IN VIVO
CITELINE COMMERCIAL

11 Oct 2022 | **Analysis**

AI In Biologics Discovery: An Emerging Frontier

by Madura Jayatunga, Lotte Bruens, Ludwig Ruder, Ulrik Schulze, Christoph Meier

AI is beginning to transform biologics discovery. The power of algorithms used in biologics discovery has increased over the last decade, and today between 50 and 60 AI-enabled biologics are in different stages of discovery, preclinical and clinical development. We expect the number of AI-enabled biologics to continue to grow rapidly, driven by advances in AI technology and algorithms, growing computing power, increasing availability of data, and evolving discovery workflows. We show that the volume of data used for training algorithms in biologics discovery is increasing exponentially over time, a trend reminiscent of Moore's Law in computer technology.

Artificial intelligence (AI) and machine learning are revolutionizing drug discovery. Over the last five to ten years the main focus has been on small molecules, where we are seeing strong growth in the number of compounds discovered using AI *technologies*.

In parallel, but less well-known, AI is beginning to transform other areas of drug discovery, most importantly biologics (protein and peptide-based medicines). Over the past decade many AI algorithms and tools enabling the discovery and optimization of biologics have been developed. Perhaps the most well-know of these is AlphaFold, an algorithm which predicts protein structures in a highly accurate way, and which facilitates one of the key steps in biologics discovery. (Also see "*AI Accelerates In Drug Discovery*" - In Vivo, 4 Aug, 2021.)

In parallel with technological progress, several biologics companies have been founded for which AI is central to their discovery process (the "AI-natives"). And there is now a growing pipeline of AI-enabled biologics at different stages of discovery, pre-clinical development, and even clinical development.

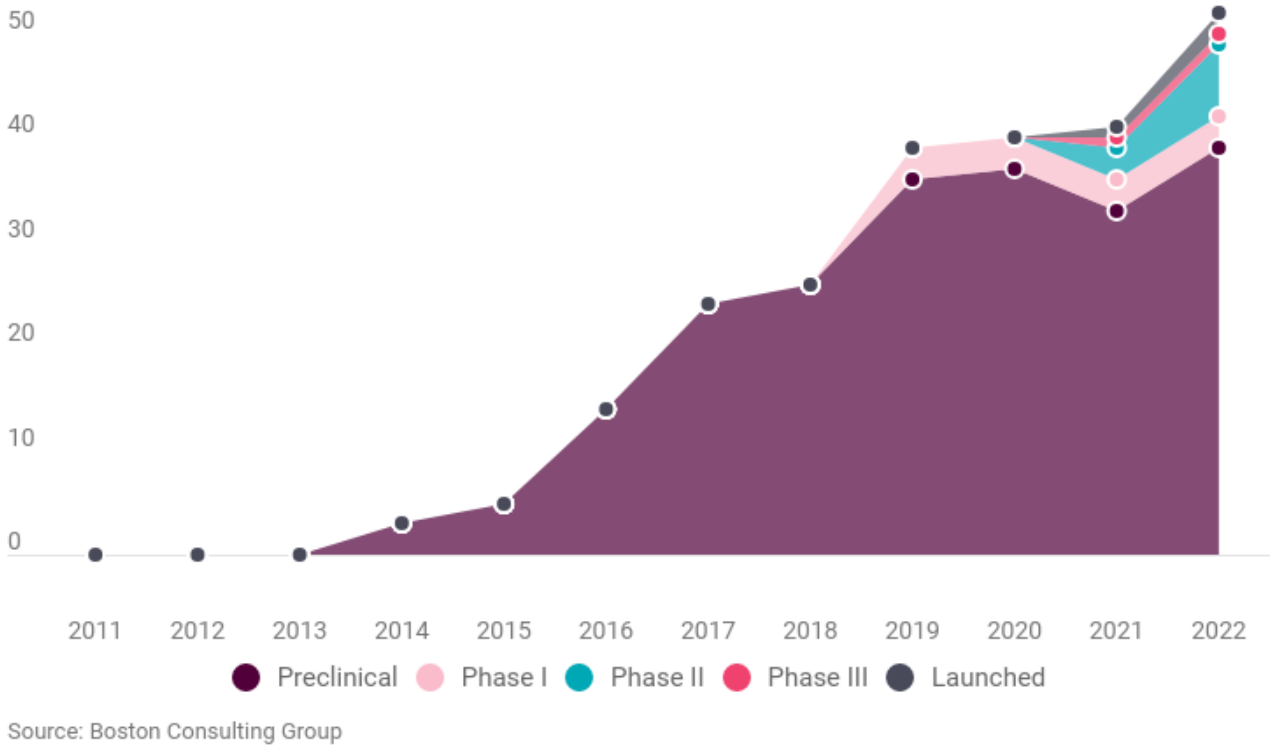
The wider industry is beginning to embrace AI in biologics: Several large pharma companies have entered partnerships with AI-native biologics companies, have invested in their own internal capabilities, and there is substantial venture capital and other investment in the space. For example, in 2021 there were 30 deals between pharma companies and AI-native biologics companies worth a total of $1.2bn, while in 2022 year-to-date there were nine deals worth $4.5bn, indicating a strong growth in deal size, according to *BioMedTracker*.

