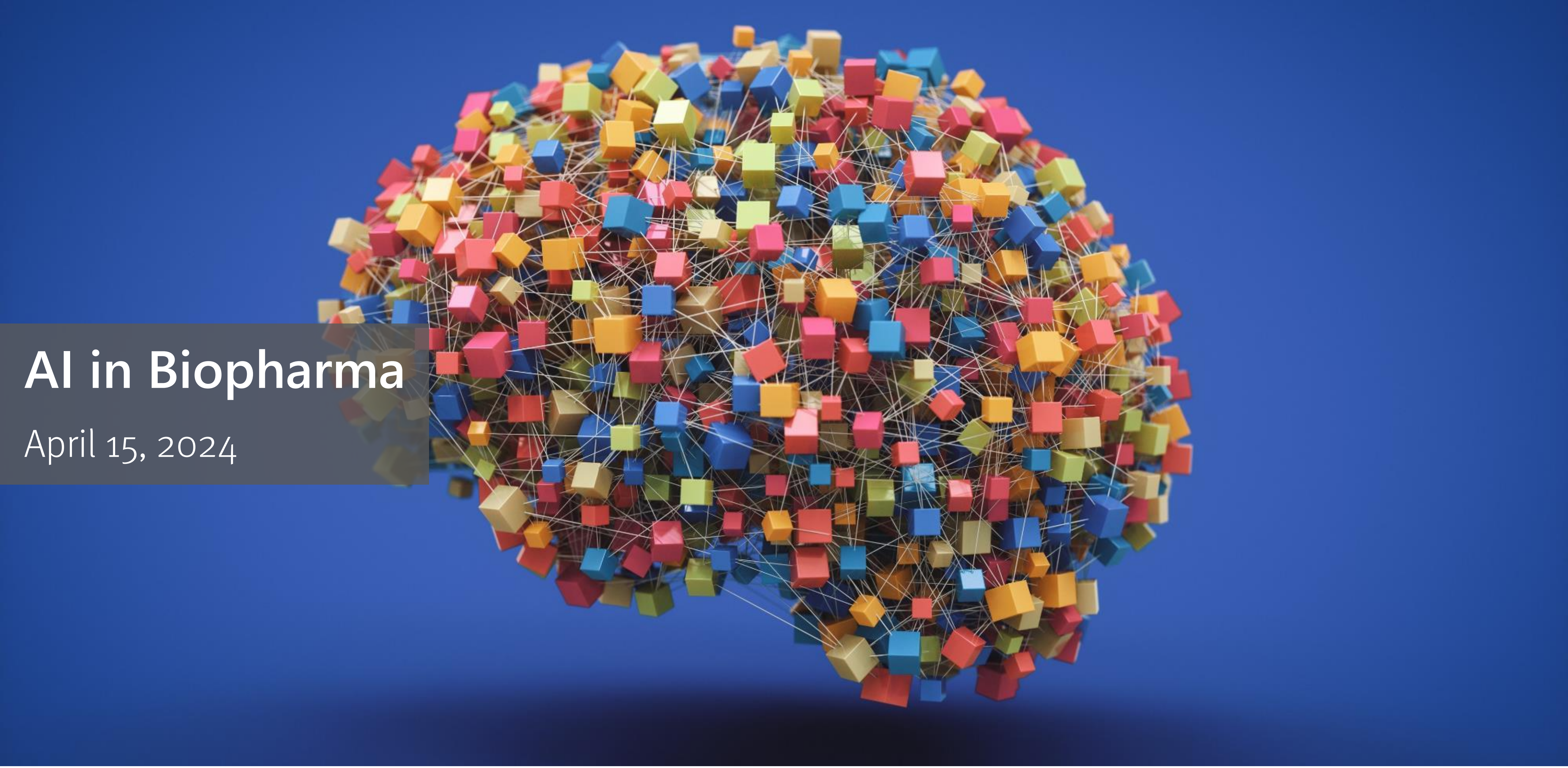


AI in Biopharma

April 15, 2024

STIFEL | Healthcare

Table of Contents

STIFEL | Healthcare

787 7th Avenue, New York NY 10019, +1 (212) 887-7777
web: www.stifel.com

Introduction

This review surveys how developments in artificial intelligence and machine learning (AI/ML) are changing the pharmaceutical sector.

We go over the basics of AI and then delve into key opportunities for the use of AI in drug discovery, clinical trials and other settings.

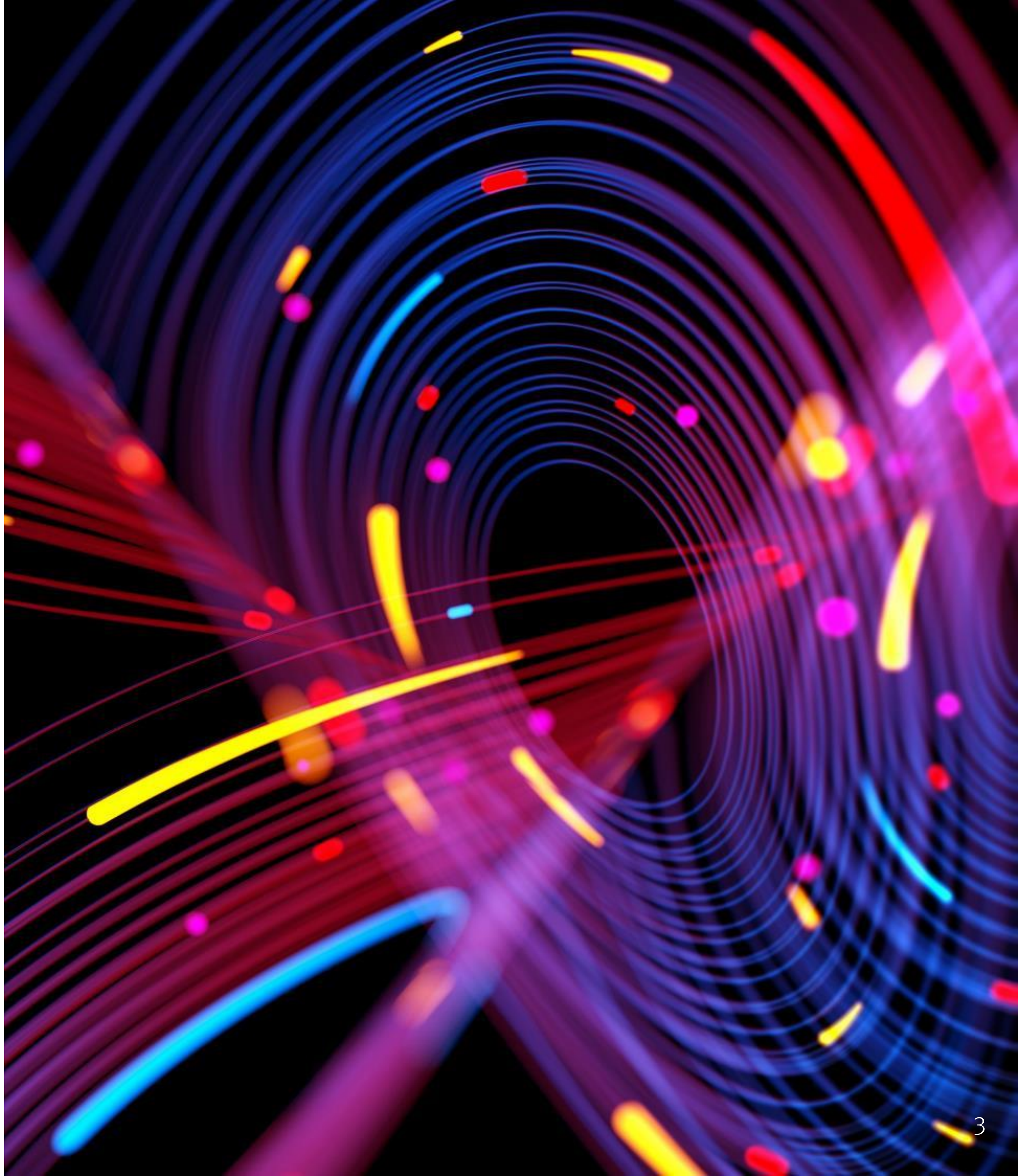
We review in detail areas that really matter in today's AI/Pharma landscape including:

1. Major approaches to target identification using machine learning
2. Approaches to dimensionality reduction when using AI in small and large molecule drug discovery. These include physics-based approaches, cell-based approaches and quantitative SAR studies.
3. Ways of solving the "dirty data" problem that besets the industry
4. What is going on with AI in clinical drug development
5. Guidelines for investing in AI companies in biopharma and
6. Big pharma efforts in AI/ML

We wish to acknowledge fantastic help from the dozens of people who helped to inform us and comment on this review.

Tim Opler, *Managing Director*, Healthcare, Stifel (oplert@stifel.com)

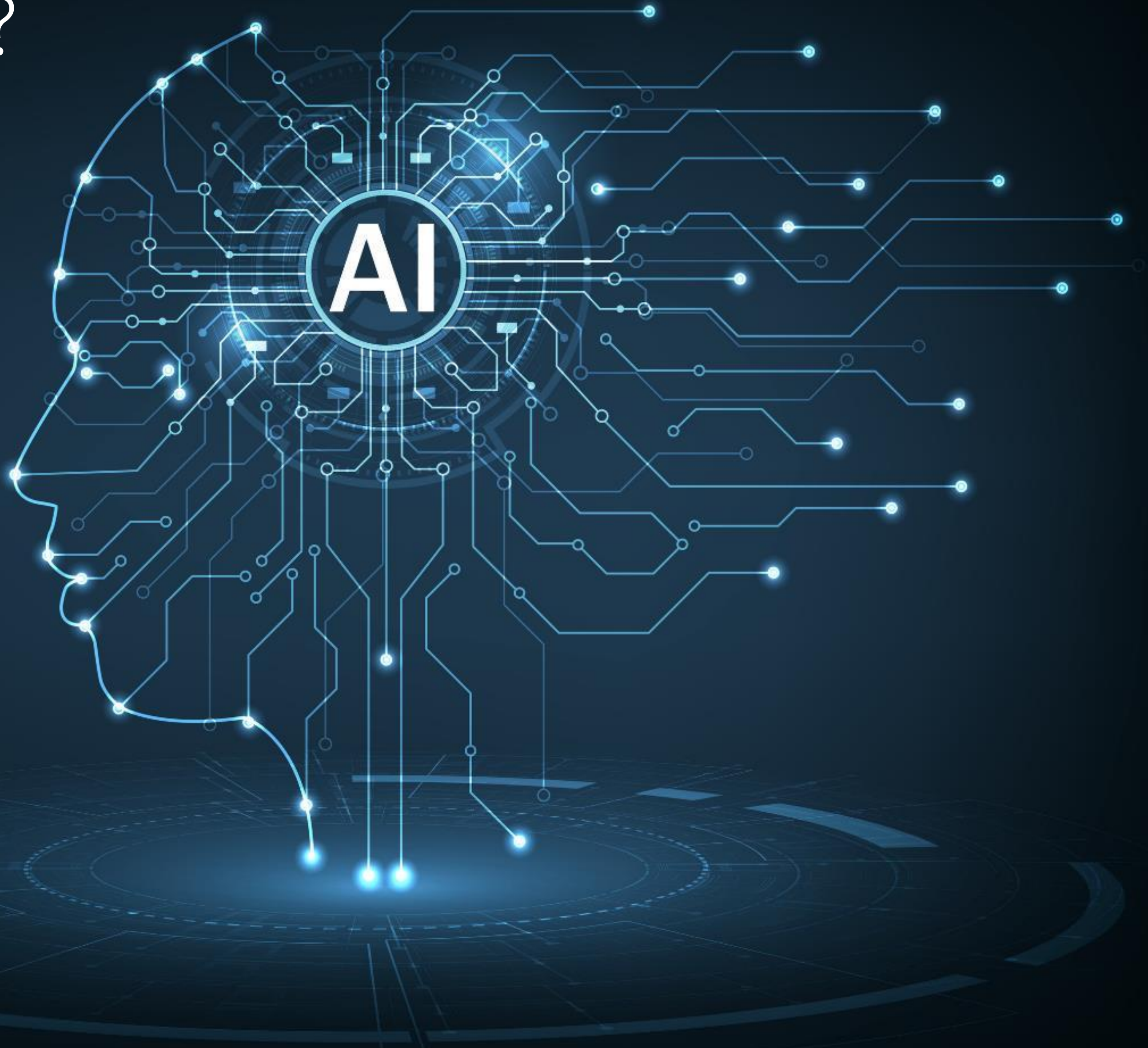
To subscribe to our regular weekly updates and reports click here and go to page 4 for details.



AI Basics

What is Artificial Intelligence?

Artificial Intelligence is the ability for a computer to think, learn and simulate human mental processes, such as perceiving, reasoning, and learning.



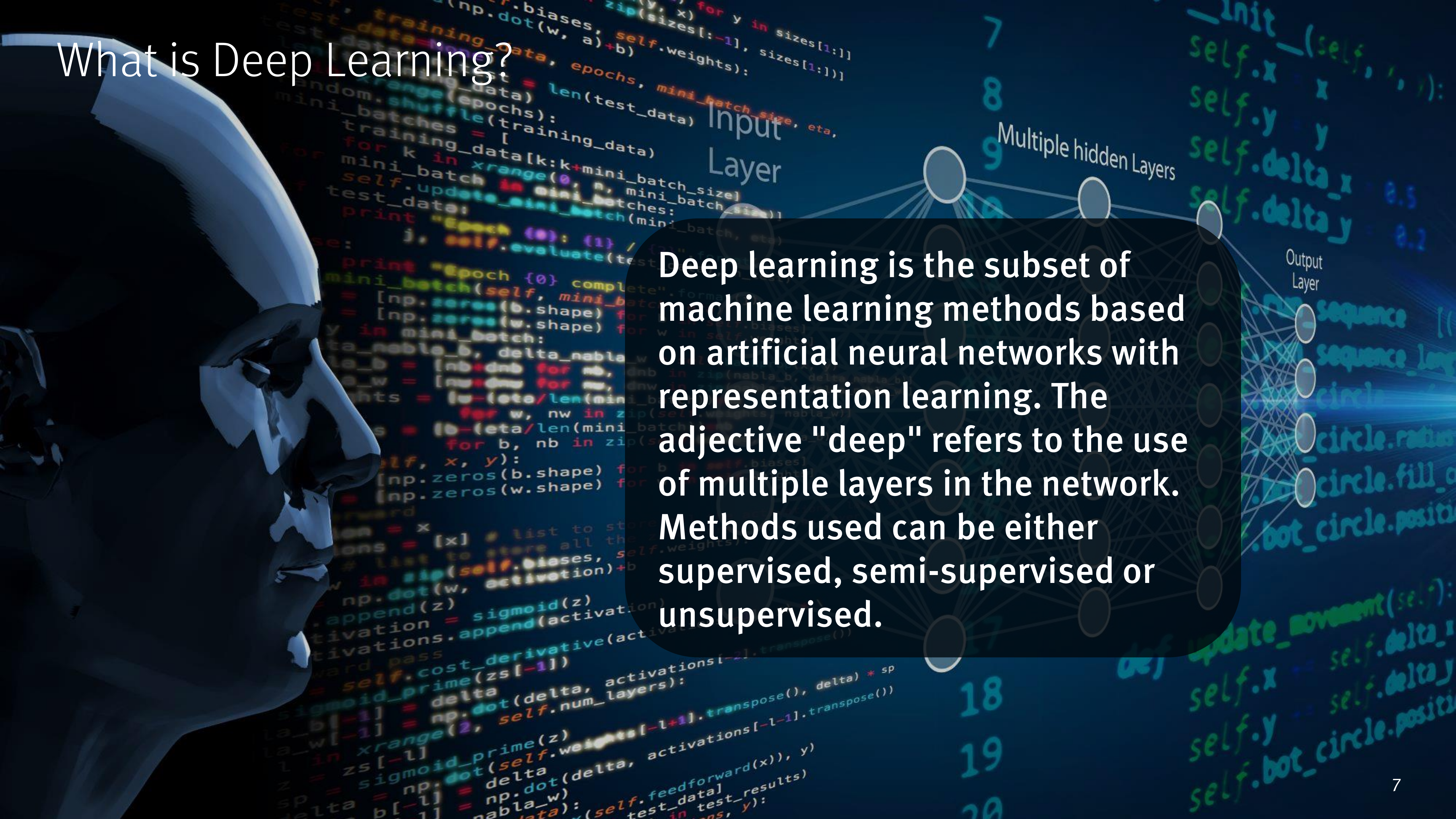
What is Machine Learning?

While artificial intelligence encompasses the idea of a machine that can mimic human intelligence, machine learning does not.

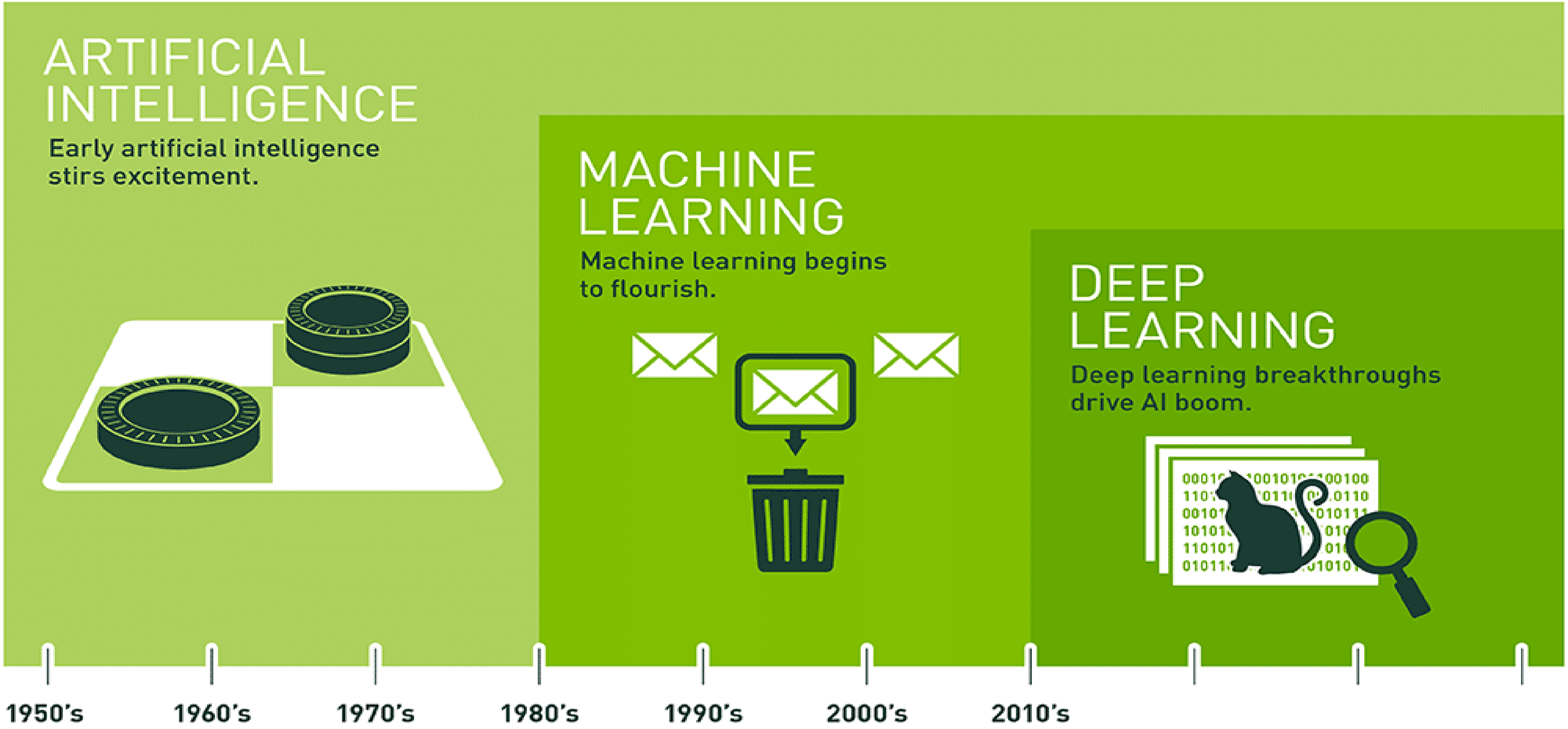
Machine learning aims to teach a machine how to perform a specific task and provide accurate results by identifying patterns.

What is Deep Learning?

Deep learning is the subset of machine learning methods based on artificial neural networks with representation learning. The adjective "deep" refers to the use of multiple layers in the network. Methods used can be either supervised, semi-supervised or unsupervised.



The Current Excitement With AI Has Been Driven by Deep Learning



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

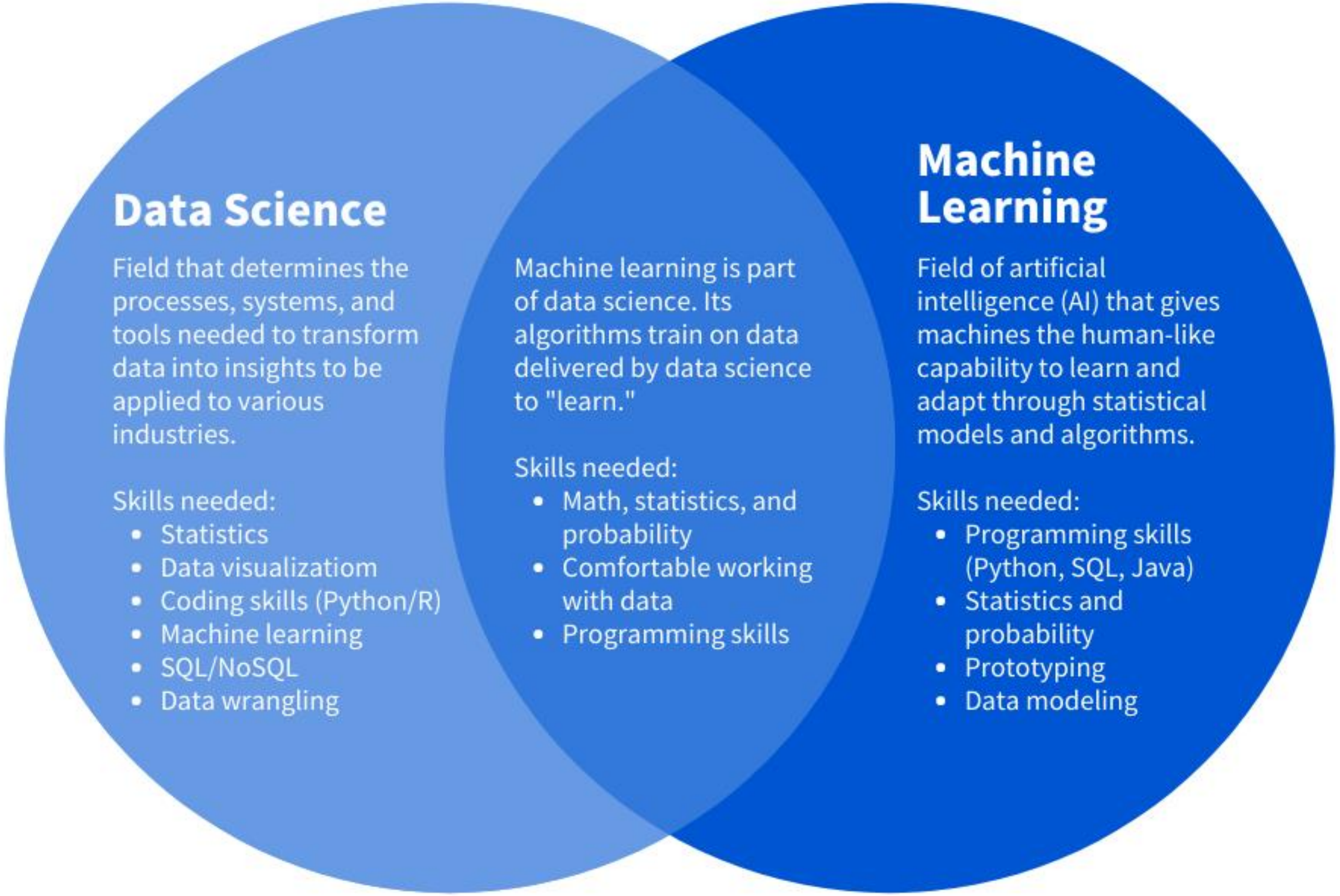
DEEP LEARNING

Deep learning breakthroughs drive AI boom.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: Nvidia

Machine Learning Versus Data Science



Data Science

Field that determines the processes, systems, and tools needed to transform data into insights to be applied to various industries.

Skills needed:

- Statistics
- Data visualizatiom
- Coding skills (Python/R)
- Machine learning
- SQL/NoSQL
- Data wrangling

Machine learning is part of data science. Its algorithms train on data delivered by data science to "learn."

Skills needed:

- Math, statistics, and probability
- Comfortable working with data
- Programming skills

Machine Learning

Field of artificial intelligence (AI) that gives machines the human-like capability to learn and adapt through statistical models and algorithms.

Skills needed:

- Programming skills (Python, SQL, Java)
- Statistics and probability
- Prototyping
- Data modeling

Data science is the process of developing systems that gather and analyze disparate information to uncover solutions to various business challenges and solve real-world problems. Machine learning is used in data science to help discover patterns and automate the process of data analysis. Data science contributes to the growth of both AI and machine learning. This article will help you better understand the differences between AI, machine learning, and data science as they relate to careers, skills, education, and more.

It’s important to consider how data science, machine learning and AI intersect. Fundamentally, machines can't hope to mimic humans' cognitive processes without information -- and Data scientists are tasked with "feeding" machines accurate, empirical data and statistical models that enable machines to learn autonomously. By constantly improving machine learning, society comes closer to realizing true artificial intelligence (AI).

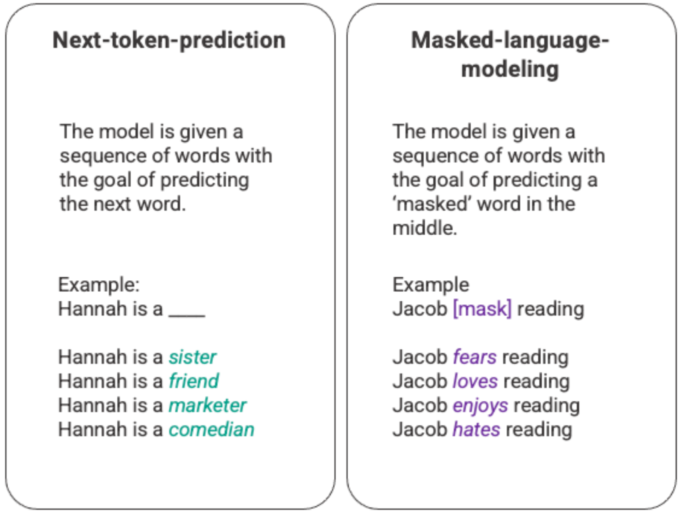
Sources: https://www.coursera.org/articles/data-science-vs-machine-learning, https://csweb.rice.edu/academics/graduate-programs/online-mds/blog/data-science-vs-ai-and-ml

How LLM’s and ChatGPT Work

Molly Ruby, *Toward Data Science*, January 2023

ChatGPT is an extrapolation of a class of machine learning Natural Language Processing models known as Large Language Model (LLMs). LLMs digest huge quantities of text data and infer relationships between words within the text. These models have grown over the last few years as we’ve seen advancements in computational power. LLMs increase their capability as the size of their input datasets and parameter space increase.

The most basic training of language models involves **predicting a word in a sequence of words**. Most commonly, this is observed as either next-token-prediction and masked-language-modeling.



Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:
Hannah is a ____

Hannah is a *sister*
Hannah is a *friend*
Hannah is a *marketer*
Hannah is a *comedian*

Masked-language-modeling

The model is given a sequence of words with the goal of predicting a ‘masked’ word in the middle.

Example
Jacob [mask] reading

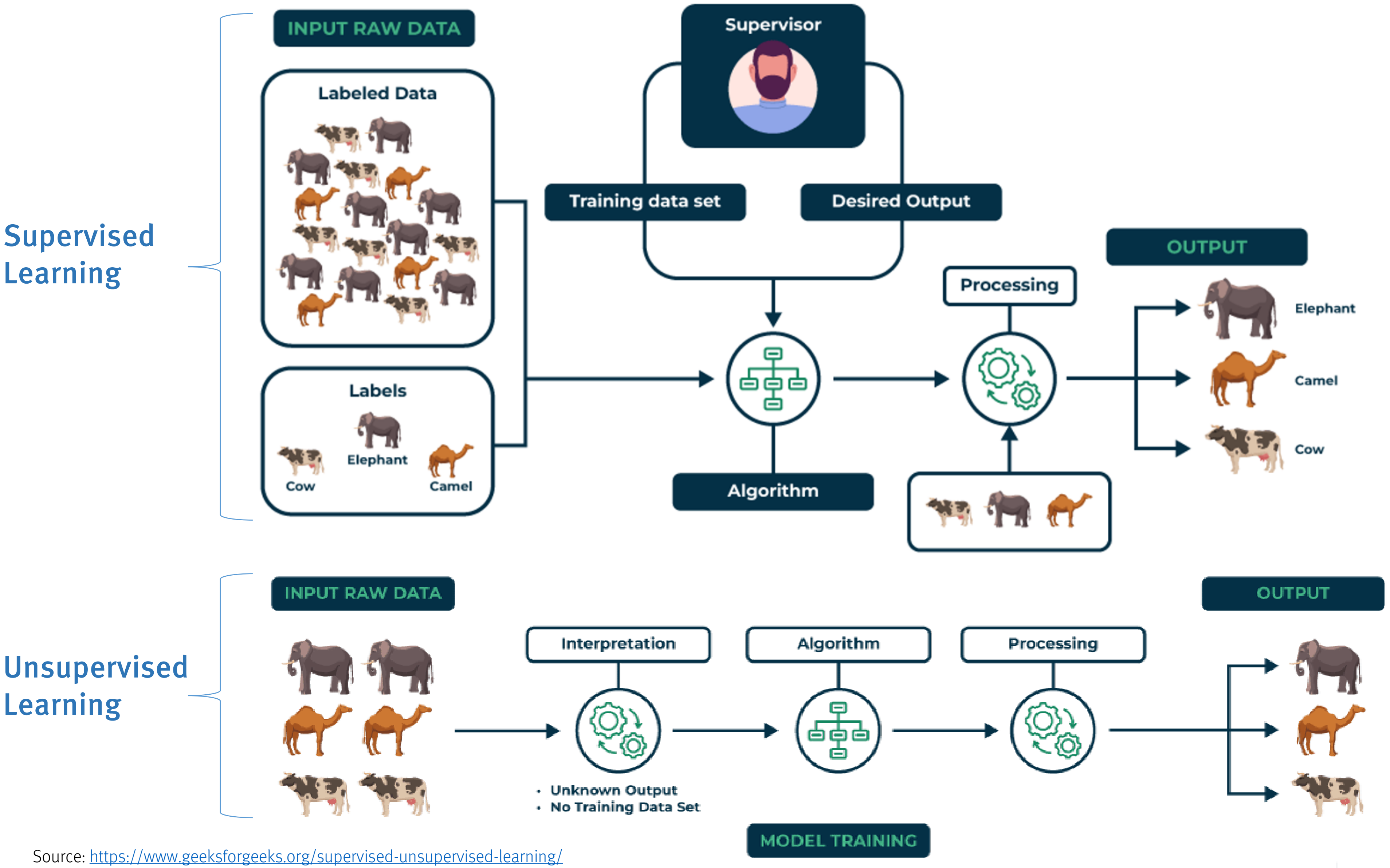
Jacob *fears* reading
Jacob *loves* reading
Jacob *enjoys* reading
Jacob *hates* reading

All GPT models largely follow the Transformer Architecture established in “Attention is All You Need” (Vaswani et al., 2017), which have an encoder to process the input sequence and a decoder to generate the output sequence. Both the encoder and decoder in the original Transformer have a multi-head self-attention mechanism that allows the model to **differentially weight parts of the sequence to infer meaning and context**.* As an evolution to original Transformer, GPT models leverage a decoder-only transformer with masked self-attention heads. The decoder-only framework was used because the main goal of GPT is to generate coherent and contextually relevant text. Autoregressive decoding, which is handled by the decoder, allows the model to maintain context and generate sequences one token at a time.

* The explanation for these technical terms can be found at https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853.

Source: https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Supervised Versus Unsupervised Learning



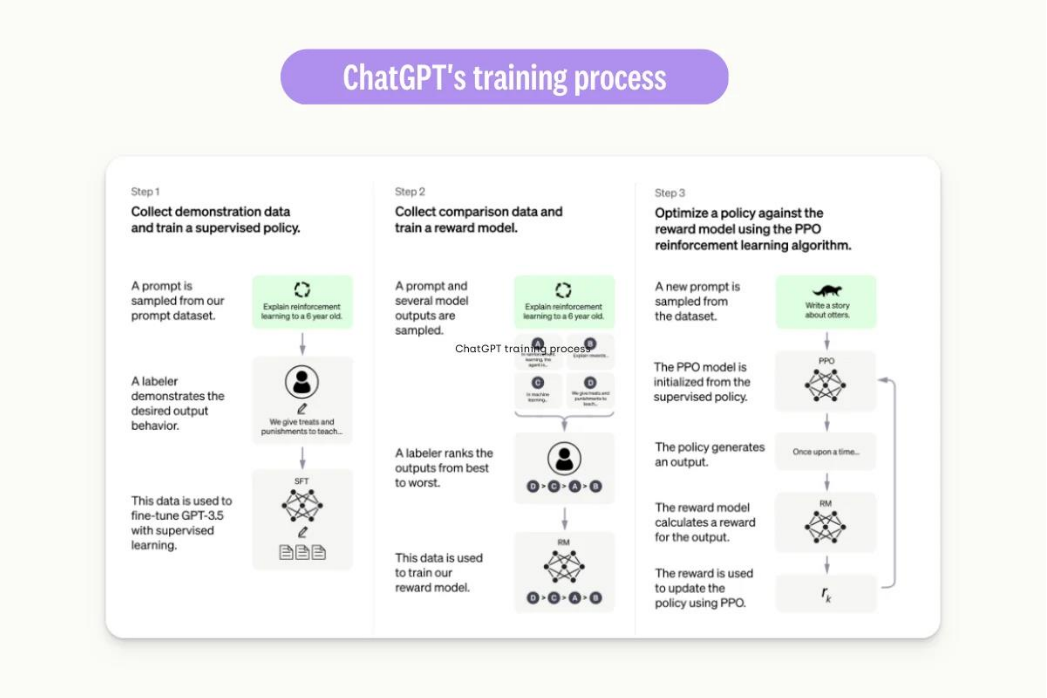
Source: https://www.geeksforgeeks.org/supervised-unsupervised-learning/

Supervised learning works from labelled data.

Labelling data makes it immensely easier for the computer to identify relationships between inputs and outputs.

Modern LLM’s use a mix of labelling with non-supervised learning. But, LLMs have a massive advantage which is that they are working from word sequences in natural language.

Today’s LLM’s Use What’s Called a PPO Model



ChatGPT's training process

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
ChatGPT training process
A
B
C
D
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, Proximal Policy Optimization Algorithms, *arXiv*, Aug 28, 2017 (abstract)

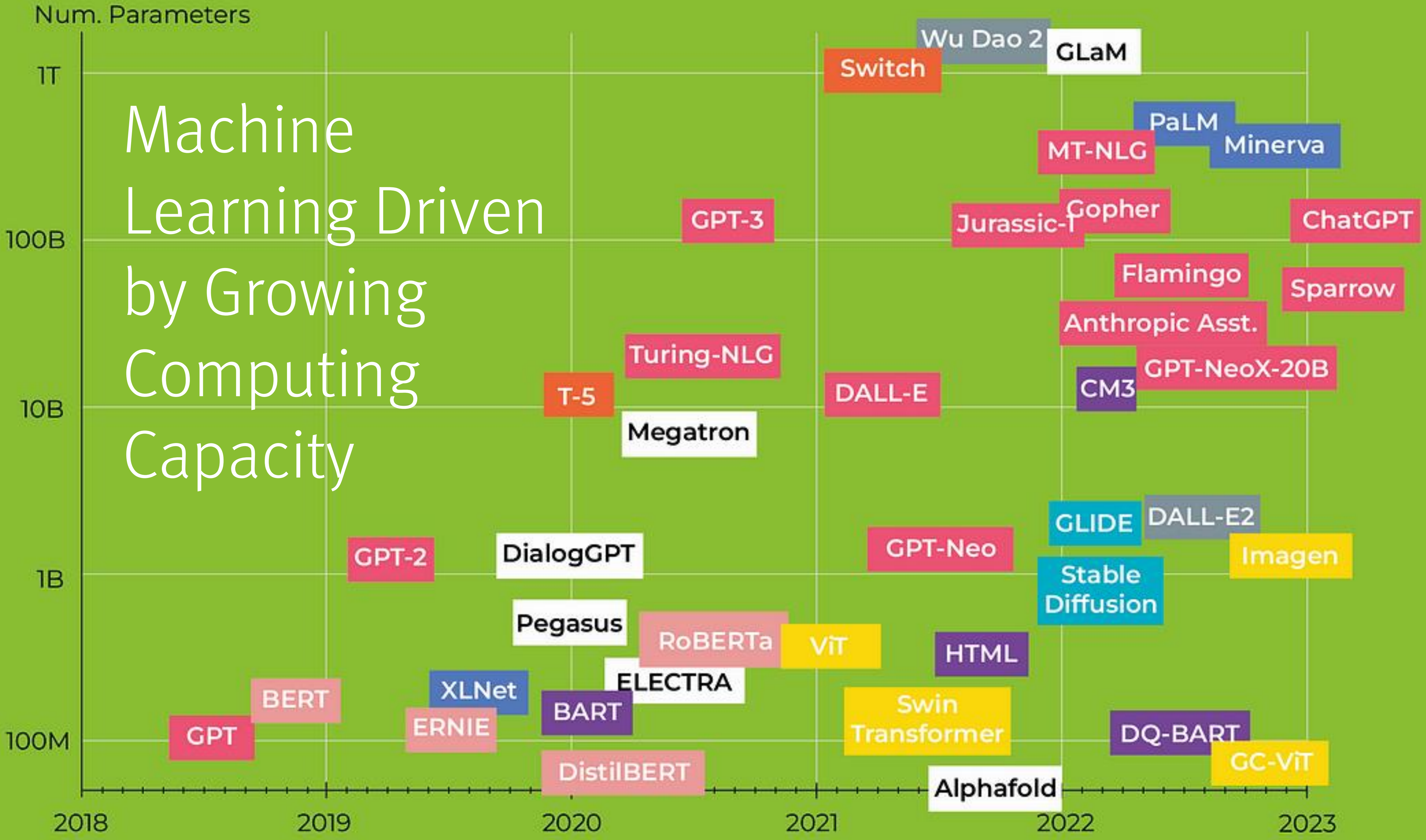
“We propose a new family of policy gradient methods for reinforcement learning, which alternate between sampling data through interaction with the environment and optimizing a "surrogate" objective function using stochastic gradient ascent. Whereas standard policy gradient methods perform one gradient update per data sample, we propose a novel objective function that enables multiple epochs of minibatch updates. The new methods, which we call proximal policy optimization (PPO), have some of the benefits of trust region policy optimization (TRPO), but they are much simpler to implement, more general, and have better sample complexity (empirically). Our experiments test PPO on a collection of benchmark tasks, including simulated robotic locomotion and Atari game playing, and we show that PPO outperforms other online policy gradient methods, and overall strikes a favorable balance between sample complexity, simplicity, and wall-time.”

What’s important here is that LLM builders such as OpenAI (maker of ChatGPT) have designed non-supervised reinforcement learning methods that are highly efficient in building and guiding transformers. Further, the prevalent PPO approach has the advantage of being able to learn from human labelling and input. openAI is aggressive in using human input to improve its LLM’s (supervised learning).*

Sources: https://arxiv.org/abs/1707.06347, https://openai.com/research/openai-baselines-ppo, https://kili-technology.com/data-labeling/how-to-boost-your-data-labeling-with-gpt-and-large-language-models, https://www.linkedin.com/pulse/openais-pioneering-contributions-reinforcement-learning-subhajit-naha

* See https://www.forbes.com/sites/aparnadhinakaran/2023/03/27/ten-questions-with-openai-on-reinforcement-learning-with-human-feedback/

Machine Learning Driven by Growing Computing Capacity



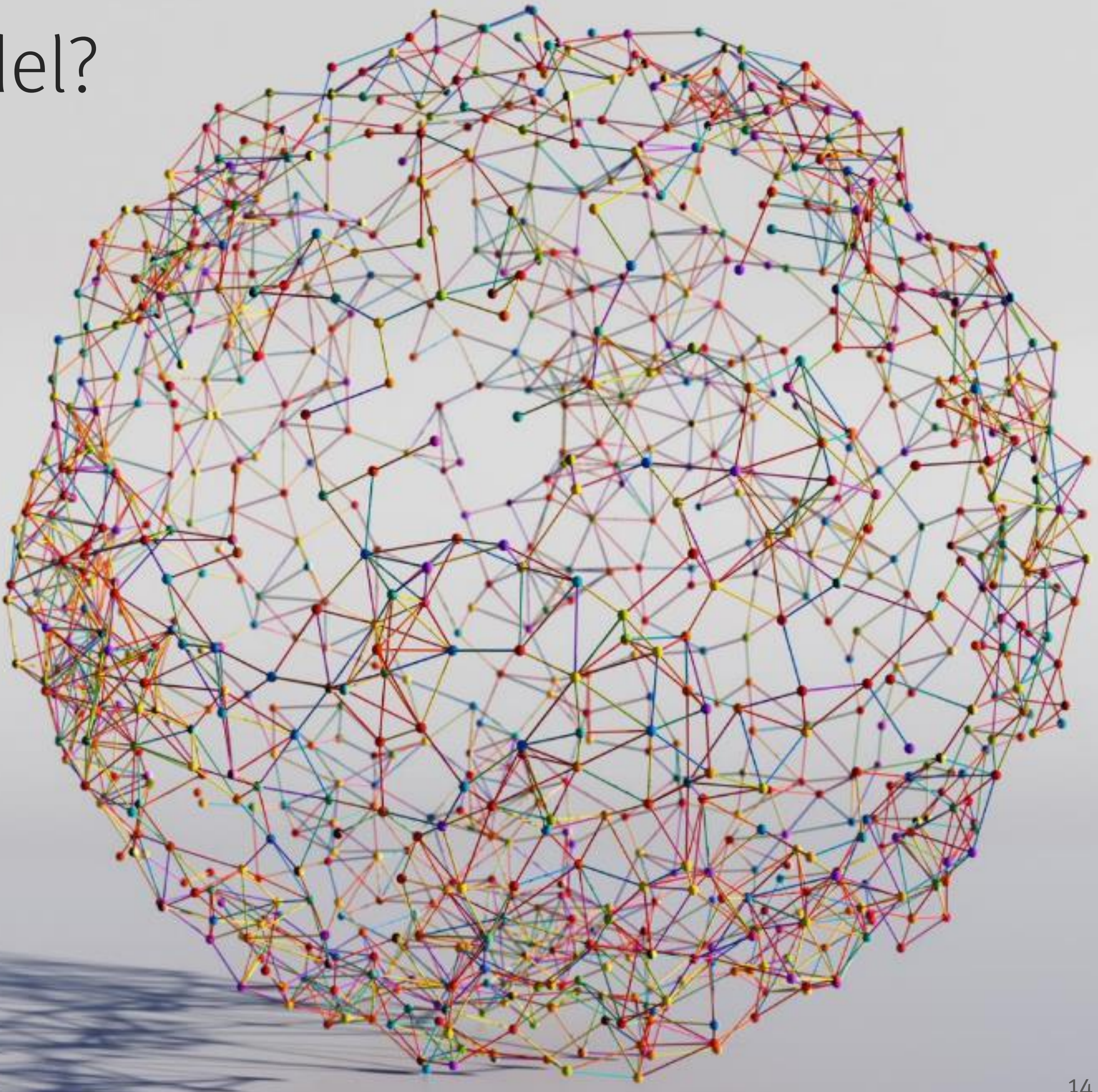
Source: https://medium.com/@vlad_zh/a-comprehensive-review-of-transformers-from-bert-to-chatgpt-f7dfe2b23043

What is a Foundation Model?

A Foundation Model is a form of generative artificial intelligence. It generates output from one or more inputs (prompts) in the form of human language instructions. Models are based on complex neural networks including transformers and variational encoders.

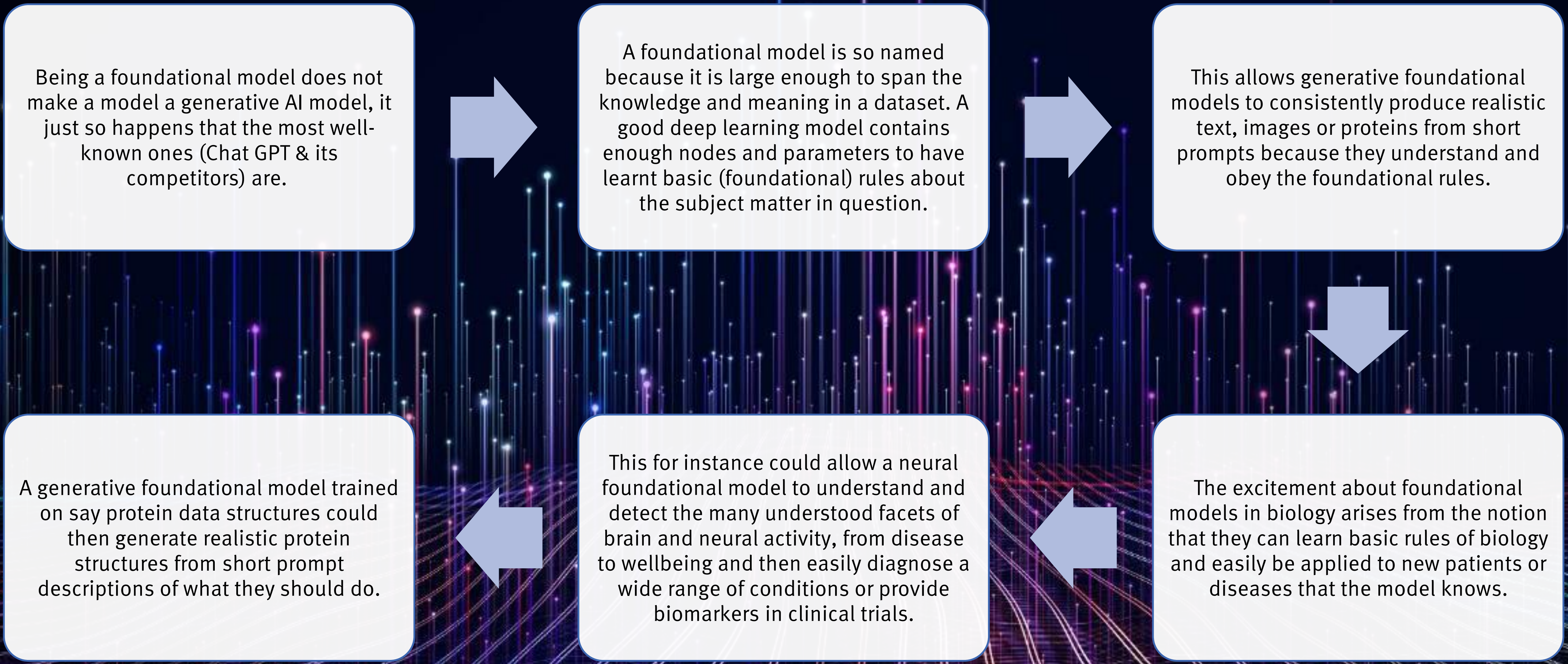
Although each type of network functions differently, the principles behind how they work are similar. In general, an FM uses learned patterns and relationships to predict the next item in a sequence. For example, with image generation, the model analyzes the image and creates a sharper, more clearly defined version of the image. Similarly, with text, the model predicts the next word in a string of text based on the previous words and its context. It then selects the next word using probability distribution techniques.

Foundation models use self-supervised learning to create labels from input data. This means no one has instructed or trained the model with labeled training data sets.



Source: https://aws.amazon.com/what-is/foundation-models

On Foundational Models, LLM’s and Biology



Being a foundational model does not make a model a generative AI model, it just so happens that the most well-known ones (Chat GPT & its competitors) are.

A foundational model is so named because it is large enough to span the knowledge and meaning in a dataset. A good deep learning model contains enough nodes and parameters to have learnt basic (foundational) rules about the subject matter in question.

This allows generative foundational models to consistently produce realistic text, images or proteins from short prompts because they understand and obey the foundational rules.

The excitement about foundational models in biology arises from the notion that they can learn basic rules of biology and easily be applied to new patients or diseases that the model knows.

This for instance could allow a neural foundational model to understand and detect the many understood facets of brain and neural activity, from disease to wellbeing and then easily diagnose a wide range of conditions or provide biomarkers in clinical trials.

A generative foundational model trained on say protein data structures could then generate realistic protein structures from short prompt descriptions of what they should do.

Big Potential of AI to Transform Biology

Jensen Huang, CEO, Nvidia, Talk at Stanford University, March 2024

We saw that if we could reduce the marginal cost of computing down to approximately zero then we might use it to do something insanely amazing: Large language models.

To literally extract all digital human knowledge from the internet and put it into the computer and let it go figure out what the knowledge is – that idea of scraping the entire internet and put it into one computer and then the computer figure out what the program is – is an insane concept.

…We have figured out how to use the computer to understand the *meaning* of almost all digital knowledge. Which means, anything we can digitize, we can understand its meaning. As an example, gene sequencing is digitizing genes. Now with large language models we can go learn the *meaning* of that gene.* We can also do that with amino acids and proteins without a large investment in things like Cryo-EM. And, now we can understand the meaning of a cell.

From a computer's perspective this is no different than digitizing a page full of words. We are all on the cusp of this. AI… is probably the single greatest invention of the technology age. This will likely be the single greatest invention of the 21st Century.

Source: https://www.youtube.com/watch?v=cEg8cOx7UZk

* See https://www.fiercebiotech.com/medtech/nvidia-unveils-generative-ai-programs-drug-molecule-design-protein-predictions

How AI Can Augment the Biopharma Industry

AI's role in drug discovery marks a pivotal shift towards more efficient, faster, and potentially less expensive development processes for new medications. Here's a breakdown of how AI is revolutionizing this field:

Target Identification: AI algorithms can sift through vast datasets of biological information to identify potential targets for new drugs. These targets are usually proteins or genes associated with a disease. By understanding the structure and function of these targets, researchers can design drugs that specifically interact with them.

Drug Design and Optimization: Once a target is identified, AI can help in the design of new drug molecules. Using techniques like deep learning, AI systems can predict how different chemical structures will interact with the target, allowing for the rapid generation and optimization of potential drug candidates. This process can significantly reduce the time and cost compared to traditional methods.

Synthesis Prediction: AI can also predict the most efficient chemical synthesis routes for new drug compounds. This capability is crucial for producing drugs at scale and can help in identifying more environmentally friendly and cost-effective synthesis pathways.

Screening and Predictive Toxicology: High-throughput screening of large compound libraries is an essential step in drug discovery. AI enhances this process by predicting the biological activity of compounds and their potential toxicity before they are synthesized and tested in the lab. This predictive power can lead to safer drug candidates and reduce the reliance on animal testing.

Clinical Trials: AI algorithms can optimize the design of clinical trials, including patient selection and monitoring. By analyzing data from electronic health records, genetic information, and other sources, AI can identify the most suitable candidates for trials and predict patient responses to a drug. This targeted approach can improve the efficiency of trials and the likelihood of successful outcomes.

Personalized Medicine: AI is paving the way for personalized medicine, where treatments are tailored to the individual patient based on their genetic makeup, lifestyle, and other factors. By analyzing large datasets, AI can help in identifying which patients are likely to benefit from specific treatments, thereby improving efficacy and reducing side effects.

Real-world Evidence: AI can analyze real-world data from various sources, including wearables, electronic health records, and social media, to monitor drug safety and efficacy in the broader population. This analysis can provide insights that are not always evident in clinical trials.

In summary, AI is a powerful tool in drug discovery and development, offering the potential to make the process faster, more efficient, and tailored to individual needs. Its impact is expected to grow as the technology advances, leading to the discovery of new therapies and treatment approaches.

Source: Stifel research

Sequoia Partner on AI: ‘This Time It’s Not Vaporware’

Bloomberg

By Edward Ludlow and Caroline Hyde
February 16, 2023 at 4:29 PM EST

The artificial intelligence industry is experiencing a “Cambrian explosion” of applications, said Sequoia Capital Partner Sonya Huang.

Sonya Huang *Source: Bloomberg*

Source: https://www.bloomberg.com/news/articles/2023-02-16/sequoia-partner-on-ai-this-time-it-s-not-vaporware

What AI Could Mean for Society

Joseph DeAvila, *Wall Street Journal*, April 10, 2024 (excerpt)

Google Chief Executive Sundar Pichai has said AI could be **more profound than the invention of fire or electricity**.

Vinod Khosla, founder of venture-capital firm Khosla Ventures, declared last year that within 10 years, AI will take on “80% of 80% of the jobs that exist today.”

Musk, who is chief executive of Tesla and also runs his own AI company, said **AI was the fastest-advancing technology** he’s ever seen. He predicted it will probably surpass the collective intelligence of humans in five years.

“My guess is that we’ll have **AI that is smarter than any one human probably around the end of next year**,” Musk said in an interview Monday with Nicolai Tangen, CEO of Norges Bank Investment Management, Norway’s $1.6 trillion sovereign fund and one of the largest investors in Tesla. The interview was broadcast on Musk’s social-media platform X.

Source: https://www.wsj.com/tech/ai/elon-musk-and-jamie-dimons-ai-predictions-and-what-they-mean-for-the-future-of-humanity-d100b0c8

Sundar Pichai, CEO, Google

What's Next? Better GPT's and Better Foundation Models

Saqib Shah, *The Standard*, Mar 26, 2024 (excerpt)

OpenAI is reportedly gearing up to release a more powerful version of ChatGPT in the coming months.

The new AI model, known as GPT-5, is slated to arrive as soon as this summer, according to two sources in the know. Ahead of its launch, some businesses have reportedly tried out a demo of the tool, allowing them to test out its upgraded abilities.

The tech forms part of OpenAI's futuristic quest for artificial general intelligence (AGI), or systems that are smarter than humans.

In the case of GPT-4, the AI chatbot can provide human-like responses, and even recognize and generate images and speech. Its successor, GPT-5, will reportedly offer better personalization, make fewer mistakes and handle more types of content, eventually including video.

Others such as Google and Meta have released their own GPTs with their own names, all of which are known collectively as large language models.

Source: https://www.standard.co.uk/news/tech/chatgpt-5-release-date-details-openai-chatbot-b1130369.html

Learning Models for LLM's are Getting a Lot Better

This paper from Chelsea Finn's lab at Stanford proposes direct preference optimization (DPO) for aligning large language models (LLM) to human preferences without using reinforcement learning from human feedback. This significantly simplifying the training process – making LLM's far more useful for specific tasks in areas like biology.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," ***arXiv*****, Dec 13, 2023**

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call Direct Preference Optimization (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

Chelsea Finn
Assistant Professor, Computer Science and EE
Stanford University

Sources: https://arxiv.org/abs/2305.18290, https://ai.stanford.edu/~cbfinn/

What's Next? Quantum Computing

Jordan Scott, *HealthTech Magazine*, Sep 27, 2023 (excerpt)

Quantum computing harnesses the laws of quantum mechanics to solve complex problems that are can't be solved by traditional computers, including today's supercomputers, according to IBM. Quantum technology can consider numerous variables that interact with each other in complicated ways. In healthcare this offers the potential to advance precision medicine, drug discovery and diagnoses through complex analyses.

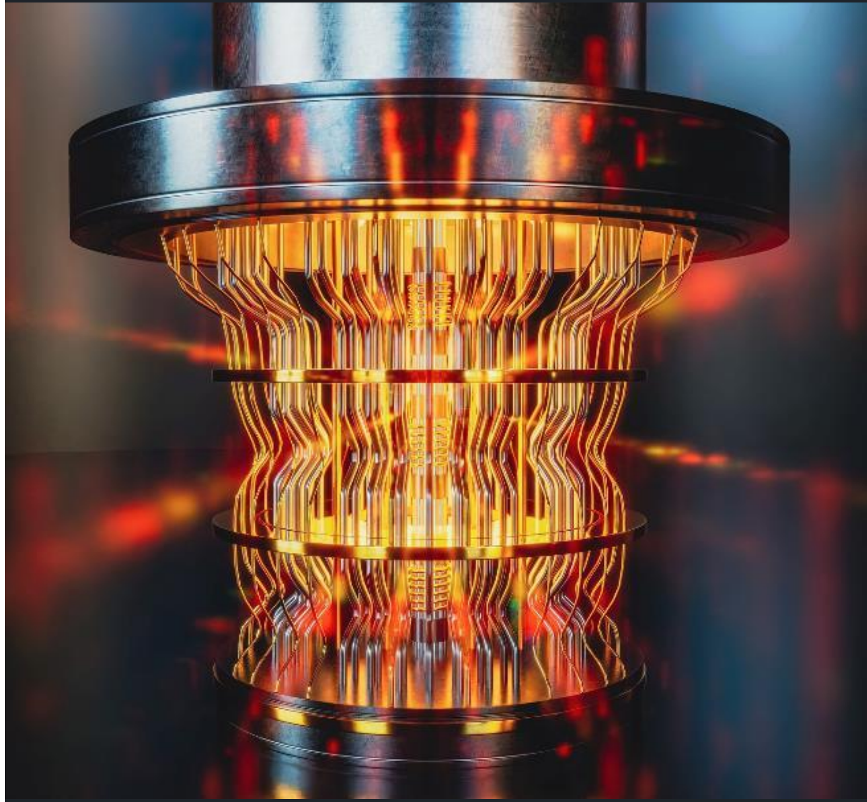
Earlier this year, Cleveland Clinic and IBM unveiled the world's first quantum computer dedicated to healthcare research onsite at Cleveland Clinic's main campus. The organization's goal is to accelerate biomedical discovery.

"Quantum computing is a unique way of doing computing that flips traditional computational principles on their head," says Dr. Lara Jehi, chief research information officer at Cleveland Clinic and executive lead of the organization's partnership with IBM. She describes the jump from traditional to quantum computers as similar to advancing from fire to lightbulbs. "It accomplishes the same purpose, but the foundation is completely different."

Traditional computing is based on deterministic computing. In essence, the 1's and 0's, or bits, translate to specific outputs. Quantum computing is based on probabilistic computing.

"Each unit of computation, or qubit, has a continuous probability of being between 0 and 1," says Jehi. "So, it can store a lot more information compared with a classical computer, and it can process computation much faster because it can do it in parallel rather than sequentially. They are fundamentally different."

Source: https://healthtechmagazine.net/how-is-quantum-computing-being-used-in-healthcare-perfcon



Recently a Google-designed quantum processor called Sycamore completed a task in 200 seconds that, by Google's estimate, would take 10,000 years on the world's fastest supercomputer. Honeywell, which once sold massive mainframes, predicts the performance of its quantum computers will grow by a factor of 10 every year for each of the next five years -- meaning they'd be 100,000 times faster in 2028.

Additional Resources to Learn More About AI and Pharma

AI Basics

DataCamp: AI for Beginners
Google's AI Video Course
Matt Turck: The AI/ML Landscape
MIT: AI Basics
Nvidia: Jensen Huang on AI
SimpliLearn: AI Bootcamp
WSJ: AI Predictions and Future of Humanity

AI in Pharma Blogs and News

Andreesen Horowitz: Bio + Health Themes
BenchSci: https://blog.benchsci.com/tag/artificial-intelligence-in-drug-discovery
Decoding Bio Blog: https://www.decodingbio.com
Endpoints AI Channel: https://endpts.com/channel/ai/
Eric Topol on AI and health: https://erictopol.substack.com/
GSK on AI: https://gsk.ai
Jacob Oppenheim Blog: https://medium.com/@jnoppenheim
Kinomoto mag: https://medium.com/kinomoto-mag
Nature Machine Intelligence: https://www.nature.com/natmachintell/
Pat Walters Blog: http://practicalcheminformatics.blogspot.com
Stat News on AI: https://www.statnews.com/topic/artificial-intelligence/

Articles on AI in Pharma

Bain: How to Scale Generative AI in Pharma
BCG: Three Steps to Value with AI
Bloomberg: World's Pharma Giants Bet on AI
Economist: AI is Taking Over Drug Development
Endpoints: AI Drugs Fall Short in the Clinic
Inc.: 1910 Genetics Story
McKinsey: Generative AI in the pharmaceutical industry
MIT Technology Review: AI is Dreaming up Drugs We've Never Seen
NEJM AI: https://ai.nejm.org/
New Biotech: AI for Life
PharmaVoice: Is AI Hype or a Real Revolution in Pharma?
Politico: AI is About to Remake the Pharma Industry
Rand Corp: Using Quantum Computers in the Life Sciences
SciLife: AI in the Pharmaceutical Industry
SoltDB: Will Supercomputers drive pharma breakthroughs?
WBUR: How Biotechnology Companies are Using AI

AI Company Trackers

AI Industry Analytics: Guide to AI companies in drug discovery
Matt Turck's: 2024 MAD (ML, AI & Data) Landscape

Expert Perspectives on AI in Pharma

Findings from Expert Interviews

As part of this review of the AI field, we spoke to eighteen experts on AI and its application in pharma. These included seven people in AI biotechs, three investors, three technologists, a few service providers and six big pharma representatives. Several interviewees agreed to be quoted by name and are cited but most preferred to stay anonymous. There are dozens of use cases in pharma from the application of AI. These range from drug discovery to technical writing. We focused our conversations almost entirely on the drug discovery and development use cases.

Key conclusions emerging from expert interviews included:

1. We Need Better Data Annotation in Biology
2. Target ID May Be More Important than Drug Discovery
3. Early Results from AI in Drug Discovery Have Been Disappointing
4. Cell Perturbation Technologies are the Real Deal
5. AI is Likely More Valuable to Biologics Discovery Than Small Molecule Discovery
6. Multimodal “Development in a Loop” Approaches Highly Promising for Drug Discovery
7. Heavy Investment is Required in Clinical Development AI/ML

Findings from Expert Interviews

Point 1: We Need Better Data Annotation in Biology

If there was one area where all parties agreed, it was on the need for better data. Machine learning is great at taking structured datasets and discerning any embedded patterns in the data that a human might miss. There was a sense that both drug discovery applications and drug development applications are not progressing as fast as one might hope due to the lack of well annotated data.

One astute observer noted that human language is so structured that the use of ChatGPT as a natural language processor of words and paragraphs found in text sources on the internet works well. Those who use ChatGPT are universally amazed by how "human" the computer can be. But this is very much linked to the structure of language itself.

In contrast, biological datasets do not necessarily come with the same natural structure and built in annotation. There are numerous cool companies emerging with an AI focus and an interest in drug discovery and development. Many of them are focused on digitizing aspects of biology including cell morphology, protein binding and the like. But these efforts are still in early days.

Point 2: Target ID May Be More Important than Drug Discovery

With the explosion in biological data, one large opportunity for AI in drug discovery involves finding new targets linked to disease. There are quite a few groups working on finding new targets. And some companies have had good results such as Insilico Medicine of Hong Kong. But, broadly speaking, efforts to find new targets have not wowed investors and other observers so far.

We heard widespread belief that while AI can be helpful in drug discovery, our industry is already pretty good in this area. In contrast, multiple interviewees felt that we humans could do a lot better at target discovery and that machine learning tools could prove to be highly valuable.

One large pharma executive explained that their company has gotten many of their small molecule groups up and running with state-of-the-art cell models ("perturbation models at scale") with full single cell transcriptomics based on gene edits combined with analysis of organoids. They are positioned obtain deep biological insights from this data. This same pharma has been reanalyzing past clinical trials with deep multiomics analysis, full molecular analysis of responders vs. nonresponders and is doing both "dry

Expert Views (continued)

lab" and "wet lab" iterative work on biology of interest that is emerging from cell models.

This executive says that AI in drug discovery is helpful but that target ID could completely change their R&D productivity.

Ultimately, if you are the first to find an interesting target and keep the finding confidential while developing drugs against the target, it can provide a major leg up in an otherwise competitive industry.

Point 3: Early Results from AI in Drug Discovery Have Been Disappointing

What rapidly becomes apparent is that billions of dollars have been spent on AI in pharma, but all parties feel that field remains quite nascent.

Several interviewees remarked that there is not a long list of drugs that have been approved due to AI deployment. It's not even clear that there is a better pipeline emerging from AI-focused drug development companies.

One interviewee said: "We're also seeing clinical failures with AI-generated drugs. Was this supposed to happen? Obviously, it remains early days and this could change in the years to come."

Several parties felt that the productivity of R&D is going to improve because of machine learning and AI techniques in time. At the same time, thus far, the technology is, at best, incremental. One point we heard made is that a large pharma may benefit more from AI than a biotech.

The point was that if you could improve the drug pipeline's success rate by a small percent (say 10%) and speed up its time to market by 10% that would be a huge win given how large big pharma pipelines are.

This person said that AI in small molecule drug discovery turned out not to be the "amazing opportunity we all thought it was". The idea that "we could solve all of these undruggable targets turned out not to work as planned". The reason was "far worse combinatorics than generally understood and intrinsic disorder in most targets." But, they noted, "suppose that there are ten properties you are optimizing for in a small molecule being designed against a known target. One can do a much better job with computation and machine learning solving this optimization problem than humans. This is a big deal for us."

Expert Views (continued)

Point 4: Cell Perturbation Technologies are the Real Deal

Interviewees shared widespread excitement about cell perturbation approaches for target discovery. Nobel Prize winner Jim Rothman has long said that we can learn a great deal about biology by studying what happens in live cells. His vision is becoming more and more of a reality every day with the advent of powerful computing and advanced machine learning techniques.

Companies like Recursion can study cells in detail and change one thing at a time about those cells and observe what happens. The effort to learn from the data gathered from cells is going to take time but appears to be quite exciting. Eikon Therapeutics has drawn particularly high interest because it can label proteins and track what happens to them after a cell sees a drug added. Over time, it will be possible to tag and track more and more analytes within a cell which will expand the power of this approach. One party noted that Vevo is exciting as it gathers masses of structured data from many live cells which gives a much wider aperture with which to apply machine learning to data generated from an intervention.

The party noted that that AI is enabling a whole new toolset to be added in between traditional in vitro drug testing and in vivo animal pharmacology studies. Because it is increasingly possible to structure data resulting from cell experiments, machine learning is a natural in this area.

Point 5: AI is Likely More Relevant to Biologics Discovery Than Small Molecule Discovery

We picked up universal excitement regarding new emerging AI biologics companies from VC's and pharma players.

Google's AlphaFold was a big breakthrough. Starting in July 2021 it became possible to solve for any protein structure if you knew the protein's sequence. This technology was not that useful but has since progressed massively. New companies that are generating high excitement include Charm Therapeutics (can solve for small molecule structures that will bind to a protein based on tech out of David Baker's lab) and Aikium which is able to generate large molecules on certain extracellular targets where mAbs have not traditionally worked. AbSci and BigHat were also often mentioned favorably.

Expert Views (continued)

One large pharma observer explained the importance of ML models in biologics discovery. They indicated that it's easy to get an antibody made with phage display the old way. But it's hard to get a really good antibody made: "You want one that binds well, that has low viscosity so it can be SubQ, that doesn't need to be in a deep freezer all the time, that is heat stable and the like. It's just a lot easier to get to this by using the computer than by designing everything the old way."

Point 6: Multimodal "Development in a Loop" Approaches Highly Promising for Drug Discovery

A typical approach to AI in drug development involves a process that begins with a target, then generates drug leads against the target, tests the drug leads against the target, gathers data from all the testing then using machine learning to infer which drug characteristics matter. This then loops back with synthesis of new drug candidates based on what's been learned. This process goes iteratively until one comes up with something really good, ideally "best-in-class" or "first-in-class" against the target.

A number of experts we spoke to indicated that while most AI companies follow this loop-based development approach, some are a lot better than others. The quality of the tools used in the loop are critical, the objective functions used in the ML optimization matter a lot and the core ML software is critical.

Views we heard expressed were that the use of multi-modal data in objective functions mattered greatly and that distinguishing tools were the use of DNA-encoded libraries, massive parallelization in measuring target engagement, automated synthesis of drug in nanoliter quantities, robotics to transport samples from the synthesizer to the equipment to measure binding and the like.

There is not consensus at all about physics-based versus quantitative SAR approaches in small molecule design. Time will tell which drug discovery methods yield the best results.

Expert Views (continued)

Point 7: Heavy Investment is Required in ML in Clinical Development

Interviewees were widely concerned that AI in drug discovery isn't going to do a lot of good unless we can speed up clinical development. Stig Hansen of Kimia, for example, said:

> "If we can't speed up clinical development all our investment in accelerating drug discovery could be diminished. Unfortunately, little has changed with clinical development. How do we overcome today's expensive, slow clinical trial process?"

Several technologists we spoke to were particularly concerned about the clinical development process as the main bottleneck slowing down the pharmaceutical industry.

The view is that the real money gets spent on clinical development and it's getting slower and less efficient all the time, particularly in disease areas where there is a lot of clinical trial activity.

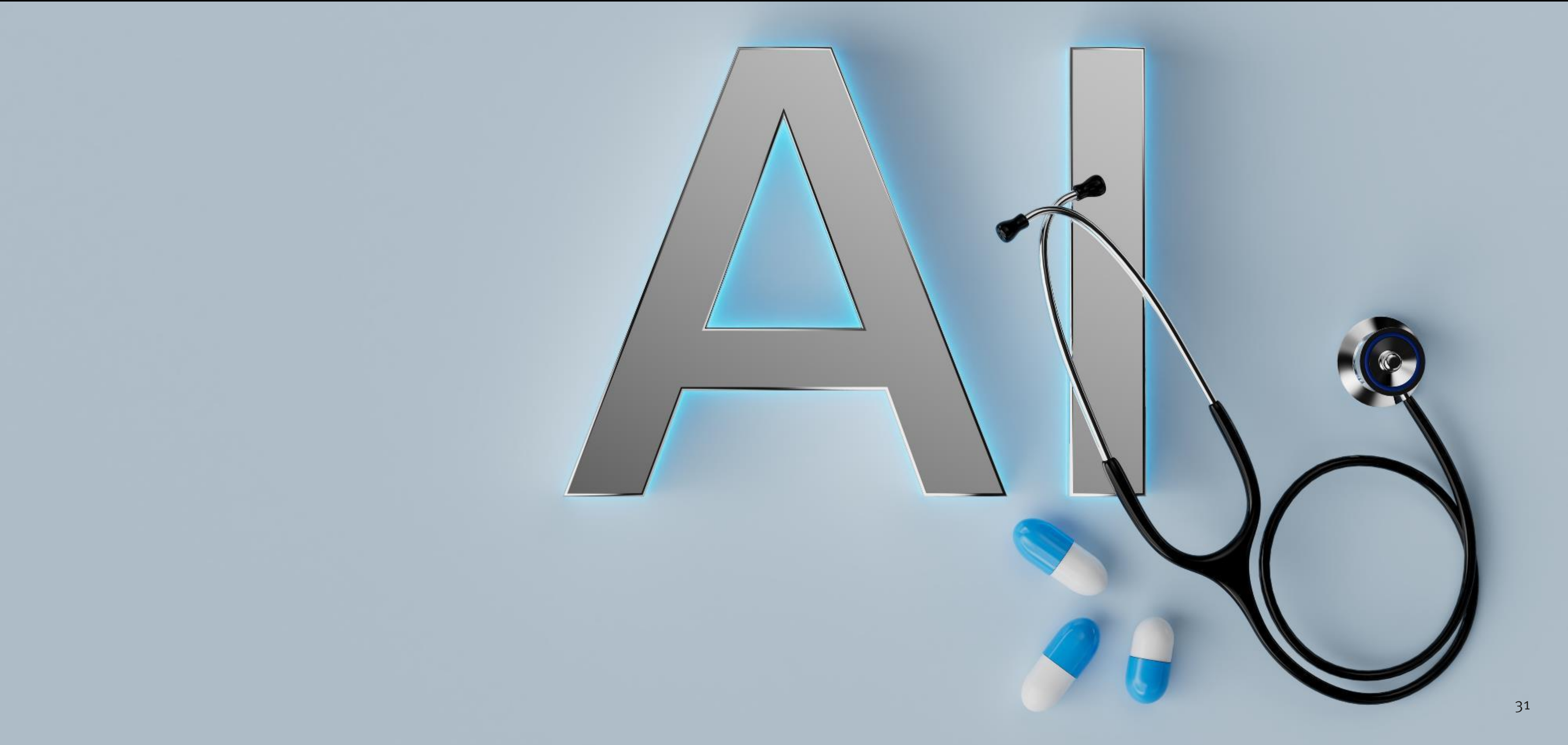
Elena Viboch of General Catalyst was more optimistic about the potential of using AI/ML to improve the clinical trial process.

General Catalyst has just invested, for example, in Paradigm with Arch. Paradigm sets up clinical trials as a care option thereby embedding study enrollment in physician workflows as opposed to requiring busy medical centers to separately hunt for patients. General Catalyst has also invested in Faro Health which uses real world evidence to allow pharmas to understand how each part of a trial protocol will speed up or slow down enrollment of a trial.

Viboch's view is that AI/ML will ultimately be quite helpful in speeding up today's highly inefficient clinical trial process and indicated that the most interesting applications of AI in clinical trials are (1) enrolling patients faster by improving the process of finding patients, (2) designing the right studies to enroll patients faster and (3) using diagnostic and patient selection strategies to have shorter trials with smaller patient counts due to larger treatment effects.

This will all time to materialize but there is obviously cause for optimism that AI/ML is going to improve the clinical stage of developing drugs.

How AI Will Impact the Pharma Industry



AI and Pharma: An Essay

In a January 2024 meeting of pharma CEO's there was broad agreement that AI has profound potential to change the pharmaceutical sector.

While virtually every company at the meeting was taking steps to understand the relevance of AI, there was also a tangible sense from many CEOs that it was important to "catch up" with rapidly changing technology. Most pharma executives don't feel ready for the future.[1]

Without doubt, ever more powerful computation combined with machine learning software can fundamentally change many aspects of the pharmaceutical industry and the broader healthcare sector.

There are countless articles on AI in drug discovery and numerous reports put out by consulting firms on the topic. Most consulting reports seem designed to create anxiety by arguing that pharmaceutical companies risk obsolescence if they don't incorporate artificial intelligence methods into their processes.

Some large pharma have made greater investments in the AI/ML area. In carrying out this review, we were very struck by the substantial organic investments in AI/ML made by AZ, Bayer, GSK and Roche. These companies are heavily focused on improving R&D productivity through investments in data infrastructure, tools to simplify the use of generative AI for their scientists, drug discovery collaborations using AI/ML and

Many pharmaceutical industry executives believe that AI advancements will profoundly change the industry.

[1] Indegene has done good work on this. See, for example, https://endpts.com/pharma-companies-arent-future-ready-as-pressure-mounts-from-ai-advances-and-ira-legislation-survey-finds/. Also see Paul Hudson of Sanofi's article on this topic: https://www.statnews.com/2023/09/08/artificial-intelligence-pharmaceutical-industry-paul-hudson-sanofi/

Tangible Progress by Pharma in AI/ML

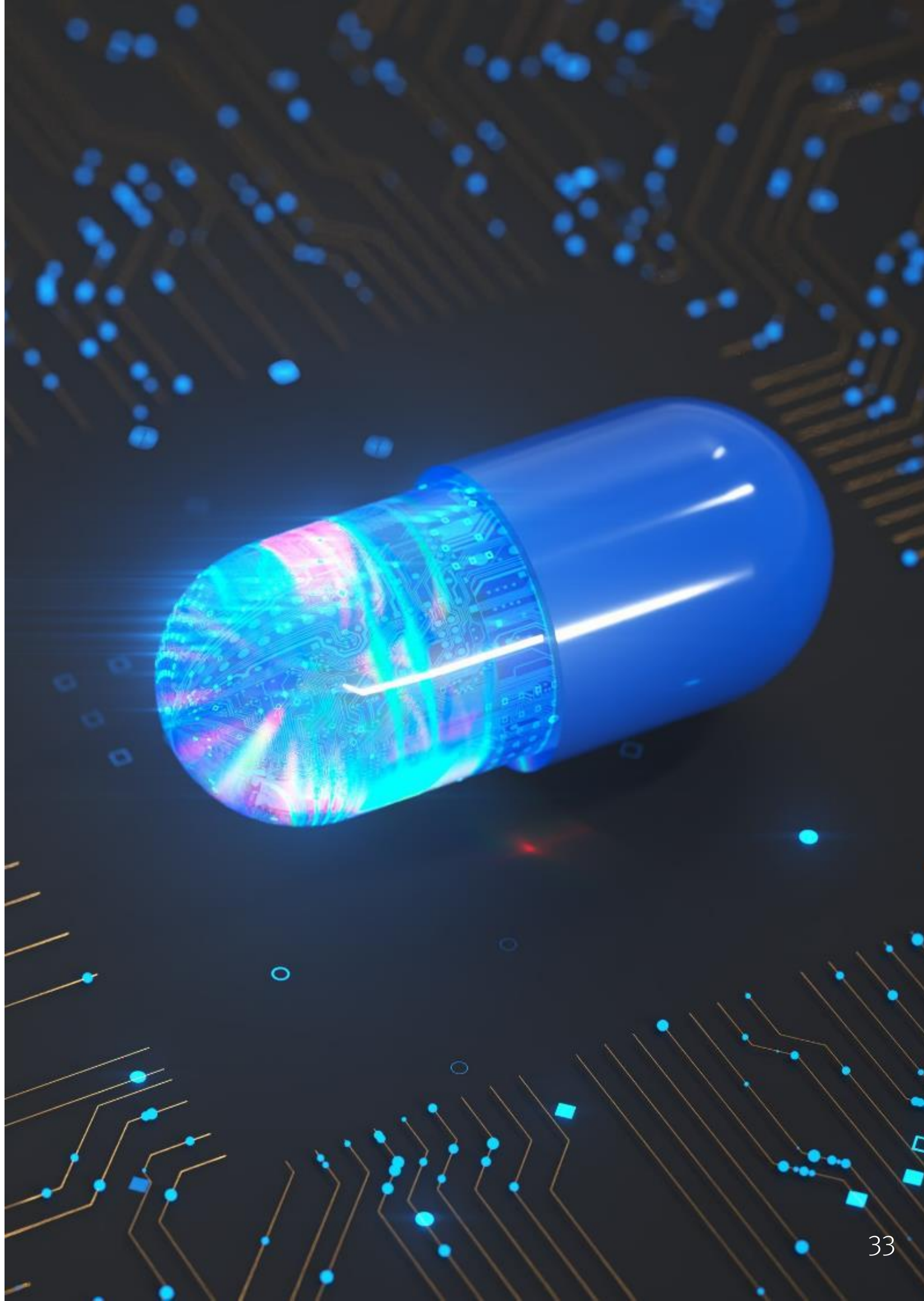
and people. A LinkedIn search revealed, for example, that GSK has over 230 employees with the phrase "artificial intelligence" or "machine learning" in their job title. GSK's website (gsk.ai) is fascinating and provides a window into their efforts to streamline target identification and drug discovery with machine learning.

We most enjoyed going through Roche's webinar held late last year on how digitalization is changing their world. The work at Genentech on informatics in drug discovery is impressive and gives one a tangible sense of the power of machine learning to accelerate drug discovery. Roche is making a meaningful investment in attacking hard problems in biology with new technologies. Their webinar highlights deep strengths across the organization in data science. Interestingly, Roche's new CEO, Thomas Schinecker, hails from the diagnostics world and comes to the job with a deep understanding of informatics. It's becoming a lot clearer why Roche's board chose him for the role, despite lacking a pharma background.

Another recent interesting discussion of AI was in Moderna's March investor event. Moderna was much more focused on AI as a clinical development tool rather than an aid to enterprise-wide productivity.

Overall, our data show a dramatic and highly tangible scale up of investment by both big pharma, venture capitalists and many scientists from academia and biotechs in the use of machine learning to improve pharma processes over the last 60 months. The nascent transformation of the biopharma industry to being one that is mastering AI/ML is very clear after going through this review. As you might imagine, there is a lot of enthusiasm for AI in healthcare. Technologists imagine that a pharma switch from "analog to digital" will yield far better insights and higher R&D productivity.

(continued)



Silicon Valley Thinking Arriving

To quote Jensen Huang, CEO of Nvidia, now the world's third most valuable company:

> *"Biology has the opportunity to be engineering not science. When something becomes engineering not science it becomes...exponentially improving, it can compound on the benefits of previous years."*

You get the idea. Once we digitize all this messy biological data the computer is going to figure how the biology works in order to arrive at better medicines more quickly. There are countless entrants into biotech investment from the tech world who use beguiling phrases like "full stack bio", "techbio", "computational biology", "biology at scale" etc., implying that we are the dawn of an explosion in innovation that will be led by Silicon Valley types.[2] The trajectory of change is forecast to wow us in the years ahead.

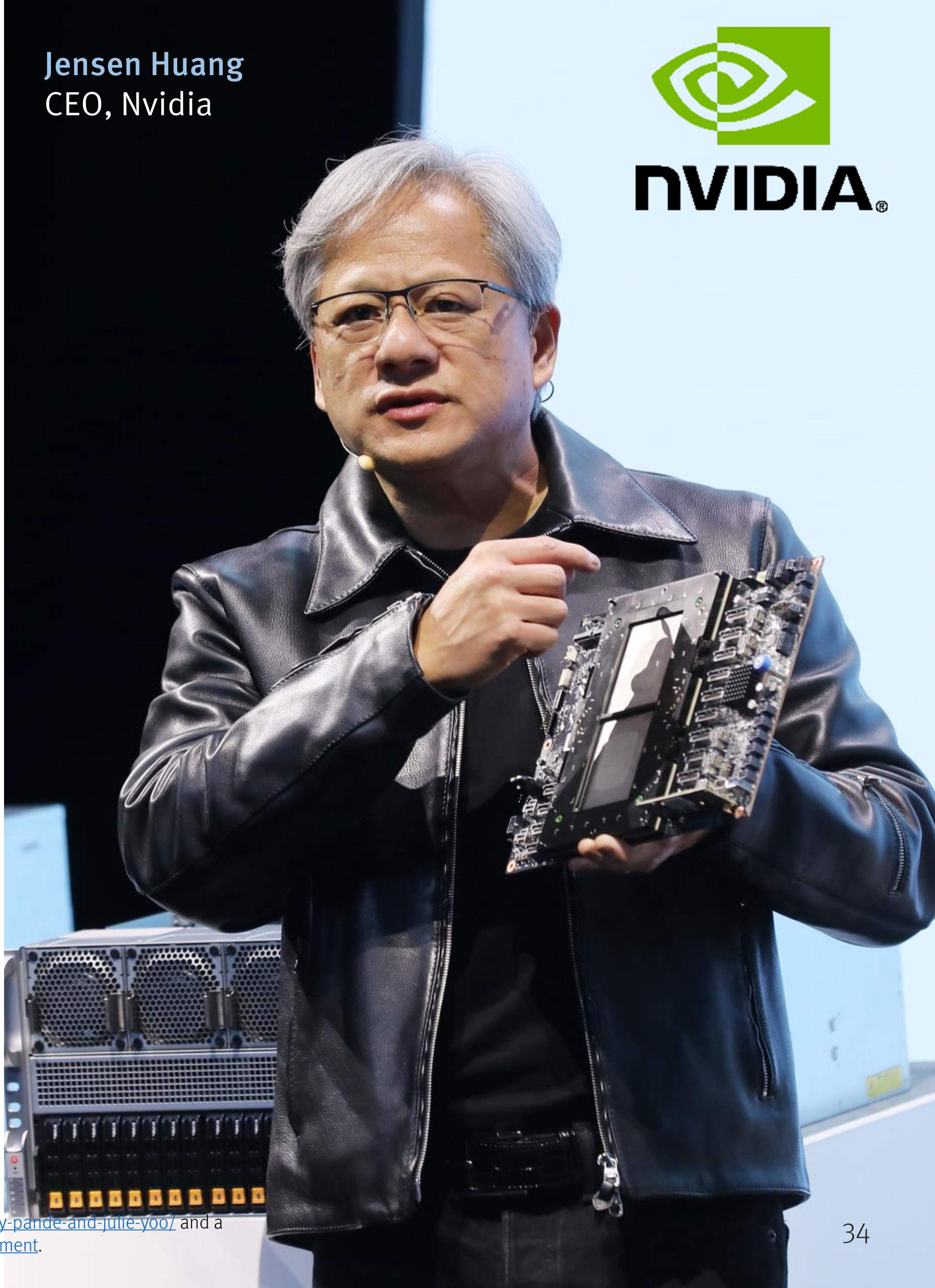
No doubt there is some truth in all of this and, perhaps, some hype has crept in as well.

The field of AI in drug discovery has its share of skeptics, some of whom are recounted in this review.

If AI is supposed to make drug discovery quicker, better and cheaper then where are all the approved drugs?

We count, for example, 24 AI-oriented biotechs that have been around for ten years or longer and 34 that have been around for eight years or longer. The 34 companies have raised $12 billion. These companies have managed to produce some solid pipeline.[3]

(continued)


Jensen Huang
CEO, Nvidia

[2] See, for example, https://www.synthego.com/blog/full-stack-genome-engineering and https://a16z.com/podcast/grand-challenges-in-healthcare-ai-with-vijay-pande-and-julie-yoo/ and a very nice survey last month in the *Economist:* https://www.economist.com/technology-quarterly/2024/03/27/artificial-intelligence-is-taking-over-drug-development.

[3] See a helpful review at https://www.biopharmatrend.com/ai-drug-discovery-pipeline/

Listening to the Skeptics

According to Deloitte it costs $2.3 billion to make a drug the old-fashioned way.[4] So, with twelve billion raised, we should be seeing something like five new drugs approved if AI was not more efficient than the old way. But it was supposed to be a lot better. Thus, the real number should be ten or, even, twenty.[5]

By this point, there should have been enough time to figure out if AI could result in improved R&D productivity and drug approvals, although truly generative AI in drug discovery is more recent and too early to evaluate. By our count, there are only *two* drugs that have been approved that used AI/ML in their development. These are TIBSOVO (credit to Schrödinger) and IDHIFA (credit to Schrödinger).[6] Each of these drugs is a *tour de force* of chemistry but none has gone on to anything near blockbuster status (at least yet). Thus, while there are some success stories the skeptics would say that the yield from AI has been disappointing overall. Further, despite the emerging pipeline there have also been multiple clinical failures with AI generated compounds.[7] Was this supposed to happen? We note the studied indifference of some pharmas, think Regeneron, towards AI. We sat on a panel this January with a Regeneron representative who said that they don't see AI as something that will change their R&D productivity in an important way. Regeneron has already assembled a multi-modal, well-annotated 2 million+ exome database that they are mining every day for insights to use in developing novel drugs.

