SPOTLIGHT feature

LIMS & Lab Automation

Deciphering the code of chemical and biological processes: The role of AI in laboratory research

Oliver King-Smith, smartR AI

This article explores the transformative role of artificial intelligence (AI) in deciphering the complex molecular and genetic codes that govern biological and chemical processes. From molecular interactions to genetic regulation, AI is enabling researchers to uncover new insights, identify patterns, and optimise laboratory workflows. By applying machine learning (ML) algorithms and leveraging large language models (LLMs), AI is offering a powerful tool to accelerate research in both the biological and chemical sciences. Expert contributors in this field discuss the current advancements and the potential of AI to revolutionise our understanding of chemical and biological systems.

Introduction

Artificial Intelligence (AI) has emerged as one of the most transformative technologies of our time, revolutionising fields from autonomous vehicles to healthcare. At its core, AI refers to computer systems that can perform tasks that typically require human intelligence – including pattern recognition, learning from experience, and making complex decisions. These systems use various approaches, from rule-based algorithms to sophisticated machine learning models that can identify patterns in vast amounts of data.

The evolution of AI has been particularly remarkable in recent years, driven by advances in computing power, the availability of massive datasets, and breakthroughs in machine learning architectures. Deep learning, a subset of machine learning inspired by the human brain's neural networks, has enabled computers to process and analyse information in increasingly sophisticated ways. Large language models (LLMs) like GPT-4 have demonstrated unprecedented capabilities in understanding and generating human-like text, while specialised AI models have achieved superhuman performance in specific tasks such as image recognition and game playing.

In the scientific realm, AI is proving to be an invaluable tool for researchers, offering new ways to analyse complex data, automate routine tasks, and uncover patterns that might be invisible to human observers. Laboratory research, in particular, stands at the frontier of AI application, where the technology is helping scientists decipher the intricate codes of biological and chemical processes. From interpreting genetic sequences to predicting molecular interactions, AI is accelerating the pace of scientific discovery and opening new avenues for investigation.

This article explores how AI is specifically transforming laboratory research in both the biological and chemical sciences. By applying machine learning algorithms and leveraging large language models, researchers are gaining new insights into molecular interactions, genetic regulation, and chemical reactions. These advances are not just incremental improvements to existing methods – they represent a fundamental shift in how scientific research is conducted, promising to accelerate discoveries that could have far-reaching implications for human health, environmental protection, and technological advancement.

In recent years, AI has made significant strides in laboratory research, providing tools that enhance the way scientists interpret complex data. One area where AI is particularly impactful is genomics. Techniques like basecalling—decoding the electrical signals from DNA sequencing into actual DNA sequences—have become industry standards, and AI is central to improving their accuracy. Neural networks, for example, have been used to interpret sequencing data, and new AI-powered techniques combined with Duplex Basecalling have improved accuracy by an order of magnitude across various applications.

AI's Role in Genomics: From Basecalling to Cell Type Annotation

AI's potential in genomics extends beyond basecalling. One of the most exciting recent applications of AI is in the annotation of cell types using large language models (LLMs). The software library GPTCelltype uses GPT-4, a powerful LLM, to automatically annotate cell types from single-cell RNA sequencing data. By utilising marker gene information, GPT-4 generates annotations that align closely with manual annotations, which can significantly reduce the time and expertise required for this task. Evaluated across hundreds of tissue types and cell types, the model has shown impressive accuracy and efficiency. As James Prendergast, Professor of Bioinformatics at the Roslin Institute, explains, "There are an ever-increasing number of papers using LLMs in research. For example, we are using them to understand which bits of the genome are functional and shape important livestock traits."

Genomic Language Models: Unlocking hidden patterns in DNA

The success of LLMs in genomics has inspired the development of 'gLMs' (Genomic Language Models), which are specifically designed to handle the structure of DNA. Given that DNA can be viewed as a long list of letters, much like text, gLMs are trained to predict the next base in a sequence, similar to how LLMs predict the next word in a sentence. These auto-regressive models study billions of base pairs across multiple species to identify hidden patterns that may not be immediately obvious to researchers. Applications of gLMs are already emerging in areas such as fitness prediction, sequence design, and transfer learning. As James Prendergast, Professor of Bioinformatics at the Roslin Institute, explains: "There are an ever-increasing number of papers using LLMs in research. For example, our lab is using them to understand which bits of the genome are functional and that shape important livestock traits."

One of the challenges for gLMs is ironically limited data. Normally you don't think of DNA as 'limited' but the latest LLMs are often getting trained on 10's trillions of tokens, where each token containing 100s of times more information than a base pair. So, to build an equivalent dataset that is used to train the latest state of the art LLMs would require more than Quadrillion 1,000,000,000,000,000,000 base pairs.

Furthermore, gLMs must contend with the limited 'context window' of the data they process - currently, LLMs can handle up to 128,000 tokens, but for some long-range interactions in genomic sequences, this window is too small. Basepairs may interact megabases apart, which won't fit into the context windows of state-of-the-art LLMS.

