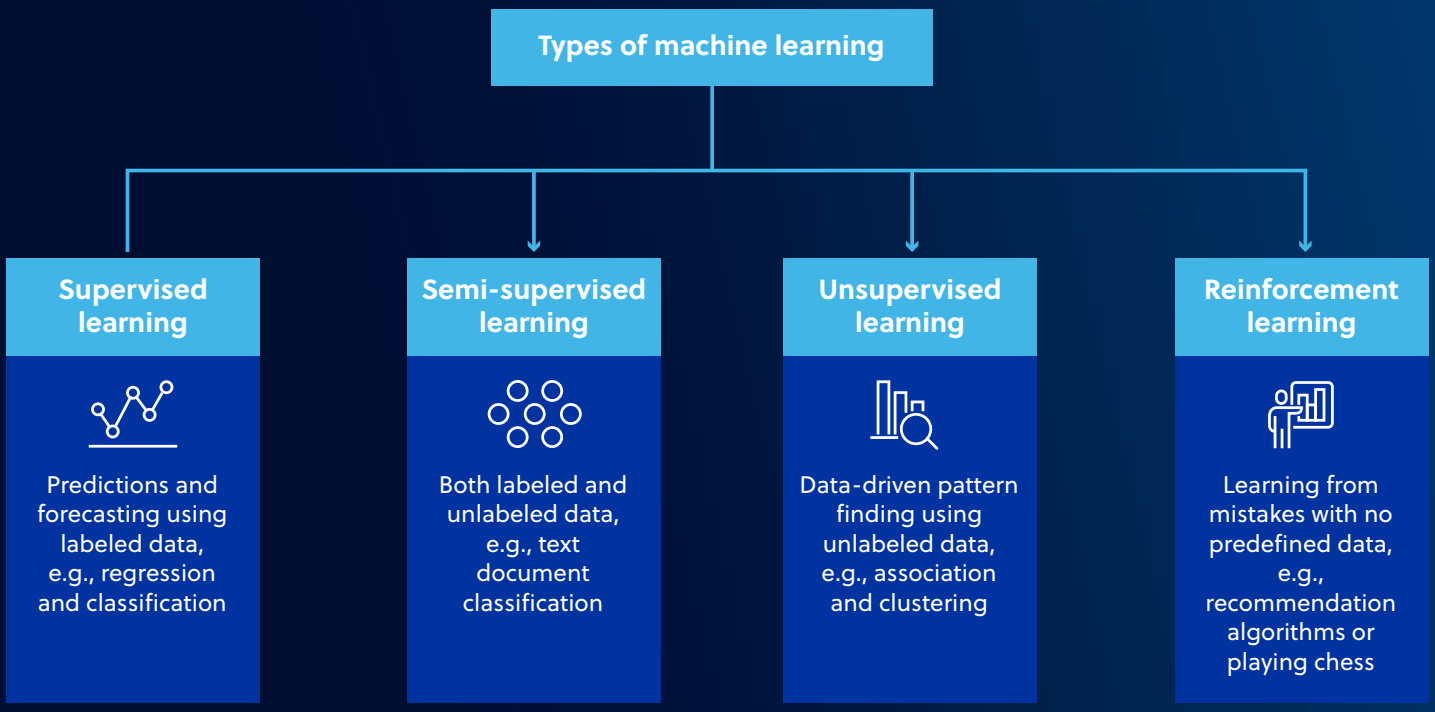


Figure 1. Overview of the types of machine learning

CAS Content Collection ™ and the growing role of AI in chemistry

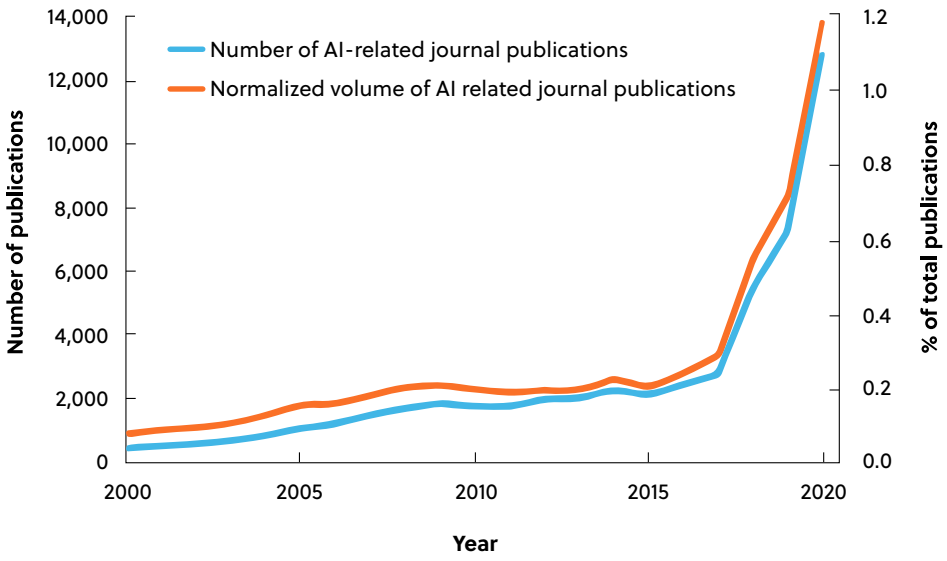


Figure 2. Annual publication volume of AI-related chemistry journal articles from 2000 to 2020

CAS is a leader in scientific information solutions – curating, connecting, and analyzing the valuable data disclosed in scientific literature globally to help accelerate breakthroughs. Our team of scientists and AI experts are at the heart of our collection, covering over 150 years of discoveries to build the highest quality and most up-to-date collection of scientific information in the world – the CAS Content Collection.

The CAS Content Collection is the largest library of chemical information, and we used this to contextualize the current AI landscape by classifying and quantifying all chemistry publications related to AI between 2000 and 2020.

Worldwide, scientific publications are growing at a rate of around 8% per year, equivalent to a two-fold increase of scientific output every nine years. Interestingly, the growing number of AI-related journal publications in chemistry, relative to that of all scientific journal publications (as seen in Figure 2), suggests that this topic is rapidly gaining momentum in the research community. There are several reasons for the explosion of AI publications in chemistry since 2015, including the introduction of open-source machine-learning frameworks, such as TensorFlow and PyTorch, and deep learning and image recognition demonstrations that were widely circulated online – all of which have increasingly drawn interest from the scientific community.

In fact, 50% of all chemistry publications related to AI have been published in the last 4 years. Countries leading the field are the US and China, accounting for over 40% of journal publications worldwide. India, Iran, the UK, and Germany together account for 20% of all published articles.

> The expert-curated content from CAS is suitable for the quantitative analysis of publications against variables such as time, country, research area, and the details of the substance studied.

