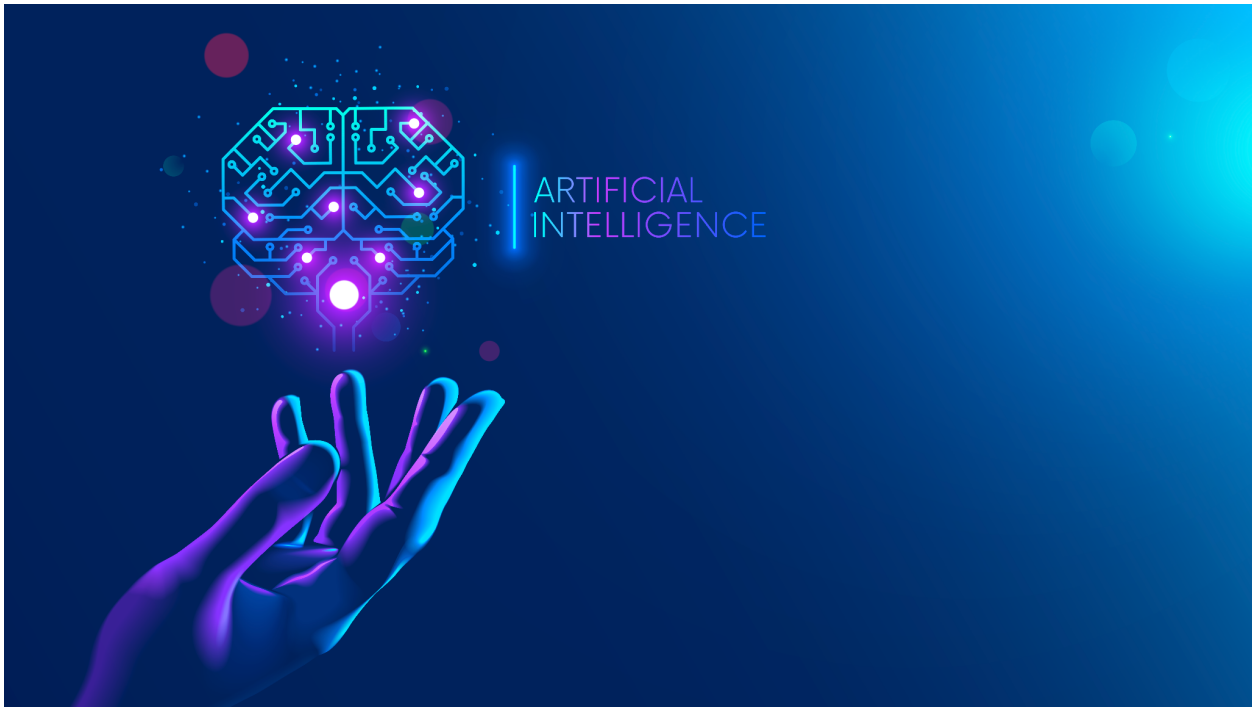
How to maximise the impact of AI on your research



Artificial Intelligence (AI) or Machine Learning (ML) is no longer optional if you want to be competitive in life sciences research and development. Whether you are identifying potential drug targets or trying to increase crop yields, machine learning algorithms can help. In this article, we will take a close look and understand the latest breakthroughs enabled by AI:

- Why AI/ML will accelerate your research
- How others are doing it
- How you should do it

Why do AI/ML

The amount of biological data has increased exponentially in recent years. This is largely due to advancement in high-throughput sequencing techniques, growing adoption of single-cell technologies, and biobank-scale data generation initiatives.

Data has also become more complex. It can suffer from a low signal-to-noise ratio, rare and barely identifiable cell types in complex heterogeneous systems, and a lack of structure. Therefore the application of machine learning in life sciences becomes more critical. ML can deal with messy data and gain insights that cannot be detected using traditional approaches.

Doing science the traditional way can also cost a lot of extra money. Take drug discovery for example, only 1 in 1,000 molecules identified as potential treatments make it to human clinical trials - trials that then take an average of five years and eliminate another 90% of chemical candidates. The entire process of developing a drug from preclinical research to marketing can take 12 to 18 years and often costs between $2.0 and $3.0 billion, with only about 10% of candidates successfully completing clinical trials and receiving approval. [1] Artificial Intelligence can accelerate this research and make discovery time cut from the standard of four to five years to between eight to fourteen months.[2] For example, an A2 receptor antagonist designed to help T cells fight solid tumours was found in just 8 months by harnessing an AI design platform [3].

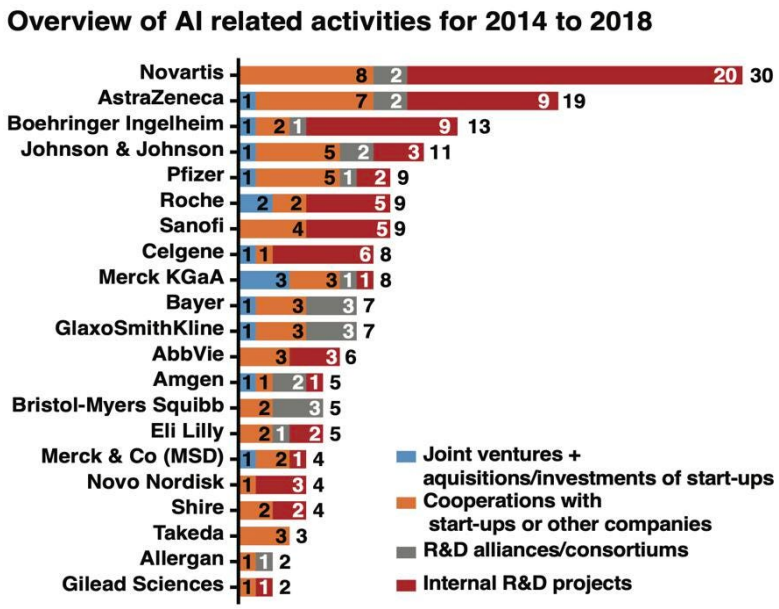
AI is more accessible today than ever before. Very little specialised training is required to perform machine learning. This is possible because there are a large number of different ready-to-use frameworks and libraries for most programming languages. For example, if you are familiar with Python, you can use SciKit Learn, PyTorch and SciPy to implement machine learning with just a few lines of code [4]. Of course, some complex algorithms still require mathematical training to develop, but the majority of tasks can be done using a combination of existing solutions.