(continued)

If AI is so amazing, then where are all those approved new drugs from it?

[4] See https://www.genengnews.com/gen-edge/the-unbearable-cost-of-drug-development-deloitte-report-shows-15-jump-in-rd-to-2-3-billion/

[5] Critics could argue that this is not fair. We haven't waited enough time to see the approvals yet and some of this investment is foundational. We'd counter that there is very little Phase 3 pipeline to look at and that getting the ultimate approvals of the pipeline will cost *even more* money. Recursion, for example, forecast a hundred drug candidates in the clinic with a decade. In reality they got four candidates there after a decade. Another ROI metric is money in/money out analysis. There have not been major M&A deals (yet) in AI. But we calculate that pharma has put $1.1 billion into AI companies via partnership upfronts. That pales in comparison to the money VC's have put in to keep these companies afloat.

[6] Moderna's ML platform was helpful in accelerating its Covid-19 vaccine development, but Moderna does not claim it as an AI victory. Similarly, a precursor to Kimia's ML platform was used to design Amgen's LUMAKRAS® but it's not clear that AI/ML was critical in the design of the molecule.

[7] See https://endpts.com/first-ai-designed-drugs-fall-short-in-the-clinic-following-years-of-hype/.

Head Scratching by Some

They just don't think that machine learning and AI is going to add that much to the insights that they are already getting.

We are sympathetic to Regeneron's perspective. Doubtlessly, others are thinking this and don't dare utter their belief given all the AI euphoria sweeping the world right now.[8]

Imagine that scientists discover a receptor tomorrow that when inhibited makes all diabetes go away. We wouldn't doubt that smart scientists at virtually every biopharma company would be able to develop drugs that would work against this receptor using existing technologies.

Perhaps AI/ML would make the right drug faster but, last we heard, most pharmas and many biotechs are quite good at small molecule discovery.

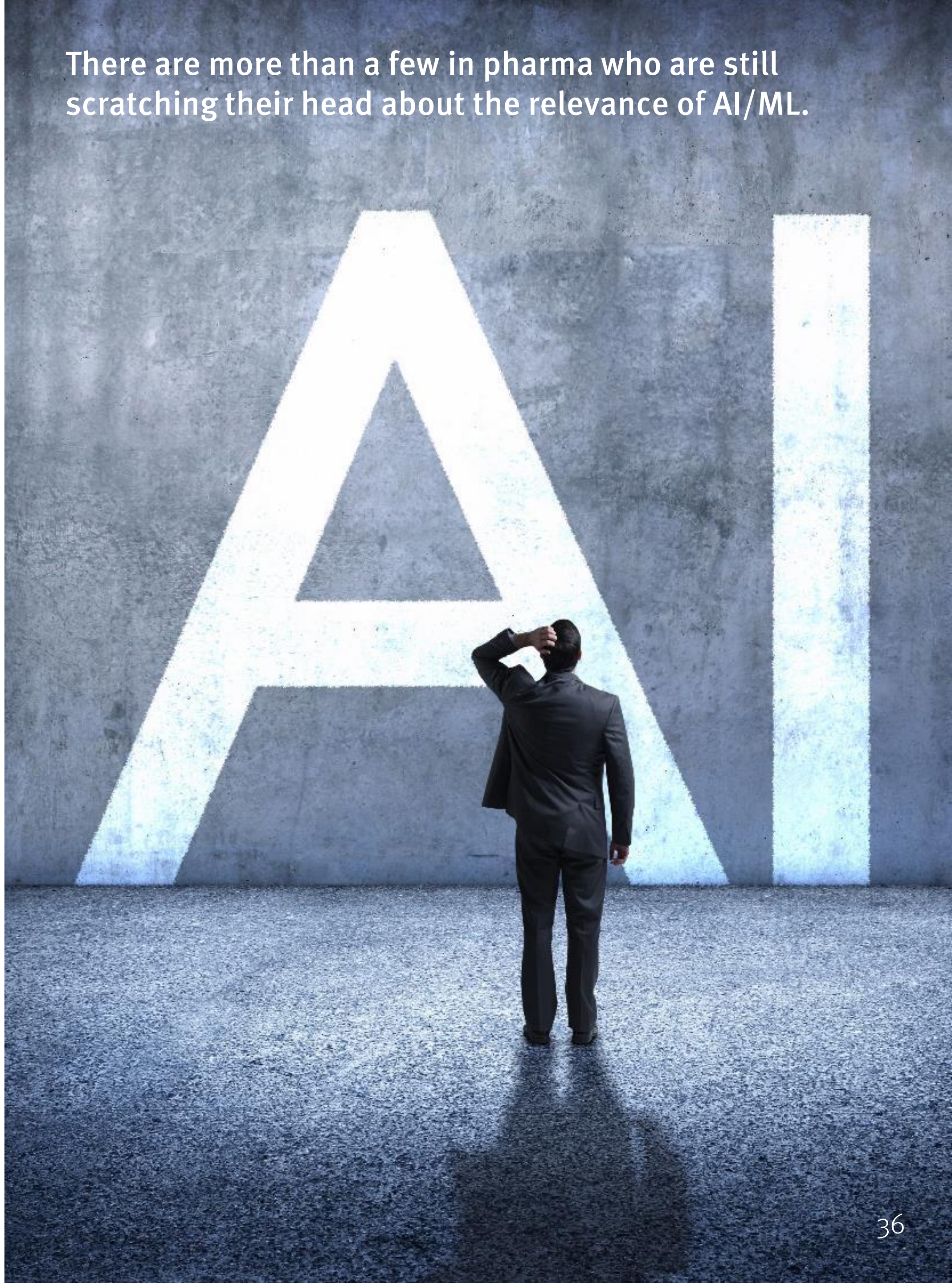
We spoke to Stig Hansen of Kimia about all of this (and reprint his thoughtful comments later in this publication). He should know as he and his team have achieved two major drug discovery victories using AI. His main point is that first generation AI efforts have been hindered by poor data and suboptimal compound search strategies.

He notes that it's often the case that current chemistry approaches are not able to design effective and safe drugs, hence the many failures of drugs in the clinic associated with off-target toxicities, poor binding and poor drug-like properties.

He is highly optimistic about the future of AI and drug discovery and would argue that AI/ML allows to get *better* drugs against targets such as that miracle diabetes receptor.

(continued)

[8] A discussion of AI implications for humanity and associated hyperbole appeared recently in the Wall Street Journal.



There are more than a few in pharma who are still scratching their head about the relevance of AI/ML.

Cause for Optimism

Hansen argues that, ultimately, AI in drug discovery will be a very big deal. It will change pharma and can be far more productive than current approaches. Others such as Daphne Koller and Eric Topol share this view.[9]

Stig Hansen argues that we will have a lot to learn and that the key is to take a focused approach that goes after important hard problems in drug design.

Our own view is threefold:

First, high-powered computing, data science and machine learning is going to become part of the woodwork in pharma. Our industry is undergoing continuous change and improvement in available research tools. AI/ML will become part of the toolkit that every drug developer uses. Developers that ignore AI risk obsolescence. To quote Jim Weatherall, VP of Data Science at AstraZeneca:

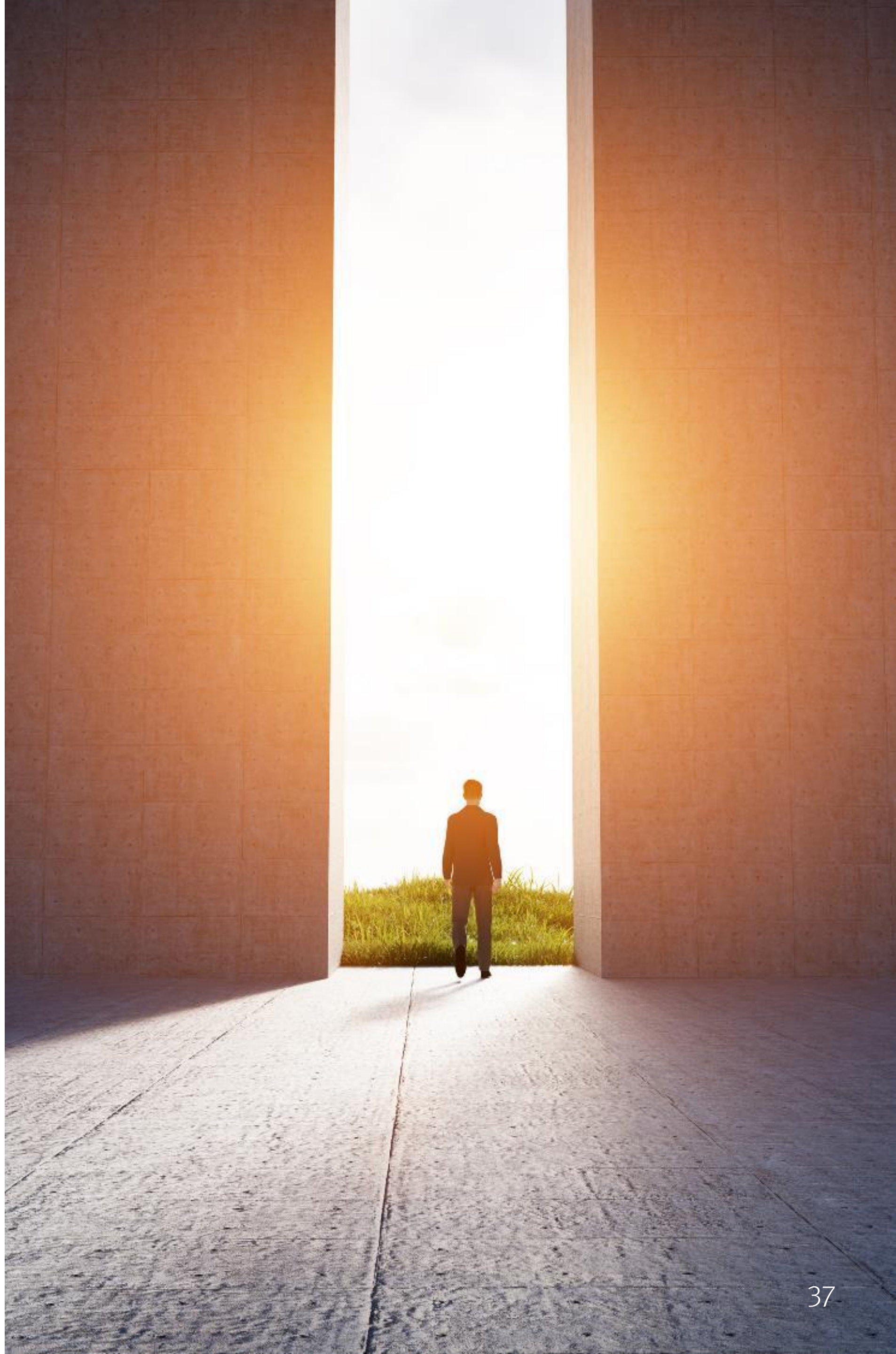
> *"AI will not replace drug hunters, but drug hunters who don't use AI will be replaced by those who do."*

Nonetheless, we view the advent of AI/ML in pharma as evolutionary – providing our industry with the tools to leverage the great datasets possible with modern genomic, single cell and microscopy datasets.

Second, the real value add is going to come from increasing our ability to go after diseases and associated targets that have henceforth been undruggable or hard to drug.

(continued)

[9] See, for example, https://erictopol.substack.com/p/daphne-koller-the-convergence-of



Raising the Bar

MYC in cancer, for example has been seen as an undruggable target. In an interview in early April, Abraham Heifets, CEO of Atomwise, argued:

> "The two ways to really change the trajectory of patient care are first-in-class or best-in-class medicines. I think we tend to overcomplicate things in this industry, but I believe the essence boils down to those two."[10]

To date, advocates for AI in drug discovery have noted the ability to accelerate approvals and reduce costs for the current way of doing things.

We have heard and read countless pitches from AI in healthcare companies and reproduce some of the best material in this review. What stands out are companies that are taking fresh approaches to tough problems. Examples of such companies include Eikon, Insitro, Isomorphic, Molecular Health, Recursion, Valo and xTalPi.

The idea of digitizing and automating cellular perturbation experiments, combining such experiments with CRISPR editing etc. is cool and is generating fresh biological insights.

(continued)

The name of the game in the future will be to use AI/ML to radically raise the bar of what's possible in biopharma industry R&D.

[10] See, https://www.drugdiscoverytrends.com/forget-efficiency-focus-on-shifting-the-standard-of-care-with-ai-in-drug-discovery/

The Future is Unpredictable

What we mean is that it's not about getting *more* drugs. The real value contribution of AI/ML would be to help our society get the *right* drugs.

Further, AI/ML will enable radical change in healthcare itself. Heifets above is getting at the issue. Why are we so focused on doing better at what we *already do* in drug discovery and healthcare with AI?

The French artist Villemard designed postcards in 1910 on what the future would look like in 2000. We have reproduced some of his illustrations at right.

What is so striking is that Villemard got some of the possibilities right. Sort of. Indeed, we would travel by air, not just by airship. We would get audiobooks, but they wouldn't be used to teach in school. OK, we never got anywhere near that makeup machine thing.

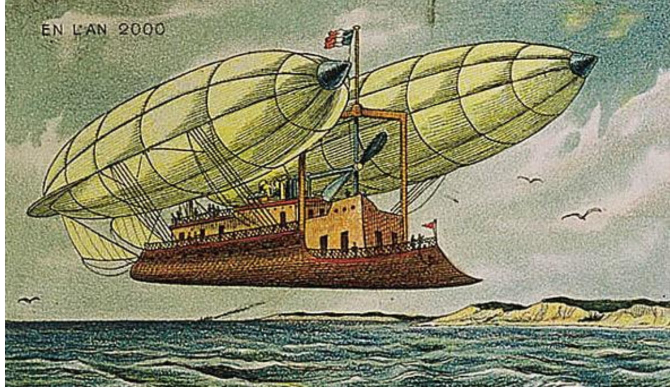
Joshua Keating writes in *Foreign Policy* that:

> "Visions of the future generally reveal more about the time in which they are created than the time they are predicting. Take for example this film of rocket pioneer Werner von Braun describing humanity's space-bound future, made during the early days of the space race."[11]

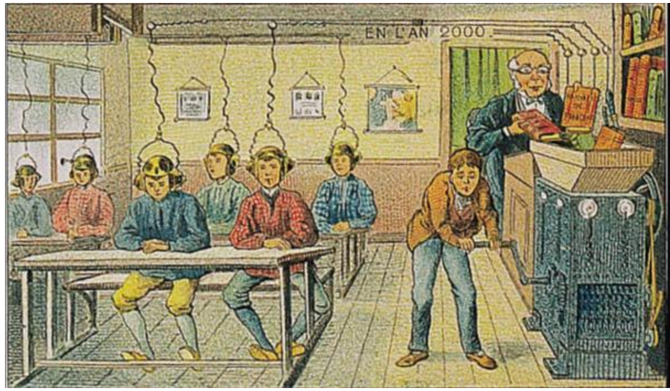
We believe that current visions of what AI/ML will mean for the biopharma industry are shaped by where we've come from and miss the larger opportunity ahead – which is to rewrite the rules for how therapeutics and healthcare can work.

One hopes that an incumbent company could be the one to envision what's possible and implement it. There is no doubt that many of the big pharma are trying to do just that.

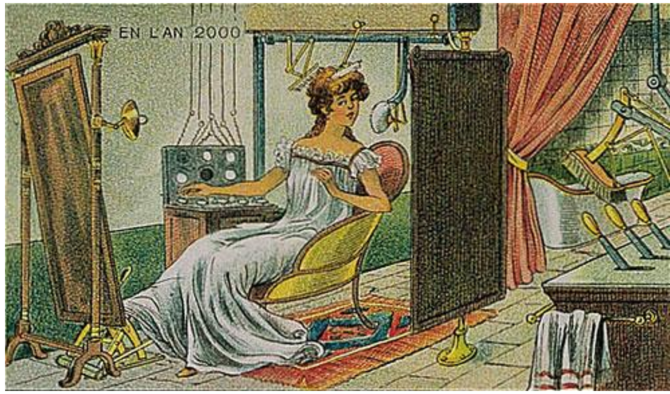
(continued)



We would travel by airship.



School would be taught by audio books



Makeup would be applied by machine

[11] See https://foreignpolicy.com/2007/09/21/why-its-hard-to-predict-the-future/

AI/ML Challenging to Incumbents

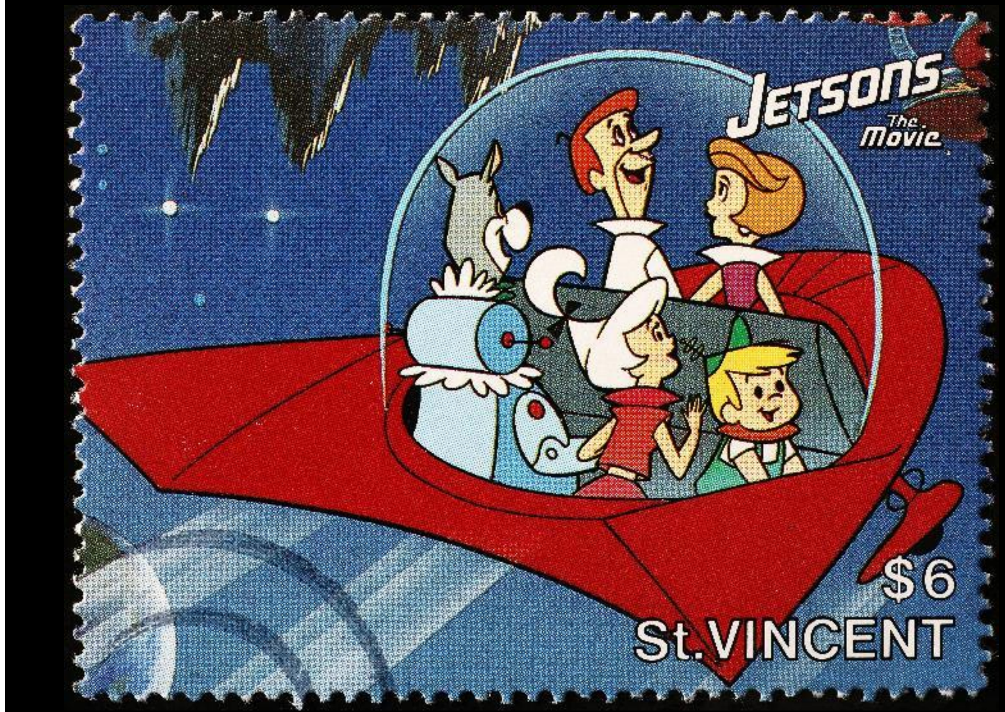
But history would suggest that it is likely to be a new company that introduces the new ideas and news ways of using AI/ML to reshape healthcare.

Daphne Koller of Insitro recently commented:

> “And we've seen that in every technological revolution that the native companies that were born in the new age move faster, incorporate the technology much more deeply into every aspect of their work, and they end up being dominant players if not the dominant player in that new world. And you could look at the internet revolution and realize that Google did not emerge from the yellow pages. Netflix did not emerge from Blockbuster, Amazon did not emerge from Walmart. Some of those incumbents did make the adjustment and are still around, some did not and are no longer around. And I think the same thing will happen with drug discovery and development where there will be a new crop of leading companies to work, I think, maybe together with some of the incumbents that were able to make the adjustment.”[12]

The real question is what ideas or approaches to pharma and healthcare that are being or could be adopted today will end up being real and relevant in, say 2050? And which gee-whiz ideas will end up on the dreaded trash heap of history – like that Jetson’s “flying car” highlighted at right?

We would start by noting that new technologies can fundamentally reshape the medium. In the same sense that Blockbuster could not envision what a streaming service might be, it’s important to think very fundamentally about what disease is, how the current



We’re still waiting to get those flying cars highlighted in episodes of the Jetsons.

[12] See https://erictopol.substack.com/p/daphne-koller-the-convergence-of

(continued)

Are We Fully Thinking Through the Possibilities?

medical system works and why do we use drugs to treat it.

To analogize, many of the earliest television programs were modified versions of well-established radio shows. Lacking ideas of what to do with TV, early broadcasts involved reading radio plays over TV.[13] It is hard to believe given the amazing content now available on modern video streaming that TV began with such uncreative applications.

When presented with a new medium we often don't know what to do.

This leads us to ask might we be able to better treat disease and maintain health using the leverage provided by better computing, machine learning techniques and improving healthcare datasets?

Our gut is that there should be a whole lot more here than just getting the computer to do billions of binding calculations between targets and ligands in hours rather than lifetimes. In the same sense that we didn't really know what to do with TV in 1940, it feels to us like we haven't begun to think about how AI is going to change medicine and the drug industry.

We encourage you to think of this question and don't be shy about sharing your thoughts with us and others.

Broadcasting a radio play at NBC, soon to be one of the largest television stations in the U.S.

Here are a few thoughts to begin the conversation. We should note that we don't have high conviction that X or Y will specifically happen. We just want to outline some possibilities.

The world of pharma is largely separated from the world of healthcare. Pharmaceutical companies build novel chemical entities that can be administered by physicians, nurses or even patients themselves to reduce the effect of a disease or even get rid of the disease entirely.

(continued)

[13] See https://historycooperative.org/the-first-tv-a-complete-history-of-television/

AI, Pharma and the Delivery of Care

Pharma companies protect themselves with patents that prohibit others from making copies of their chemical inventions and seek to profit from exploiting those patents in the market.

The factors that determine the value of a pharma company are how many good products they can invent, how long the patents last, how much you can sell a product for, the margins of the products and the quantity of product sold.

It's an interesting business because while companies can get excellent margins from temporary monopolies created by patenting, there is intense competition to get those monopolies. Hence, with few exceptions, you don't see pharma companies get valued anything like the biggest tech companies.

AI, Dimensionality and the Delivery of Healthcare

We believe that AI will have much more impact on how healthcare is delivered than on drug development.

Later in this review we talk about problem dimensionality and where computation can help. Low dimensional problem solving doesn't need computation. We humans can reason our way through most binary choices, for example.

But then there are more complicated problems with thousands but not billions of permutations. Like, what's the best way to get from my house to this restaurant on the other side of the city? This a medium-dimensional problem. We can all do this effortlessly now with Google Maps.

Not too long ago, you needed years of training to learn to navigate all the streets of London. This was called "The Knowledge" by London taxi drivers. Now, anyone can do just as well as the best London cabbie by using the computer in seconds. For free.

After this, there are high dimensional problems. Like what are all the dishes I could make with the food in my fridge and cupboard? (answer: an incomputable large number).

Diagnosing and treating patients as a doctor is a medium dimensional problem.

Delivering care and learning to be a good doctor is harder than learning to be a good London cab driver - but not that *much* harder. Decades of clinical experience help even further. To boot, doctors will often see patients where they don't know what to do and refer those out to specialists with the requisite knowledge.

Thus, it should not take long before the computer can do just as good of a job in taking care of patients as a doctor can.[14]

(continued)

[14] There is evidence that this is happening now. See https://www.mcpdigitalhealth.org/article/S2949-7612(23)00070-6/fulltext.

AI Will Completely Change the Delivery of Medicine

This turns out to be highly controversial. There are endless missives on this in places like *JAMA* and *NEJM* arguing that, at best, we are going to see computers help doctors rather than replace them.

The skeptics say that medicine is an art and that there is no way doctors will never be replaced by a computer.[15] Interestingly, every one of these anti-AI articles we've found is written by a doctor who sees AI as *augmenting* not *replacing* physician-centric care.[16]

Our own view is that this is whistling in the wind. It's sort of like saying that the cab drivers will just use those Garmin's when they get lost.

The reality is that a reasonably well-trained medical assistant will visit a patient with an iPad hooked up to a generative AI program and do a perfectly good job of diagnosing and recommending treatment for 95%+ of complaints. And, in time, this might get done without a human present.

Why on earth, then, would a health insurance company pay for the current way of providing care when this alternative approach works? Of course, this will all play out in time. We'd say, less than twenty years. Maybe five. One prominent VC has said we'll see the number of physicians used in everyday medical care cut by half or more by 2032.

We aren't trying to pick a fight here or even delve too far into this area. Our topic in this review is how AI is going to change pharma, not care delivery. So, coming back to the topic, how will AI most profoundly change pharma?

We ask you to indulge us and imaging that doctors still exist but there are a lot fewer of them in the future and, instead, there are human medical guides that are good at observing patients and analyzing them with computer help.

They also have at least average bedside manner and aren't afraid to spend time with patients.[17] These guides might have a year or less of hard-core medical training and would cost a small fraction of what doctors cost today.

We'd say pharma companies will need to change a lot as the entire medical substrate that they are grafted onto is going to change.

How and when drugs are prescribed could change. It might be possible to use fewer drugs and more alternative interventions. Instead of lobbying doctors to use your blood pressure pill, one will need to understand what is driving the computer to recommend it (or not). For what it's worth, we think the rational care algorithms of the future will be good for pharma. Ultimately, *drugs aren't being prescribed enough – not the other way around.*

(continued on next page)

[15] See, for example, https://time.com/6306922/artificial-intelligence-medicine-doctors/

[16] See, for example, https://insight.kellogg.northwestern.edu/article/will-ai-replace-doctors, https://postgraduateeducation.hms.harvard.edu/trends-medicine/how-artificial-intelligence-disrupting-medicine-what-means-physicians, https://www.economist.com/technology-quarterly/2024/03/27/can-artificial-intelligence-make-health-care-more-efficient

[17] Google's AI Chatbot for medicine was rated as having better manner than real doctors and did a better job at diagnosis in a recent test. See https://www.nature.com/articles/d41586-024-00099-4. Further, Microsoft is developing a program that listens in on patient/doctor conversations and runs ChatGPT "in reverse" to interpret patient comments and provide suggestions to the doctor by integrating large databases and patient data with what is being said to help provide a diagnosis or recommendation for next steps.

AI and Care Guidelines

The diagnostic algorithm is going to be driven by a foundational model of medicine that is going to be far smarter than current doctors.

To amplify on this idea, we have argued elsewhere that in the future there will be substantial value to be had from proprietary healthcare treatment algorithms and that these can be incorporated into wholly or partially owned treatment settings. Further, these can include proprietary pharma products. To some degree, this happens now. Fresenius, for example, owns numerous kidney dialysis centers in the U.S. Their doctors follow treatment guidelines set up by Fresenius. Fresenius also owns part of a pharma JV company called Fresenius Vifor. The effect is that Fresenius makes substantial incremental margins if guidelines incorporate pharmaceuticals from Fresenius Vifor.

Trust us. They do. You won’t find a Fresenius dialysis clinic using a competitive iron product, for example.

Our view is that AI makes it dramatically easier to design “mass customized” treatment plans for each patient. Now, with AI that uses customized treatment plans linked to customized drugs (that could be wholly or partly owned) it becomes a lot easier to do a great job for the patient.

And, when this happens, it should feed back and save money for the system and help all parties involved as positive word of mouth from the patient spreads through the system.

(continued)

Ownership of Foundation Models and Languages

What we're suggesting is that pharmas could get a lot more involved in care and could use algorithms to enhance outcomes for patients.

The economic opportunity is huge. If you could get paid an extra thousand dollars for every Type 2 diabetic in the U.S. because of your pharma guided treatment algorithm, that would work out to a $30 billion revenue stream.

All of this is Jetsons stuff. Nowhere near reality today. Flying cars sound easy in comparison.

We agree, but we think this vision of the system is ultimately feasible and would be beneficial to almost all parties in the system.

We think this type of vision would really work well in a world where machine learning, big data and high-powered computing are linked together in the service of patients.

Ownership of Foundational Models, Physiologic Languages and Operating Systems

Recall that foundational model is a large-scale summary of a bunch relationships such that If you give the model an input, it will give you a specific output.

You can go to https://chat.openai.com or https://copilot.microsoft.com/ and try out ChatGPT today which runs off a very sophisticated large language model called a GPT. A GPT is a foundational model as it is a large framework that will give a good answer for any query (if it's any good).

Foundation models have taken the world of AI by storm. These pre-trained powerhouses have revolutionized natural language processing, computer vision, and speech processing, making remarkable advancements in various domains.

With their ability to understand language, images, or multimodal data at a deep level, foundation models have paved the way for cutting-edge AI applications and accelerated development timelines.

To prepare this review we read countless PowerPoint pitch decks/PDF's, read through over 100 corporate websites of AI / techbio companies and reviewed many dozen articles. With some exceptions, few companies aspired to create foundation models for an aspect of biology. That is, we found that most companies aspired to learn a key piece of biology and then relate it to disease and then develop a drug that would impact that disease by exploiting the biological insight. Many companies were super specific. Such as "we are doing protein engineering for neurodegeneration with ML based on the principles of quantum physics".

(continued)

Few Companies are Chasing Foundation Models in Biology

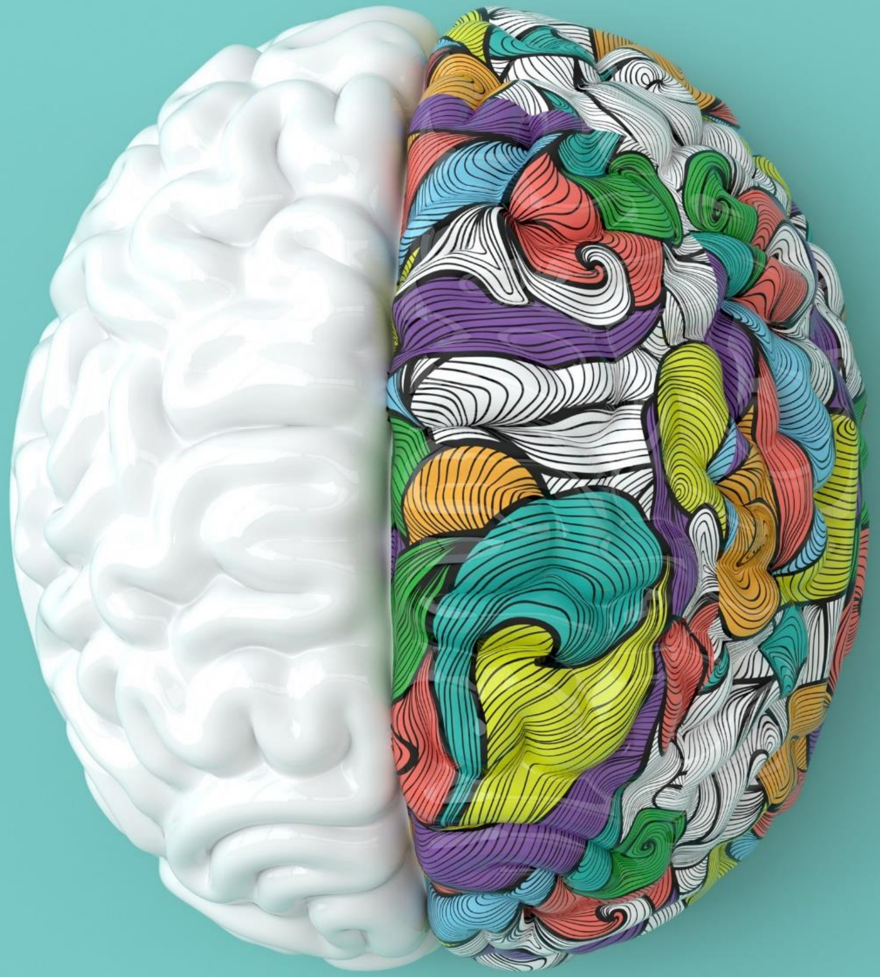
A few companies seemed to grasp this larger opportunity – which is to create a translator that would take biologic inputs from an organ system or other part of the body, deconvolute those inputs into something with meaning and then allow one to feed back into that part of the body.

A highly sophisticated biological Rosetta Stone.

An example of an interesting company we encountered is Soley Therapeutics who say:

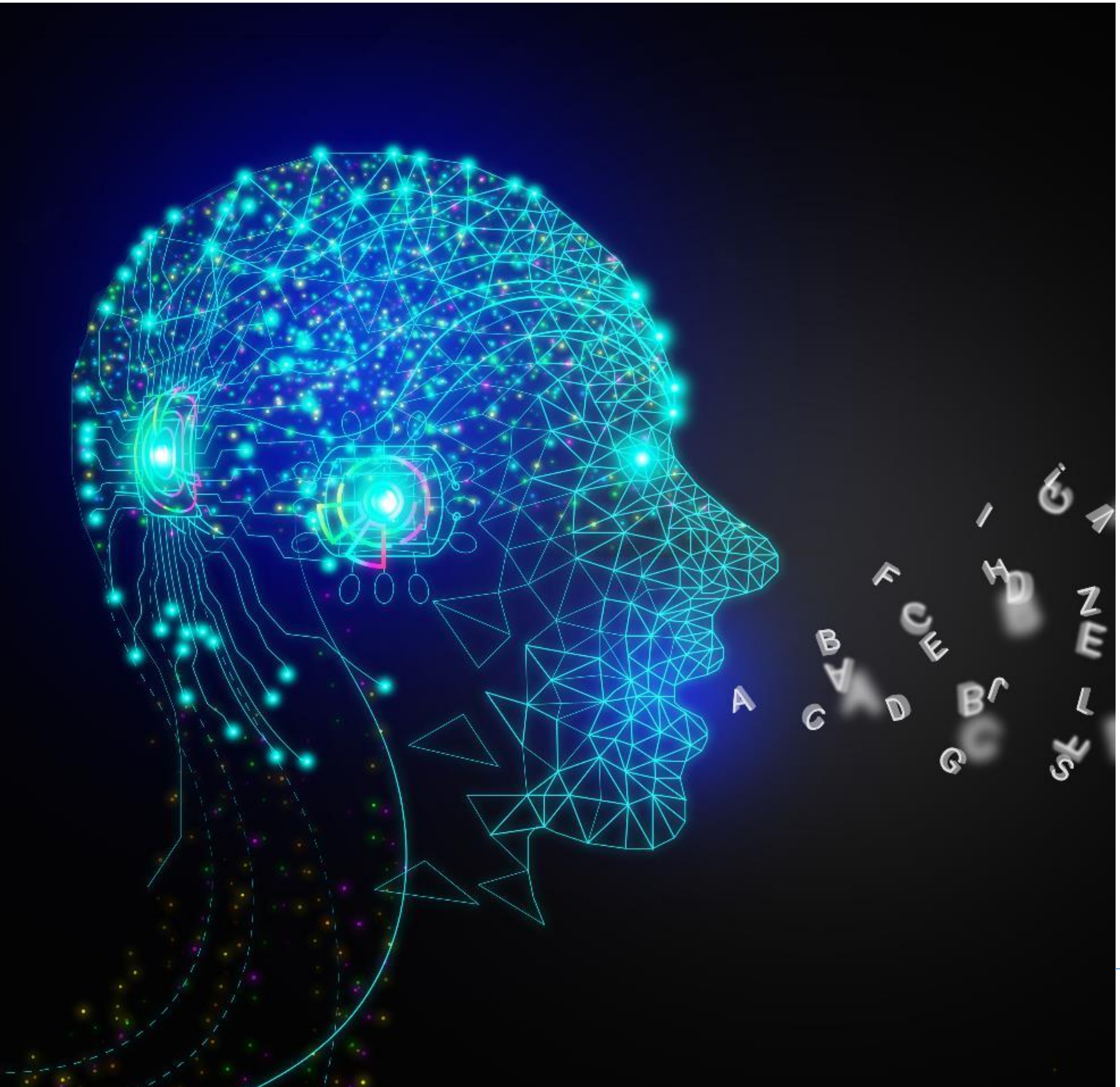
> "Cells are the most powerful computers ever created. Over billions of years of evolution, cells from vastly different organisms have learned to adapt and respond to their environment in a remarkably similar fashion. By pinpointing and interpreting these commonalities, we begin to decode the language of cells. Through our founding research, we have defined and validated a common Cellular Language across an array of diseases. As we become more fluent in this Cellular Language, we are able to prognosticate cell fate with unprecedented proficiency."[18]

Heady stuff, right? Soley is arguing, perhaps heroically, that they can create a generalized input / output model of cells. Cells, as you know signal in many ways and can be observed using microscopy and analyzed with histology.



[18] See https://soleytherapeutics.com/our-science/

AI Constructs Can Radically Shift the Pharma Business Model



Without saying it, Soley is trying to create a foundation model for human cells. Some of the more ambitious companies we reviewed seem to have similar long-term goals for aspects of the body (especially Recursion and Valo).

If it's not obvious, we are suggesting that one can break out of the entire pharma business model by building foundation models for all or parts of human physiology. Obviously, a home web appliance like Alexa takes inputs (your questions or comments) and turns them into outputs. This is, technically, a physiologic system (aural / brain) that links a living organism to a larger outside world. First generation, open-source foundation models in bio have included GenePT for DNA and BigRNA for RNA.

We wish to define a few terms here before going too much further. Again, a **foundation model** is a generalized data/analytical framework that converts inputs into outputs in a rational and predictable way. A **language**, in contrast, is a stream of signals (e.g., words) that have a specific meaning within a specific context. Normally, one thinks of a language as a human language like Mandarin.[19]

But, a language, could just as easily be a computer language (e.g., machine code for an Intel PC) that could be read and written. Languages can be **programmable**. To contrast, an **operating system** is a piece of software that interfaces with and controls the various parts of a mechanical or biologic organism.

(continued)

[19] You won't be surprised to hear that the topic of how one can take a sequenced jumble of symbols and assign meaning to them has long been studied in the field of semiotics. Best known is Charles Sanders Pierce's work on signs and semiotics dating from the 1860s (see https://www.degruyter.com/document/doi/10.1515/sem-2012-0035/html).

Biology, Foundation Models, Languages and Operating Systems

So, coming back to Soley Therapeutics, the idea of creating a foundation model that could read any cell and, even better, give feedback to that cell through any number of mechanisms would be really cool. The signals in and out can be deconvoluted and thought of as a language. And one could imagine an organism controlling all its cells through a central unit (perhaps the brain). That control would need to work through a well coordinated set of procedures that would themselves be encoded somewhere and could be called an operating system.

So, is this all Villemard-like fantasy – perhaps a great idea for what we should be doing in a hundred years?

We'd argue not.

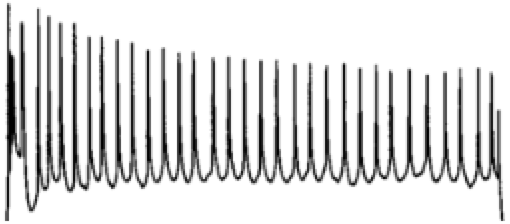
We've all seen examples of programmable cell companies.

Cells can be engineered in very specific ways using something a lot like a language. Early practitioners of this art include Autolus and ArsenalBio and a new generation of companies such as Asimov, bit.bio, GC Therapeutics and Ginkgo Bioworks are all working on far more advanced ways of programming human cells.

Perhaps the most relevant place today for this computer / body / language model is the field of bioelectronics.

The idea is straightforward. Our body is 90%+ saltwater held in cells and organs connected to the brain via electrical signals. Those signals are generated in the brain and sent to organs and cells via the central nervous system (CNS). Organs and cells can respond to the signals and send signals back to the brain.

It's not too hard to put an electrode on a nerve and get an output. It will look something like this:



A lot of electrode readings look like noise or gibberish and have no specific meaning to us. This is where AI comes in. Through well established methods of signal deconvolution, it's absolutely possible to digitize electrical signals at different frequencies. One can then use ML to associate these with physiologic data (e.g., how does the readout change when you stick a needle in someone's arm?).

Once you have figured out the meaning and can interpret these signals, it's possible to write these signals back into specific nerves in the CNS.

(continued)

Examples of Using ML to Augment Biologic Systems

Elon Musk’s company, Neuralink, has exactly this ambition. Neuralink is focused on building a generalized interface between computer and the human brain – referred to as a brain-machine-interface (BMI). The idea is to both read signals from the brain but also to feed signals back in.

This is obviously an ambitious idea and one that might take more than a little time to achieve its hoped-for potential.

But we would note that others are working on problems in this area that are much more achievable in the near term.

Some favorite companies of ours in this field are BIOS Health and Beacon Biosignals.

Beacon Biosignals is specialized in reading EEG signals from the brain and associating those with neurophysiologic states. Think depression, epilepsy, Parkinson’s etc. You might say, how can we do that since phrases like “depression” refer to broad and ill-characterized phenotypes.

We’d say: exactly! That’s why it’s so much more helpful to look at EEG patterns directly. With enough good data and ML modelling one can describe perhaps a range of depressive states and interventions that could impact them.

(continued)

Obvious Questions About Evolution of Pharma

We think the implications could be profound as ownership of a generalized language for reading and writing a major human physiologic system would be far more valuable say than full ownership of mRNA vaccine technology or an ADC platform.

It's quite interesting to see what's happening with AI and cell biology today at companies like Eikon, Insitro and Recursion. Recursion has come out and said that their purpose is to build a generalized foundation model for biology and chemistry.

These AI/ML companies speak of "multi-modal" data which can be multi-omics based, optical, electrical etc. These companies are on a path to be more than pharmas because they are building the data and tools require to make foundational biologic models.

Our view is that there is a way to evolve from today's AI/ML companies in the drug discovery field who are largely focused on generating drugs to something much bigger and more important. As in any Darwinian battle for survival, knowledge, imagination, creativity, resources and luck will be very important.

There are obvious questions here. Many, such as Daphne Koller, suggest that it will be difficult for an incumbent to adapt quickly enough to the new possibilities and that it is likely to be a new entrant that will become dominant in pharma. History is certainly on her side.

Charles Darwin.

Rapid computation, AI and machine learning are raising questions regarding the survival of incumbents in many industries including pharmaceuticals.

Looking to the Future

We would ask a different question than Koller. Will we have anything like today's world of big pharma *at all* in an AI-driven pharma world? Today, we cover over a dozen companies as "big pharmas". There is no one company that is so much bigger than the rest to be *the big pharma*. Yet, in other technology-driven industries (e.g., search engines, operating systems, online commerce), there is usually a single player or, at most, two players that come to dominate a field.

Our instinct is that a company that can build good foundation models for drugs and/or provide medicine (with drugs) via sophisticated algorithms could be in such a privileged position versus competitors. The more progress one makes on such a model the easier it gets to make the model better. The usual diseconomies of scale in pharma R&D would flip.[20] It's for this reason, that competition against social media companies and search engines has been so difficult.[21]

The implications could be profound as ownership of a generalized language for reading and writing a major physiologic system would be far more valuable say than full ownership of mRNA vaccine technology or the ADC technology.

We encourage you to think about AI's potential in pharma as something that can transform therapeutics broadly and to not focus as much on who has how much drug pipeline and whether it's any good.

We started with skepticism and end with optimism. We're hugely positive about AI/ML in biology and pharma and look forward to the next several decades of development.

[20] See https://www.baybridgebio.com/blog/rd_bigpharma_startup for a discussion of how small biotechs run circles around big pharma in R&D productivity.
[21] See, for example, https://www.youtube.com/watch?v=doIjuiwkJC4

AI in Drug Target Identification

Big Pharma R&D Business Models and Target Identification

A critical factor in R&D resource allocation goes beyond the selection of what disease to work on.

Perhaps the single most basic element of R&D strategy is what targets to work on.

It's fair to say that there are three basic approaches that pharma companies take: (1) find novel targets and develop drugs against those targets, (2) develop drugs against targets that are already known to be relevant for a disease state or (3) develop drugs without knowing the specific target (but knowing for other reasons that a drug will work).

The pace of work in industry and academic on novel target identification has accelerated massively in the last decade.

A critical aspect of the acceleration of work on target identification comes from machine learning and AI.

Initially, the main driver was emerging understanding of the genetics of disease and efforts to exploit such understanding with novel drug constructs.

More recently, a key driver has been emerging insights from the fields of transcriptomics, proteomics, epigenomics and metabolomics and disease.

For simplicity, scientists refer to these emerging disciplines (combined with genomics) as 'OMICs or multi-omics.

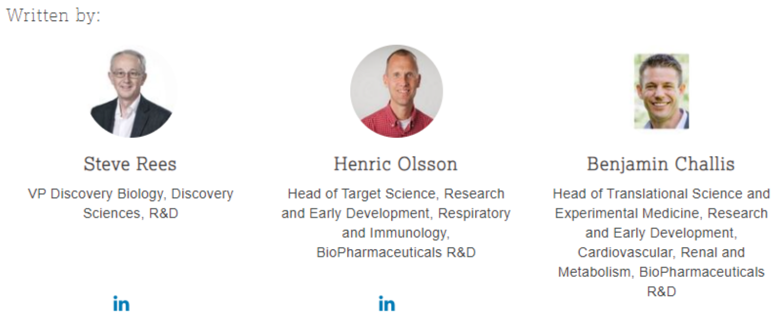
It appears to us that we are in early days of leveraging 'OMICs insights in drug development – and related areas like diagnostics.

Another critical tool is cellular perturbation analysis. How does a cell change when an aspect of its biology is changed. For example, a mutation is created using CRISPR/Cas9 editing.

AI driven cellular perturbation analysis is becoming a mainstay of the pharma industry's work on target identification.

Structuring literature searches with AI has also become increasingly important.

Target ID: Harder Than You Might Think



Written by:

Steve Rees
VP Discovery Biology, Discovery Sciences, R&D

Henric Olsson
Head of Target Science, Research and Early Development, Respiratory and Immunology, BioPharmaceuticals R&D

Benjamin Challis
Head of Translational Science and Experimental Medicine, Research and Early Development, Cardiovascular, Renal and Metabolism, BioPharmaceuticals R&D

AstraZeneca

2021

“Target identification lies at the heart of modern drug discovery. On paper, the process sounds simple - find a biological target that plays a role in disease, then find a therapeutic that interacts with it - yet this belies the complexity of the task.

The challenge of discovering and validating targets is reflected in the failure rate of drug candidates in the clinic, where promising treatments fail to show efficacy even in relatively late-stage trials. The reason for this failure is usually that the underlying hypothesis - that this drug activates or inhibits a target and modulates the disease in a particular way in a particular patient population - turns out to be wrong.

While our success rate of pipeline molecules advancing from pre-clinical investigation to completion of Phase III clinical trials is higher than industry averages, we endeavour to do better. For this reason, over the past few years we have invested in multiple technologies to help improve target discovery.

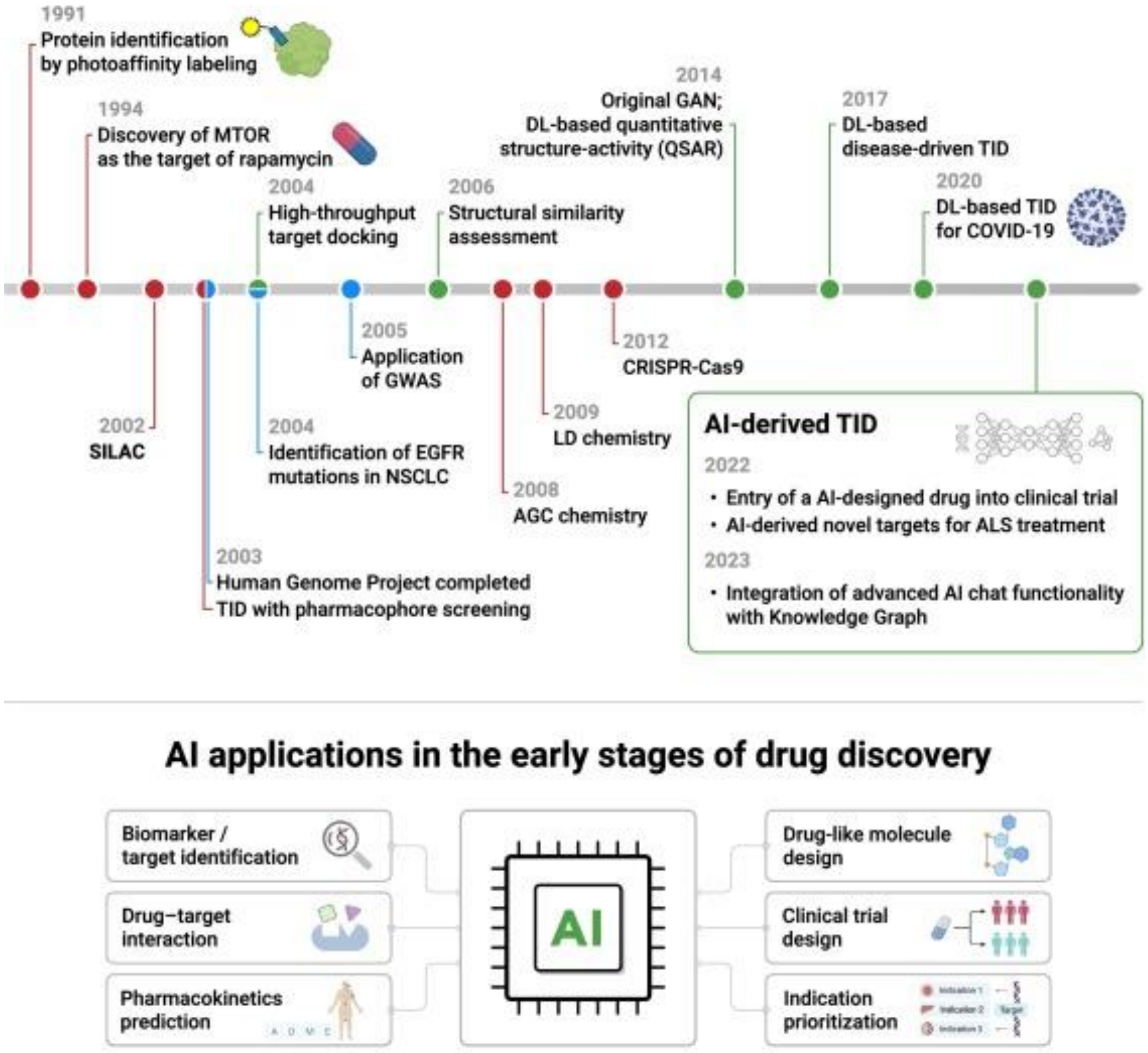
In the past, most of our drug targets have been found by combing the published scientific literature for insights into molecular pathways or genetic variants linked to disease. We’re now aiming to get ahead of the curve by focusing on the identification of original novel targets through our recent investments in genomics, functional genomics, and machine learning and artificial intelligence (ML/AI).

The journey to discovering better targets starts with building a deep understanding of biology. Increasingly, this comes from genomic insights, whether from patients and public biobanks or from tissue and tumour samples, aiming to identify genetic alterations underpinning disease.

Through our Centre for Genomics Research, we’re aiming to analyse 2 million genomes by 2026, drawn from diverse populations and covering a wide range of diseases and clinical trials.”

Source: https://www.astrazeneca.com/what-science-can-do/topics/disease-understanding/Hitting-the-bullseye-finding-better-targets-for-drug-development.html

AI / Computational Methods are Taking Over Target ID



1991
Protein identification by photoaffinity labeling
1994
Discovery of MTOR as the target of rapamycin
2002
SILAC
2003
Human Genome Project completed
TID with pharmacophore screening
2004
High-throughput target docking
2004
Identification of EGFR mutations in NSCLC
2005
Application of GWAS
2006
Structural similarity assessment
2008
AGC chemistry
2009
LD chemistry
2012
CRISPR-Cas9
2014
Original GAN;
DL-based quantitative structure-activity (QSAR)
2017
DL-based disease-driven TID
2020
DL-based TID for COVID-19
AI-derived TID
2022
• Entry of a AI-designed drug into clinical trial
• AI-derived novel targets for ALS treatment
2023
• Integration of advanced AI chat functionality with Knowledge Graph
AI applications in the early stages of drug discovery
Biomarker / target identification
Drug–target interaction
Pharmacokinetics prediction
A D M E
AI
Drug-like molecule design
Clinical trial design
Indication prioritization
Indication 1
Indication 2
Indication 3
Target

Trends in Pharmacological Sciences

Pun.et.al., "AI-Powered Therapeutic Target Discovery," *Trends in Pharmacological Sciences*, Sep 2023 (excerpt)

The emergence of artificial intelligence (AI) in early drug development.

(Upper panel) Key technological advances in the history of target identification are classified into three types: experiment-based (red), multiomic (blue), and computational (green) approaches. Traditionally, experiment-based methods have been the go-to approach for discovering therapeutic targets. However, with the rise of big data, integrated analysis of multiomic data has become a more efficient strategy for target identification. In addition, recent advances in AI-driven biological analysis have identified novel targets and AI-designed drugs are now entering clinical trials. (Lower panel) AI applications in the early stages of drug discovery. Abbreviations: AGC chemistry, affinity-guided catalyst chemistry; ALS, amyotrophic lateral sclerosis; DL, deep learning; EGFR, epidermal growth factor receptor; GAN, generative adversarial network; GWAS, genome-wide association study; LD chemistry, ligand-directed chemistry; MTOR, mammalian target of rapamycin; NSCLC, non-small cell lung cancer; SILAC, stable isotope labeling with amino acids in cell culture; TID, target identification.

Source: https://www.cell.com/trends/pharmacological-sciences/fulltext/S0165-6147(23)00137-2

Selected Biotechs Using AI for Target Discovery

CELL-BASED TARGET DISCOVERY

AITIA, ALTO NEUROSCIENCE, Atomic AI, boltzmann, celsius

CytoReason, deep genomics, DRUG FARM, EMPRESS

eikon therapeutics, engine Biosciences, Exscientia, healx, Hexagon Bio

Insilico Medicine, insitro, InveniAI Innovate with Intelligence, OCTANT, PHENOMIC AI, PATHOS

RECURSION, TransitionBio, VERGE genomics, Vesalius Therapeutics

PROTEOMIC TARGET DISCOVERY

AITIA, compugen FROM CODE TO CURE, congruence TX, CytoReason

DRUG FARM, eikon therapeutics, engine Biosciences, Health Outlook

healx, immunai, Insilico Medicine, insitro, Juvena THERAPEUTICS

nucleai, Protai, REZO, RECURSION

TransitionBio, VERGE genomics, Vesalius Therapeutics, Valo

31

GENETIC TARGET DISCOVERY

DRUG FARM

celsius, eikon therapeutics, EVAXION, Exscientia

immunai, insitro, Juvena THERAPEUTICS, RECURSION, REZO

SOLEY THERAPEUTICS, SPRING, TransitionBio, Valo, VERGE genomics

Source: Stifel Research

LITERATURE BASED TARGET DISCOVERY

AITIA, BenevolentAI, CytoReason

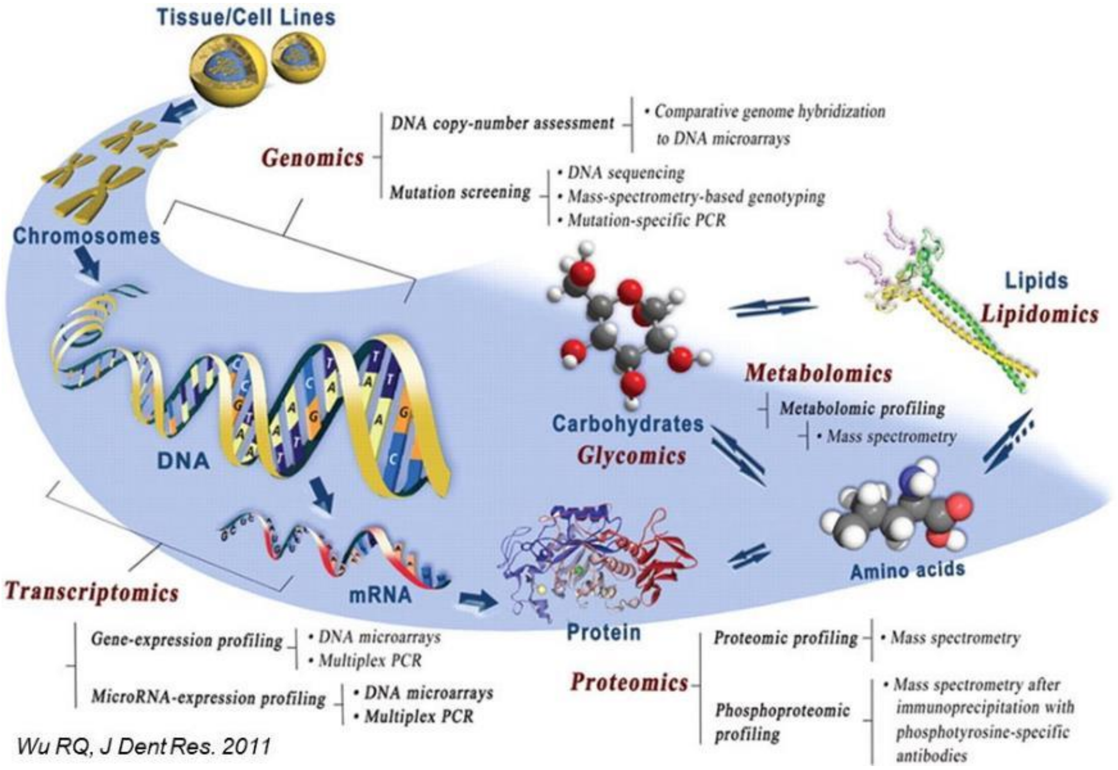
DRUG FARM, healx, InveniAI Innovate with Intelligence

LanternPharma, MOLECULAR HEALTH, Valo

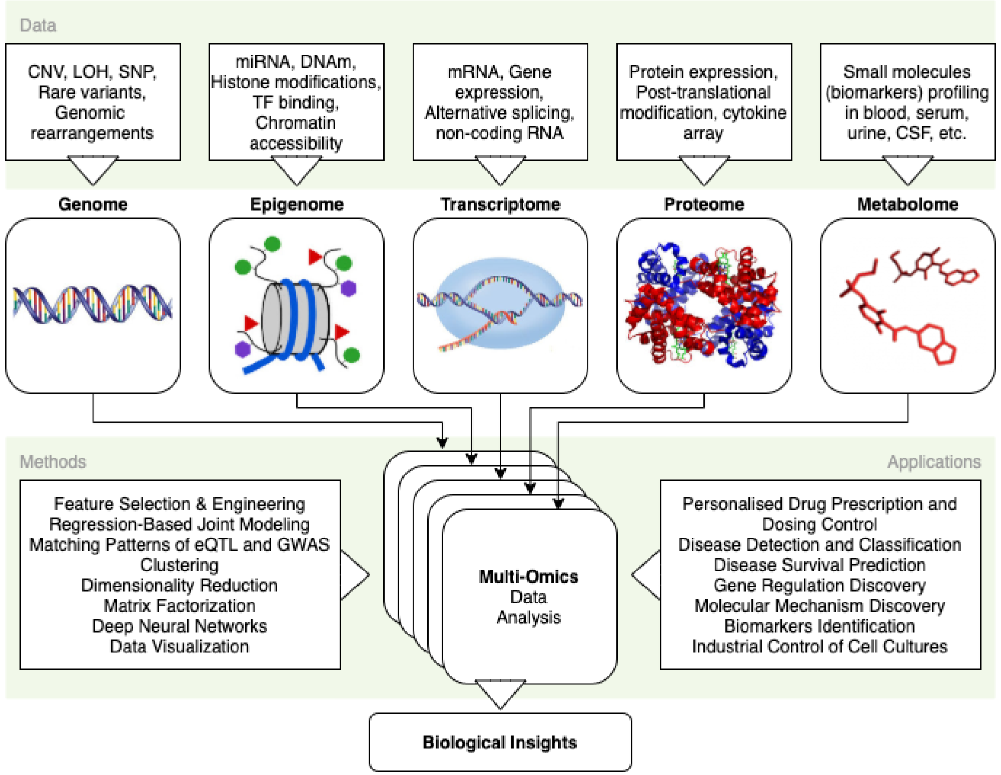
Subsection

Genomics, Proteomics and Related Approaches to Target ID

‘OMICS Technologies Are One of the Main Methods Used to Obtain Biological Insights Today



Wu RQ, J Dent Res. 2011



Data

Genome	Epigenome	Transcriptome	Proteome	Metabolome
CNV, LOH, SNP, Rare variants, Genomic rearrangements	miRNA, DNAm, Histone modifications, TF binding, Chromatin accessibility	mRNA, Gene expression, Alternative splicing, non-coding RNA	Protein expression, Post-translational modification, cytokine array	Small molecules (biomarkers) profiling in blood, serum, urine, CSF, etc.

Methods

Feature Selection & Engineering
Regression-Based Joint Modeling
Matching Patterns of eQTL and GWAS
Clustering
Dimensionality Reduction
Matrix Factorization
Deep Neural Networks
Data Visualization

Multi-Omics Data Analysis

Applications

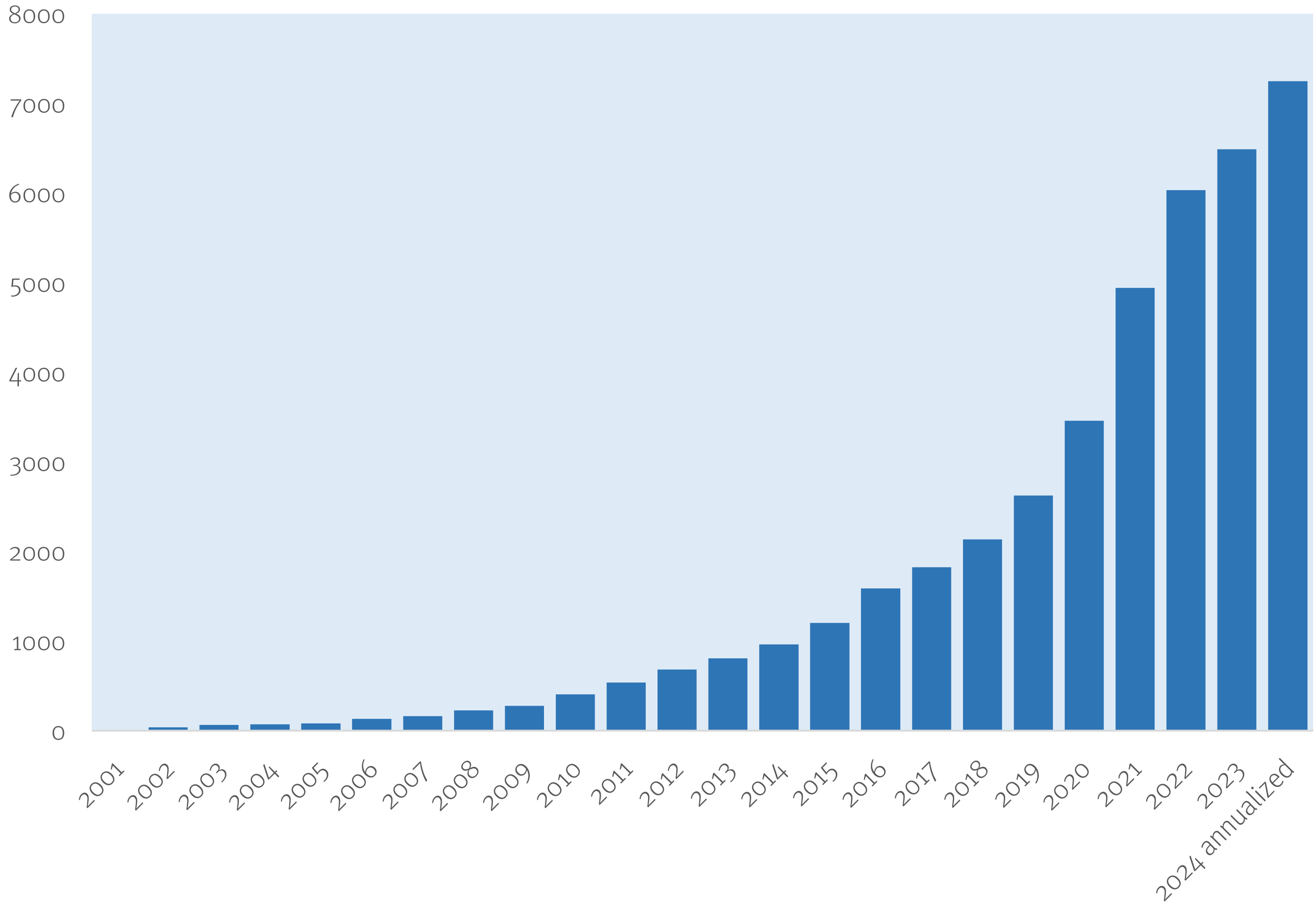
Personalised Drug Prescription and Dosing Control
Disease Detection and Classification
Disease Survival Prediction
Gene Regulation Discovery
Molecular Mechanism Discovery
Biomarkers Identification
Industrial Control of Cell Cultures

Biological Insights

Sources: https://www.sciencedirect.com/book/9780124159556/soil-microbiology-ecology-and-biochemistry, https://pharmafeatures.com/insights-on-the-latest-developments-on-oncology-with-dr-laurent-audoly-co-founder-ceo-and-chairman-of-the-board-of-parthenon-therapeutics/

Explosion in ‘OMICS and ML Research in Last Decade

Publications in Pubmed Mentioning the Word 'Omics, 2001 to 2024



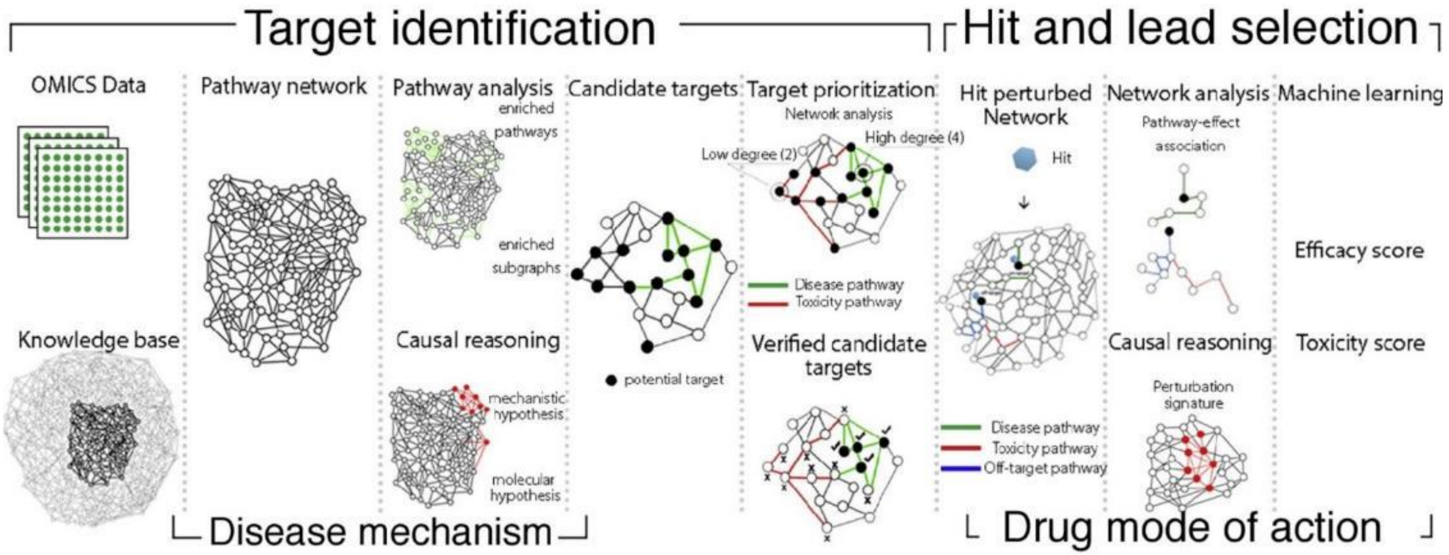
Source: Pubmed

1. Explosion in ‘OMICS research and disease insights over the last two decades (chart at left)
2. Over half of recent publications mention AI or deep learning and ‘Omics
3. OMICS factors (particularly proteins and metabolites) have major insights for medicine as the association of proteins with phenotypes often reveals relevant mechanistic information.
4. Non-genetic differences in protein and metabolite expression can account for vastly more variation in mortality.
5. It’s not that genes don’t matter. Rather, its that proteins / metabolites are ever changing so they give you a good picture of disease.

‘OMICS Data Needs to be Combined with Biological Analysis and Pathway Analysis to Get to Drug Targets

Fotis C, Antoranz A, Hatziavramidis D, Sakellaropoulos T, Alexopoulos LG. Network-based technologies for early drug discovery. *Drug Discov Today*. 2018 Mar;23(3):626-635.

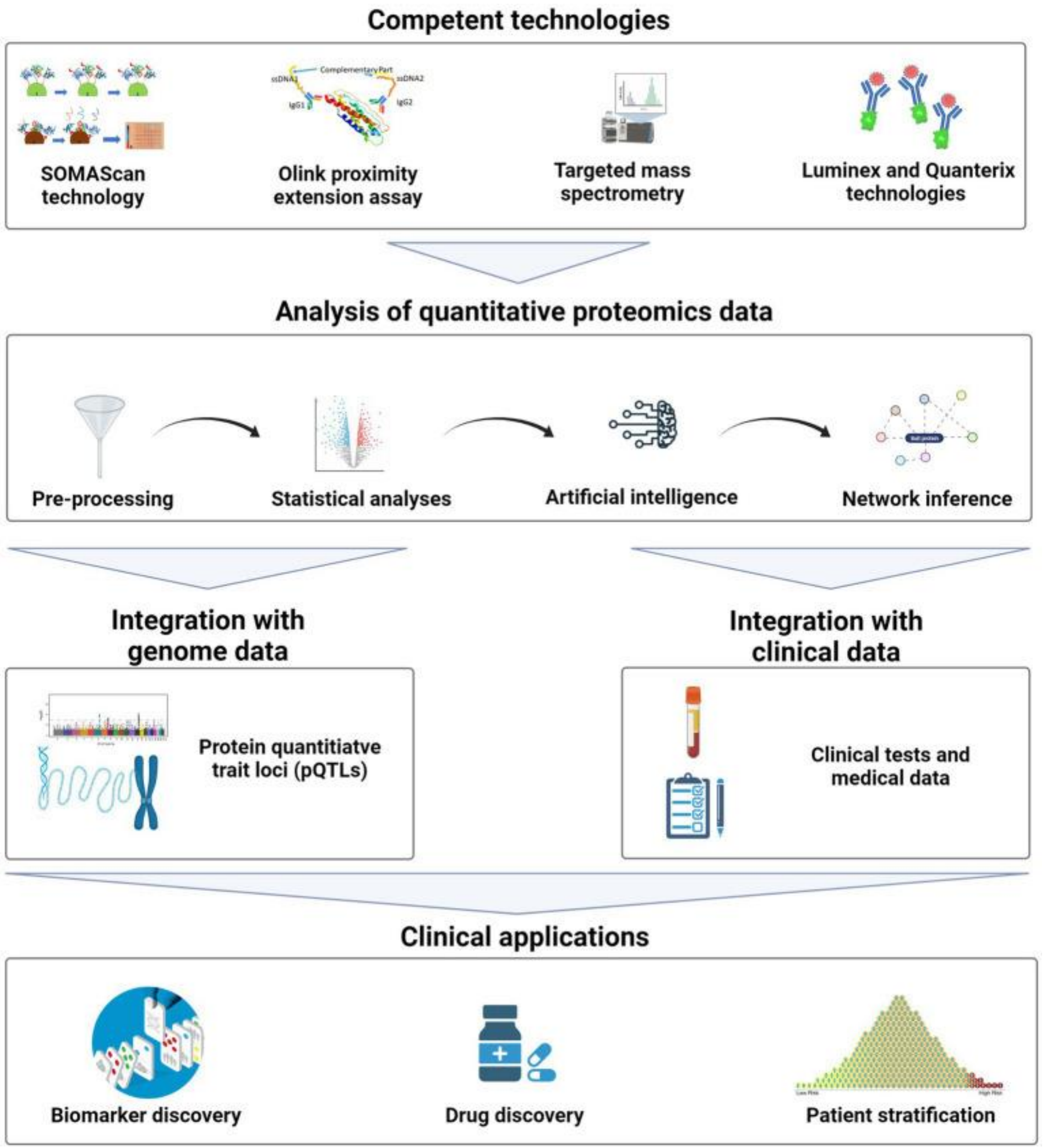
Although the traditional drug discovery approach has led to the development of many successful drugs, the attrition rates remain high. Recent advances in systems-oriented approaches (systems-biology and/or pharmacology) and ‘omics technologies has led to a plethora of new computational tools that promise to enable a more-informed and successful implementation of the reductionist, one drug for one target for one disease, approach. These tools, based on biomolecular pathways and interaction networks, offer a systematic approach to unravel the mechanism(s) of a disease and link them to the chemical space and network footprint of a drug. Drug discovery can draw upon this holistic approach to identify the most-promising targets and compounds during the early phases of development.



Source: https://www.sciencedirect.com/science/article/abs/pii/S1359644617305949

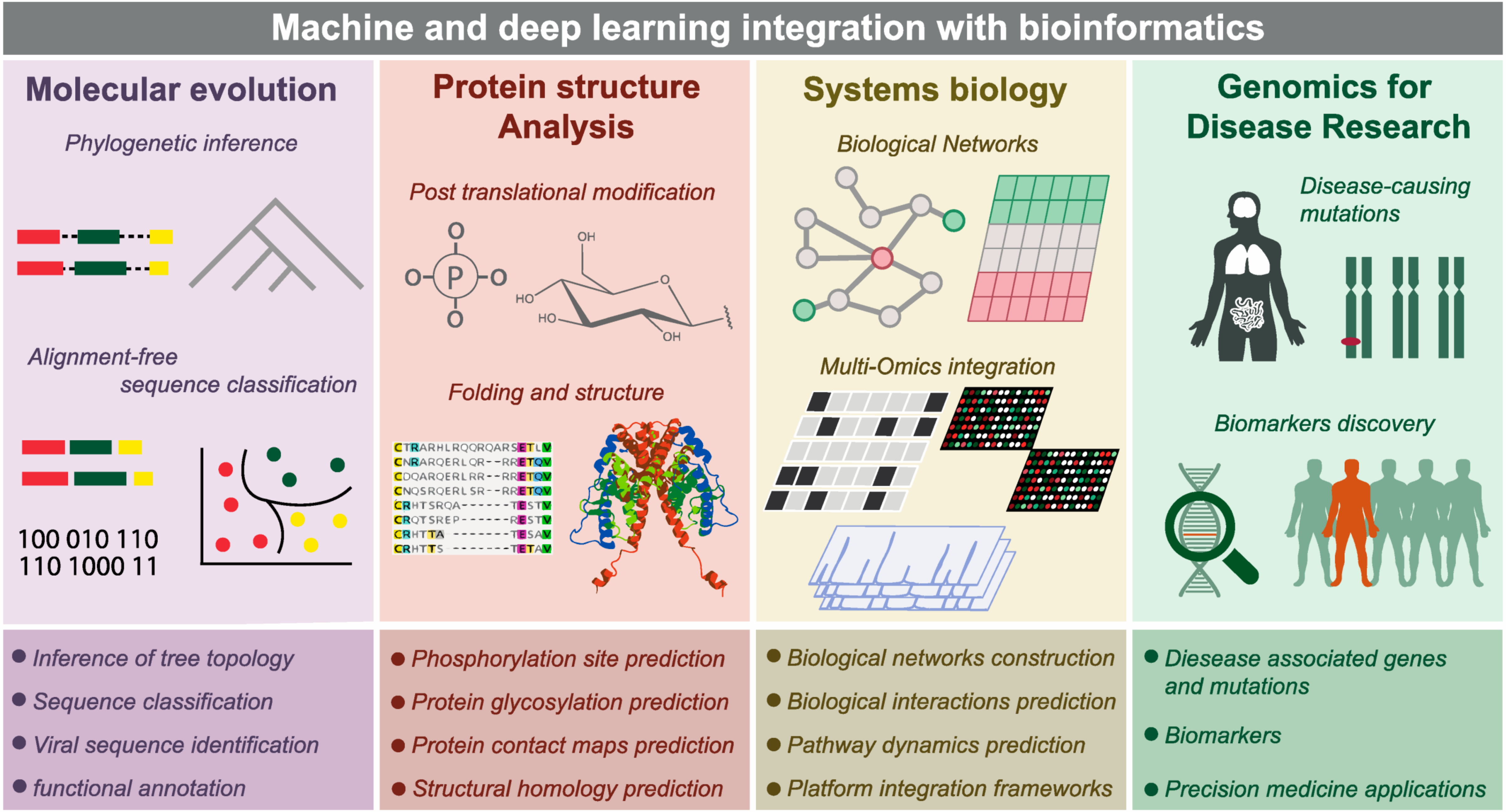
Typical Workflow Used to Integrate Proteomic Data with Genetic and Phenotype Data for Insights

General workflow for quantitative proteomics. The figure describes the different types of targeted technologies, and the common methodologies to analyze quantitative proteomics data. These analyses potentially provide clinical applications in biomarker and drug discovery and patient stratification.



Competent technologies
SOMAScan technology
Complementary Part
ssDNA1
ssDNA2
IgG1
IgG2
Olink proximity extension assay
Targeted mass spectrometry
Luminex and Quanterix technologies
Analysis of quantitative proteomics data
Pre-processing
Statistical analyses
Artificial intelligence
Network inference
Integration with genome data
Protein quantitiatve trait loci (pQTLs)
Integration with clinical data
Clinical tests and medical data
Clinical applications
Biomarker discovery
Drug discovery
Patient stratification

Source: Correa Rojo A, Heylen D, Aerts J, Thas O, Hooyberghs J, Ertaylan G, Valkenborg D. Towards Building a Quantitative Proteomics Toolbox in Precision Medicine: A Mini-Review. *Front Physiol.* 2021 Aug 26;12:723510.



Machine and deep learning integration with bioinformatics

Molecular evolution

Phylogenetic inference

Alignment-free sequence classification

- *Inference of tree topology*
- *Sequence classification*
- *Viral sequence identification*
- *functional annotation*

Protein structure Analysis

Post translational modification

Folding and structure

- *Phosphorylation site prediction*
- *Protein glycosylation prediction*
- *Protein contact maps prediction*
- *Structural homology prediction*

Systems biology

Biological Networks

Multi-Omics integration

- *Biological networks construction*
- *Biological interactions prediction*
- *Pathway dynamics prediction*
- *Platform integration frameworks*

Genomics for Disease Research

Disease-causing mutations

Biomarkers discovery

- *Diesease associated genes and mutations*
- *Biomarkers*
- *Precision medicine applications*

Source: https://www.mdpi.com/1422-0067/22/6/2903

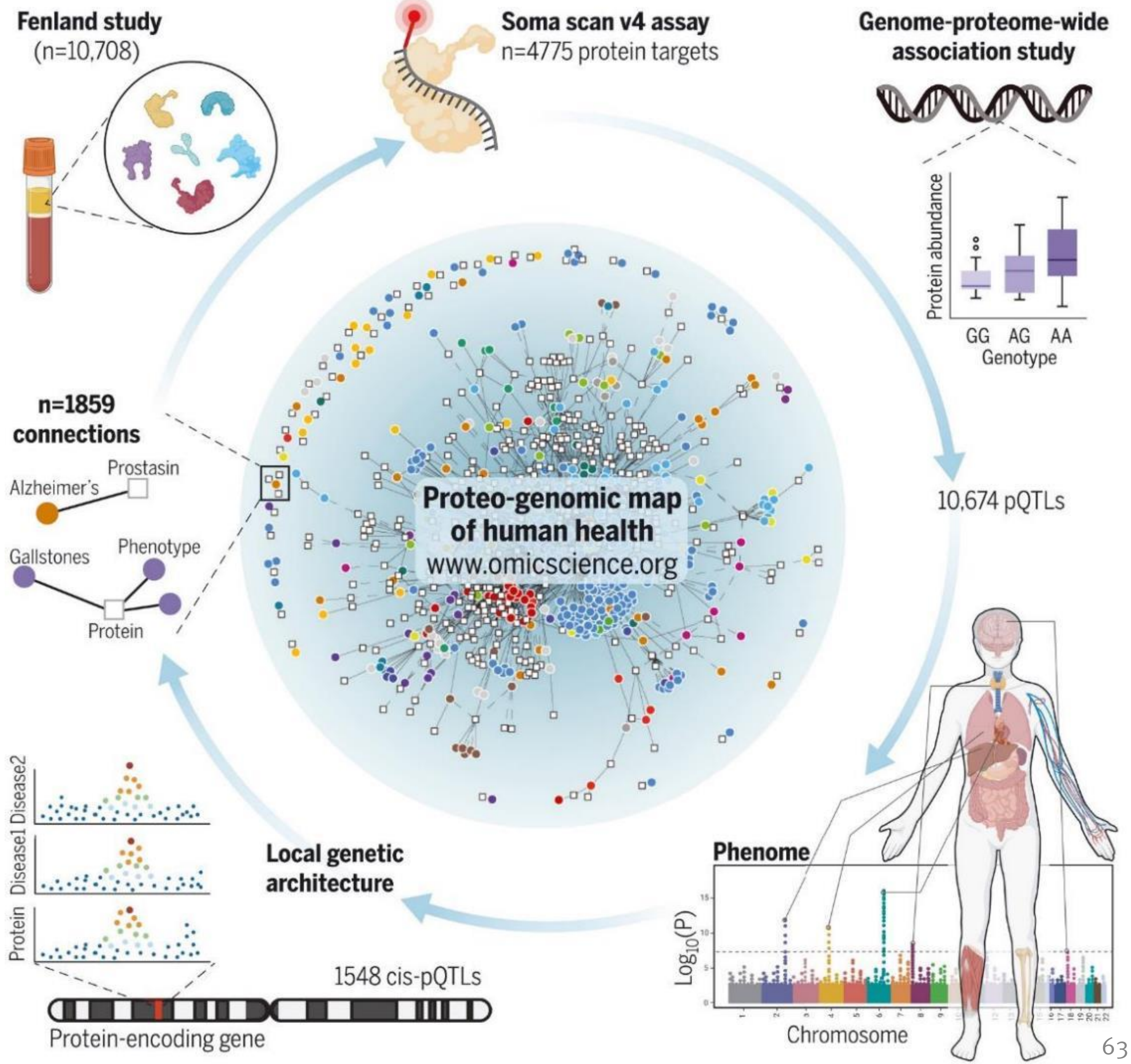
Illustrative Target ID with Omics

High Throughput Studies of pQTL's, Proteins and Phenotype (Pietzner Paper)

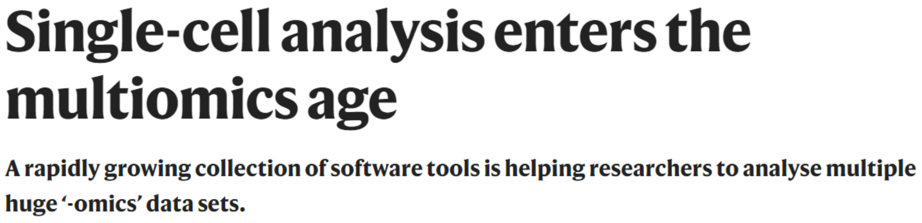
"Many diseases are at least partially due to genetic causes that are not always understood or targetable with specific treatments. To provide insight into the biology of various human diseases as well as potential leads for therapeutic development, Pietzner *et al.* undertook detailed, genome-wide proteogenomic mapping. The authors analyzed thousands of connections between potential disease-associated mutations, specific proteins, and medical conditions, thereby providing a detailed map for use by future researchers.

They also supplied some examples in which they applied their approach to medical contexts as varied as connective tissue disorders, gallstones, and COVID-19 infections, sometimes even identifying single genes that play roles in multiple clinical scenarios."

Source: Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, Wörheide MA, Oerton E, Cook J, Stewart ID, Kerrison ND, Luan J, Raffler J, Arnold M, Arlt W, O'Rahilly S, Kastenmüller G, Gamazon ER, Hingorani AD, Scott RA, Wareham NJ, Langenberg C. Mapping the proteo-genomic convergence of human diseases. Science. 2021 Nov 12;374(6569):eabj1541. doi: 10.1126/science.abj1541. Epub 2021 Nov 12



Single Cell Proteomic, Genomic and Transcriptomic Data Used to Create a Spatial Map of What's Going in to Glean Biology Insights



Single-cell analysis enters the multiomics age

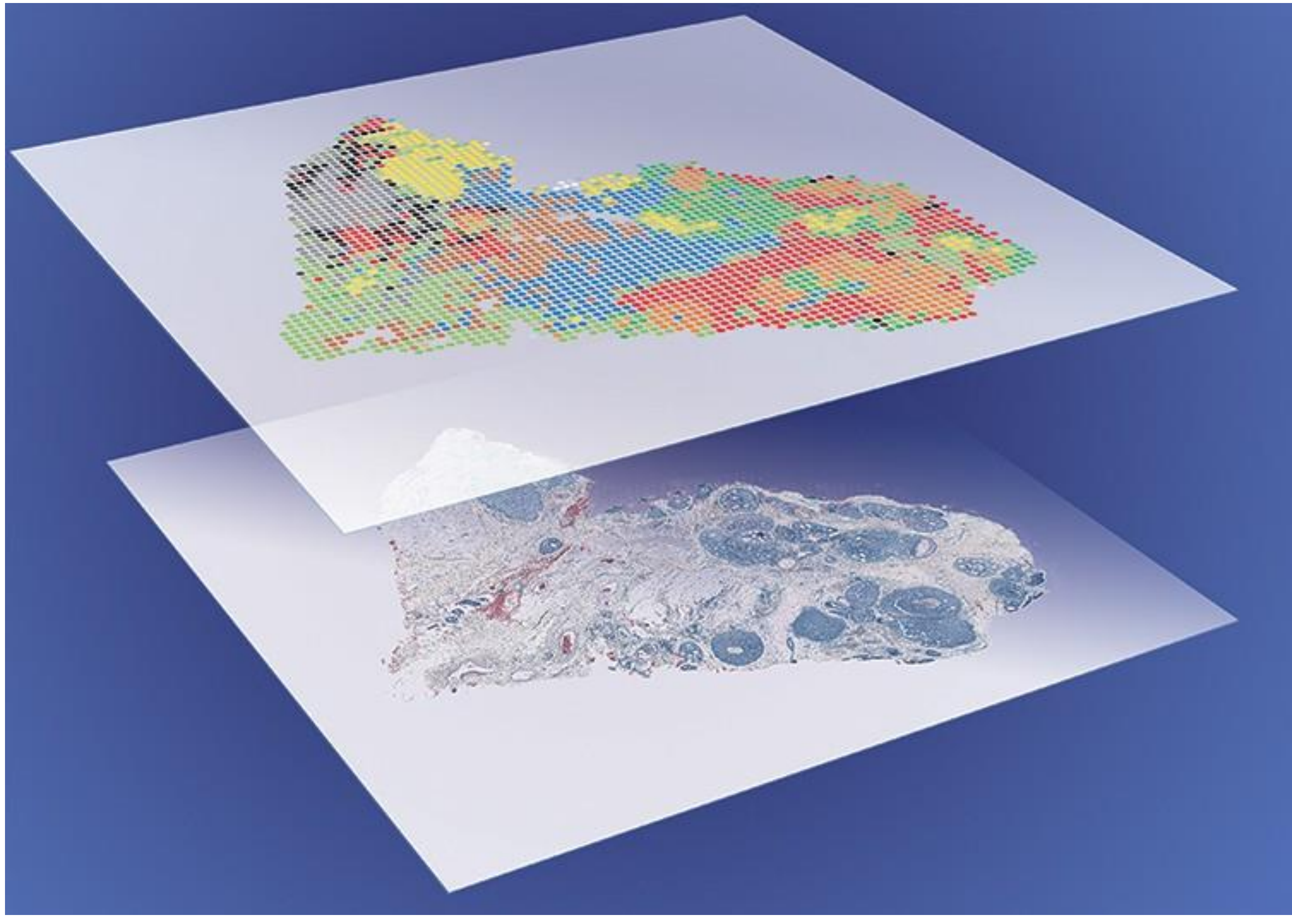
A rapidly growing collection of software tools is helping researchers to analyse multiple huge '-omics' data sets.