The chemistry disciplines embracing the AI revolution

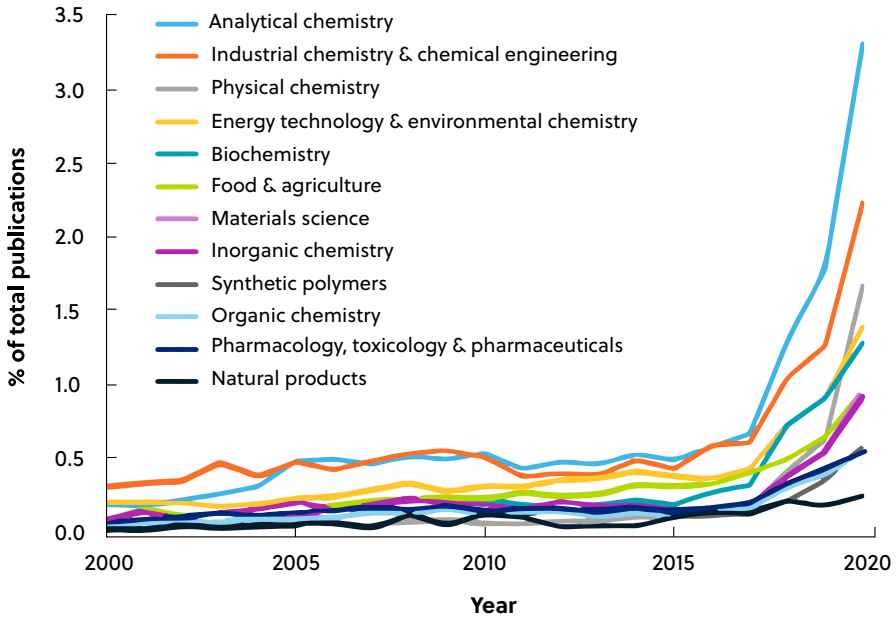
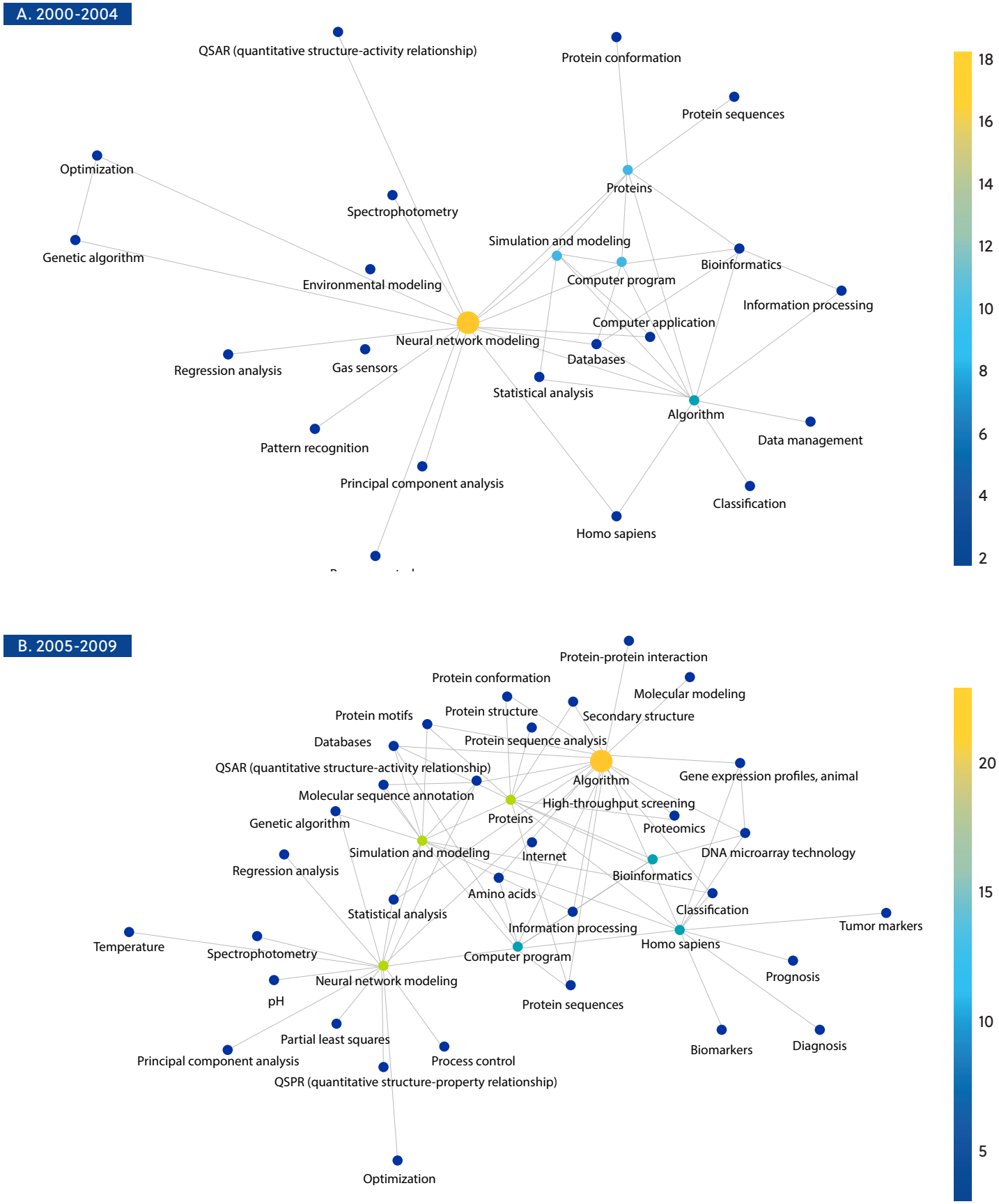