Similar to other areas of drug discovery, AI in biologics discovery offers the promise of (i) reducing timelines and costs; (ii) improving the quality and novelty of molecules discovered; and (iii) increasing the probability of success of R&D programs. Here we look at AI-enabled biologics discovery along these dimensions, aiming to identify proof points for the value drivers based on publicly available information. Specifically, we explore the applications and use cases of AI in biologics discovery, and we assess the underlying drivers and enablers.

Impact Of AI On Biologics Discovery So Far

We identified 14 AI-native biologics companies which are working on antibodies and other proteins and peptides, and for which AI is central to their discovery process. For these companies we were able to reconstruct their pipeline between 2010 and 2022 using public databases, see Figure 1 below. During this time, AI-native biologics companies experienced rapid pipeline growth, with a year-on-year growth rate of approximately 40%. As of 2022, the combined pipeline of these companies contains at least 52 R&D programs and assets. These findings suggest that AI in biologics discovery represents an emerging frontier in R&D.

Figure 1: Number Of AI-Enabled Biologics Projects in Clinical, Preclinical and Discovery



Source: Boston Consulting Group

Figure 1: Number of R&D programs and assets over time, showing an emerging frontier of AI-enabled biologic molecules (antibodies & other proteins).
Companies with in-house AI-enabled biologics discovery pipelines were included in this analysis. These comprise AbCellera Biologics, Adagene, Adimab, Evaxion Biotech, Hummingbird Bioscience, and RubrYc Therapeutics.
Companies without in-house discovery pipelines were not included in the analysis. These comprise AbSci, Antiverse, BigHat Biosciences, Biomatter Designs, Generate Biomedicines, LabGenius, MAbSilico, and OmniAb Technologies.

Pipeline information was reconstructed using Citeline Pharmaprojects. Additional projects mentioned only on company websites were included manually.

For comparison, across the top 20 pharmaceutical companies, there are approximately 580 ongoing biologics R&D programs and assets. AI-native biologics companies today thus have a

combined pipeline equivalent to just under 10% of 'big pharma', indicating that AI technology in biologics is still in its early days. However, if the rapid growth of AI-enabled biologics continues, the technology is likely to play an increasingly prominent role in the biologics space in the coming years.

Of the AI-enabled biologics R&D programs and assets we identified, ca. 90% are monoclonal antibodies. This is unsurprising given that monoclonal antibodies make up the majority (>70%) of all biologics R&D programs across the industry.

So far, 13 AI-derived biologics have reached the clinical stage, with 11 currently in clinical trials and two launched. The COVID pandemic has had a major impact, with four of the most mature assets targeting SARS-CoV-2. Other major therapy areas include oncology (eight AI-derived biologics in clinical trials) and neurology (1 AI-derived biologic in the clinic). It remains to be seen how many more AI-derived biologics reach the clinical trial stage, and how successful they will be in clinical trials.