Jeffrey Perkel, *Nature*, July 19, 2021

"The past decade has witnessed an explosion in single-cell genomics. Single-cell RNA sequencing (RNA-seq), which profiles gene expression, is the most common technique. Other methods detail processes such as methylation, genetic variation, protein abundance and chromatin accessibility.

Now, researchers are increasingly combining these methods — and the resulting layers of data — in 'multiomics' experiments. Argelaguet, for instance, combined gene-expression profiling, methylation and chromatin accessibility in a technique called scNMT-seq. Another technique, CITE-seq, profiles both transcription and protein abundance. And G&T-seq captures both genomic DNA and RNA.

Whatever the acronym, all these techniques aim to glean complex biological insights that might be undetectable using any single method. But the task is computationally challenging, and making sense of the resulting data even more so. A fast-growing suite of software tools can help."



Illustrative Multi-omic Spatial Analysis Using 10X Genomics Technology

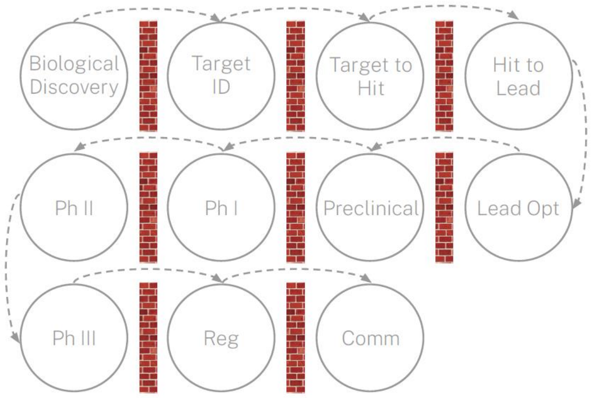
Source: https://www.nature.com/articles/d41586-021-01994-w

Integrating ‘OMICs with AI: Valo Health

Valo

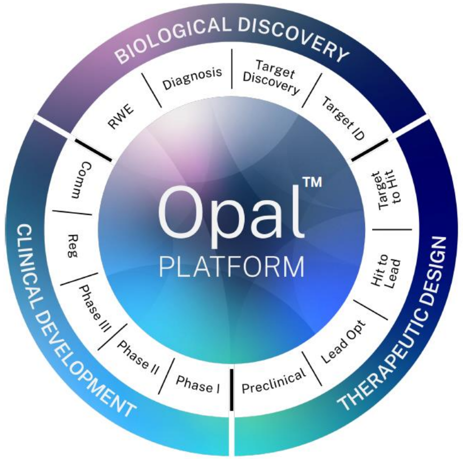
Valo Health is a technology company built to transform the drug discovery and development process using human-centric data and artificial intelligence-driven computation. Valo aims to fully integrate human-centric data across the entire drug development life cycle into a single unified architecture, thereby accelerating the discovery and development of life-changing drugs while simultaneously reducing costs, time, and failure rates. The company’s Opal Computational Platform™ is an integrated set of capabilities designed to transform data into valuable insights that may accelerate discoveries and enable Valo to advance a robust pipeline of programs across cardiovascular, metabolic, renal, oncology, and neurodegenerative disease.

LEGACY BIOPHARMA MODEL[1,2]



LOCALIZED[3] **| DISINTEGRATED**[3]
SURROGATE-DEPENDENT[4] **| SERIAL**[1]

VALO DRUG ACCELERATION MODEL



UNIFIED | INTEGRATED
HUMAN-CENTRIC | PARALLEL

Target ID = Target Identification; RWE = Real World Evidence; Lead Opt = Lead Optimization; Reg = Regulatory; Comm = Commercial; AI = Artificial Intelligence
[1] Paul, Steven M., et al. "How to improve R&D productivity: the pharmaceutical industry's grand challenge." Nat Rev Drug Discov 9, 203–214 (Mar 2010). [2]Hughes, James P., et al. "Principles of Early Drug Discovery." *British Journal of Pharmacology* 162.6, 1239-1249 (Mar 2011). [3] Konersmann, Todd., et al. "Innovating R&D with the Cloud: Business Transformation Could Require Cloud-Enabled Ecosystems, and Services." Deloitte

Multidimensional -’omics

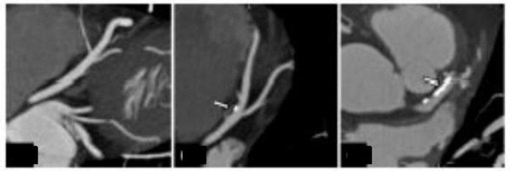
Exclusive access to one of the largest prospective studies spanning pan-omics, imaging, and medical records

>22.5T
Whole genome sequencing data points

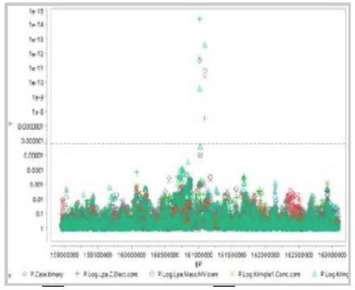
>210M
mRNA sequencing data points

>21M
Metabolomic and/or proteomic data points

>320K
Blood sample aliquots



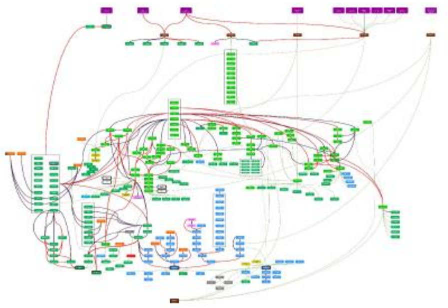
>13K images
paired with related scoring data



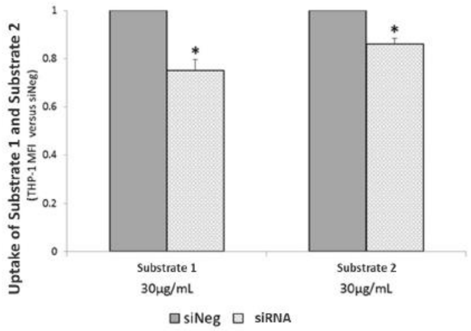
Source: https://valohealth.com

Aitia's Gemini Digital Twins successfully predict clinical outcomes, validating their use in target discovery and clinical trial design

Discovery of a Novel Target in Atherosclerotic Disease



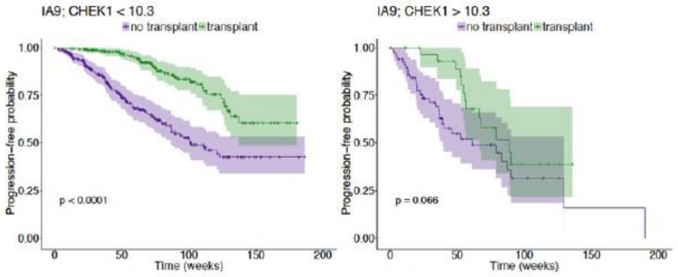
Discovered a **specific receptor** that eliminates larger Lp(a) particles which have been linked to Atherosclerosis



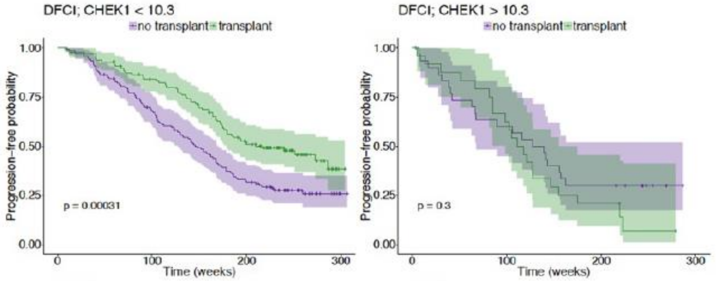
Circulating levels of **large Lp(a),** not small Lp(a), were associated with the **gene encoding the novel receptor** and was **validated by siRNA knockdowns**

Part of the research effort tied to the discovery of biomarkers driving atherosclerosis (Voros et al., 2023)

Discovery and Validation of Novel Driver of Response to SCT for Multiple Myeloma Extending PFS by 20 months



CHEK1 as the **top driver** predicting progression-free survival **(PFS) benefit from Stem Cell Therapy (SCT)**

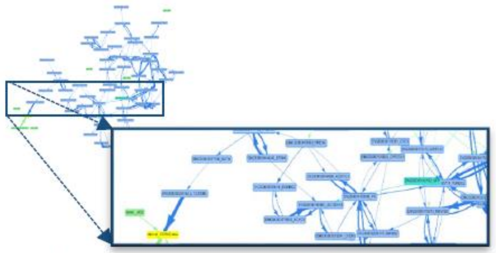


Results were validated out of cohort in a randomized control trial at Dana Faber

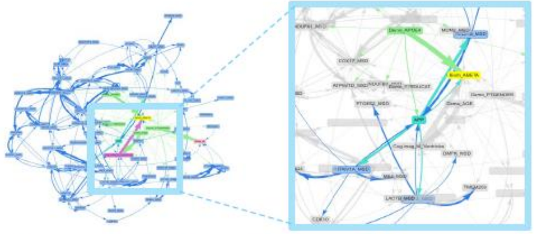
Results published at the 59th ASH Annual Meeting

Understanding Mechanisms of Failure for BACE Inhibitors in Alzheimer's Disease

At least 17 clinical trials of BACE inhibitor compounds have failed due to lack of efficacy or side effects



*Brain tissue-based causal network shows reducing APP expression may positively affect the **accumulation of a-β in the brain** and may have some gross **impact on cognitive effects at end-stage disease***



Blood-based causal network shows that despite mechanistic connections between APP and the outcomes, there is **no causal impact of reducing APP on cognition for early-stage disease, validating results of earlier clinical trial failures**

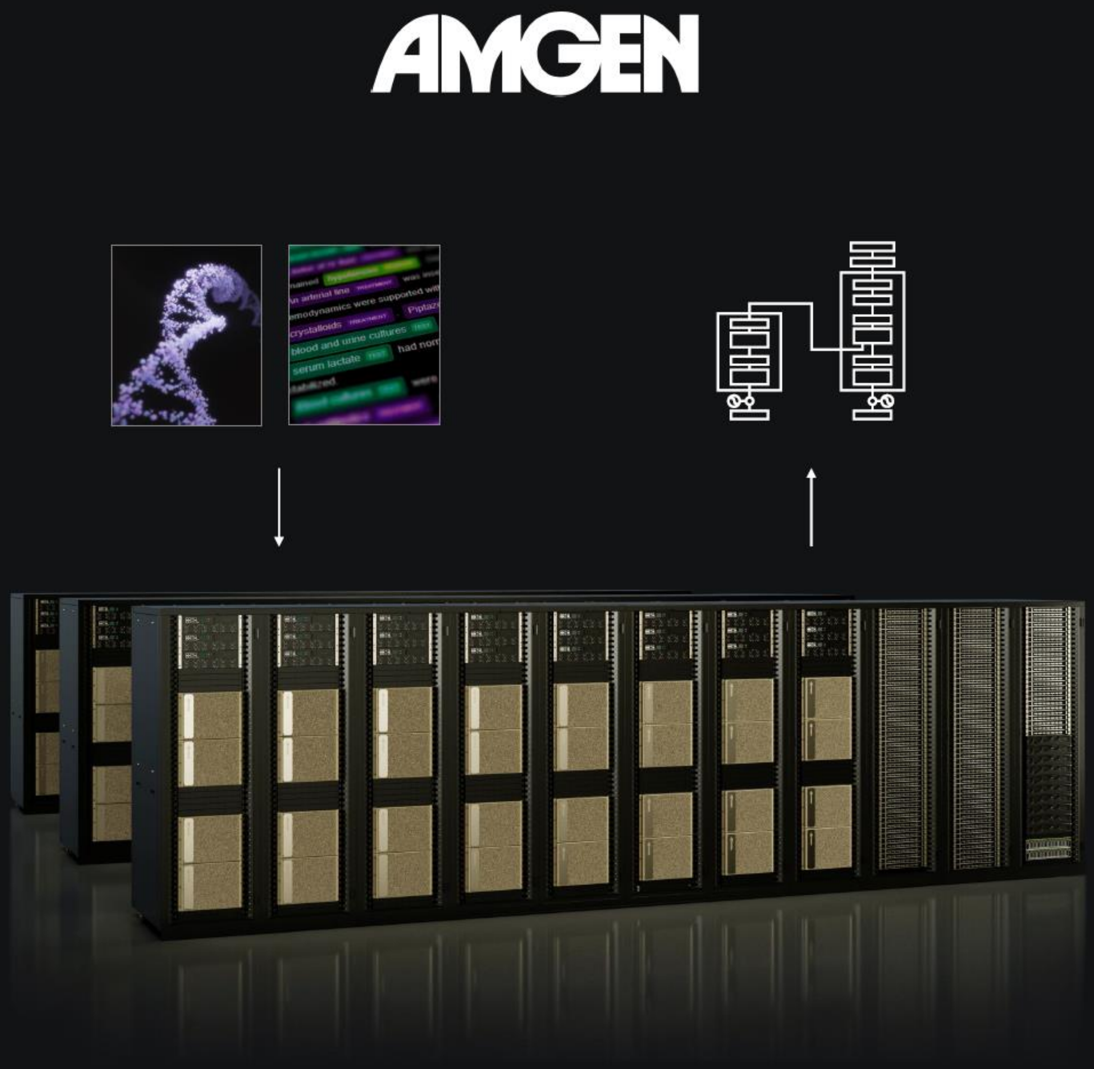
AITIA

Source: Aitia Health Investor Presentation, April 2024

Amgen Partnering with Nvidia for Target ID Using its Large Genomic/Proteomic Datasets

Announcing Amgen Building Generative AI Models in the Search for Novel Human Data Insights and for Drug Discovery

Powered by —
NVIDIA DGX H100 and BioNeMo



Source: https://s201.q4cdn.com/141608511/files/doc_presentations/2024/Jan/08/NVIDIA-Accelerating-Healthcare-Innovation-with-AI.pdf

Amgen/Nvidia Collaboration Focused on Decode Data

Casey Ross, "AI can speed up drug discovery but don't expect it to cure cancer, yet," *Stat+*, Feb 7, 2024.

Working with the computing giant Nvidia, Amgen is planning to rapidly build models that can crunch data compiled by Decode, a company Amgen acquired in 2012 that has collected de-identified information on 3 million people. Its repository contains what in the business is called "multi omics" data — genomic data that describes DNA; proteomic data that explains the structure and function of proteins; and transcriptomic data that catalogs RNA and the operation of genes.

All of that detail is combined with a dense thicket of information — 10,000 different traits — that describe the people who contributed the data, from hair color to the size of their livers.

"What we can now do is use machine learning and artificial intelligence to begin querying these datasets to understand disease on a basic level," said Amgen's David Reese (Chief Technology Officer). The goal is to break down a diagnosis such as lupus, an autoimmune disease characterized by a wide range of symptoms, into more precise molecular subgroups. "That gives you drug development hypotheses," Reese said. "How can we then intervene to reset that immune system or block that immune system that is overactive against normal tissue?"

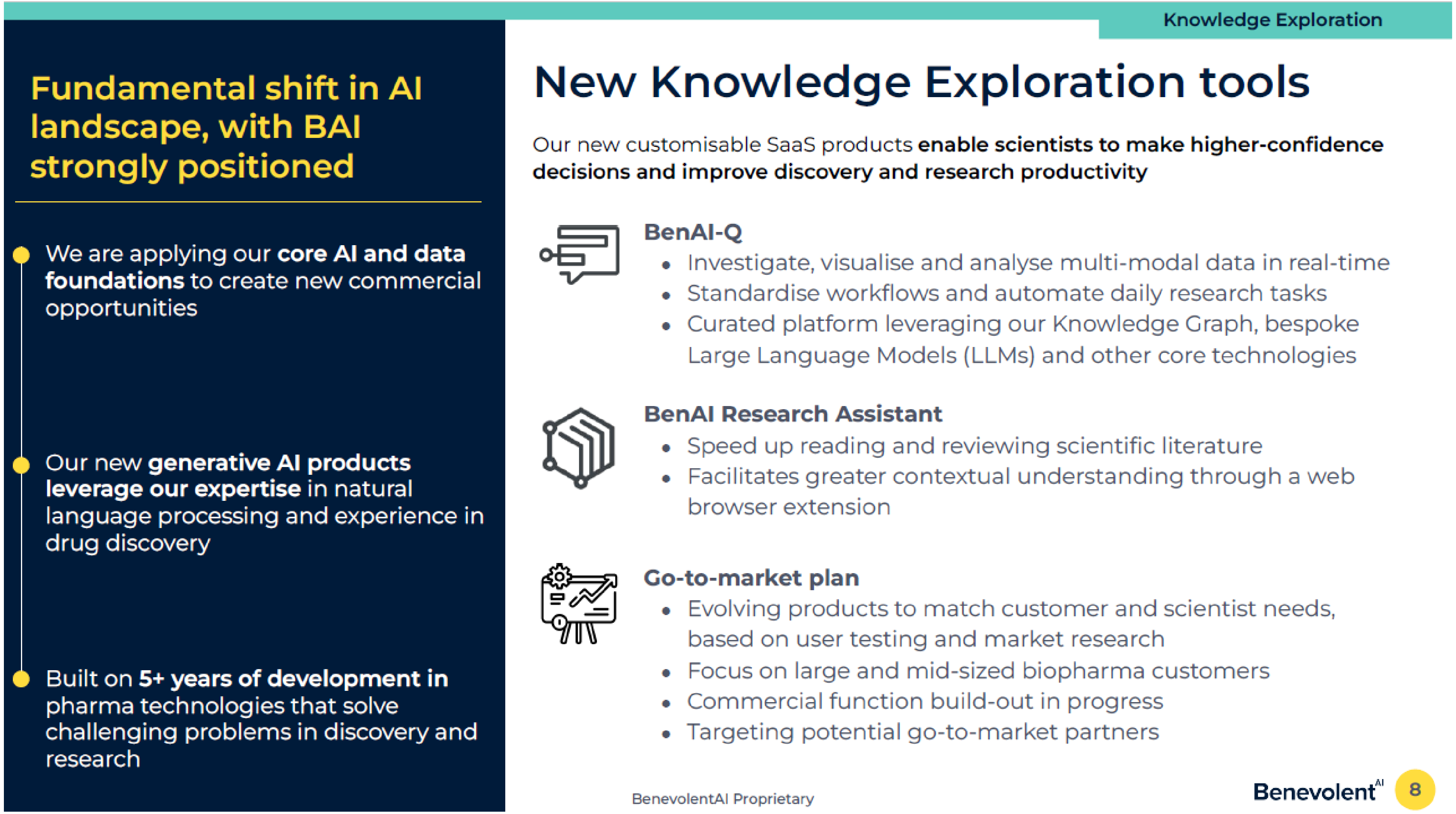
Biologic Samples in Cold Storage, Decode

Source: https://www.statnews.com/2024/02/07/ai-drug-discovery-novartis-amgen-sanofi/

Subsection

Literature and Cell-Screening Approaches to Target ID

Benevolent AI Offers Literature Review AI Tools and Multi-Modal Data Visualization



Knowledge Exploration

Fundamental shift in AI landscape, with BAI strongly positioned

- We are applying our **core AI and data foundations** to create new commercial opportunities
- Our new **generative AI products leverage our expertise** in natural language processing and experience in drug discovery
- Built on **5+ years of development in** pharma technologies that solve challenging problems in discovery and research

New Knowledge Exploration tools

Our new customisable SaaS products **enable scientists to make higher-confidence decisions and improve discovery and research productivity**

BenAI-Q

- Investigate, visualise and analyse multi-modal data in real-time
- Standardise workflows and automate daily research tasks
- Curated platform leveraging our Knowledge Graph, bespoke Large Language Models (LLMs) and other core technologies

BenAI Research Assistant

- Speed up reading and reviewing scientific literature
- Facilitates greater contextual understanding through a web browser extension

Go-to-market plan

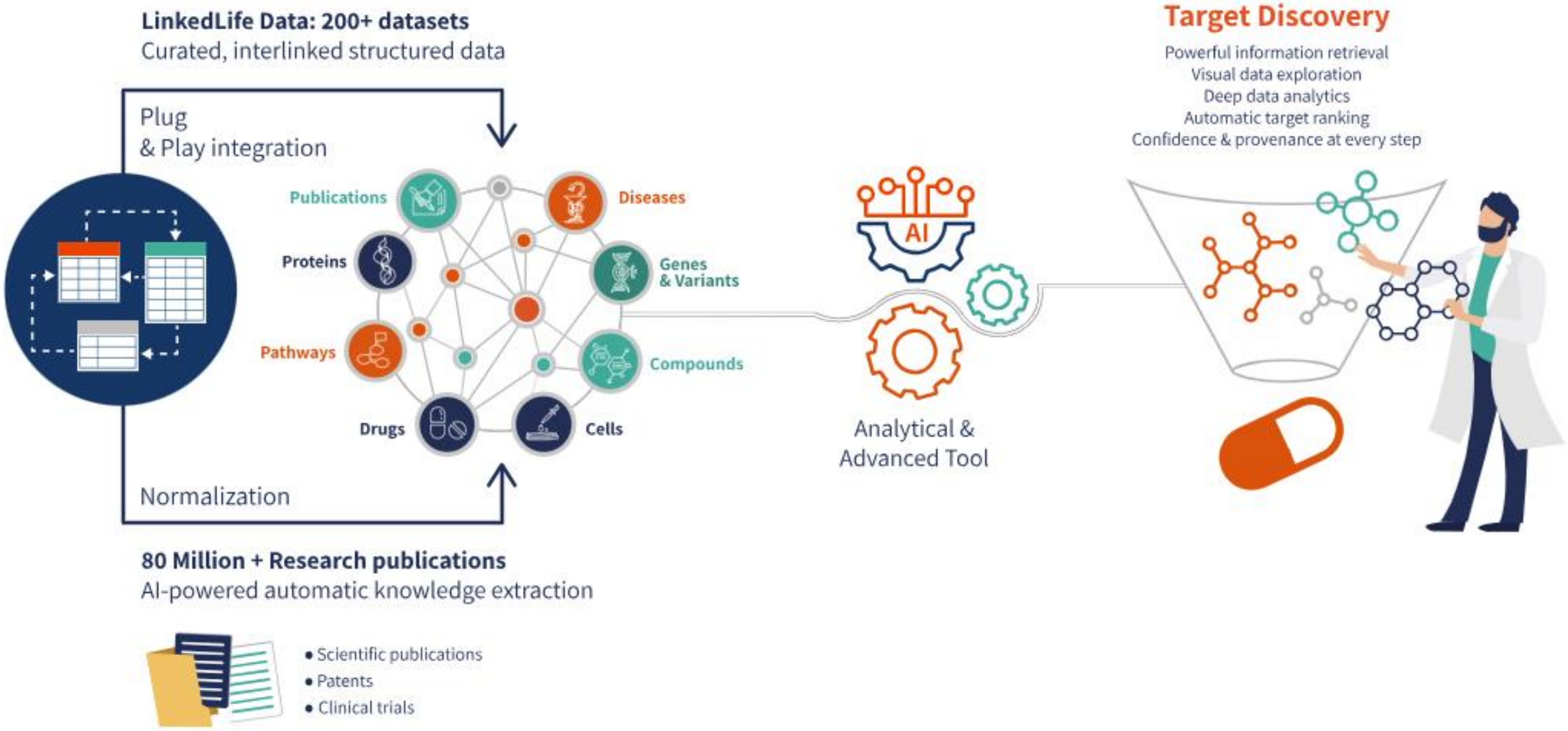
- Evolving products to match customer and scientist needs, based on user testing and market research
- Focus on large and mid-sized biopharma customers
- Commercial function build-out in progress
- Targeting potential go-to-market partners

BenevolentAI Proprietary

Benevolent^AI 8

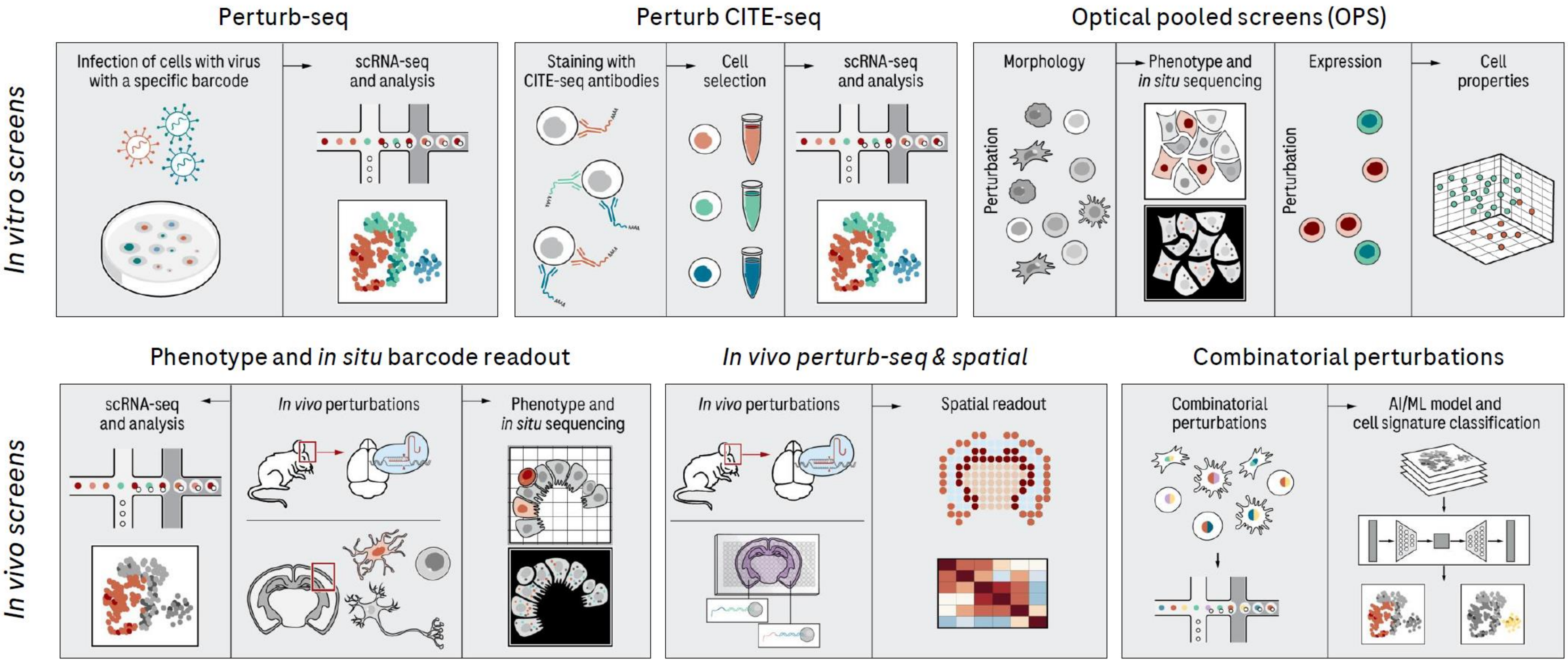
Source: https://www.benevolent.com/application/files/1317/0498/7775/BenevolentAI_-_J.P._Morgan_Healthcare_Conference_San_Francisco_2024_presentation.pdf

OntoText Also Employs Literature + Multimodal Datasets for Target Identification



Source: https://www.ontotext.com/blog/turbocharging-target-identification-ontotexts-ai-powered-solution-at-work/

Genentech: Massively Parallel, Pooled Cell Screens for Function


Perturb-seq
Infection of cells with virus with a specific barcode
scRNA-seq and analysis
Perturb CITE-seq
Staining with CITE-seq antibodies
Cell selection
scRNA-seq and analysis
Optical pooled screens (OPS)
Morphology
Phenotype and in situ sequencing
Expression
Cell properties
Perturbation
Perturbation
In vitro screens
Phenotype and in situ barcode readout
scRNA-seq and analysis
In vivo perturbations
Phenotype and in situ sequencing
In vivo perturb-seq & spatial
In vivo perturbations
Spatial readout
Combinatorial perturbations
Combinatorial perturbations
AI/ML model and cell signature classification
In vivo screens

Dixit et al., Cell 2016; MacFarland et al., Nature Communications 2020; Jin et al., Science 2020; Frangieh et al., Nature Genetics 2021; Ursu et al., Nature Biotechnology 2021; Paulsen et al., Nature 2022; Yao et al., Nature Biotechnology, 2023, Geiger, Eraslan et al., Biorxiv 2023

Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 60

Insitro Uses Cell Models to Understand How Genetic Variations Link to Disease

insitro™

Daphne Koller, CEO, Insitro, Nov 16, 2022

Our focus is "de-convoluting" the biology of human disease. Often, clinicians tackle disease without really understanding what the disease even is. Disease is often defined by coarse-grained symptomatic manifestations, some of which use classifications that date back 50 years or more. These are typically filtered through a subjective lens of both the patient and the clinician, so we end up with a mishmash that really doesn't speak to the underlying biological causes of the disease.

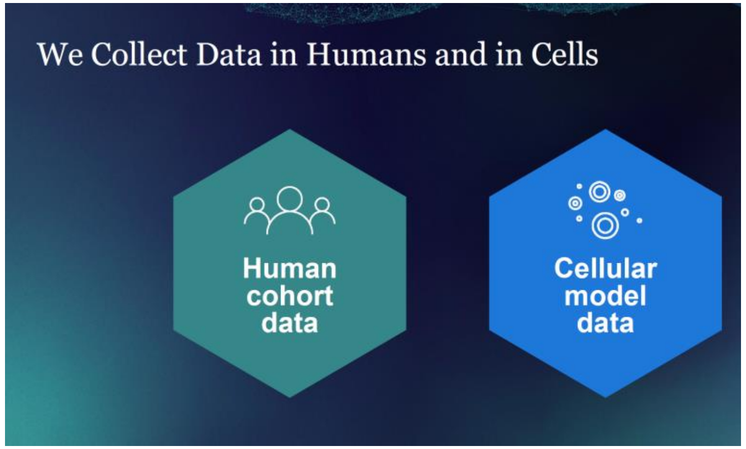
At insitro, we collect high-content data to help us understand underlying biological processes that correspond to disease. Some of those data sets come from patients. For example, we collect imaging data, such as MRI and histopathology; various molecular measurements; and other data that allow us to identify, via machine learning, subtle patterns to disentangle distinct patient subsets.

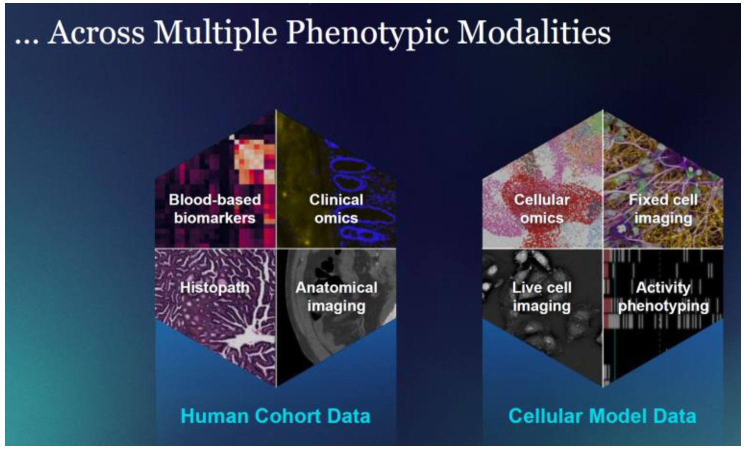
At the same time, we generate in our lab large amounts of human-derived cells, called induced pluripotent stem cells. These are human cells that were reverted to stem cell status, from which we then create neurons or hepatocytes that carry our genetics. We can further introduce into those cells genetic variations that we know are likely to cause disease. Then we can measure those cells and interrogate—with microscopy or RNA sequencing—what disease looks like at the cellular level. This system gives us a rapid approach for testing therapeutic interventions that could potentially work in humans.

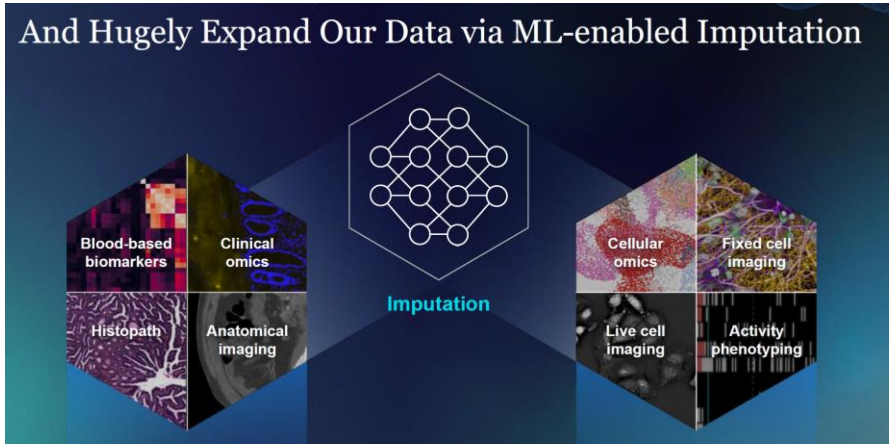
Source: https://www.mckinsey.com/industries/life-sciences/our-insights/it-will-be-a-paradigm-shift-daphne-koller-on-machine-learning-in-drug-discovery

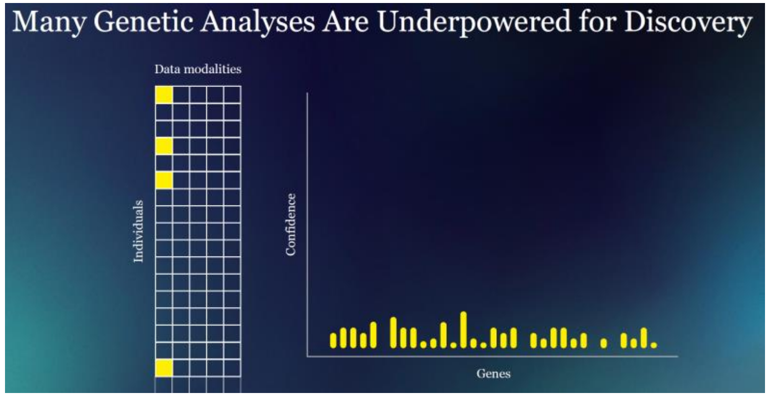
Daphne Koller, CEO, Insitro

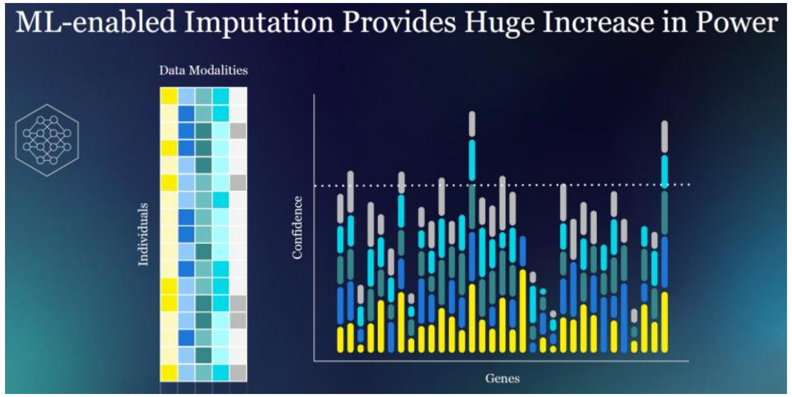
Insitro Idea: Use Multiple Modalities + ML to Triangulate Disease Insights from Data

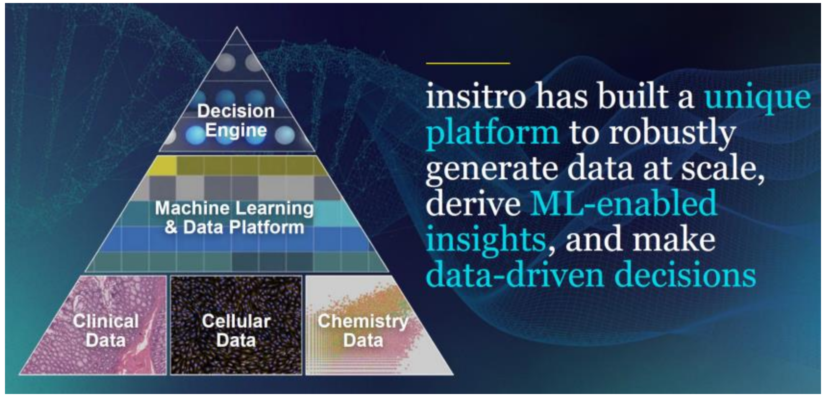

We Collect Data in Humans and in Cells


… Across Multiple Phenotypic Modalities


And Hugely Expand Our Data via ML-enabled Imputation


Many Genetic Analyses Are Underpowered for Discovery


ML-enabled Imputation Provides Huge Increase in Power

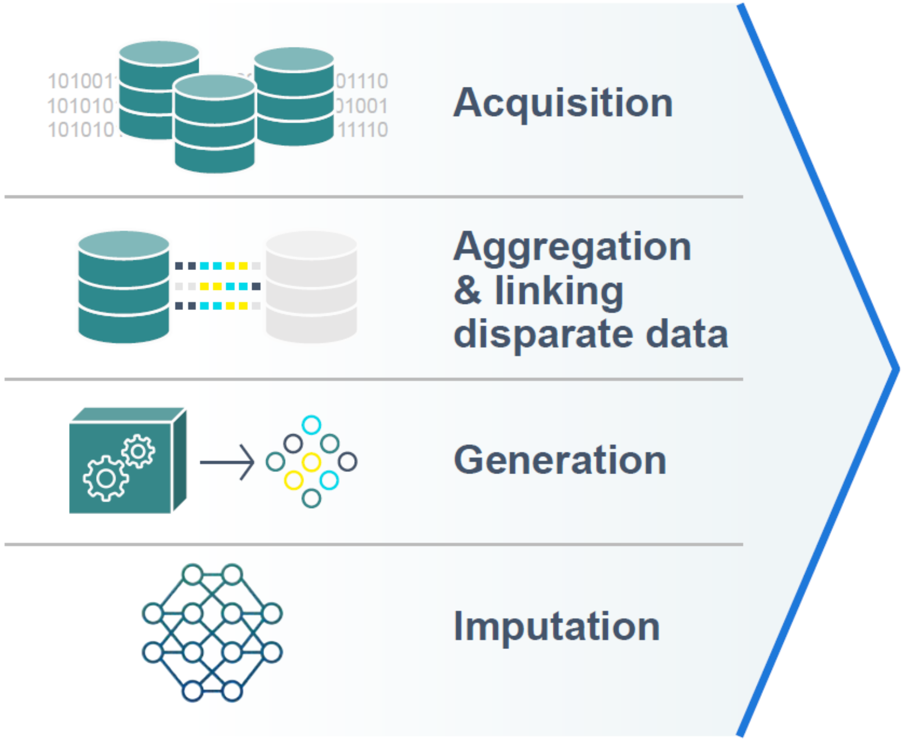


insitro has built a unique platform to robustly generate data at scale, derive ML-enabled insights, and make data-driven decisions

Source: https://www.insitro.com/wp-content/uploads/2024/01/insitro_2024JPM_npweb.pdf

Multi-pronged Approach to Scale High-content Clinical Data Drives Platform Flywheel

We are building a rich data corpus spanning multiple modalities & disease contexts

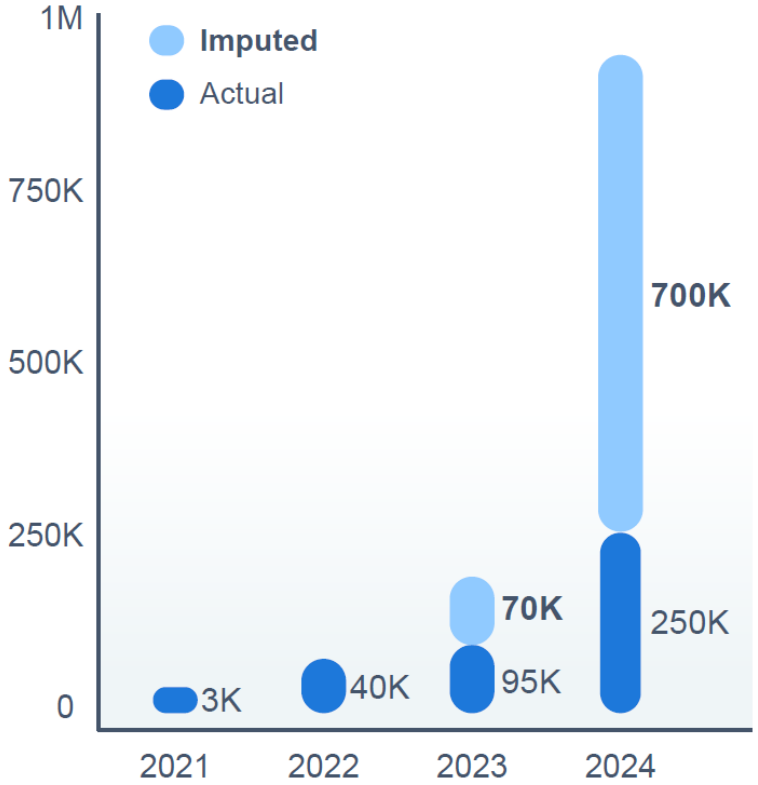


Acquisition

Aggregation & linking disparate data

Generation

Imputation

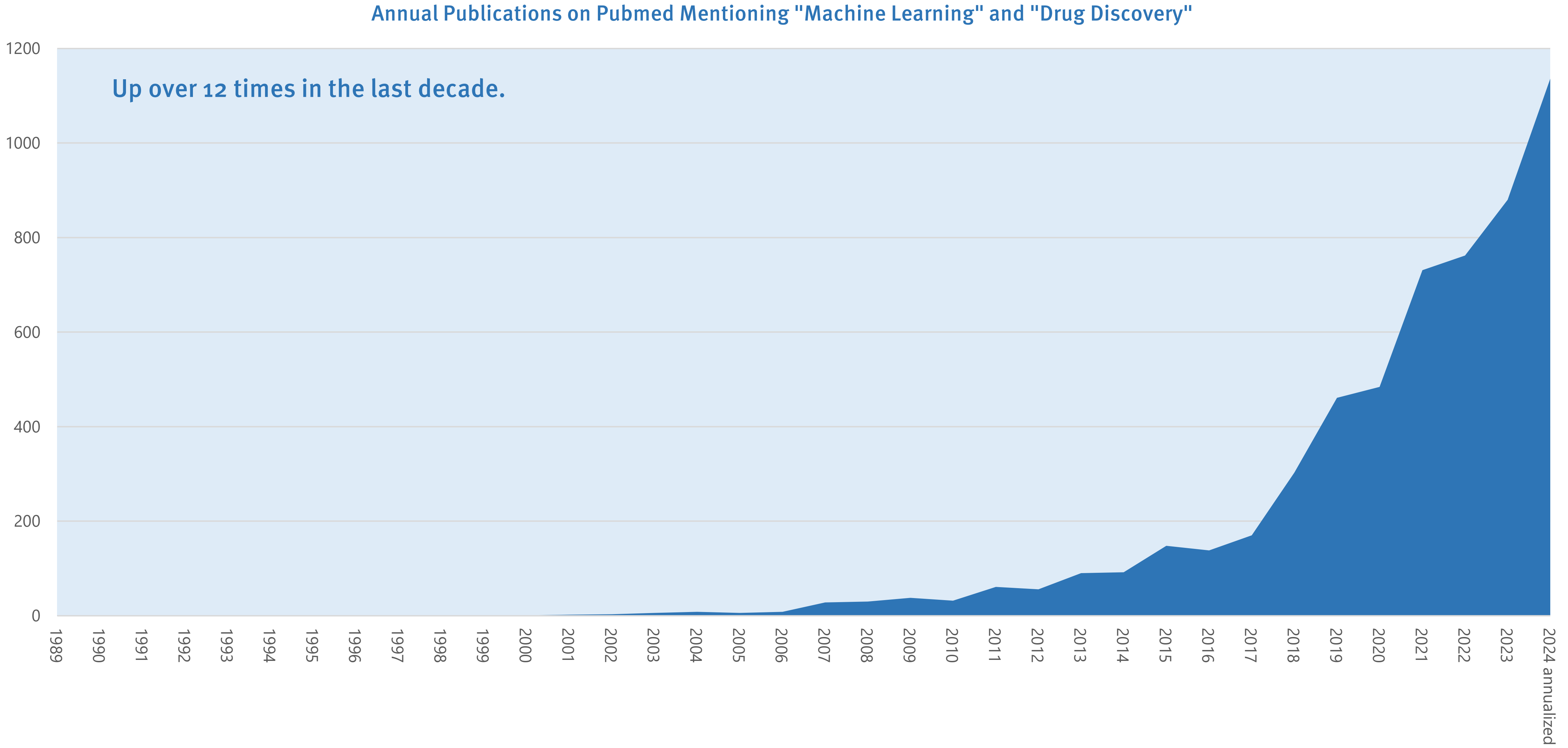


- ~100,000 curated cases with extensive multimodal data today
- Scaling to ~1M multimodal cases with imputation
- More data drives better and broader imputation
- Platform flywheel to massively increase prediction quality *without* massive costs

Source: https://www.insitro.com/wp-content/uploads/2024/01/insitro_2024JPM_npweb.pdf

AI in Drug Discovery

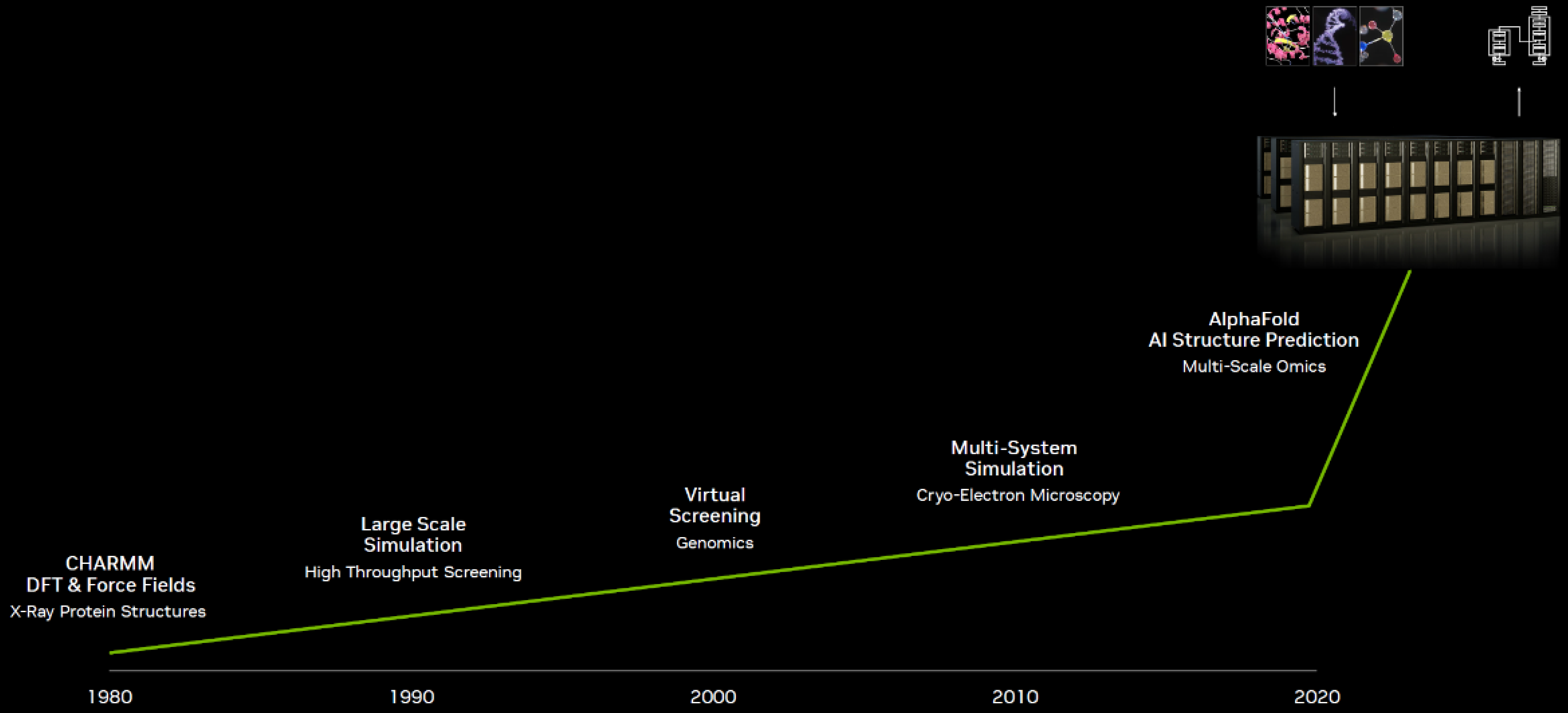
Explosion of Interest in Using Machine Learning in Drug Discovery



Source: Pubmed

Drug Discovery Is at an Inflection Point

Computer Aided Drug Discovery is Expanding Exponentially



CHARMM
DFT & Force Fields
X-Ray Protein Structures

Large Scale Simulation
High Throughput Screening

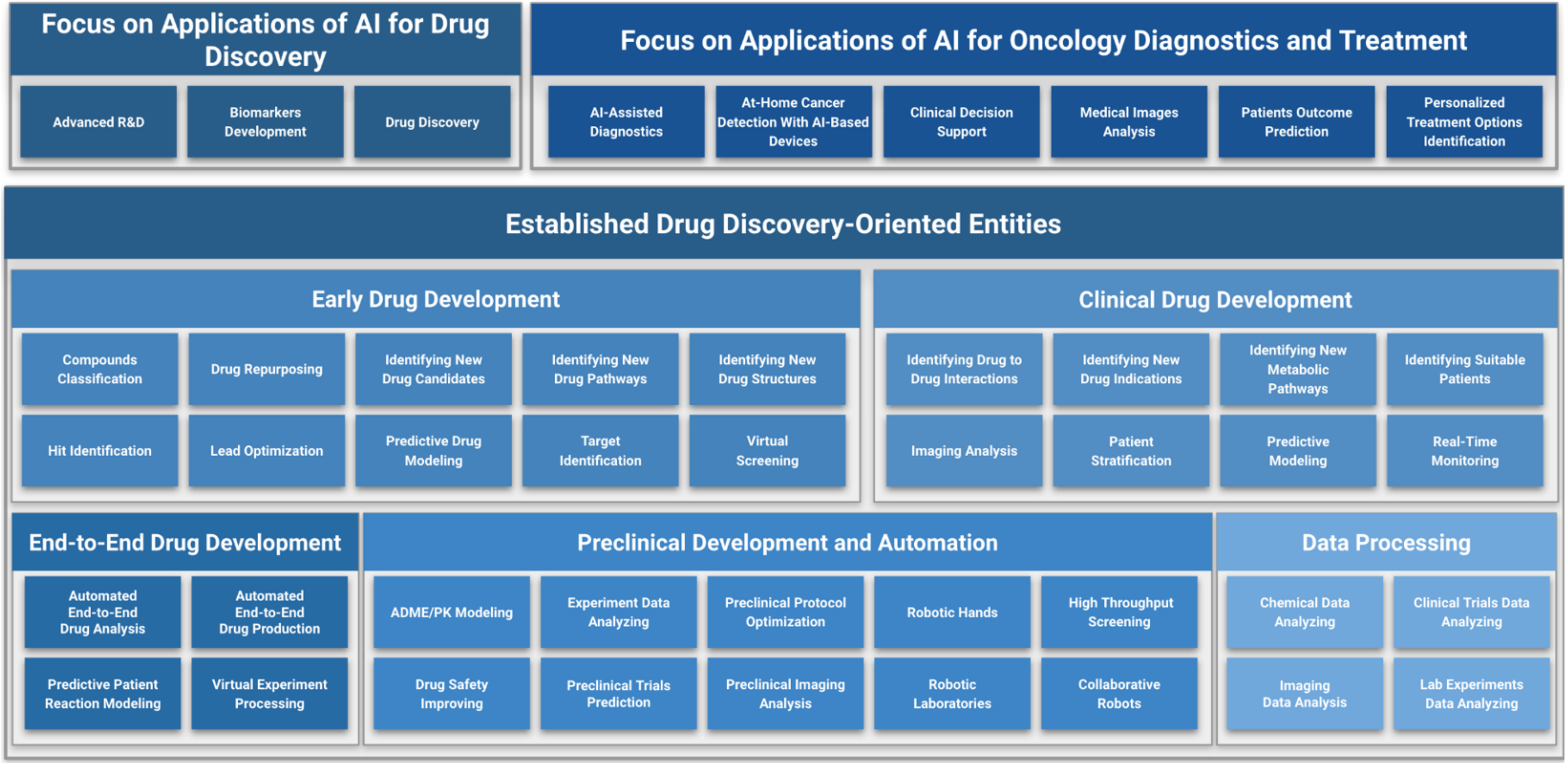
Virtual Screening
Genomics

Multi-System Simulation
Cryo-Electron Microscopy

AlphaFold
AI Structure Prediction
Multi-Scale Omics

Source: https://s201.q4cdn.com/141608511/files/doc_presentations/2024/Jan/08/NVIDIA-Accelerating-Healthcare-Innovation-with-AI.pdf

Highly Diverse Set of Companies Working on Drug Discovery Using AI Methods



Focus on Applications of AI for Drug Discovery

- Advanced R&D
- Biomarkers Development
- Drug Discovery

Focus on Applications of AI for Oncology Diagnostics and Treatment

- AI-Assisted Diagnostics
- At-Home Cancer Detection With AI-Based Devices
- Clinical Decision Support
- Medical Images Analysis
- Patients Outcome Prediction
- Personalized Treatment Options Identification

Established Drug Discovery-Oriented Entities

Early Drug Development

- Compounds Classification
- Drug Repurposing
- Identifying New Drug Candidates
- Identifying New Drug Pathways
- Identifying New Drug Structures
- Hit Identification
- Lead Optimization
- Predictive Drug Modeling
- Target Identification
- Virtual Screening

Clinical Drug Development

- Identifying Drug to Drug Interactions
- Identifying New Drug Indications
- Identifying New Metabolic Pathways
- Identifying Suitable Patients
- Imaging Analysis
- Patient Stratification
- Predictive Modeling
- Real-Time Monitoring

End-to-End Drug Development

- Automated End-to-End Drug Analysis
- Automated End-to-End Drug Production
- Predictive Patient Reaction Modeling
- Virtual Experiment Processing

Preclinical Development and Automation

- ADME/PK Modeling
- Experiment Data Analyzing
- Preclinical Protocol Optimization
- Robotic Hands
- High Throughput Screening
- Drug Safety Improving
- Preclinical Trials Prediction
- Preclinical Imaging Analysis
- Robotic Laboratories
- Collaborative Robots

Data Processing

- Chemical Data Analyzing
- Clinical Trials Data Analyzing
- Imaging Data Analysis
- Lab Experiments Data Analyzing

Source: https://www.deep-pharma.tech/post/framework

Selected Biotechs Focused on Drug Discovery and AI

SMALL MOLECULE DISCOVERY WITH AI/ML

1859
A2A PHARMA
A-ALPHA BIO
Atomwise
boltzmann
deep apple
congruence TX
DEEPCURE
dewpoint$_x$
DRUG FARM
engine Biosciences
Enveda BIOSCIENCES
evotec
Exscientia
Genesis Therapeutics
Iambic
IKTOS
Insilico Medicine
Inceptive
insitro
KIMIA
Montai Health
METiS
OCTANT
PAULING.AI
Predictive Oncology
RELAY THERAPEUTICS
RECURSION
Regor Therapeutics Group
StoneWise 望石智慧
Schrödinger
Valo
VANTAI
vevo
XtalPi

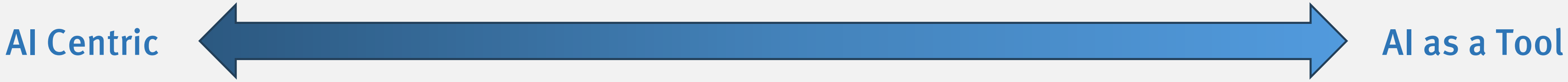
Source: Stifel research

BIOLOGICS / AAV / NUCLEIC ACID DISCOVERY

absci
AI PROTEINS
AI AIKIUM
Atomic AI
BigHat
CHARM THERAPEUTICS
compugen
DYNO THERAPEUTICS
EVOZYNE NATURAL MACHINES
eikon therapeutics
Exscientia
Genesis Therapeutics
Generate:Biomedicines
LabGenius
Isomorphic Labs
menten.AI
Peptilogics
Profluent
RELAY THERAPEUTICS
SEISMIC THERAPEUTIC
VILYA
XtalPi

Flavors of Approaches to Drug Discovery with AI

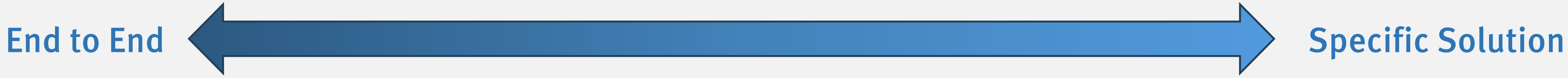
Some companies do everything with AI including in silico drug generation, testing etc. while others use AI but is not a central part of their value proposition. Rather AI is another tool used as part of the drug discovery and development process.



AI Centric

AI as a Tool

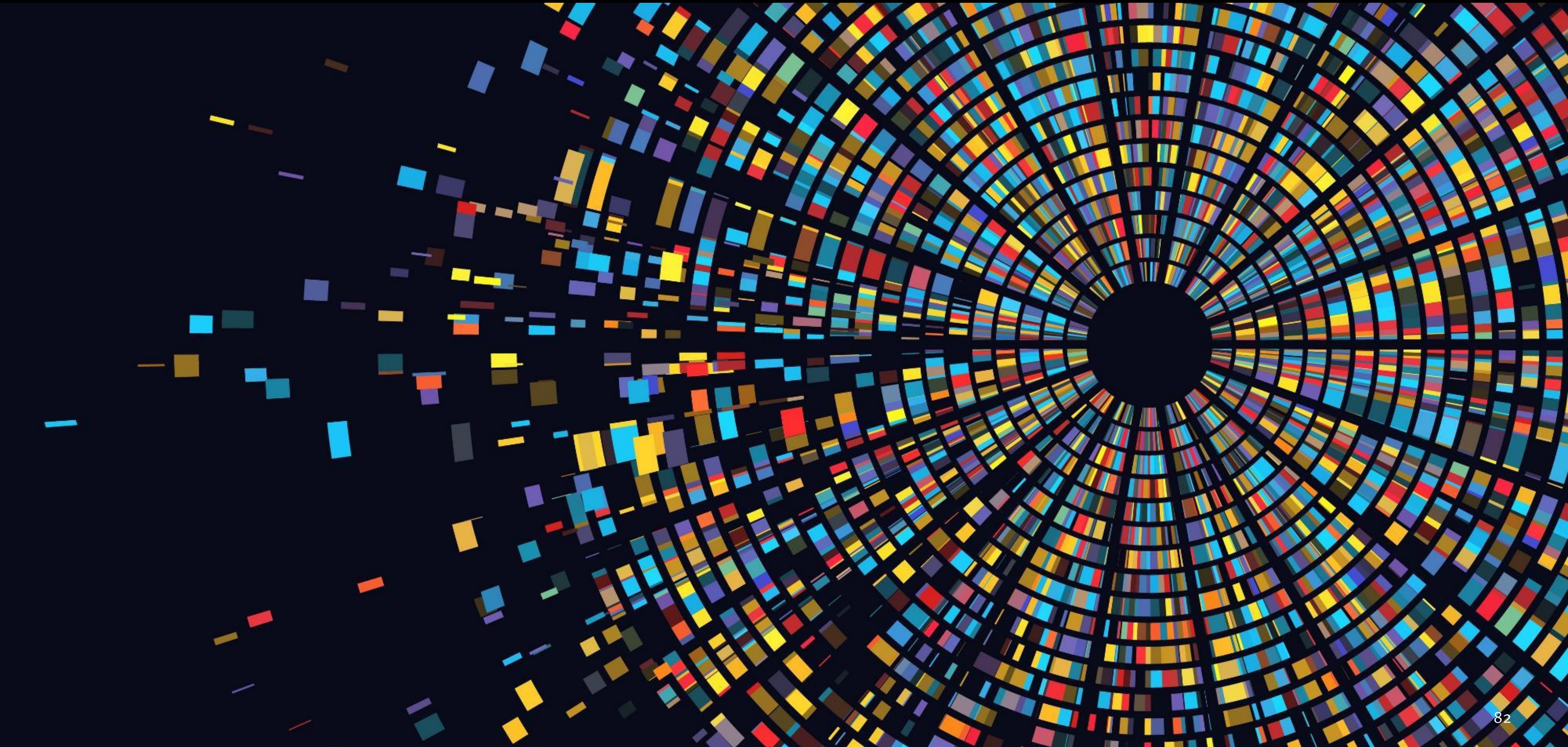
Some companies position themselves as a new type of pharmaceutical company that uses AI and machine learning in every aspect of their business ranging from target discovery, drug generation and clinical trials. Others focus on a specific area (e.g., DNA encoded libraries) for target binding.



End to End

Specific Solution

Key Issues with AI in Drug Discovery



It’s Early Days in AI and Drug Discovery

“It’s still very early from the standpoint of really integrating AI models in the experimental drug design and discovery pipeline.”

Marinka Zitnik
Professor of Bioinformatics
Harvard Medical School

Source: https://www.statnews.com/2024/02/07/ai-drug-discovery-novartis-amgen-sanofi/

AI Skeptic: Carl Hansen, CEO of Abcellera

Brian Buntz, "AI in antibody drug discovery as a tool, not a magic wand," *Drug Discovery & Development*, Sep 25, 2023 (excerpt)

AI in drug discovery is a topic that gets an outsized amount of attention, observes Carl Hansen, CEO of AbCellera, a company specializing in antibody drug discovery. "It's as if people are saying, 'AI is here, it's going to save us. Thank God, we're finally gonna be able to create drugs," he said. "To me, that's implicitly saying, 'We don't know what the heck we're doing.'"

But human researchers have been tirelessly working behind the scenes, making incremental advances that lay the foundation for these so-called AI "breakthroughs," according to Hansen, who holds a Ph.D. in applied physics/biotechnology from Caltech. "I'm much longer on human intelligence than artificial intelligence," he added.

While generative AI engines such as ChatGPT may have fueled more public interest in the capabilities of artificial intelligence, Hansen said the approach to training the large language model contrasts sharply with biotech requirements. "ChatGPT has basically read every single word on the internet," he said. And consequently, the large language model can generate textual responses based on its training and iteratively refine its approach based on new data or feedback, going through waves of iterations to optimize its accuracy. **To do something similar in biotech would require a huge amount of data to start. "It would have to be harmonized, which basically means you have to do it internally. It can't just be vacuumed up from the world," Hansen said.**

The notion that AI alone can revolutionize biotech is misguided, according to Hansen: "If you rank order all of the hard things, the AI model part is the easiest part." The challenges relate more to data generation. "It takes years and years to build that capability," he said.

Pointing to claims that generative AI models are spitting out de novo sequences of antibodies that recognize the target, without any experimental technique, Hansen expressed skepticism. "For one, I think that's beyond what the technology can do today," he said. "Two, I think the problem is ill-posed." Pointing to the complex nature of antibody drug development, Hansen noted that when presented with a drug target, the ideal binding region for an antibody isn't necessarily clear immediately.

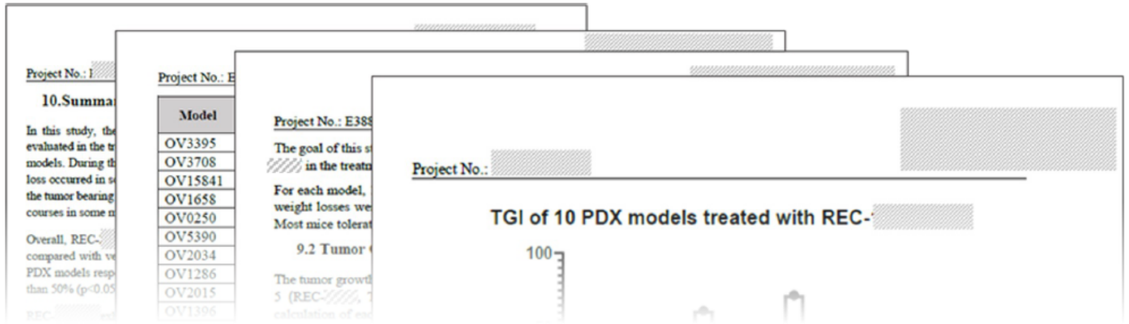
Traditionally, researchers would immunize animals to produce antibodies that bind to multiple potential regions of a target molecule. They then experimentally test the antibodies to identify the most effective. "That's one of the most important steps in developing a drug," Hansen said. "But the way these AI companies frame it is, 'Here's the molecule, and I will make an antibody that binds on this region.'" The result is a dismissal of an array of potential binding options, potentially overlooking more innovative drugs. Emphasizing the stakes in drug development, Hansen underscored the importance of innovation. "In drug development, you don't get a ribbon for second place. It's all about innovation and novelty," Hansen said.

Source: https://www.drugdiscoverytrends.com/ai-antibody-drug-discovery-abcellera/

Data Gaps are Everywhere in Academia / Pharma

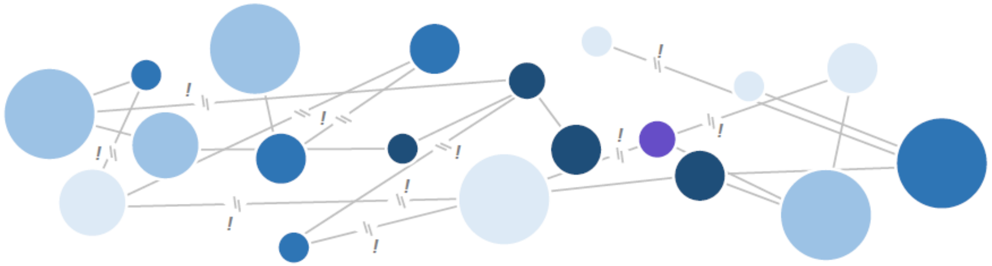
Analog Standard

The fax machine is alive and well in medicine, while in biopharma, study results from CROs are still often reported as PDFs or scanned printouts


Project No.: [redacted]
10.Summar
Project No.: E
Model
OV3395
OV3708
OV15841
OV1658
OV0250
OV5390
OV2034
OV1286
OV2015
Project No.: E38
9.2 Tumor
Project No.: [redacted]
TGI of 10 PDX models treated with REC-[redacted]
100

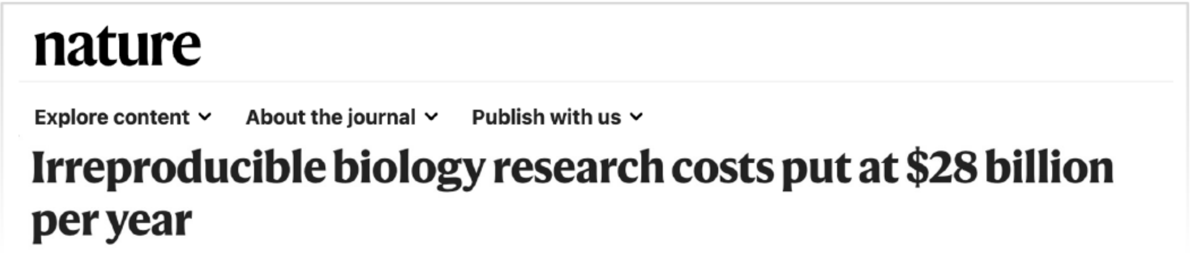
Siloed Data in Pharma

Biopharma has 100s of petabytes of scientific data stored on a project-by-project basis without the meta-data or annotation needed to relate it to other projects or questions in biology



Reproducibility Crisis

Multiple studies have shown that the vast majority of published academic literature cannot be recapitulated



nature

Irreproducible biology research costs put at $28 billion per year

Trademarks are the property of their respective owners and used for informational purposes only.

Baker, M. Irreproducible biology research costs put at $28 billion per year. Nature (2015). https://doi.org/10.1038/nature.2015.17711

Data Quality is Critical

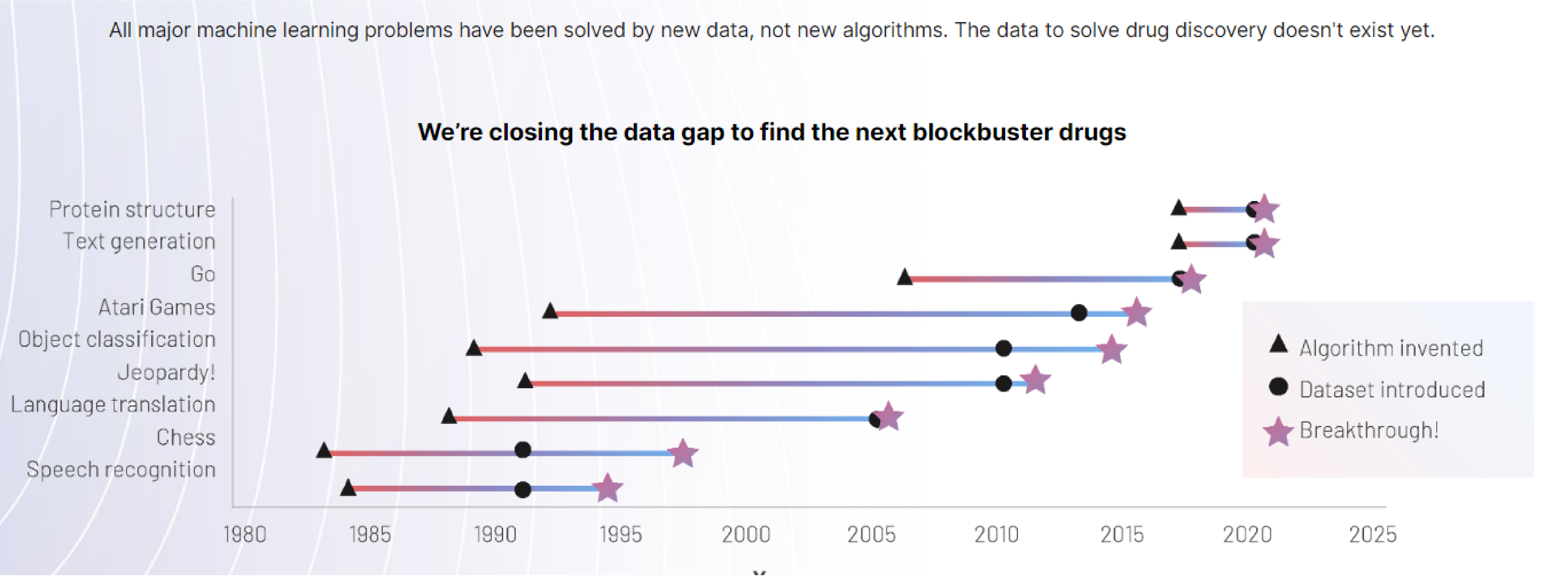
Dr. Dave Latshaw, founder and CEO of BioPhy

“The success of AI algorithms is heavily influenced by the quality and diversity of the data used for training. Ensuring that the data is accurate, representative, and free from biases helps create AI tools that perform well across different applications and user groups.

Algorithms are great but data makes the model.”

Source: https://www.forbes.com/sites/cindygordon/2024/02/23/using-ai-to-modernize-drug-development-and-lessons-learned/

It’s The Data that Really Matters



All major machine learning problems have been solved by new data, not new algorithms. The data to solve drug discovery doesn't exist yet.

We’re closing the data gap to find the next blockbuster drugs

Source: https://www.leash.bio/#data

Success is All About Solving Data Gaps

VERGE GENOMICS

Alice Zhang and Victor Hansson-Smith, Verge Genomics, *Timmerman Report*, Feb 9, 2023

“ChatGPT is a hot topic across many industries. Some say the technology underpinning it – called generative AI – has created an “AI arms race.” However, relatively little attention is given to what is needed to fully leverage the promise of generative AI in healthcare, and specifically how it may help accelerate drug discovery and development.

In short, our *belief* is that AI will identify better targets, thus reducing clinical failures in drug development and leading to new medicines. Generative AI will play a role. However, the fundamental *challenge* in making better medicines a reality comes down to closing the massive data gaps that remain in drug development today.”

Source: https://timmermanreport.com/2023/02/first-we-need-to-generate-the-right-data-then-ai-will-shine/

The Lack of Good Datasets is a Major Issue for AI in Drug Discovery

Blanco-González A, Cabezón A, Seco-González A, Conde-Torres D, Antelo-Riveiro P, Piñeiro Á, Garcia-Fandino R. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *Pharmaceuticals* (Basel). 2023 Jun 18;16(6):891.

Despite the potential benefits of AI in drug discovery, there are several challenges and limitations that must be considered. **One of the key challenges is the availability of suitable data. AI-based approaches typically require a large volume of information for training purposes.** In many cases, the amount of data that is accessible may be limited, or the data may be of low quality or inconsistent, which can affect the accuracy and reliability of the results.

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10302890/

Core Problem #1:

We lack good quality data for target identification and compound generation and selection.

Further Issue: Problem Dimensionality vs. Computing Power

Solving an optimization problem depends on how many branches need to be analyzed and scored.

Humans are generally pretty good at solving low dimensional problems.

For example,

> *"I could have pizza or roast chicken for lunch. Hmmm. I will pick the chicken today because I had pizza recently and didn't love it. Plus, it's not nearly as healthy for me."*

A computer could help but wouldn't add much value really.

Us humans quickly learned to use "rules of thumb" or other simplification methods to reduce the dimensionality of problems with many permutations.

To illustrate, "While there are 2,000 restaurants that I could go to in Cleveland tonight, I am know I will prefer one that I can get to in five minutes or less so I will just focus on those that are within a block of where I am now." This is called dimensionality reduction.

There are also quantitative methods for reducing the dimensionality of a computing problem such as that used in the Simplex algorithm used to solve the famous Traveling Salesman problem.

In general, the computer is going to be much better than a human at solving a problem with 10^1 to 10^{14} permutations. By the way, that's a lot of freaking permutations. Computers are really amazing!

A binary choice

OR

is a low dimensional problem

Dimensionality and Computing: London Taxis

At the Human Bounds of Problem Solving: Being a London Cab Driver

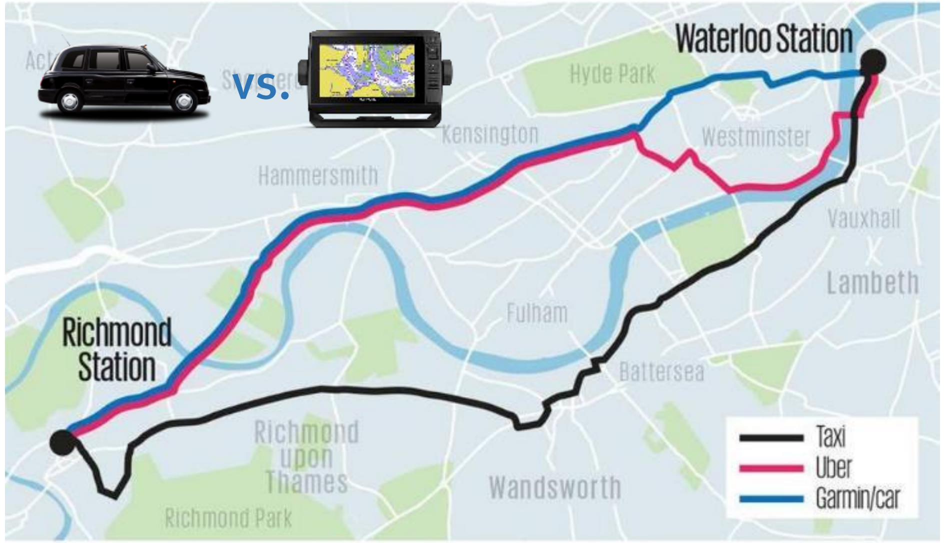
There are over 60,000 streets or roads within the 6-mile radius of Central London - with all of their one-way and restricted turn intricacies - plus over 100,000 places of note that the potential London cab driver has to learn to be licensed. Knowing this is called **The Knowledge**.

The Knowledge requires candidates to learn a total of 320 routes that crisscross London and are specifically designed to leave no gaps.[#] Taxi drivers have to also remember all places of interest or note *en route*: embassies, colleges, buildings, municipal offices and all other public buildings, hotels, theatres, stations, hospitals, museums, restaurants - and the list goes on.

It takes a candidate between two to four years to master the knowledge.

Then the Garmin "SatNav" arrived. By 2005, anyone could use a Garmin to navigate London

Today, it's trivially easy. In one recent experiment which pitted a London taxi driver against a Garmin and Uber, the Garmin matched the taxi driver and was much less expensive of a ride.[*] The computer solves a medium dimension problem in seconds that takes a human years to figure out.



Why this works

The Garmin is able to match years of human work in seconds because it is working off a well-labelled dataset and a bounded optimization problem.

If the data in the Garmin had Streets that were jumbled up and not assembled into a map and conveniently labelled, the human would easily beat it.

Today, the best computer can do around 400 petaflops a second (see next page). That's around 10^{16}.

That's a whole lot, but there are problems in biology that require much, much more.

Sources: https://www.the-london-taxi.com/london_taxi_knowledge, https://www.standard.co.uk/lifestyle/motors/theknowledgeubersatnav-a3694061.html

[#] An obvious dimensionality reduction maneuver.
[*] The Uber was slower for reasons that aren't entirely clear.

What a 400 Petaflops Supercomputer Looks Like

CAMBRIDGE-1 AI SUPERCOMPUTER

AstraZeneca
gsk
KING'S College LONDON
NHS
Guy's and St Thomas' NHS Foundation Trust
Oxford NANOPORE Technologies

80 NVIDIA DGX A100 | 400 PETAFLOPS AI COMPUTE

Source: https://nvidianews.nvidia.com/news/nvidia-building-uks-most-powerful-supercomputer-dedicated-to-ai-research-in-healthcare

Biological Datasets Can Get Really, Really Big

RECURSION

Eleanor Laise, *MarketWatch*, Sep 8, 2023

Salt Lake City-based Recursion Pharmaceuticals Inc. is looking to tackle that problem with its artificial-intelligence models for drug discovery and **25-petabyte biological and chemical dataset**. With a warehouse full of robots running millions of experiments per week continuously adding to that trove of data, Recursion is working to build "a foundational model of how biology and chemistry work and interact" that can profoundly change the drug discovery process, Recursion CEO and co-founder Chris Gibson told MarketWatch.

eikon therapeutics

Max Bayer, *FierceBiotech*, Jun 1, 2023

Meanwhile, Eikon is accruing massive amounts of data, derived from the company's single molecule tracking technology that allows the tracking of behavior and movements of individual proteins in a cell. That results in a lot of data—roughly **one and a half petabytes every week**, or 1.5 million gigabytes. Perlmutter says the largest team in the company of roughly 300 people are software engineers that are using machine learning to help manage all of this information.

The issue here is an obvious one:

While today's best supercomputers can just barely handle the data coming out of Eikon and Recursion, that data would be have to be really well annotated to be usable in machine learning applications.

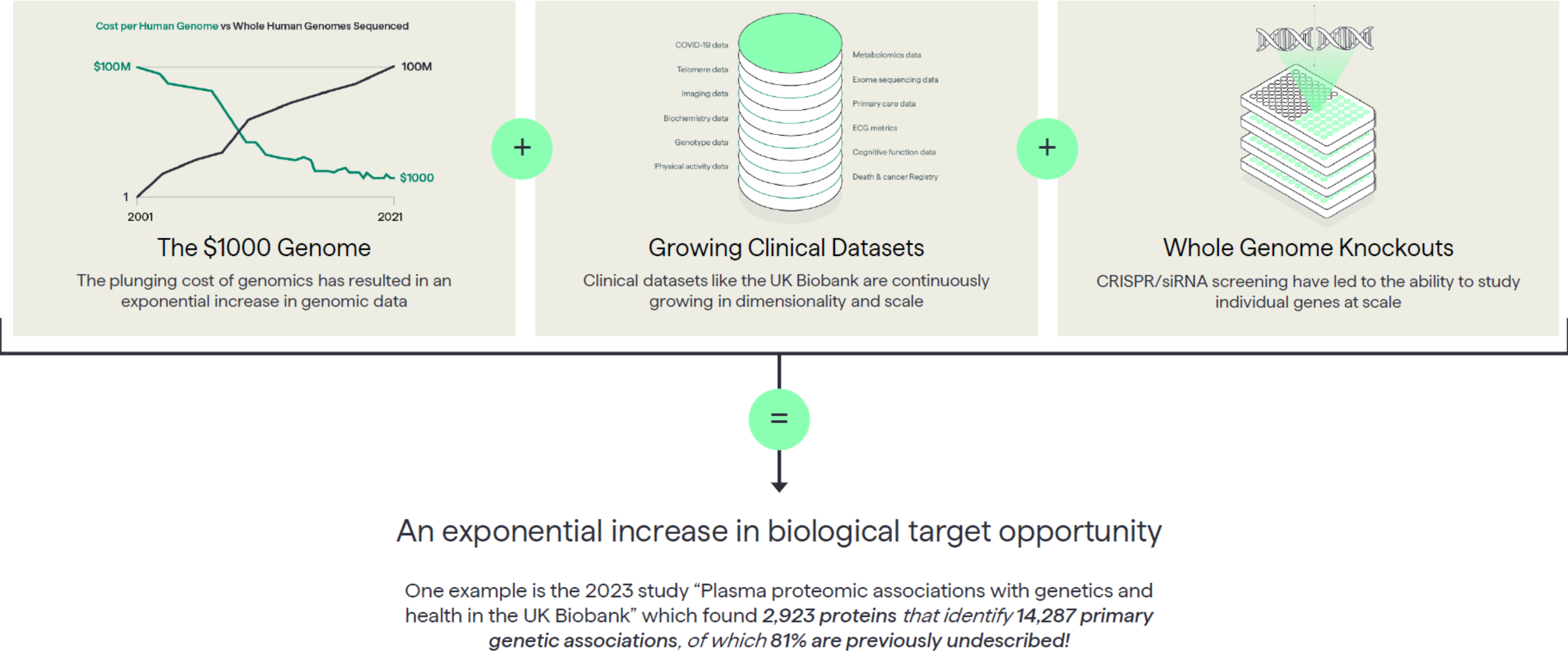
Our conversations with numerous observers and participants in the computational biology field indicate that the *quality of datasets* is the main issue.

The key need is to find datasets that are well annotated in a manner that is relevant to a problem in pharmacology.

Source: https://www.marketwatch.com/story/recursion-pharmaceuticals-has-a-prescription-for-high-drug-costs-ai-and-25-petabytes-of-data-8cb3d4ec, https://www.fiercebiotech.com/biotech/eikon-goes-prodigy-biotech-clinical-stage-buyer-flash-platform-remains-beating-heart

The Biologist Dream: Explosion in Known Biology Targets

(Slide from Terray Investor Deck)



The $1000 Genome

The plunging cost of genomics has resulted in an exponential increase in genomic data

Growing Clinical Datasets

Clinical datasets like the UK Biobank are continuously growing in dimensionality and scale

Whole Genome Knockouts

CRISPR/siRNA screening have led to the ability to study individual genes at scale

An exponential increase in biological target opportunity

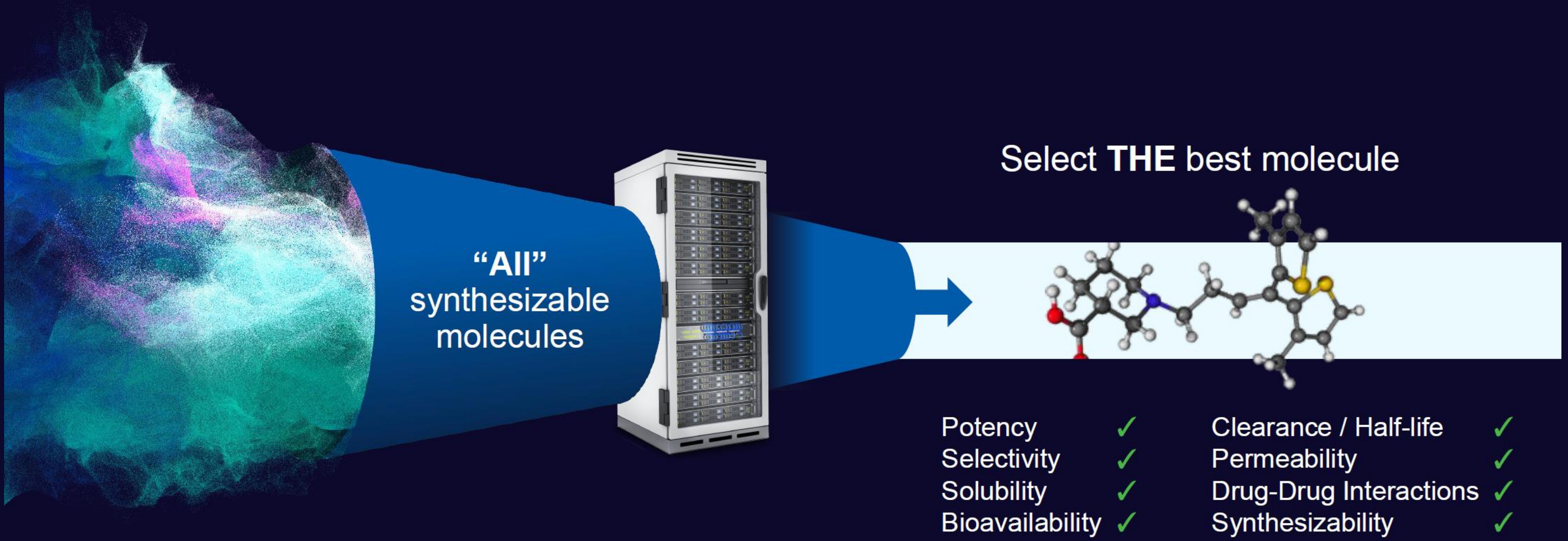
One example is the 2023 study “Plasma proteomic associations with genetics and health in the UK Biobank” which found ***2,923 proteins** that identify **14,287 primary genetic associations**, of which **81% are previously undescribed!***

Source: Terray Therapeutics Investor Presentation, April 2024, (https://www.terraytx.com/).

