LUMA GROUP

From Hippocrates to Hyperspeed: How AI Will Catalyze Advancements in Life Science

Scientific Addendum

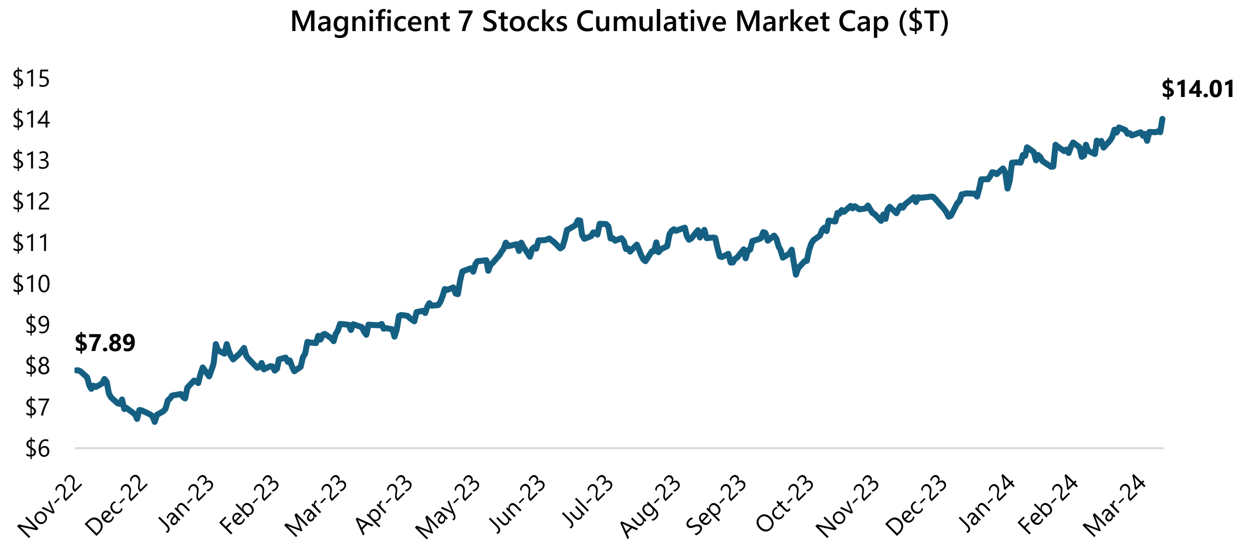
The AI revolution's impact on society, particularly healthcare, is set to surpass the transformative effects of the Industrial and Information Revolutions. Real-world evidence of this can be seen in the application of AI-driven diagnostics and therapeutics, which have significantly enhanced the accuracy and speed of disease detection and rate of targeting and curing disease. The forthcoming role of AI in healthcare is ambitious yet achievable but requires a level of collaboration, innovation, and dedication to a brighter and healthier future.

April 2024

The journey from a research lab to a patient for a single drug spans an average of 12 years. The accompanying cost is by some estimates over $2 billion. Even still, the odds of success remain minuscule, standing at one in 5,000.[1,2] The biotechnology industry, in its current state, appears inefficient; however, the integration of artificial intelligence ("AI") with exemplary scientific practices presents a promising avenue for significant improvement.

The advent of AI models and applications has sparked a wave of breakthroughs, culminating in the creation of $6 trillion in market capitalization for the "Magnificent 7" stocks alone since the launch of ChatGPT (Figure 1).

Figure 1: The "Magnificent 7" stocks – Microsoft, Amazon, Meta, Apple, Google parent Alphabet, Nvidia and Tesla – drastically outperformed the major U.S. indexes in 2023 as the market recognized them as AI winners. Outperformance YTD has continued with more innovation in the AI space expected in 2024.



Source: Public market data as of April 10, 2024.