AI-driven virtual labs: Accelerating scientific discovery

AI is also making its mark in other areas of biological research, such as protein engineering. A recent study created an AI-driven virtual lab to design nanobodies that could bind to the SARS-CoV-2 virus. This 'actor/critic' AI framework involves several agents working together to simulate and predict nanobody efficacy. The Immunology Agent focuses on understanding the immune response and designing the nanobodies, the Computational Biology Agent models interactions between the nanobodies and the virus, and the Machine Learning Agent develops algorithms to predict their effectiveness. This collaborative virtual lab accelerated the discovery of nearly 100 nanobody structures in a fraction of the time it would have taken a human team. The rapid success of such AI-driven models highlights the growing role of machine learning in optimising scientific workflows.

AI's expanding role in chemistry

While AI's application in genomics and biology is making headlines, it is also gaining traction in the field of chemistry. Computational chemistry has traditionally been a time-consuming and expensive endeavour, with significant reliance on human intuition. Despite advances in the last 50 years, computational chemistry has not seen the same dramatic speed improvements as biochemistry or genomics. However, AI holds promise in reshaping this field by providing deeper insights into chemical processes.

Chemistry presents unique challenges for AI applications. Unlike DNA sequences or proteins, small organic molecules cannot easily be described as linear objects.

There are techniques like SMILES strings, but it can be hard to canonically describe molecules using this technique. These molecules lack the clear 'beginning' and 'end' sequences that characterise proteins and genetic codes, and their 3D representations are not standardised. As a result, chemists must develop new AI techniques that can effectively process molecular data. The data required to train AI models is also scarce and fragmented, often hidden in decades of academic journals and custom diagrams that are not easily amenable to analysis.

New approaches to AI in chemistry: Graph neural networks

One promising AI technique for chemistry is graph neural networks (GNNs), which have shown great potential in predicting chemical reactions. GNNs operate by representing molecules as graphs, where atoms are nodes, and bonds are edges. This method allows the network to learn about the relationships between nearby atoms and predict key properties such as reactivity and toxicity. As Dr Emma King-Smith, Chancellor's Fellow at the University of Edinburgh, explains: "Chemistry is a really interesting field because our models for reactivity are pretty rudimentary, but chemists are shockingly good at getting reactions to work. I think what this says is that we have a solid foundation of the chemistry basics, but there's still a lot of intuition involved." Dr King-Smith's research has explored using message-passing neural networks to predict important chemical properties, even with limited data. "Even with less than 1,000 data points, it can predict a vast array of important chemistry, such as acute toxicity, odour profile, and how well a molecule will perform in chemical reactions," she notes.

The future of AI in laboratory research

The potential of AI in both biological and chemical research is vast. In genomics, LLMs and the more specialised gLMs are helping researchers decode genetic information and uncover hidden patterns, which can lead to breakthroughs in areas like personalised medicine and crop engineering. In chemistry, AI models are beginning to assist in understanding molecular reactivity, predicting reaction outcomes, and even designing new compounds with specific properties. As Dr King-Smith puts it: "From one model, we can see such a variety in the prediction tasks, which suggests that AI has the potential to revolutionise many facets of chemical research."

While challenges remain - such as the need for larger datasets, better model interpretability, and overcoming domain-specific complexities - AI is undeniably opening new frontiers in laboratory research. By continuing to refine these models and exploring novel applications, AI has the potential to transform the way we understand and manipulate the natural world. The integration of AI into laboratory practices is not just a trend, but a fundamental shift that will accelerate discoveries and optimise workflows across disciplines.

Conclusion

As AI technologies continue to evolve, their applications in laboratory research are expanding rapidly. From genetic sequence analysis to the modelling of complex chemical reactions, AI is enabling scientists to push the boundaries of what is possible. By automating tedious tasks, uncovering hidden patterns, and providing new ways to simulate and predict outcomes, AI has the potential to revolutionise both the biological and chemical sciences, paving the way for faster, more efficient discoveries that could have a lasting impact on fields ranging from medicine to agriculture and beyond.

About the author

Oliver holds a PhD in Mathematics from UC Berkeley and an executive MBA from Stanford, and is an innovator with expertise in Data Visualization, Statistics, Machine Vision, Robotics, and AI. As a serial entrepreneur, he has founded three companies and contributed to two successful exits. At his latest company, smartR AI, Oliver King-Smith spearheads innovative patent applications harnessing AI for societal impact, including advancements in health tracking, support for vulnerable populations, and resource optimisation. Throughout his career, Oliver has been dedicated to developing cutting-edge technology to address challenges, and today smartR AI is committed to providing safe AI programs within your own secure and private ecosystems.

LinkedIn profile: https://www.linkedin.com/in/oliverkingsmith/

Email: oliverks@smartr.ai