The Chemist’s Dream (slide from Schrödinger Investor Deck)

Vision for the Future of Drug Discovery

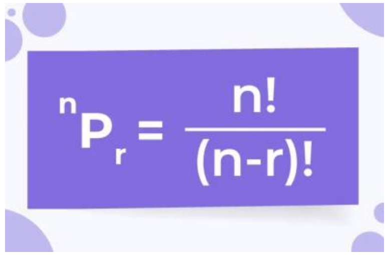
If all properties can be calculated with perfect accuracy, designing drugs would have a much **higher success rate**, be much **faster** and **cheaper**, and would produce much **higher-quality** molecules.



Potency	✓	Clearance / Half-life	✓
Selectivity	✓	Permeability	✓
Solubility	✓	Drug-Drug Interactions	✓
Bioavailability	✓	Synthesizability	✓

Is It Possible to Solve for Every Synthesizable Chemical Structure?

Question: If there are 118 known elements and you can’t have more than 100 atoms in a single drug, how many permutations of the drug would exist?



$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

where “n” is the total number of items in the set,
“r” is the number of items to be chosen, and
“!” denotes factorial, which is the product of all positive integers from 1 to the given number.

Answer - This number:

7310000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000

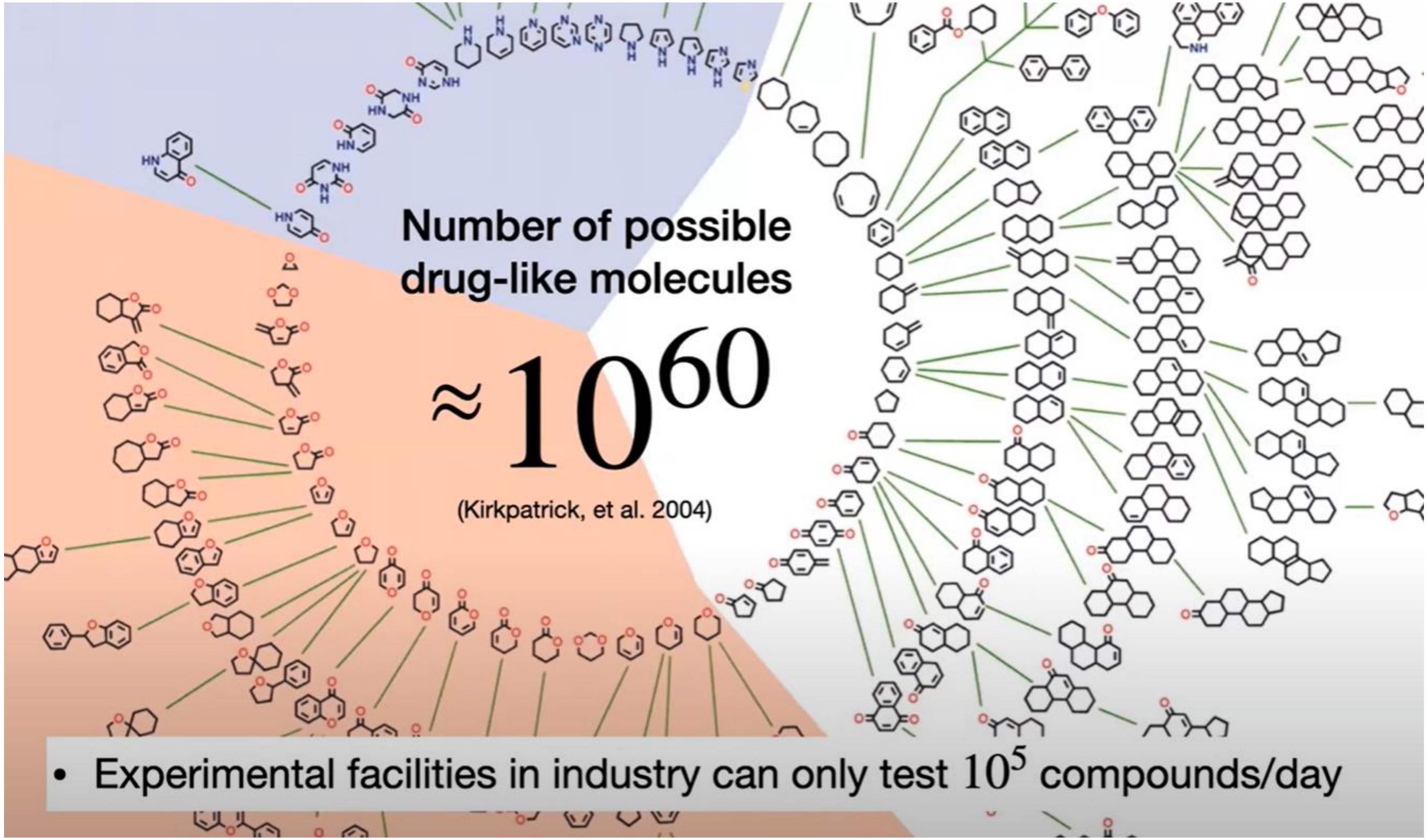
In shorthand:

7.31e+178 or a number with 163 extra zeros after quadrillion

Unfortunately, not all of the Nvidia GPU’s in the world can handle the problem of designing these drugs – much less testing them *in silico* against some objective. For all the progress made in computation, we are far from being able to design and analyze the universe of possible small molecule class of drugs.

Source: Info on the formula for computing number of permutations: https://www.cuemath.com/npr-formula/

Even if you Limit to Drug-Like Molecules, there are Far too Many for Modern Computation to Handle



Number of possible drug-like molecules

$\approx 10^{60}$

(Kirkpatrick, et al. 2004)

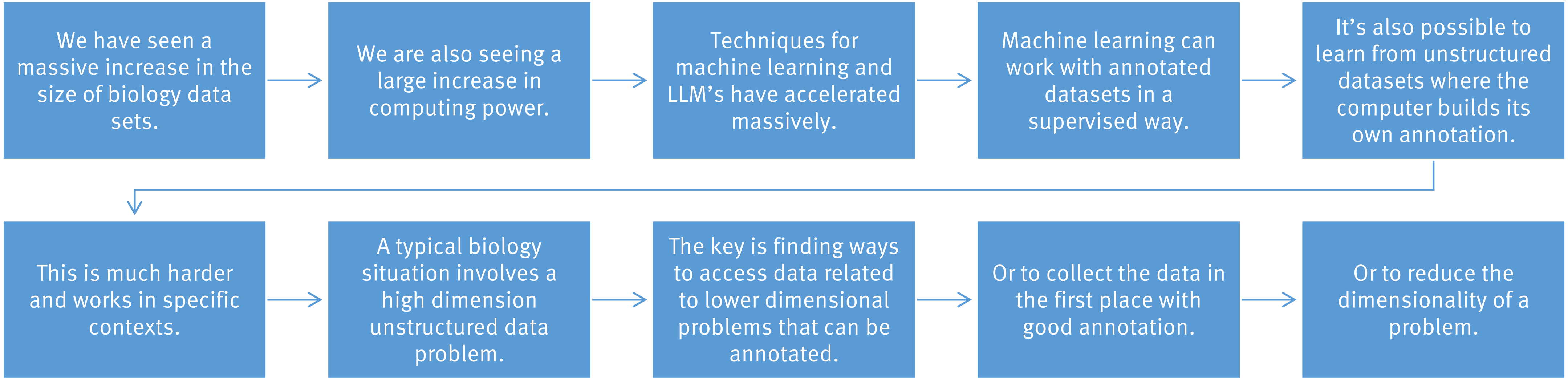
- Experimental facilities in industry can only test 10^5 compounds/day

See https://www.nature.com/articles/432823a, https://www.pnas.org/doi/10.1073/pnas.0503647102, https://www.youtube.com/watch?v=AHVJv5RNqKs

Core Problem #2:

We need good strategies to find the right compounds given a target in our lifetime with the computing resources that we have.

Bottom Line: Achieving Breakthroughs in Computational Biology is Really Hard



We have seen a massive increase in the size of biology data sets.

We are also seeing a large increase in computing power.

Techniques for machine learning and LLM’s have accelerated massively.

Machine learning can work with annotated datasets in a supervised way.

It’s also possible to learn from unstructured datasets where the computer builds its own annotation.

This is much harder and works in specific contexts.

A typical biology situation involves a high dimension unstructured data problem.

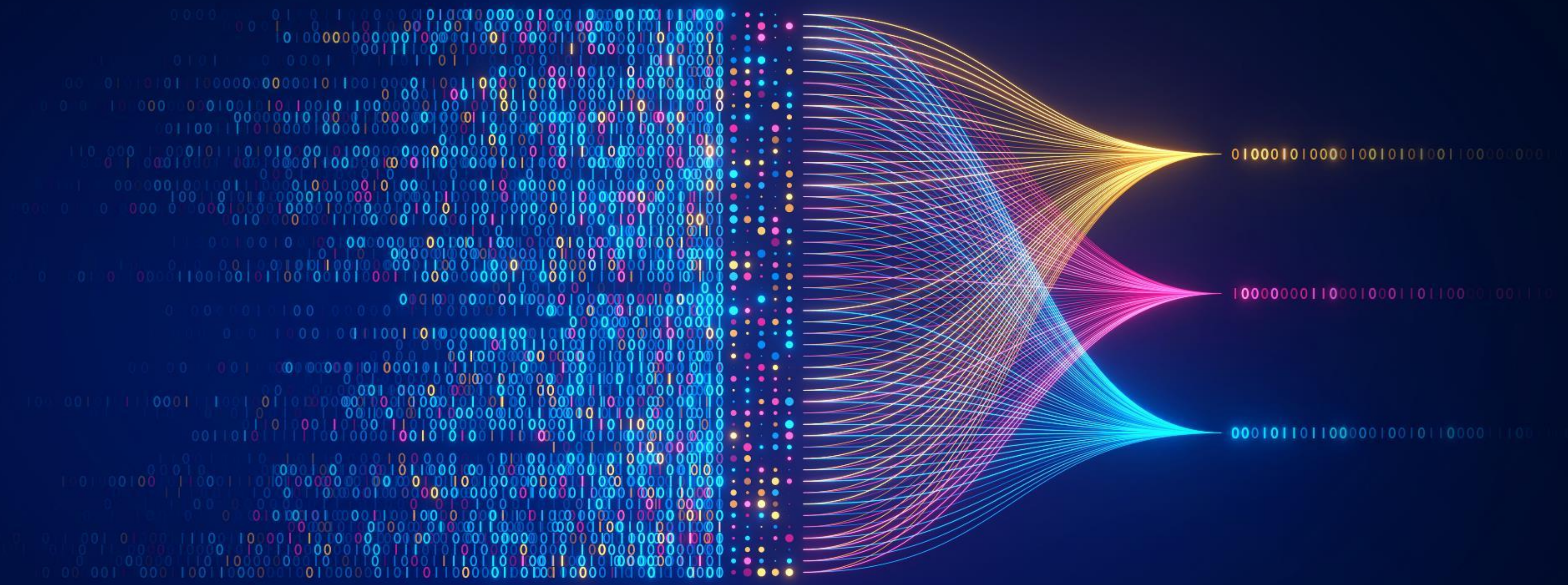
The key is finding ways to access data related to lower dimensional problems that can be annotated.

Or to collect the data in the first place with good annotation.

Or to reduce the dimensionality of a problem.

Progress in AI and drug discovery will require attacking low dimensional problems or simultaneously adopting methods that reduce problem dimensionality while improving data quality.

Approaches To Solving the Dimensionality Problem



Computational Chemistry Approaches Used for Dimensionality Reduction in Finding Small Molecule Drug Candidates with ML

	Physics First Approach	Quantitative SAR (Qsar)	Cellular Perturbation
How it's done:	Start with a good crystal model of your target. Then generate a series of drug leads using any number of traditional methods. Then screen using physics-based first principles to identify new drugs for targets that are aligned with good drug like properties. For example, applying physics means running a molecular dynamics simulation to compute the solubility of a molecule in water, or the affinity of the molecule for a particular protein, or its permeability.	Take a series of drug candidates and study their binding and drug-like properties using high-throughput assays. Then associate the results of these assays with drug structure using your favorite machine learning approach to identify positive aspects of a pharmacophore (structure activity relationships). Then use machine-recommended novel chemical structures to iteratively look for a drug that has good binding, specificity, and drug-like properties.	Find a tissue-specific cell model that is relevant to your disease target. For example, it might be liver stellate cells that excrete collagen causing fibrosis when subject to stress. Then introduce a wide range of pharmacologic interventions to the cell. Collect data on what happens using transcriptomics (RNAseq) or other outputs. One might even go so far as to tag proteins in a cell as does Eikon and then photograph what happens to those proteins with an intervention.
Illustrative Companies:	Iambic Qubit PHARMACEUTICALS RELAY THERAPEUTICS Schrödinger XtalPi	A2A PHARMA Atomwise KIMIA PostEra TERRAY	eikon therapeutics Genentech A Member of the Roche Group insitro KIMIA RECURSION immunai TransitionBio vevo
Comment:	Well-suited to finding drugs against targets where X-ray or Cryo-EM crystal structures exist. Not helpful in other contexts. Iambic's approach can also fish for ligand-protein binders.	Well-suited to contexts where crystal structures don't exist and where one would like to still find a good binding molecule. Not well suited to situations where a drug target is unknown.	Well-suited to contexts where biology is complex or unknown. If there is a disease-relevant cellular output one can study the effect of drug candidates without knowing a target.

Subsection

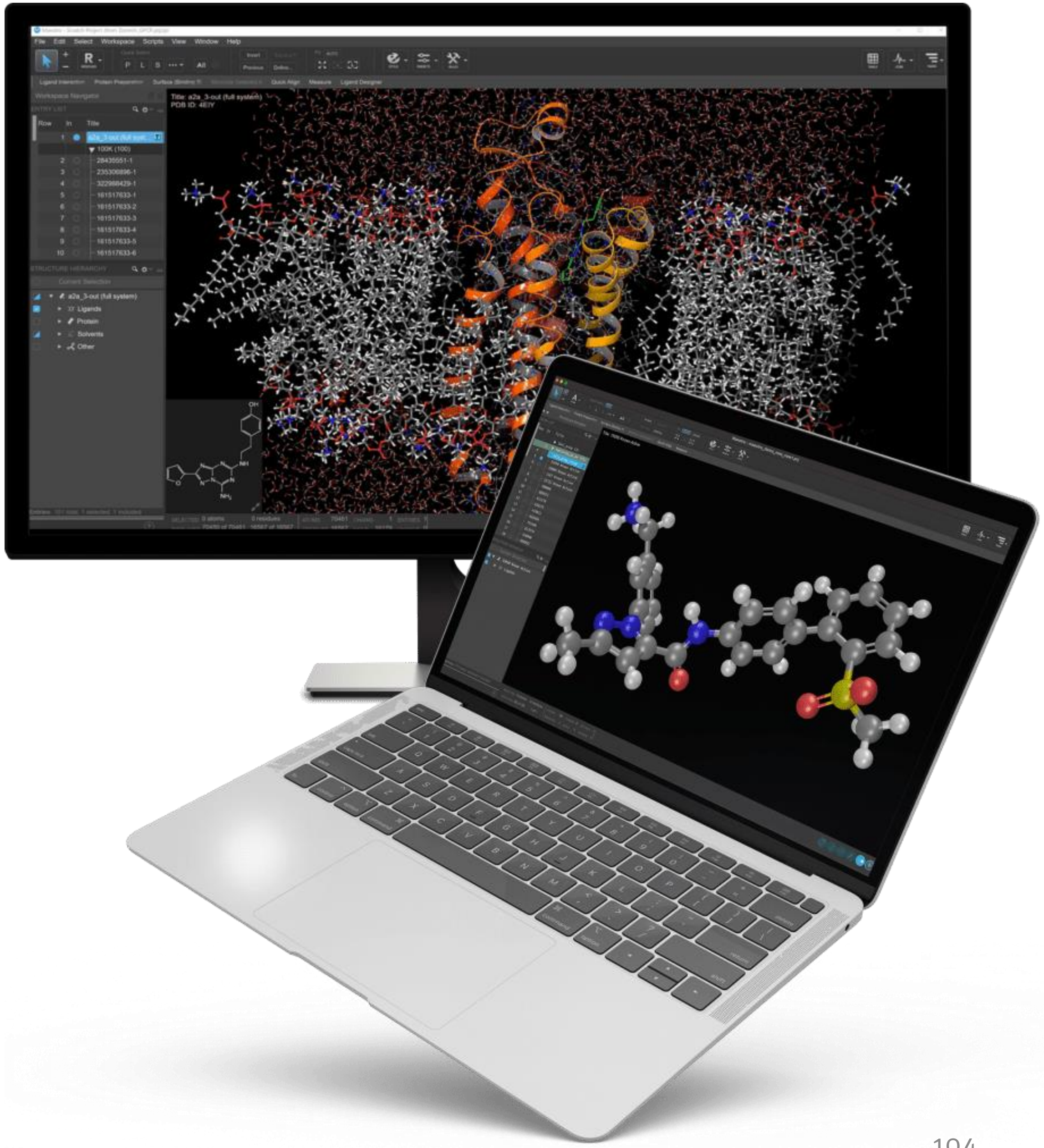
Examples of Companies that Take the Physics First Approach

Schrödinger

Computational Physics Platform to Find the Right Molecules from a Vast Chemical Space

Accelerating the convergence of physics, machine learning and enterprise informatics

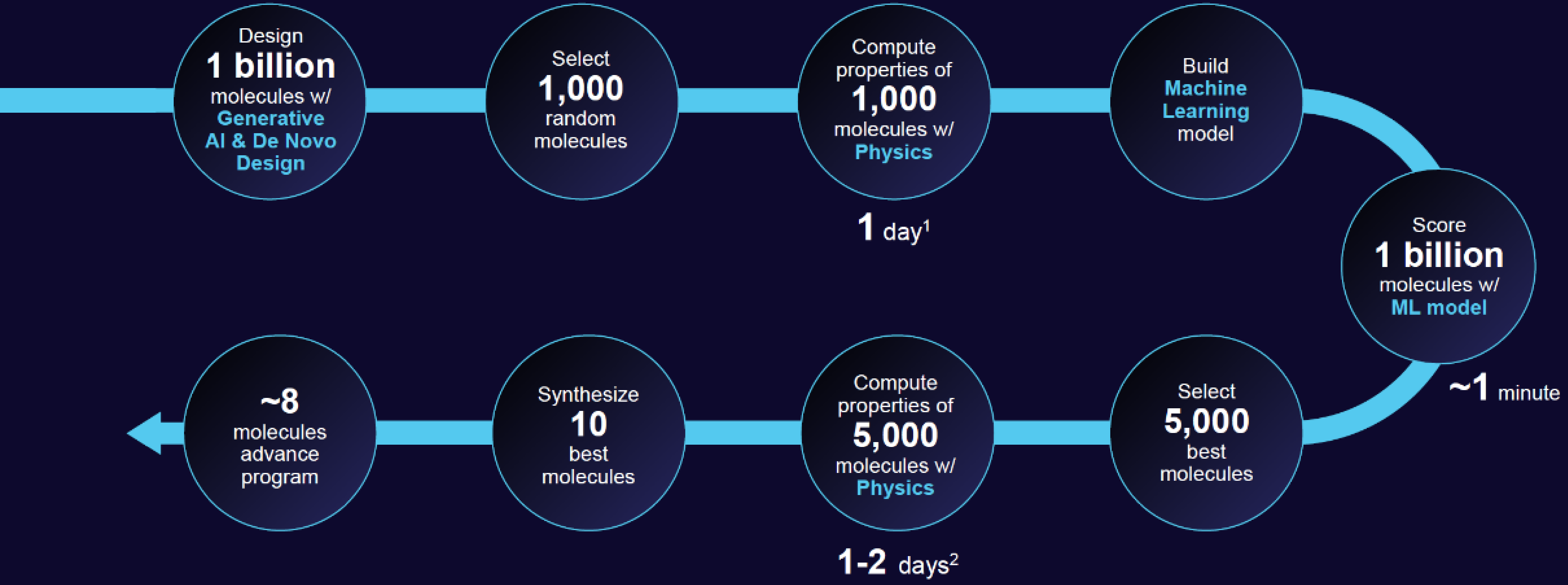
Built upon more than 30 years of R&D, our industry-leading computational platform is transforming the way therapeutics and materials are discovered by enabling highly accurate in silico predictions of key molecular properties across vast chemical space.



Source: https://newsite.schrodinger.com/platform/

Schrödinger Chemistry Approach Involves Massive Dimensionality Reduction

Physics used to produce sufficiently large representative training set for Machine Learning



Design **1 billion** molecules w/ **Generative AI & De Novo Design**

Select **1,000** random molecules

Compute properties of **1,000** molecules w/ **Physics**

1 day[1]

Build **Machine Learning** model

Score **1 billion** molecules w/ **ML model**

~1 minute

Select **5,000** best molecules

Compute properties of **5,000** molecules w/ **Physics**

1-2 days[2]

Synthesize **10** best molecules

~8 molecules advance program

[1] Would take ~**1 year** to do experimentally

[2] Would take ~**5 years** to do experimentally

Schrödinger

Schrödinger Uses Atomic Force Field Data and Free Energy Perturbation (FEP+) to Profile Possible Compounds for Drug-Like Properties

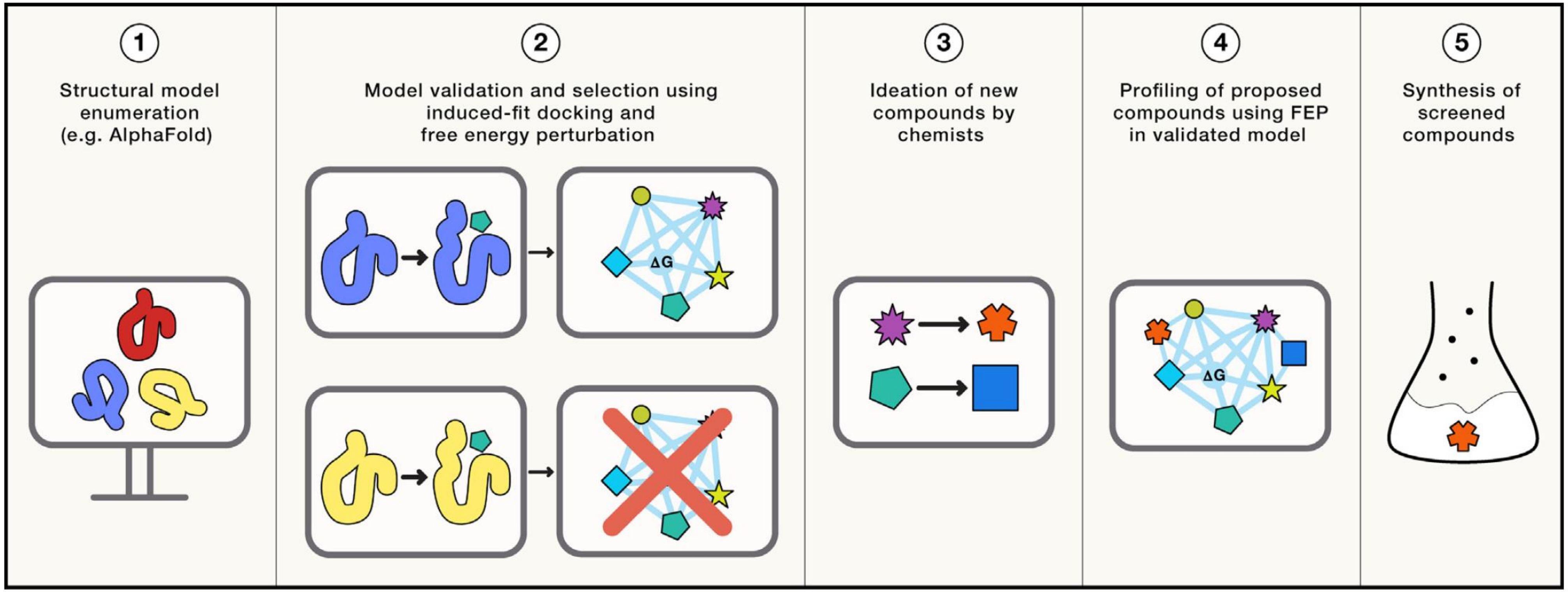

1
Structural model enumeration (e.g. AlphaFold)
2
Model validation and selection using induced-fit docking and free energy perturbation
ΔG
3
Ideation of new compounds by chemists
4
Profiling of proposed compounds using FEP in validated model
ΔG
5
Synthesis of screened compounds

Figure 2. Conceptual illustration of free energy perturbation for validation of a predicted structure and prospective deployment to screen novel chemical matter

(1) Predicted protein conformations are enumerated. Any method can be utilized here, for example, AlphaFold, or even alternatively refined crystal structures, as the structure will be subsequently validated.

(2) An existing set of congeneric ligands with an associated experimentally determined functional activity or binding affinity is used for model validation. One or more ligands from this series are induced-fit docked into the initial receptor. The correct model will yield a recapitulation of the experimental affinity data via FEP. Models that fail to satisfactorily reproduce experimental measurements are discarded.

(3) Visual inspection of the FEP validated model is used to guide ideation of novel chemical matter.

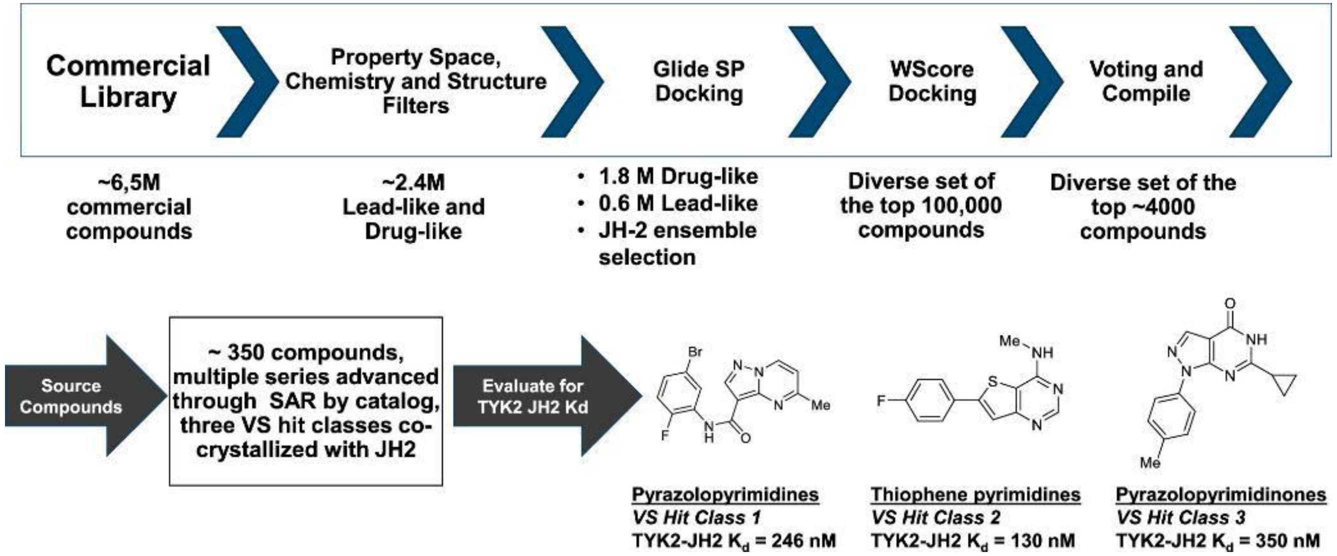
(4) Proposed compounds are prospectively screened in FEP, predicting binding affinity in advance of experiment.

(5) Molecules that are predicted to be satisfactory are sent to be synthesized and assayed for experimental validation.

Source: https://www.cell.com/cell/fulltext/S0092-8674(23)01444-7

Nimbus and Schrödinger Use X-Ray Co-Crystal Structure, Compound Screens and FEP+ Physics to Find a Great TYK2 Inhibitor

Leit S, Masse CE. et.al. Discovery of a Potent and Selective Tyrosine Kinase 2 Inhibitor: TAK-279. *J Med Chem*. Aug 10, 2023;66(15):10473-10496.



Commercial Library → Property Space, Chemistry and Structure Filters → Glide SP Docking → WScore Docking → Voting and Compile

~6,5M commercial compounds

~2.4M Lead-like and Drug-like

- 1.8 M Drug-like
- 0.6 M Lead-like
- JH-2 ensemble selection

Diverse set of the top 100,000 compounds

Diverse set of the top ~4000 compounds

Source Compounds → ~ 350 compounds, multiple series advanced through SAR by catalog, three VS hit classes co-crystallized with JH2 → Evaluate for TYK2 JH2 Kd

Pyrazolopyrimidines
VS Hit Class 1
TYK2-JH2 K_d = 246 nM

Thiophene pyrimidines
VS Hit Class 2
TYK2-JH2 K_d = 130 nM

Pyrazolopyrimidinones
VS Hit Class 3
TYK2-JH2 K_d = 350 nM

At the outset of this work, we leveraged X-ray co-crystal structures reported for compounds bound to the JH2 domain of TYK2 for a structure-based virtual screen (VS) of a large database of commercially available compounds. To limit the screen to lead-like and drug-like compounds, a library of ~6.5 million commercially available compounds was first narrowed to about 2.4 million compounds by structure filter restriction of property space (see Supporting Information for details). This smaller library was next assessed in a virtual screen (VS) with a JH2 ensemble structural selection developed from reported X-ray crystal structures from PDB ID codes using the Glide SP docking program with a hinge constraint. Next, a diverse set of the top ~100,000 compounds was assessed using WScore docking for a more accurate treatment of water molecules in the binding pocket. The WScore triage results were compiled, and a voting system was used to rank a structurally diverse set of the top ~3500 compounds. We then sourced ~340 compounds and measured their respective TYK2 JH2 domain binding potencies. Promising hits were progressed through SAR by catalog to identify similar neighbors. Ultimately, 3 structurally distinct hit classes were identified, including a pyrazolopyrimidine class, a thiophene–pyrimidine class, and a pyrazolopyrimidinone class (see figure). Although all three of these classes were promising leads, we were especially intrigued by the steric and electronic similarity between the pyrazolopyrimidine core heterocycle and the imidazopyridazine core of compound 6a.

To follow-up on the possibility of incorporating a scaffold-hop strategy to replace the imidazopyridazine ring of 6a with a pyrazolopyrimidine core similar to the core identified in VS hit class 1 from Figure 2, we applied the **physics-based computationally guided structure-based drug design** (SBDD) tool free energy perturbation (FEP+) to guide compound design in this scaffold-hop series. We used FEP+ to compute the binding potency of designed ligands for the JH2 and JH1 binding sites of TYK2, taking into account the entropic and enthalpic effects of ligands in solvent and in protein, as well as the local dynamics of protein residues and second-shell effects.

Source: https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.3c00600

XtalPi Predicts Molecule Success Using Automated Lead Optimization

ID4 Platform

Our comprehensive technology platform, *Inclusive Digital Drug Discovery and Development* (**ID4**) focuses on hit identification and lead optimization to produce validated preclinical candidates. Our platform is empowered by AI computation, laboratory experiments and research expertise in medicinal chemistry. ID4 includes molecule ideation, drug-like properties evaluation and optimization, ADMET properties prediction, chemical synthesis and biological functional studies. The ID4 platform aims to revolutionize traditional pharmaceutical R&D by delivering candidate compounds with improved speed, scale, novelty and diversity.

We recognize that every project is unique, and we look forward to working with each collaboration partner to provide a custom solution, leveraging parts or the entirety of our ID4 platform, that best suit the project's needs.

AI + Physics-based Model

We utilize AI to process data and generate predictions at scale. Built upon virtually limitless cloud computing resources, we have constructed over 200 proprietary AI models to evaluate key drug-like properties. We also embed AI within our physics-based algorithms to improve calculation efficiency without sacrificing accuracy. We are able to customize AI models as appropriate to improve the performance of our in silico predictions based on each individual project's unique needs.

Our AI + physics-based models are optimized with a cloud architecture that allows us to benefit from the security, scalability, flexibility and efficiency of cloud computing. The cloud architecture is designed for multi-cloud across geolocations and supported by leading public cloud service providers, and can scale up capacity to millions of cores to accelerate simulations. In addition, we adopt a cloud-native design of our computing architecture, which allows us to quickly update our software in response to the evolving industry requirements.

Source: https://www.xtalpi.com/en/DigitalDD

XtalPi Using 100 Robots to Perform Automated Lab Workflows in Shanghai Facility

How Quantum Physics and AI Are Disrupting Drug Discovery & Development

Pfizer AI Team, 2024:

But now thanks to a recent strategic research collaboration with XtalPi, a U.S.-China pharmaceutical tech company, Pfizer scientists are performing these calculations in a matter of days. Pioneered by a group of quantum physicists from MIT, the XtalPi technology leverages artificial intelligence and cloud computing to perform these complex equations. “A process that used to take us so much time that we almost didn’t event attempt it has become the norm, and now we can try it on almost every small molecule project,” says Bruno Hancock, Global Head of Materials Science at Pfizer’s Groton, Connecticut research site.

As one of the early collaborators with XtalPi, Pfizer, bringing its deep experience in the field, helped establish new techniques for early drug screening. “This collaboration is already changing the way Pfizer performs its screening work and has the potential to disrupt the industry as a whole,” says Geoff Wood, a Principal Scientist, also based at Pfizer’s Groton site.

While electrons are one of the smallest pieces of matter, predicting their movement involves a massive amount of computational power. To perform a single crystal structure prediction requires the computing power equivalent to one million laptops, says Hancock. “When they (cloud provider) do one of these calculations for us, XtalPi becomes the single biggest user of those services in the USA at that moment in time,” says Hancock.

At the microscale, what they’re calculating is the properties of electrons in a molecule. But since molecules have many electrons that are constantly moving and changing, they must perform multiple calculations simultaneously. “It can take billions of calculations to come up with a final answer,” says Hancock.

Once scientists can calculate the predicted 3-D structure of a drug molecule, they can use it to predict its mechanical and chemical properties, such as its shape, solubility and melting point, and how it binds with a protein receptor.

Source: https://www.pfizer.com/news/articles/how_quantum_physics_and_ai_is_disrupting_drug_discovery_development

Helping Pfizer Accelerate Development of Novel COVID-19 Oral Antiviral Drug PAXLOVID®

XtalPi Website, 2024:

“PAXLOVID is the world's first FDA-approved oral COVID-19 drug developed by Pfizer and sold around the world. To expedite the development of PAXLOVID (PF-07321332), Pfizer and XtalPi worked closely together, combining XtalPi's digital prediction algorithm and experimental validation. It took only six weeks for the teams to complete mutual validation and precise matching of drug crystal structure prediction against the experimental results, making possible the subsequent development and production.

XtalPi's computational prediction provided powerful evidence of the crystal structure designed by Pfizer being the most stable crystal structure under room temperature, thus making it suitable for scale-up and production. In this way CMC scientists were able to rapidly make research decisions and begin the process without delay.

As an oral drug, PAXLOVID was developed in solid state to facilitate storage and transportation. Patients can self-administer the drug at home, leading to greater compliance, thus helping to alleviate the tremendous strain on the medical system at the height of the COVID-19 pandemic.”

Source: https://www.xtalpi.com/en/about

Stonewise Labs Generates Molecule Binders Using Physics Principles

StoneWise 望石智慧

Wang, L., Bai, R., Shi, X. et al. A pocket-based 3D molecule generative model fueled by experimental electron density. Sci Rep 12, 15100 (2022).

We report for the first time the use of experimental electron density (ED) as training data for the generation of drug-like three-dimensional molecules based on the structure of a target protein pocket. Similar to a structural biologist building molecules based on their ED, our model functions with two main components: a generative adversarial network (GAN) to generate the ligand ED in the input pocket and an ED interpretation module for molecule generation. The model was tested on three targets: a kinase (hematopoietic progenitor kinase 1), protease (SARS-CoV-2 main protease), and nuclear receptor (vitamin D receptor), and evaluated with a reference dataset composed of over 8000 compounds that have their activities reported in the literature. The evaluation considered the chemical validity, chemical space distribution-based diversity, and similarity with reference active compounds concerning the molecular structure and pocket-binding mode. Our model can generate molecules with similar structures to classical active compounds and novel compounds sharing similar binding modes with active compounds, making it a promising tool for library generation supporting high-throughput virtual screening.

Source: https://www.nature.com/articles/s41598-022-19363-6

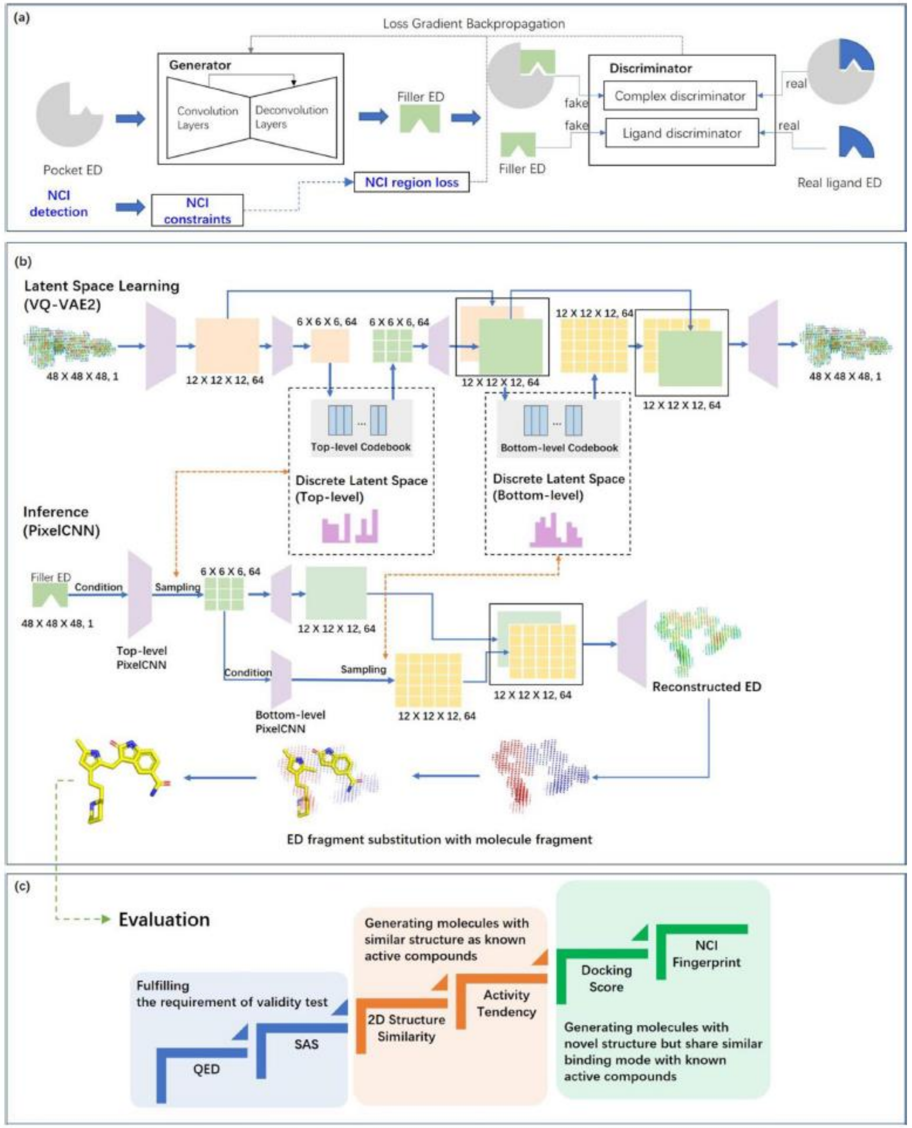


Figure 1. Model architecture. (**a**) The GAN for generating filler ED based on pocket ED. (**b**) ED interpretation module for molecule generation. VQ-VAE2 and PixelCNN used for latent space construction and autoregressive sampling, as well as the subsequent process of ED fragment substitution are illustrated. (**c**) The evaluation framework for generated molecules.

Subsection

Examples of Companies that Take the QSAR + ML Approach

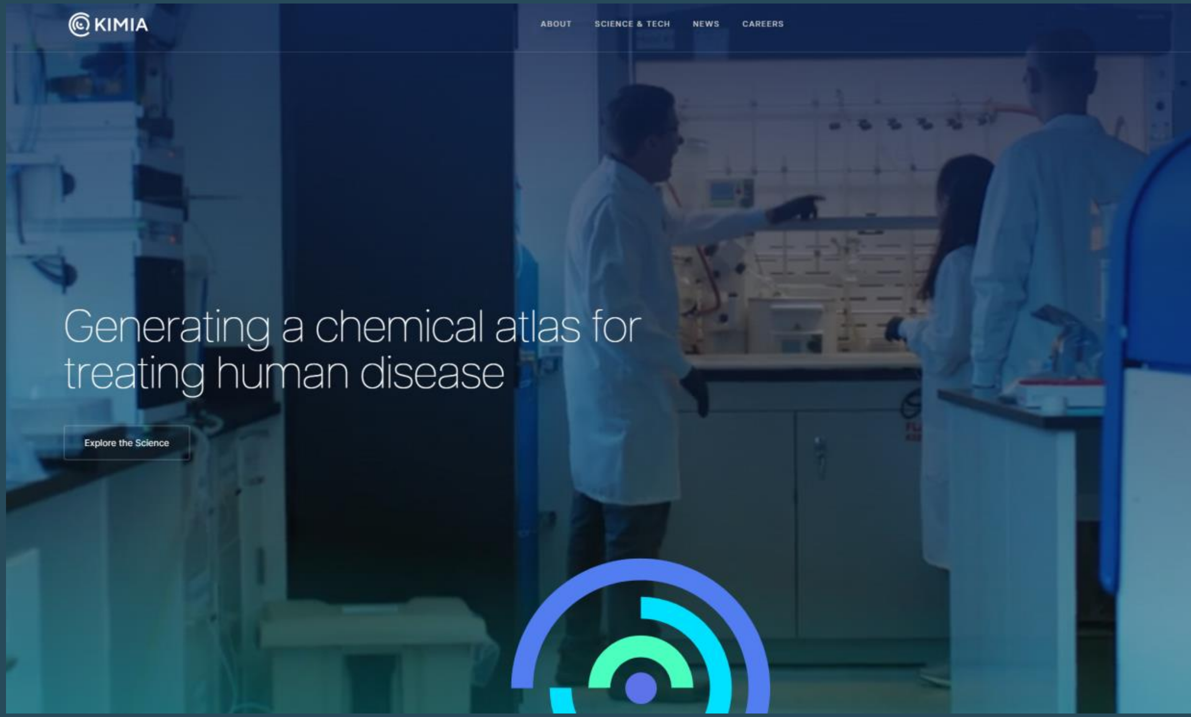
Kimia Has Had High Success By Generating Compound Candidates Using a QSAR Based Approach Linked to Machine Learning

Kimia can rapidly explore a large chemical space for compounds that bind to a target biological molecule through automated chemical synthesis and screening combined with active learning algorithms.

Chemical synthesis is performed at nanogram-scale (as opposed to milligram-scale that is industry standard) using building blocks and reaction conditions that are informed by ML models trained on internally generated data.

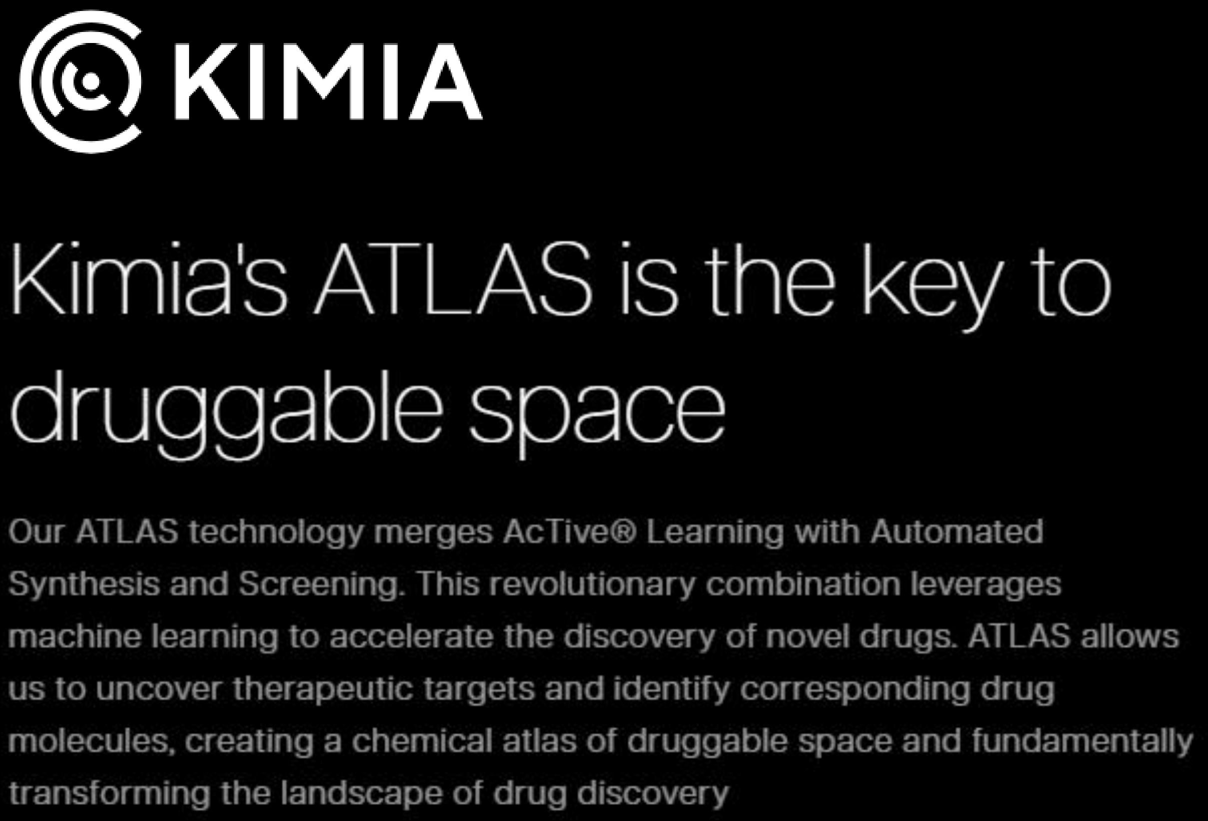
Screening is done biochemically and in cells against known or novel (unknown) targets.

Chemical synthesis, assay data (QSAR), ADME data etc are then used to train an active learning model that predicts the next round of compounds for library synthesis and screening. This is an iterative process that continues until a candidate is identified.



Generating a chemical atlas for treating human disease

A Key Aspect of Kimia’s Long-Term Vision is to Create a Chemical Map of the Druggable Proteome to Accelerate Drug Discovery


KIMIA

Kimia's ATLAS is the key to druggable space

Our ATLAS technology merges AcTive® Learning with Automated Synthesis and Screening. This revolutionary combination leverages machine learning to accelerate the discovery of novel drugs. ATLAS allows us to uncover therapeutic targets and identify corresponding drug molecules, creating a chemical atlas of druggable space and fundamentally transforming the landscape of drug discovery

Kimia’s platform is called ATLAS, an acronym for AcTive Learning through Automated Synthesis and screening.

A Conversation with Kimia's CEO, Stig Hansen

Q: Why haven't we seen more success from drug discovery with AI in recent years?

A: It's really hard to find drugs with AI without good data. You need lots of high-quality data to get it done. It's very hard to get AI/ML to work if you just pull datasets on drugs out of the literature. That's a fallacy. It's why we haven't seen more success. People in the field haven't been generating their own well-curated data at scale.

Q: Can you elaborate a bit more?

A: If you train ML models on datasets under vastly different conditions, the ML models may pick up artifacts that are not due to small molecule / target interactions but rather something linked to environmental factors. It's why we do it all in-house at Kimia and try not to get fooled.

Q: So, what exactly does Kimia do to enhance the chances of success?

A: We synthesize compounds in nanoliter quantities using our automated high-throughput chemistry platform combined with automated assays. We can do this with thousands of compounds a day which allows us to train machine learning models using consistent internally generated datasets very quickly.

Stig Hansen, CEO, Kimia Therapeutics

Source: Stifel interview with Stig Hansen, April 2024

Stig Hansen Conversation (continued)

Q: How would you compare Kimia to other AI drug discovery companies?

Without naming names, some of the early implementers have run through a lot of money and have not been that successful. Weren't those companies supposed to be more efficient? A number of these companies have raised $100's of millions. Where are the results? Show me the approved molecules. It was going to be a revolution, right? Many companies have not been able to overcome the key bottlenecks in drug developments. Most importantly, many early implementers have not generated their own data and have instead worked with messy external data.

Q: How about Schrödinger?

Schrödinger have gotten good drugs approved and is now well into Phase 3. Schrödinger is very structure driven. Their focus is understanding chemical attributes of the binding site and the physical properties. That's great when you have structural information available. Such as CRYO-EM or X-ray crystallography. Obviously, many times you don't have such information. That's where a company like Kimia can excel.

Q: How about groups like Eikon and Recursion?

Recursion does something completely different than what we do. They start by looking at how known molecules perform in cellular assays that replicate disease processes. Arguably, Eikon is even more advanced by tracking protein motion. These cell-based companies are best positioned to make breakthroughs in areas involving complex biology and improve our understanding of proteins/pathways in disease. But how will this translate to new and better drugs? Where are the actual drug molecules going to come from? At the end of the day, the money is in the molecule.

Q: What do you see happening next?

There are several exciting next generation companies like Isomorphic. What they do will likely revolutionize how we identify new binding sites and find chemical starting points for drug discovery. In the long-term, how can we all make things better? If we can't speed up clinical development all our investment in accelerating drug discovery could be diminished. Unfortunately, little has changed with clinical development. How do we overcome today's expensive, slow clinical trial process?

Stig Hansen and Colleagues at the Whiteboard

Stig Hansen Conversation (continued)

Q: What is your outlook on AI/ML in drug discovery overall?

A: I am very bullish on the space. If you take a well thought out approach and have high quality input data, ML can deliver amazing results. We are seeing this at Kimia already where we combine empirical data with computational analytics.

Q: Why does the Kimia approach work?

A: Kimia employs a strategy also known as QSAR – but on steroids because it is performed at massive scale. We generate biological data for structurally related – not random – compounds. . This allows us to understand the relationship between structure and activity.

Maybe, the way to think about it, is that you can start with a tool compound or a suboptimal compound from the literature. This is what we did the with Carmot obesity molecule that Roche bought. Using our compound evolution approach, the computer quickly gets much more diversity to work with, and importantly this is relevant diversity, not random diversity. This allows you to generate a predictive model. Then, the computer can make and test new molecules in the computational sphere. A lot of them (a billion even). You then have the computer prioritize, and we make, say 20,000 out of the pile. We can make and test them in a matter of days. Even if hit rate is only 1% you win. And with each round of training or active learning, the computer model becomes better at predicting.

We also apply this approach in cells where we can generate high resolution QSAR data for disruption of disease pathways even if we don't know what the target is. The algorithms will be able to cluster compounds in a target specific manner. We then use select compounds to de-convolute and identify the target. The result is a cell active molecule for a novel target or a preciously known target in a new context.

Q: So, what do you end up with?

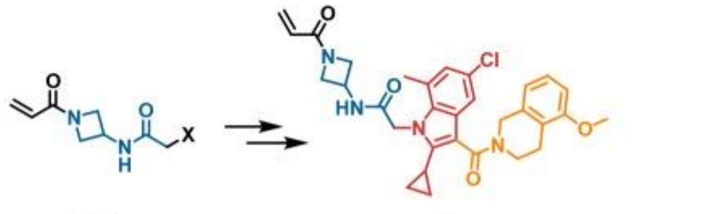
A: We end up with a Pharmacophore Model. It will tell you how a drug binds and why. For example, we might note that all active drugs have an aromatic ring in one position and a hydrophobic moiety in another. While we don't have that co-crystal structure, we get a very good picture of what a binding site can accommodate.

We can then use that information to drive optimization through computational strategies. Ideally, we would like to use dynamic three-dimensional simulations to train models, but the computing power is still too expensive.

Kimia Predecessor Company, Carmot, Generated Amgen's LUMAKRAS® Using Its Chemical Evolution and QSAR Approach

Shin Y, Jeong JW, Wurz RP, Achanta P, Arvedson T, Bartberger MD, Campuzano IDG, Fucini R, Hansen SK, Ingersoll J, Iwig JS, Lipford JR, Ma V, Kopecky DJ, McCarter J, San Miguel T, Mohr C, Sabet S, Saiki AY, Sawayama A, Sethofer S, Tegley CM, Volak LP, Yang K, Lanman BA, Erlanson DA, Cee VJ. Discovery of N-(1-Acryloylazetidin-3-yl)-2-(1H-indol-1-yl)acetamides as Covalent Inhibitors of KRASG12C. *ACS Med Chem Lett*. 2019 Aug 20;10(9):1302-1308.

KRAS regulates many cellular processes including proliferation, survival, and differentiation. Point mutants of KRAS have long been known to be molecular drivers of cancer. *KRAS p.G12C*, which occurs in approximately 14% of lung adenocarcinomas, 3–5% of colorectal cancers, and low levels in other solid tumors, represents an attractive therapeutic target for covalent inhibitors. **Herein, we disclose the discovery of a class of novel, potent, and selective covalent inhibitors of KRAS^{G12C} identified through a custom library synthesis and screening platform called Chemotype Evolution and structure-based design. Identification of a hidden surface groove bordered by H95/Y96/Q99 side chains was key to the optimization of this class of molecules.** Best-in-series exemplars exhibit a rapid covalent reaction with cysteine 12 of GDP-KRAS^{G12C} with submicromolar inhibition of downstream signaling in a KRAS^{G12C}-specific manner.



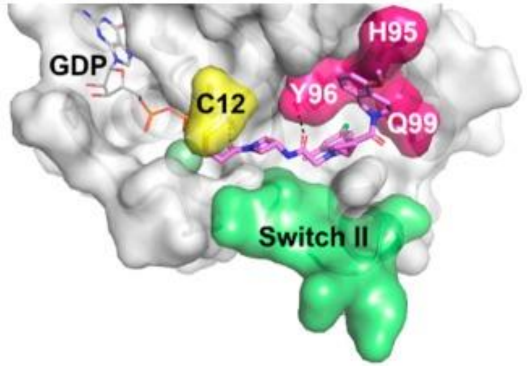
Bait

1

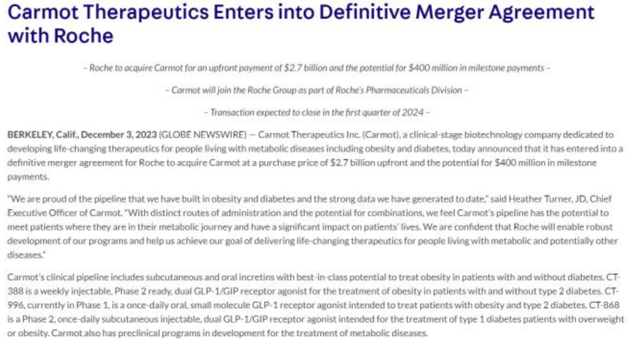
KRAS^{G12C} $k_{obs}/[I]$ = 2640 $M^{-1}s^{-1}$

MIA PaCa-2 p-ERK 4 h IC_{50} = 0.219 µM

MIA PaCa-2 Viability 72 h IC_{50} = 0.067 µM



Kimia Platform Used to Design Compounds that led to $2.7bn purchase of Carmot Therapeutics by Roche in 2023



Carmot Therapeutics Enters into Definitive Merger Agreement with Roche

– Roche to acquire Carmot for an upfront payment of $2.7 billion and the potential for $400 million in milestone payments –

– Carmot will join the Roche Group as part of Roche's Pharmaceuticals Division –

– Transaction expected to close in the first quarter of 2024 –

BERKELEY, Calif., December 3, 2023 (GLOBE NEWSWIRE) — Carmot Therapeutics Inc. (Carmot), a clinical-stage biotechnology company dedicated to developing life-changing therapeutics for people living with metabolic diseases including obesity and diabetes, today announced that it has entered into a definitive merger agreement for Roche to acquire Carmot at a purchase price of $2.7 billion upfront and the potential for $400 million in milestone payments.

"We are proud of the pipeline that we have built in obesity and diabetes and the strong data we have generated to date," said Heather Turner, JD, Chief Executive Officer of Carmot. "With distinct routes of administration and the potential for combinations, we feel Carmot's pipeline has the potential to meet patients where they are in their metabolic journey and have a significant impact on patients' lives. We are confident that Roche will enable robust development of our programs and help us achieve our goal of delivering life-changing therapeutics for people living with metabolic and potentially other diseases."

Carmot's clinical pipeline includes subcutaneous and oral incretins with best-in-class potential to treat obesity in patients with and without diabetes. CT-388 is a weekly injectable, Phase 2 ready, dual GLP-1/GIP receptor agonist for the treatment of obesity in patients with and without type 2 diabetes. CT-996, currently in Phase 1, is a once-daily oral, small molecule GLP-1 receptor agonist intended to treat patients with obesity and type 2 diabetes. CT-868 is a Phase 2, once-daily subcutaneous injectable, dual GLP-1/GIP receptor agonist intended for the treatment of type 1 diabetes patients with overweight or obesity. Carmot also has preclinical programs in development for the treatment of metabolic diseases.

Sources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6746093/, https://carmot.us/carmot-therapeutics-enters-into-definitive-merger-agreement-with-roche/

Atomwise Uses ML to Predict Drug Efficacy Based on Target Chemical Analysis

The “key and lock” model gives us a way to categorize the techniques that AtomNet® technology is inspired by. Lock-oriented techniques, which are called “structure-based” algorithms, look to the composition of the target protein to guide their predictions. This approach is appealing because, in principle, it works for totally novel targets. Accordingly, a variety of software packages have been introduced, such as Dock, AutoDock, and Glide. The main limitation of these technologies is their accuracy. In general, these methods have a high rate of false-positives (saying a molecule is a good candidate when it is not). Many researchers remain skeptical of their usefulness.

An ideal solution then would be a structure-based rational drug design system that is also highly accurate. Such a system could predict candidate molecules for new and challenging drug discovery targets, and have a reasonable chance of those predictions proving correct – giving researchers what they need from a virtual drug discovery method. We think AtomNet® technology is a big step in that direction.

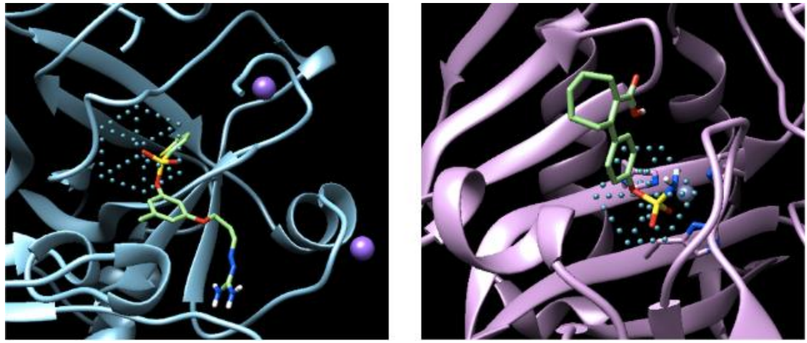
AtomNet® technology is the first drug discovery algorithm to use a deep convolutional neural network. This type of network came to prominence only a few years ago and has a unique property: it excels at understanding complex concepts as a combination of smaller and smaller pieces of information. This property is a key reason why convolutional networks have produced the world’s best results for image classification, speech recognition, and other longstanding problems. For example, a convolutional model can learn to recognize faces by first learning a set of basic features in an image, such as edges. Then, the model can learn to identify parts such as noses, ears, and eyes by combining the edges. Finally, the model can learn to recognize faces by combining those parts.

Similarly, AtomNet® technology might learn that proteins and ligands are made up of a variety of specialized chemical structures. This would be an exciting result because it would suggest that AtomNet® model was learning fundamental concepts in organic chemistry. Intriguingly, this is what AtomNet® platform does. When we examine different neurons on the network we see something new: AtomNet® platform has learned to recognize essential chemical groups like hydrogen bonding, aromaticity, and single-bonded carbons.

Critically, no human ever taught our AtomNet® technology the building blocks of organic chemistry. Our AtomNet® model discovered them itself by studying vast quantities of target and ligand data. The patterns it independently observed are so foundational that medicinal chemists often think about them, and they are studied in academic courses. Put simply, AtomNet® technology is teaching itself college chemistry.

Source: https://blog.atomwise.com/introducing-atomnet-drug-design

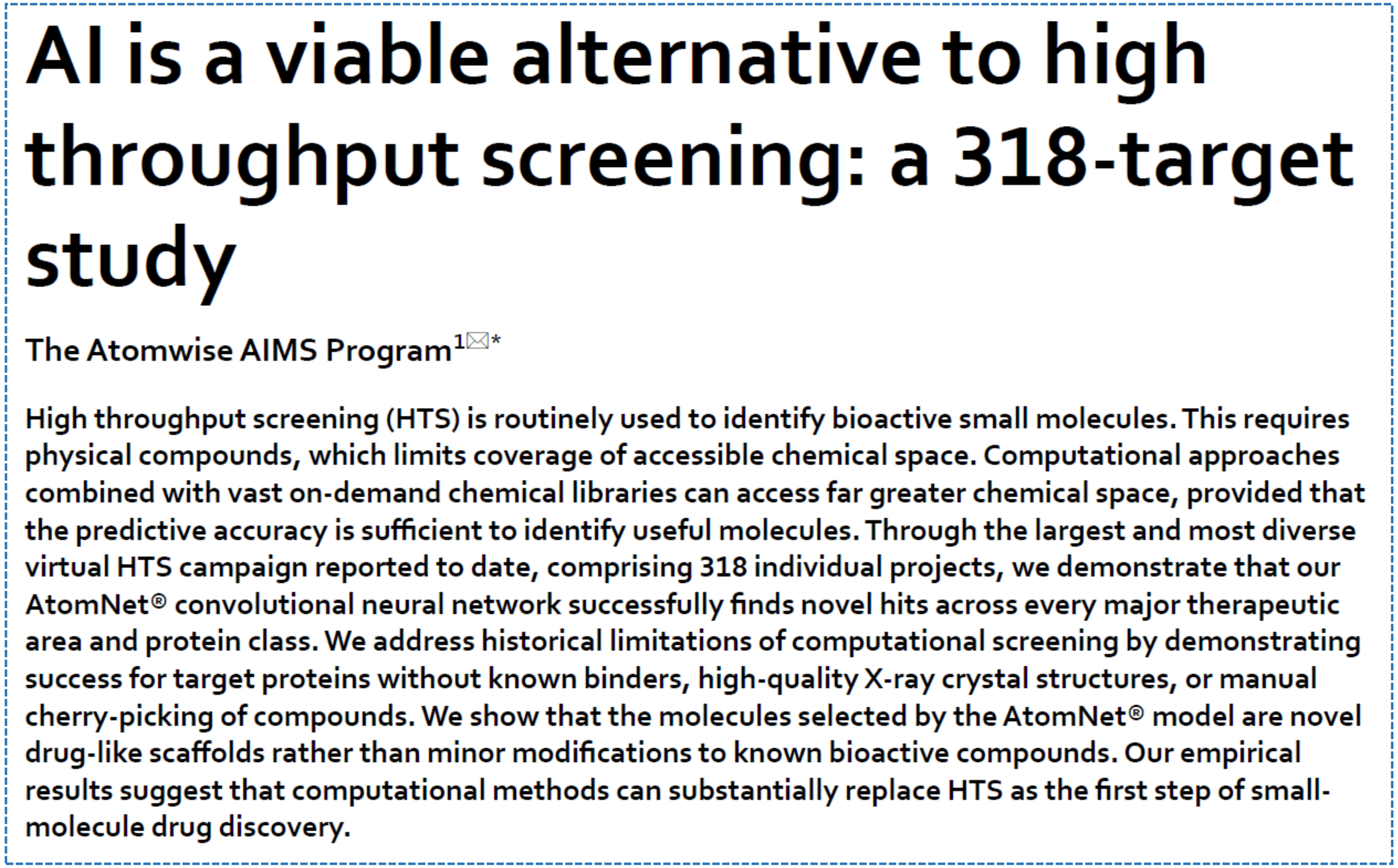
Atomwise



AtomNet® model learning to recognize sulfonyl groups – a structure often found in antibiotics.

Atomwise Study Argues that AI Small Molecule Drug Design Beats Traditional HTS Methods

Nature Scientific Reports, April 2, 2024



AI is a viable alternative to high throughput screening: a 318-target study

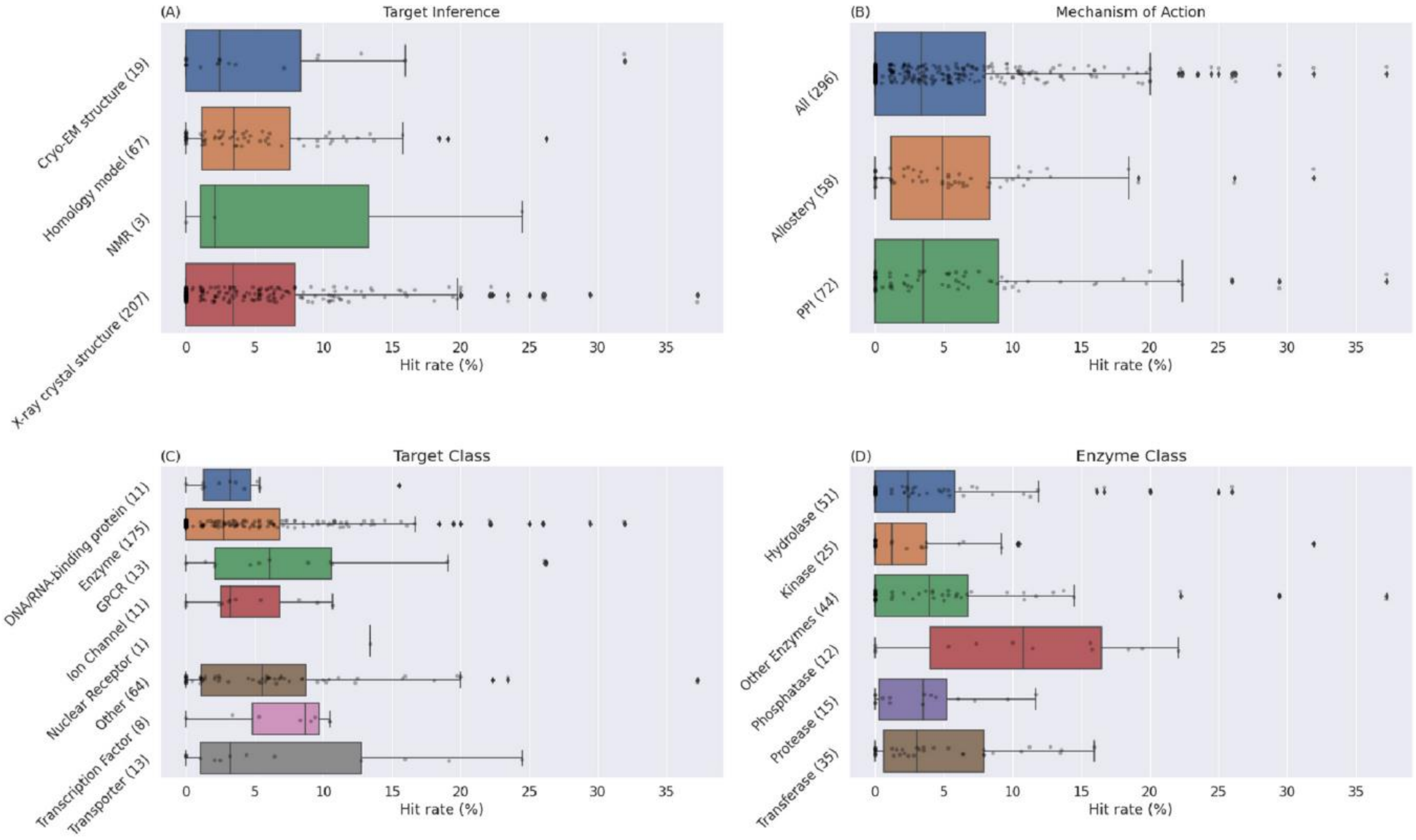
The Atomwise AIMS Program[1]✉*

High throughput screening (HTS) is routinely used to identify bioactive small molecules. This requires physical compounds, which limits coverage of accessible chemical space. Computational approaches combined with vast on-demand chemical libraries can access far greater chemical space, provided that the predictive accuracy is sufficient to identify useful molecules. Through the largest and most diverse virtual HTS campaign reported to date, comprising 318 individual projects, we demonstrate that our AtomNet® convolutional neural network successfully finds novel hits across every major therapeutic area and protein class. We address historical limitations of computational screening by demonstrating success for target proteins without known binders, high-quality X-ray crystal structures, or manual cherry-picking of compounds. We show that the molecules selected by the AtomNet® model are novel drug-like scaffolds rather than minor modifications to known bioactive compounds. Our empirical results suggest that computational methods can substantially replace HTS as the first step of small-molecule drug discovery.

In 215 projects, we identified at least one bioactive compound for the target in a biochemical or cell-based assay. **This 73% success rate substantially improves over the ~50% success rate for HTS.** On average, we screened 85 compounds per project and discovered 4.6 active hits, with an average hit rate of 5.5%. For the subset of targets where we found any hits, the average was 6.4 hits per project. Thus, we achieved an average hit rate of 7.6%, which again compares favorably with typical HTS hit rates.

Source: https://www.nature.com/articles/s41598-024-54655-z

Atomwise Study Finds Reasonable Compound Hit Rates No Matter What the MOA, Enzyme Class, Target Class or Target Inference Method Used



Hit rates obtained for the 296 AIMS projects. (A) A comparison of hit rates using X-ray crystallography, NMR, Cryo-EM, and homology for modeling the structure of the proteins. Each point represents a project with the x-axis denoting the hit rate of the project (the percentage of molecules tested in the project that were active). The number of projects of each type is given in parentheses. We observed no substantial difference in success rate between the physical and the computationally inferred models. We achieved average hit rates of 5.6%, 5.5%, and 5.1% for crystal structures, cryo-EM, and homology modeling, respectively. The number of projects using NMR structures is too small to make statistically-robust claims. (B) A comparison of hit rates observed for traditionally challenging target classes such as protein–protein interactions (PPI) and allosteric binding. Of the 296 projects, 72 targeted PPIs and 58 allosteric binding sites. The average hit rates were 6.4% and 5.8% for PPIs and allosteric binding, respectively. (C) Comparison of hit rates observed for different target classes and (D) enzyme classes. No protein or enzyme class falls outside the domain of applicability of the algorithm.

Source: https://www.nature.com/articles/s41598-024-54655-z

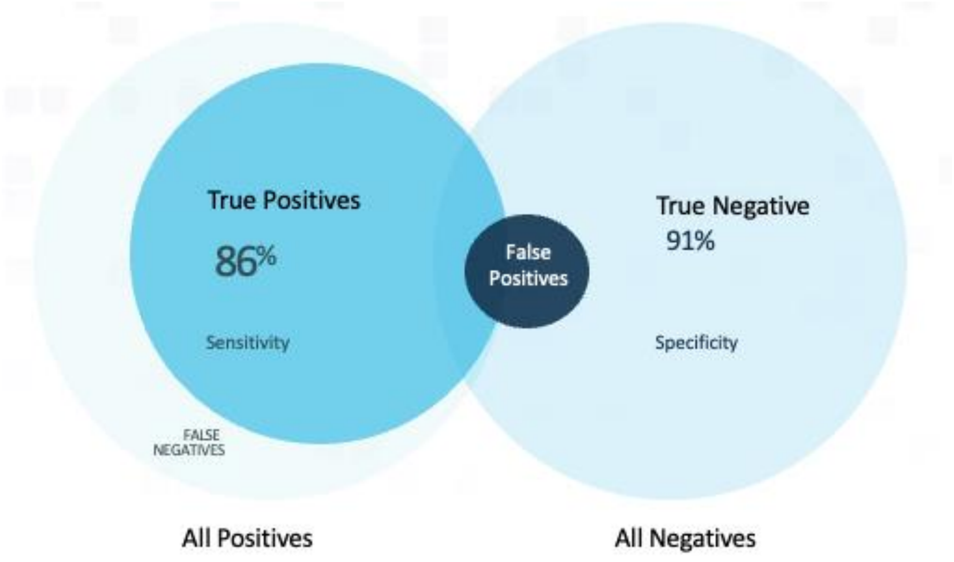
GATC Health Optimizes for Compounds with Good Druglike Properties

GATC Health is an AI-based technology company that optimizes drug discovery and development to improve pre-clinical and clinical successes while reducing time and cost. GATC's Multiomics Advanced Technology™ (MAT) AI platform predicts molecular target activation for novel chemical entities, including detailed predictions of efficacy, safety, and non-obvious off-target effects, greatly de-risking and expediting preclinical process and optimizing clinical work.

GATC's MAT AI platform turns around the high-risk drug development model by rapidly discovering and predicting the safest and most effective novel compound to address specific diseases: two studies demonstrating greater than 80% success rate vs industry's 90% failure rate (11x over current standards). GATC reduces the set of candidate drugs for clinical testing to several promising compounds in months, instead of the typical 5,000 compounds that require years of vetting and testing.

GATC Health has a renowned Board of Advisors including Herb Boyer, co-founder of Genentech, the first biotech to go public; Tomas Philipson, PhD, former Acting Chairman of the Council of Economic Advisers in the White House and Advisor to FDA, and others.

GATC Health



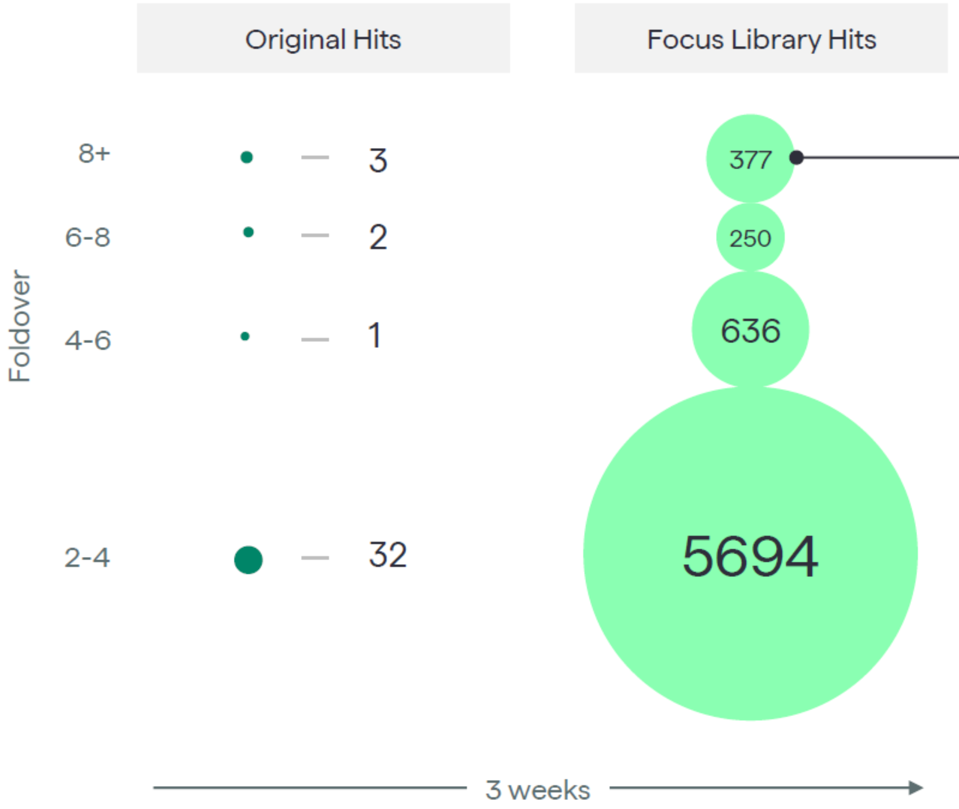
In a double-blinded *in-silico* study, GATC identified drugs that were successful in clinical trials with 86% accuracy and identified drugs that failed in clinical trials at a rate of 91%.

Source: GATC Health

Terray Doing QSAR with MultiComponent Binding Optimization

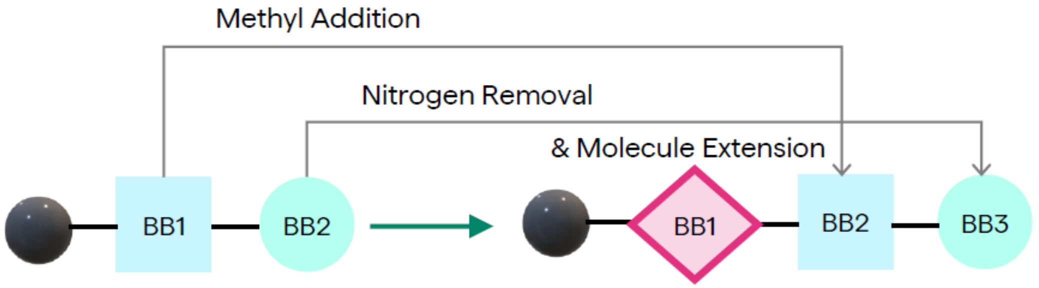
Iteration: 10,000x Faster than Traditional Approaches

In <1 month, we test over 2 million new AI-designed molecules *per program*.



From 38 original hits to ~7,000

As an example, one of the 377 datapoints from the 8+ foldover range is shown below:

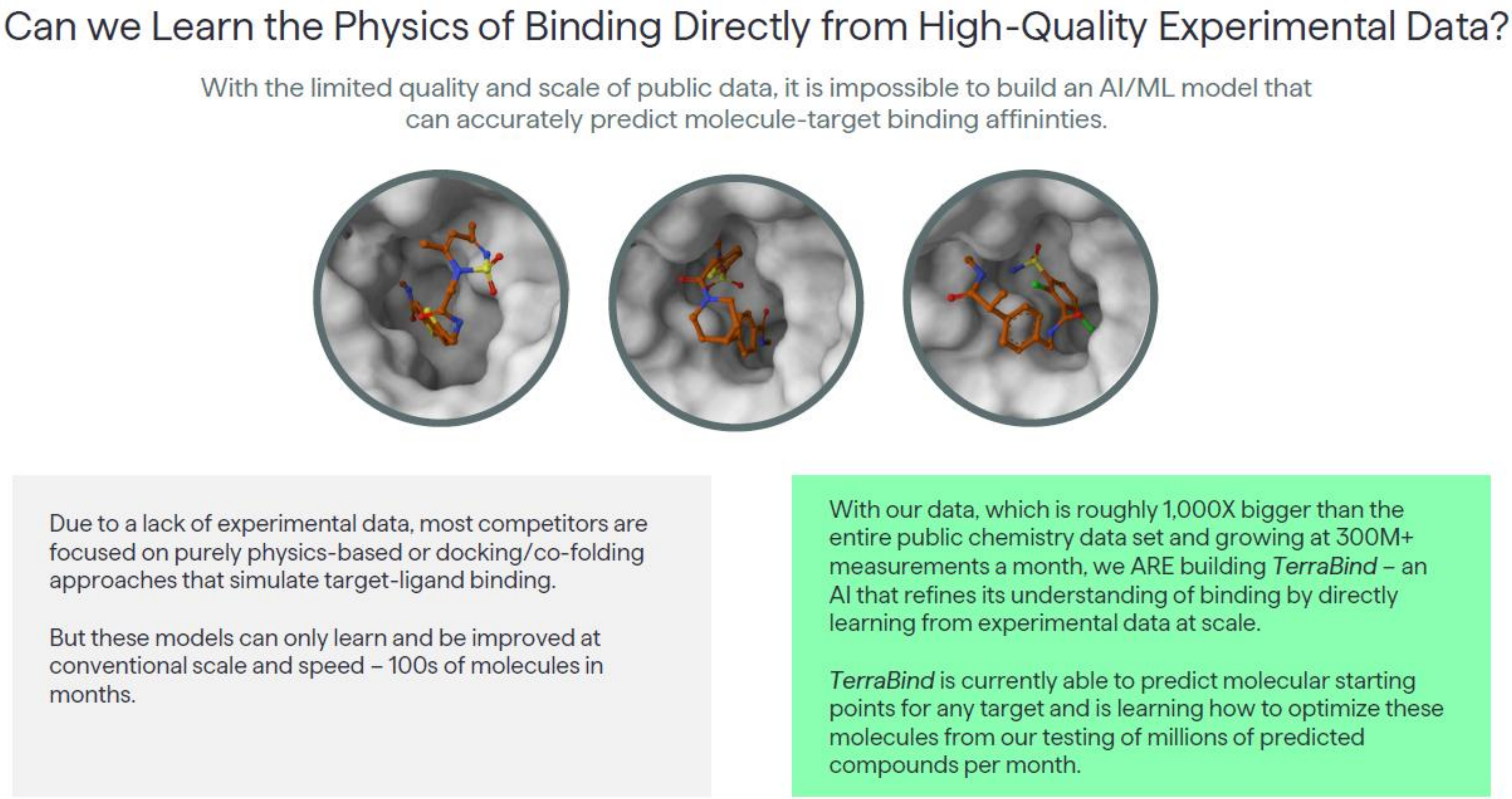


	Original 2-cycle	New 3-cycle	Improvement
Foldover	1.34	22.63	16.9x
IC50	7.32 µM →	89 nM	82x

82x improvement in biochemical IC50 in 1 iteration.
Simultaneous optimization of 3 components.

Source: Terray Therapeutics Investor Presentation, April 2024, (https://www.terraytx.com/).

Terray Argues that High Throughput QSAR Better than Physics Approach



Can we Learn the Physics of Binding Directly from High-Quality Experimental Data?

With the limited quality and scale of public data, it is impossible to build an AI/ML model that can accurately predict molecule-target binding affinities.

Due to a lack of experimental data, most competitors are focused on purely physics-based or docking/co-folding approaches that simulate target-ligand binding.

But these models can only learn and be improved at conventional scale and speed – 100s of molecules in months.

With our data, which is roughly 1,000X bigger than the entire public chemistry data set and growing at 300M+ measurements a month, we ARE building *TerraBind* – an AI that refines its understanding of binding by directly learning from experimental data at scale.

TerraBind is currently able to predict molecular starting points for any target and is learning how to optimize these molecules from our testing of millions of predicted compounds per month.

Source: Terray Therapeutics Investor Presentation, April 2024, (https://www.terraytx.com/).

A2A's QSAR / AI Engine Designs Biomea's Covalent Menin Inhibitor

SCULPT

OVERVIEW

- **S**ystematic **C**ombinatorial **U**nification of fragments into **L**ibraries against a **P**harmacological **T**arget (SCULPT^{TM})
- Fragment-based drug design platform used to develop candidates against intractable targets
- Designs and evaluates up to 100M+ novel compounds
- Iterated until candidates with optimal properties and matches to target features are obtained
- **Machine learning** and **artificial intelligence** tools target specific workflows in the iterative process

THE ADVANTAGES

✓ AI driven integration of ligand-based data into a primarily structure-based approach

✓ 3D dynamic hotspot template optimization to enable proprietary, target specific, novel library creation

✓ Proprietary ADMET property optimization in each round of in silico iteration

✓ Minimization of wet-lab synthesis and testing with computational pre-optimization of complete synthetically feasible molecules rather than just fragments

✓ Target specific workflow optimization to enable application to protein-protein interactions, kinase and degrader approaches across multiple diseases

IMPRESSIVE RESULTS

MLL-Menin (JV with Biomea)
- 12 molecules synthesized
- **4 Potent hits**
- Lead molecule – Phase 1 initiated

TACC3, TYK2
- Selective, potent hits across lead optimization targets

YAP-TEAD
- 28 molecules synthesized
- **9 Potent hits**
- IC50 0.2– 20 µM

KRAS
- 5 molecules synthesized
- **5 Potent binder hits**
- 400-700 nM K_D

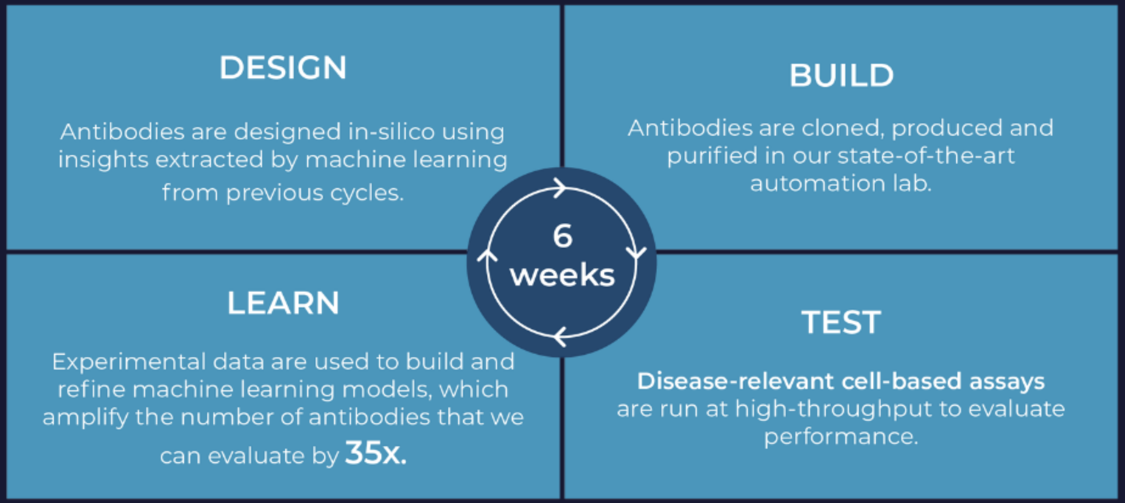
Source: A2A Pharma Corporate Presentation, February 2024

LabGenius Using Iterative QSAR Like Approach to Antibody Design

LabGenius

We intelligently navigate the design space using an iterative process called active learning

LabGenius' unique search capabilities come from the deep integration of **disease-relevant cell-based assays, robotic automation and machine learning**.