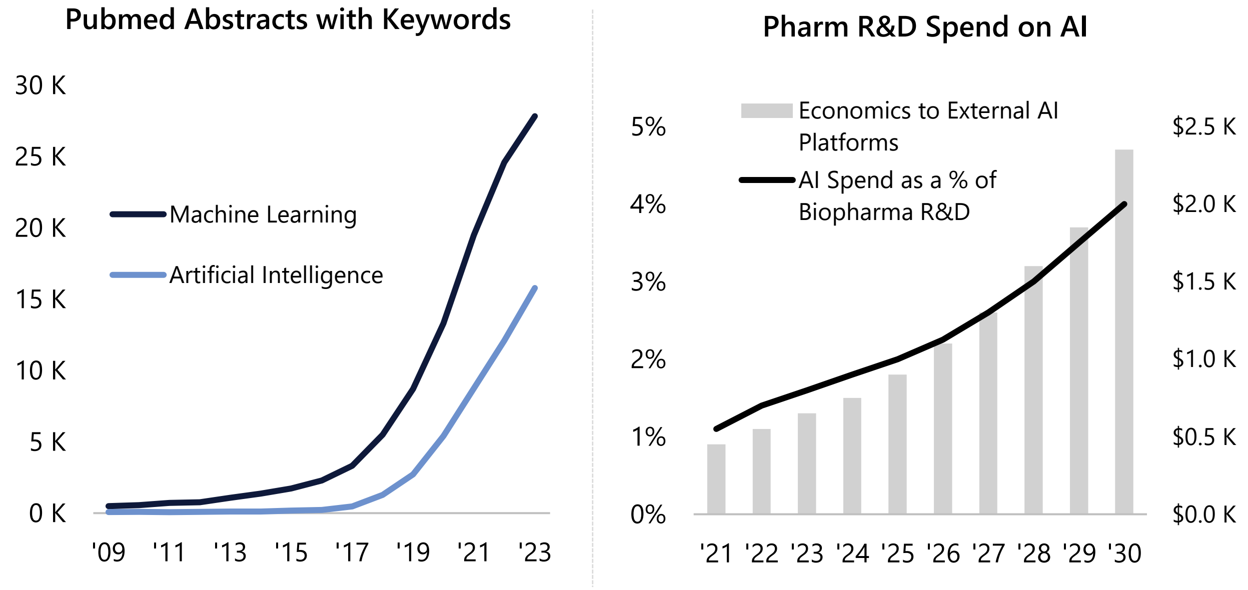
Unlike tech stocks, AI's impact is not yet baked into healthcare valuations. AI has already begun generating new ideas, discoveries and potential opportunities to accelerate the healthcare sector to become more efficient. We can observe this trend by looking at the proliferation of AI mentions in scientific publications and healthcare AI expenditures (Figure 2). The healthcare industry spent approximately $13 billion on AI hardware and software in 2023, with projections to rise to $47 billion by 2028.[3] CB Insights reports that between 2019 and 2022, investors injected $31.5 billion in equity funding into AI healthcare ventures.

[1] https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/transforming-healthcare.html

[2] https://www.biopharmadive.com/news/new-drug-cost-research-development-market-jama-study/573381/

[3] https://www.economist.com/technology-quarterly/2024/03/27/ais-will-make-health-care-safer-and-better

Figure 2: The growing focus in AI/ML can be seen in the exponential growth in its presence in academic publications, which is expected to feed into continued industry investment in AI-enabled products in both R&D budgets and M&A activity.



Source: Pubmed, Biospace.

Broadly on AI: what is it, and why should I care?

AI, in its essence, encompasses the development of computer systems that can perform tasks that typically require human intelligence. These tasks include learning, decision-making, language understanding, and visual perception. The dawn of AI marks a significant leap forward in our ability to process information, automate complex processes and solve intricate problems across almost all sectors and industries. The potential societal impact of AI cannot be overstated. These technologies promise to revolutionize every sector of the economy, from manufacturing to healthcare, education to finance, by enhancing efficiency, unlocking new insights and opening avenues for innovation that were previously unimaginable.

At the core of recent advancements in AI are neural networks inspired by the human brain's method of processing data inputs and generating predictive outputs. These networks, essential to AI's evolution, frequently derive their architecture from biological structures, particularly the brain's complex neural pathways. This biomimetic approach is far from being merely symbolic; it's fundamentally practical, equipping AI systems to analyze and understand complex information similarly to how biological entities do. Moreover, the deployment of AI in biological contexts provides crucial insights that feedback into AI development, refining algorithms to better emulate natural neural processes. This continuous feedback loop not only drives technological advancements in healthcare but also enhances our comprehension of the biological inspirations behind these AI innovations.

The intersection of biology and technology

In biotechnology, the application of AI spans research and development in discovering and developing new drugs, diagnostics, manufacturing and personalized medicine approaches. It serves as a catalyst for innovation and efficiency, which is a cornerstone of AI's promise in biotech. The conventional drug development journey proves laborious and resource-heavy. As highlighted in the introduction, it represents a protracted and costly journey towards eventual commercialization. AI has the real potential to significantly reduce cost and time scales, as highlighted in the following examples (see box 1):

Drug Discovery and Development

AI-driven algorithms can analyze vast amounts of biological and chemical data to identify potential drug candidates more efficiently than traditional methods. According to a report by McKinsey, AI can reduce drug discovery times by up to 50% and decrease research and development costs by 60%.[4]

Precision Medicine

AI is enhancing the effectiveness of precision medicine by analyzing genomic data to tailor treatments to individual patients. For instance, a study published in Nature Medicine found that machine learning algorithms could predict patient responses to cancer treatments with up to 85% accuracy.[5]