Figure 3. Journal publication trends of AI in specific research areas from 2000 to 2020

We analyzed the CAS Content Collection to understand the growth and distribution of AI-related publications and patents in chemistry. We then delved further to identify the topics studied, notable publications and patents,and the types of chemical substances most frequently involved.

Our analysis revealed that the major contributors to AI-related journal publications include analytical chemistry, industrial chemistry and chemical engineering, and physical chemistry. areas with a less rapid adoption include natural products, organic chemistry and pharmacology, toxicology and pharmaceuticals. For an accurate comparison, we normalized the numbers of AI-related publications in each area to that area's respective total year publication volume (Figure 3). Interestingly, when looking at patents, biochemistry is among the fields most represented in AI-related applications alongside analytical chemistry. However, in terms of journal publications, its proportion is relatively moderate. This incentive to patent AI technologies could potentially be due to the use of biochemistry in drug research and development.

In addition to examining the numbers, the connections between frequently used concepts of research topics and AI algorithms have been explored over the last 20 years to understand the problems that AI is helping to solve (Figure 4).



2000–2009 – In the years 2005–2009, the term 'homo sapiens' was a popular term as AI was used to explore diagnoses and prognoses related to human-based diseases. In addition to 'homo sapiens', other related concepts appeared, including 'biomarkers', 'tumor markers prognosis' and 'diagnosis'.