DESIGN

Antibodies are designed in-silico using insights extracted by machine learning from previous cycles.

BUILD

Antibodies are cloned, produced and purified in our state-of-the-art automation lab.

6 weeks

LEARN

Experimental data are used to build and refine machine learning models, which amplify the number of antibodies that we can evaluate by **35x.**

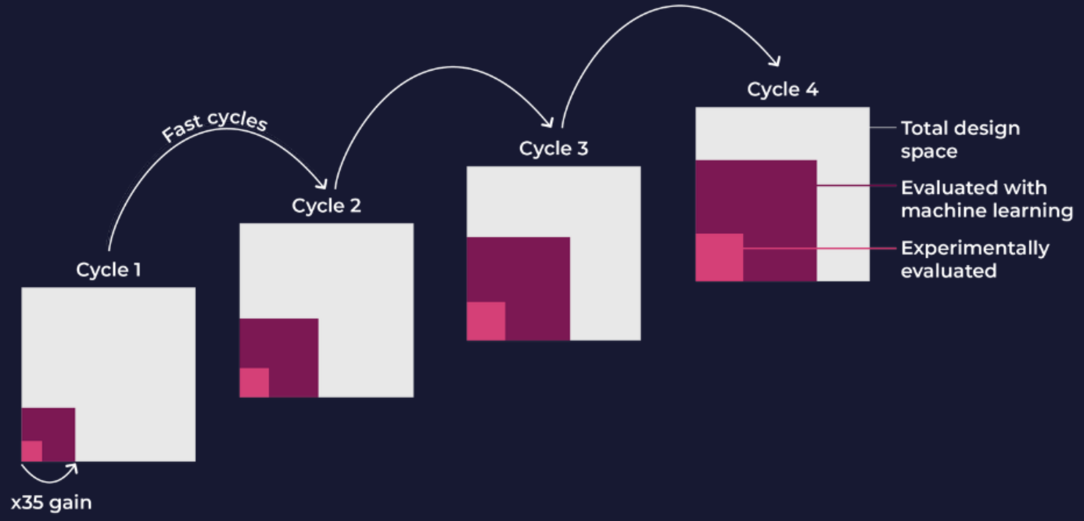
TEST

Disease-relevant cell-based assays are run at high-throughput to evaluate performance.

With this approach, for every design we experimentally evaluate, we're able to **predict the performance of up to 35x more**.

Our approach increases the likelihood of finding high-performing antibodies, faster

Over successive cycles, active learning allows us to intelligently search more of the antibody design space. This means that **we're more likely to identify high-performing antibodies in a shorter space of time versus conventional methods.**

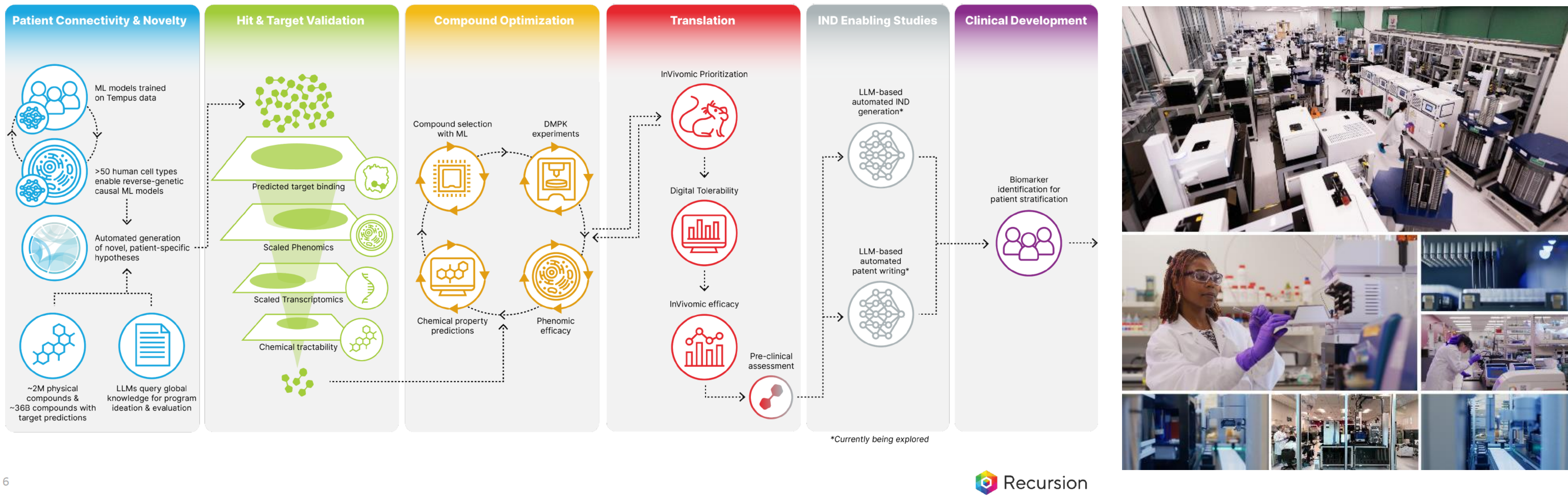


Source: LabGenius Web site

Subsection

Examples of Companies that Take a Cell Perturbation Approach

Recursion Building an End-to-End System for Industrial Drug Discovery



Patient Connectivity & Novelty

ML models trained on Tempus data

>50 human cell types enable reverse-genetic causal ML models

Automated generation of novel, patient-specific hypotheses

~2M physical compounds & ~36B compounds with target predictions

LLMs query global knowledge for program ideation & evaluation

Hit & Target Validation

Predicted target binding

Scaled Phenomics

Scaled Transcriptomics

Chemical tractability

Compound Optimization

Compound selection with ML

DMPK experiments

Chemical property predictions

Phenomic efficacy

Translation

InVivomic Prioritization

Digital Tolerability

InVivomic efficacy

Pre-clinical assessment

IND Enabling Studies

LLM-based automated IND generation*

LLM-based automated patent writing*

Clinical Development

Biomarker identification for patient stratification

*Currently being explored

6

Recursion

Source: Recursion Investor Presentation, April 2024

Recursion Focused on Changes in Cell-Based Phenomena in Biology

AUTOMATION

High-throughput screening

Our highly automated wet-labs systematically capture images of human cells in response to different perturbations

Up to
2.2M experiments
conducted every week

PROFILING SYSTEMS

Diverse biological and chemical inputs

We manipulate human cells with CRISPR/Cas9-mediated gene knockouts, compounds, and other reagents

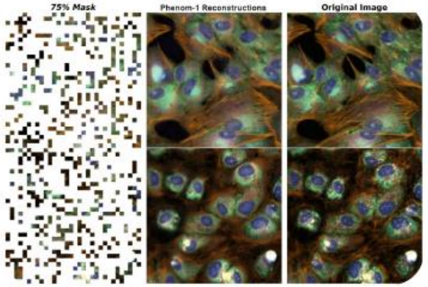
>50 human cell types
~2M physical compounds
Whole-genome CRISPR knockouts

Phenomics

FOUNDATION MODELS

Phenom-1

Groundbreaking models trained on >1 billion images and hundreds of millions of parameters learn to extract biologically meaningful signals from cell images



DIGITIZATION

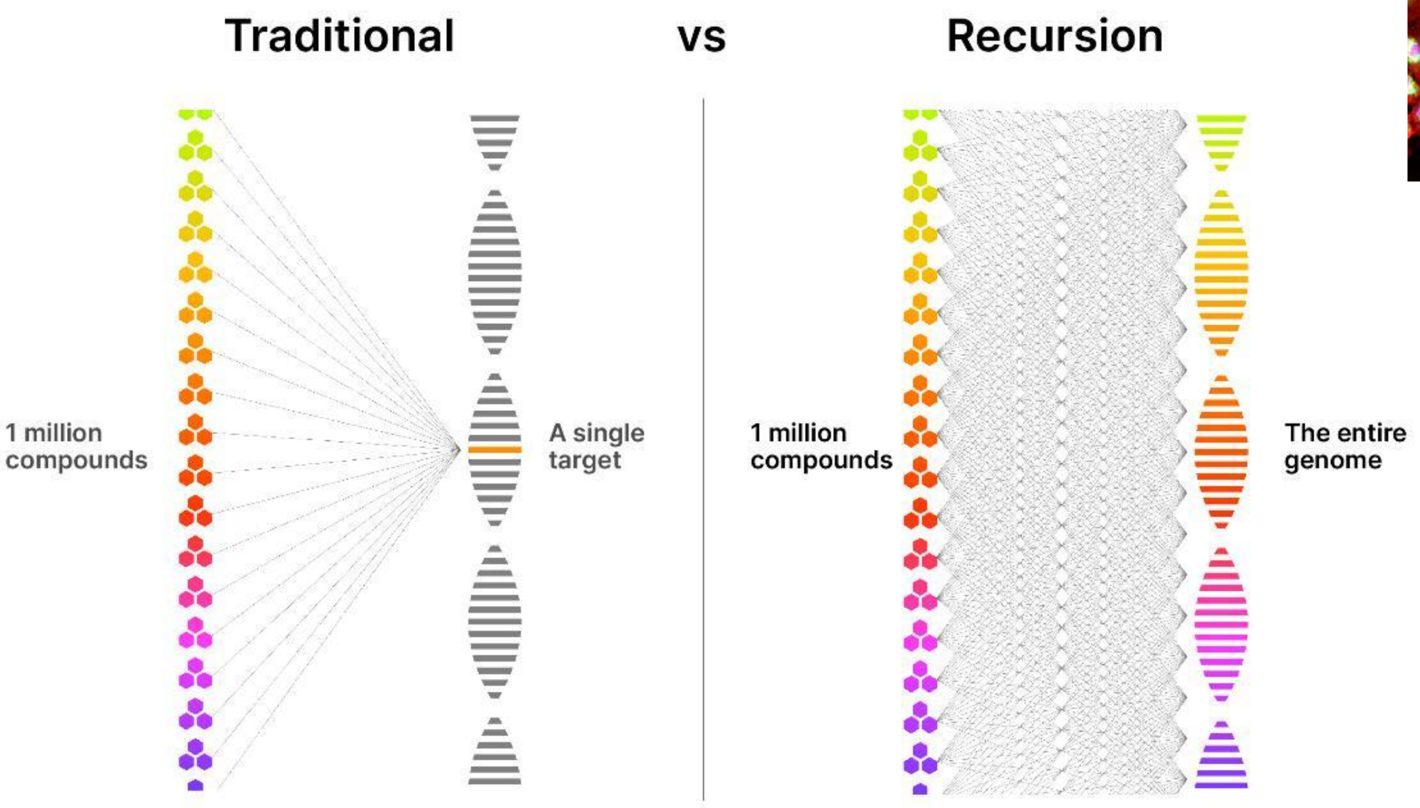
Maps of Biology & Chemistry

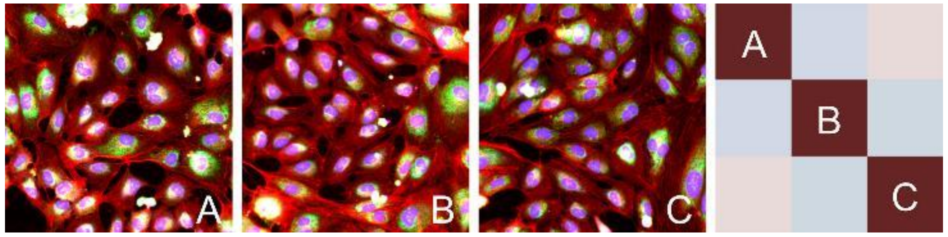
Models infer relationships between all possible combinations of genes and compounds, recapitulating known biology and revealing novel insights

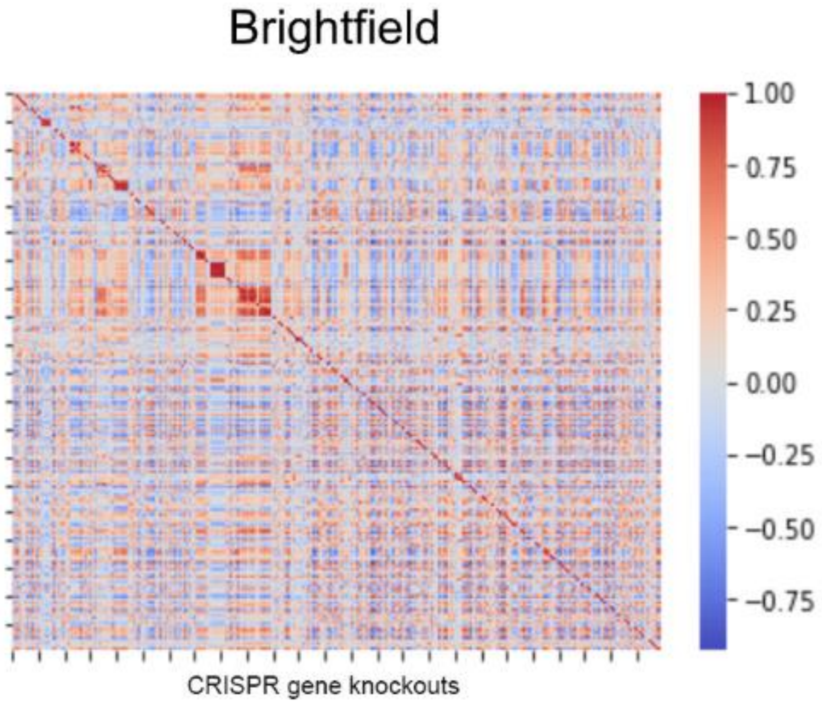
>5 trillion relationships
across multiple biological and chemical contexts

Source: Recursion Investor Presentation, April 2024

Recursion Technology Can Test Millions of Compounds in Cells with Successive CRISPR Gene Knockouts

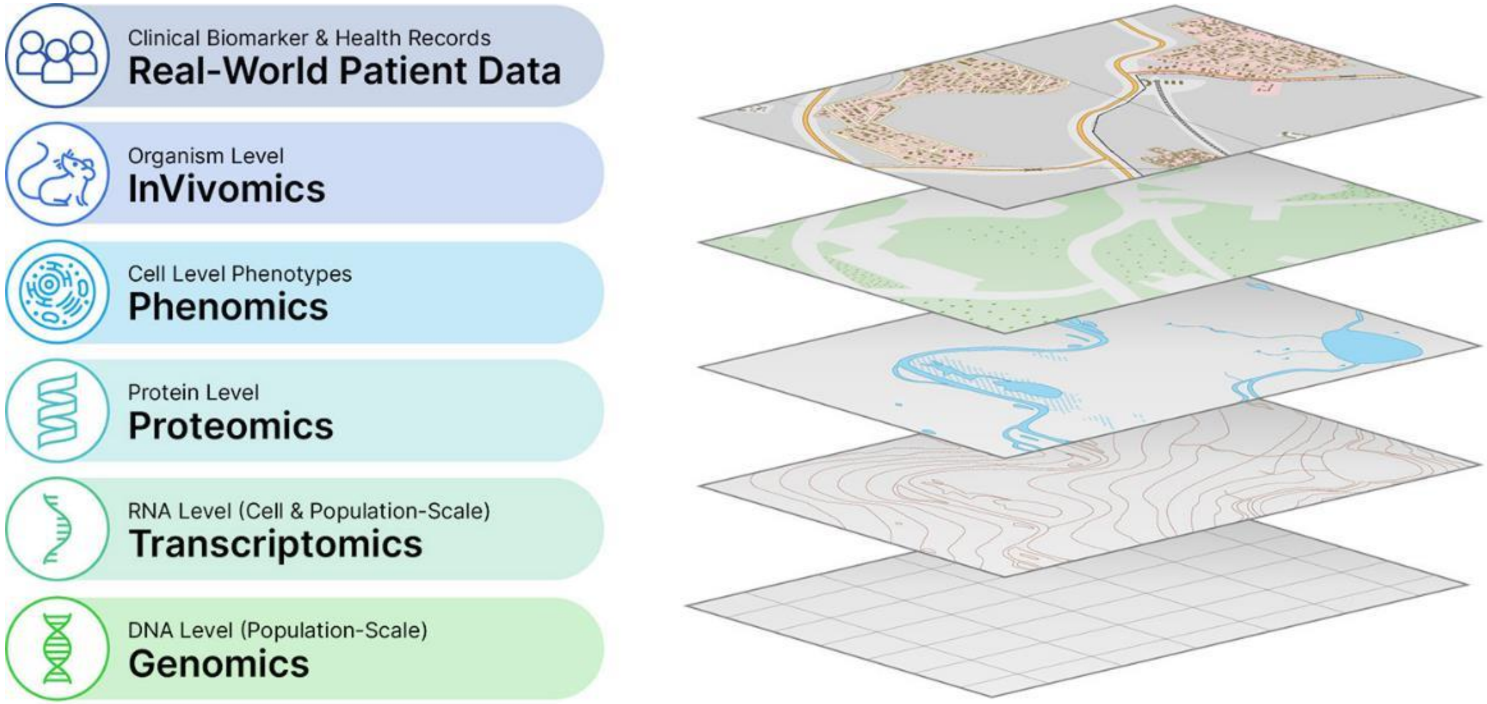






Source: https://www.sec.gov/Archives/edgar/data/1601830/000160183024000019/rxrx-20231231.htm

Recursion Uses Diverse Datasets in its AI Drug Development Programs



Clinical Biomarker & Health Records
Real-World Patient Data

Organism Level
InVivomics

Cell Level Phenotypes
Phenomics

Protein Level
Proteomics

RNA Level (Cell & Population-Scale)
Transcriptomics

DNA Level (Population-Scale)
Genomics

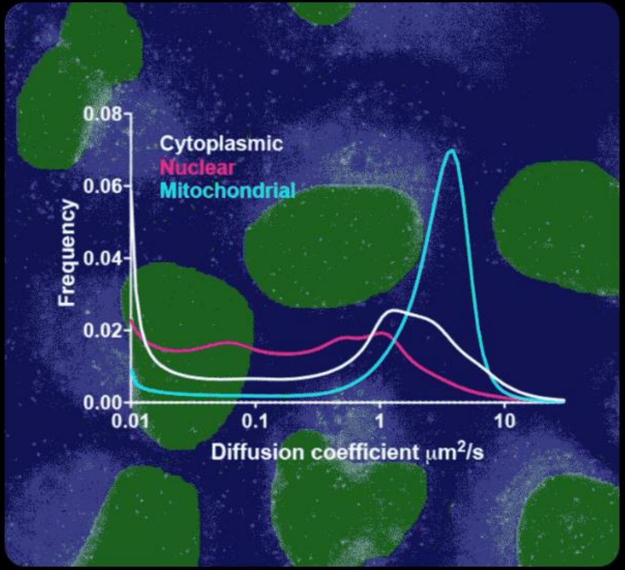
Data utilized by the Recursion OS spans staining and multi-timepoint live-cell phenomics (brightfield), transcriptomics, proteomics, InVivomics, ADME assays, as well as predicted protein-ligand relationships. Recursion also has a physical library of over 1.7 million compounds, including over 1 million new chemical entity (NCE) starting point substances, a large library of known chemical entities which can serve as guideposts, and more than 500,000 compounds belonging to our collaborators. Further, Recursion has generated a custom whole-genome arrayed CRISPR guide library. Together, these tools allow Recursion to explore millions of different biological perturbations in our own wet labs. We have executed over 200 million phenomics and over 700,000 whole transcriptomics experiments across different biological and chemical contexts in multiple human cell types. In 2023, with the completion of our automated DMPK module, we have now conducted tens of thousands of ADME experiments. Our tissue culture facility has scaled the production of over 50 human cell types and has also enabled work at scale in co-cultures and complex iPSC-derived cell types. Since 2022, for example, Recursion generated more than 1 trillion hiPSC-derived neuronal cells for our partnered work with Roche and Genentech - a scale achieved by few other companies in the world.

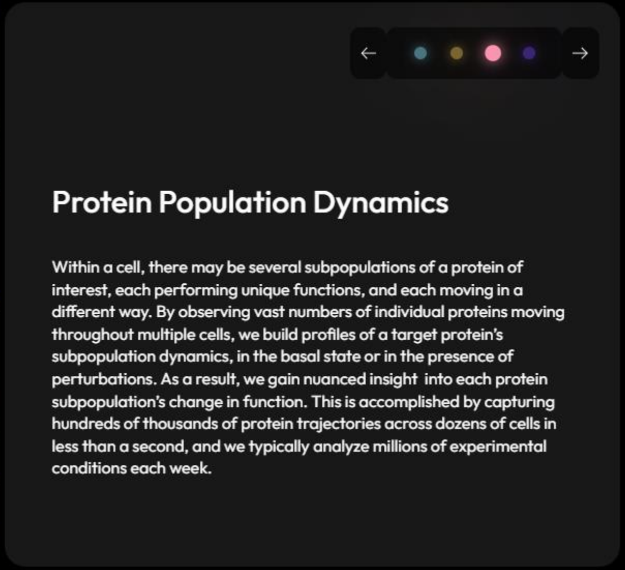
Source: https://www.sec.gov/Archives/edgar/data/1601830/000160183024000019/rxrx-20231231.htm

Eikon Uses Single Molecule Tracking Technology (SMT) To Study the Effect of Drugs in Cells

Super Resolution Microscopy

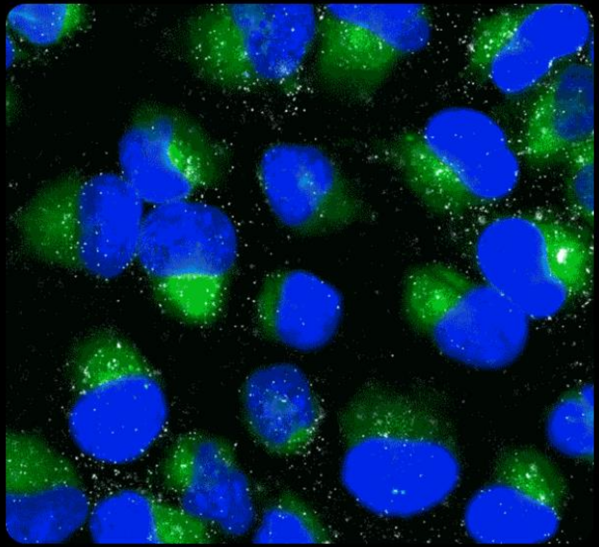
Recognizing the need to visualize the dynamics of macromolecules within live cells, our team is building on Nobel Prize-winning super-resolution fluorescence microscopy tools (developed by Eikon co-founder Eric Betzig). Our Single Molecule Tracking (SMT) systems enable exquisitely precise measurement of the dynamic behavior of individual protein molecules in living cells, on a massive scale.

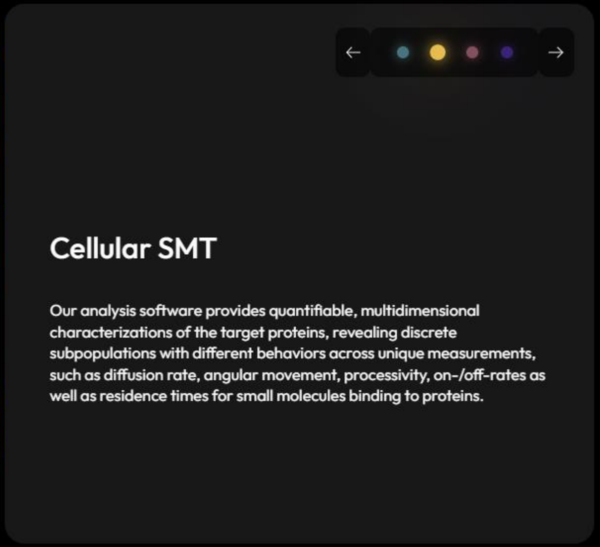




Protein Population Dynamics

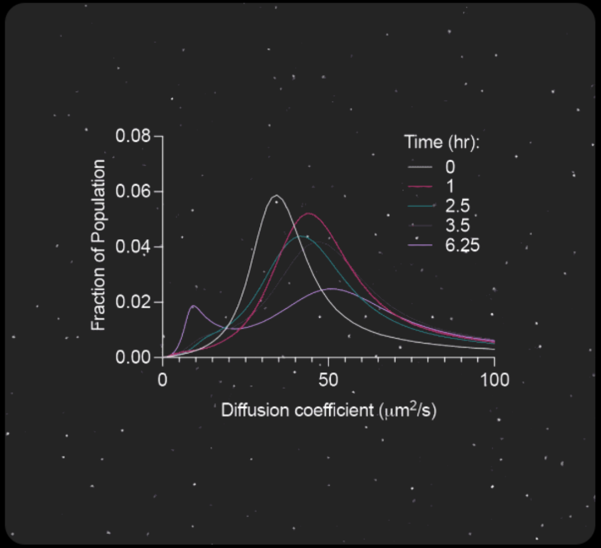
Within a cell, there may be several subpopulations of a protein of interest, each performing unique functions, and each moving in a different way. By observing vast numbers of individual proteins moving throughout multiple cells, we build profiles of a target protein's subpopulation dynamics, in the basal state or in the presence of perturbations. As a result, we gain nuanced insight into each protein subpopulation's change in function. This is accomplished by capturing hundreds of thousands of protein trajectories across dozens of cells in less than a second, and we typically analyze millions of experimental conditions each week.

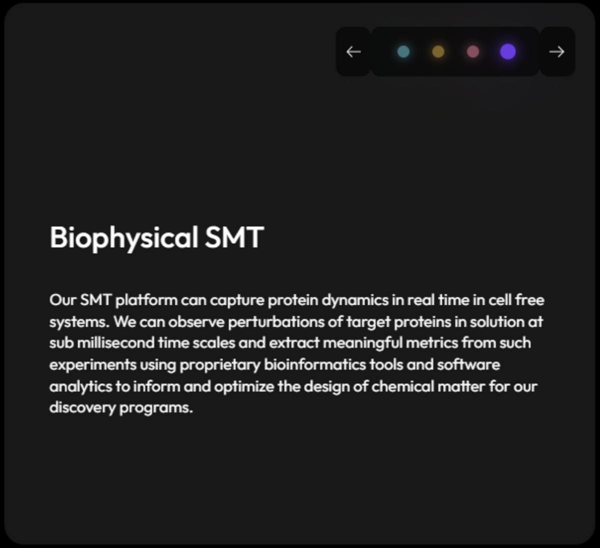




Cellular SMT

Our analysis software provides quantifiable, multidimensional characterizations of the target proteins, revealing discrete subpopulations with different behaviors across unique measurements, such as diffusion rate, angular movement, processivity, on-/off-rates as well as residence times for small molecules binding to proteins.



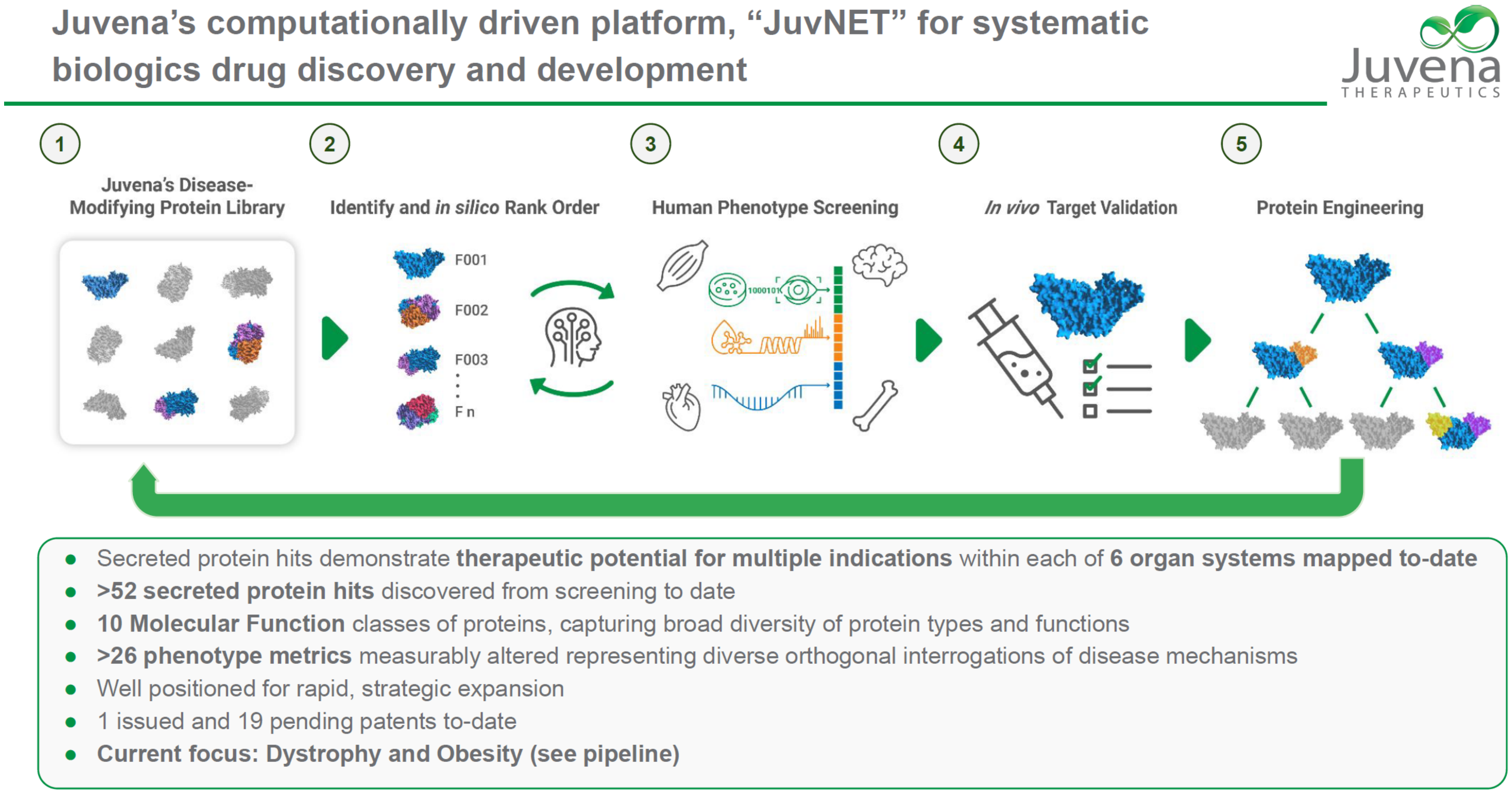


Biophysical SMT

Our SMT platform can capture protein dynamics in real time in cell free systems. We can observe perturbations of target proteins in solution at sub millisecond time scales and extract meaningful metrics from such experiments using proprietary bioinformatics tools and software analytics to inform and optimize the design of chemical matter for our discovery programs.

Source: Eikon Therapeutics Web Site, April 2024

Juvena Focuses on Cell Screening of Secreted Proteins And Has Identified Multiple Productive Hits for Neuromuscular Disease and Obesity



Juvena's computationally driven platform, "JuvNET" for systematic biologics drug discovery and development

Juvena THERAPEUTICS

1 Juvena's Disease-Modifying Protein Library

2 Identify and *in silico* Rank Order

3 Human Phenotype Screening

4 *In vivo* Target Validation

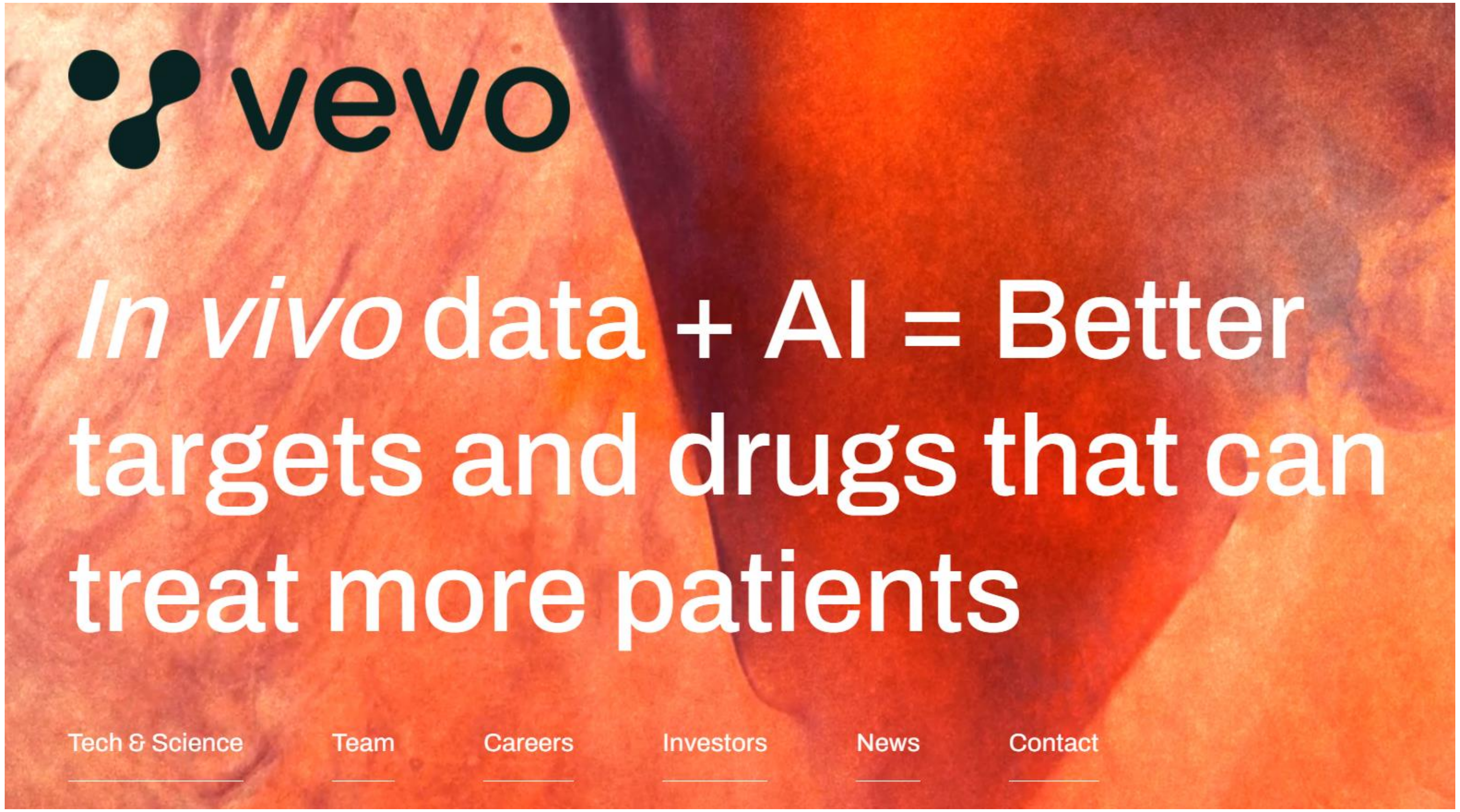
5 Protein Engineering

- Secreted protein hits demonstrate **therapeutic potential for multiple indications** within each of **6 organ systems mapped to-date**
- **>52 secreted protein hits** discovered from screening to date
- **10 Molecular Function** classes of proteins, capturing broad diversity of protein types and functions
- **>26 phenotype metrics** measurably altered representing diverse orthogonal interrogations of disease mechanisms
- Well positioned for rapid, strategic expansion
- 1 issued and 19 pending patents to-date
- **Current focus: Dystrophy and Obesity (see pipeline)**

Juvena's approach is an interesting contrast to the cell perturbation approach taken by others shown here. They hypothesize that the protein secretome is much more likely to have pharmaceutically relevant hits and restrict their phenotypic screening to proteins from this group. Their results to date have been promising.

Source: Juvena Investor Presentation, April 2024

Vevo is Going Beyond Perturbation of a Single Live Cell



In vivo data + AI = Better targets and drugs that can treat more patients

In a single in vivo experiment, Vevo's Mosaic platform can measure how a drug impacts cells from hundreds of patients, generating millions of datapoints on drug-induced changes in gene expression.

Vevo is performing thousands of Mosaic experiments to create what is deemed impossible today: an in vivo atlas of how chemistry perturbs biology. Vevo's AI models will be trained on this atlas to uncover novel targets and drugs undetectable by other technologies.

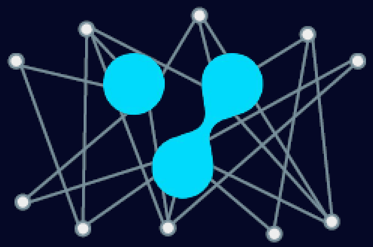
Source: https://www.vevo.ai/

Single-Cell Foundation Models

Learning context-dependency of gene function from 100s of millions of single-cell data, taken from a myriad of disease-relevant biological contexts

The ■ of the old ■ was rough and covered in ■.

■	0	3	■	■	45



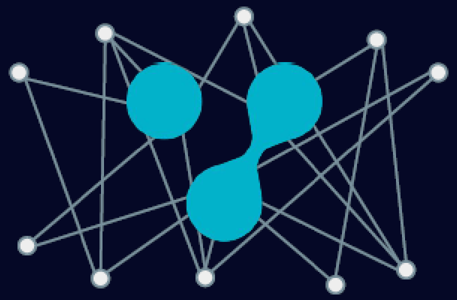
Self-supervised foundation models, trained on single-cell transcriptomic data

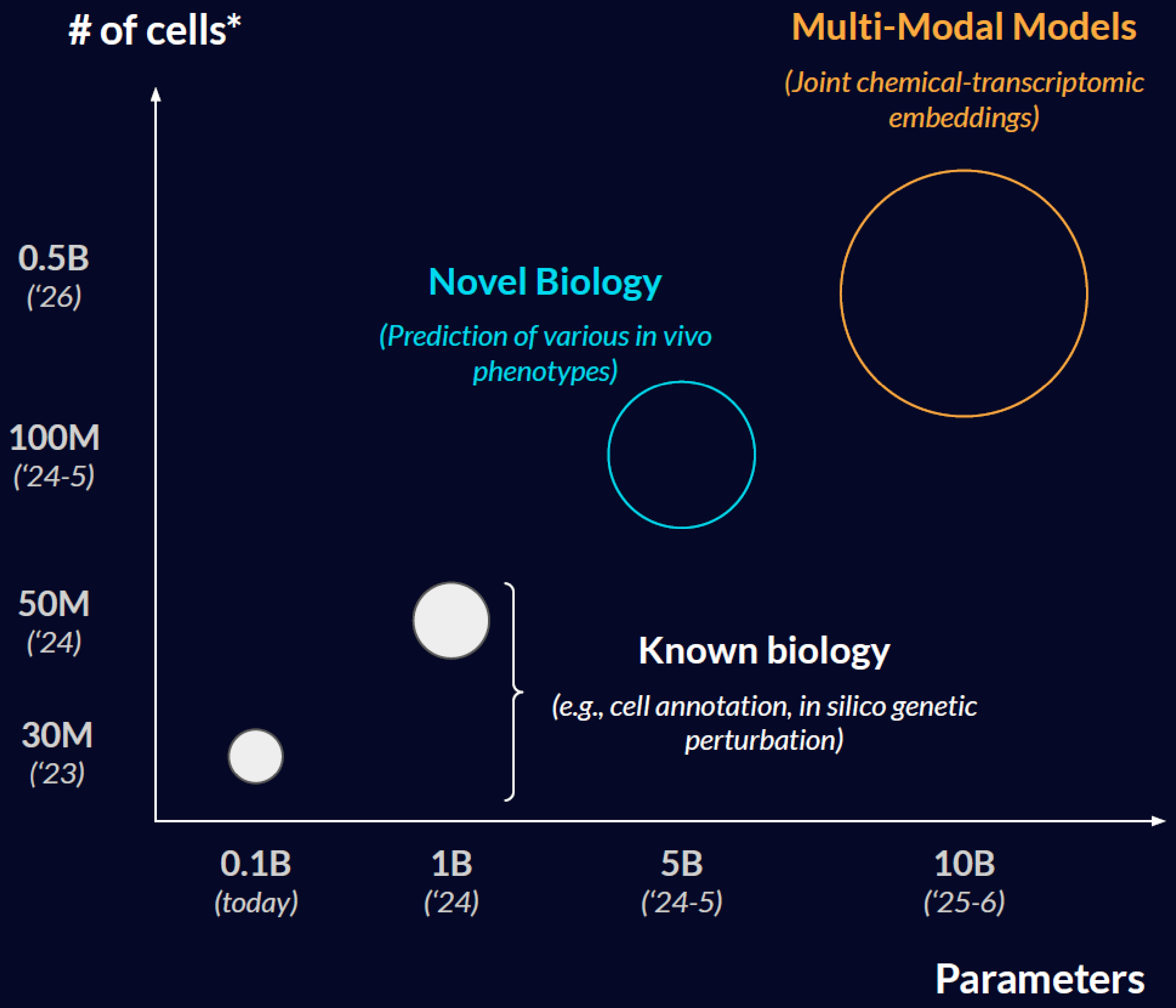
The **bark** of the old **tree** was rough and covered in **moss**.

8	0	3	**7**	**0**	45

Source: Vevo Investor Presentation, April 2024

Scaling our single-cell foundation models as we grow our atlas





# of cells*
0.5B
('26)
100M
('24-5)
50M
('24)
30M
('23)
0.1B
(today)
1B
('24)
5B
('24-5)
10B
('25-6)
Parameters
Known biology
(e.g., cell annotation, in silico genetic perturbation)
Novel Biology
(Prediction of various in vivo phenotypes)
Multi-Modal Models
(Joint chemical-transcriptomic embeddings)

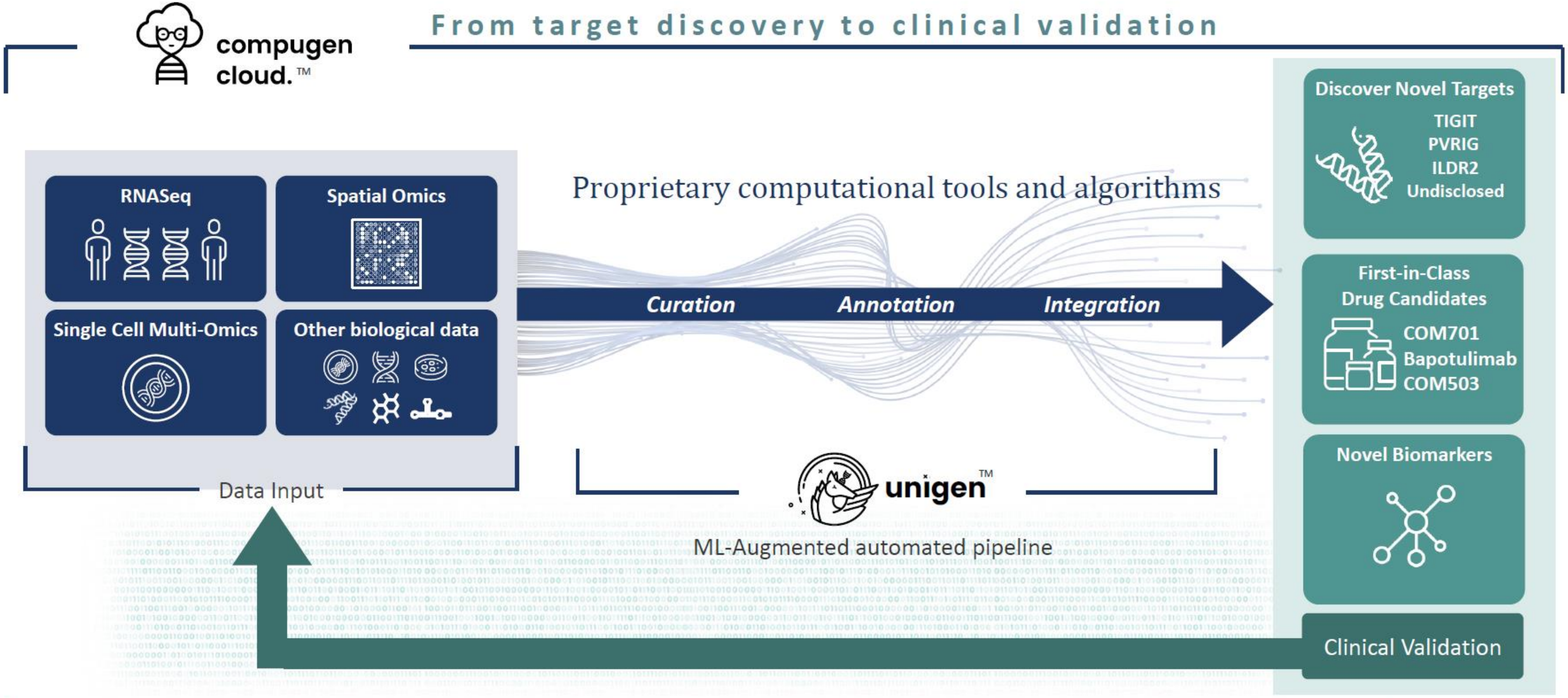
* Each cell corresponds to ~1-2K genes (tokens). 1Bn cells correspond to 1-2Tn tokens

Source: Vevo Investor Presentation, April 2024

Subsection

Examples of Companies that Take a Hybrid Approach

Compugen Uses Multiple Types of Biologic Data for Novel Target ID and Then Uses the Computer to Help Design Appropriate Biologic Drugs



compugen cloud.™

From target discovery to clinical validation

RNASeq

Spatial Omics

Single Cell Multi-Omics

Other biological data

Data Input

Proprietary computational tools and algorithms

Curation

Annotation

Integration

unigen™

ML-Augmented automated pipeline

Discover Novel Targets

TIGIT
PVRIG
ILDR2
Undisclosed

First-in-Class Drug Candidates

COM701
Bapotulimab
COM503

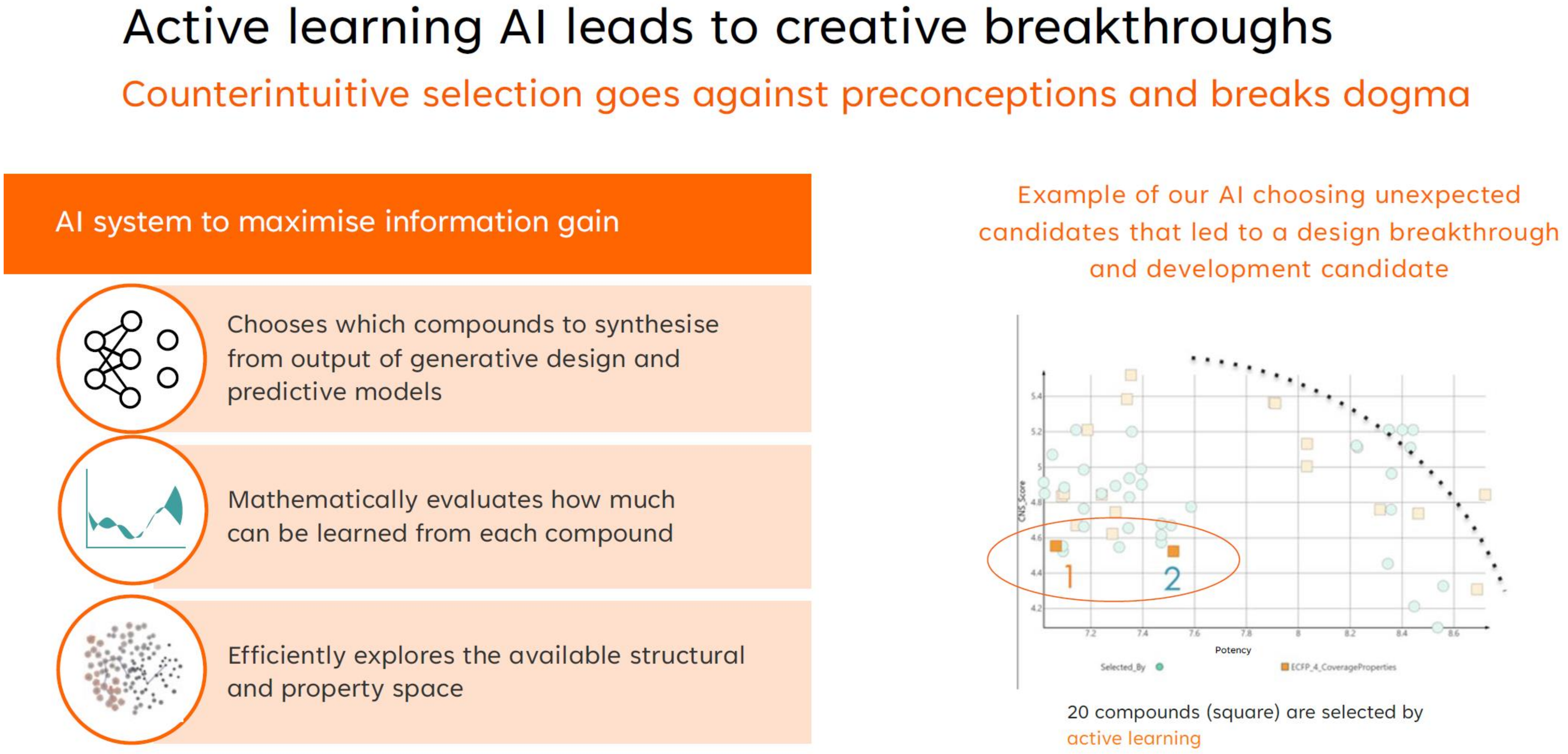
Novel Biomarkers

Clinical Validation

Source: https://ir.cgen.com/Corporate-Presentation/default.aspx

compugen
FROM CODE TO CURE®

Exscientia Uses Machine Learning For Dimensionality Reduction



Active learning AI leads to creative breakthroughs

Counterintuitive selection goes against preconceptions and breaks dogma

AI system to maximise information gain

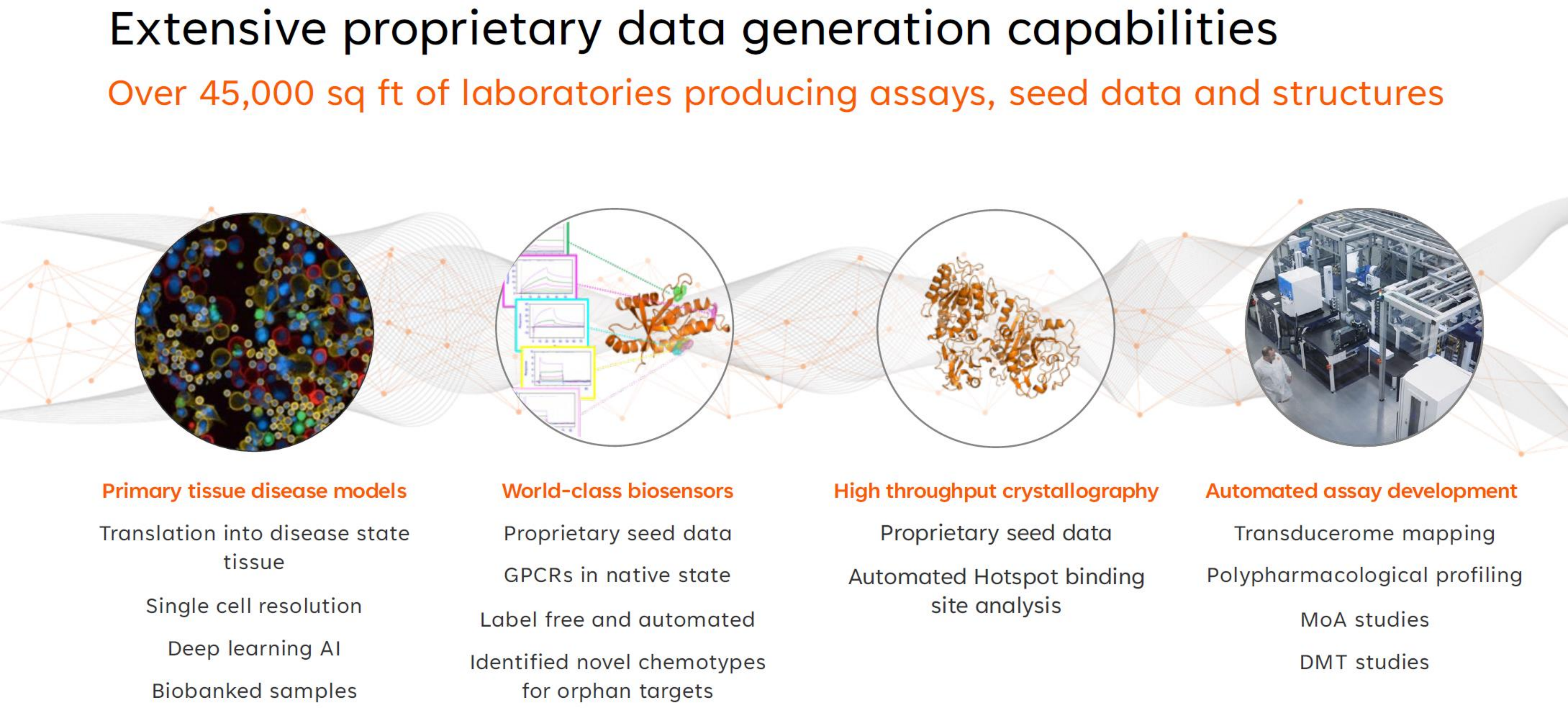
- Chooses which compounds to synthesise from output of generative design and predictive models
- Mathematically evaluates how much can be learned from each compound
- Efficiently explores the available structural and property space

Example of our AI choosing unexpected candidates that led to a design breakthrough and development candidate

20 compounds (square) are selected by active learning

Source: Exscientia Investor Presentation, April 2024

Exscientia Uses Machine Learning For Dimensionality Reduction



Extensive proprietary data generation capabilities

Over 45,000 sq ft of laboratories producing assays, seed data and structures

Primary tissue disease models

Translation into disease state tissue

Single cell resolution

Deep learning AI

Biobanked samples

World-class biosensors

Proprietary seed data

GPCRs in native state

Label free and automated

Identified novel chemotypes for orphan targets

High throughput crystallography

Proprietary seed data

Automated Hotspot binding site analysis

Automated assay development

Transducerome mapping

Polypharmacological profiling

MoA studies

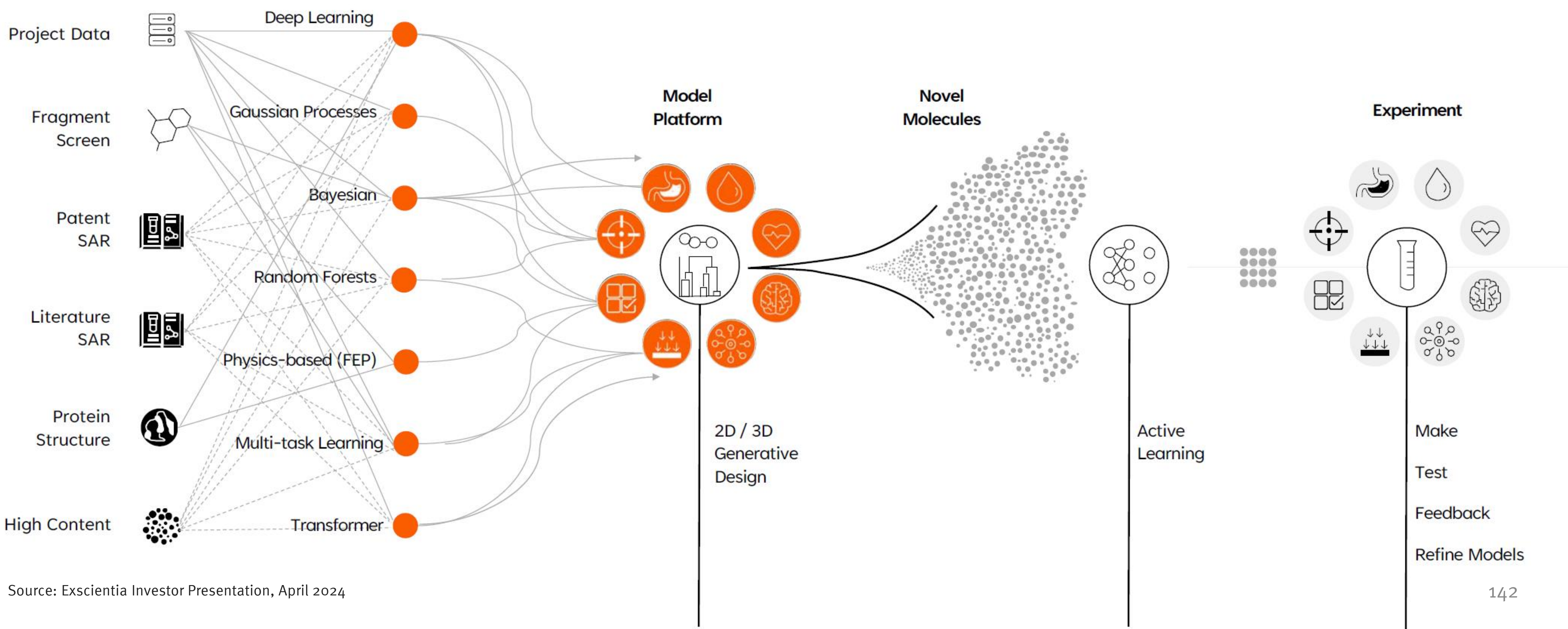
DMT studies

Source: Exscientia Investor Presentation, April 2024

Exscientia Predicts Molecule Success Using Multiple Inputs

Data and model agnostic

Our AI design platform can optimise complex drugs from diverse starting data



Source: Exscientia Investor Presentation, April 2024

Exscientia Automating Chemical Synthesis and Testing

Our new automation facility

Automation of laboratory processes

4,500 sq ft automation studio

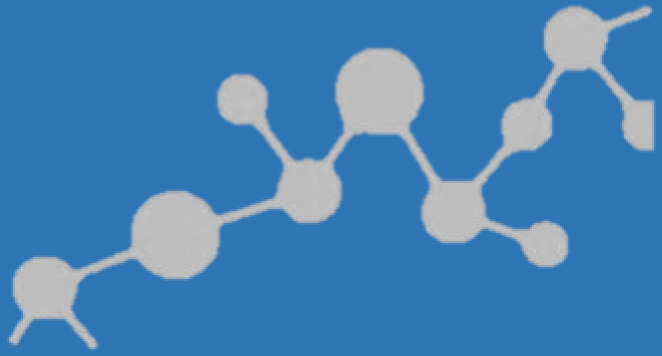
Encode & automate laboratory workflows

Building modular & scalable platforms

Chemical Synthesis ❯ Compound Management ❯ Biological Testing

Source: Exscientia Investor Presentation, April 2024

Subsection



AI and Biologics: Examples of Companies that Solve Find Binders to Proteins and Solve for Protein Structures

Alphabet's Isomorphic Working on the Next Generation of AlphaFold

Isomorphic Labs October 2023

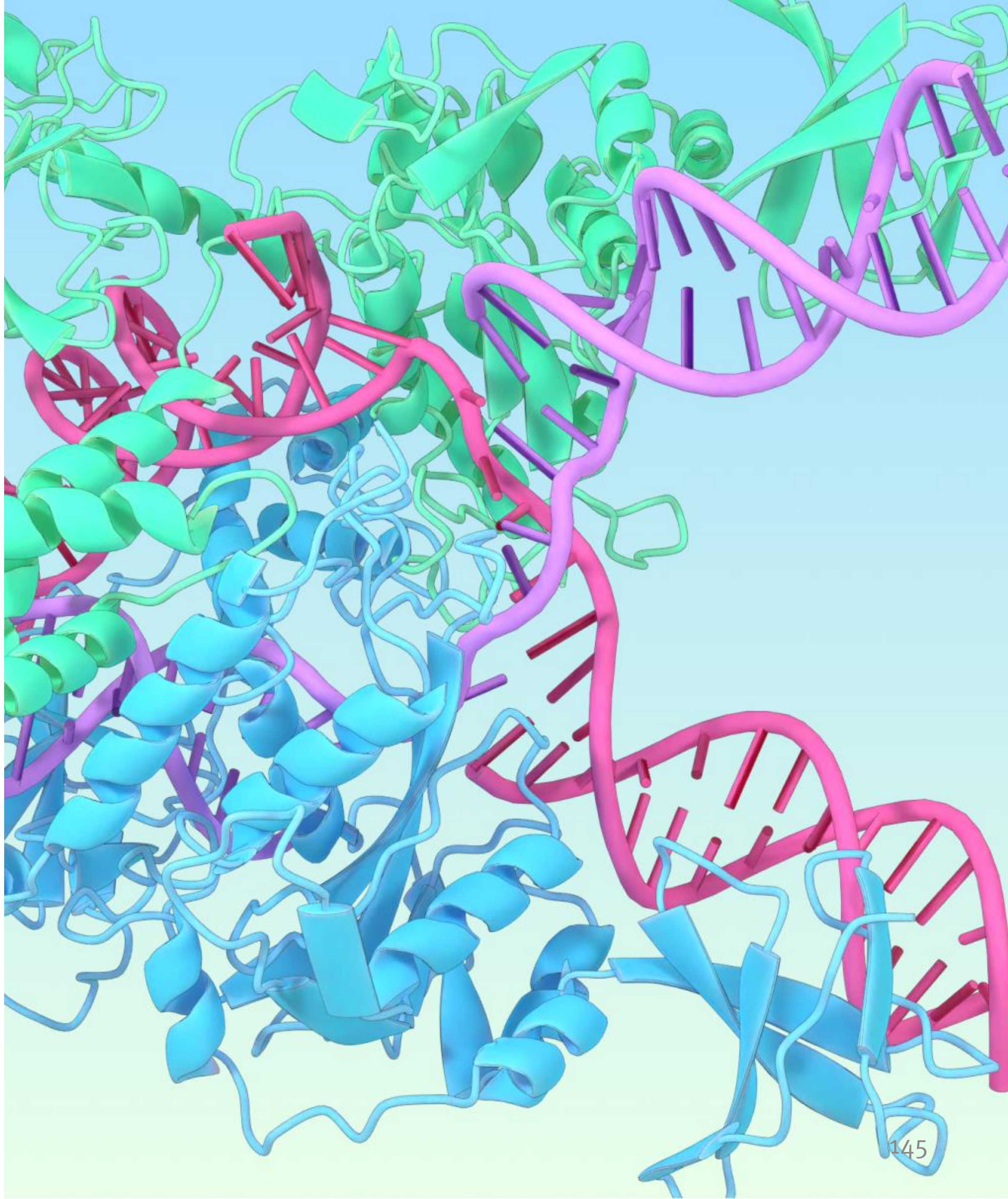
Since its release in 2020, AlphaFold has revolutionised how proteins and their interactions are understood. Isomorphic Labs and Google DeepMind have been working together to build the foundations of a more powerful AI model that expands coverage beyond just proteins to the full range of biologically-relevant molecules.

Today we're sharing an update on progress towards the next generation of AlphaFold. Our latest model can now generate predictions for nearly all molecules in the Protein Data Bank (PDB), frequently reaching atomic accuracy.

It unlocks new understanding and significantly improves accuracy in multiple key biomolecule classes, including ligands (small molecules), proteins, nucleic acids (DNA and RNA), and those containing post-translational modifications (PTMs). These different structure types and complexes are essential for understanding the biological mechanisms within the cell, and have been challenging to predict with high accuracy.

The model's expanded capabilities and performance can help accelerate biomedical breakthroughs and realise the next era of 'digital biology' - giving new insights into the functioning of disease pathways, genomics, biorenewable materials, plant immunity, potential therapeutic targets, mechanisms for drug design, and new platforms for enabling protein engineering and synthetic biology.

Source: https://www.isomorphiclabs.com/articles/a-glimpse-of-the-next-generation-of-alphafold



Generate BioMedicines Uses Chroma Model to Develop Proteins

Ingraham, J.B., Baranov, M., Costello, Z. et al. Illuminating protein space with a programmable generative model. *Nature* 623, 1070–1078 (2023).

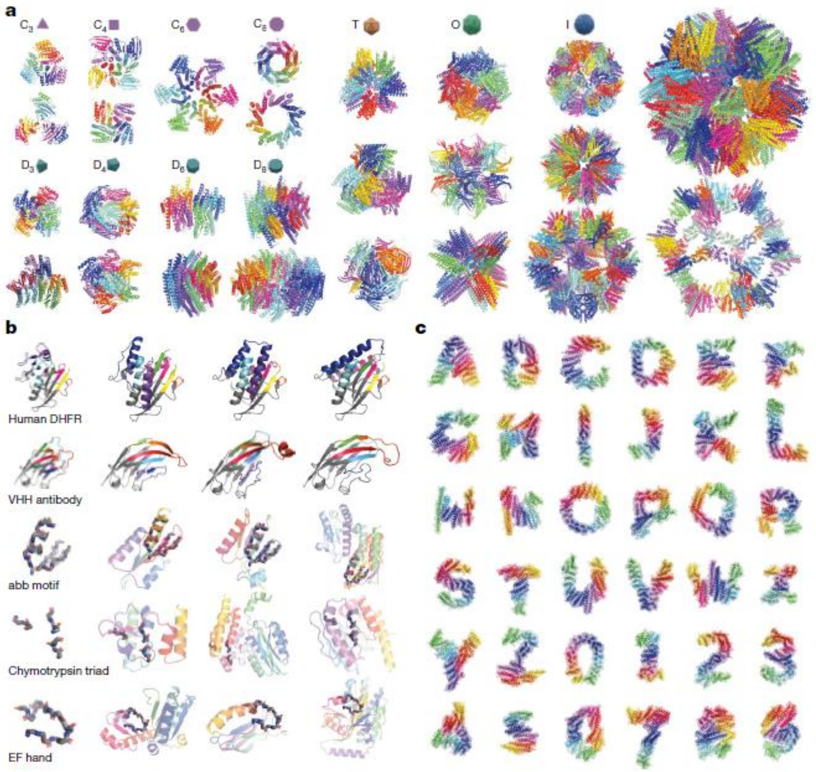
Three billion years of evolution has produced a tremendous diversity of protein molecules, but the full potential of proteins is likely to be much greater.

Accessing this potential has been challenging for both computation and experiments because the space of possible protein molecules is much larger than the space of those likely to have functions.

Here we introduce Chroma, a generative model for proteins and protein complexes that can directly sample novel protein structures and sequences, and that can be conditioned to steer the generative process towards desired properties and functions.

To enable this, we introduce a diffusion process that respects the conformational statistics of polymer ensembles, an efficient neural architecture for molecular systems that enables long-range reasoning with sub-quadratic scaling, layers for efficiently synthesizing three-dimensional structures of proteins from predicted inter-residue geometries and a general low-temperature sampling algorithm for diffusion models. Chroma achieves protein design as Bayesian inference under external constraints, which can involve symmetries, substructure, shape, semantics and even natural-language prompts. The experimental characterization of 310 proteins shows that sampling from Chroma results in proteins that are highly expressed, fold and have favourable biophysical properties. The crystal structures of two designed proteins exhibit atomistic agreement with Chroma samples (a backbone root-mean-square deviation of around 1.0 Å). With this unified approach to protein design, we hope to accelerate the programming of protein matter to benefit human health, materials science and synthetic biology.

Source: https://www.nature.com/articles/s41586-023-06728-8

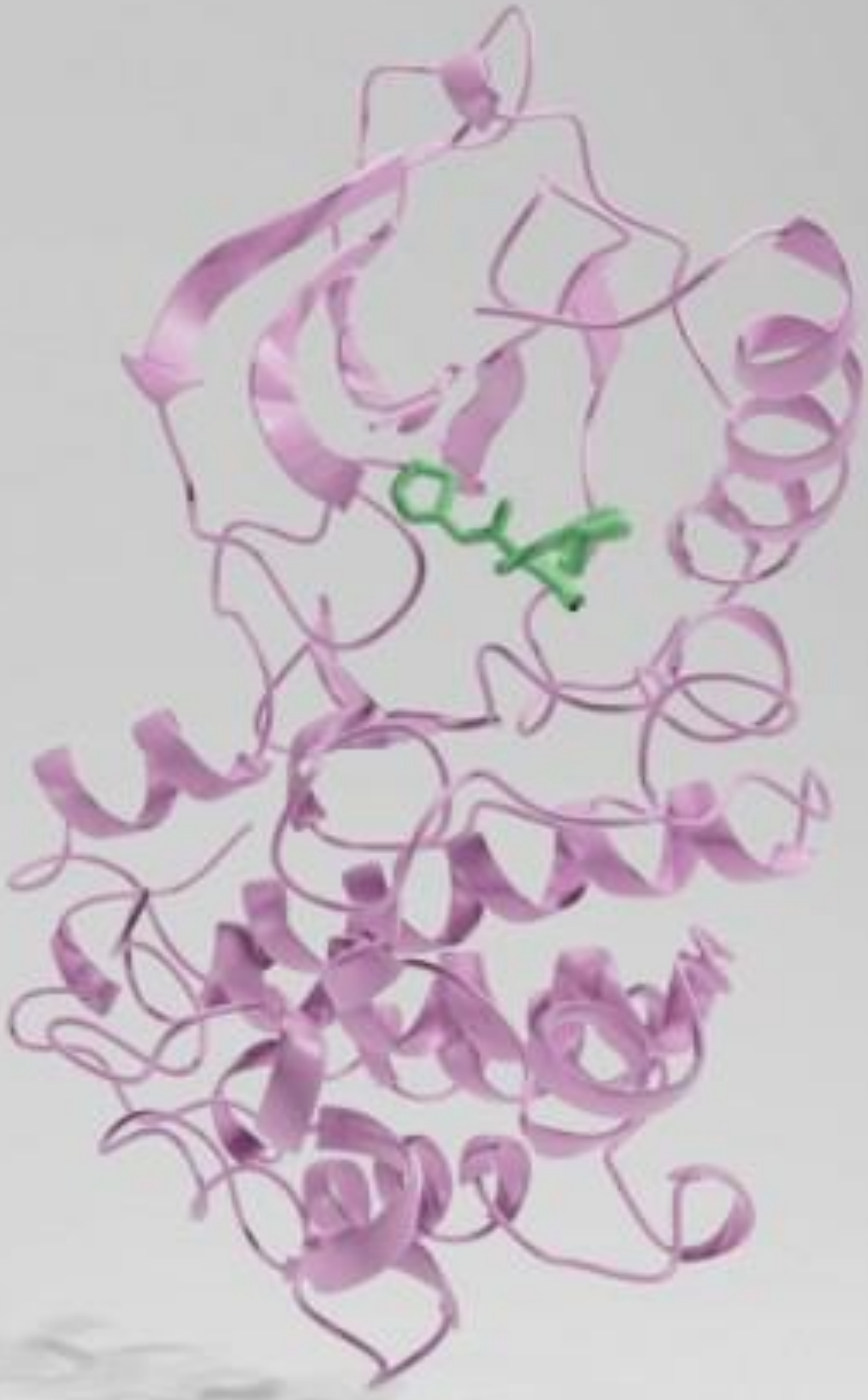


a, Sampling oligomeric structures with arbitrary chain symmetries is possible by using a conditioner that tessellates an asymmetric subunit in the energy function. Cyclic (Cn), dihedral (Dn), tetrahedral (T), octahedral (O) and icosahedral (I) symmetry groups can produce a wide variety of possible homomeric complexes. The right-most protein complex contains 60 subunits and 60,000 total residues, which is enabled by leveraging symmetries and using our subquadratically scaling architecture. b, Conditioning on partial substructure (monochrome) enables protein infilling or outfilling. The top two rows illustrate regeneration (colour) of half a protein (the enzyme DHFR, first row) or complementarity-determining region loops of a VHH antibody (second row). The next three rows show conditioning on a predefined motif. The order and matching location of motif segments is not prespecified here. c, Conditioning on arbitrary volumetric shapes is exemplified by the complex geometries of the Latin alphabet and Arabic numerals. All structures were selected from protocols with high rates of in silico refolding

CHARM THERAPEUTICS

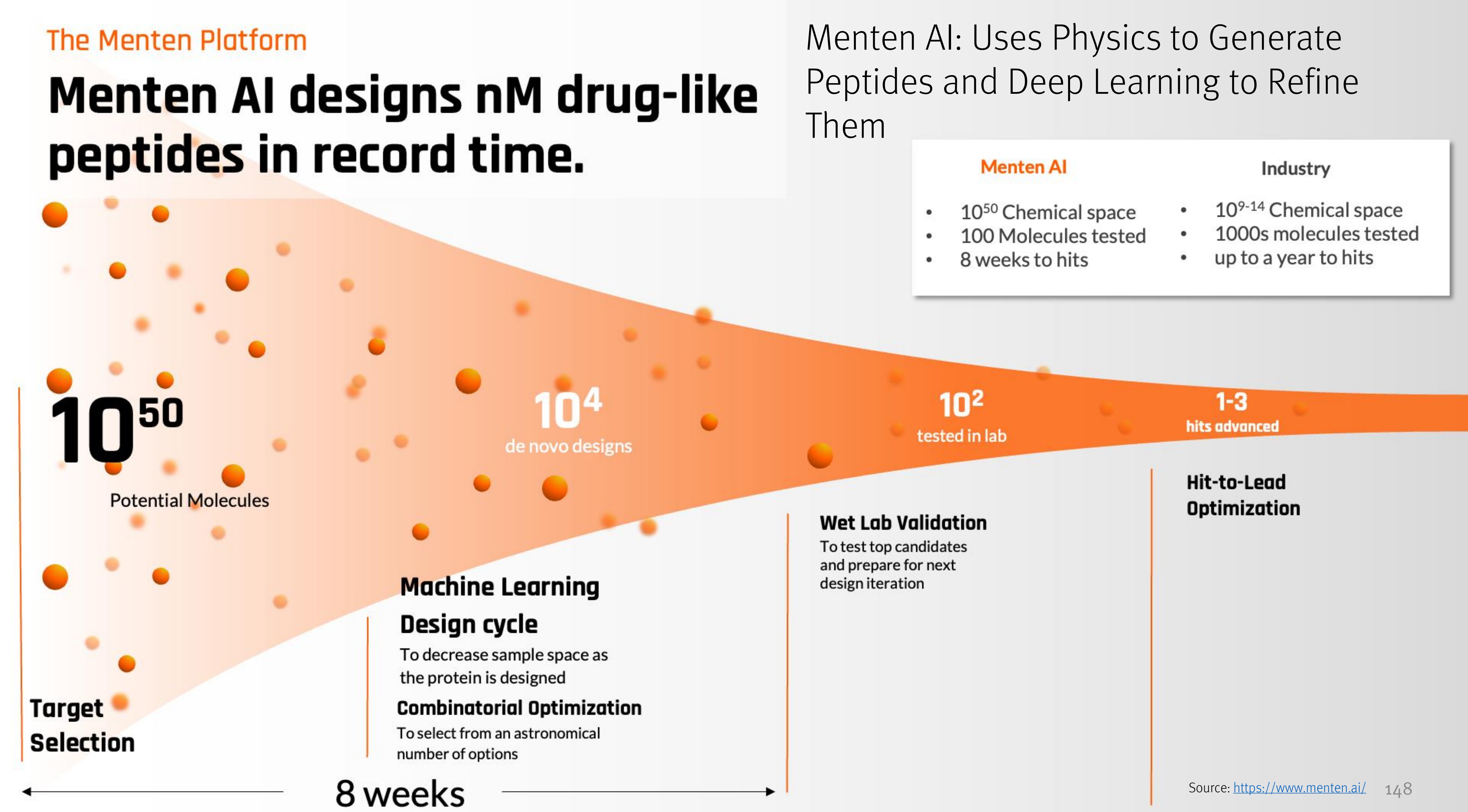
DragonFold

Charm Therapeutics applies proprietary expertise to the challenge of protein ligand co-folding to discover and develop a new wave of novel small molecule medicines in cancer and other indications. Its DragonFold technology, inspired by the breakthroughs of our cofounder and Wiley Prize winner David Baker, is the first rapid, accurate protein/ligand co-folding algorithm. The company's approach is to identify small molecules algorithmically through its co-folding algorithm.



Source: https://charmtx.com/about/

Menten AI: Uses Physics to Generate Peptides and Deep Learning to Refine Them



The Menten Platform

Menten AI designs nM drug-like peptides in record time.

Menten AI	Industry
• 10^{50} Chemical space	• 10^{9-14} Chemical space
• 100 Molecules tested	• 1000s molecules tested
• 8 weeks to hits	• up to a year to hits

10^{50}
Potential Molecules

Target Selection

10^{4}
de novo designs

Machine Learning Design cycle
To decrease sample space as the protein is designed

Combinatorial Optimization
To select from an astronomical number of options

10^{2}
tested in lab

Wet Lab Validation
To test top candidates and prepare for next design iteration

1-3
hits advanced

Hit-to-Lead Optimization

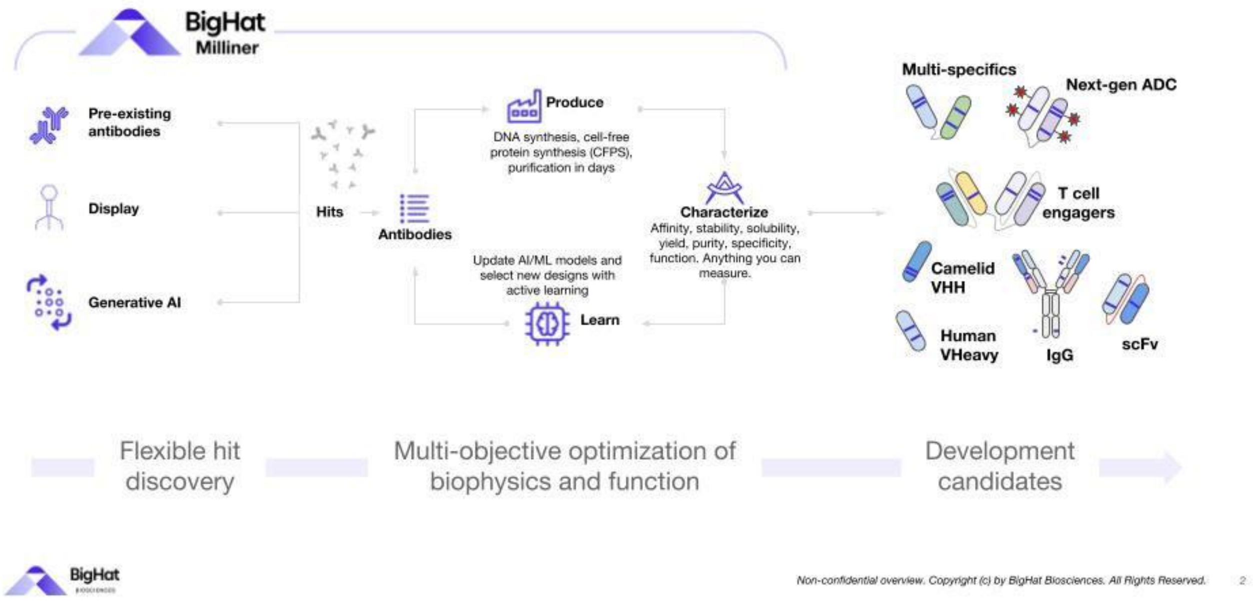
Source: https://www.menten.ai/

BigHat Biosciences Has Modern AI Biologics Generation Platform

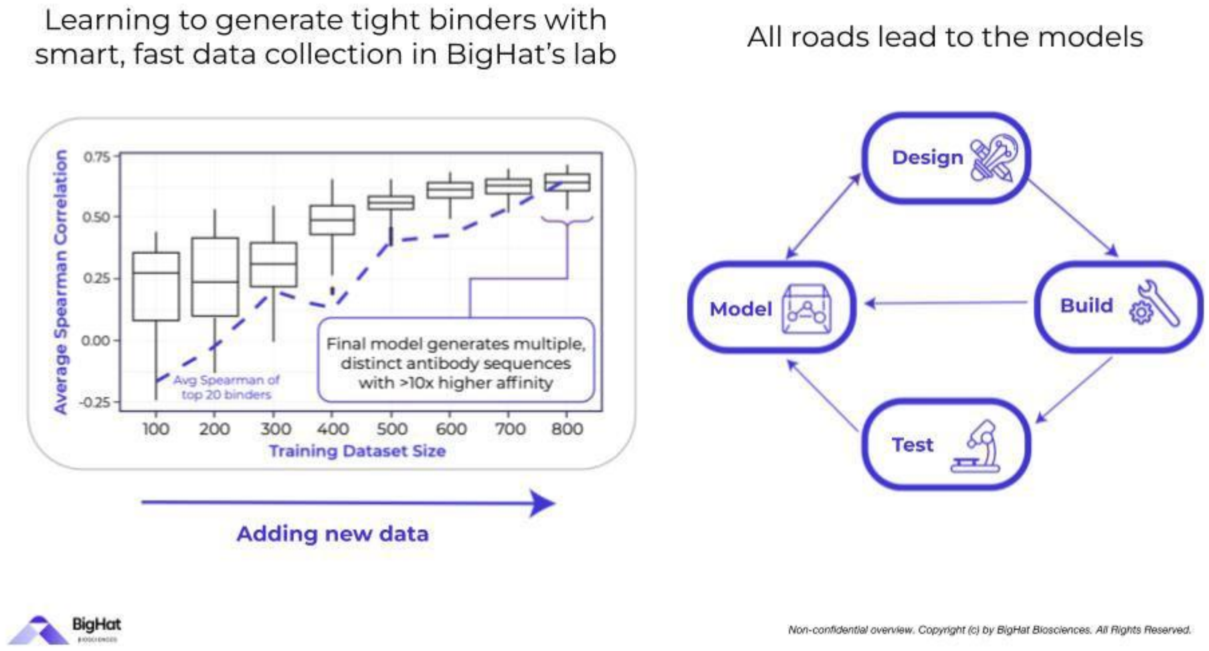
Barry Davidson, *Genetic Engineering and Biotechnology News*, October 25, 2023 (excerpt)

Several recent scientific and technical innovations in synthetic biology and ML enabled the duo to build BigHat's Milliner platform. "It was the synthetic biology advances, in the early 2010s, that allowed us to build the rapid iteration lab that we needed to make the ML methods work," said Greenside, who serves as BigHat's chief scientific officer. "We use synthetic biology techniques throughout our platform, from DNA synthesis to cell-free protein synthesis."

Milliner is a highly automated, closed-loop, integrated ML high-speed wet lab dedicated to antibody discovery and development. "Like classic tech platforms, BigHat's platform is reusable and scalable, but to a greater extent," DePristo said. "It's why we can pursue so many therapeutic programs in parallel in a lean organization."


BigHat
Milliner
Pre-existing antibodies
Display
Generative AI
Hits
Antibodies
Produce
DNA synthesis, cell-free protein synthesis (CFPS), purification in days
Characterize
Affinity, stability, solubility, yield, purity, specificity, function. Anything you can measure.
Update AI/ML models and select new designs with active learning
Learn
Multi-specifics
Next-gen ADC
T cell engagers
Camelid VHH
Human VHeavy
IgG
scFv
Flexible hit discovery
Multi-objective optimization of biophysics and function
Development candidates
BigHat

Non-confidential overview. Copyright (c) by BigHat Biosciences. All Rights Reserved. 2


Learning to generate tight binders with smart, fast data collection in BigHat's lab

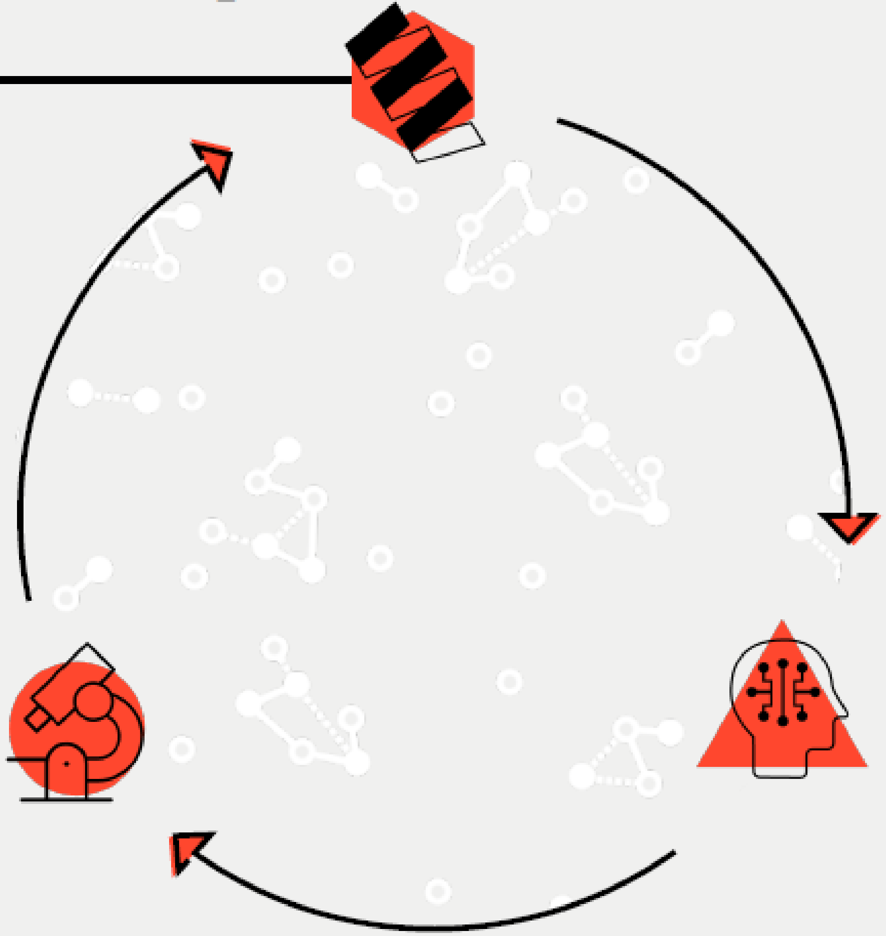
All roads lead to the models

Non-confidential overview. Copyright (c) by BigHat Biosciences. All Rights Reserved.