Diagnostic Imaging

AI algorithms are improving the accuracy and speed of medical image analysis, leading to better diagnoses. For example, a study published in Nature reported that AI outperformed human radiologists in detecting breast cancer from mammograms, reducing false positives by 5.7% and false negatives by 9.4%.[6]

Diagnostic Imaging

AI is optimizing bioprocesses in biomanufacturing, leading to higher yields and reduced production costs. For instance, researchers at GlaxoSmithKline used AI algorithms to optimize cell culture conditions, resulting in a 40% increase in cell productivity and a 35% reduction in process development time.[7]

[4] https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/cloud-based-artificial-intelligence.html

[5] https://www.nature.com/articles/nm.4463

[6] https://www.nature.com/articles/s41467-020-19334-3

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8043889/

Generative AI models are beginning to impact and revolutionize drug discovery and development. These generative models are sophisticated algorithms designed to analyze extensive, intricate datasets and generate new content that simulates the original data's structure and features. This capability enables the algorithm to "generate" and fill in missing data with high potential accuracy, bypassing the time and resource-intensive empirical methods. This technology is incredibly powerful in drug discovery and development as understanding protein structure is fundamental to the field. Our knowledge has traditionally been constrained by empirical methods such as x-ray crystallography and cryo-electron microscopy, which, while accurate, are slow and have deciphered the structures for only about 30% of the human proteome.[8] Meanwhile, generative AI like DeepMind's AlphaFold has predicted structures for an estimated 98.5% of human proteins, representing a massive leap forward. In the 2020 competition, AlphaFold 2 achieved over 92.4% accuracy in positioning atoms, matching experimental methods like X-ray crystallography.[9] These AI-predicted structures, while not as rigorously procured as with empirically validated structures through experimentation, have been highly accurate and swiftly produced, proving useful in the development of therapeutics, including those against diseases like COVID-19. Generative AI's prediction of the SARS-CoV-2 spike protein structure has been crucial for accelerating vaccine and antiviral drug development. These advances showcase how generative AI models are becoming instrumental tools in the field of biotechnology.

Conversely, non-generative AI approaches provide invaluable statistical insights that have far-reaching implications in healthcare. Unlike generative models that fill in gaps, these models process large and complex datasets to reveal trends and predictors that would not have been discovered without AI approaches. For example, AI has been instrumental in oncology where it aids in predicting patient responses to various cancer treatments. By analyzing medical records, genetic information and treatment outcomes, AI can more efficiently identify which patients are more likely to benefit from specific therapies. Another compelling application is in diagnostics, where non-generative AI models assist in analyzing complex diagnostic data. These models have improved the detection or progression of diseases for early intervention and likely better health outcomes (see Curve case study for Luma example, Figure 6). Both generative and non-generative AI models are making strides in advancing human health. By integrating diverse datasets including OMICs (genomics, proteomics, metabolomics and others) data, clinical data and electronic health records, AI is facilitating advancements in healthcare, enhancing patient outcomes and streamlining drug development processes.

Drug discovery and development: generative or not generative, one size does not fit all

Applications of AI in drug discovery and development are not new to the field, but new generative approaches have stirred the imagination of researchers and given them a new toolset to develop new therapies. Even with these advancements, there are areas within drug development that are

[8] https://onlinelibrary.wiley.com/doi/full/10.1002/pro.4038

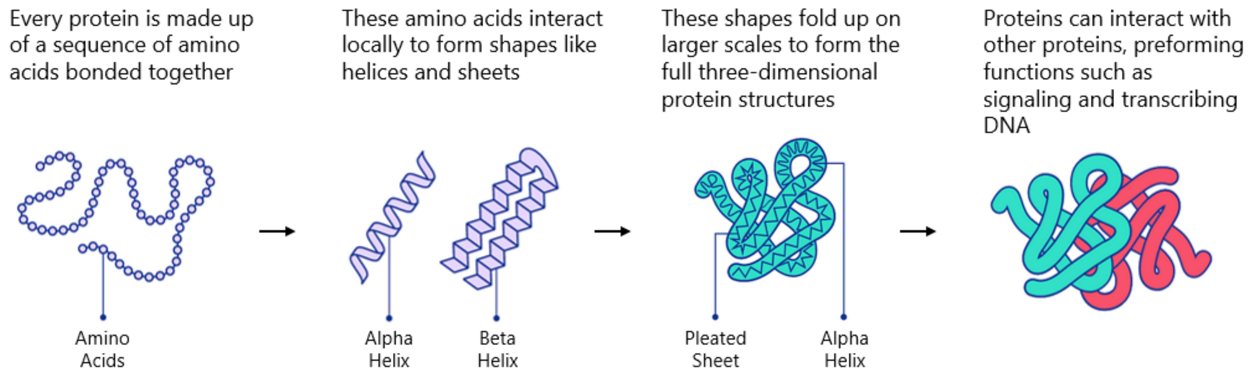
[9] https://www.nature.com/articles/d41586-020-03348-4

currently better suited for generative approaches, some for non-generative approaches and some that are still in the earliest stages of development.