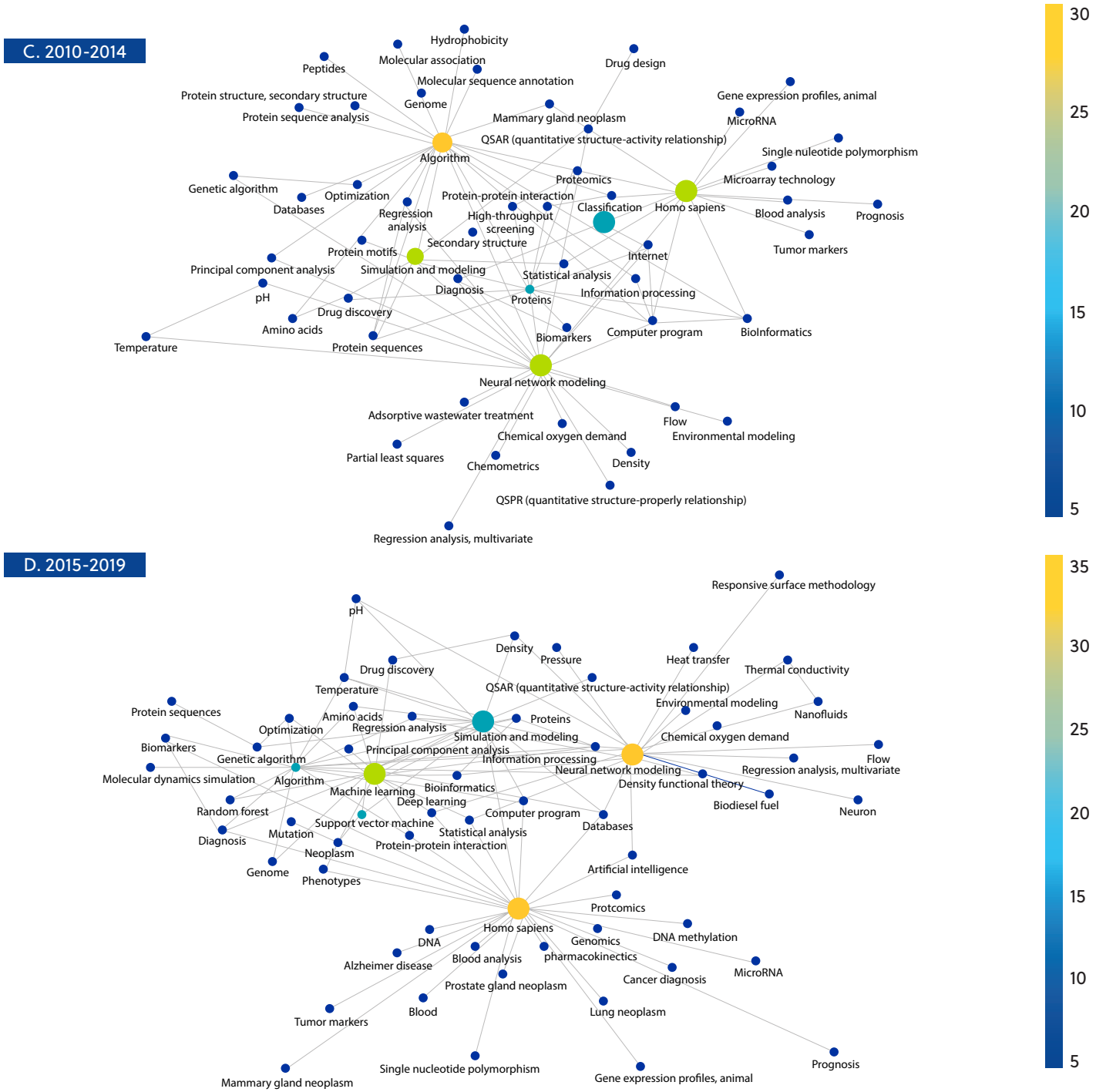


Figure 4: Evolution of co-occurring terms in AI-related chemistry patents (A: 2000-2004, B: 2005-2009, C: 2010-2014, D: 2015-2019)

2010–2014 – During this period, genome related concepts became more prevalent, including 'genome', 'single nucleotide polymorphism', and 'biomedical fields'. AI was used to explore concepts such as 'drug discovery and design', as well as 'blood analysis', 'neoplasm', and 'microRNA'.

2015–2019 – Within the last six years, the use of AI has been focused more on solving problems related to DNA methylation, mutation, nanofluids, heat transfer, and biodiesel fuel, with frequent reference to cancer and Alzheimer's disease. Since the beginning of 2020 and the COVID-19 pandemic, the increase in terms related to drug discovery, diagnosis, and disease tracking illustrate the driving force of AI in these areas. Similar trends were seen in the patent literature, moving from organic compounds in the early 2000s to DNA sequencing and terms related to machine learning closer to the modern day.

Distribution of substances in AI literature

Some of the barriers to AI implementation in chemistry include challenges in substance representation and data availability. A review of the distribution of journal and patent publications by substance class may be indicative of areas in which researchers have, in some instances, been able to overcome such challenges. Therefore, the distribution of AI-related research has been investigated by studying the number of documents involving some of the most frequently occurring substance classes.

As seen in Figure 5 below, publications containing small molecule substances are the highest in number, followed by those containing elements and manual registration (large biomolecules) substances. It's likely that the relative simplicity and ease of modeling contributes heavily to the high volumes of research and invention of AI involving these classes compared to substances in other classes, such as coordination compound and polymer.

Patent literature has also been analyzed (Figure 6), showing a similar trend to the number of journal publications and substance counts seen in Figure 5. However, patent publications in nucleic acid sequences and peptide sequences are highest in number, likely because patents containing these often contain large numbers of sequences per document.