Source: https://www.genengnews.com/topics/artificial-intelligence/ai-created-antibodies-drive-innovation-at-bighat-biosciences/

The Solution

At Absci, the future is now with our Integrated Drug Creation™ platform



DATA TO TRAIN

Proprietary wet-lab assays generate massive quantities of high-quality data for generative AI model training

AI TO CREATE

Advanced generative AI models used to “create” antibodies and next-gen biologics

WET LAB TO VALIDATE

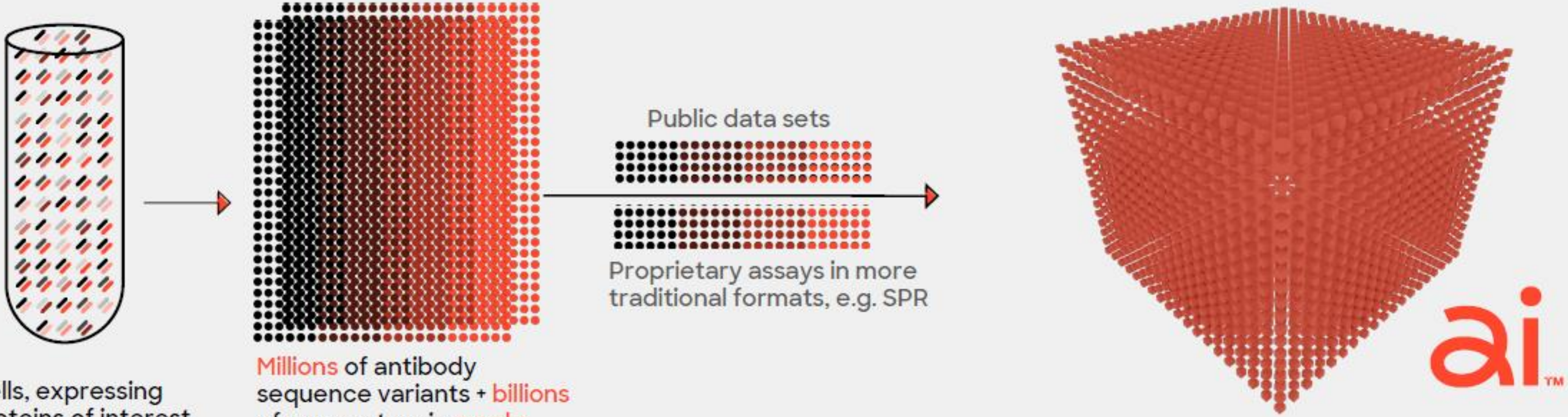
Scalable wet-lab infrastructure capable of validating millions of unique AI-generated designs a week

Source: https://investors.absci.com/news-and-events/events-and-presentations

Absci’s scalable biological data enables true generative AI for biologics drug discovery

Absci’s ACE Assay™ generates data at >4,000x the throughput of traditional HT assays

Massive Training Data Sets



Absci’s ACE Assay™

Cells, expressing proteins of interest

Millions of antibody sequence variants + billions of parameters in weeks

Public data sets

Proprietary assays in more traditional formats, e.g. SPR

Source: https://investors.absci.com/news-and-events/events-and-presentations

THE POWER OF GENERATIVE AI

By *creating* the antibodies with generative AI, we can *design* candidates with desired attributes

Instead of the long, iterative process of sequentially optimizing parameters one by one, our platform is engineered to design an antibody with all of the desired attributes from the start.

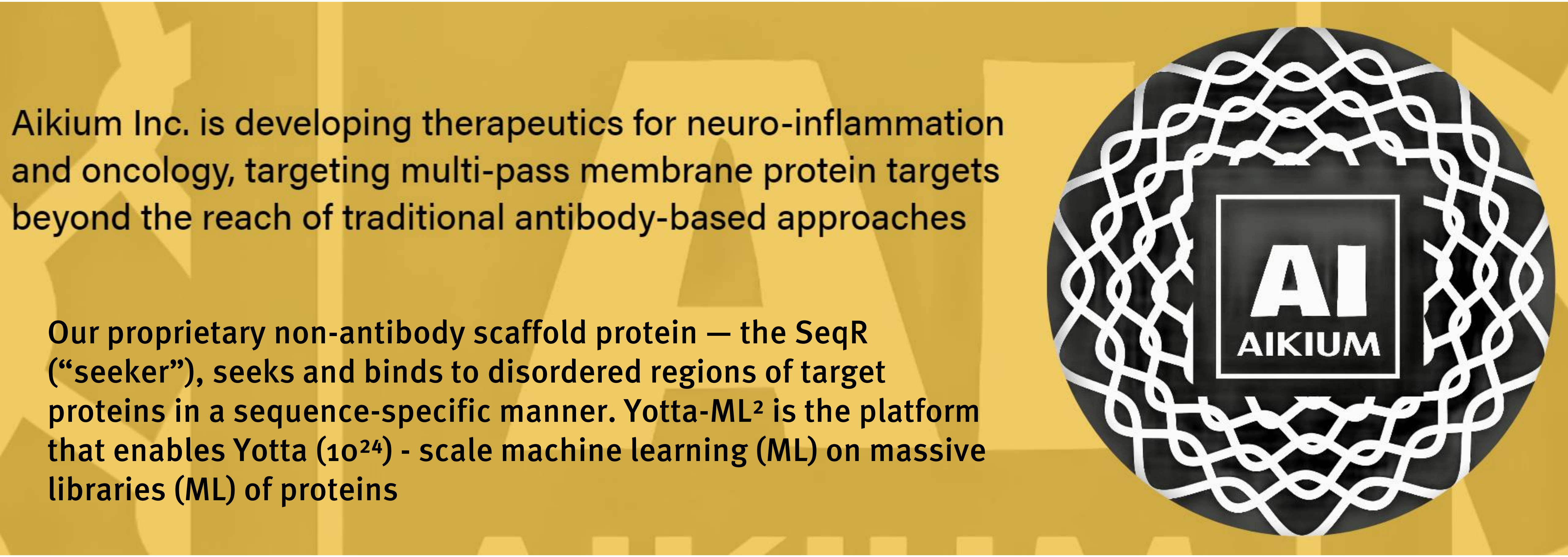
- This workflow would potentially reduce the time to clinic, lower the cost of discovery work, and lead to a higher ultimate probability of success.

This multiparametric optimization allows us to design for:

- Target - antibodies bind to specific foreign substances in the body, such as proteins on surface of bacteria, viruses, or cancer cells to help protect against infection and fight disease
- Affinity - the strength of the bond between an antibody and its target
- Epitope - the region on an antigen (e.g., virus, bacteria, cancer cell) recognized and bound by an antibody
- Developability - the ease with which an antibody can be developed into a drug for use in humans or other animals
- Manufacturability - the ease with which an antibody can be produced in large quantities
- Immunogenicity (inverse) - the ability of an antibody to trigger an immune response in the body

Source: https://investors.absci.com/news-and-events/events-and-presentations

Aikium Using a Novel LLM to Build Drugs Against Very Tough Protein Targets



Aikium Inc. is developing therapeutics for neuro-inflammation and oncology, targeting multi-pass membrane protein targets beyond the reach of traditional antibody-based approaches

Our proprietary non-antibody scaffold protein — the SeqR (“seeker”), seeks and binds to disordered regions of target proteins in a sequence-specific manner. Yotta-ML^2 is the platform that enables Yotta (10^{24}) - scale machine learning (ML) on massive libraries (ML) of proteins

Source: https://www.aikium.com/

The Aikium LLM is the First Protein Builder that uses DPO Rather than PPO Reinforcement Learning

Pouria Mistani and Venkatesh Mysore, “Preference optimization of protein language models as a multi-objective binder design paradigm,” *arXiv*, Mar 7, 2024 (extract)

We present a multi-objective binder design paradigm based on instruction fine-tuning and direct preference optimization (DPO) of autoregressive protein language models (pLMs). Multiple design objectives are encoded in the language model through direct optimization on expert curated preference sequence datasets comprising preferred and dispreferred distributions. We show the proposed alignment strategy enables ProtGPT2 to effectively design binders conditioned on specified receptors and a drug developability criterion. Generated binder samples demonstrate median isoelectric point (pI) improvements by 17% to 60%.

Drug discovery and development is a multi-objective optimization process. Beyond binding affinity, numerous other factors need to be considered for therapeutic development such as expressibility, synthesizability, stability, immunogenicity, solubility and bioavailability. Our goal is to develop a framework for binder design beyond binding affinity, where downstream properties and experimental heuristics from human experts can be readily incorporated in a multi-objective optimization framework. Interestingly, techniques such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Bai et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) can instill desired behaviors in the responses generated by large language models; e.g., see (Park et al., 2023) for unconditional small molecule generation. In this study, we optimize autoregressive pLMs to capture diverse preferred and undesired protein sequence distributions, while conditioned on target receptor sequences. This approach enables development of computational frameworks for multi-objective drug design. We use positive and negative data distributions for protein-peptide binding affinity as well as peptide isoelectric points (pI) to show that the model can learn to generate novel sequences that simultaneously respect these different objectives.

Aikium founding team outside their lab at UC Berkeley. Still in seed funding mode this is one of the hottest AI/biotech companies in existence.

Source: https://arxiv.org/html/2403.04187v1

Approaches to Solving the "Dirty Data" Problem

Subsection

Examples of Companies that Are Working to Improve Chemistry Binding Data

Iambic Excels in Predicting Protein – Ligand Binding

Qiao, Z., Nie, W., Vahdat, A. et al. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat Mach Intell* 6, 195–208 (2024).

The binding complexes formed by proteins and small molecule ligands are ubiquitous and critical to life. Despite recent advancements in protein structure prediction, existing algorithms are so far unable to systematically predict the binding ligand structures along with their regulatory effects on protein folding.

To address this discrepancy, we present NeuralPLexer, a computational approach that can directly predict protein–ligand complex structures solely using protein sequence and ligand molecular graph inputs. NeuralPLexer adopts a deep generative model to sample the three-dimensional structures of the binding complex and their conformational changes at an atomistic resolution. The model is based on a diffusion process that incorporates essential biophysical constraints and a multiscale geometric deep learning system to iteratively sample residue-level contact maps and all heavy-atom coordinates in a hierarchical manner.

NeuralPLexer achieves state-of-the-art performance compared with all existing methods on benchmarks for both protein–ligand blind docking and flexible binding-site structure recovery. Moreover, owing to its specificity in sampling both ligand-free-state and ligand-bound-state ensembles, NeuralPLexer consistently outperforms AlphaFold2 in terms of global protein structure accuracy on both representative structure pairs with large conformational changes and recently determined ligand-binding proteins. NeuralPLexer predictions align with structure determination experiments for important targets in enzyme engineering and drug discovery, suggesting its potential for accelerating the design of functional proteins and small molecules at the proteome scale.

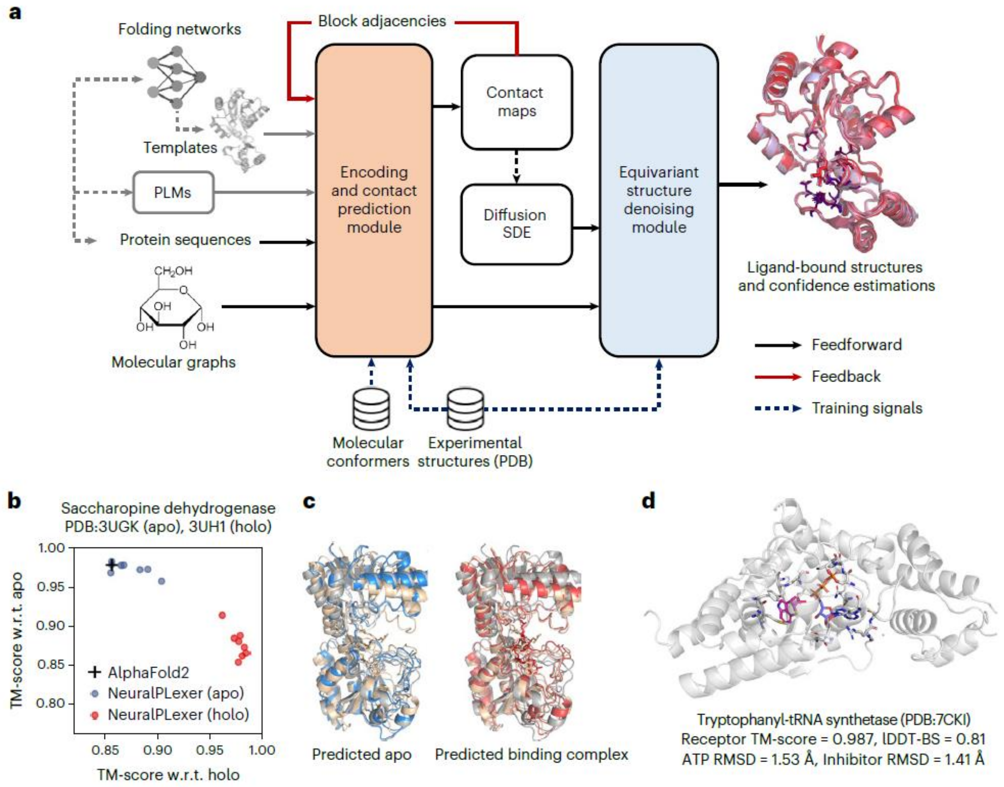
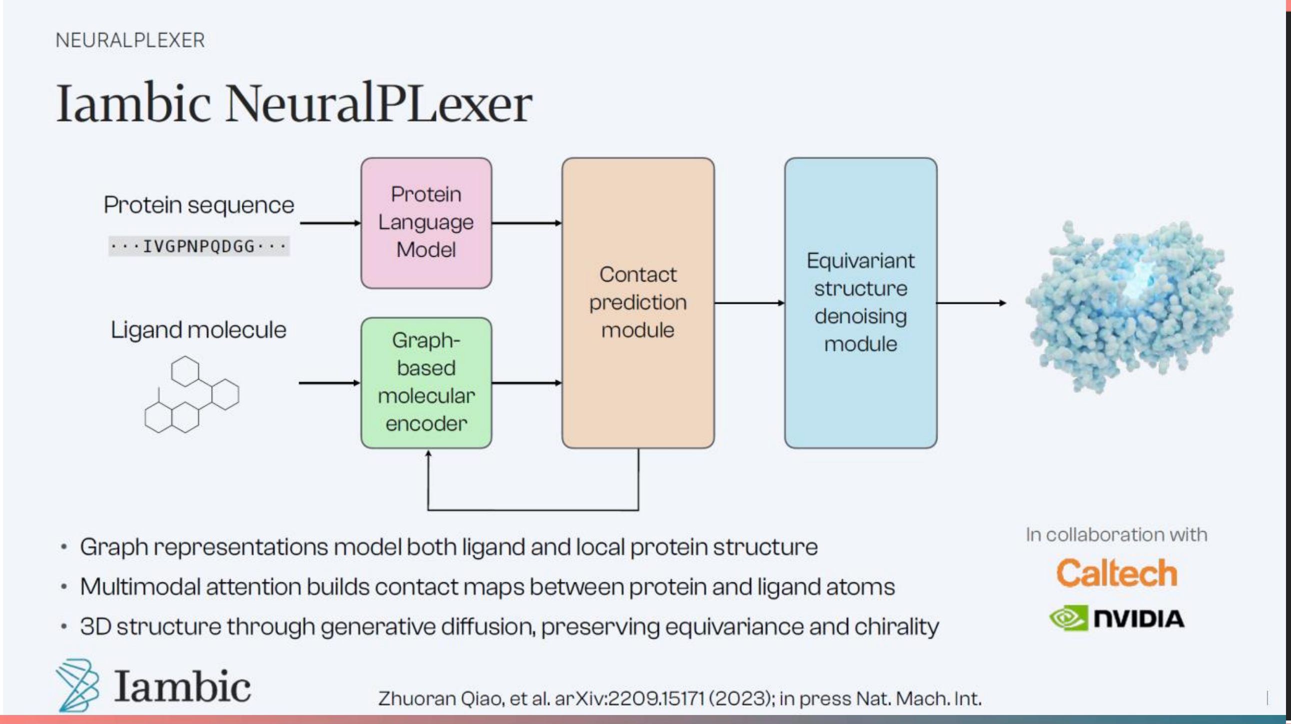

a
Folding networks
Block adjacencies
Templates
PLMs
Protein sequences
CH2OH
OH
OH
OH
OH
Molecular graphs
Encoding and contact prediction module
Contact maps
Diffusion SDE
Equivariant structure denoising module
Ligand-bound structures and confidence estimations
Feedforward
Feedback
Training signals
Molecular conformers
Experimental structures (PDB)
b
Saccharopine dehydrogenase
PDB:3UGK (apo), 3UH1 (holo)
TM-score w.r.t. apo
1.00
0.95
0.90
0.85
0.80
AlphaFold2
NeuralPLexer (apo)
NeuralPLexer (holo)
0.85
0.90
0.95
1.00
TM-score w.r.t. holo
c
Predicted apo
Predicted binding complex
d
Tryptophanyl-tRNA synthetase (PDB:7CKI)
Receptor TM-score = 0.987, lDDT-BS = 0.81
ATP RMSD = 1.53 Å, Inhibitor RMSD = 1.41 Å

Fig. 1 | NeuralPLexer enables accurate prediction of protein–ligand complex structure and conformational changes. a, Method overview. To perform predictions, the input protein sequence is first used to retrieve PLM features and structure templates; NeuralPLexer then combines the set of PLM and template features with molecular graph representations of the input ligands to directly sample an ensemble of binding complex structures via a multiscale generative model. The main network of NeuralPLexer is composed of a coarse-grained, auto-regressive CPM and an atomistic, diffusion-based ESDM. **b**,**c**, Prediction example on a target with large-scale domain motions upon ligand binding (UniProt:P38998). **b**, The structure similarities against experimental apo (that is, ligand-free protein, PDB:3UGK) and holo (that is, ligand-bound protein, PDB:3UH1) structures measured by TM-score are plotted for AF2 predictions (grey), ligand-free NeuralPLexer predictions (blue) and ligand-bound NeuralPLexer predictions (red). **c**, Visualizations of representative NeuralPLexer-predicted structures (blue for apo, red for holo) are overlaid with the experimental apo structure (grey) and the holo structure (light yellow). **d**, Visualization of a prediction example (PDB:7CKI, UniProt:P00953) for which NeuralPLexer achieves high atomic accuracy for both the ATP (blue) and an inhibitor bound to the tryptophan site (magenta) upon an induced-fit structure rearrangement. w.r.t., with respect to.

Source: https://www.nature.com/articles/s42256-024-00792-z



NEURALPLEXER

Iambic NeuralPLexer

Protein sequence

···IVGPNPQDGG···

Ligand molecule

Protein Language Model

Graph-based molecular encoder

Contact prediction module

Equivariant structure denoising module

- Graph representations model both ligand and local protein structure
- Multimodal attention builds contact maps between protein and ligand atoms
- 3D structure through generative diffusion, preserving equivariance and chirality

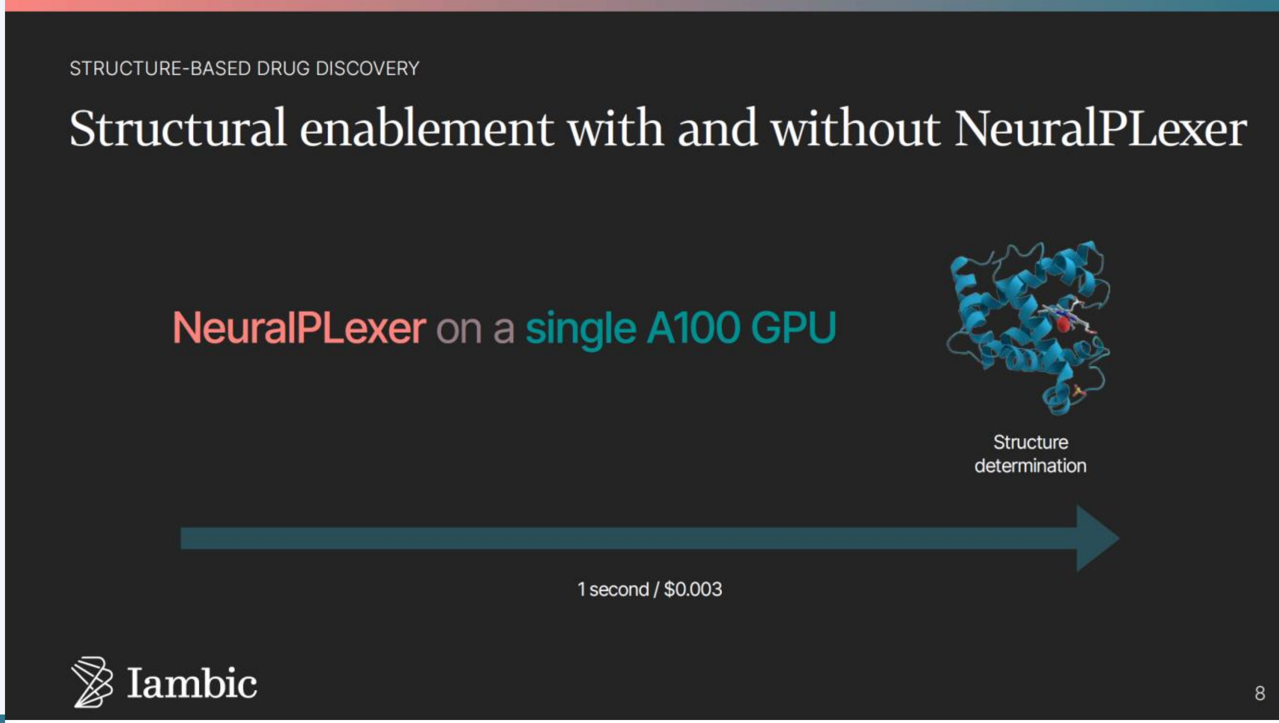
In collaboration with

Caltech

NVIDIA

Iambic

Zhuoran Qiao, et al. arXiv:2209.15171 (2023); in press Nat. Mach. Int.



STRUCTURE-BASED DRUG DISCOVERY

Structural enablement with and without NeuralPLexer

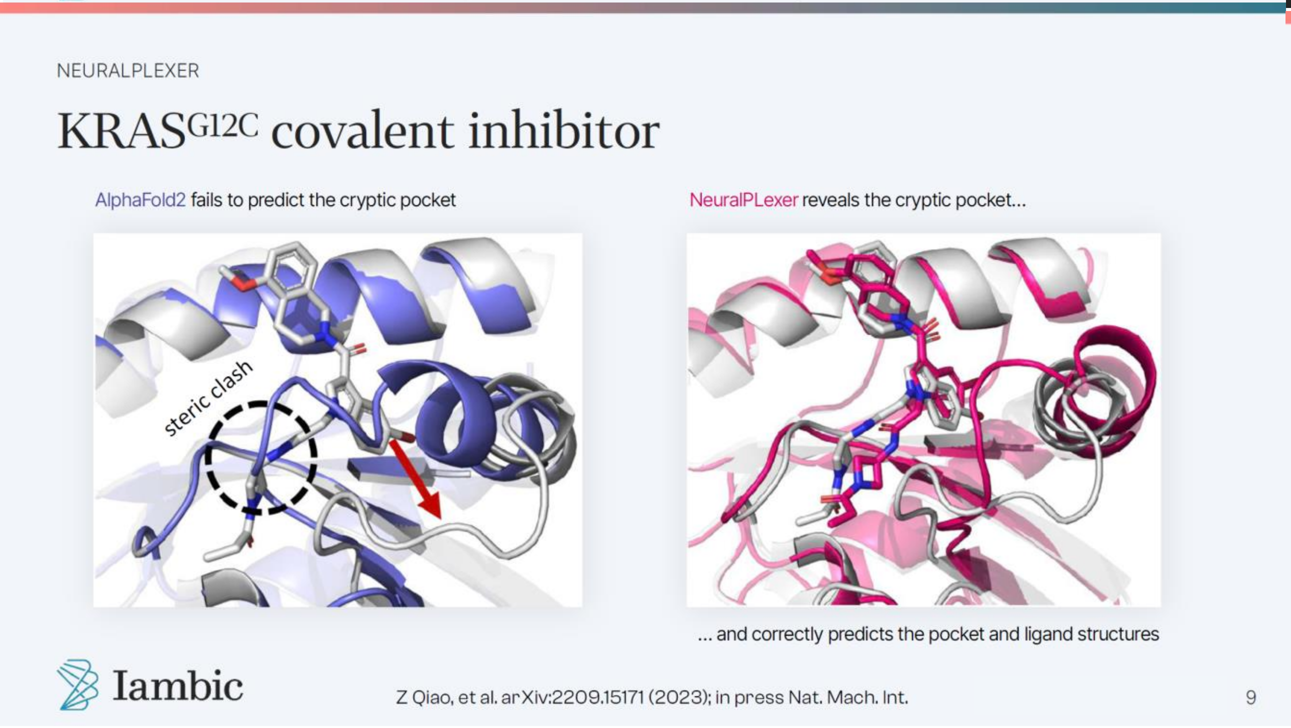
NeuralPLexer on a single A100 GPU

Structure determination

1 second / $0.003

Iambic

8



NEURALPLEXER

KRAS^{G12C} covalent inhibitor

AlphaFold2 fails to predict the cryptic pocket

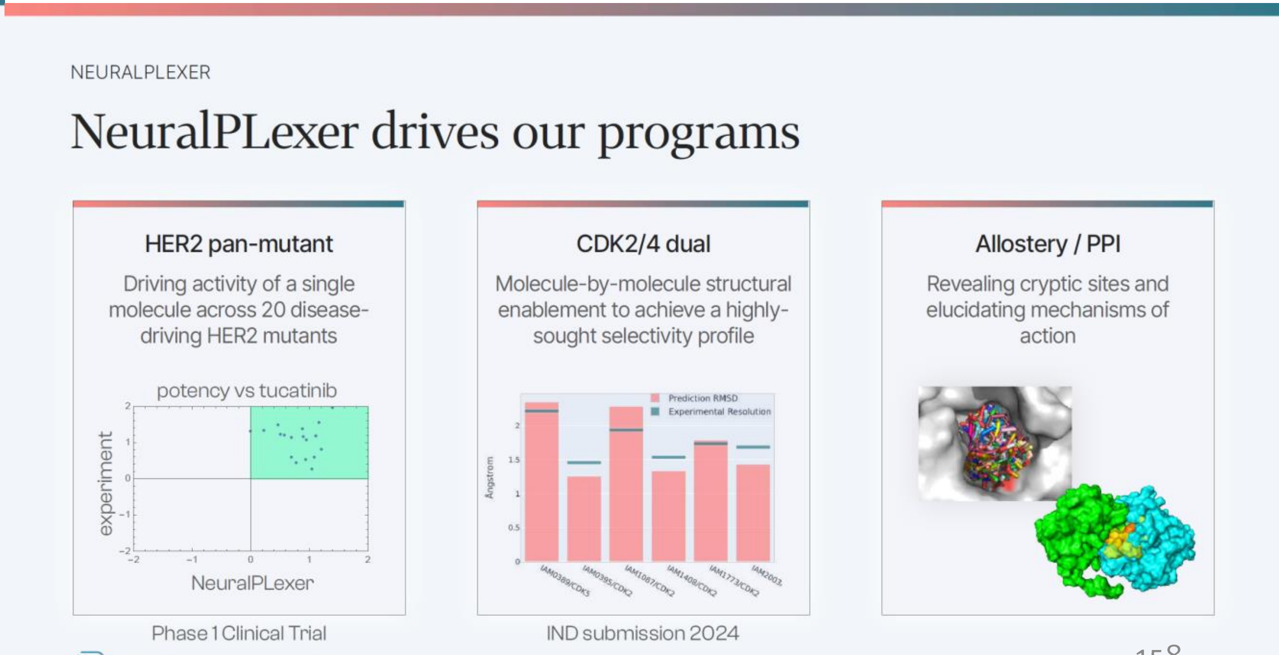
NeuralPLexer reveals the cryptic pocket...

... and correctly predicts the pocket and ligand structures

Iambic

Z Qiao, et al. arXiv:2209.15171 (2023); in press Nat. Mach. Int.

9



NEURALPLEXER

NeuralPLexer drives our programs

HER2 pan-mutant

Driving activity of a single molecule across 20 disease-driving HER2 mutants

Phase 1 Clinical Trial

CDK2/4 dual

Molecule-by-molecule structural enablement to achieve a highly-sought selectivity profile

IND submission 2024

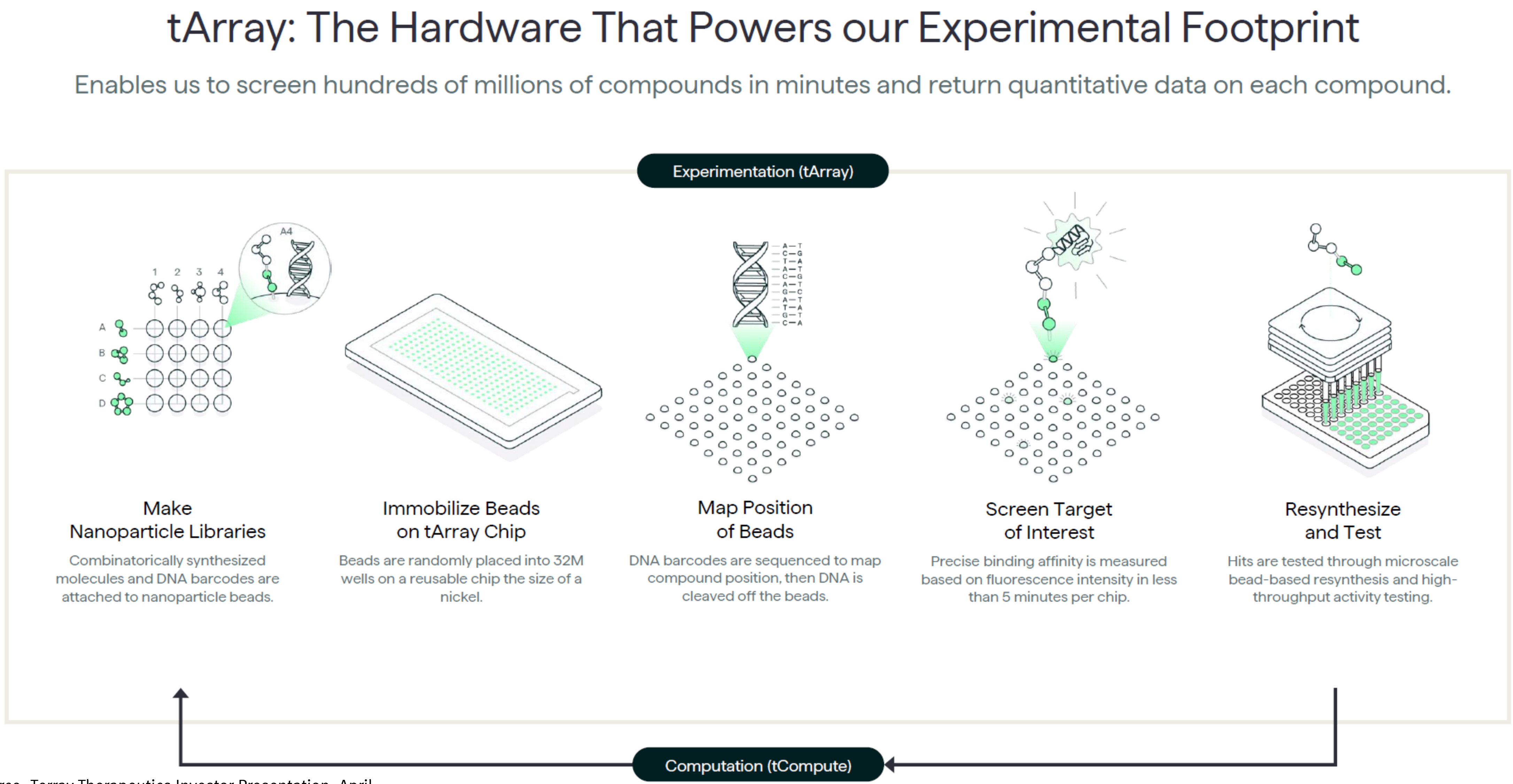
Allostery / PPI

Revealing cryptic sites and elucidating mechanisms of action

Iambic

Source: Iambic Investor Presentation, April 2024

Terray Uses a Bead Approach and a Custom Chip to Find Compound / Target Binding Characteristics



tArray: The Hardware That Powers our Experimental Footprint

Enables us to screen hundreds of millions of compounds in minutes and return quantitative data on each compound.

Experimentation (tArray)

Make Nanoparticle Libraries
Combinatorically synthesized molecules and DNA barcodes are attached to nanoparticle beads.

Immobilize Beads on tArray Chip
Beads are randomly placed into 32M wells on a reusable chip the size of a nickel.

Map Position of Beads
DNA barcodes are sequenced to map compound position, then DNA is cleaved off the beads.

Screen Target of Interest
Precise binding affinity is measured based on fluorescence intensity in less than 5 minutes per chip.

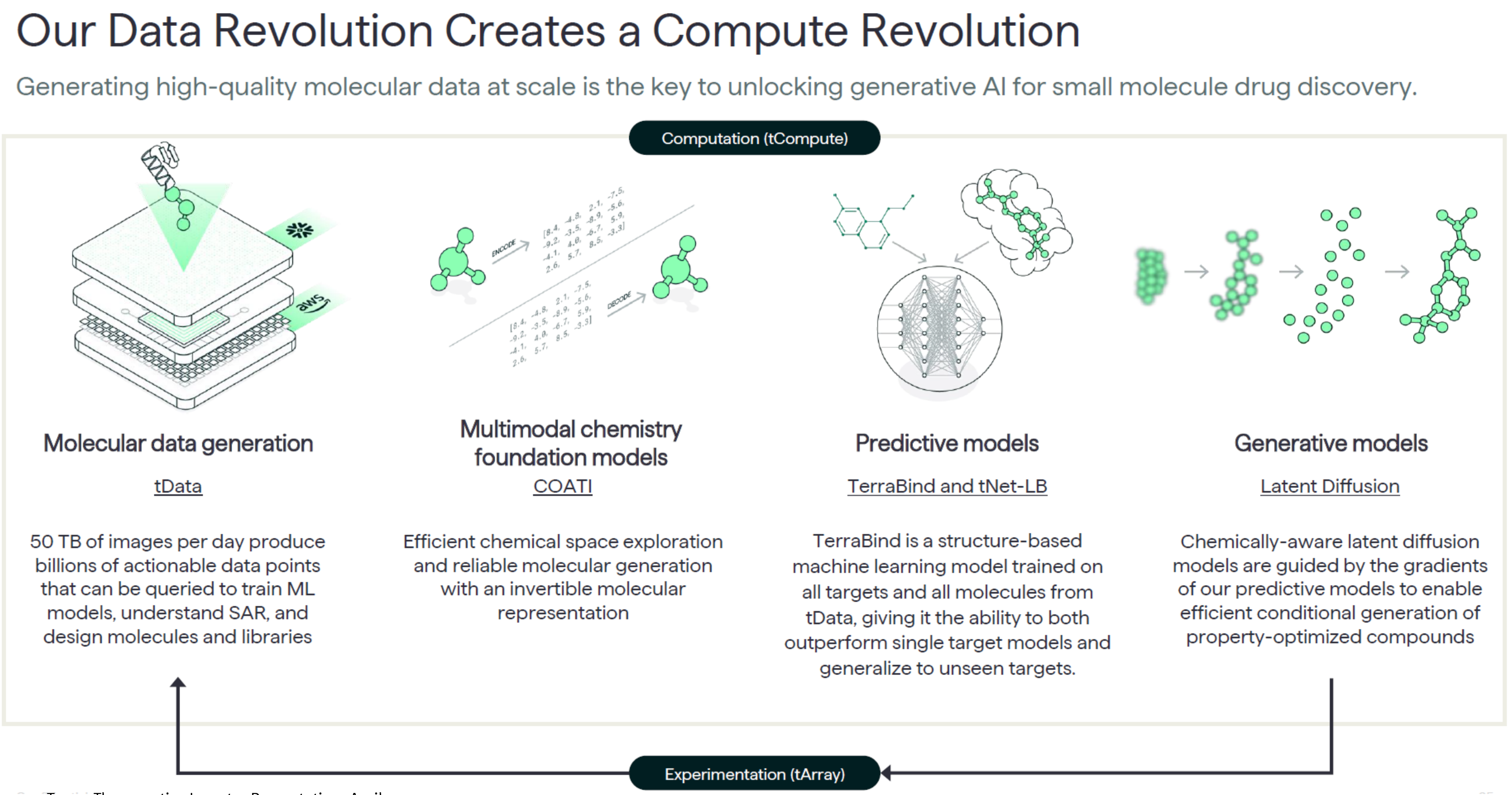
Resynthesize and Test
Hits are tested through microscale bead-based resynthesis and high-throughput activity testing.

Computation (tCompute)

Source: Terray Therapeutics Investor Presentation, April 2024, (https://www.terraytx.com/).

TERRAY

It’s All About the Data



Our Data Revolution Creates a Compute Revolution

Generating high-quality molecular data at scale is the key to unlocking generative AI for small molecule drug discovery.

Computation (tCompute)

Molecular data generation

tData

50 TB of images per day produce billions of actionable data points that can be queried to train ML models, understand SAR, and design molecules and libraries

Multimodal chemistry foundation models

COATI

Efficient chemical space exploration and reliable molecular generation with an invertible molecular representation

Predictive models

TerraBind and tNet-LB

TerraBind is a structure-based machine learning model trained on all targets and all molecules from tData, giving it the ability to both outperform single target models and generalize to unseen targets.

Generative models

Latent Diffusion

Chemically-aware latent diffusion models are guided by the gradients of our predictive models to enable efficient conditional generation of property-optimized compounds

Experimentation (tArray)

Source: Terray Therapeutics Investor Presentation, April 2024, (https://www.terraytx.com/).

DELs in a Nutshell: AI Meets DEL: Is This The Most Powerful Combo in Modern Drug Discovery?

Receptor.AI, September 23, 2023 (excerpt)

DNA-encoded libraries are revolutionizing modern drug discovery by allowing an unprecedented amount of small molecules to be screened automatically. The technology is based on the conjugation of small molecules with unique DNA tags, which could be used to identify those which bind to the protein target of interest.

In contrast to traditional HTS, the DEL technology offers the possibility to screen an overwhelming amount (hundred of millions) of molecular species in a single experiment.

The traditional screening techniques rely mostly on biochemical assays, which allow some kind of automatic readout. Each compound must be placed on the dedicated cell on the plate, severely limiting the number of molecules that could be evaluated simultaneously. Even the most advanced screening robots are limited by the physical size of the plates and the latency of the corresponding biochemical reactions. Even if physical detection techniques, such as surface plasmon resonance or capacitance sensors, are employed instead of biochemical assays, the “one cell — one compound” restrictions still hamper the scaling of the screening.

The DELs benefit from the DNA tags that could be read out by modern sequencing techniques with an overwhelming sensitivity and selectivity. In theory, even a single tagged molecule could be detected among the millions of others. At the same time, the number of tags with particular sequences present in the sample could also be determined quantitatively. All this allows for employing radically different screening paradigms based on physical binding rather than any kind of functional activity.

The idea under all DELs screening techniques is to immobilize the target proteins and incubate the DNA-tagged small molecules with them. Those molecules which bind to the target will be trapped, while the rest could be washed out. Next, the tags are sequenced, and the corresponding species of molecules binding to the target are identified and quantified.

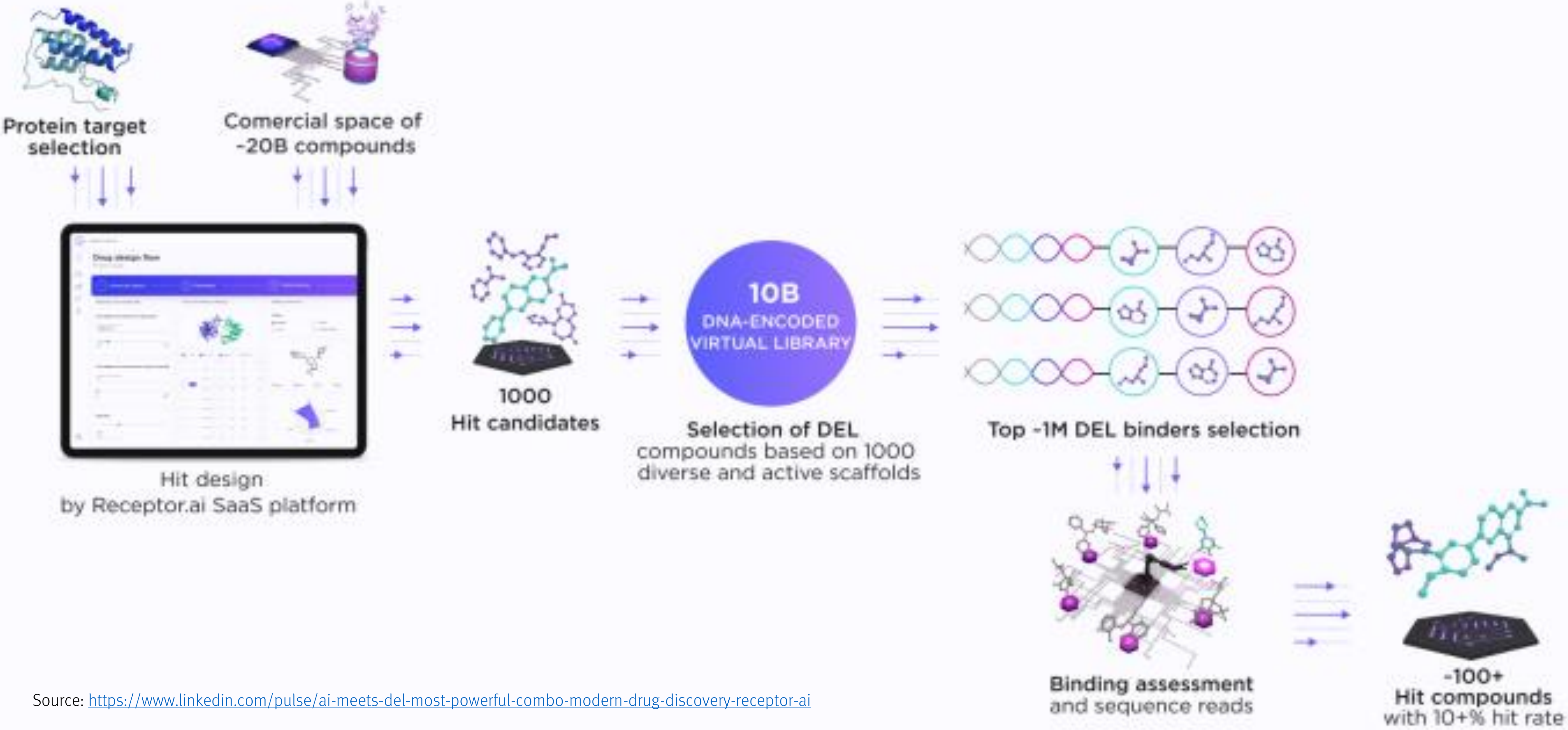
The target proteins could be immobilized on magnetic beads, porous resin or various solid surfaces. It is even possible to utilize the proteins expressed on the cell’s surface.

DELs are all about big data, which makes them perfectly compatible with machine learning. The ML is unable to mitigate the inherent physical limitations of DELs, but it can help with data management at all important stages.

Source: https://www.linkedin.com/pulse/ai-meets-del-most-powerful-combo-modern-drug-discovery-receptor-ai

AI FIRST, DEL SECOND

Applying of Receptor.ai SaaS platform for initial selection of potent DEL libraries for novel hit discovery



Protein target selection

Comercial space of ~20B compounds

Hit design by Receptor.ai SaaS platform

1000 Hit candidates

10B DNA-ENCODED VIRTUAL LIBRARY

Selection of DEL compounds based on 1000 diverse and active scaffolds

Top ~1M DEL binders selection

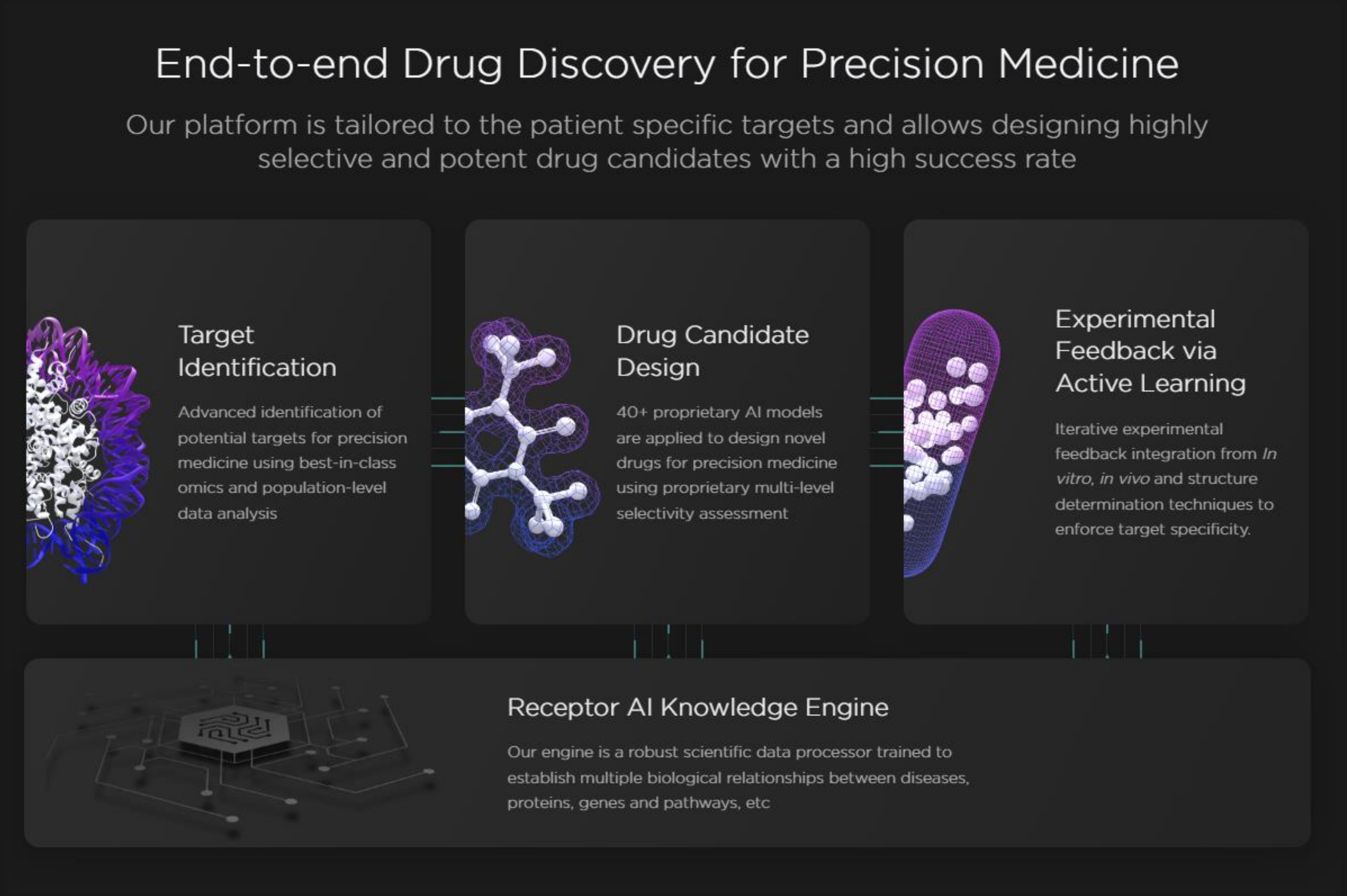
Binding assessment and sequence reads

~100+ Hit compounds with 10+% hit rate

Source: https://www.linkedin.com/pulse/ai-meets-del-most-powerful-combo-modern-drug-discovery-receptor-ai

RECEPTOR.AI

Predicts binding of small molecule ligands in protein pockets at speed and accuracy



End-to-end Drug Discovery for Precision Medicine

Our platform is tailored to the patient specific targets and allows designing highly selective and potent drug candidates with a high success rate

Target Identification

Advanced identification of potential targets for precision medicine using best-in-class omics and population-level data analysis

Drug Candidate Design

40+ proprietary AI models are applied to design novel drugs for precision medicine using proprietary multi-level selectivity assessment

Experimental Feedback via Active Learning

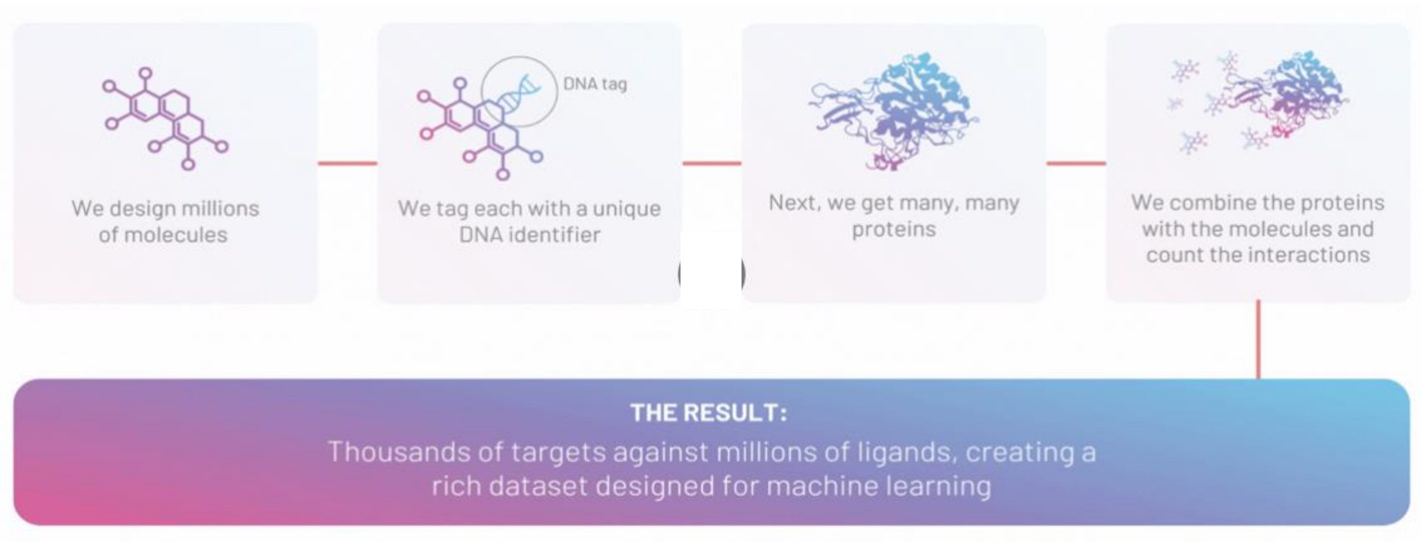
Iterative experimental feedback integration from *In vitro, in vivo* and structure determination techniques to enforce target specificity.

Receptor AI Knowledge Engine

Our engine is a robust scientific data processor trained to establish multiple biological relationships between diseases, proteins, genes and pathways, etc

Source: Receptor.ai website, April 2024

Leash Rents Out Giant DNA Encoded Library (DEL) on Protein Binding Partners



We design millions of molecules

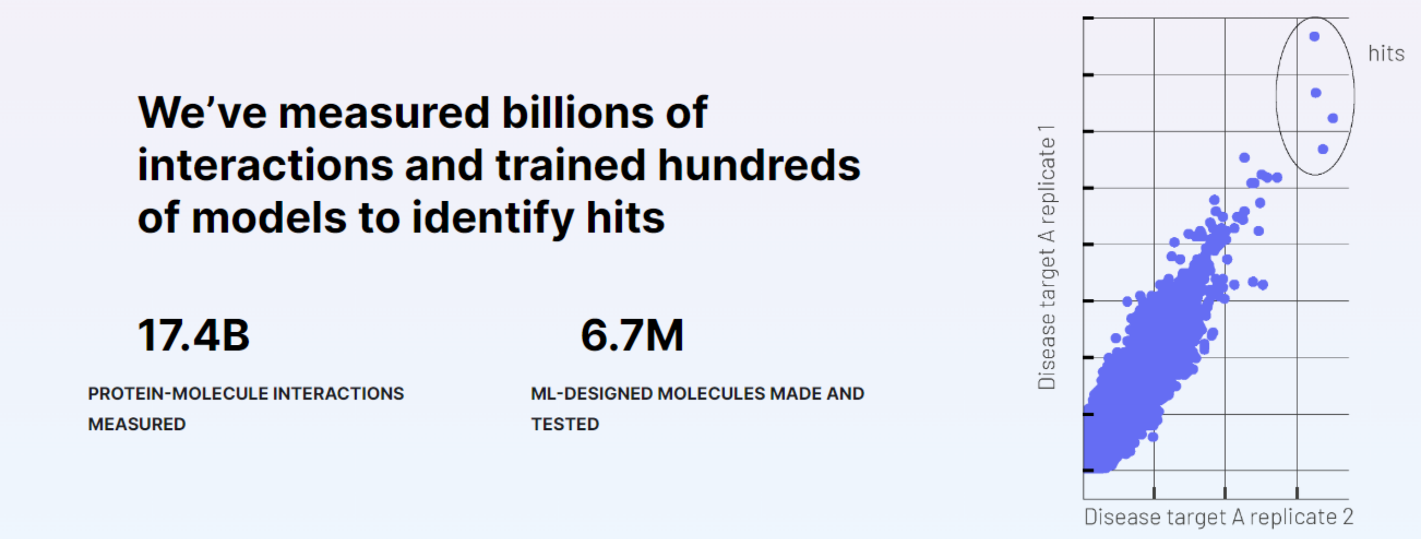
We tag each with a unique DNA identifier

Next, we get many, many proteins

We combine the proteins with the molecules and count the interactions

THE RESULT:

Thousands of targets against millions of ligands, creating a rich dataset designed for machine learning



We've measured billions of interactions and trained hundreds of models to identify hits

17.4B

PROTEIN-MOLECULE INTERACTIONS MEASURED

6.7M

ML-DESIGNED MOLECULES MADE AND TESTED

Andrew Dunn, *Endpoints News*, April 4, 2024

Leash's dataset includes about 3.6 billion biological activities, generated by testing DNA-encoded small molecules. Quigley compared the method to fishing. They attach unique pieces of DNA to tens of millions of molecules in a tube. They then drop a target protein into the tube, like a fish hook, and yank it out. The stuff sticking to that hook is the molecules binding to that protein. Leash sequences those compounds and keeps repeating this process, building a massive molecule library.

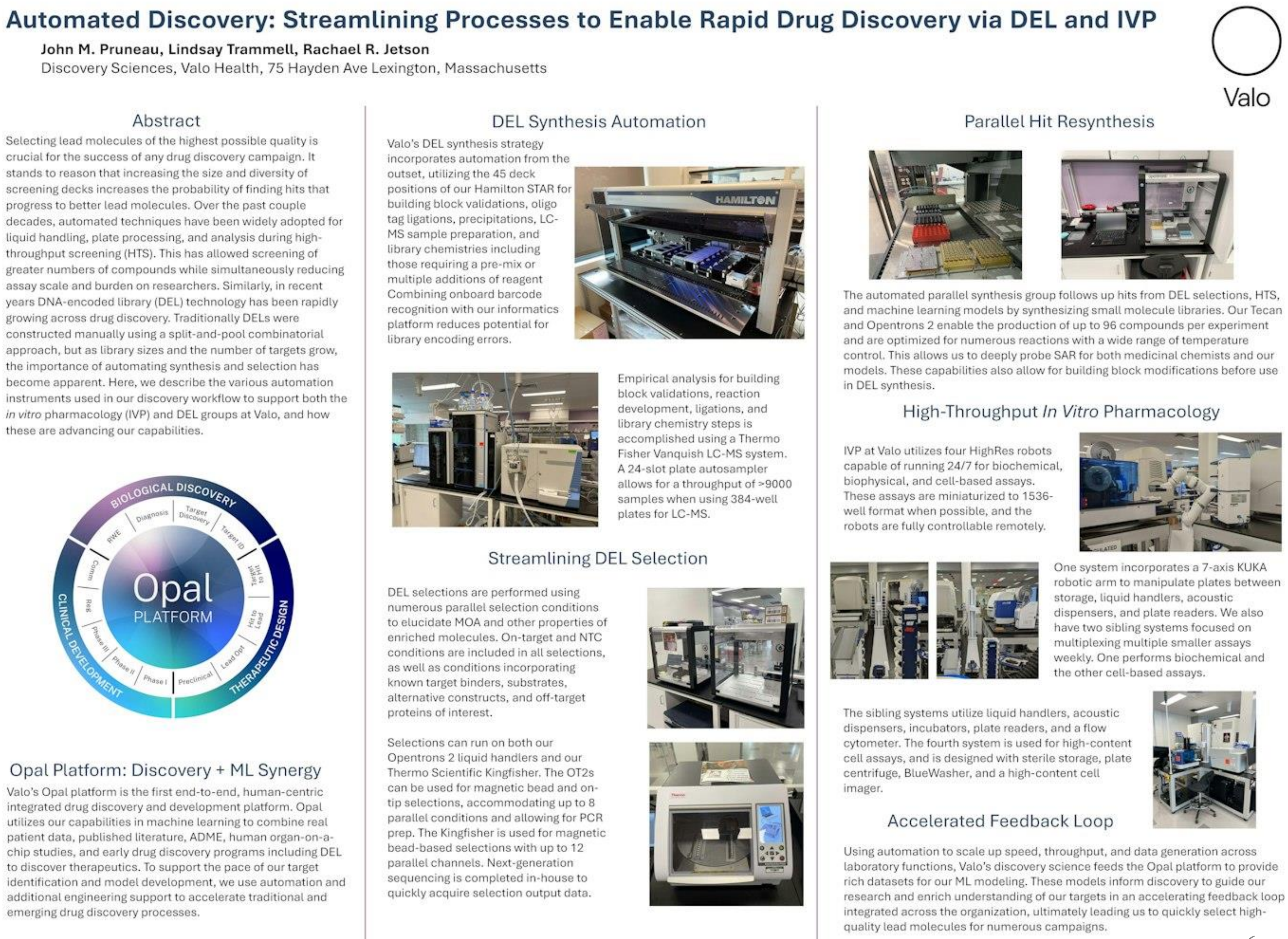
Leash's dataset comes from screening 133 million molecules against three protein targets. Quigley pointed to other relevant public datasets like BindingDB, a curated set of nearly three million data points, and PubChem BioAssay database, which has about 294 million activities.

Leash's long-term goal is to generate enough data to make reliable binding predictions for any small molecule and any protein — a key part of the early drug R&D process. Quigley said Leash is also advancing two cancer drug programs in the earliest stages, declining to share timelines or further details on the pipeline for now.

Source: https://endpts.com/exclusive-ex-recursion-team-launches-leash-bio-challenging-ai-field-with-its-bet-on-needing-more-data/, https://leash.bio

Valo Also Uses DELS

Selecting lead molecules of the highest possible quality is crucial for the success of any drug discovery campaign. It stands to reason that increasing the size and diversity of screening decks increases the probability of finding hits that progress to better lead molecules. Over the past couple decades, automated techniques have been widely adopted for liquid handling, plate processing, and analysis during high throughput screening (HTS). This has allowed screening of greater numbers of compounds while simultaneously reducing assay scale and burden on researchers. Similarly, in recent years DNA encoded library (DEL) technology has been rapidly growing across drug discovery. Traditionally DELs were constructed manually using a split and pool combinatorial approach, but as library sizes and the number of targets grow, the importance of automating synthesis and selection has become apparent. Here, we describe the various automation instruments used in our discovery workflow to support both the in vitro pharmacology (IVP) and DEL groups at Valo, and how these are advancing our capabilities.



Automated Discovery: Streamlining Processes to Enable Rapid Drug Discovery via DEL and IVP

John M. Pruneau, Lindsay Trammell, Rachael R. Jetson
Discovery Sciences, Valo Health, 75 Hayden Ave Lexington, Massachusetts

Valo

Abstract

Selecting lead molecules of the highest possible quality is crucial for the success of any drug discovery campaign. It stands to reason that increasing the size and diversity of screening decks increases the probability of finding hits that progress to better lead molecules. Over the past couple decades, automated techniques have been widely adopted for liquid handling, plate processing, and analysis during high-throughput screening (HTS). This has allowed screening of greater numbers of compounds while simultaneously reducing assay scale and burden on researchers. Similarly, in recent years DNA-encoded library (DEL) technology has been rapidly growing across drug discovery. Traditionally DELs were constructed manually using a split-and-pool combinatorial approach, but as library sizes and the number of targets grow, the importance of automating synthesis and selection has become apparent. Here, we describe the various automation instruments used in our discovery workflow to support both the *in vitro* pharmacology (IVP) and DEL groups at Valo, and how these are advancing our capabilities.

Opal Platform: Discovery + ML Synergy

Valo's Opal platform is the first end-to-end, human-centric integrated drug discovery and development platform. Opal utilizes our capabilities in machine learning to combine real patient data, published literature, ADME, human organ-on-a-chip studies, and early drug discovery programs including DEL to discover therapeutics. To support the pace of our target identification and model development, we use automation and additional engineering support to accelerate traditional and emerging drug discovery processes.

DEL Synthesis Automation

Valo's DEL synthesis strategy incorporates automation from the outset, utilizing the 45 deck positions of our Hamilton STAR for building block validations, oligo tag ligations, precipitations, LC-MS sample preparation, and library chemistries including those requiring a pre-mix or multiple additions of reagent Combining onboard barcode recognition with our informatics platform reduces potential for library encoding errors.

Empirical analysis for building block validations, reaction development, ligations, and library chemistry steps is accomplished using a Thermo Fisher Vanquish LC-MS system. A 24-slot plate autosampler allows for a throughput of >9000 samples when using 384-well plates for LC-MS.

Streamlining DEL Selection

DEL selections are performed using numerous parallel selection conditions to elucidate MOA and other properties of enriched molecules. On-target and NTC conditions are included in all selections, as well as conditions incorporating known target binders, substrates, alternative constructs, and off-target proteins of interest.

Selections can run on both our Opentrons 2 liquid handlers and our Thermo Scientific Kingfisher. The OT2s can be used for magnetic bead and on-tip selections, accommodating up to 8 parallel conditions and allowing for PCR prep. The Kingfisher is used for magnetic bead-based selections with up to 12 parallel channels. Next-generation sequencing is completed in-house to quickly acquire selection output data.

Parallel Hit Resynthesis

The automated parallel synthesis group follows up hits from DEL selections, HTS, and machine learning models by synthesizing small molecule libraries. Our Tecan and Opentrons 2 enable the production of up to 96 compounds per experiment and are optimized for numerous reactions with a wide range of temperature control. This allows us to deeply probe SAR for both medicinal chemists and our models. These capabilities also allow for building block modifications before use in DEL synthesis.

High-Throughput *In Vitro* Pharmacology

IVP at Valo utilizes four HighRes robots capable of running 24/7 for biochemical, biophysical, and cell-based assays. These assays are miniaturized to 1536-well format when possible, and the robots are fully controllable remotely.

One system incorporates a 7-axis KUKA robotic arm to manipulate plates between storage, liquid handlers, acoustic dispensers, and plate readers. We also have two sibling systems focused on multiplexing multiple smaller assays weekly. One performs biochemical and the other cell-based assays.

The sibling systems utilize liquid handlers, acoustic dispensers, incubators, plate readers, and a flow cytometer. The fourth system is used for high-content cell assays, and is designed with sterile storage, plate centrifuge, BlueWasher, and a high-content cell imager.

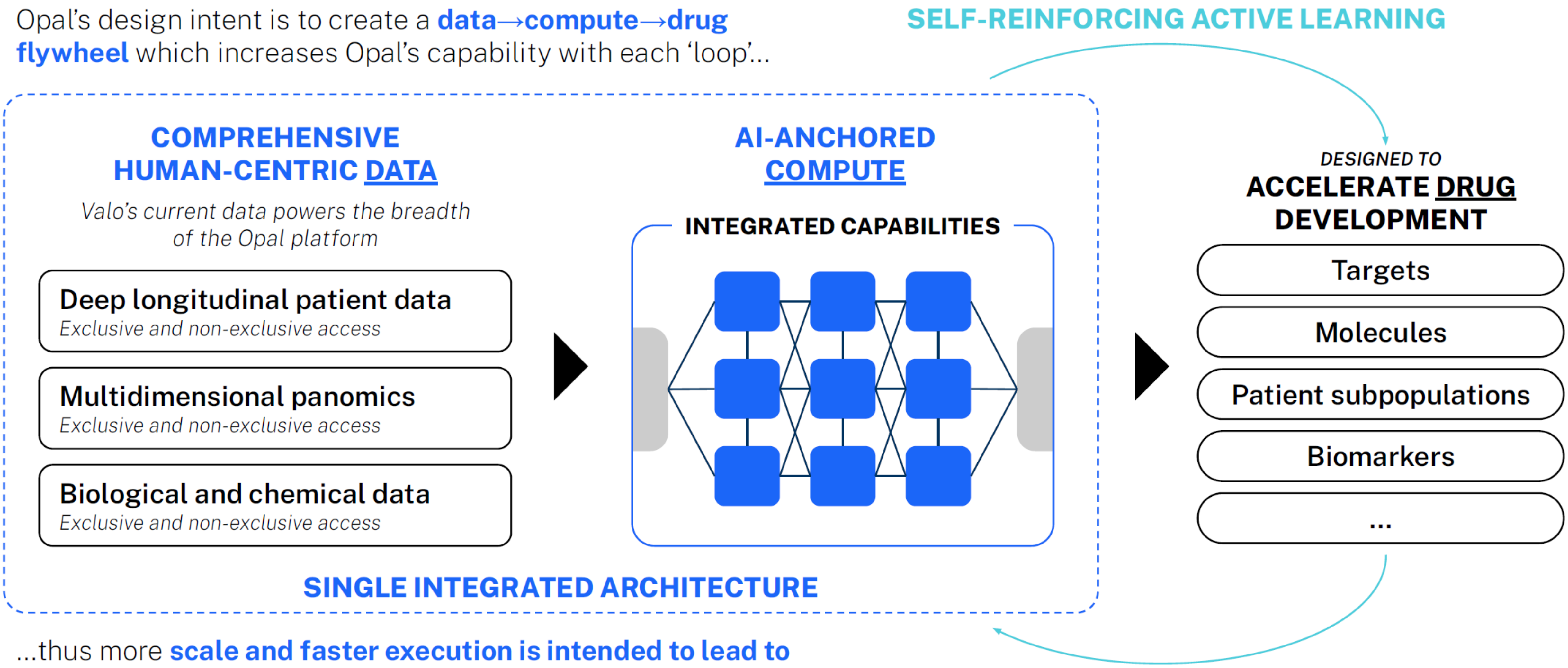
Accelerated Feedback Loop

Using automation to scale up speed, throughput, and data generation across laboratory functions, Valo's discovery science feeds the Opal platform to provide rich datasets for our ML modeling. These models inform discovery to guide our research and enrich understanding of our targets in an accelerating feedback loop integrated across the organization, ultimately leading us to quickly select high-quality lead molecules for numerous campaigns.

Source: https://www.valohealth.com/publications/poster-automated-discovery-streamlining-processes-to-enable-rapid-drug-discovery-via-del-and-ivp

Valo's Opal platform consists of an integrated set of capabilities designed to transform data into valuable insights that may accelerate discoveries

Opal's design intent is to create a **data→compute→drug flywheel** which increases Opal's capability with each 'loop'...



SELF-REINFORCING ACTIVE LEARNING

COMPREHENSIVE HUMAN-CENTRIC DATA

Valo's current data powers the breadth of the Opal platform

Deep longitudinal patient data
Exclusive and non-exclusive access

Multidimensional panomics
Exclusive and non-exclusive access

Biological and chemical data
Exclusive and non-exclusive access

AI-ANCHORED COMPUTE

INTEGRATED CAPABILITIES

SINGLE INTEGRATED ARCHITECTURE

DESIGNED TO
ACCELERATE DRUG DEVELOPMENT

- Targets
- Molecules
- Patient subpopulations
- Biomarkers
- ...

...thus more **scale and faster execution is intended to lead to increasing capability and competitive advantage**

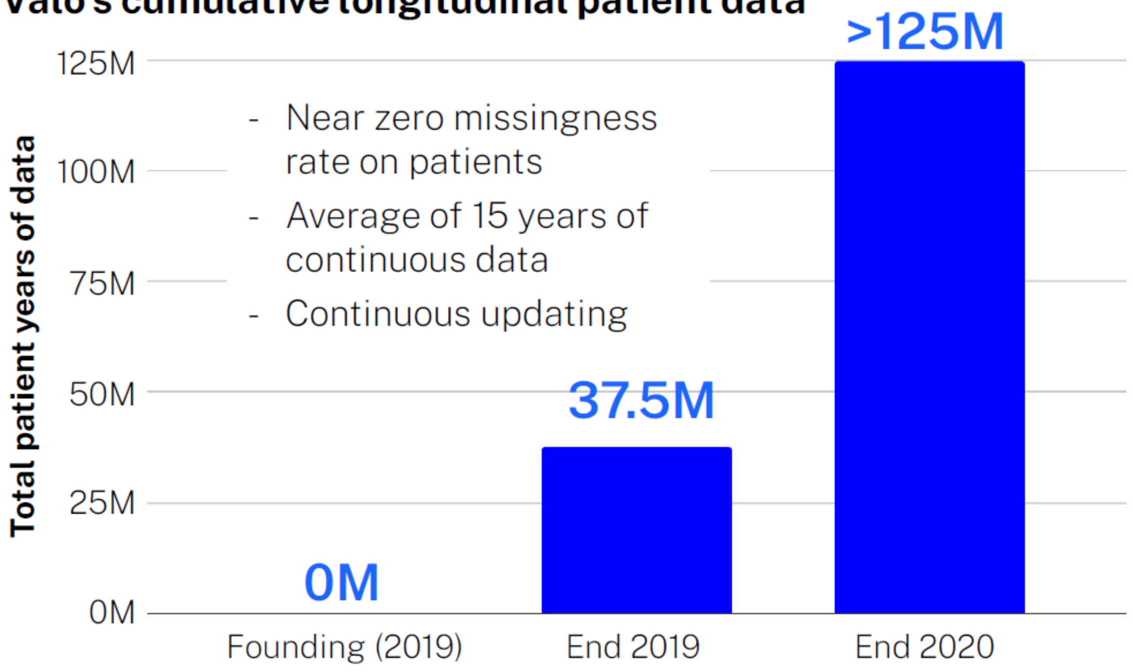
Source: Valo Corporate Presentation

Opal is built upon a differentiated, human-centric, and high quality data foundation

DATA

>125M years of longitudinal patient data

Valo's cumulative longitudinal patient data



- Near zero missingness rate on patients
- Average of 15 years of continuous data
- Continuous updating

Multidimensional -'omics

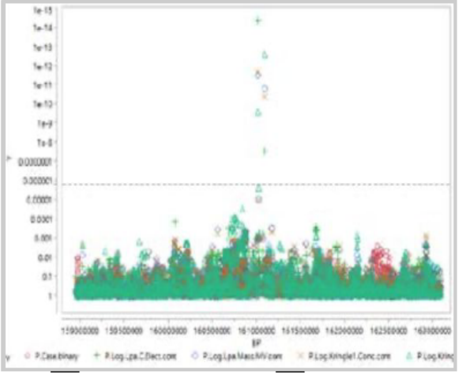
Exclusive access to one of the largest prospective studies spanning pan-omics, imaging, and medical records

>22.5T
Whole genome sequencing data points

>210M
mRNA sequencing data points

>21M
Metabolomic and/or proteomic data points

>320K
Blood sample aliquots



>13K images
paired with related scoring data

Opal fuses Valo's novel and/or exclusive longitudinal and 'omics data using proprietary methodologies designed to enable intelligent imputation, the upgrade of public and semi-private data, and the generation of novel insights

Source: Valo Corporate Presentation

Subsection

Examples of Companies that Are Working to Improve Clinical and Disease Datasets

Rich Ecosystem Emerging to Meet Need for Quality AI Data

Literature Analytics

CytoReason

AITIA

Benevolent AI

causaly

Exscientia

Health Outlook

ontotext

RECURSION

Real World Evidence / EMR

AITIA

CuriMeta

nference

ConcertAI

TEMPUS

TRUVETA

Valo

VERGE GENOMICS

XCURES

Clinical Trials Analytics

antidote

Benevolent AI

IQVIA

MOLECULAR HEALTH

Genomic Data

AMGEN

biobank uk
Improving the health of future generations

Benevolent AI

CARIS LIFE SCIENCES

deep genomics

EMPRESS

MOLECULAR HEALTH

scipher MEDICINE

VERGE GENOMICS

Proteomic Data

celsius

Health Outlook

immunai

insitro

TEMPUS

VERGE GENOMICS

Multi-OMICs Data

biobank uk
Improving the health of future generations

GATC Health

nference

nucleai

TEMPUS

Valo

vevo

Histology / Cell Data

aiforia
AI for image analysis

nucleai

OWKIN

PAIGE

Predictive Oncology

SPRING

Epidemiology Data

CuriMeta

Notable Labs

nference

TEMPUS

Chemical Binding Data

absci

Leash

RECEPTOR.AI

Iambic

TERRAY

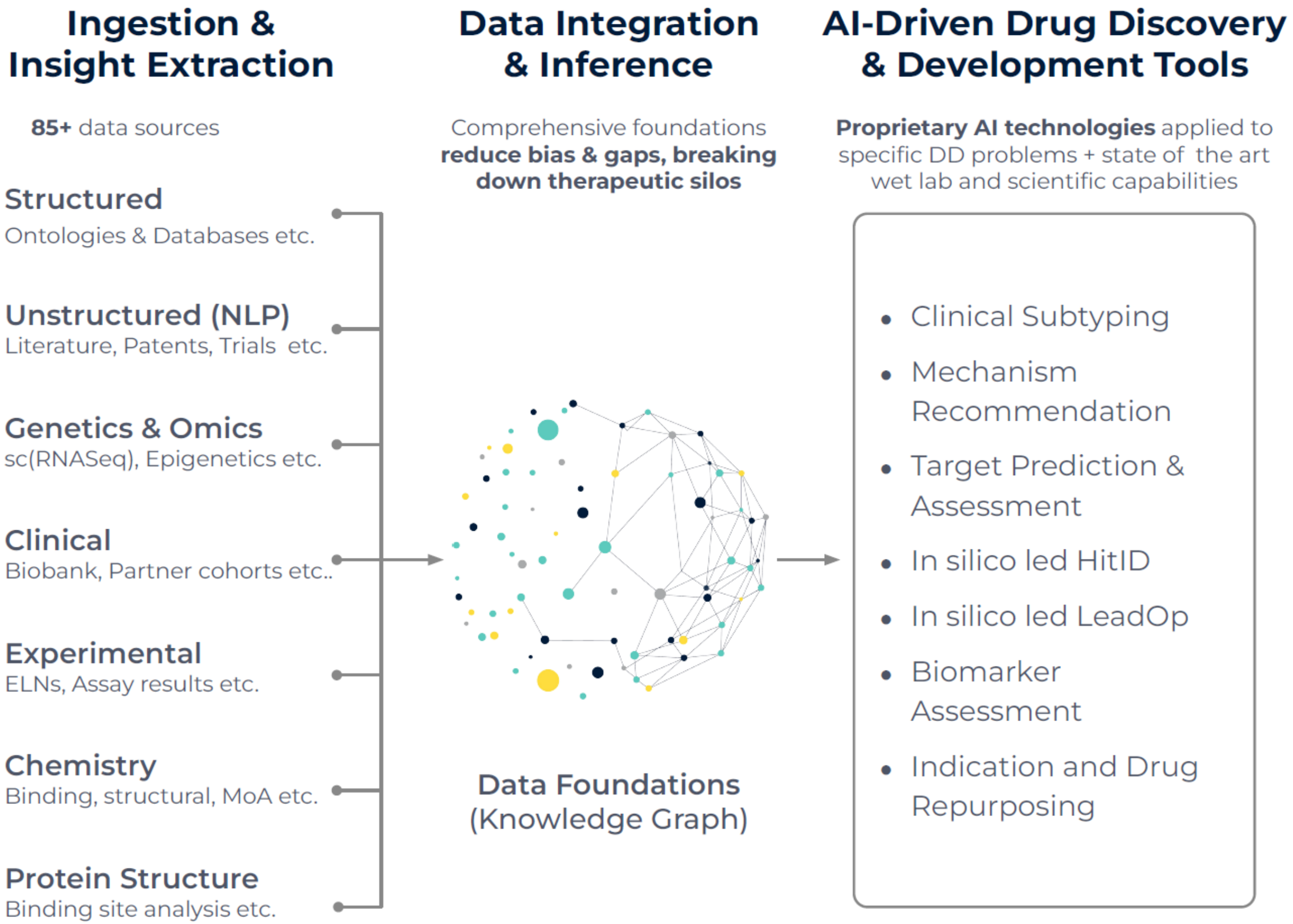
WuXi Biologics
Global Solution Provider

Source: Stifel Research

Data Knowledge Graphs Can Be Built from a Range of Biological Data Types

BenevolentAI

Example



Ingestion & Insight Extraction

85+ data sources

Structured
Ontologies & Databases etc.

Unstructured (NLP)
Literature, Patents, Trials etc.

Genetics & Omics
sc(RNASeq), Epigenetics etc.

Clinical
Biobank, Partner cohorts etc..

Experimental
ELNs, Assay results etc.

Chemistry
Binding, structural, MoA etc.

Protein Structure
Binding site analysis etc.

Data Integration & Inference

Comprehensive foundations **reduce bias & gaps, breaking down therapeutic silos**

Data Foundations
(Knowledge Graph)

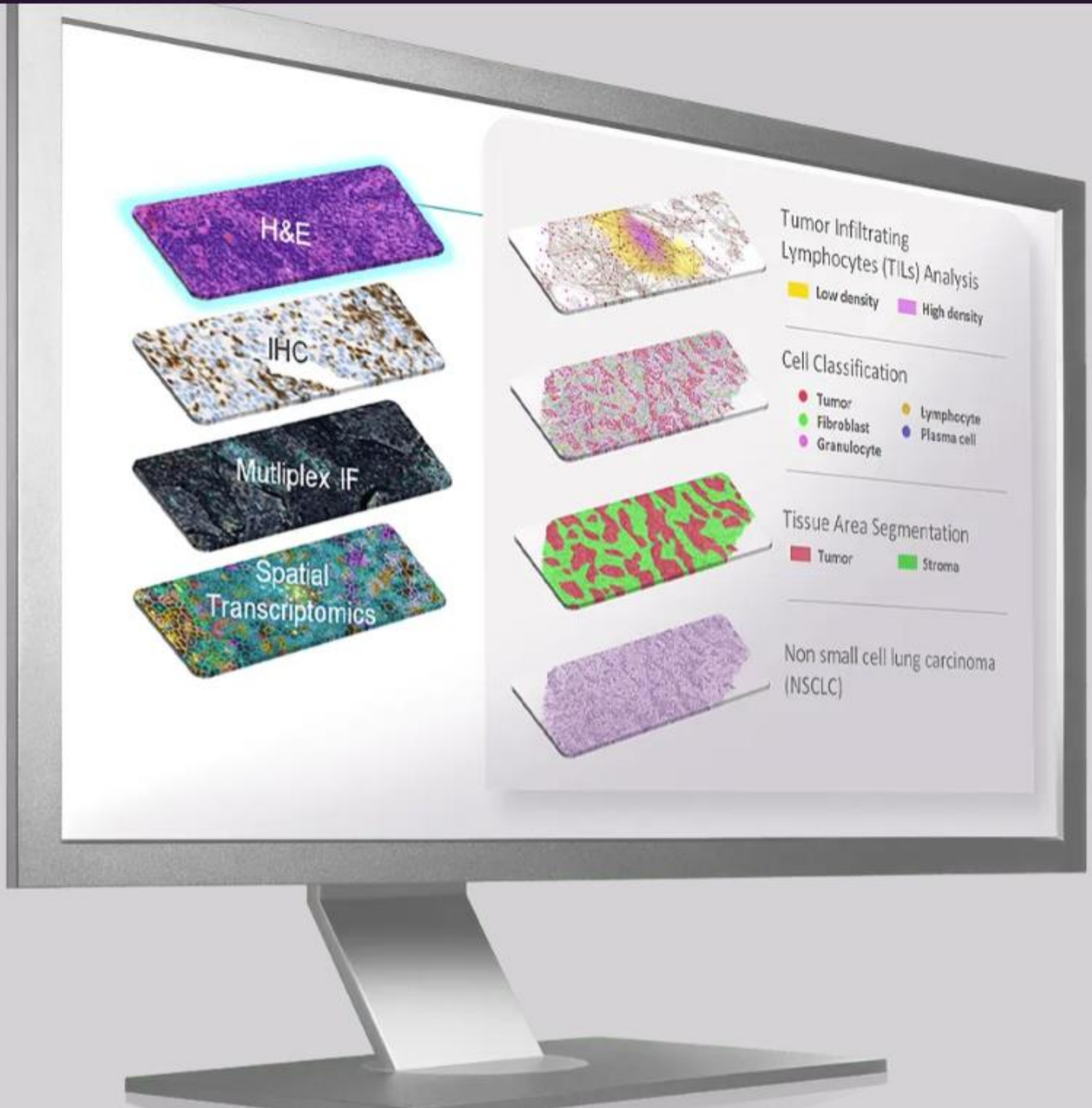
AI-Driven Drug Discovery & Development Tools

Proprietary AI technologies applied to specific DD problems + state of the art wet lab and scientific capabilities

- Clinical Subtyping
- Mechanism Recommendation
- Target Prediction & Assessment
- In silico led HitID
- In silico led LeadOp
- Biomarker Assessment
- Indication and Drug Repurposing

Source: https://www.benevolent.com/investors/reports-and-presentations/

Spatial Biomarker Analysis for Clinical Trials and Diagnostics



Our AI Technology

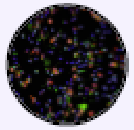
Our platform pioneers AI-driven spatial biomarker analysis, offering a first-of-its-kind multimodal solution. It ingests images from various modalities (H&E, IHC, multiplex immunofluorescence, and spatial transcriptomics), conducts advanced spatial analysis, and delivers actionable insights. These insights optimize biomarker scoring, determine biomarker prevalence and mechanism of action (MoA), and predict response to therapy, revolutionizing our approach to disease diagnosis and treatment.

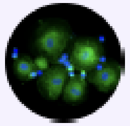
Source: Nucleai Investor Presentation, April 2024

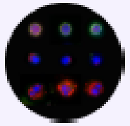
Spring accelerates discovery and development of novel therapeutics with leading AI tools, starting with a single-cell image analysis engine

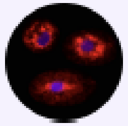
AI-driven engine for fast single-cell analysis

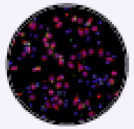
Spring's technology unlocks **hundreds of novel and interpretable signals hiding in biological images**

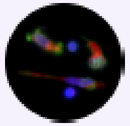

Cell populations


Actin dynamics


Death pathways


Metabolism

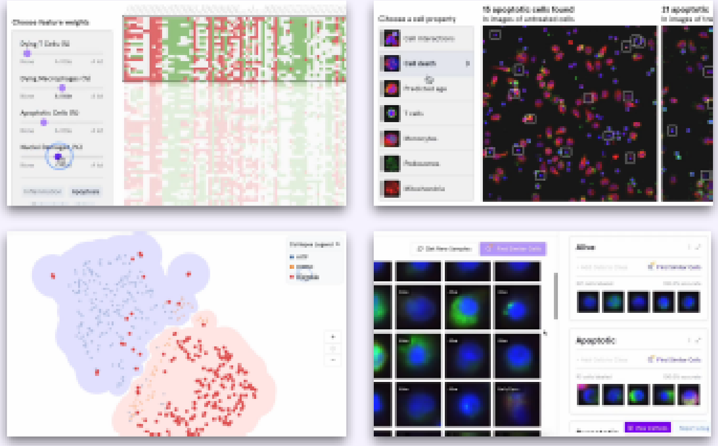

Cell-to-cell interactions


Polarization

Cytokine production

& more...

Spring's engine leverages AI to **identify novel mechanisms, rank compounds, and discover new therapies**



Spring's out of the box platform is easy to use and purpose built to **unblock key R&D milestones while creating efficiencies for computational teams**

> ***"This is an amazing platform for single-cell phenotypes. I haven't seen anything else on the market like it."***
>
> – Head of Imaging, Big Pharma

Source: Spring Investor Presentation, April 2024

Spring's platform has scaled and demonstrated success across multiple areas of focus

>700 terabytes of high-quality, AI-ready imaging data generated using primary human samples

>1.21M wells processed over 6 years of AI-driven discovery programs, including screens of 10,000+ compounds

20+ cell types and 30+ whole tissue types successfully analyzed

Industry-leading performance on key imaging benchmarks such as phenotypic compound similarity and target ID

SPRING

Trusted by scientists around the world...

GILEAD

novo nordisk®

BROAD INSTITUTE

UCSF

BILL & MELINDA GATES foundation

MERCK

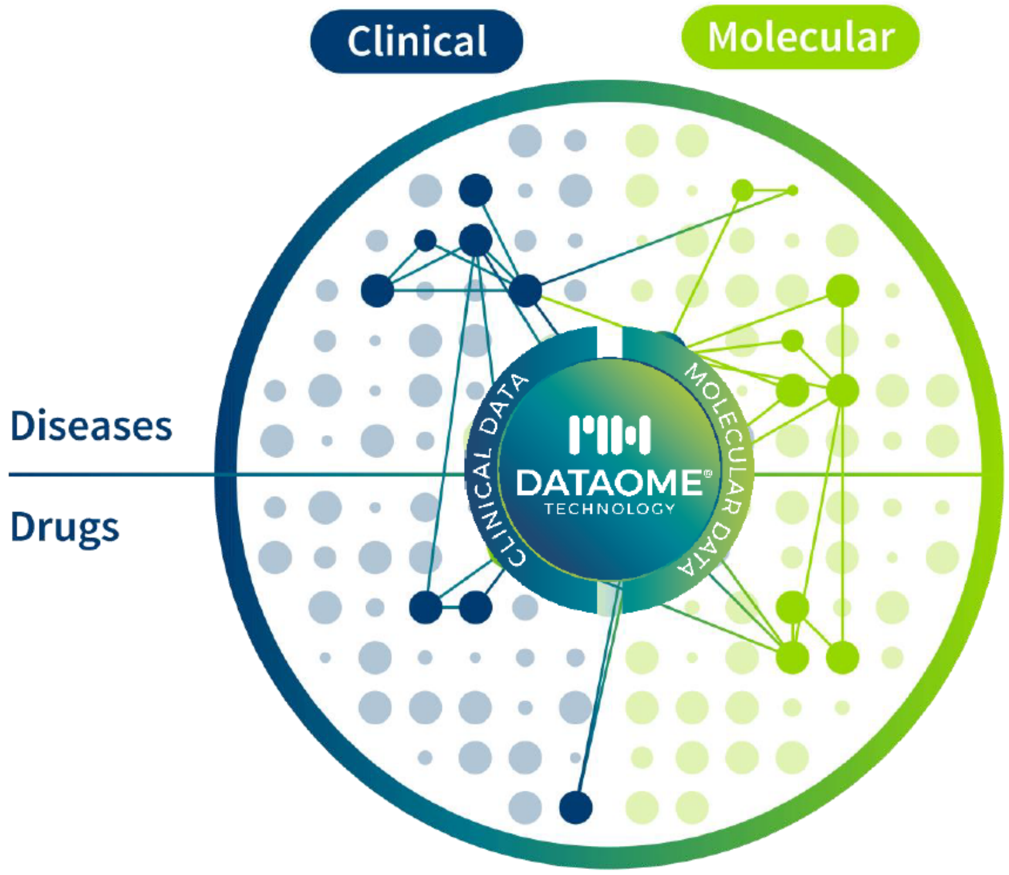
UNIVERSITY OF TORONTO

NVIDIA.