Generative applications

The generative AI work out of the academic laboratories of luminaries like David Baker, Michael Bronstein and others has been nothing but revolutionary. Their work primarily focuses on generative protein structures, and it is a step function advancement in the field. As mentioned above, they have taken determining protein structures from an arduous and labor-intensive process to one that can be done *in silico* in a matter of minutes to hours (Figure 3). Anyone worldwide may now use their algorithms to take any genetic or protein sequence and generate an accurate 3D structure with it.

Figure 3: Determining protein structure from amino acid sequence.


Every protein is made up of a sequence of amino acids bonded together
These amino acids interact locally to form shapes like helices and sheets
These shapes fold up on larger scales to form the full three-dimensional protein structures
Proteins can interact with other proteins, preforming functions such as signaling and transcribing DNA
Amino Acids
Alpha Helix
Beta Helix
Pleated Sheet
Alpha Helix

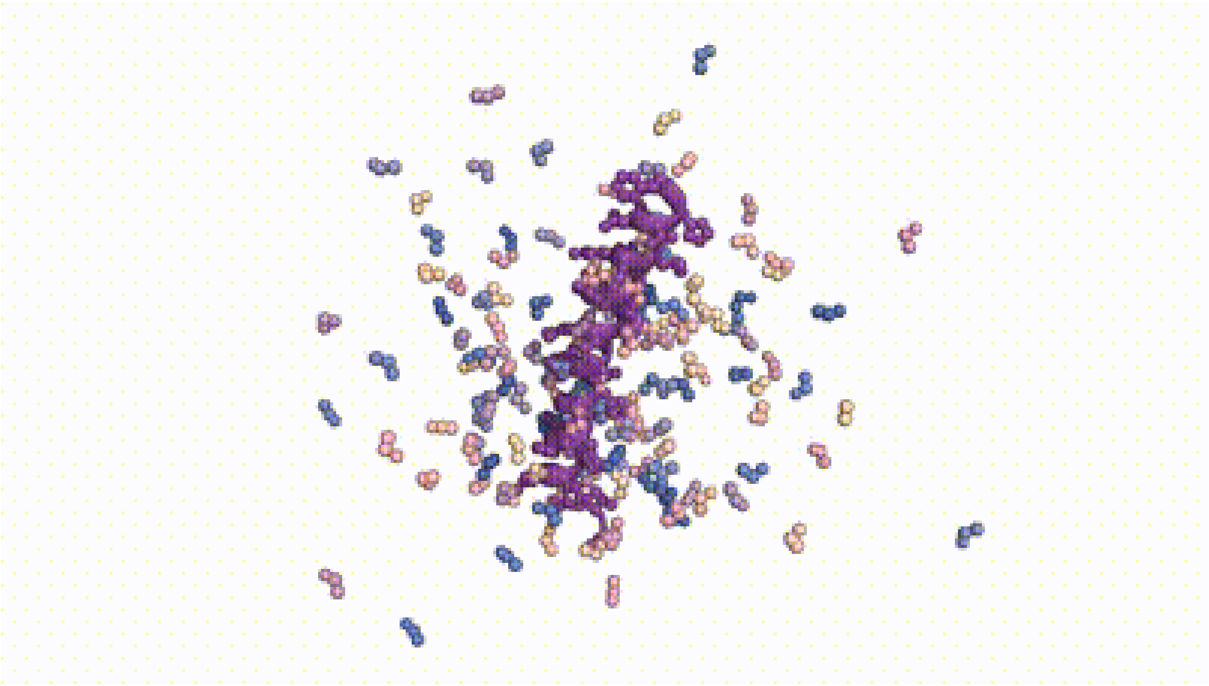
Source: Baker Lab.

Generative protein design has the right blend of large, empirically determined datasets (think datasets of 200,000 protein or fragment structures) to train on and the proper constraints to make accurate generative predictions on 3D structures.

Having a deep knowledge of protein structure is a powerful tool in both basic and pharmaceutical research. The structure of a protein directly correlates with its function. This enables researchers to not only understand the behavior and impact of a protein's structure on its function but also to leverage this knowledge to find ways to modulate its structure/function for therapeutic benefits. For example, kinases are enzymes that serve dozens of different functions in the cell, and individual kinases have been linked to regulating important cellular functions in both disease and health. These proteins have enzymatic pockets that allow them to catalyze chemical reactions in the cell. Some of these targets, such as EGFR, HER2 and cMet, drive cell growth; by blocking their enzymatic pocket with a therapeutic, you can halt cell growth in cancerous cells. This illustrates the power of knowing a protein's 3D structure. Without a 3D structure of the target, it is extremely hard to make and optimize therapeutics that fit within these pockets.