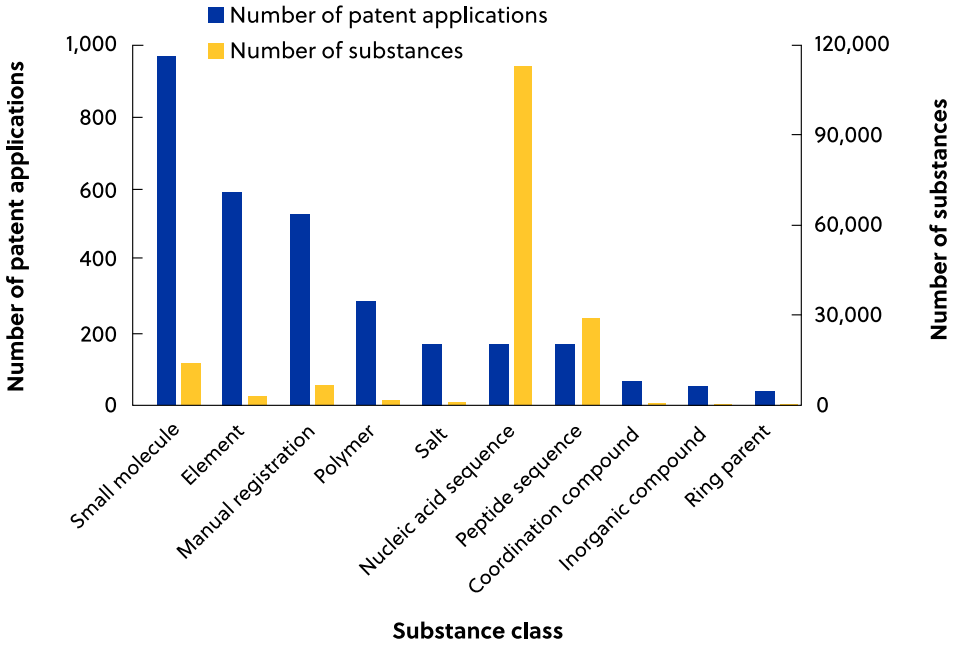


Figure 5. Number of AI-related journal publications and number of substances associated with each class

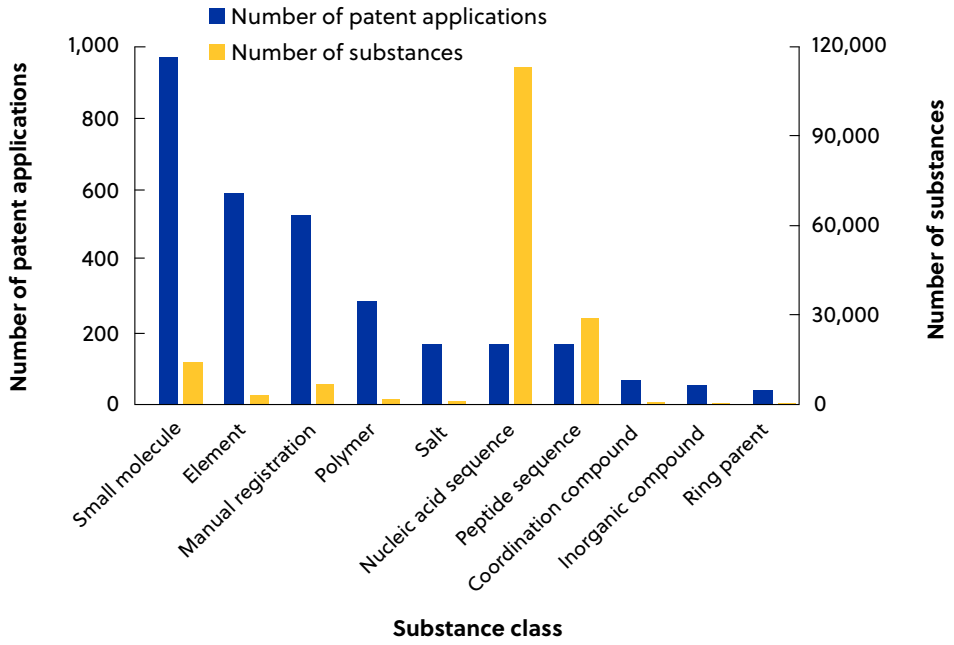


Figure 6. Number of AI-related patent publications and number of substances associated with each substance class