Google

Source: Spring Investor Presentation, April 2024

Dataome™: Proprietary databases and tools allow for clinico-molecular interconnectivity for advanced analytics on drug outcomes



MH has developed one of the largest and strictly curated clinico-molecular repositories in the world: Dataome™

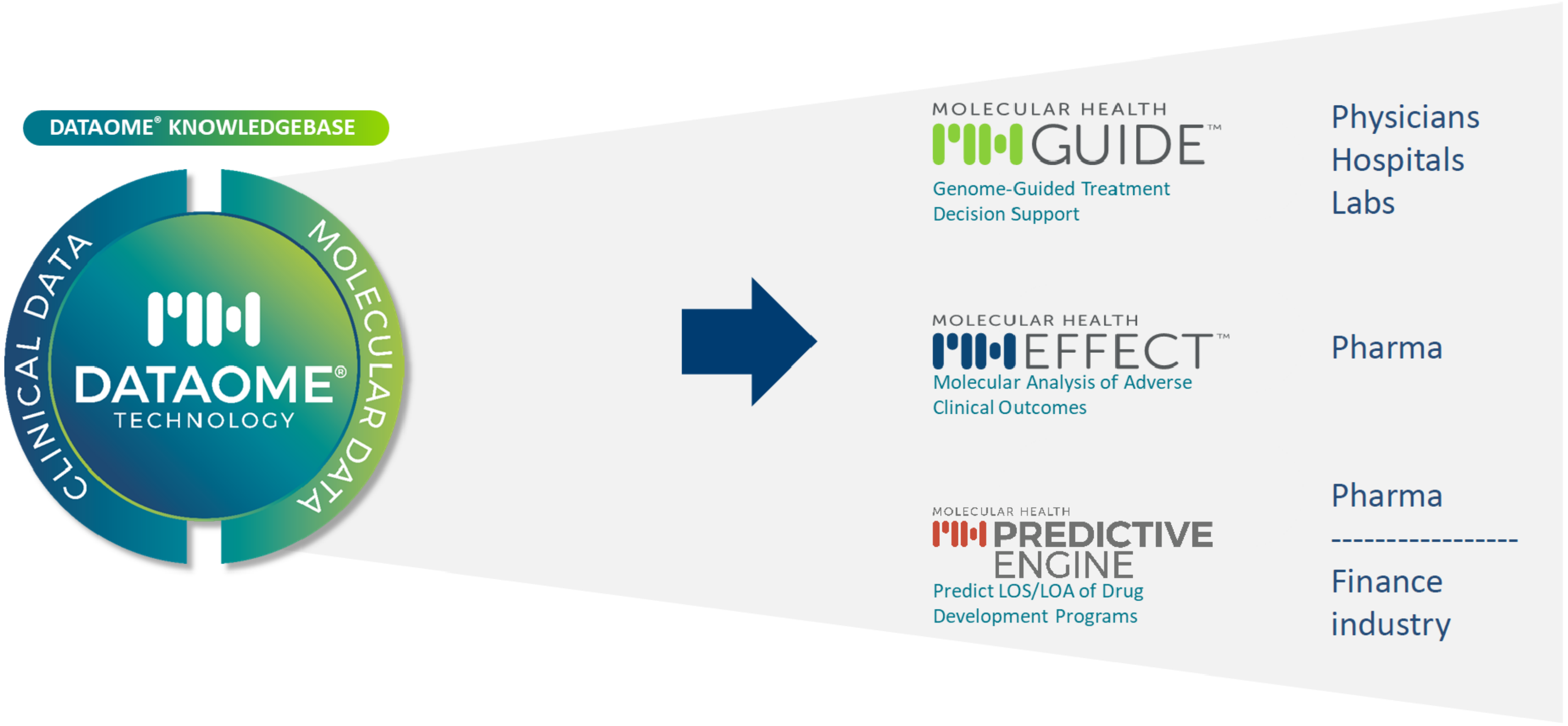
- >100 cleaned, curated and integrated public databases
- >20 proprietary entity dictionaries
- Text-mined medical literature mapped into Dataome™
- Proprietary RW data models and integration workflow
- >10 proprietary databases and interoperable data tools:
  - Drugs
  - Disease genetic variants
  - Biomarkers
  - Drug – Target relations
  - Disease pathways
  - Drug MoA
  - De-biased reference genomes

Source: Molecular Health Corporate Presentation

MOLECULAR HEALTH™

Creates a Set of Tools Based on its Heavily Labelled Datasets

Data integration enables specialized clinico-molecular analytics and AI predictive models to strongly improve drug R&D & treatment



MOLECULAR HEALTH GUIDE™
Genome-Guided Treatment Decision Support

Physicians
Hospitals
Labs

MOLECULAR HEALTH EFFECT™
Molecular Analysis of Adverse Clinical Outcomes

Pharma

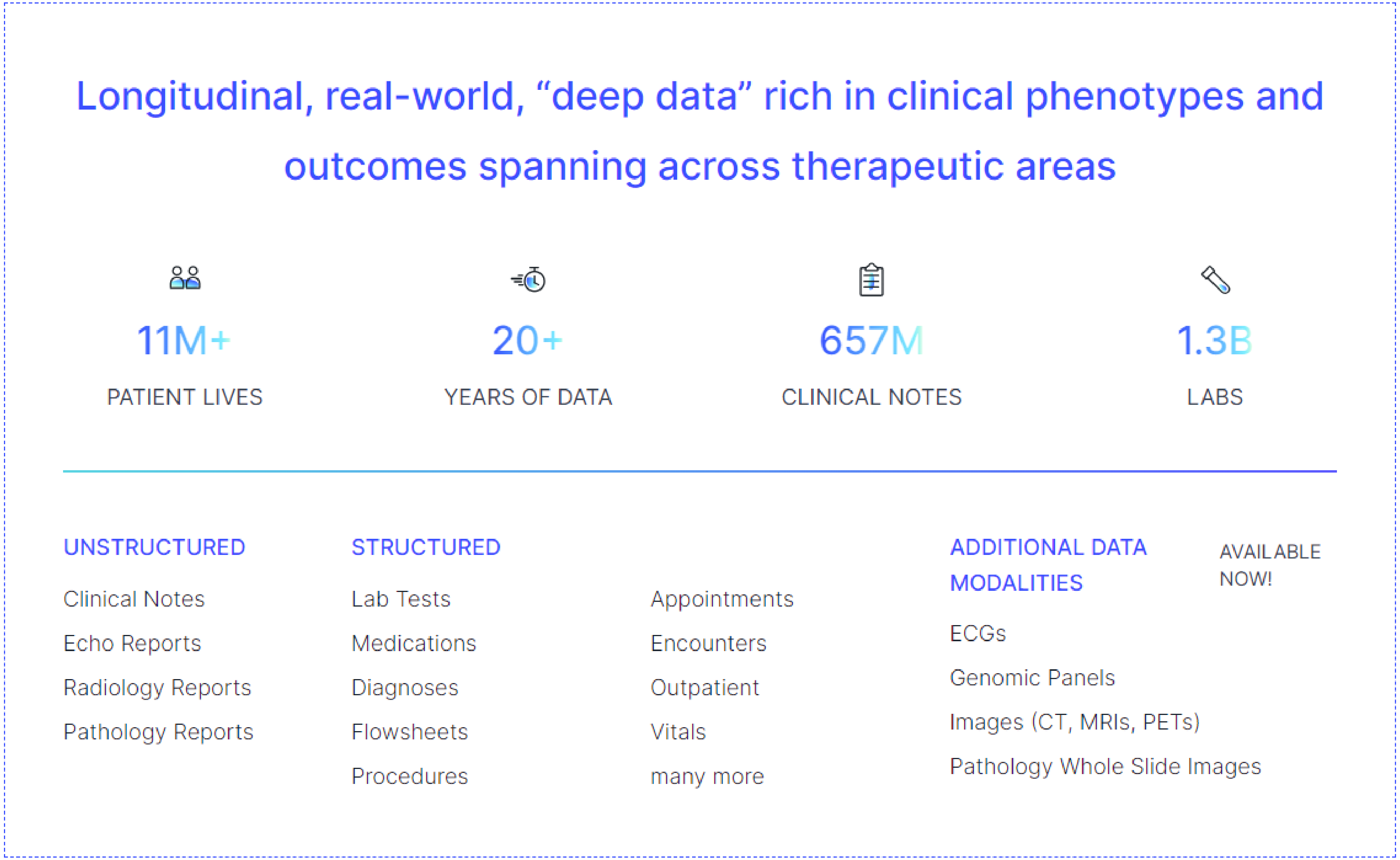
MOLECULAR HEALTH PREDICTIVE ENGINE
Predict LOS/LOA of Drug Development Programs

Pharma

Finance industry

Source: Molecular Health Corporate Presentation

nference® Is a top provider of phenotypic and clinical data for AI



Longitudinal, real-world, “deep data” rich in clinical phenotypes and outcomes spanning across therapeutic areas

11M+	20+	657M	1.3B
PATIENT LIVES	YEARS OF DATA	CLINICAL NOTES	LABS

UNSTRUCTURED

Clinical Notes

Echo Reports

Radiology Reports

Pathology Reports

STRUCTURED

Lab Tests

Medications

Diagnoses

Flowsheets

Procedures

Appointments

Encounters

Outpatient

Vitals

many more

ADDITIONAL DATA MODALITIES

AVAILABLE NOW!

ECGs

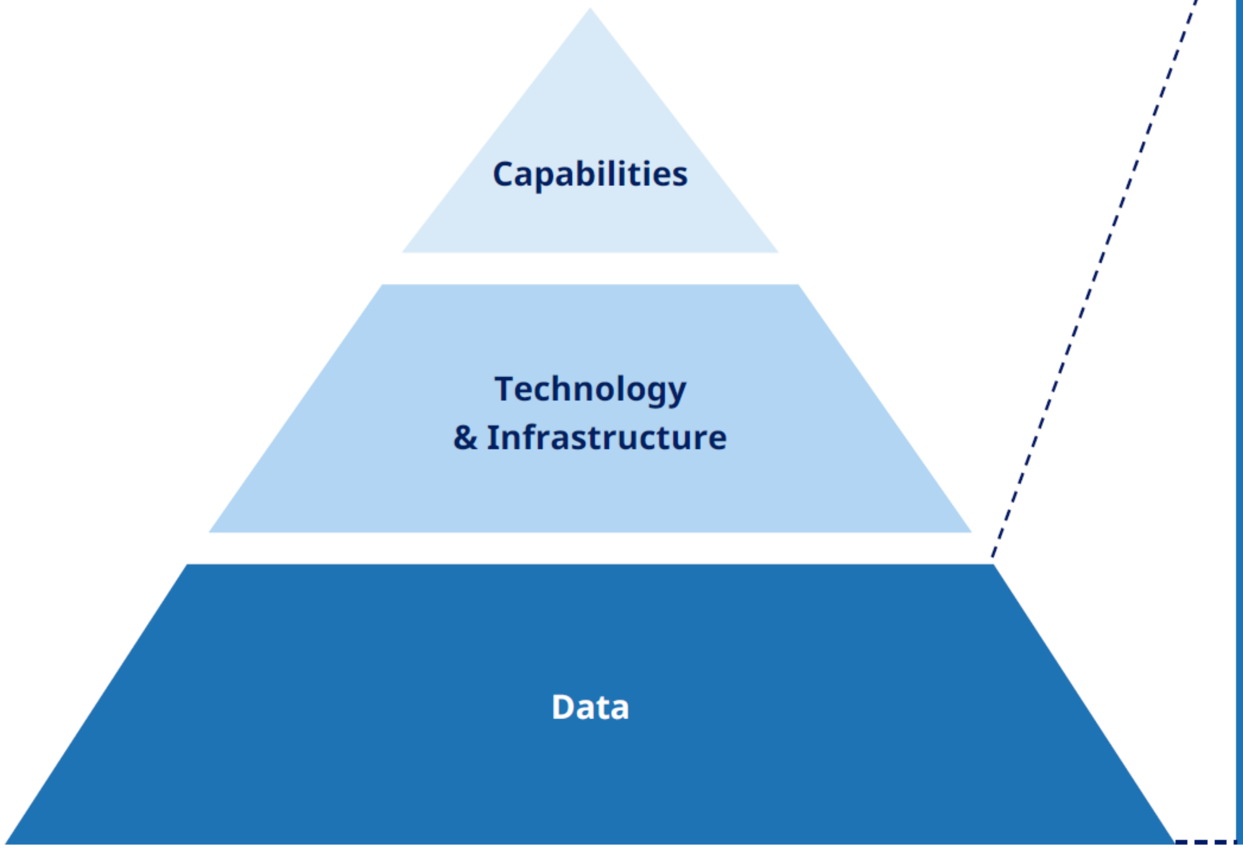
Genomic Panels

Images (CT, MRIs, PETs)

Pathology Whole Slide Images

Source: https://nference.com/

Novo Nordisk utilises best-in-class clinical data within the cardiometabolic space



Capabilities

Technology
& Infrastructure

Data

Deploying data to enable AI

Harmonising data from ~1,600 clinical trials

SELECT and STEP results providing best-in-class cardiometabolic data for new disease insights, patient stratification and drug targets

Expanding data modalities for better drug understanding including omics, imaging and wearable data

Combining real-world data with clinical study data

Automating wet-lab processes for more research data with higher quality

Source: Novo Nordisk Capital Markets Day, March 8, 2024 (https://www.novonordisk.com/investors/capital-markets-day-2024.html)

FounData makes billions of clinical data points accessible across the entire value chain

Unleashing the power of data for secondary use

FounData — **A seamless infrastructure** where all data from completed clinical trials are pooled and prepared for insights-generation

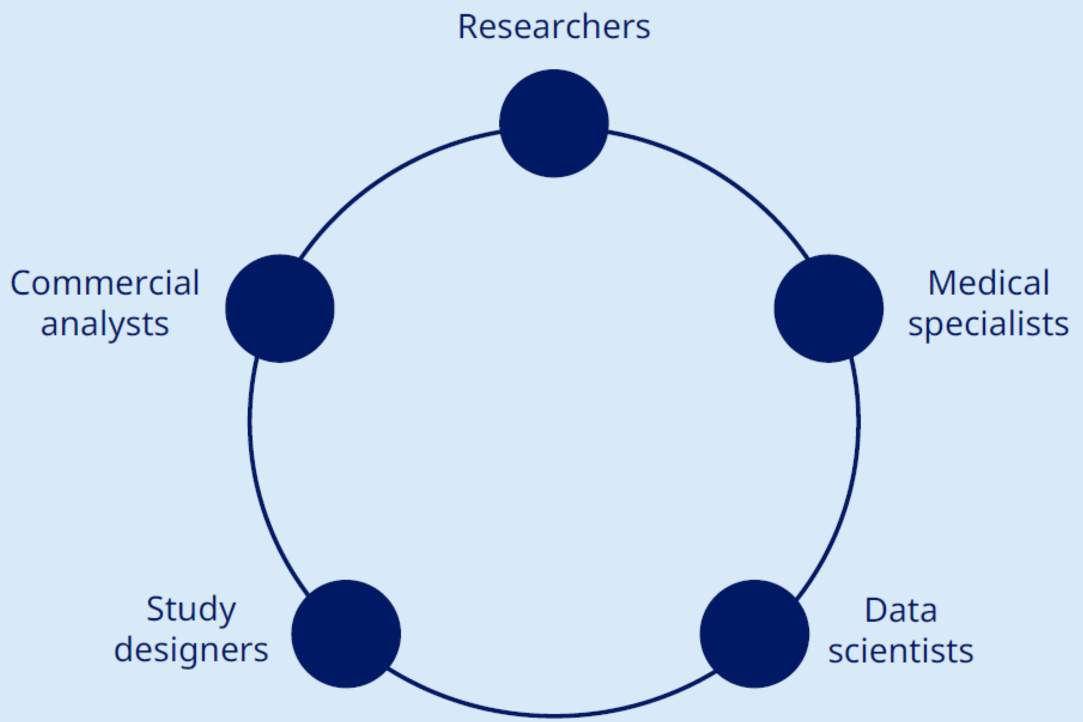
Easy to Find
available data from clinical trials including real-world data, omics and imaging

Easy to access
clinical trial data in a seamless yet controlled way

Easy to connect
multiple data types and sources across the organisation

Easy to solve
for advanced clinical data insights

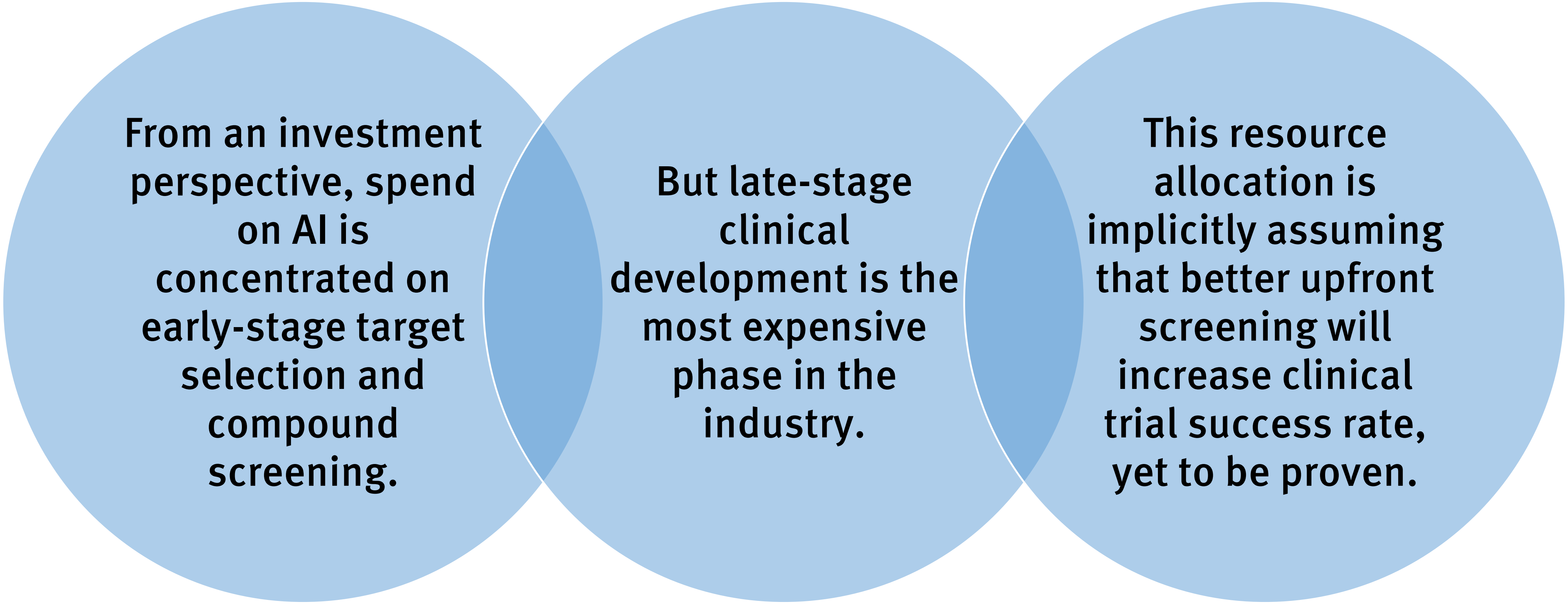
Democratise data across the Novo Nordisk value chain



Source: Novo Nordisk Capital Markets Day, March 8, 2024 (https://www.novonordisk.com/investors/capital-markets-day-2024.html)

AI in Clinical Stage Drug Development

Are AI Investment Dollars Allocated Properly?



From an investment perspective, spend on AI is concentrated on early-stage target selection and compound screening.

But late-stage clinical development is the most expensive phase in the industry.

This resource allocation is implicitly assuming that better upfront screening will increase clinical trial success rate, yet to be proven.

Most costs and time are lost in the last stages of clinical trials (Phase II, Phase III).

	Target Validation	Compound screening	Lead optimization	Pre-clinical	Phase I	Phase II	Phase III	Approval to launch
% Cost per molecule	~3%	~6%	~17%	~7%	~15%	~21%	~26%	~5%
Probability of success (per stage)	80%	75%	85%	69%	54%	34%	70%	91%
Probability of reaching FDA approval (per stage)	4%	5%	7%	8%	12%	22%	64%	91%
Endpoints	Disease models, Target identification, Target validation	Visual screening HTS	SAR, Drug-like properties, Solubility, Permeability, ADME, Plasm PK, Efficacy, Toxicity		PK, Dose escalation, Toxicity	Dose, Efficacy, Toxicity		Market fit, Geographical relevance
Cycle time without AI	5-7 years				5-7 years			
Cycle time with AI	3-5 years				5-7 years			

Source: Acta Pharmaceutica Sinica B, 2022

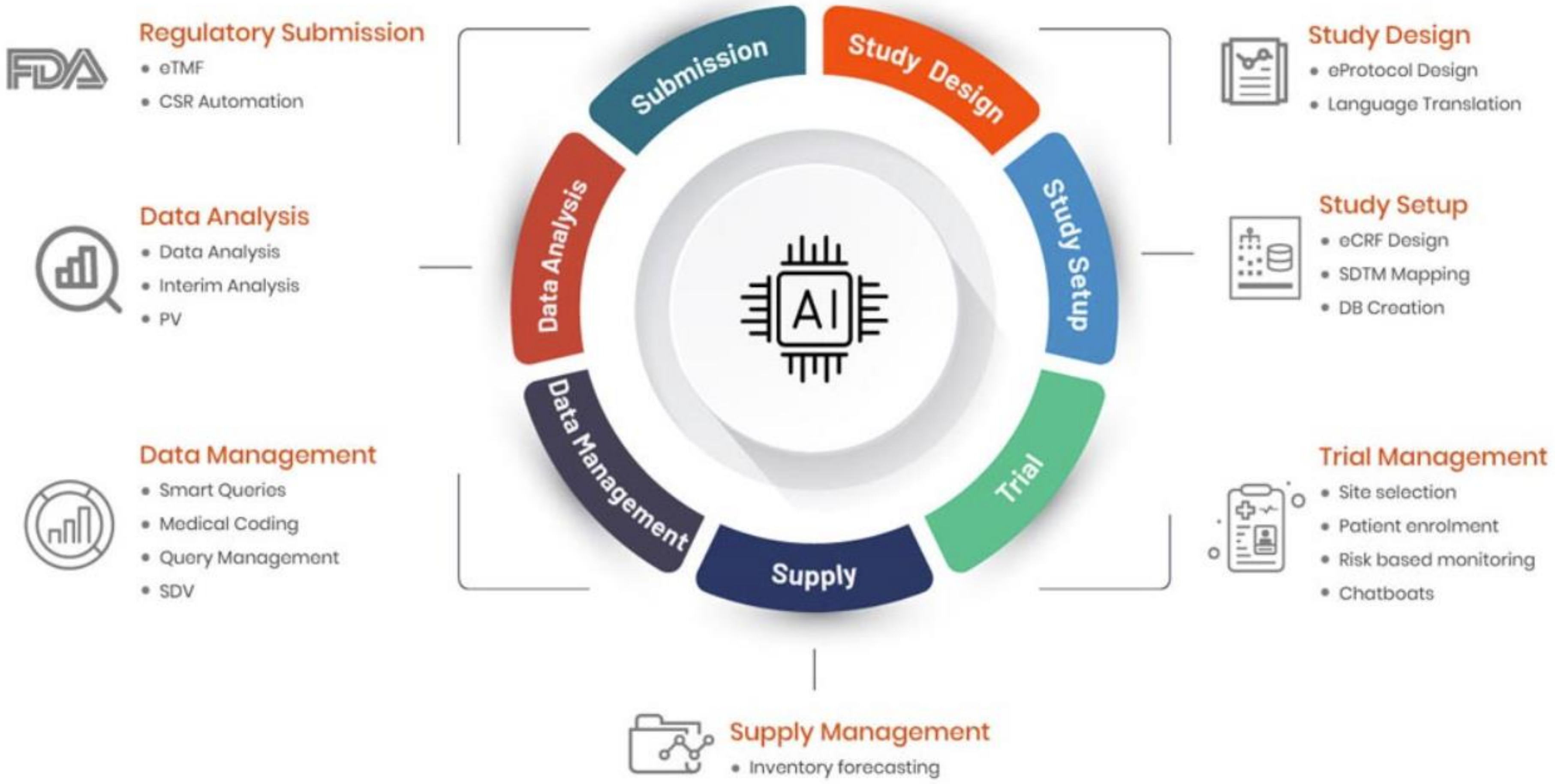
Source: https://dealroom.co/reports/techbio-x-drug-discovery

Key Point

AI today is mainly focused on discovery of drugs and less time is being spent on the real causes of high drug development costs: clinical trials. If we can’t lower clinical trial costs, AI in drug discovery will not have the profound change hoped for in the pharma industry.

Numerous Ways to Use AI to Improve Clinical Trial Efficiency

AI & ML in Clinical Trials



Regulatory Submission
- eTMF
- CSR Automation

Study Design
- eProtocol Design
- Language Translation

Data Analysis
- Data Analysis
- Interim Analysis
- PV

Study Setup
- eCRF Design
- SDTM Mapping
- DB Creation

Data Management
- Smart Queries
- Medical Coding
- Query Management
- SDV

Trial Management
- Site selection
- Patient enrolment
- Risk based monitoring
- Chatboats

Supply Management
- Inventory forecasting

Source: https://www.clinion.com/insight/ai-and-automation-in-clinical-trials/

Ecosystem of AI Companies Focused on Improving Clinical Trials Cost, Speed and Success Rates

Finding Patients / Selecting Patients / Optimizing Site Selection / Optimizing Protocols / Accelerating Enrollment / Medical Logic / EDC

AMGEN BAYER ConcertAI DEEP 6 AI IQVIA™ faro

MENDEL NOVARTIS nucleai Paradigm Reify HEALTH

saama Science 37® teckro THREAD® VIAL

Synthetic Control Arms / Digital Twins / Real World Data

ALTIS CytoReason IQVIA™

phesi Patient-Centric Data Science saama TEMPUS

TriNetX TRUVETA UNLEARN

Source: Stifel Research

Trial Software / Patient Analytics / Outcomes / Data Analysis

AiCure® altoida Andaman7

BIOPHY Medable VIAL

MOLECULAR HEALTH™ Veeva

Paradigm Launches to Get More Patients into Clinical Research, Accelerate Drug Trials

Paradigm

Heather Landi, FierceHealthcare, Jan 23, 2023 (excerpt)

A new tech startup wants to disrupt the clinical trial process and landed $203 million to help scale up its technology.

Paradigm, a clinical trials data and patient-matching platform, aims to open up access to research to boost patient recruitment and speed up drug development. The company was conceived by Arch Venture Partners and co-incubated with General Catalyst. Last year, Paradigm also acquired Deep Lens, a clinical trial patient recruitment tech platform focused on oncology, for an undisclosed sum.

Paradigm CEO Kent Thoelke has a long career in life sciences, having spent more than 25 years overseeing drug development and clinical trials at healthcare companies. He saw firsthand the inefficiencies in the current clinical trial process.

"The clinical trials process takes years, and, what was frustrating especially, was that my dad died of a brain tumor and I have multiple family members who have cancer. If a lot of those drugs could come to market years sooner, then think about all those patients' lives that could have been saved," Thoelke said in an interview. He recognized the need to "change the paradigm" of clinical research, but there was inertia in the industry to make any significant changes. Then COVID-19 hit, and life sciences companies quickly pivoted to using technology for remote clinical trials.

"But they were all kind of temporary measures, and they weren't really deployed at scale," he noted. "He said, 'What if we just change the whole model? Let's just blow up the model that exists and just reimagine it,'" he said.

The idea behind Paradigm is to build a technology-enabled, scaled clinical research ecosystem to tackle inefficiencies and enable more patients to have access to clinical research to drive down the timelines and the cost. The Paradigm platform was designed to reduce the operational burden on physicians and healthcare provider organizations and improve access for patients.

There are significant barriers to patient participation in clinical trials, including lack of access to trials in community settings, restrictive eligibility criteria, burdensome trial protocols and lack of financial support. Healthcare providers, particularly community clinics, are under-resourced to support clinical research at scale and lack an adequate portfolio of therapeutic trials to serve their populations, which exacerbates enrollment disparities.

Source: https://www.reuters.com/technology/big-pharma-bets-ai-speed-up-clinical-trials-2023-09-22/

Amgen’s ATOMIC Model

Natalie Grover and Martin Coulter, Reuters, Sep 22, 2023

Before AI, Amgen would spend months sending surveys to doctors from Johannesburg to Texas to ask whether a clinic or hospital had patients with relevant clinical and demographic characteristics to participate in a trial.

Existing relationships with facilities or doctors would often sway the decision on which trial sites are selected.

However, Deloitte estimates about 80% of studies miss their recruitment targets because clinics and hospitals overestimate the number of available patients, there are high dropout rates or patients don't adhere to trial protocols.

Amgen's AI tool, ATOMIC, scans troves of internal and public data to identify and rank clinics and doctors based on past performance in recruiting patients for trials.

Enrolling patients for a mid-stage trial could take up to 18 months, depending on the disease, but ATOMIC can cut that in half in the best-case scenario, Amgen told Reuters.

Amgen has used ATOMIC in a handful of trials testing drugs for conditions including cardiovascular disease and cancer, and aims to use it for most studies by 2024.

The company said by 2030, it expects AI will have helped it shave two years off the decade or more it typically takes to develop a drug.

Source: https://www.reuters.com/technology/big-pharma-bets-ai-speed-up-clinical-trials-2023-09-22/

Atomic: Follow the Data

Amgen article on ATOMIC, October 12, 2022

Fifteen percent of clinical trial sites never enroll a single participant. And the sites that do find participants often take years to complete enrollment. Almost half the time spent on bringing a drug through clinical trials is during the enrollment phase. This causes serious delays in getting potential new drugs to patients who need them now.

Amgen's solution? Follow the data.

A highly collaborative, cross-functional group of data scientists, engineers and analysts at Amgen have teamed up to create a tool that can sort through and analyze vast amounts of data to find potential clinical trial sites that are most likely to complete a study more quickly through faster enrollment of patients. The project is called the Analytical Trial Optimization Module, better known as ATOMIC. It leverages machine learning (ML) to pull hundreds or even thousands of pieces of data related to clinical trial sites and determines which data is most relevant to high-enrolling sites. It then puts out a ranked list of sites, predicted enrollment rates at those sites and other relevant data about the country, site and related investigators.

"We wanted to be able to design clinical trials to be faster and have a higher likelihood of success," said Matt Austin, executive director of Data Sciences in the Center for Design & Analysis (CfDA). "ATOMIC allows us to make predictions in an informed way and build a ranked list of sites that may perform best."

Amgen's clinical trial teams have been able to sort through parts of this information successfully, but site selection has traditionally been a manual, time-consuming, inconsistent process that isn't always able to follow all the data available. Often it is based on the trial team member's experience with a site or investigator, combined with user-directed computer research. There is no centralized data source and no consistent method for analyzing that data.

"There are only so many parameters that the human mind can calculate," said Scott Skellenger, vice president of Information Systems. "And often what happens is you're selecting sites based on human relationships, past experience and anecdotal input," he added. "ATOMIC represents a new class of technology that's involved in using both leading-edge data and predictive modeling to be able to make choices that are more difficult for people to make on their own."

Source: https://www.amgen.com/stories/2022/10/follow-the-data

ATOMIC Story (continued)

Improving, accelerating and diversifying trials

ATOMIC can help identify key drivers of enrollment, leading to the prediction of the most successful sites. By integrating so much disparate data, ATOMIC also provides a single, comprehensive, automated information source for clinical trial team members.

“We are truly at a historic moment, where we have a confluence of unprecedented access to data and advanced analytics,” said Rob Lenz, senior vice president of Global Development. All this data calls for new analytic approaches to predict its value and improve decision-making, said Lenz.

The ATOMIC project team has been working with Amgen’s Global Development Operations (GDO), Center for Design & Analysis (CfDA), Center for Observational Research (CfOR) and Representation in Clinical Research (RISE) teams on ways to increase the diversity and improve the representation of clinical trial investigator and participant populations. They have pulled anonymized and/or aggregated data such as racial and ethnic demographics and geography about clinical sites from Amgen and other sponsors’ trials and aggregated it into a scorecard. The data are then incorporated into ATOMIC’s pipeline for use by any team looking to diversify or increase representation in a trial. This may support the enrollment of a patient population that is more closely representative of the real-world population typically afflicted by the disease being studied, including race, ethnicity, sex and age. It is also important more broadly across Amgen to help improve patient access to our medicines.

Many kinds of data can be fed into ATOMIC to help inform site selection, such as anonymized real-world data (RWD), which includes vast amounts of data derived from electronic health records (EHRs) and medical claims. These data can be leveraged to better understand where populations with certain clinical and demographic characteristics seek care. This information can be included with other data types in the ATOMIC platform. For example, the ATOMIC team incorporated RWD that included diversity data on providers and their patients in the RISE diversity scorecard.

By examining RWD, the ATOMIC project team and CfOR were also able to identify trial sites likely to be found near clusters of patients with high Lipoprotein(a), or Lp(a) levels. Lp(a) levels are associated with risk of cardiovascular disease and levels differ between some populations. Using these data can help site investigators, who are testing an investigational Lp(a)-lowering drug, screen 50% fewer patients to find one participant with elevated Lp(a) for the trial.

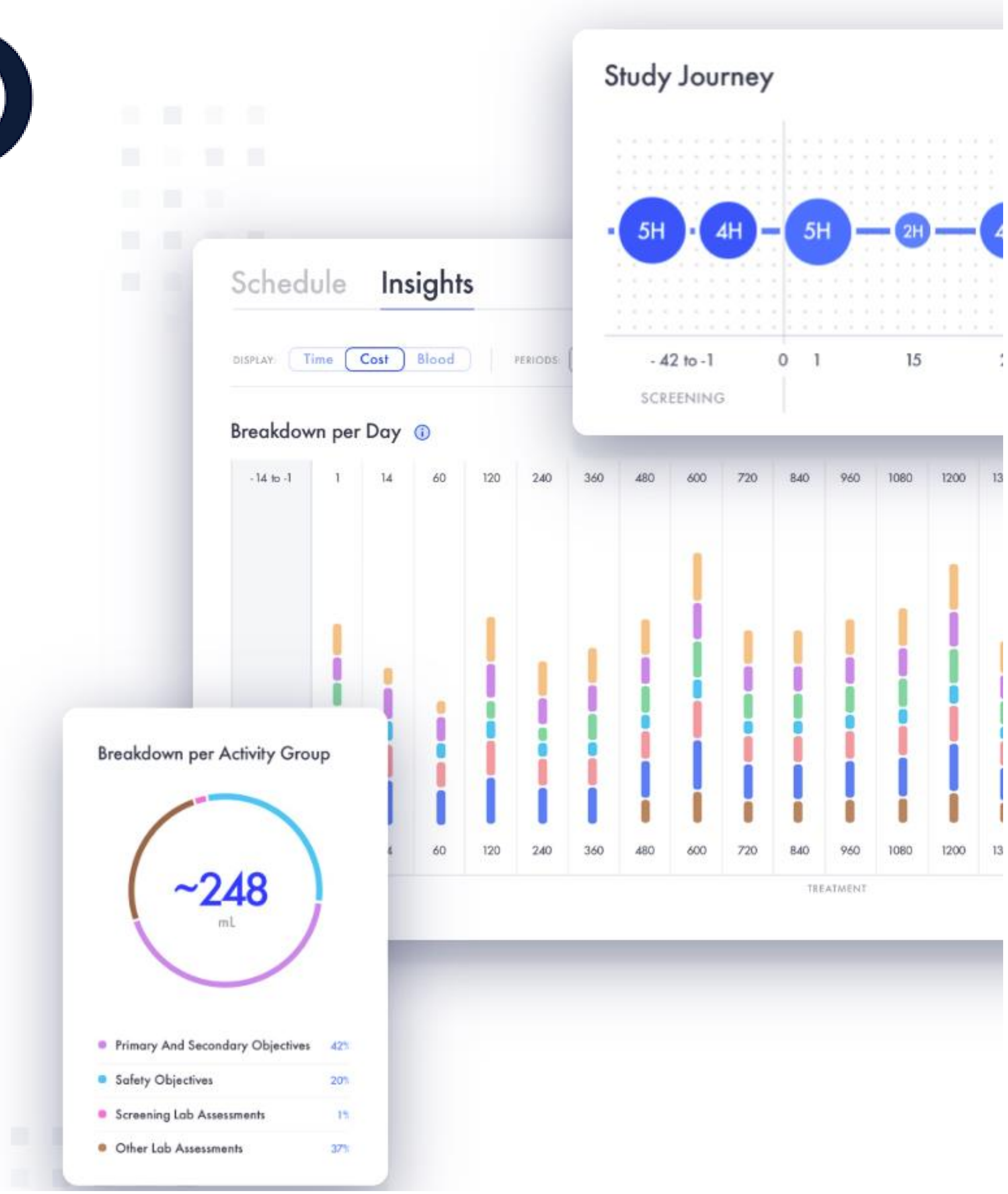
Source: https://www.amgen.com/stories/2022/10/follow-the-data

faro

Clinical Development for the Digital Age

Faro is a software platform that orchestrates complex clinical development with a single source of truth. It brings words, data and teams together on one flexible surface empowering teams to focus on centricity while balancing the complexity of protocol design.

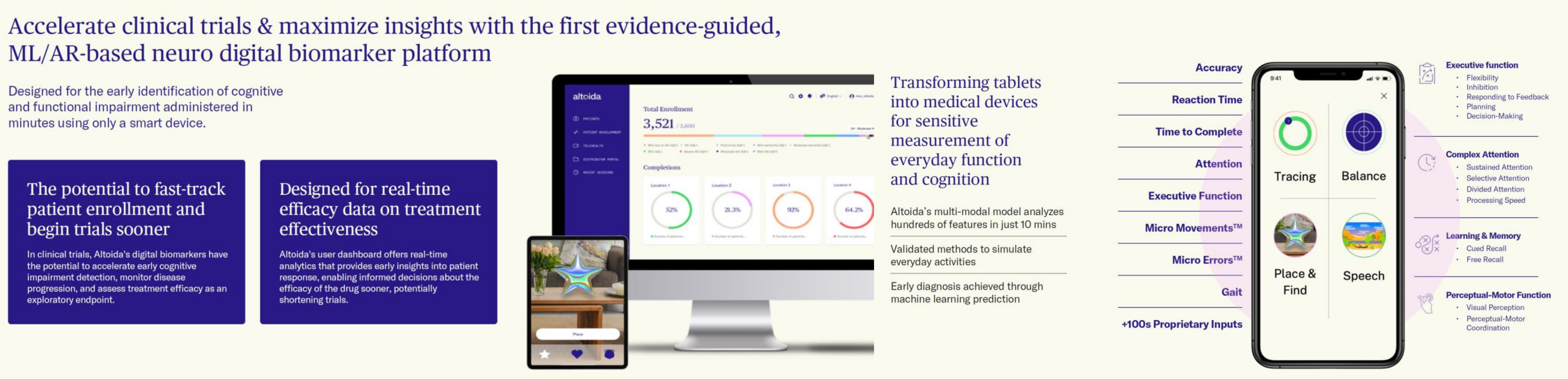
Faro Health does digital protocol design to help pharma understand in real time the implications of study design decisions which leads to studies that can enroll faster and cost less to execute.



● Primary And Secondary Objectives	42%
● Safety Objectives	20%
● Screening Lab Assessments	1%
● Other Lab Assessments	37%

Source: https://farohealth.com/

Altoida Enables Adoption of Digital Biomarkers in Clinical Trials



Accelerate clinical trials & maximize insights with the first evidence-guided, ML/AR-based neuro digital biomarker platform

Designed for the early identification of cognitive and functional impairment administered in minutes using only a smart device.

The potential to fast-track patient enrollment and begin trials sooner

In clinical trials, Altoida's digital biomarkers have the potential to accelerate early cognitive impairment detection, monitor disease progression, and assess treatment efficacy as an exploratory endpoint.

Designed for real-time efficacy data on treatment effectiveness

Altoida's user dashboard offers real-time analytics that provides early insights into patient response, enabling informed decisions about the efficacy of the drug sooner, potentially shortening trials.

Transforming tablets into medical devices for sensitive measurement of everyday function and cognition

Altoida's multi-modal model analyzes hundreds of features in just 10 mins

Validated methods to simulate everyday activities

Early diagnosis achieved through machine learning prediction

Accuracy

Reaction Time

Time to Complete

Attention

Executive Function

Micro Movements™

Micro Errors™

Gait

+100s Proprietary Inputs

Executive function
- Flexibility
- Inhibition
- Responding to Feedback
- Planning
- Decision-Making

Complex Attention
- Sustained Attention
- Selective Attention
- Divided Attention
- Processing Speed

Learning & Memory
- Cued Recall
- Free Recall

Perceptual-Motor Function
- Visual Perception
- Perceptual-Motor Coordination

Altoida is developing first-of-kind, digital neurological biomarkers designed to assess cognition in clinical trials. The platform will involve a tablet-based application that uses augmented reality (AR) to replicate activities of daily living (iADLs) measuring cognitive and motoric functions via multimodal digital biomarkers.

The motoric and AR tasks in the Altoida assessment are designed to extract multimodal features, such as micro-movements, micro-errors, speed, reaction times, and navigation trajectories, which are used to train specific machine-learning models, which are known as Digital Neuro Signatures (DNS).

Source: https://altoida.com/our-science/

Investing in AI Biology Companies

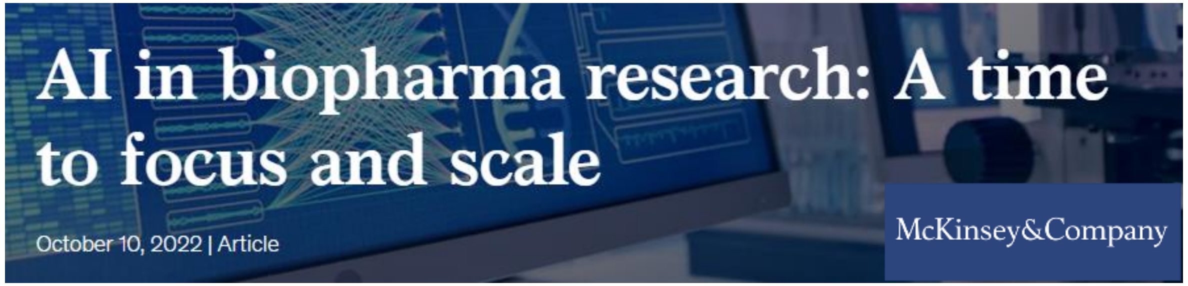
AI in Healthcare is a Huge Market Opportunity

Press Release, June 15, 2023

“Bridge Market Research analyses that the artificial intelligence in healthcare market, which is $9.6 billion in 2022, is expected to reach $272.9 billion by 2030, at a CAGR of 51.9% during the forecast period 2023 to 2030.”

Source: https://www.digitaljournal.com/pr/news/cdn-newswire/artificial-intelligence-in-healthcare-market-estimated-to-witness-exponential-growth-of-usd-272-91-billion-by-2030. Photo from Getty Images.

Over 250 Companies Working on AI and Drug Discovery

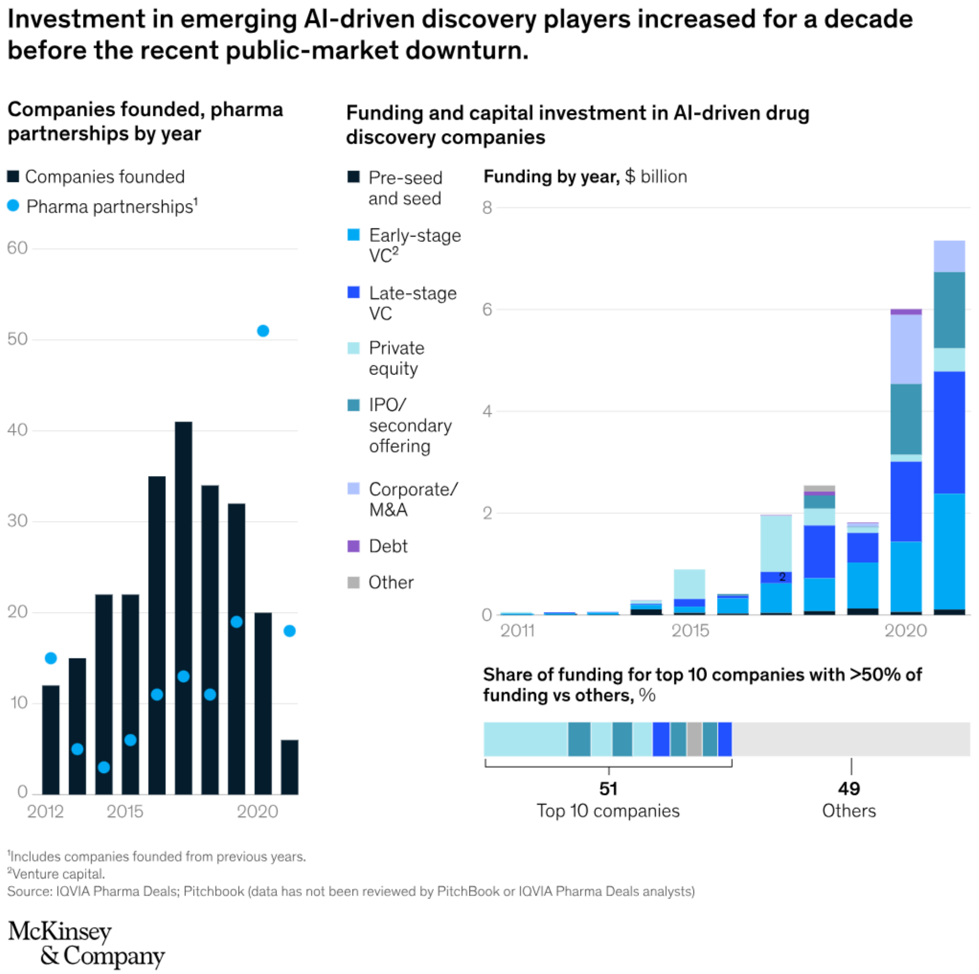


AI in biopharma research: A time to focus and scale

October 10, 2022 | Article

McKinsey&Company

Our research has identified nearly **270 companies** working in the AI-driven drug discovery industry, with more than 50 percent of the companies based in the United States, though key hubs are emerging in Western Europe and Southeast Asia. The number of AI-driven companies with their own pipeline is still relatively small today (approximately 15 percent have an asset in preclinical development). Those with new molecular entities (NMEs) in clinical development (Phase I and II) have predominantly in-licensed assets or have developed assets using traditional techniques.



Investment in emerging AI-driven discovery players increased for a decade before the recent public-market downturn.

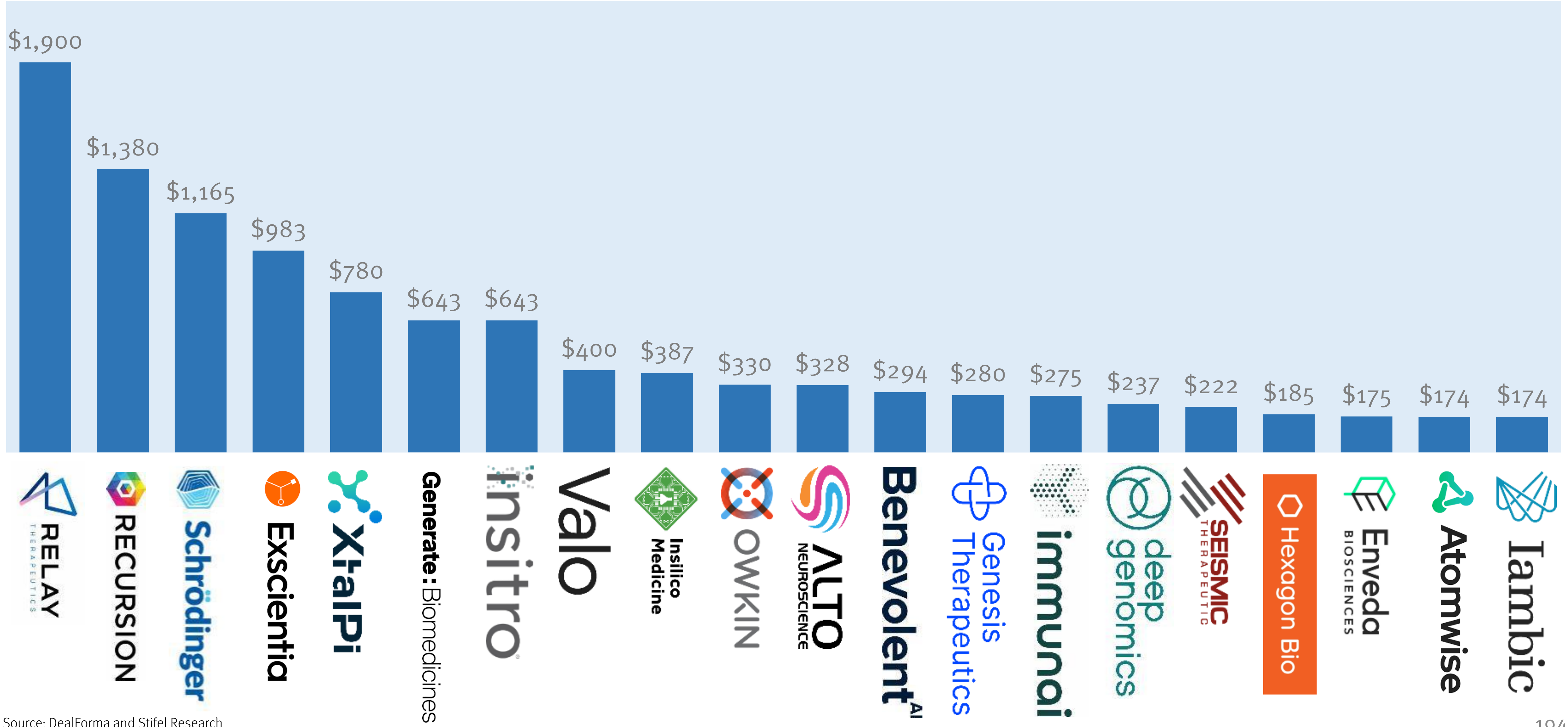
[1] Includes companies founded from previous years.
[2] Venture capital.
Source: IQVIA Pharma Deals; Pitchbook (data has not been reviewed by PitchBook or IQVIA Pharma Deals analysts)

McKinsey & Company

Source: https://www.mckinsey.com/industries/life-sciences/our-insights/ai-in-biopharma-research-a-time-to-focus-and-scale

Top 20 AI-Centric Biotechs by Capital Raised to Date

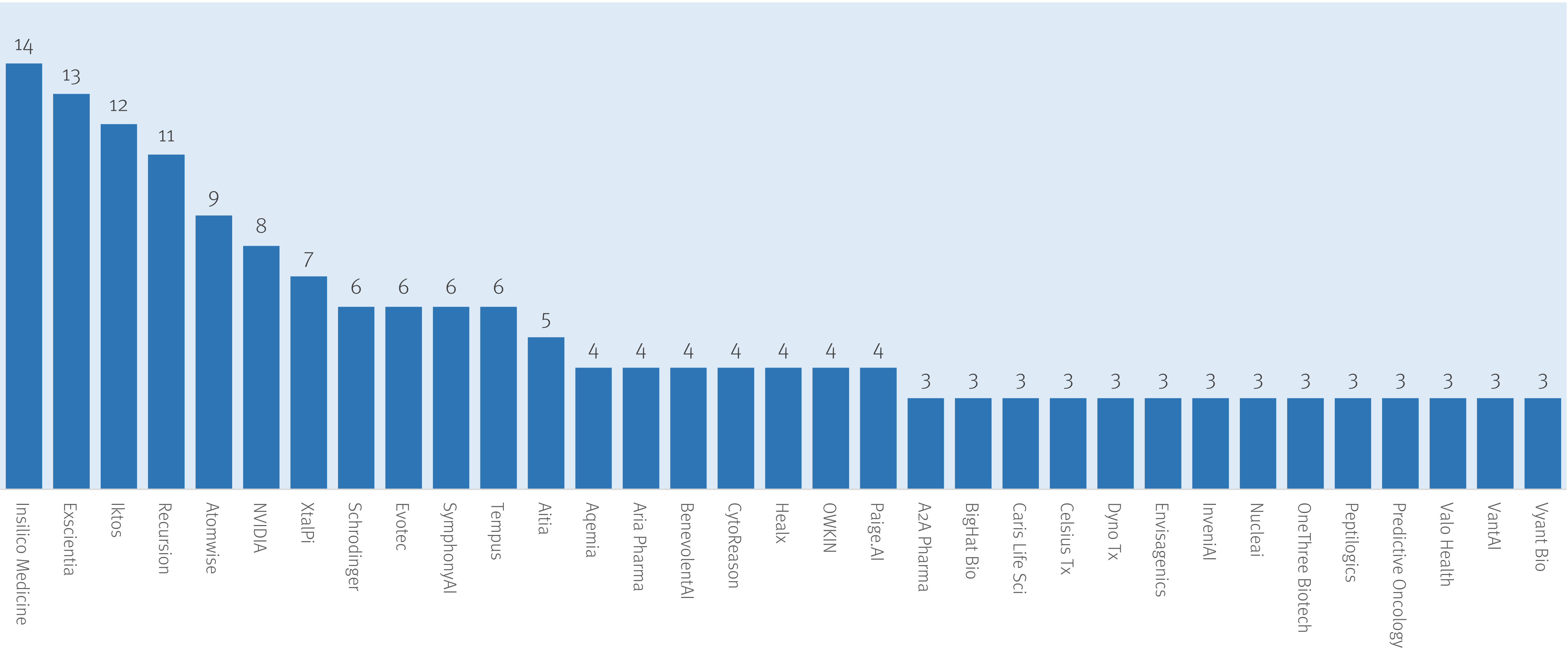
Total Capital Raised by AI-Centric Biotech Companies ($ millions), 2010 to 2024



Source: DealForma and Stifel Research

Insilico, Exscientia, Iktos and Recursion Lead in Number of Platform Collaborations and License Deals in Last 15 Years

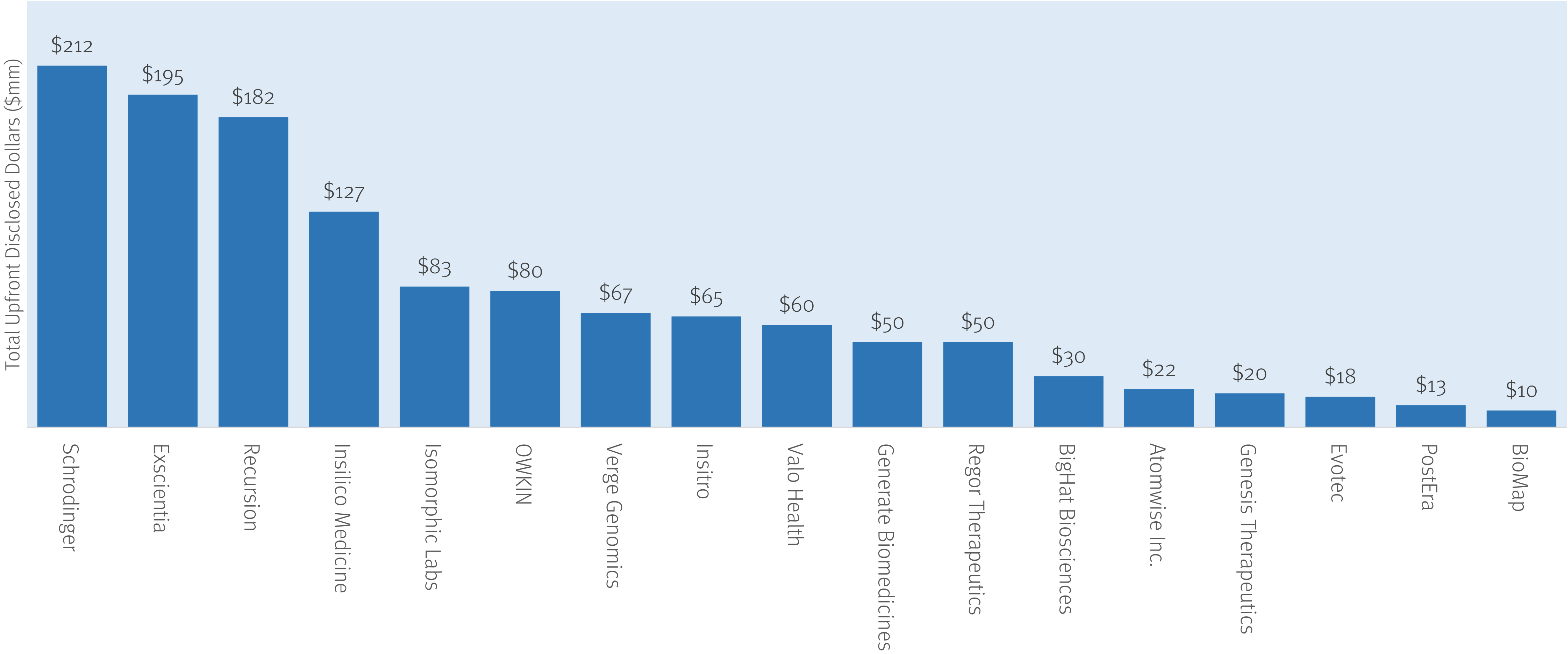
Most Active AI-Centric Biotechs and Data Players by Licensing and Collaboration Deal Count, 2010 to 2024



Source: DealForma

Schrodinger, Exscientia, Recursion, Insilico, Isomorphic and OWKIN Lead in Dollars from Platform and License Deals in Last 15 Years

Most Active AI Biotechs by Dollars In from Licensing Deals and Platform Collaborations, 2010 to 2024 ($mm)



Source: DealForma and Stifel Research.
Note: Schrodinger dollars in includes their share of proceeds from Nimbus / Takeda deal and Nimbus / Gilead deal.

Roche R&D Phenomap Neuroscience Collaboration Agreement with Recursion the Largest AI Licensing Deal so Far ($150mm Upfront)

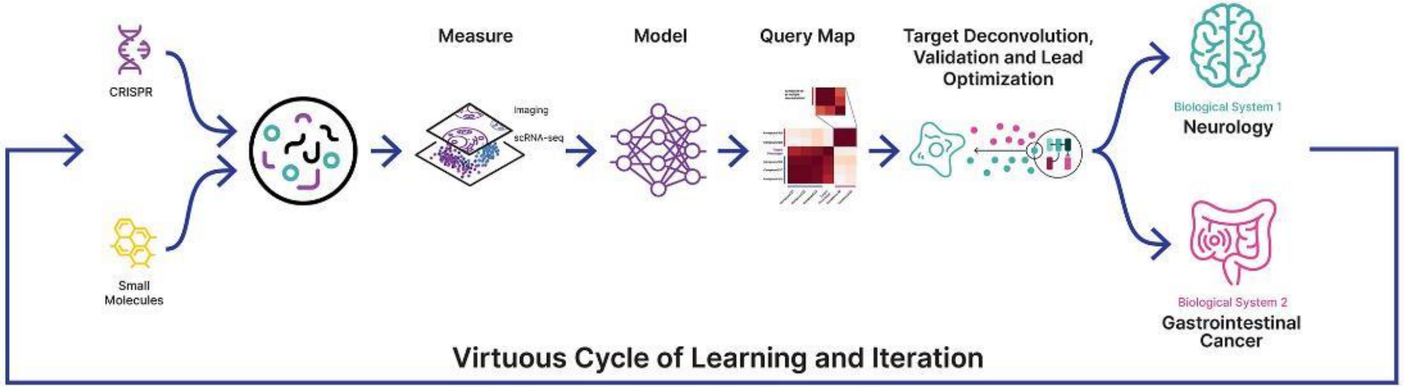
Recursion (2024 10K): On December 5, 2021, Recursion entered into a Collaboration and License Agreement with Genentech and Roche pursuant to which we will construct, using our imaging technology and proprietary machine learning algorithms, unique maps of the inferred relationships amongst perturbation phenotypes in a given cellular context. We will together create multimodal models and maps to further expand and refine such inferred relationships, in both cases with the goal to discover and develop therapeutic small molecule programs in a gastrointestinal cancer indication and in key areas of neuroscience. In January 2022, Roche paid us an upfront cash payment of $150.0 million.

Phenomap Creation, Acceptance and Access. Under the Collaboration Agreement, we are responsible for creating a certain number of phenomaps in each of the Exclusive Fields. We will also provide Roche with limited access to our pre-existing human umbilical vein endothelial cells (HUVEC) phenomap. Roche will have specified rights to query or access the phenomaps to generate novel inferences that may lead to the discovery or development of therapeutic products.

Phenomap-Related Options. Each of the phenomaps requested by Roche and created by Recursion may be subject to either an initiation fee, acceptance fee or both. Such fees could exceed $250.0 million for sixteen (16) accepted phenomaps. In addition, for a period of time after Roche's acceptance of certain phenomaps, Roche will have the option to obtain, subject to payment of an exercise fee, rights to use outside the collaboration the raw images generated in the course of creating those phenomaps. If Roche exercises its External Use Option for all twelve (12) eligible phenomaps, Roche's associated exercise fee payments to Recursion could exceed $250.0 million.

Collaboration Programs and Roche Options. Roche and Recursion will collaborate to select certain novel inferences with respect to small molecules or targets generated from the phenomaps for further validation and optimization as collaboration programs. Roche and Recursion may also combine sequencing datasets from Roche with Recursion's phenomaps and collaborate to generate new algorithms to produce multimodal maps from which additional collaboration programs may be initiated. For every collaboration program that successfully identifies potential therapeutic small molecules or validates a target, Roche will have an option to obtain an exclusive license to develop and commercialize such potential therapeutic small molecules or to exploit such target in the applicable Exclusive Field. In October 2023, Roche exercised its Small Molecule Validated Hit Option to further advance our first partnership program in GI-oncology.

Under Recursion's collaboration with Roche & Genentech, Recursion is creating multimodal maps of cellular biology in neurology and gastric cancer to elucidate novel targets and starting points.



Source: https://www.sec.gov/Archives/edgar/data/1601830/000160183024000019/rxrx-20231231.htm

M&A History in the Biotech AI Field (2010 to 2024)

Incumbents in the AI field have been the main source of acquisitions to date in AI/ML. There have been zero big pharma M&A deals to date involving AI/ML biotechs. The largest two deals were BioNTech / Instadeep and Ginkgo / Zymergen. Valo Health has been the most active acquirer with four transactions to date. Ginkgo and Recursion have all done two acquisitions thus far.

Announced	Seller	Buyer	Upfront Value ($mm)	Total Deal Value ($mm)
01/10/2023	InstaDeep	BioNTech	$551	$734
07/25/2022	Zymergen	Ginkgo Bioworks	$300	$300
04/16/2021	ZebiAI Therapeutics	Relay Therapeutics	$85	$170
05/20/2019	Just Biotherapeutics	Evotec	$60	$90
06/15/2021	Allcyte GmbH	Exscientia	$59	$59
05/08/2023	Valence Discovery	Recursion Pharmaceuticals	$48	$48
05/08/2023	Cyclica	Recursion Pharmaceuticals	$40	$40
08/19/2019	PointR Data	Mateon Therapeutics	$15	$165
06/06/2022	Bitome	Ginkgo Bioworks Inc.	NA	NA
07/27/2021	Courier	Valo Health	NA	NA
10/18/2017	DeepCrystal Technologies	Valo Health	NA	NA
03/19/2020	EnEvolv	Zymergen	NA	NA
11/15/2019	Numerate	Valo	NA	NA
09/24/2020	Forma Therapeutics AI Assets	Valo	NA	$30
10/22/2020	Haystack Sciences	Insitro Inc.	NA	NA
02/28/2024	Patch Biosciences Inc.	Ginkgo Bioworks Inc.	NA	NA
06/14/2021	Totient	AbSci Corporation	NA	NA

Source: DealForma and Stifel Research.

BioNTech's $551 Million Acquisition of Instadeep the Largest M&A Deal in Biotech AI So Far

MAINZ, Germany, and LONDON, United Kingdom, January 09, 2023:

BioNTech and InstaDeep Ltd. today announced that they have entered into an agreement under which BioNTech will acquire InstaDeep, a leading global technology company in the field of artificial intelligence ("AI") and machine learning ("ML"). The transaction includes a total upfront consideration of approximately £362 million in cash and BioNTech shares to acquire 100% of the remaining InstaDeep shares, excluding the shares already owned by BioNTech. In addition, InstaDeep shareholders will be eligible to receive additional performance-based future milestone payments up to approximately £200 million. The transaction follows BioNTech's initial equity investment as part of InstaDeep's Series B financing round in January 2022.

The acquisition supports BioNTech's strategy to build world-leading capabilities in AI-driven drug discovery and development of next-generation immunotherapies and vaccines to address diseases with high unmet medical need. The transaction will combine two organizations with a common culture and is expected to add approximately 240 highly skilled professionals to BioNTech's workforce, including teams in AI, ML, bioengineering, data science, and software development. Through the acquisition, BioNTech will grow its network of global research collaborators in the field and expand its footprint in key talent hubs across the United States, Europe, Africa and the Middle East.

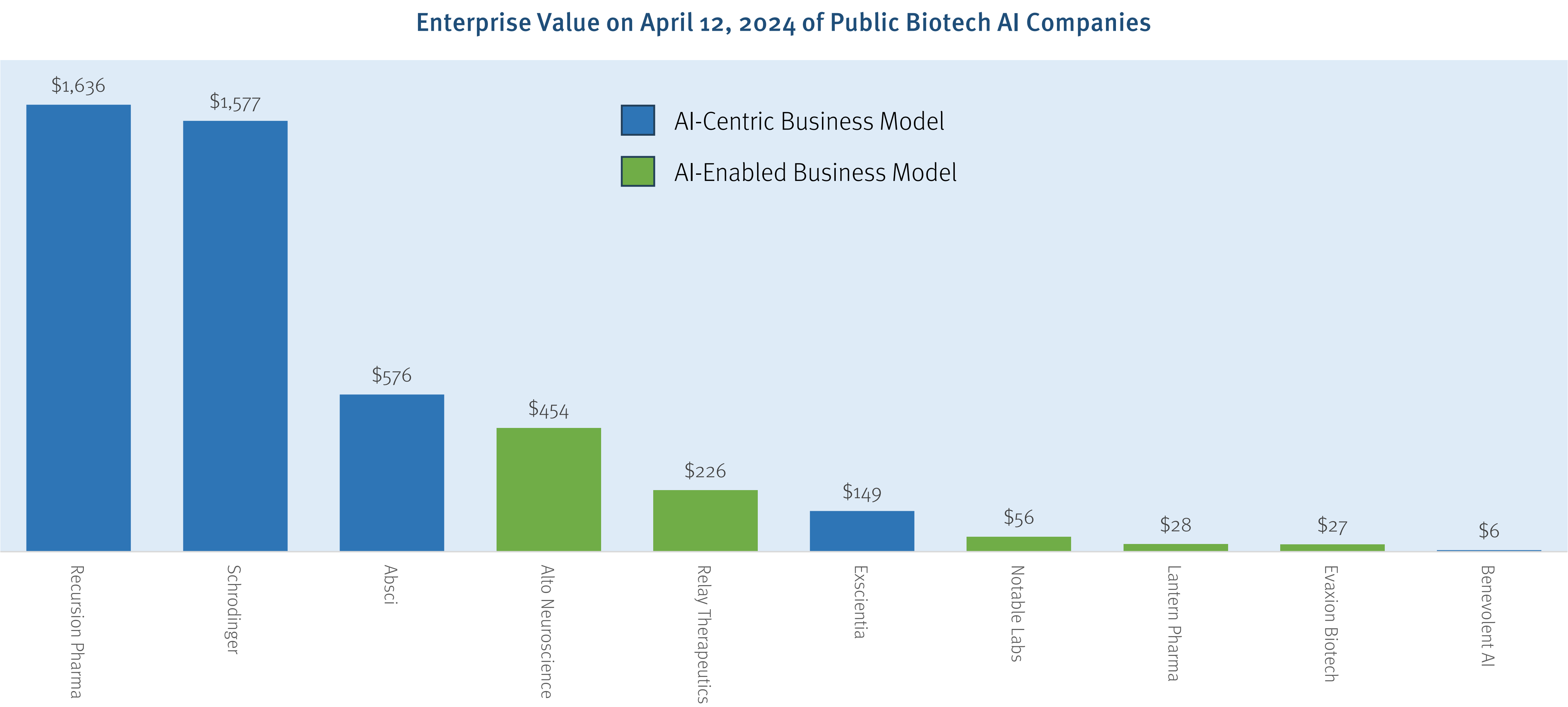
With the acquisition of InstaDeep, validated and novel BioNTech-trained AI- and ML-based models are planned to be embedded across BioNTech's discovery platforms and connected, through InstaDeep's DeepChain™ platform, to an integrated automated lab infrastructure. This has the objective to enable high-throughput design and testing of novel drug candidates at scale. In addition, BioNTech plans to develop novel AI and ML technology solutions which it aims to apply across key strategic and operational functions.

The acquisition builds on a successful track record of increasing collaboration between the two companies since 2019: In November 2020, the companies announced a multi-year strategic collaboration and joint AI Innovation Lab to apply the latest advances in AI and ML technology to develop novel medicines for a range of cancers and infectious diseases. The companies have jointly developed multiple end-to-end AI-based applications trained on public and proprietary datasets across a wide variety of scientific domains. These include projects to enhance neoantigen selection, ribological sequence optimization for BioNTech's RiboCytokine® and RiboMab® platforms as well as the development of an Early Warning System to detect and monitor high risk SARS-CoV-2 variants, based on their ability to escape immune defenses and transmissibility potential, defined as fitness, which was announced in January 2022.

"Since the inception of BioNTech, we have focused on leveraging computational solutions to create personalized immunotherapies that can reach a wide patient population," said Prof. Ugur Sahin, M.D., CEO and Co-founder of BioNTech. "The acquisition of InstaDeep allows us to incorporate the rapidly evolving AI capabilities of the digital world into our technologies, research, drug discovery, manufacturing, and deployment processes. Our aim is to make BioNTech a technology company where AI is seamlessly integrated into all aspects of our work."

Source: https://www.instadeep.com/2023/01/biontech-to-acquire-instadeep-to-strengthen-pioneering-position-in-the-field-of-ai-powered-drug-discovery-design-and-development/

Only Two Public AI Companies Are Valued Over $1 Billion

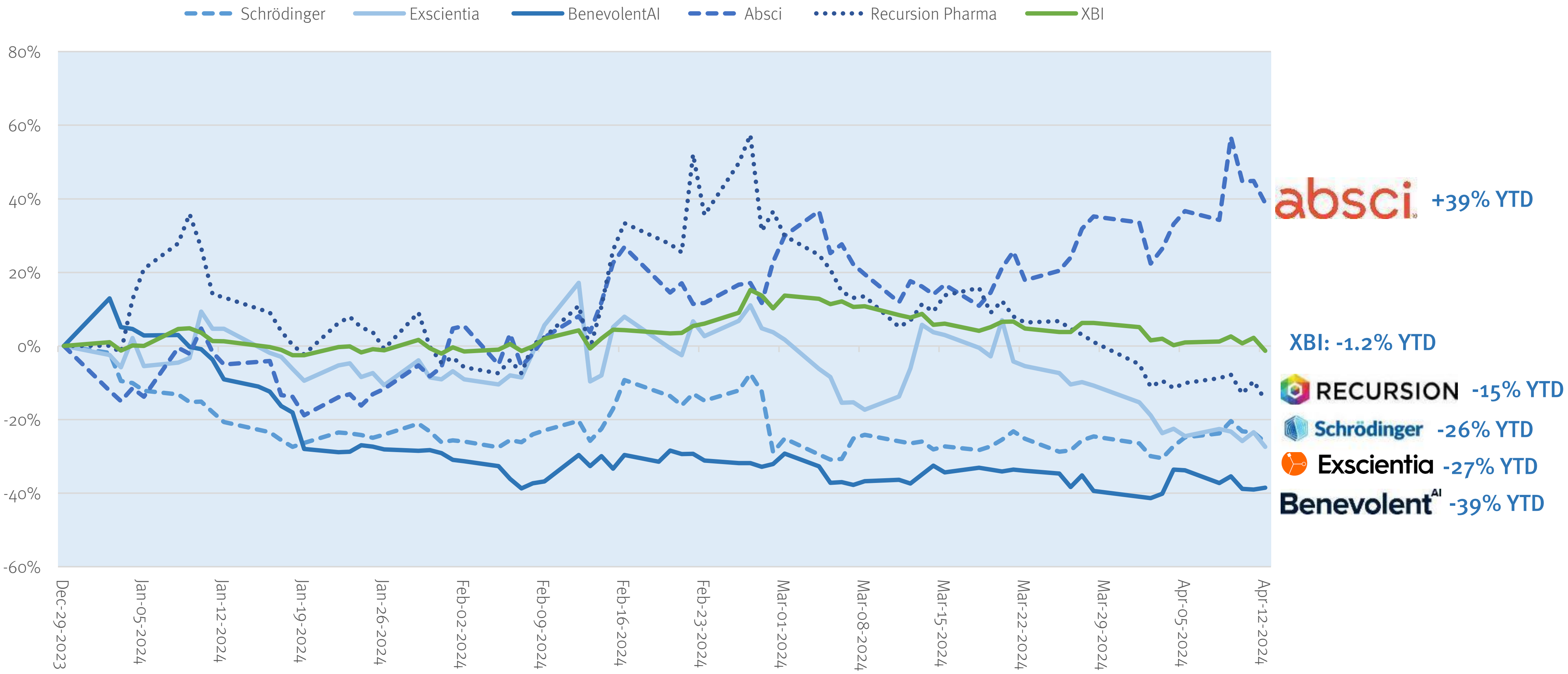


Source: S&P CapitalIQ and Stifel Research

AI-Centric AI Biotech Stocks Largely Down in 2024

We are in a drug and data centric stock market for biotech. AI companies without data or a strong story in immunology or obesity are not performing well this year. The long exception is Absci which has built up its own pipeline well in the field of immunology.

AI Biotech Stock Performance vs. the XBI, Dec 30, 2023 to April 12, 2024



Source: CapitalIQ

Advice on Biotech AI Investments from Digitalis Ventures

- A key point that cannot be overemphasized: We lack consistent and quality data for both training and validation across most applications in drug discovery. While large corpora of publicly available unstructured text and images have driven the recent explosion of LLMs and generative AI, the development of foundation models in chemistry and biology will require the intentional curation of bespoke hiqh-quality, well-annotated, and robust datasets.
- We are beginning to see the same debates from the broader tech world – particularly about the merits of open-source models – echoed in the life sciences. We expect proprietary datasets paired with new data management and computational tools to drive advantage even as publicly available tools proliferate.
- None of this is cheap. The costs associated with the GPUs necessary to train, run, and refine AI models are becoming quite significant, and the cost of data generation itself can be unsustainable without new infrastructure. As AI adoption increases, these costs should be considered alongside other standard elements of drug development.
- And in the end, the data matters only insofar as it can be leveraged – with reasonable investment, on a reasonable timescale – to bring safe and effective drugs to patients. It's worth noting that many of the most exciting biotechs we see today are led by seasoned medicinal chemists or protein engineers who are perhaps understated in their use of cutting-edge computational advances, but nonetheless use AI technology as one of many discovery tools to maximize clinical impact. That is the metric by which all these efforts should ultimately be measured.

Comments at left from . . .

Sam Bjork
Partner
Digitalis Ventures

Travis Hughes
Senior Associate
Digitalis Ventures

Web: www.digitalisventures.com

Advice to Founders and Investors on AI/ML from Elena Viboch of General Catalyst

- Focus on the problem you're trying to solve and avoid developing a solution looking for a problem or getting too attached to a single approach. The best small molecule drug developers use every tool needed to make medicines. They'll leverage fragment-based screens, chemoproteomics and DNA-encoded libraries to find novel chemical matter.
- Be interdisciplinary! We look for teams who bring together expertise across multiple verticals, are curious, great communicators and fearlessly tackling hard problems.
- We see companies making incredible progress by generating wet lab data, and then training their AI to guide iterative cycles of that wet lab work. They make more progress than is possible using solely wet lab or solely computational approaches.
- Use your compute to go after hard problems. We want to support founders leveraging technology to make meaningful medicines. If it's something that can be done in a wet lab quickly or cheaply, then keep using the wet-lab based approach. Find a way to apply your AI to make a transformational solution.

Elena Viboch
Partner
General Catalyst
(https://generalcatalyst.com)

On Biotech AI Investments from Tal Zaks of Orbimed

- Some AI companies are really good in certain parts of the value chain. Like Charm Therapeutics can do biologics very well. I like a focused story like that where I know it is going to work.
- So many companies come and pitch on how many petabytes of data they have. Funny – then you look at their pipeline and the TPP doesn't make sense.
- Being a chemist doesn't make you a drug developer. There is so much more too it. You need to know biology. You need to know how you are going to develop your drug in the clinic. All of this feeds back into how well you are going to design drugs for patients in need. We fail in our industry in clinical development. I would love to see more pitches that link how they are designing drugs to how they will be able to get through clinical development efficiently.
- It's really hard to invest in AI-enabled drug discovery: I get pitched the goose that will lay the golden egg and is priced accordingly; but since the goose is AI, which translates into "black box by design", I can only judge that value once I actually see a golden egg. And I haven't seen any yet…

Tal Zaks
Partner
OrbiMed
(https://orbimed.com)

Does Investing in AI/ML Stories in Biotech Make Sense at All?

The data in this section is not so encouraging. AI biotech stocks have not done well lately at all. If one looks back three or four years, the picture is no better. It's no surprise we have been in a market where there have not been AI/ML IPO's in biotech. There are so many good private companies that could go public. Think Insitro, Valo or XtalPi. These three companies are highly substantive with end-to-end platforms and strong investor support, but we haven't seen any IPO's yet from them.

What it's going to take is obvious from the comments from VC's. Investors want to see "golden eggs" – not the "golden goose". Companies won't do well in AI unless and until they can convince investors that they can deliver interesting drugs to the market.

This is a "Show Me" market. Not a "Wow Me" market.

So, does it make sense to invest in AI/ML stories in biotech at all?

Obviously, one must be highly selective about deploying capital in this area. Many stories sound good in the pitch but don't quite work out in the execution.

But there are good reasons to consider investing in AI and biotech. AI is an important tool that can knock investor's socks off and stocks up.

(continued)

The Case *for* Investing in Biotech and AI

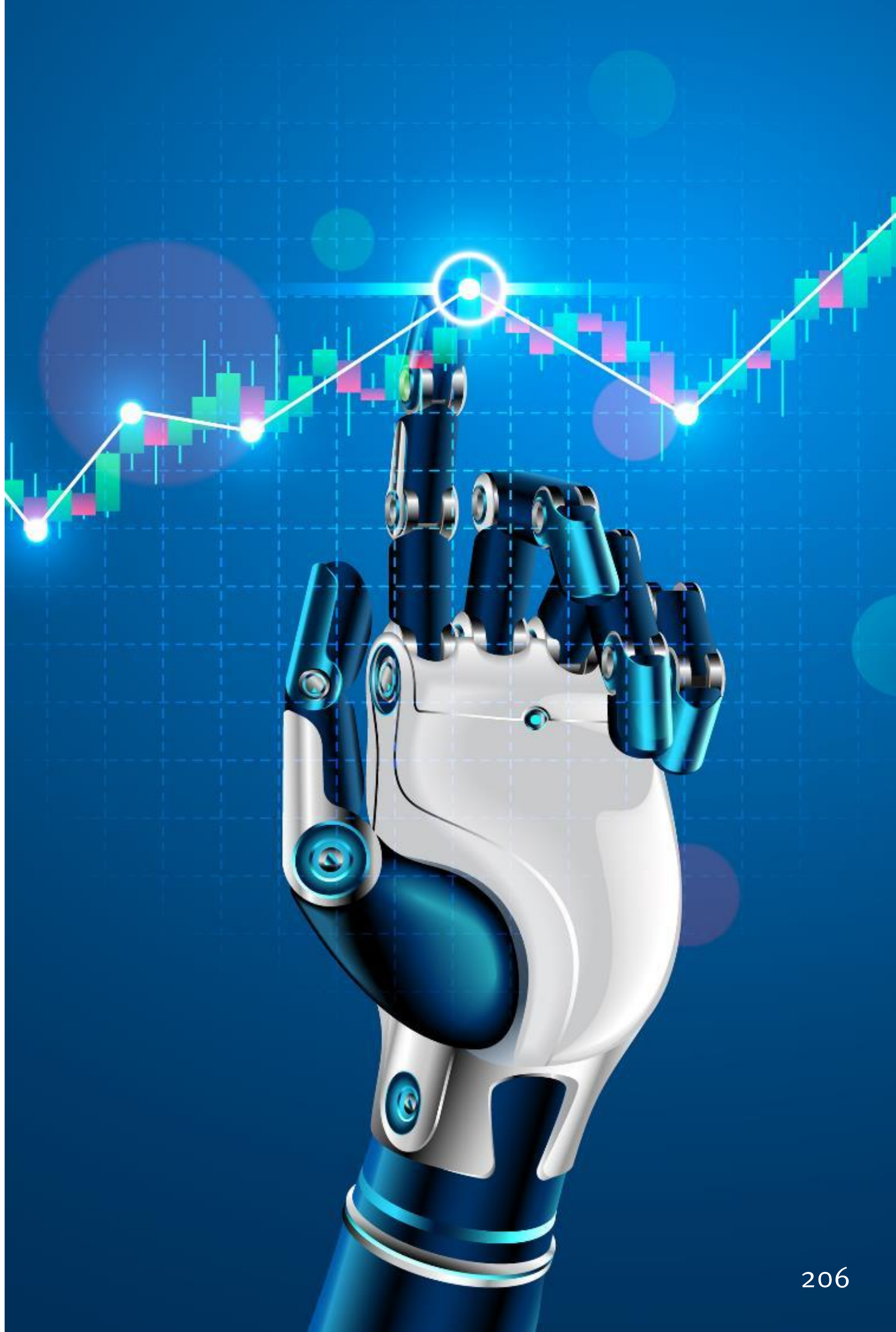
Here are some factors to consider.

First, good money has already been made from AI in biotech and will likely continue to be made. Nimbus contracted with Schrödinger to make it's TYK2 inhibitor, took it through clinical POC and then sold it off to Takeda for over $4 billion. While the details of the ROI calculation are not public, it was obviously a stellar deal for Nimbus shareholders. Also, the investors in Carmot (which was absolutely a computational QSAR shop) did quite well. The prudent use of AI/ML to get good drugs into the clinic is likely to continue to generate high payoffs for investors. The sticky part has been that many AI platforms have not yielded great drugs for the clinic – at least so far.

Second, M&A is going to happen in AI and biotech. We can't tell you which companies get bought and when, but we can tell you that M&A is very likely in the cards for one or more AI/Biotech companies. This was obvious in talking to the various pharma companies. They are the ones most likely to benefit from AI/ML tools and many are clearly under resourced. Historically, when important new technologies begin to mature, the independent platform companies get bought up. This was true for antibodies. It was true for RNAi and will be true for AI platforms. The questions is which platforms get bought and why. We noted highest pharma interest in biologics and target ID. Our gut is that some of the data-rich companies like nference, Recursion and Valo would be tempting targets.

Third, the Deep Learning in Biologics Theme is Working. It's hard to say if we are getting systematically better compounds out of the many small molecule AI platforms but

(continued)



Keep an Eye on the Industry Disruption Theme

it's obvious that *in silico* design of biologics using generative AI models works. It's a lower dimensional problem than small molecule AI design. Ironic, actually, given that biologics are larger molecules. There is one promising public platform in this area, Absci, and they are the lone positive performer among the public crop of AI-Centric biotechs this year. The crop of private companies in this area is really strong as we have discussed. We think that it is likely that one or more of the private companies gets snapped up for M&A in the months ahead.

Fourth, the disruption theme is the big one. We discussed three major ideas for how AI is going to disrupt the healthcare ecosystem and how pharma could participate. To remind, these are (1) AI to create highly conditioned treatment algorithms – facilitating back integration of pharma into healthcare and vice-versa, (2) the use of AI to create foundation models that go beyond creating drugs – if you have a foundation model for the cell, for example, it should be powerful in any number of settings and (3) applications of AI in organ systems where direct read/write access is possible could be highly disruptive. Elon Musk's Neuralink is getting ever closer to its IPO day and has been most recently valued at $5 billion based on private stock trades. We are bigger fans of bioelectronics + AI or in vivo cell engineering + AI than drug development + AI.

In many ways the venture landscape is the more exciting one today for AI investment. There is so little that is public in the hotter parts of the AI market thus far.

We are, ultimately, optimistic about AI as a place to invest and note that the quality of each story and team is critical in how well companies will do.

Big Pharma and AI

Big Pharma Using AI in Drug Discovery, Trials and Operations

Cindy Gordon, *Forbes*, Feb 23, 2024 (excerpt)

Identifying and accelerating drug development is big business. The costs in this industry are significant and finding pathways to optimize using AI methods is top of mind in this fast and evolving industry.

Many biopharmaceutical companies are using AI to speed up drug development. For example, machine-learning models are trained using information about the protein or amino-acid sequence or 3D structure of previous drug candidates, and about properties of interest.

Of course, every large pharmaceutical company is embracing AI and have been for some time. How efficient their integrated value chain processes - well that is a story likely for another day. Let's look at the drug process logic more closely.