Biological therapeutics is an area where generative models currently have a significant impact on drug discovery and development. These biological therapeutics, composed of amino acids, either natural or synthetic, interact with their target protein through protein-protein interactions. Generative models excel in this space, helping to discover and optimize their interaction to develop a drug. This is because, like proteins, the function of biological therapeutics is also determined by its structure. Biological therapies typically involve either engineering a natural peptide to have drug-like properties or brute-force screening of thousands to millions of peptides or antibodies to identify ones that bind to a target region for development. Though these top-down approaches have been productive, they are very labor- and time-intensive. However, newer generative models, like RFdiffusion and others, can provide a focused bottom-up approach that allows for focused screening based on predicted structures (Figure 4). If researchers can model the specific region or pocket of the target protein to bind to modulate its activity for therapy, then generative models could predict the shape the therapeutic must have to fit within that area.

Figure 4: RF diffusion protein interaction mapping.



Source: RFdiffusion the Baker Lab.

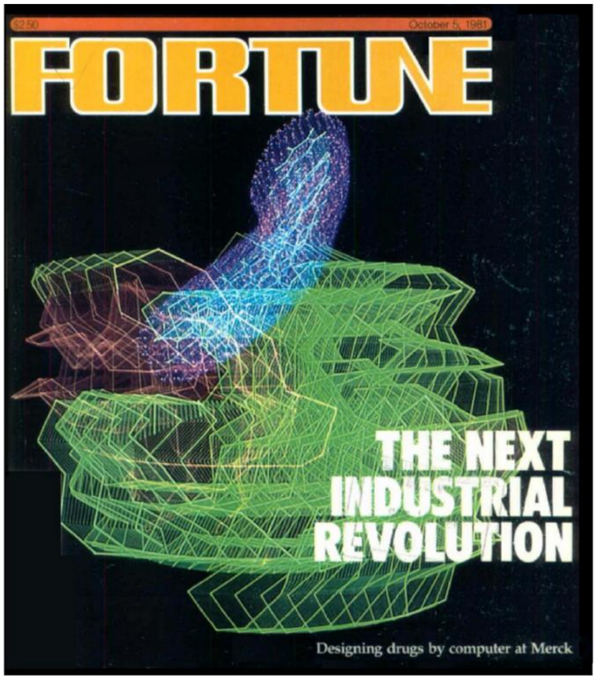
This eliminates the need to 'fish out' candidates that bind to their target region and focuses on predictive structures that fit within that region. We are starting to see companies adopt this generative approach for developing biologic drug discovery.

Even with recent advancements, significant limitations and gaps persist, particularly in terms of computing power and the availability of high-quality data. While improvements in GPU technology have somewhat alleviated the computational demands for running certain models, the need for greater computing power remains a critical bottleneck, especially for executing more complex models or conducting multiple protein structure analyses simultaneously. Furthermore, there is a critical need to broaden our datasets to encompass structures that defy traditional analytical methods. Notably, existing models predominantly rely on data derived from structures suited to X-ray crystallography, neglecting many proteins with crucial unstructured regions that resist crystallization. To bridge this gap, alternative techniques such as cryo-electron microscopy and sophisticated Nuclear Magnetic Resonance (NMR) methods are indispensable. These approaches can unveil the structures of these elusive proteins, thereby enriching our datasets and enabling the development of next-generation models capable of accurately predicting unstructured protein regions. At Luma, we believe these applications are at the intersection of Here and Hope, and the next couple of years of refinement of these applications will likely shift from Hope to Here.

Non-Generative applications

In the realm of biotechnology, while generative models have garnered considerable attention recently, non-generative models have been silently revolutionizing the field for decades (Figure 5). These models have carved out a distinct suite of applications that do not overlap with those of generative models, indicating a diverse technological landscape. For instance, the integration of AI into the fabric of small molecule drug design and development stands as a testament to this evolution, leveraging computational datasets not merely as a supplement but as a cornerstone of innovation. These methodologies go beyond filling predictive gaps, offering granular insights that propel the sector from theoretical frameworks to tangible advancements. This field has been gradually advancing for decades from better models, and faster/cheaper computation and not generative gap filling. This narrative extends beyond drug discovery to include diagnostic applications, as will be elaborated in our Curve case study.

Figure 5: Fortune magazine cover from 1981.