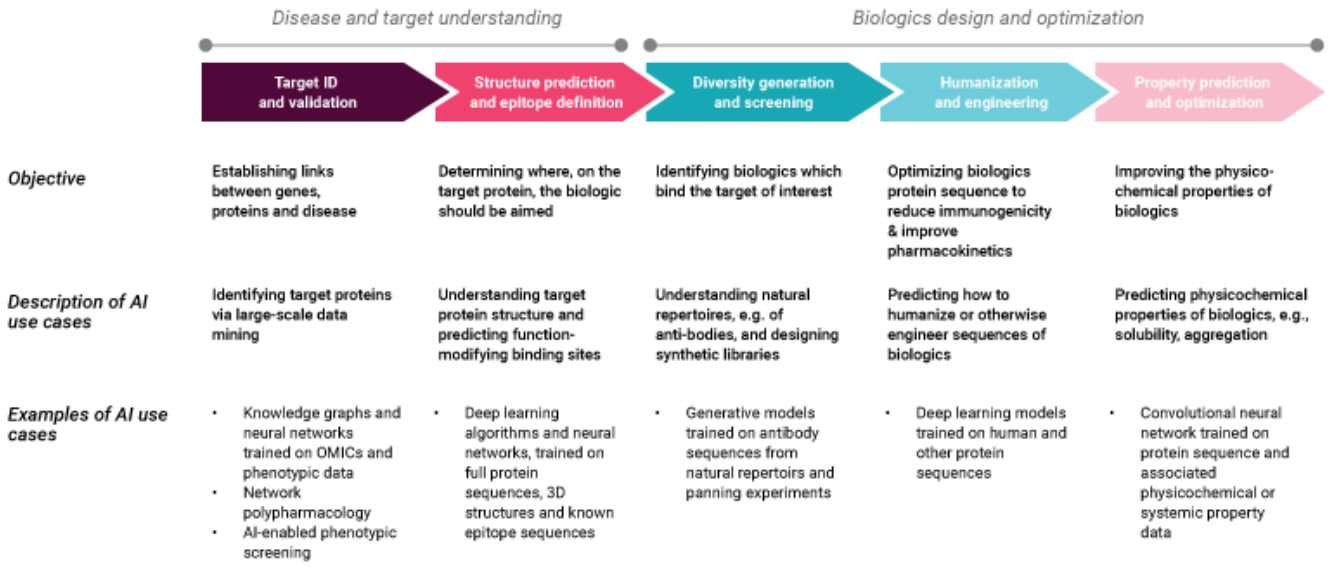
Technology Drivers And Enablers Of AI In Antibody Discovery

For small molecule drugs (size typically between 200 and 700 Da) AI can be used to design and optimize the molecular structure directly, e.g. using generative neural networks.

In contrast, for biologics (size between 5,000 and 200,000 Da) limitations of computational capacity mean that direct, atom-by-atom design using AI is typically not feasible. Instead, our analysis suggests, AI is used differently in the context of biologics. Across the discovery value chain, we see five main use cases for AI:

1. Target discovery and validation
2. Target structure prediction and epitope definition
3. Diversity generation and screening
4. Humanization and engineering of biologics
5. Property prediction and optimization

Figure 2: Typical Value Chain And Overview Of AI Use Cases In Biologics Discovery



One of the AI use cases, *target discovery and validation*, is modality-agnostic and can in principle be used as part of any drug discovery program. However in the case of biologics, additional constraints typically apply – namely that the targets identified need to be extracellular or otherwise accessible to biologics. In recent years we have seen several success stories of novel targets identified using AI and molecules subsequently entering the clinic, such as the *AstraZeneca PLC* and *BenevolentAI Limited* partnership. (Also see "*AstraZeneca And BenevolentAI Advance AI Partnership With Second Novel Target*" - Scrip, 16 Dec, 2021.)

In contrast, all the other AI use cases are highly specific to biologics and address distinct aspects of the discovery workflow.

In *structure prediction and epitope definition*, AI can help to identify optimal epitopes on target proteins. Historically, epitopes were often identified using experimental approaches. For example, for antibodies *in vivo* or *in vitro* methods are used which often rely on naturally occurring immunogenic hotspots on target proteins. With recent AI-based methods, such as the AlphaFold algorithm, it is now beginning to be possible to understand target protein structures better and to select epitopes more rationally. This improves the likelihood of finding truly function-modifying biologics, with the first examples targeting such AI-selected epitopes now in *clinical trials*.

In *diversity generation and screening*, we observe that AI is rarely used for primary screening of

biologics libraries. This is due to the enormous number of sequences that would need to be screened (hundreds of millions of variants, or more) which would be computationally very expensive and time-consuming. Instead we see AI being used in the optimization of naturally occurring repertoires of biologics and in the design of novel libraries, especially for *antibodies*.