Successful examples

“By the end of this decade, I expect all drugs entering the clinic to be designed with AI.”

Andrew Hopkins,
founder of pharmaceutical startup Exscientia [2]

Evidence has shown that pharma is increasing its activity to adopt AI over the past few years with this trend expected to continue as the benefits of the technology are realised [5]. Merck, Pfizer and Teva for example have teamed up with a startup that invests in artificial intelligence and computational biology. [6]



Overview of AI Related Activities 2014-2018 by Big Pharma Player Modified from Schuhmacher et al, The upside of being a digital pharma player (2020), Drug Discovery Today [1]

Verge Genomics develops drugs by automating their discovery process. They use automated data collection and analysis to develop treatments for diseases as serious as ALS and Alzheimer's disease.

Bayer and Merck & Co have received Breakthrough Device Designation from the FDA for AI software aimed to support clinical decision-making in chronic thromboembolic pulmonary hypertension.

Another example is Novartis, which is currently using machine learning to classify digital images of cells, each treated with different experimental substances. The machine learning algorithms collect and group compounds that have similar effects before passing the clean data on to researchers who can decide how to use these findings in their work.

There are many more examples of machine learning methods used by large pharmaceutical companies. So it's safe to say that AI/ML is a big trend in drug development these days.

Some tips for successful AI/ML applications

1. Make sure your data is of good quality.

Garbage in, garbage out. And it’s well known that data collection and curation can be the most time-consuming part of data science. High-quality data brings business benefits in the form of more informed and faster decisions, higher revenues and lower costs and makes scaling a lot easier. It is especially important for machine learning, because the effectiveness of prediction models critically depends on the quality of the data.

“Data quality problems can cause projects to fail, result in lost revenues and diminished customer relationships, and customer turnover” [7]

Clean up your data. Eliminate any missing values and harmonise the fields. However, if you have more than a few records, this task can take a lot of time. So make sure your data management system makes this mundane task easier. The Genestack platform, for example, allows you to curate your studies with a few simple clicks. It also validates your data against scientific ontologies and vocabularies, points out potential problems and puts everything into a consistent format. The latter is especially important if you want to use public data.

2. Build a sustainable system of data storage and management.

Creating an SPoT - Single Point of Truth - is an important task when it comes to consistent data flow. If you have more than one lab, it's only a matter of time before you get to the point where your data is scattered everywhere. So make sure you have a consistent data management system that helps you connect all the data points from different sources and send them to downstream systems for analysis. At Genestack, we believe that a good API is a must. It helps you connect your data sources and work with them much more efficiently. That's why we have

made sure that our platform can become such an SPoT for your data and connect to all the existing services you already use.

3. Make use of public data.

Modern biology has become a data-intensive discipline and the amount of omics data in the public domain is increasing every year, because it is considered to enable reproducibility, which is a good practice for any scientific field.

For a full list of scientific databases, see Wikipedia page [8]. Some of the best known include ArrayExpress [9], TCGA [10] and GEO [11]. You can also find a specific database for a model organism you are working with: WormBase for C. elegans [12], TAIR [13] for Arabidopsis and MGI [14] for mouse. Most of these databases focus on genomics data, but you can also explore proteins in the Protein Data Bank [15] or metabolic pathways in MetaCyc [16].

However, the numerous databases with different interfaces can be overwhelming when it comes to discovering, integrating, and reusing the right datasets. So, make sure you download them first and integrate them into your own data management system. In earlier paragraphs we already talked about the importance of SPoT and here it becomes even more important. To ensure high-quality public data, you have to harmonise and/or even reprocess them. The Genestack platform can help you a lot in this regard. It can become the perfect single point of truth to provide high-quality dataset for your AI/ML application

The increasing amount of data in the life sciences today makes AI/ML extremely useful. You too can benefit from it if you have high quality data and build a sustainable system to store it in. So let Genestack take care of the hardest part of managing your datasets so you can focus on the most exciting part of the journey. Contact us to learn how our platform can accelerate the adoption of AI in your business and unleash the power of your data.

References

1. Big Pharma is forging an increased number of partnerships with artificial intelligence vendors for drug discovery services
2. All drugs will be designed by computers by 2030
3. Tapping into the drug discovery potential of AI
4. Review and comparative analysis of machine learning libraries for machine learning
5. Data hunters will be Big Pharma's next prey
6. Four big pharma companies team up in AI initiative focused on drug R&D
7. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations
8. List of biological databases
9. ArrayExpress
10. TCGA
11. GEO
12. WormBase
13. TAIR
14. MGI
15. Protein Data Bank
16. MetaCyc