$2.50
October 5, 1981
FORTUNE
THE NEXT INDUSTRIAL REVOLUTION
Designing drugs by computer at Merck

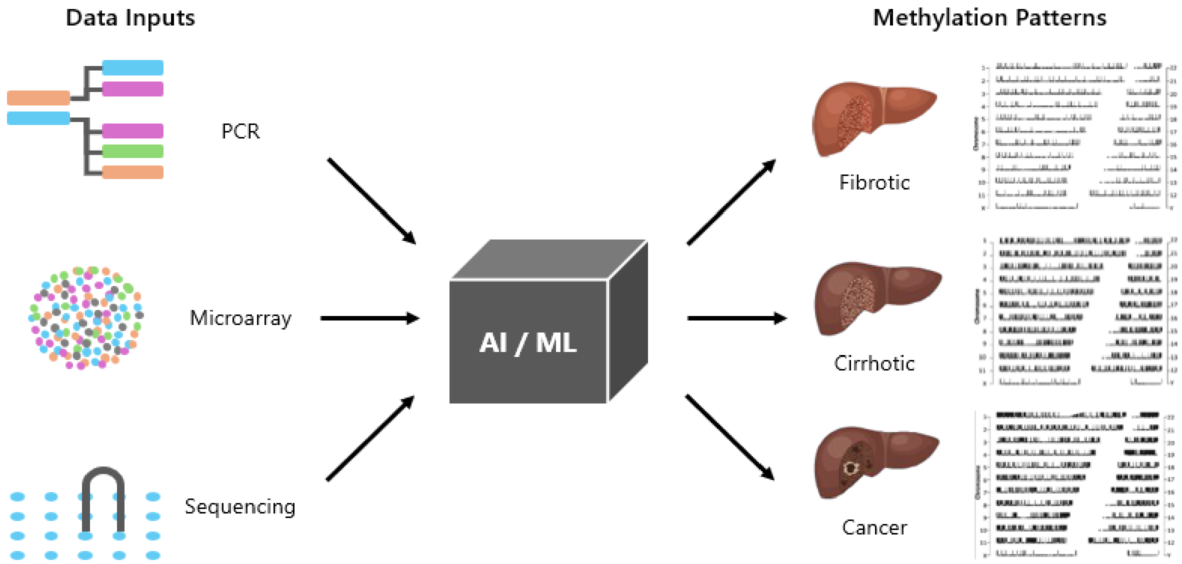
Source: Fortune Magazine.

The journey of classical small molecule drug discovery, akin to that of biologic drugs, is notoriously long and laborious. Yet, non-generative AI offers a beacon of efficiency, potentially eclipsing traditional methodologies. In a standard process, an exhaustive screening of thousands to millions of molecules culminates in the identification of a minuscule fraction of compounds that effectively bind to and modulate target proteins. These initial 'hits' are embryonic in their drug potential, requiring years, extensive financial investment and relentless scientific effort to mature into viable drug-like molecules through lead optimization. For decades, non-generative AI has been pivotal in streamlining this process, significantly economizing time, money and labor. Beyond cost reduction, non-generative AI has been a catalyst for innovation, pushing the boundaries of traditional drug discovery.

Curve Biosciences ("Curve" or "Curve Bio")

Leveraging our industrial knowledge and the capabilities of AI in the present, our fund has already actively engaged and is primed to capitalize on the technologies that are "Here" and optimizing forecasting for future advances and innovation within the AI space. For example, in 2024, we led a Series A investment in Curve Bio. Curve Bio is utilizing AI to make novel diagnostic insights for early-stage liver disease detection and progression monitoring that cannot be made without their innovative algorithms. Specifically, Curve Bio, like other researchers, observed that methylation of cell-free DNA increases as various liver diseases progress from one stage to another. This methylation accumulation occurs in circulating cell-free DNA and occurs long before the pathology of the disease manifests and highly correlates with disease progression. This is critically important because, by the time the disease pathology progresses to a detectable point by a physician, it is far too late for meaningful interventions. Additionally, our current methods of tracking disease progression, which include AFP testing and ultrasound every six months, are not sensitive enough for early detection.

Figure 6: Curve leverages AI/ML to process large PCR, microarray and sequencing data sets to inform their screening technology.



Source: Company Materials.

Curve Bio set out to improve patient outcomes by focusing on developing a test that had both early detection capabilities and the sensitivity for easy and accurately tracking disease. To solve this problem, they turned to examining thousands of patient samples with millions of individual methylated DNA segments to find a predictive pattern. This task was tailor-made for an AI solution, due to 1) large unbiased and well-curated datasets, 2) longitudinal patient outcome data, and 3) unique and constrained classifiers. They trained on this extensive profile to determine non-obvious methylation profiles that are predictive of the stage and progression of the disease that can be obtained by simple blood draws. They have shown that their approach is significantly more sensitive and specific for early detection compared to the standard detection protocol and its nearest competitors.

At the moment, the company is focused on liver diseases; however, this approach can be applied to a wide variety of diseases that require chronic disease monitoring. The company is beginning to build and train datasets for its whole body tissues atlas to develop tests for multiple tissue types.