In *humanization and engineering*, AI can be used to fine-tune pharmacokinetics and other properties in a rapid and efficient manner, which avoids costly and time-consuming experimental methods. For example, there are now tools which can distinguish human from animal sequences and propose mutations to improve '*human-ness*'.

In *property prediction and optimization*, AI can optimize the physicochemical properties of biologics, including solubility and aggregation. Historically this required time-intensive experimental methods to reach the desired properties. With AI-based methods, it is now beginning to be possible to accelerate this *process*.

While there has been great progress with individual use cases, we believe the full potential of AI in biologics will materialize only when these tools and approaches are used together, in an end-to-end manner. For example, the combination of AI-enabled target discovery, epitope identification and molecule design could give rise to uniquely novel molecules which are inaccessible to classical biologics discovery approaches. Likewise, AI-enabled engineering, humanization and property optimization of biologics currently consist of point solutions which could be combined into a multi-parameter developability optimization.

It remains to be seen if and when this will happen, what the overall impact might be, and which companies get there first.

Evolution Of AI In Biologics Discovery

To understand where we are seeing the greatest progress of AI across biologics discovery, we analyzed the recent evolution of the technology.

A first observation is that – for many use cases – the predictive power of AI approaches has been increasing. Nowhere is this more obvious than in protein structure prediction: 10 years ago, leading-edge structure prediction methods achieved a GDT score (global distance test, a measure of accuracy) of 40%-60%. Current state-of-the-art approaches, such as AlphaFold, now achieve a GDT score of over 90%. Improvements can also be seen in other areas of biologics discovery, including humanization optimization where accuracy went from 75% to 95% in the last 15 *years*. And epitope prediction which went from 50-60% to 70% in the last five *years*.

A second observation is the technology behind AI-enabled biologics discovery is becoming increasingly sophisticated. For example, the application of leading-edge deep learning neural networks and other AI algorithms has helped to increase predictive power of AI in biologics

discovery. Growing computing power and increasing availability of data also play a key *role*. Taken together, these technological advances are beginning to enable a new workflow for discovering biologics, some of which is already visible in AI-native biologics discovery companies today.

While the sophistication of algorithms is not easy to assess objectively, other factors, such as the availability of data, can be assessed more readily. Across the five use cases described in the previous section, see Figure 2, we find that the volume of data used for training the algorithms is increasing rapidly. For example, before 2015 the median volume of data used for training algorithms in biologics discovery was ca. 30 megabytes (mb). For algorithms published between 2015 and 2020, the median data volume was ca. 120mb. And for algorithms since 2020, the median volume of data is over 1,000mb.

Analyzing this question in more detail, we find that for many use cases the volume of data used for training the algorithms is increasing exponentially over time [Figure 3]. This is particularly true for protein structure prediction and humanization, where the volume of data appears to double every 14-18 months. Epitope prediction is progressing more gradually, with data volumes doubling approximately every 50-60 months. We see intriguing parallels to the well-known Moore's Law, in computer technology, whereby the CPU capacity on microchips doubles approximately every 18 months.

Figure 3: Use Case-Level Analysis Of The Volume Of Data Used In Training Algorithms For Biologics Discovery – A New Moore's Law?