We have also studied the distribution of substances across a variety of role indicators. Role indicators are controlled vocabulary terms that describe the use of a substance within the context of a specific document. A total of 41 different role indicators are used by CAS to identify roles played by substances. For substances found in journal articles, the role indicators show a strong representation from biological and drug information that all relate to the pharma space, such as adverse effects, pharmacologic activity, and therapeutic use.

AI is also increasingly being applied in analytical roles, corresponding to studies in which detection of the substance is important.

When reviewing the data for patent publications, we saw the top performing role indicators are generally similar to those seen for journal publications, but remarkable differences are found in role indicators for diagnostic use, therapeutic use and pharmacologic activity.

In the case of diagnostic use, the high number of substances claimed in patents reflects the increasing importance of AI in medical diagnostics. On the other hand, the large quantity of Small Molecules found in journal articles combined with the prevalence of the therapeutic use and pharmacologic activity indicators suggests a strong research focus on the use of AI in Small-Molecule drug discovery.

The collaborations AI is supporting

Following analysis of nearly 70,000 journal articles, we identified each primary and secondary discipline that contributed to interdisciplinary research. The output of this analysis is a heatmap graph - the more intense the heat, the greater the number of publications related to interdisciplinary collaboration.

As seen in Figure 7 below, the 'hot spots' match those in the disciplines topping those in earlier sections, including collaborations between analytical chemistry and biochemistry, and between materials science and physical chemistry.

In the collaborations between analytical chemistry and biochemistry, AI is being used to improve the analysis of proteins, peptides, lipids, and nucleic acids. And in the collaboration between materials science and physical chemistry, AI is helping to predict functional materials, structure-property relationships and chemical process optimization.While there are significant hot spots on the graph, there are also several dark areas.

Each inter-disciplinary segment may benefit from adopting AI to help solve problems or accelerate workflows. However, many areas may also have challenges or barriers preventing easy adoption.

To better understand this, we explored the potential challenges.