What's Hype: time will tell

The rapid advancements in AI are undoubtedly reshaping the healthcare sector. However, amidst this wave of innovation, numerous intricate scenarios arise, straddling the line between current limitations and those unlikely to be overcome soon. It's essential, therefore, to grasp both the potential and the boundaries of these technological advancements comprehensively. This understanding is particularly critical for investors, where distinguishing between groundbreaking innovation and speculative hype can often be challenging.

Data matters: garbage in, garbage out

This old adage is very pertinent to generative AI due to the foundational role of data quality in training AI models – generative or non-generative. AI models, designed to deliver predictive and/or generative outputs, require training on high-quality data, depending on the integrity and relevance of these datasets to produce accurate, coherent and contextually suitable responses. This process makes their output very sensitive to the quality and complexity of the training data.

There are five key issues to consider:

1. **Quality of Training Data:** Generative models learn to mimic and create new content based on the data they are fed during the training process. If this training data is poor quality, biased or contains errors, the AI is likely to replicate these issues in its outputs. High-quality, well-curated data is essential for training effective models.
2. **Model Bias**: If the input data is biased, the AI model will inherently learn these biases and reflect them in its outputs. This can perpetuate stereotypes, unfair representations, or skewed perspectives, highlighting the importance of balanced and diverse datasets.
3. **Error Propagation**: Errors in the training data can propagate through the AI's learning process, leading to incorrect or nonsensical outputs. The model's ability to generalize or infer from the input data is compromised if that data is flawed, resulting in outputs that may be irrelevant or misleading.
4. **Reliance on Context and Subtlety**: AI often deals with nuances which require a high degree of accuracy and sophistication in the training data. Subtle inaccuracies or a lack of contextual diversity in the input data can lead to outputs that are off-mark or lack subtlety.

Feedback loops

In some instances, AI models are part of systems that use their outputs as new inputs in a continual learning process. If the initial outputs are flawed due to poor quality inputs, this can create a feedback loop that further degrades the quality of future outputs. This directly applies to applications within drug discovery. There is a wealth of disparate curated data that has been generated at varying degrees of curations and quality. Take AlphaFold as an example. Generative AI models such as AlphaFold are at the forefront of predicting protein structures, largely due to their reliance on exceptionally high-quality training datasets. These models' unparalleled accuracy in protein structure prediction is attributable to the distinct characteristics of protein data, which include highly uniform and extensive datasets amassed over decades. These datasets contain atomic-level resolution structures that have been meticulously collected in a standardized fashion. Moreover, the predictive task is somewhat constrained by the limited variability within proteins, given that all proteins are composed of the same 20 amino acids.

This inherent limitation in variability contrasts sharply with the complexities encountered in fields like small molecule drug discovery, where the diversity and potential configurations are vastly greater. Despite the success of generative models like AlphaFold in protein structure prediction, it's essential to acknowledge their limitations. These models primarily excel at identifying the most energetically stable configurations of proteins. However, this perspective overlooks the intrinsic flexibility of proteins; many, particularly enzymes, exhibit multiple structural states influenced by their environmental conditions. The most stable structure, often the focus of these models, does not necessarily represent the full spectrum of a protein's dynamic nature. Proteins frequently contain both rigid and flexible regions, including intrinsically disordered regions ("IDRs"), which lack a fixed or ordered structure. Traditional empirical methods, which form the basis of training datasets for generative models, struggle to capture the nuances of IDRs accurately. Consequently, IDRs are underrepresented in these datasets, leading to a gap in the models' ability to predict such unstructured regions accurately.

This limitation underscores a critical point: the utility of generative models in protein structure prediction is inherently tied to the content of their training datasets. While these models offer significant insights into already understood protein structures, their capacity to uncover novel insights or predict the behavior of less characterized regions, such as IDRs, is limited. Therefore, while generative AI models represent a monumental step forward in our understanding of protein structures, their current application is predominantly confined to modeling scenarios well-represented in existing datasets. This recognition highlights the ongoing need for complementary techniques that can probe the dynamic and less understood aspects of proteins, offering a broader view beyond what current generative models can achieve.

In the realm of predictive generative technologies, akin to the breakthroughs achieved by AlphaFold in protein structure prediction, the near future may not hold significant advancements for several compelling reasons. At Luma, our outlook remains hopeful for eventual progress, yet we acknowledge the current landscape is more characterized by its speculative enthusiasm than by tangible achievements. The primary obstacles stem from two critical areas: the scarcity of comprehensive data and the inherent tendencies of AI systems to seek the path of least resistance.