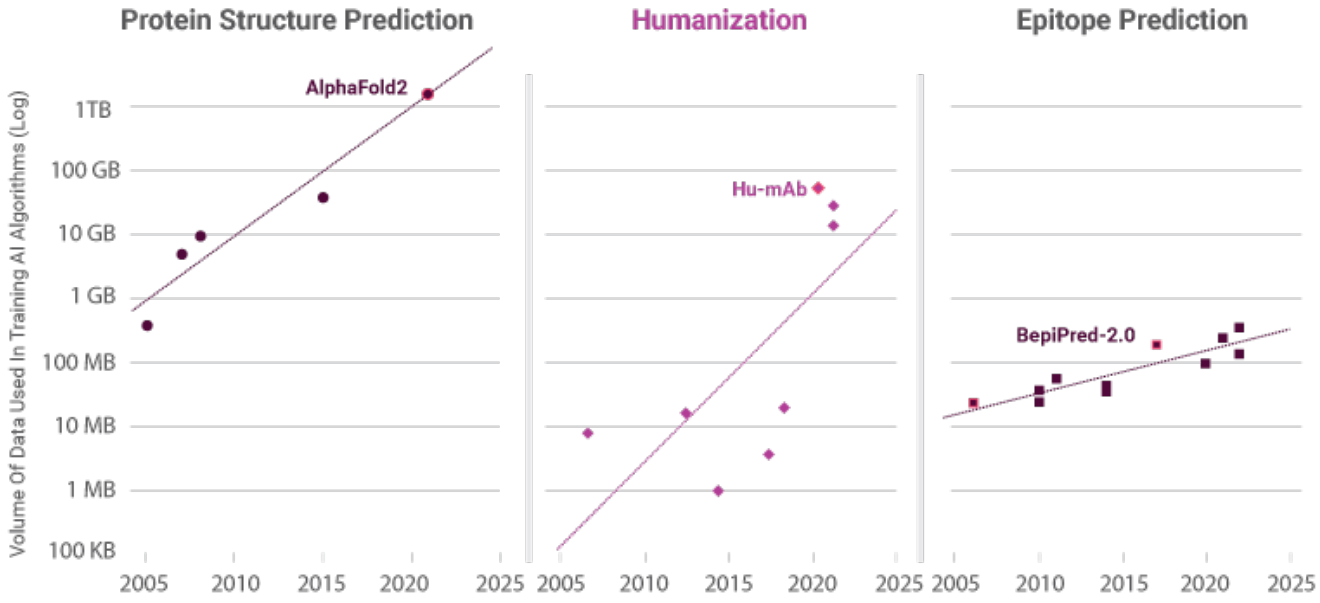


Figure 3: Use case-level analysis of the volume of data used in training algorithms for biologics discovery – a new Moore's Law?
For some use cases, the volume of data is increasing exponentially over time, enabling novel solutions (KB = kilobyte; MB= megabyte; GB = gigabyte; TB = terabyte).
For each algorithm, the volume of data was extracted from the materials and methods section of the publication. Where necessary, data volume was estimated using standardized values (protein structure = 300 kilobytes of data on average; protein sequence = 0.8 kilobyte; single gene sequence = 1 kilobyte)

However, for other use cases, the volume of data utilized is lower and we see limited growth in data volumes. For example in aggregation and solubility prediction. Interestingly, we observe that for these use cases, the predictive power also appears to progress slowly.

This suggests that, in addition to progress with algorithms and growth in computing power, more data is likely required in order to advance AI-enabled biologics discovery. Companies which have access to these larger volumes of data, either internally or through partnership, may have a strategic advantage. Already we are seeing examples of such strategic advantage. For instance, out of the 14 AI-native biologics companies described above [Figure 1] those with deep datasets and the ability to generate more data have amongst the broadest range of AI applications and use cases in the biologics space.

Going forward, we expect that biologics R&D organizations will increasingly recognize the value of their data and will push toward more end-to-end use of AI technology in biologics discovery. Given the challenges of developing AI tool and solutions, we also expect a greater need to partner with AI companies, consortia, and each other to tackle these foundational problems.

Availability of large volumes of data is clearly not the only driver of AI in biologics discovery – other advances, such as technology and algorithms, computing power, and evolving discovery workflows are likely to play a similarly important role. But our analysis suggests that availability of ever larger datasets is one of the critical enablers of AI in biologics discovery.

Conclusions And Outlook

Biologics discovery is a complex, multi-step process. Success requires a deep understanding of diseases and targets, as well as the ability to identify and optimize protein molecules. AI, with its powerful tools of solving complex problems, has the potential to dramatically improve the biologics discovery process. Our analysis indicates early signs of impact. Specifically, we have shown that the power of AI approaches to solve important problems in biologics discovery is growing enormously. And we have seen the sophistication of AI, both in terms of predictive power and in terms data volumes used for training algorithms, is increasing substantially. As a result, the industry pipeline of AI-derived biologics is growing strongly and is likely to continue to do so.

Several uncertainties remain, mainly the question how successful AI-derived biologics will be in the clinic. Also the question of how much progress AI will continue to make, especially for challenging use cases, such as immunogenicity predictions. Our expectation is that those organizations which have the largest and most comprehensive datasets are likely to win in this space.

Taken together, we see strong potential for AI in biologics discovery and, if the current momentum continues, AI is likely to become a game changer in this space.

All authors are based at European and US offices of the Boston Consulting Group.