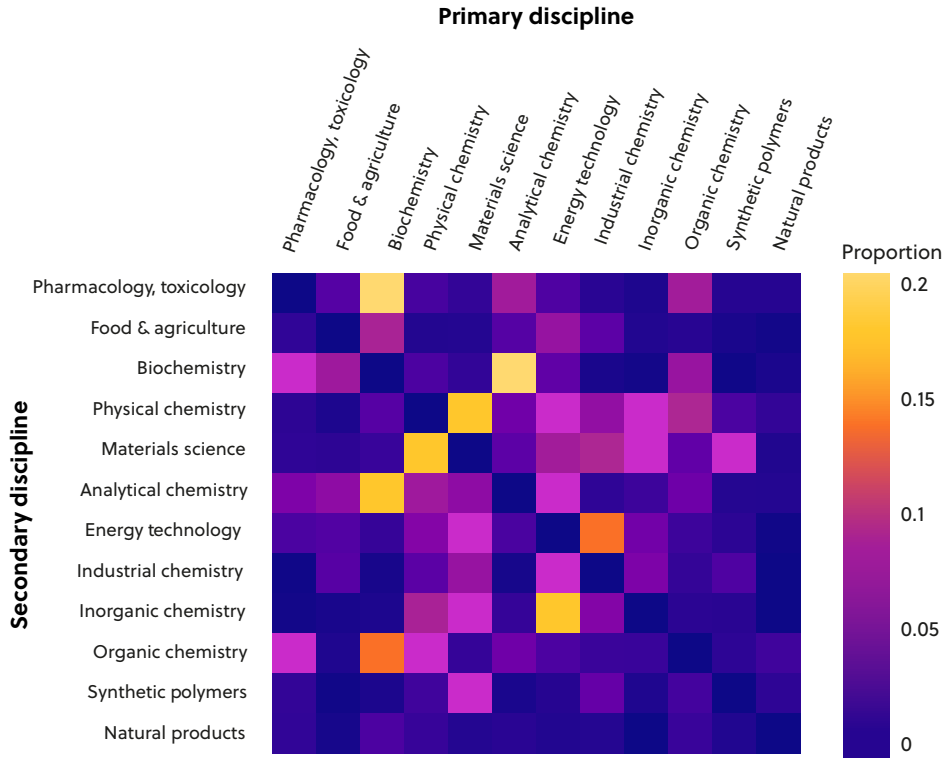


Figure 7. Relative prevalence of interdisciplinary studies published in journal articles (columns denote primary research areas, rows denote secondary research areas, and each square denotes an interdisciplinary pair of primary and secondary research areas)

Challenges to AI adoption

The stark differences in the use of AI between different chemistry disciplines, seen in patent and publications activity, may be indicative of the challenges in adopting AI. Challenges to AI adoption include:

- **Data quality:** Optimal predictions are dependent upon robust, high quality datasets that provide both positive and negative examples for training. Accessing, normalizing, and preparing the data is a significant challenge today for many organizations.
- **Technology:** While improvements are being made in computing power (quantum and cloud-based approaches), there are still perceived limitations from a user perspective. However, advances in software and user interfaces today remove programming requirements to allow more scientists to utilize machine learning in their research.
- **Talent shortages:** Data science has a well-documented talent shortage, and chemists may not understand how approachable AI is today. Increasing collaboration between chemistry and other scientific disciplines may help accelerate the integration of AI.

Although these challenges may account for the lack of growth in AI adoption across the different chemistry disciplines, future improvements in AI itself, increased awareness and acceptance, adaption of AI methods for chemistry research, and lessons from successful applications in AI combined could help increase the uptake of AI in these areas.

The bright future of AI in chemistry

Our review of the landscape shows AIapplication in chemistry-related research has become commonplace in recent years. A broad acceptance and general understanding of AI among the population at large has in turn led to greater acceptance of AI in scientific research – even in fields traditionally steeped in first principles, such as physical chemistry.

The upskilling of researchers in data preparation and the availability of public computational platforms and domain-specific datasets, which have proliferated in recent years, have also contributed greatly. In the coming decade, it is anticipated that these tools will be leveraged in various innovative applications, within an increasingly interdisciplinary research landscape. However, the increasing use of AI in chemistry by no means indicates that it is always successful. It is estimated that 75% of companies are trying to deploy AI in their organization, yet 83%of AI projects are not meeting expectations,and projects that reach deployment are only profitable 60% of the time. To improve the successful adoption of AI, challenges such as lack of/ poor data quality and talent shortages need to be addressed. For the future of AI in chemistry to be as bright as possible, there needs to be increasing collaboration between scientists and technologists alongside computational and technological improvements such as increasing computational power and investments in new tools.

For a detailed review on the growth and distribution of AI application in chemistry, see our publication in the *Journal of Chemical Information and Modelling*.*

*To view publication visit: https://pubs.acs.org/doi/10.1021/acs.jcim.1c00619

CAS is a leader in scientific information solutions, partnering with innovators around the world to accelerate scientific breakthroughs. CAS employs over 1,400 experts who curate, connect, and analyze scientific knowledge to reveal unseen connections. For over 100 years, scientists, patent professionals, and business leaders have relied on CAS solutions and expertise to provide the hindsight, insight, and foresight they need so they can build upon the learnings of the past to discover a better future. CAS is a division of the American Chemical Society.

Connect with us at cas.org

CAS

A division of the
American Chemical Society

cas.org

CASGENENGWHP100231211025