Data scarcity and constraints

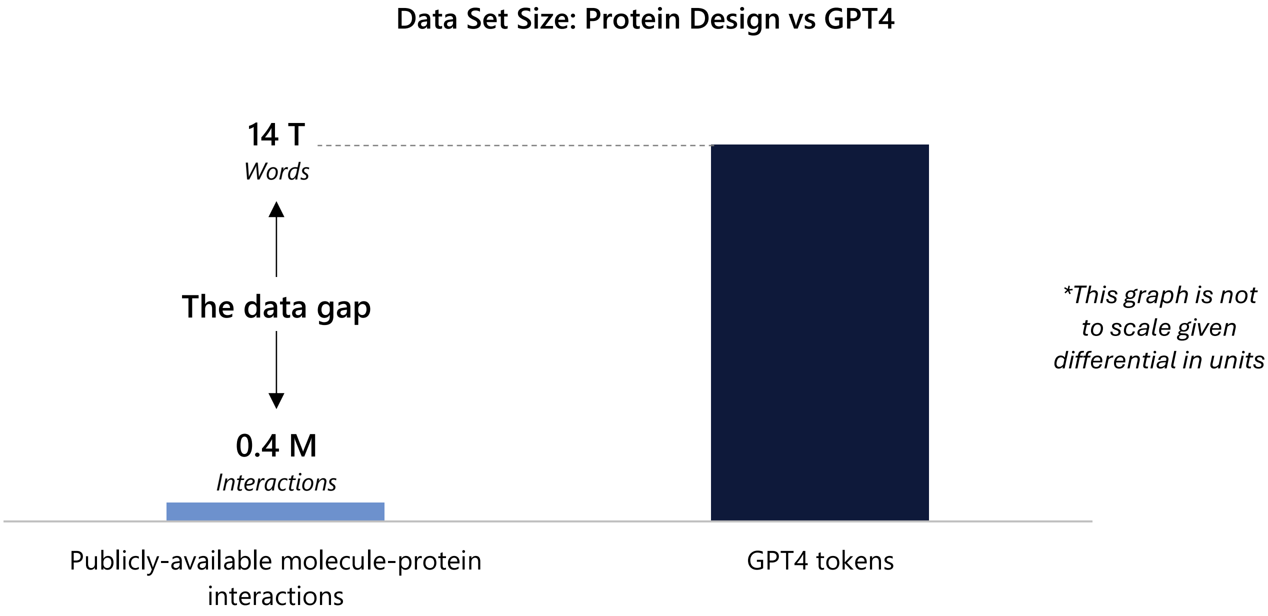
A direct relationship exists between the expansiveness of the dataset employed for AI training and the limitations imposed by variability constraints. In essence, the broader the scope of variability, the greater the need for extensive, high-quality data. Protein structure design, facilitated by generative AI like AlphaFold, benefits from three inherent advantages:

1. Composition of proteins from a finite set of 20 amino acids
2. Detailed, predictable properties of these amino acids
3. Five decades of well curated, standardized and high-resolution datasets

This allows for the utilization of smaller yet highly refined and curated datasets. Contrarily, the domain of small molecules presents a starkly different challenge. Despite their smaller size, small molecules exhibit immense variability. It's estimated that the universe of drug-like molecules spans to 10^{23}, with research from the University of Berne indicating that over 166 billion of these molecules are synthetically accessible.

The current datasets available for training in this vast chemical space are markedly inadequate. Although pharmaceutical companies have amassed millions of data points from decades of small molecule screenings, this wealth of information remains proprietary and inaccessible to the broader research community. Publicly available resources such as BindingDB — with its catalog of over a million small molecule-protein interactions — still fall significantly short of the comprehensive coverage needed to train generative models effectively. Another resource, PubMed, contains over 300 million interactions, but its data is less curated and more chaotic. For context, ChatGPT-4's training involved over 14 trillion unique data points (words), underscoring the immense data gap that exists (Figure 7). This comparison highlights the pressing need for standardized, extensive datasets akin to those in protein structure research to pave the way for groundbreaking advancements in small molecule drug discovery.

Figure 7: The size of data sets used to create popular AI model ChatGPT is several orders of magnitude larger than available data sets for drug design.



Source: Nature.

What does the future hold?

We believe that advancements in AI will profoundly influence both healthcare and overall human health. Our strategy is characterized by a rigorous and methodical approach in understanding the capabilities and developmental timelines of these technological advancements. This ensures our investments are timed perfectly to back companies equipped with the most appropriate and impactful technologies.

Despite the considerable excitement surrounding our sector, we recognize that real innovation stems from evolving approaches and methodologies. These promise to herald significant breakthroughs in both the short and long term. Our active investment in AI-centric companies, which demonstrate immediate potential (as exemplified by our Curve case study, Rome Therapeutic's work and several future Luma portfolio companies), alongside a strategic wait-and-see approach for emerging advancements, positions us to capitalize on opportunities that transform initial optimism into tangible outcomes. We are particularly impressed by initiatives like those undertaken by the CZI, focused on data standardization. Such efforts are crucial in addressing the myriad challenges that currently hinder progress in our field.