First, in the drug identification stage, what is key is quantifying biological and clinical feasibility data to act as a transparent, data-driven intermediary between biotech startups and pharmaceutical companies. By developing AI models that analyze vast amounts of information – such as scientific literature, patents, clinical trials data, and market trends, early-stage biotech startups are able to demonstrate their competitive advantages and differentiation to potential investors. As a result, these startups can increase their chances of securing funding by showcasing their unique strengths and capabilities, backed by quantitative assessments generated by AI-driven analytical tools. As a result, pharmaceutical companies are able to make better decisions in their portfolios, deploy capital more efficiently and be in a stronger position to receive approval and drive a stronger ROI.

The second area where AI can bring immense value to the drug development process is in using large language systems, to speed up critical drug development functions such as operations, quality, and regulatory. For example, in the area of regulatory intelligence, AI systems can rapidly analyze extensive documentation, guidelines, and regulations to ensure that pharmaceutical companies remain compliant and up-to-date on the latest requirements from regulatory authorities. This not only increases efficiency, but also helps to reduce the risk of non-compliance, which could lead to delays in drug development and approval processes. Customers benefiting from these AI use cases experience tangible improvements in decision-making, risk mitigation, and overall efficiency. In addition, early-stage biotech startups have found it easier to secure funding with the backing of AI-driven, quantitative assessments of their innovations, while large pharmaceutical companies have been able to expedite.

Being able to optimize the drug discovery complex value chain is being improved significantly by AI driven approaches, but what is important in building a strong AI go to market offering is to ensure there are insights that others can genuinely learn from.

Source: https://www.forbes.com/sites/cindygordon/2024/02/23/using-ai-to-modernize-drug-development-and-lessons-learned/

Insight: Big Pharma Bets on AI to Speed Up Clinical Trials

Natalie Grover and Martin Coulter, *Reuters*, Sep 22, 2023 (excerpt)

Major drugmakers are using artificial intelligence to find patients for clinical trials quickly, or to reduce the number of people needed to test medicines, both accelerating drug development and potentially saving millions of dollars.

Human studies are the most expensive and time-consuming part of drug development as it can take years to recruit patients and trial new medicines in a process that can cost over a billion dollars from the discovery of a drug to the finishing line.

Reuters interviews with more than a dozen pharmaceutical company executives, drug regulators, public health experts and AI firms show, however, that the technology is playing a sizeable and growing role in human drug trials.

Companies such as Amgen, Bayer and Novartis are training AI to scan billions of public health records, prescription data, medical insurance claims and their internal data to find trial patients - in some cases halving the time it takes to sign them up.

"I don't think it's pervasive yet," said Jeffrey Morgan, managing director at Deloitte, which advises the life sciences industry. "But I think we're past the experimentation stage."

German drugmaker Bayer said it used AI to cut the number of participants needed by several thousand for a late-stage trial for asundexian, an experimental drug designed to reduce the long-term risk of strokes in adults.

It used AI to link the mid-stage trial results to real-world data from millions of patients in Finland and the United States to predict the long-term risks in a population similar to the trial.

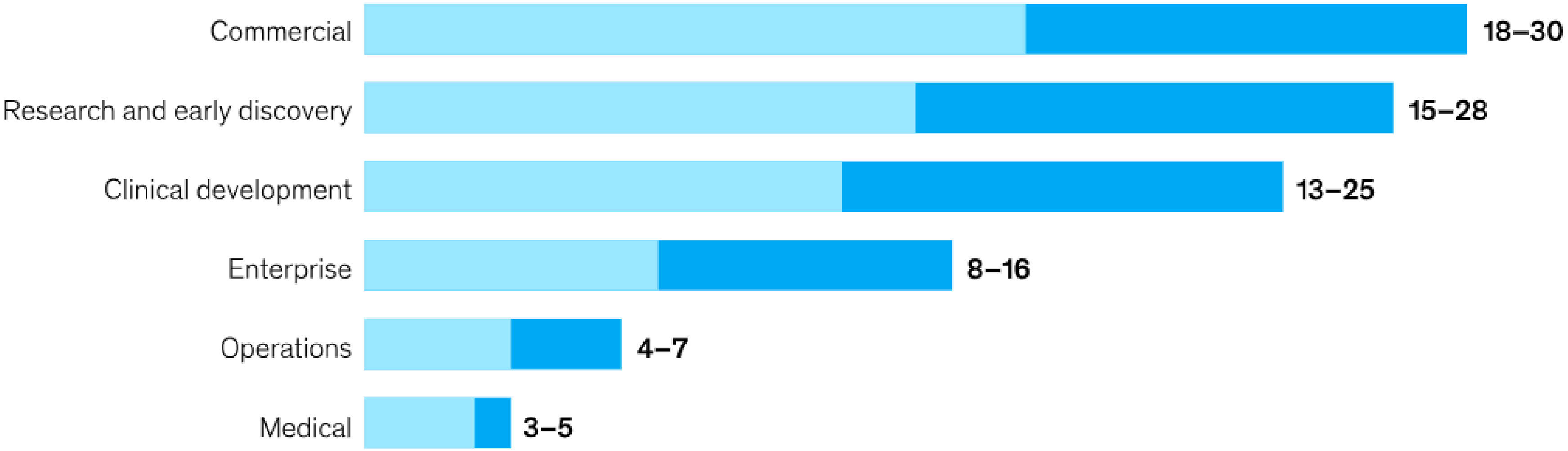
Source: https://www.reuters.com/technology/big-pharma-bets-ai-speed-up-clinical-trials-2023-09-22/

Big Value Potential from AI in Pharma

Exhibit 1 McKinsey & Company, “Generative AI in the pharmaceutical industry: Moving from hype to reality,” January 2024

Generative AI is expected to produce $60 billion to $110 billion in annual value across the pharmaceutical industry value chain.

Expected value annually, $ billion



Source: McKinsey analysis

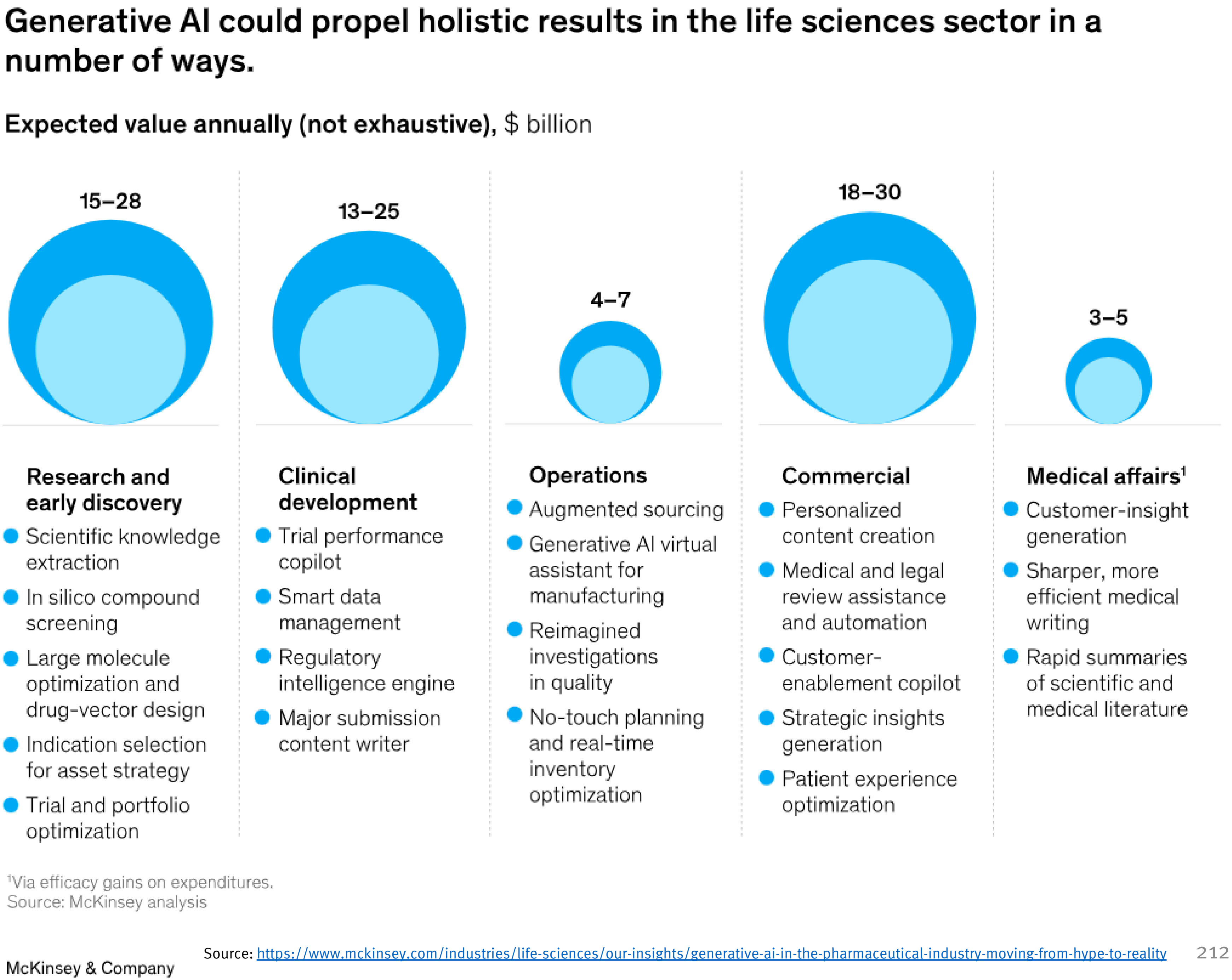
McKinsey & Company

Source: https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality

Numerous Use Cases

McKinsey & Company, "Generative AI in the pharmaceutical industry: Moving from hype to reality," January 2024

We dive into 21 individual use cases that McKinsey domain experts regard as having the greatest potential for a near-term impact across five life science domains. Many of these use cases cannot be realized unless some degree of digitalization is already in place, and not all of them will necessarily apply to all companies. While we recognize that gen AI remains an emerging technology not yet fully deployed at scale in most instances, we have also tried to estimate the potential impact for each use case.



Generative AI could propel holistic results in the life sciences sector in a number of ways.

Expected value annually (not exhaustive), $ billion

Research and early discovery

- Scientific knowledge extraction
- In silico compound screening
- Large molecule optimization and drug-vector design
- Indication selection for asset strategy
- Trial and portfolio optimization

Clinical development

- Trial performance copilot
- Smart data management
- Regulatory intelligence engine
- Major submission content writer

Operations

- Augmented sourcing
- Generative AI virtual assistant for manufacturing
- Reimagined investigations in quality
- No-touch planning and real-time inventory optimization

Commercial

- Personalized content creation
- Medical and legal review assistance and automation
- Customer-enablement copilot
- Strategic insights generation
- Patient experience optimization

Medical affairs[1]

- Customer-insight generation
- Sharper, more efficient medical writing
- Rapid summaries of scientific and medical literature

[1] Via efficacy gains on expenditures.
Source: McKinsey analysis

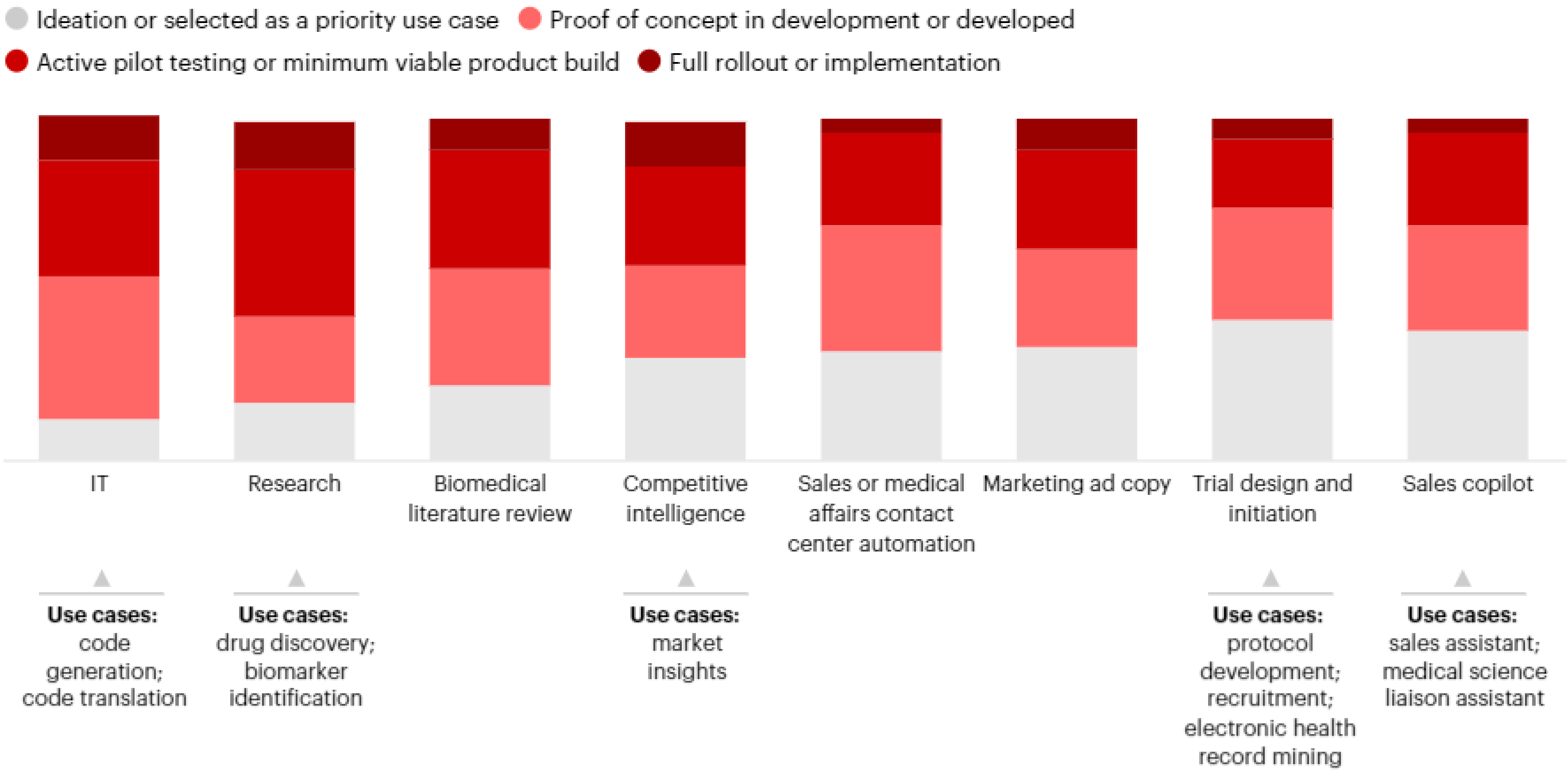
McKinsey & Company

Source: https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality

Most Pharma Uses Cases are Still in Pilot Testing Mode

Eric Berger, Robbie Sandig, and KC George, “How to Successfully Scale Generative AI in Pharma,” Bain Brief, February 12, 2024

Percentage of executives who report reaching the stage of development by use case area



Use cases: code generation; code translation

Use cases: drug discovery; biomarker identification

Use cases: market insights

Use cases: protocol development; recruitment; electronic health record mining

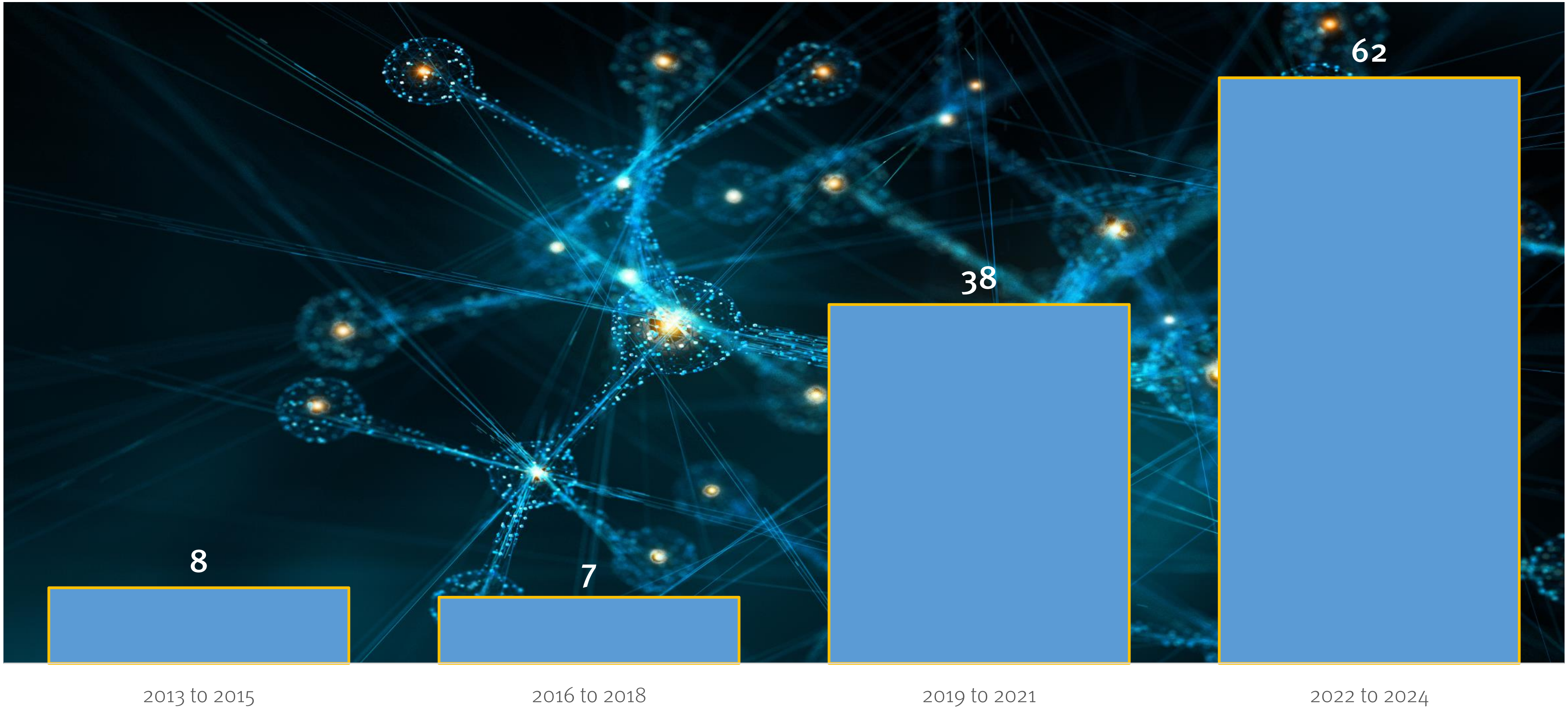
Use cases: sales assistant; medical science liaison assistant

Source: Bain Generative Artificial Intelligence in Pharma Survey, September 2023 (N=100)

Source: https://www.bain.com/insights/how-to-successfully-scale-generative-ai-in-pharma

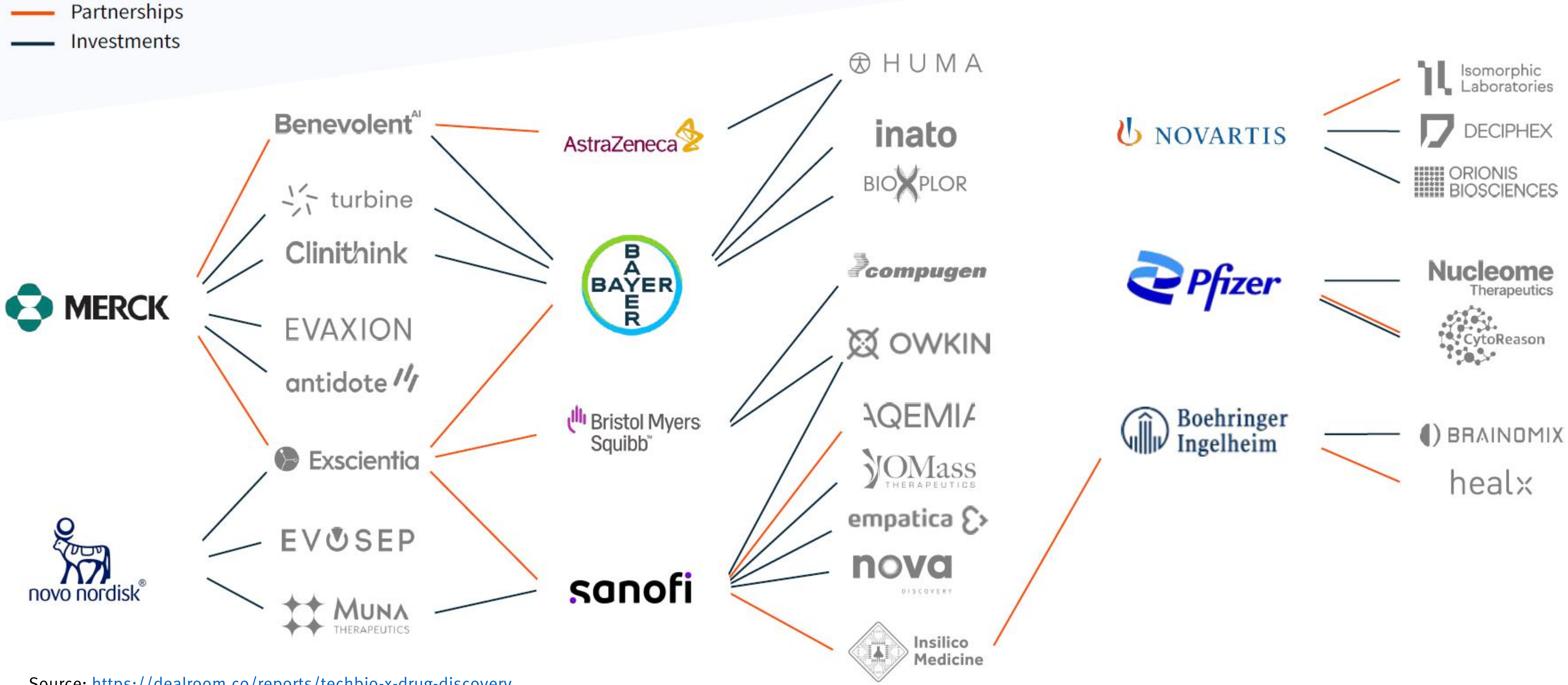
Big Pharma Interest in AI for Drug R&D Picked Up in 2019

Number of Pharma Licensing Deals Involving AI/ML Technologies and Drug Discovery, 2013 to 2024



Source: DealForma

Big pharma is playing a catch-up on AI, as evidenced by their recent surge in investments, partnerships, and strategic bets in the field.



Source: https://dealroom.co/reports/techbio-x-drug-discovery

Big Pharma Investments in AI and ML

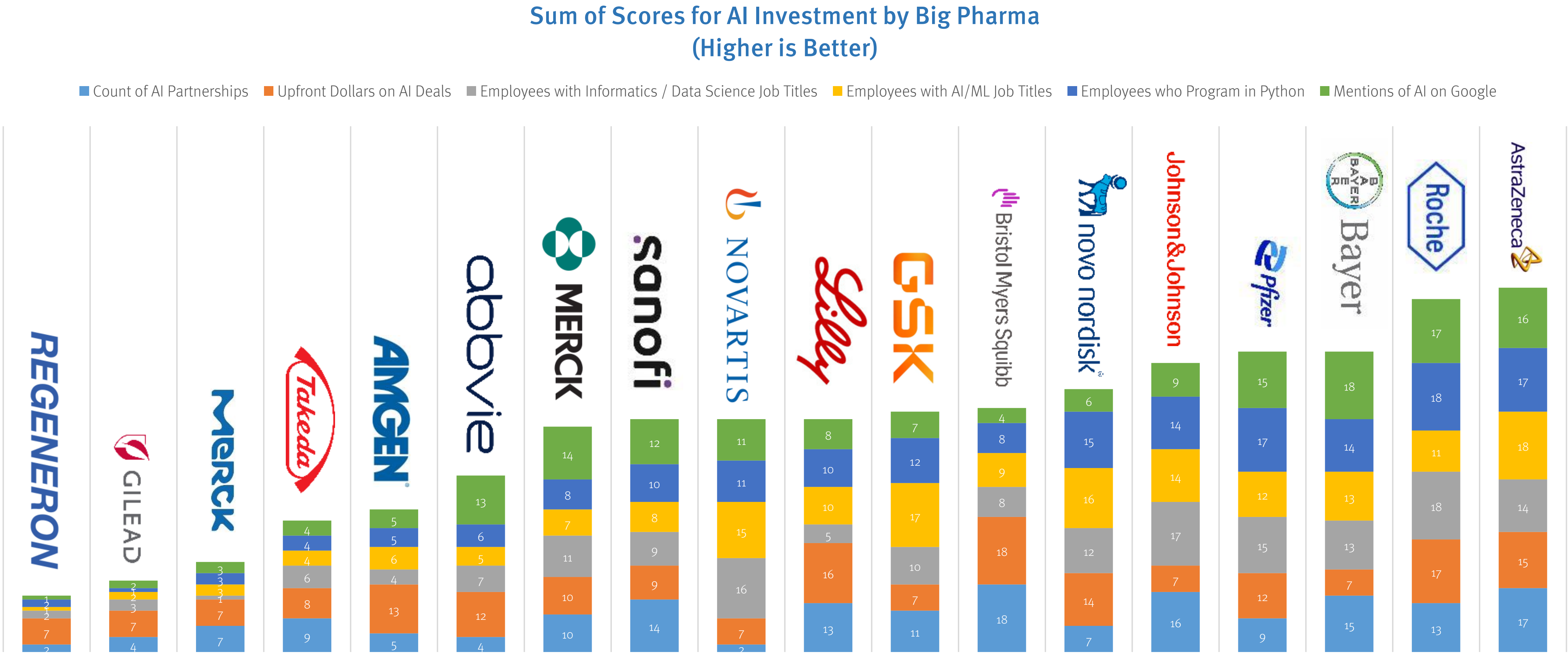
AstraZeneca leads the pack today with number of employees with AI or ML in their job titles and overall ranks across the board. BMS has spent the most on AI deal upfronts, followed by Roche and Sanofi. Roche has a whopping 2,700 employees who know how to program in Python. Bayer, Pfizer and J&J are relatively strong across the board.

Company	Overall Rank (sum of ranks)	Count of AI Partnership Deals(2)	Disclosed Upfront Dollars on AI Deals(2)	Employees with Informatics / Data Science Job Titles(3)	Employees with AI/ML Job Titles(3)	Employees who Program in Python(3)	Mentions of Company and AI on Google(4)
AstraZeneca	1	15	$81	259	304	2300	1,860,000
Roche	2	9	$159	454	146	2700	2,400,000
Bayer	3	6	$30	283	150	2300	1,590,000
Pfizer	3	12	$0	214	160	2000	7,550,000
J&J	5	13	$0	341	163	2000	554,000
Novo Nordisk	6	5	$75	178	191	2100	409,000
BMS	7	18	$433	99	68	1200	336,000
GSK	8	8	$0	117	236	1900	499,000
Eli Lilly	9	11	$4	116	62	1700	773,000
Novartis	9	3	$0	339	182	1800	634,000
Sanofi	9	9	$152	79	138	1700	518,000
Merck	12	7	$9	149	59	1200	1,030,000
AbbVie	13	1	$30	87	45	1100	796,000
Amgen	14	4	$50	66	46	1030	347,000
Takeda	15	6	$3	85	45	841	616,000
Merck Kgaa	16	5	$0	32	21	741	260,000
Gilead	17	1	$0	58	18	502	251,000
Regeneron	18	0	$0	39	14	697	128,000

Sources: (1) Stifel analysis using ranks sum of reverse ranks on an equal-weighted basis, (2) DealForma (2010 to 2024 period), (3) LinkedIn Search of people in company and affiliated subsidiaries, (4) Google search using “artificial intelligence” and company name.

Big Pharma Investments in Artificial Intelligence and Machine Learning

AstraZeneca and Roche have both made heavy investments in AI in their pharma franchises. Pfizer, J&J, Novo, BMS and GSK are not far behind. Regeneron, in contrast, has actively indicated that they do not see AI as an integral determinant of their R&D productivity.



Sources: (1) Stifel analysis using sum of reverse ranks on an equal-weighted basis (a score of 18 means the company is #1 on that metric whereas a score of 1 means a company is last on that metric), (2) DealForma (2010 to 2024 period), (3) LinkedIn Search of people in company and affiliated subsidiaries, (4) Google search using "artificial intelligence" and company name.

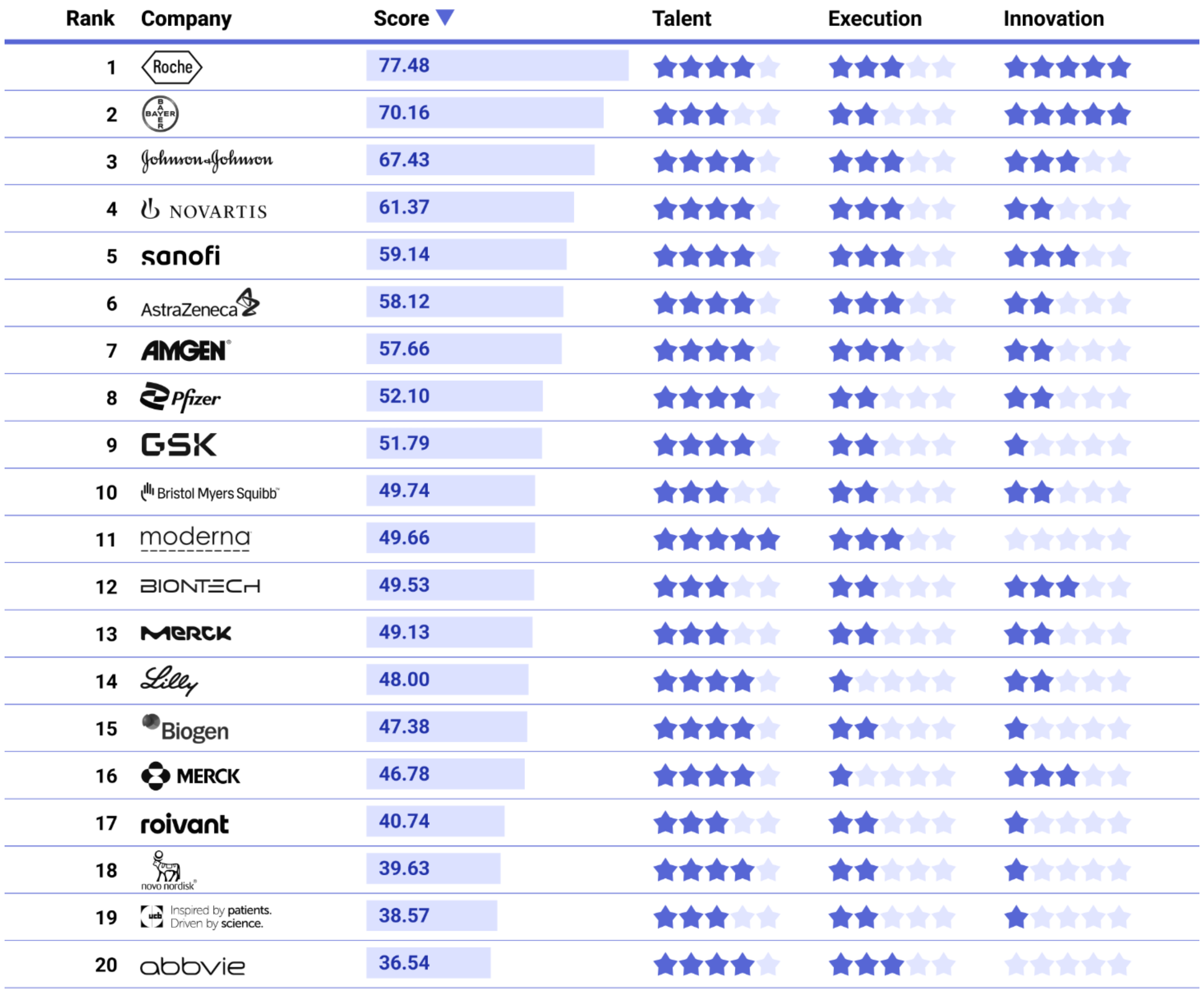
Comparison to the CB Insights Pharma AI Readiness Index

Interestingly, we found the CB Insights AI Readiness results after we had done our own analysis. It's striking that our rankings are quite similar. The biggest differences is that we rank AstraZeneca and Pfizer more highly. We and CBI both agree on the relative scale of investments made by Roche, Bayer and J&J.

CB Insights launched the Pharma AI Readiness Index in 2023. This is a ranking of the 50 largest pharmaceutical companies in the Americas and Europe by market cap, based on their demonstrated ability to attract top AI talent, execute AI projects, and innovate through R&D and investments.

The index is calculated based on CB Insights datasets including patent applications, partnership & licensing agreements, dealmaking activity, acquisitions, key people, product launches, and earnings transcripts.

Roche and Bayer are the 2 top-scoring pharma companies, primarily due to their leading levels of AI innovation via acquisitions, investments, and patents. Like Roche and Bayer, other high-scoring companies have made clear investments in AI talent and — to a lesser degree — have a demonstrated ability to execute AI initiatives.



Rank	Company	Score ▼	Talent	Execution	Innovation
1	Roche	77.48	★★★★☆	★★★☆☆	★★★★★
2	BAYER	70.16	★★★☆☆	★★☆☆☆	★★★★★
3	Johnson&Johnson	67.43	★★★★☆	★★★☆☆	★★★☆☆
4	NOVARTIS	61.37	★★★★☆	★★★☆☆	★★☆☆☆
5	sanofi	59.14	★★★★☆	★★★☆☆	★★★☆☆
6	AstraZeneca	58.12	★★★★☆	★★★☆☆	★★☆☆☆
7	AMGEN	57.66	★★★★☆	★★★☆☆	★★☆☆☆
8	Pfizer	52.10	★★★★☆	★★☆☆☆	★★☆☆☆
9	GSK	51.79	★★★★☆	★★☆☆☆	★☆☆☆☆
10	Bristol Myers Squibb	49.74	★★★☆☆	★★☆☆☆	★★☆☆☆
11	moderna	49.66	★★★★★	★★★☆☆	☆☆☆☆☆
12	BIONTECH	49.53	★★★☆☆	★★☆☆☆	★★★☆☆
13	MERCK	49.13	★★★☆☆	★★☆☆☆	★★☆☆☆
14	Lilly	48.00	★★★★☆	★☆☆☆☆	★★☆☆☆
15	Biogen	47.38	★★★★☆	★★☆☆☆	★☆☆☆☆
16	MERCK	46.78	★★★★☆	★☆☆☆☆	★★★☆☆
17	roivant	40.74	★★★☆☆	★★☆☆☆	★☆☆☆☆
18	novo nordisk	39.63	★★★★☆	★★☆☆☆	★☆☆☆☆
19	ucb Inspired by patients. Driven by science.	38.57	★★★☆☆	★★☆☆☆	★☆☆☆☆
20	abbvie	36.54	★★★★☆	★★★☆☆	☆☆☆☆☆

Source: https://www.cbinsights.com/research/ai-readiness-index-pharma/

Suggested Approach for Big Pharma in AI

Margaret Ayers, Madura Jayatunga, John Goldader, and Chris Meier, “Adopting AI in Drug Discovery,” *BCG Report*, Mar 29, 2022 (excerpt)

The AI revolution in drug discovery will not happen overnight. Even as AI-driven innovations show impressive results, established pharmaceutical companies retain many advantages. These include capital, scientific expertise, development know-how and experience, regulatory expertise, and established branding and commercial teams. That said, some of the pillars of incumbency are showing early erosion. Massive fundraising and less cost-intensive in vitro work are lowering the capital barriers for startup discovery programs. Meanwhile, AI natives are filling out their ranks with scientists and medical experts, replicating the advantages of big companies employee by employee.

It's possible—though not easy—to combine the best of both worlds. In our experience, adapting a classical drug discovery process and delivering on the promise of AI require long-term action on five strategic and operational tracks. (See Exhibit 2.)

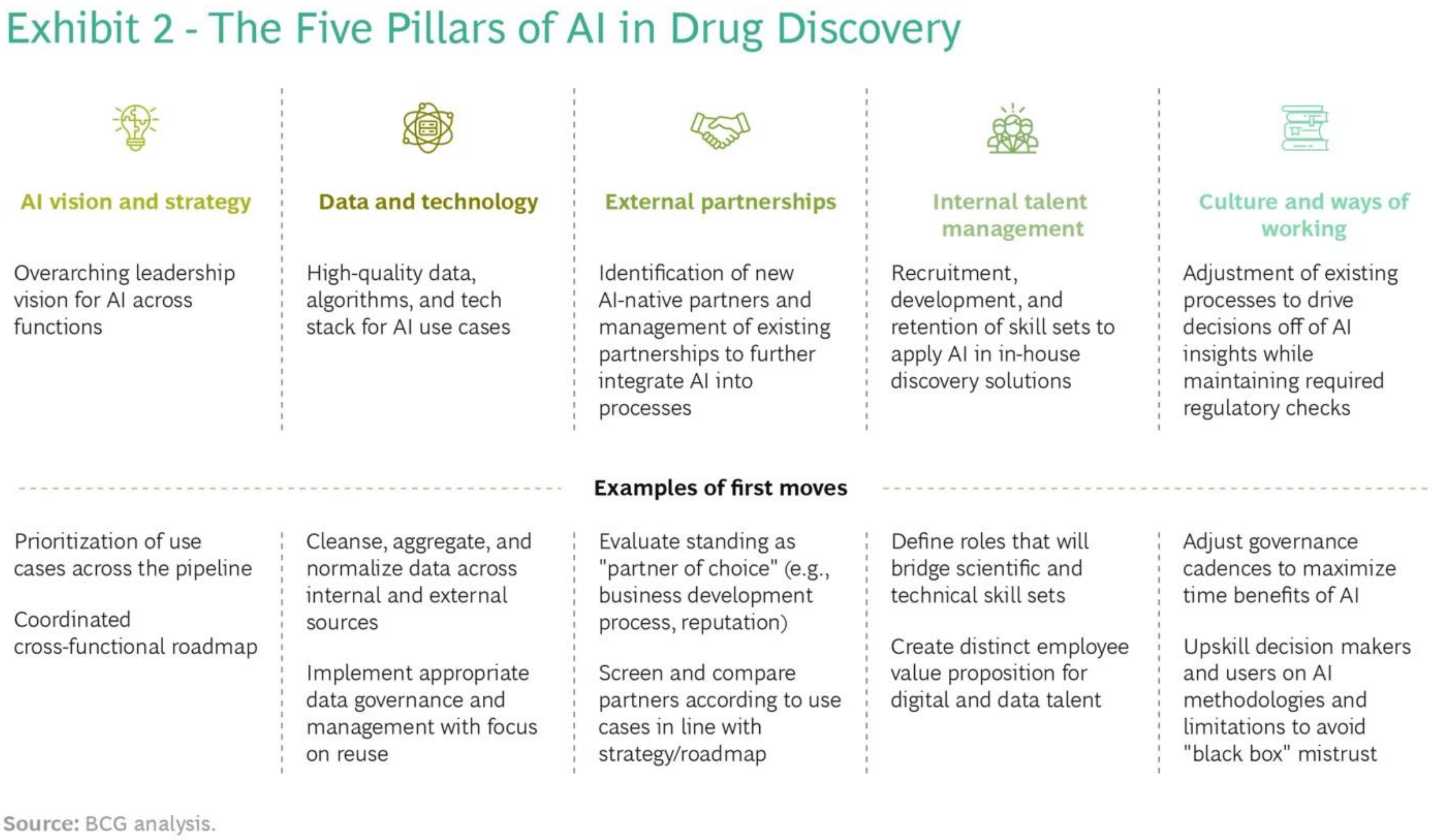


Exhibit 2 - The Five Pillars of AI in Drug Discovery

AI vision and strategy	Data and technology	External partnerships	Internal talent management	Culture and ways of working
Overarching leadership vision for AI across functions	High-quality data, algorithms, and tech stack for AI use cases	Identification of new AI-native partners and management of existing partnerships to further integrate AI into processes	Recruitment, development, and retention of skill sets to apply AI in in-house discovery solutions	Adjustment of existing processes to drive decisions off of AI insights while maintaining required regulatory checks
		Examples of first moves		
Prioritization of use cases across the pipeline Coordinated cross-functional roadmap	Cleanse, aggregate, and normalize data across internal and external sources Implement appropriate data governance and management with focus on reuse	Evaluate standing as "partner of choice" (e.g., business development process, reputation) Screen and compare partners according to use cases in line with strategy/roadmap	Define roles that will bridge scientific and technical skill sets Create distinct employee value proposition for digital and data talent	Adjust governance cadences to maximize time benefits of AI Upskill decision makers and users on AI methodologies and limitations to avoid "black box" mistrust

Source: BCG analysis.

Source: https://www.bcg.com/publications/2022/adopting-ai-in-pharmaceutical-discovery

AMGEN

Work on Artificial Intelligence and Machine Learning

AMGEN Using AI in Protein Drug Development

Amgen article, October 3, 2023

Summary:

- **Generative biology** pairs artificial intelligence (AI)/machine learning (ML) with innovations in biology and the lab to make medicines more quickly and effectively.
- **Advances in predictive and generative AI** are enabling researchers to design proteins that are more suitable to be made into drugs than those found in nature.
- **New lab techniques** such as human immune system tissue models have the potential to predict liabilities (such as unwanted immune responses) in protein drugs earlier in the drug development process.
- **Federated learning**, a data-sharing model that aims to protect companies' proprietary information while still sharing important research data, can provide much-needed high-quality protein data to help AI/ML models design better proteins more quickly.

Protein drug development is long, arduous and costly. Drug developers have typically looked to proteins in nature as starting points, and then gone through the slow, painstaking process of shaping those natural proteins into safe, effective drugs.

But the advent of artificial intelligence (AI), advanced analytical techniques and new, innovative ways to do life science research is changing all of this through a process called generative biology.

Similar to how generative AI systems (like ChatGPT) allow for the generation of new data such as text or images from existing inputs, generative biology allows for the generation of new protein-based drugs that have desired structures and properties based on existing protein data inputs.

Researchers at Amgen have started using data collected in the lab about a protein's sequence, structure and function to train machine learning (ML) algorithms to design drug candidates more quickly and with greater success than turning natural proteins into drugs. These designed protein candidates can then be evaluated using automated, high-throughput platforms in the lab, providing additional data to fine-tune the ML models in a type of generative loop.

Source: https://www.amgen.com/stories/2022/10/follow-the-data

Amgen AI and Protein Development Story (continued)

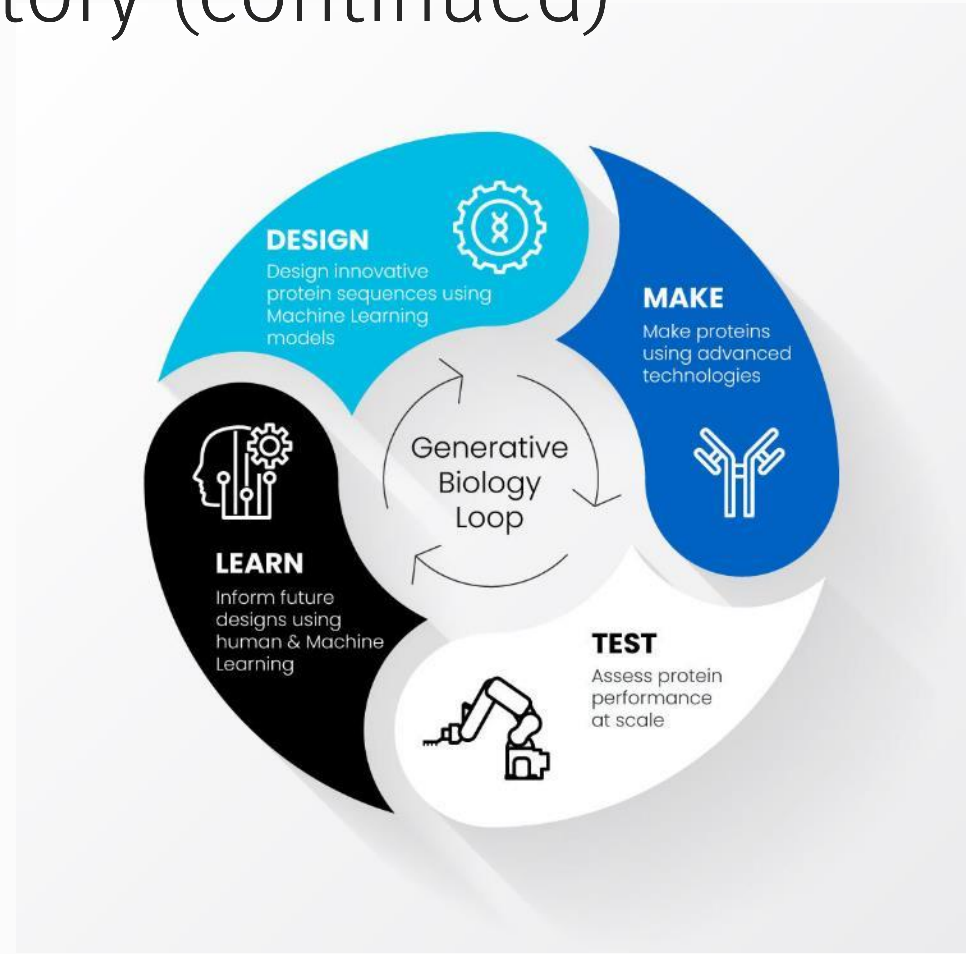
Improved computer modeling: Machine learning involves developing computer models that recognize and learn from patterns found in the data they are trained on. In the past, predicting patterns such as the structure of a protein was a big challenge for computer models. But over time, the models improved thanks to innovations in how to build them and as researchers began training them on hundreds of millions of naturally occurring protein sequences. This allowed the models to recognize complex patterns within protein sequences related to those proteins' structures and functions. With advances in generative AI, some models are even capable of generating brand-new protein designs, a much-desired advance with the potential to vastly improve drug development.

To make these ML models more amenable to drug development, researchers need functional data on whether a protein binds to the target of interest. They also need data on proteins' properties of interest to better understand their behavior. These properties may include how viscous a protein is in liquid form, how stable it is at room temperature, where it goes in the body and how the body responds to it.

To predict how a protein will behave earlier in the development process and increase chances of success, researchers are also using more computational approaches. For example, Amgen has developed an ML model to predict a protein's viscosity, an important property in drug development. If a drug is too viscous (for example, the consistency of honey), it can be too difficult to inject.

To understand a protein's properties (such as viscosity), researchers first need to consider its sequence, or the order of its amino acid building blocks. This sequence determines how a protein behaves. It has traditionally been exceedingly difficult to predict things like viscosity based on a protein's sequence because of the complexity in predicting how such large and complex molecules will behave from their individual pieces. With the advent of generative AI and more powerful ML models, it is now possible to make those predictions.

To predict protein viscosity at Amgen, researchers used sequence data from 83 antibody proteins selected from both internal and external databases. Next, they made enough of each antibody to do extensive testing. Gathering this data, they trained ML models to use the protein sequences to predict if the antibodies had high or low viscosity with greater than 80% accuracy.


DESIGN
Design innovative protein sequences using Machine Learning models
MAKE
Make proteins using advanced technologies
TEST
Assess protein performance at scale
LEARN
Inform future designs using human & Machine Learning
Generative Biology Loop

Generative biology accelerates and improves protein drug development by combining the use of machine learning-driven computational models with automated, high-throughput techniques that test protein designs in the lab. These processes of designing, making, testing and learning about proteins of interest feed into each other in a generative loop.

Source: https://www.amgen.com/stories/2023/10/how-ai-and-life-sciences-research-are-revolutionizing-protein-drug-development

AstraZeneca

Work on Artificial Intelligence and Machine Learning

Comments from:

Jim Weatherall

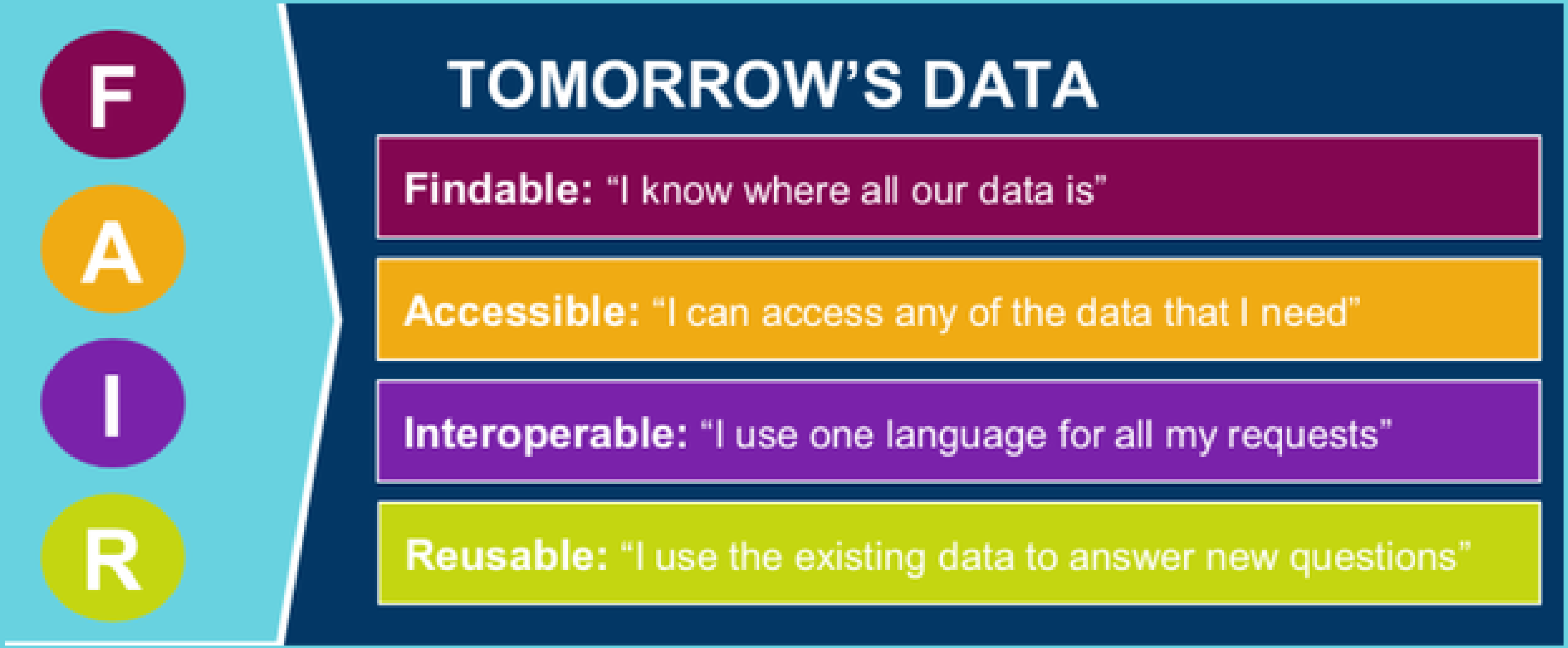
Vice President, Data Science & AI, R&D
AstraZeneca

“Data science and AI are transforming R&D, helping us turn science into medicine more quickly and with a higher probability of success.”

“AI will not replace drug hunters, but drug hunters who don’t use AI will be replaced by those who do.”

Sources: https://www.astrazeneca.com/r-d/data-science-and-ai.html , https://www.youtube.com/watch?v=j2tMVBM7fEQ

AZ Focused on Mobilizing Internal Data Resources For Scientists



TOMORROW'S DATA

Findable: "I know where all our data is"

Accessible: "I can access any of the data that I need"

Interoperable: "I use one language for all my requests"

Reusable: "I use the existing data to answer new questions"

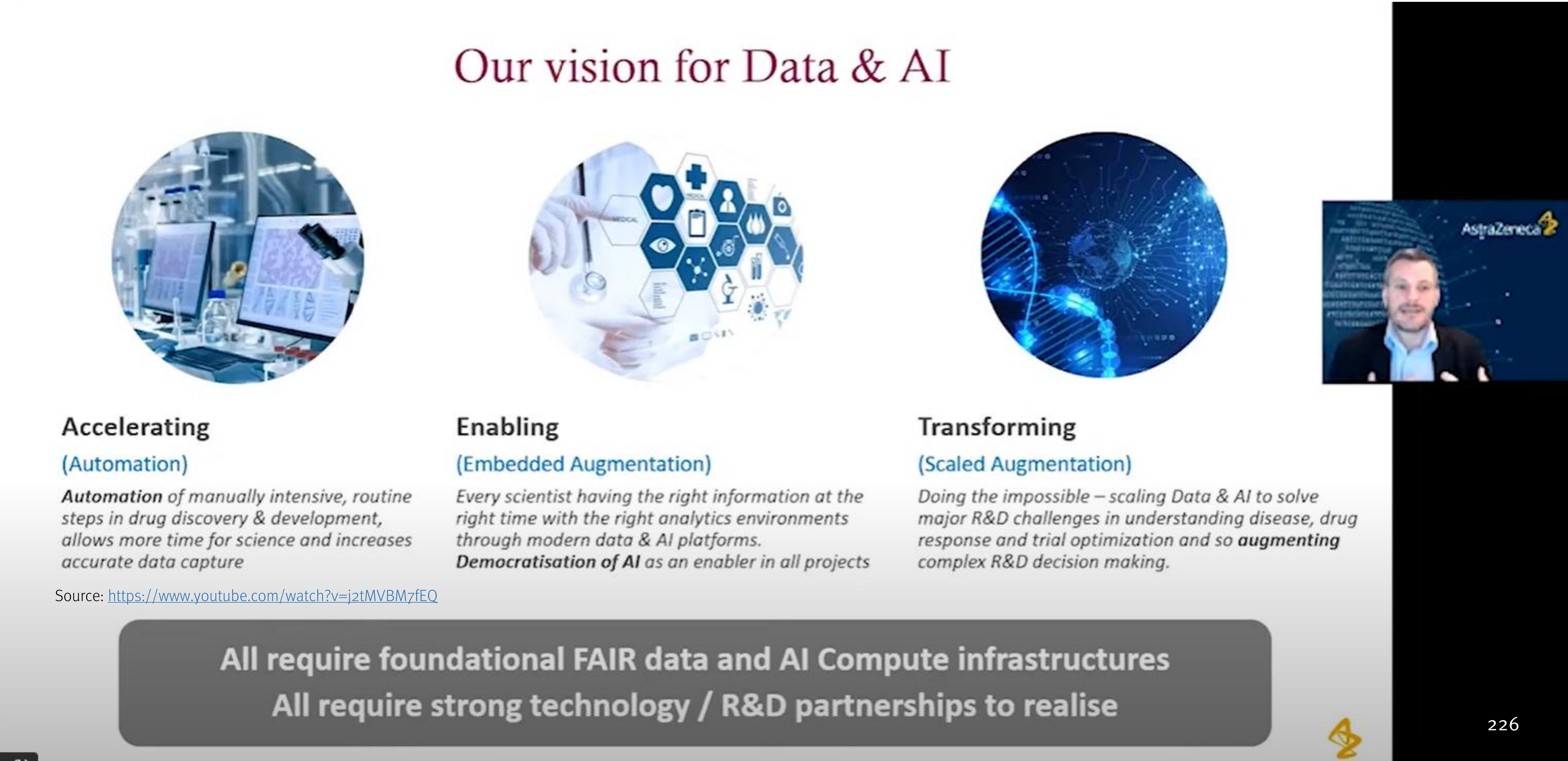
Today we are generating and have access to more data than ever before. Data and analytics have the potential to transform our business, but the true value of scientific data can only be realised if it is "FAIR" - Findable, Accessible, Interoperable and Reusable.

AstraZeneca's R&D and IT groups are working closely together to create an industry-leading enterprise data and AI architecture. This will help us answer key business questions and enhance our ability to harness new tools and technologies, such as AI and machine learning, both now and in the future.

We are also mobilising a team of data scientists, bioinformaticians, data engineers and machine learning experts from across the company to ensure we are collecting, organising and using the right data in the best way.

Jim Weatherall Slides on AZ AI Strategy

Source: https://www.youtube.com/watch?v=j2tMVBM7fEQ



Our vision for Data & AI

Accelerating

(Automation)

***Automation** of manually intensive, routine steps in drug discovery & development, allows more time for science and increases accurate data capture*

Enabling

(Embedded Augmentation)

*Every scientist having the right information at the right time with the right analytics environments through modern data & AI platforms. **Democratisation of AI** as an enabler in all projects*

Transforming

(Scaled Augmentation)

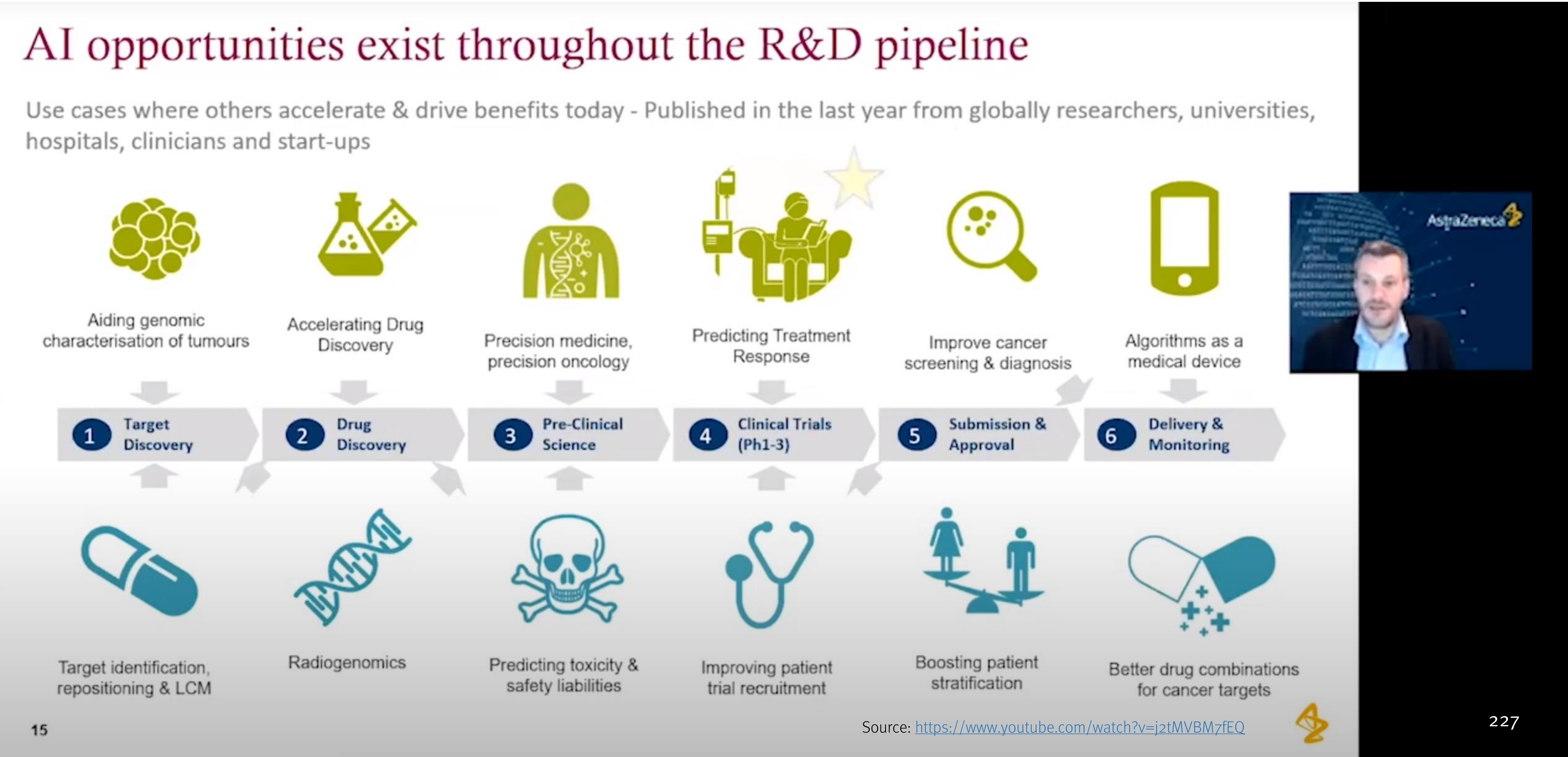
*Doing the impossible – scaling Data & AI to solve major R&D challenges in understanding disease, drug response and trial optimization and so **augmenting** complex R&D decision making.*

Source: https://www.youtube.com/watch?v=j2tMVBM7fEQ

All require foundational FAIR data and AI Compute infrastructures

All require strong technology / R&D partnerships to realise

Jim Weatherall Slides on AZ AI Strategy



AI opportunities exist throughout the R&D pipeline

Use cases where others accelerate & drive benefits today - Published in the last year from globally researchers, universities, hospitals, clinicians and start-ups

Aiding genomic characterisation of tumours

Accelerating Drug Discovery

Precision medicine, precision oncology

Predicting Treatment Response

Improve cancer screening & diagnosis

Algorithms as a medical device

1 Target Discovery

2 Drug Discovery

3 Pre-Clinical Science

4 Clinical Trials (Ph1-3)

5 Submission & Approval

6 Delivery & Monitoring

Target identification, repositioning & LCM

Radiogenomics

Predicting toxicity & safety liabilities

Improving patient trial recruitment

Boosting patient stratification

Better drug combinations for cancer targets

15

Source: https://www.youtube.com/watch?v=j2tMVBM7fEQ

Work on Artificial Intelligence and Machine Learning

Bayer Partnering with World Leaders in AI

RECURSION

Berlin, Germany and Salt Lake City, USA, Nov. 09, 2023 (GLOBE NEWSWIRE) -- Bayer and US-based Recursion Pharmaceuticals, Inc., a clinical stage TechBio company decoding biology to industrialize drug discovery, today announced that they have updated the focus of their research collaboration to precision oncology.

The oncology-focused collaboration will leverage Bayer's small molecule compound library and expertise in biology and medicinal chemistry as well as Recursion's purpose-built artificial intelligence-guided drug discovery platform. This strategic shift will enable Bayer to utilize Recursion's capabilities to initiate and advance the identification of novel therapeutic targets for challenging oncology indications with high unmet need.

"The methodology in which Recursion uses artificial intelligence (AI) in drug discovery, could be one of the most disruptive technologies of our time," said Juergen Eckhardt, M.D., Head of Business Development, Licensing & Open Innovation, Pharmaceuticals Division, Bayer AG, and Head of Leaps by Bayer. "As our collaboration and the usage of AI continue to evolve, we look forward to continuing to work with industry innovators to identify novel targets for oncology indications."

AIGNOSTICS

Berlin, Germany, March 14, 2024 – Bayer and Aignostics GmbH today announced a strategic collaboration on several artificial intelligence (AI)-powered approaches with applications in precision oncology drug research and development. Aignostics is a spin-off from one of the world's leading hospitals, Charité-Universitätsmedizin Berlin, and a global leader in using computational pathology to transform complex biomedical data into biology insights.

The partners will co-create a novel target identification platform that leverages Aignostics' technology and proprietary multimodal patient cohorts, and Bayer's deep expertise in discovering and developing novel oncology therapies. In addition, the collaboration will include the development of computational pathology algorithms powered by AI and machine learning (ML) that connect baseline pathology data, such as molecular tumor profiles, with clinical data, such as patient outcomes, to enable better patient identification, stratification, and selection for clinical trials.

Sources: https://ir.recursion.com/news-releases/news-release-details/bayer-and-recursion-focus-research-collaboration-oncology, https://www.bayer.com/media/en-us/bayer-and-aignostics-to-collaborate-on-next-generation-precision-oncology

Bayer Has Adopted an Enterprise-Wide Generative AI Team

We established a genAI catalyst team, a federated, cross-enterprise team which steers those efforts across the three divisions and functions we have. This group has a strategic, end-to-end perspective. Its members understand what we are doing, establish principles, then facilitate the adoption.

What we saw – and this is an important point – is on the one hand, this federated, cross-enterprise, cross-functional team would oversee the things we want to build centrally, while maintaining flexibility and freedom for the divisions and functions.

It's not a control tower. It's just a way of saying, "How can we be super-smart in cost, efficiency and speed while at the same time being flexible and maximizing freedom for the divisions and the functions" – the best of both worlds? How can we scale capabilities quickly?

Just to understand initially how it's working, we ran a few pilots across our business to ensure that it's somewhat low-effort while keeping risks in-check.

For example, in the U.S. we used a Microsoft GitHub co-pilot. We started developing our own internal ChatGPT, basically a GPT for internal use. The GitHub co-pilot worked well. We saw that it can augment what we're doing in coding. It's super-efficient and productive for our coders, however, it's always complementary to what people are doing.

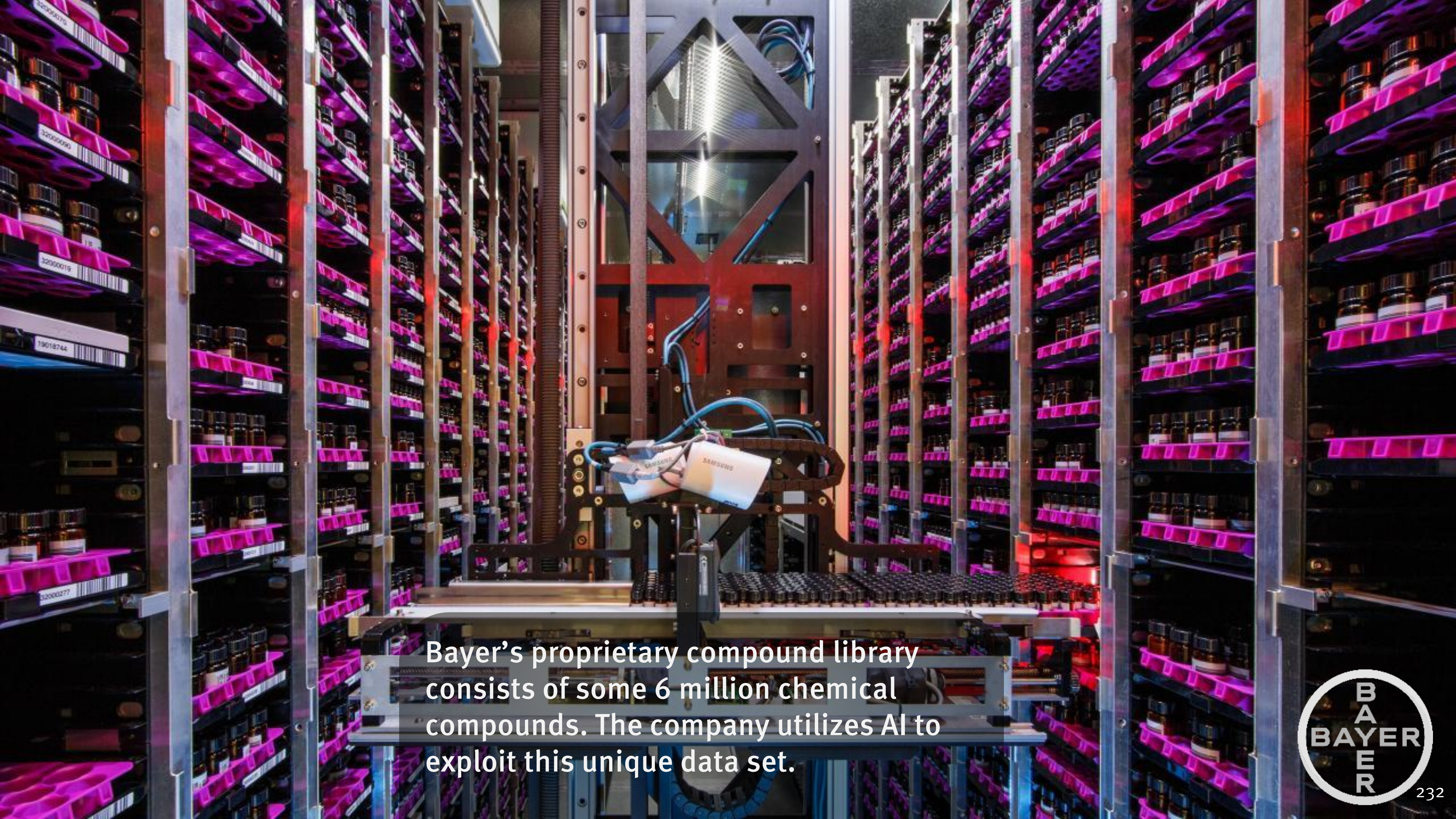
Saskia Steinacker
Bayer Head of Digital Transformation

Source: https://www.mmm-online.com/home/channel/7-day-supply/inside-bayers-embrace-of-generative-ai/

Bayer's Aalto University and Helsinki University Trial Collaboration

Future Clinical Trials, a research collaboration between Bayer in Finland, Aalto University and the Helsinki University Hospital, aims to transform the way trials are planned and conducted, improving efficiency and safety of clinical drug research. Building on real-world, high quality medical data, artificial intelligence is utilized to identify the right patients for trials. Allowing for a more efficient classification of patients and due to its ability to recognize even the rarest side effects, AI is helping to avoid potential risks and achieve better results. The partners are also investigating how AI could be used to augment the way clinical trials are designed by introducing external control groups using historical, virtual or synthetic data sets. Harnessing AI in this way has the potential to broaden our approach to clinical trial design by adding innovative and valuable methods to study the effects of new drugs – ultimately increasing the speed with which new treatments can reach patients all over the world.

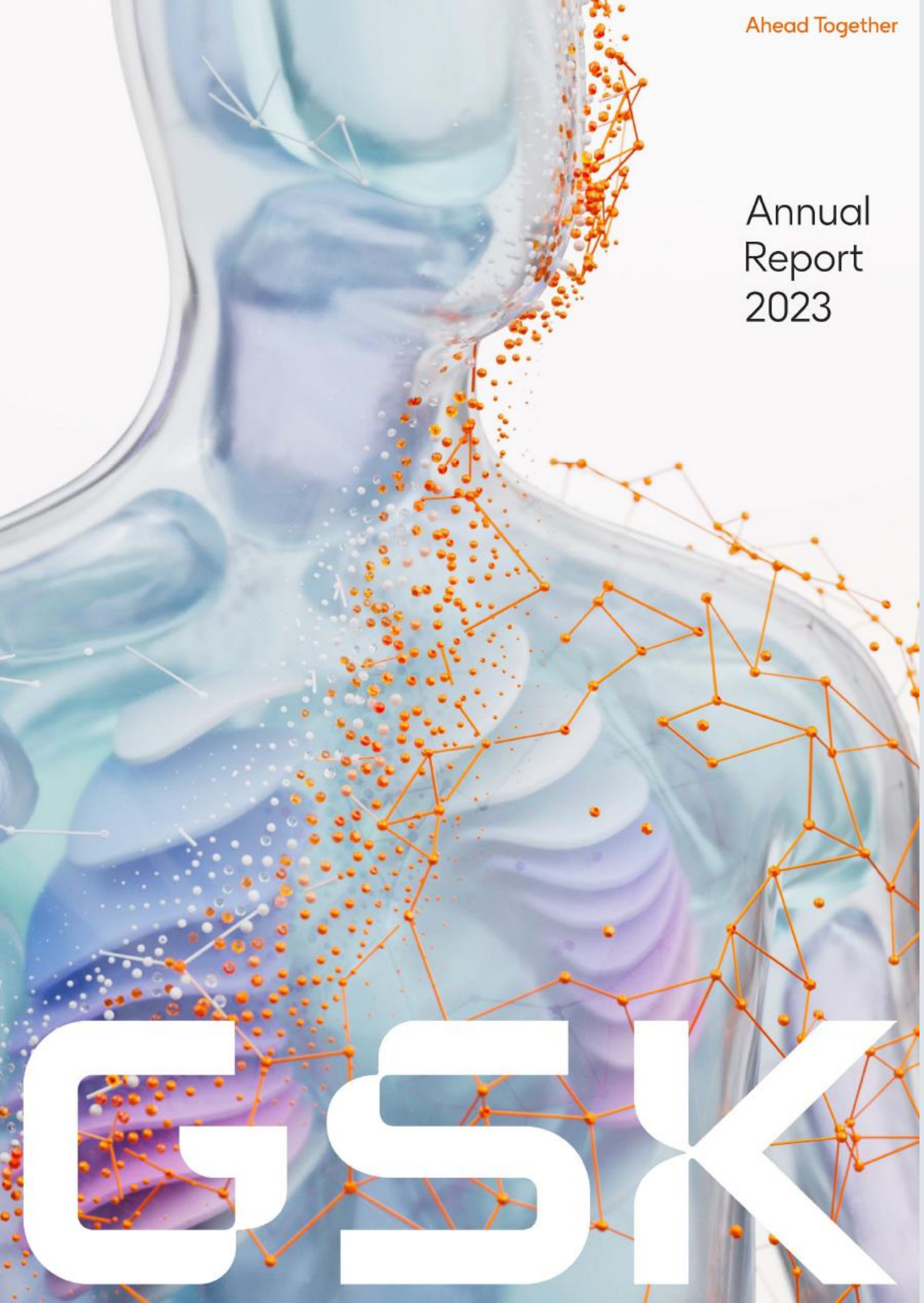
Source: https://www.bayer.com/en/pharma/artificial-intelligence#1



Bayer's proprietary compound library consists of some 6 million chemical compounds. The company utilizes AI to exploit this unique data set.

GSK

Work on Artificial Intelligence and Machine Learning



Ahead Together

Annual Report 2023

GSK

Kim Branson

Global Head of Artificial Intelligence and Machine Learning
GSK

“One of the most important things for machine learning in drug discovery is actually what to design the medicine against. It doesn't matter if you have the best medicine in the world, if you have the wrong target, you're not going to see the clinical effect.”

‘AI has already transformed medical innovation. Let’s not let fears of technology stop us from realizing its incredible potential.’

GSK

Source: https://www.gsk.com/en-gb/behind-the-science-magazine/ai-medical-innovation-ethics-responsibility/

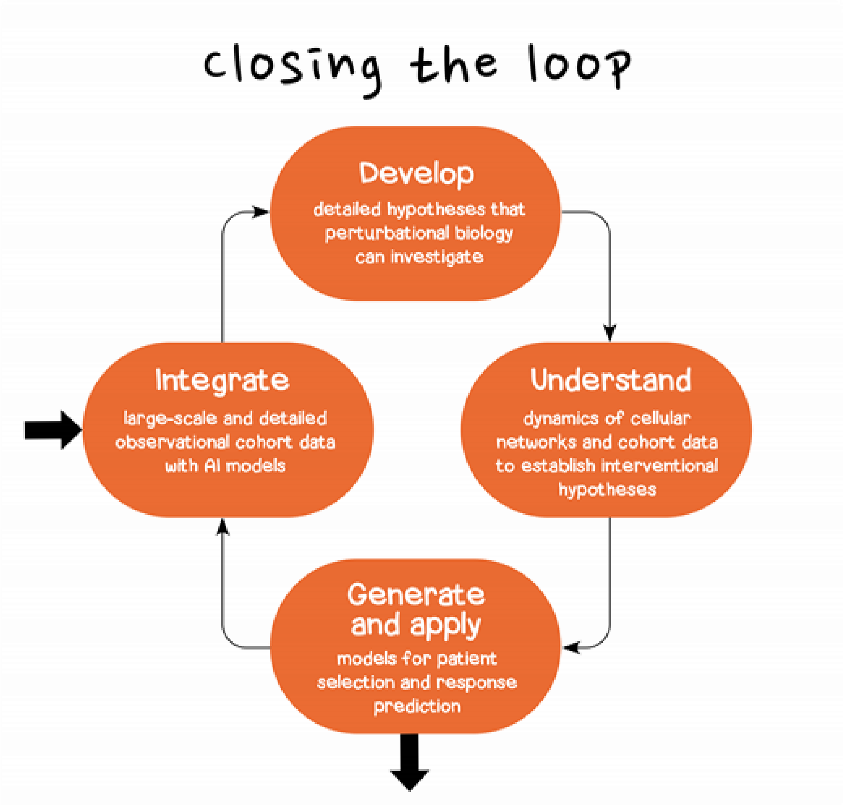
GSK Is Getting Good Value from its AI Investment

Heather McKenzie, *Biospace*, Nov 6, 2023 (excerpt)

An early adopter of AI/ML tools, GSK uses AI to determine the best point at which to intervene in the disease process, for the most patients.

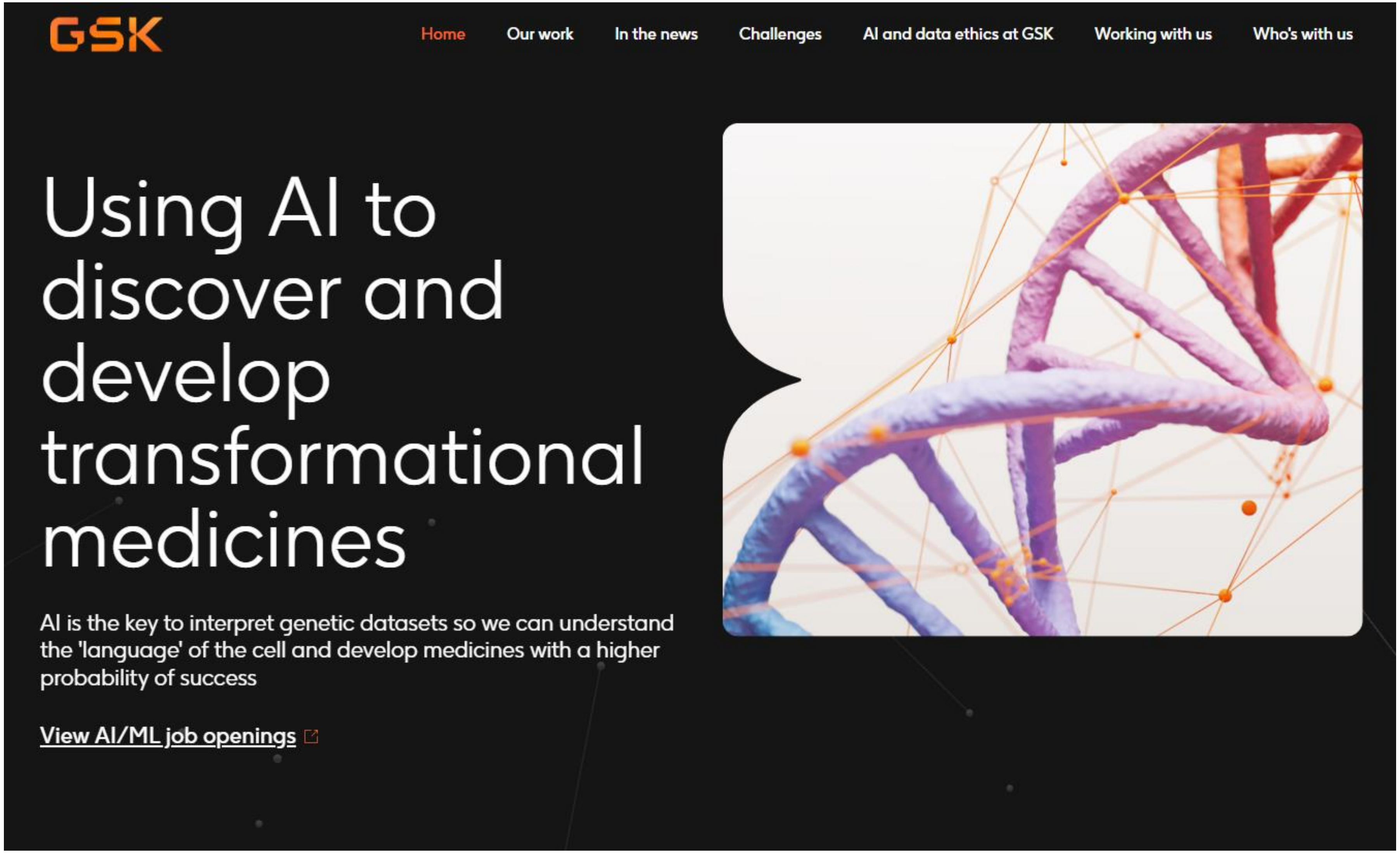
Using AI/ML, GSK has:

1. Built models to continuously monitor and quantify the amount of liver fat from patients' imaging scan data, leading to "a lot more interesting targets" for nonalcoholic steatohepatitis (NASH).
2. Built a machine learning algorithm to stratify hepatitis B patients in a Phase IIb study of bepirovirsen in order to identify which subset showed the greatest response; 9 to 10% responded so well they experienced a functional cure.
3. Developed large language models on RNA and DNA sequences to predict the effect of genetic variants on mRNA processing, identifying whether the amount of protein is increased or decreased and whether the variant changes the splicing pattern of a particular gene.


closing the loop
Develop
detailed hypotheses that perturbational biology can investigate
Understand
dynamics of cellular networks and cohort data to establish interventional hypotheses
Generate and apply
models for patient selection and response prediction
Integrate
large-scale and detailed observational cohort data with AI models

Source: https://www.biospace.com/article/deep-dive-how-four-companies-are-leveraging-ai-in-drug-development-/

GSK.ai is Consolidating GSK's Know-How in AI



Using AI to discover and develop transformational medicines

AI is the key to interpret genetic datasets so we can understand the 'language' of the cell and develop medicines with a higher probability of success

View AI/ML job openings

Focused on discovery of novel medicines starting with target identification.

Source: https://gsk.ai/

GSK’s AIML genomic sequence models for interpreting variant effects

By Stephen Young, Kalin Vetsigian, Jessica Lancaster, Caled Radens, Paul Smyth, Rob Woodruff, David Mittelman, Ferran Gonzalez, Mahmoud Hossam, Aleix Lafita, Daniel Seaton, Ari Allyn-Feuer, Shane Lewin and Kim Branson.

GSK AI/ML Team

2024

“To link variants to downstream molecular-level changes relevant to disease, we must first understand the complex patterns that exist within these sequences. This is becoming an increasing reality as new sequencing technologies enable the generation of genomic data at unprecedented scale and speed – yet the size and complexity of these datasets present both an opportunity and a challenge. The data provides a valuable resource, but its interpretation requires sophisticated computational tools. Beyond traditional approaches such as large-scale statistical methods, AIML is an obvious choice for this task.

In this article we explore how AIML sequence models, particularly those leveraging deep neural networks, can learn complex patterns over long sequences directly from large sequence datasets. We also highlight some of our published efforts in building and adapting genomic foundation models, and describe two models currently being applied at GSK for variant analysis: ExonNet and Seneca.”

Source: https://gsk.ai/blogs/gsk-s-aiml-genomic-sequence-models-for-interpreting-variant-effects/

GSK’s AIML Genomic Sequence Models (Continued)

Which variants are interesting?

Variants of interest arise from large-scale statistical genetics studies. These studies aim to connect variants found in human populations with specific disease states [1] and can lead us to novel druggable targets. At GSK, this includes variants identified through biobanks and other population-level studies with our partners. Our use of AIML also allows us to model rare or indeed arbitrary variants that have not been observed in any specific population, enabling a more comprehensive exploration of genetic landscapes.

Another category of variants of interest are those that we can actively introduce to manipulate the biology of cells in a new way. A good example of this ‘editing’ comes, for example, from certain modalities of antisense oligonucleotide (ASO) drugs. Here RNAs, which either encode proteins or have gene regulatory functions, can be altered at the single-base level to change their function in a way that alleviates disease. Some edits will be useful, others either inert or actively harmful, and we need to know which.

Why is it hard to interpret variant effects?

The consequences of variants in DNA, RNA and protein sequences invariably depend on sequence context, meaning that the sequence on either side of the variant site also influences whether and how the variant affects downstream biology. This can occur across very long-range context – sometimes hundreds of thousands of bases or longer. The longer the relevant sequence context, the greater the number of possible sequence contexts encountered and therefore the difficulty in sequence interpretation grows (e.g. for DNA, this is four – the number of bases – to the power of the sequence length).

Another definition of sequence context is that sequences do not operate in a vacuum. They exist in different cell types hosting different constellations of molecules acting that selectively and cooperatively bind to sequence motifs and trigger downstream processes. The significance of a variant depends on the cell in which it is operating because there is interplay between the modified sequence and a complex cellular apparatus.

How are foundation models being applied to this problem at GSK?

The rapid proliferation of language model AI tools has taken the world by storm. In the genomics space, over the last few years, there has been a steady increase in the development of analogous techniques, BERTs and GPTs [2] trained on biological sequences. We have also seen other models emerge that are not strictly speaking language models, but are foundational in the sense that they are a trunk model that can be repurposed effectively for multiple downstream applications. This general approach is gaining ground as it shows promise for important biological prediction tasks resistant to direct measurement.

There is an important approach emerging in sequence modelling for drug discovery that underlies several of our sequence interpretation tools. This approach has two key elements: self-supervised training of foundation language models on a large corpus of unlabelled data (during which a broad set of sequence contexts will be encountered), followed by transfer learning in which the language model is finetuned on domain-specific labelled data. In our domain, both steps are essential in the creation of specialized predictors of genetic effects, and each comes with its own unique challenges.

Source: https://gsk.ai/blogs/gsk-s-aiml-genomic-sequence-models-for-interpreting-variant-effects/

GSK's AIML Genomic Sequence Models (Continued)

In the rest of this article, we describe some previously unpublished models that are being applied at GSK for genomic variant analysis that exploit this same paradigm. The premise is that models exposed to very large numbers (millions) of molecular sequences during pre-training can learn from a huge space of relevant sequence patterns. Through this, we gain a mechanism to address the large sequence context problem. This enables the construction of high-quality internal representations of sequences that are useful for downstream prediction tasks (relative to a randomly initialized set of weights where everything about sequence patterns and their meanings is learned from a smaller labelled dataset).

ExonNet: predicting tissue-specific alternative splicing directly from sequence. Human gene expression does not simply turn genes on and off to various degrees, but can generate multiple distinct RNA and protein products from the same gene sequence. One of the main cellular processes responsible for this is splicing – as RNA is transcribed from DNA, large chunks of RNA called introns are removed, and the remaining pieces called exons are spliced together.

Most human genes (>95%) can be spliced in multiple ways within a cell or across cell types, producing multiple RNAs and proteins that may be functionally different. For example, an extra retained exon can change the protein sequence in a way that affects its enzymatic activity, localization within the cell or interactions with other proteins. The exon–intron borders and transcription start or termination sites can also shift, further increasing the diversity of products.

The complex splicing process is shaped by interactions between the transcribed RNA and RNA-binding proteins, and as such is very sensitive to variants that alter the RNA sequence. It is estimated that up to 30% of disease-causing variants are splice-altering variants, which makes this process important for interpreting human genetics data and mechanistic understanding of pathogenicity.

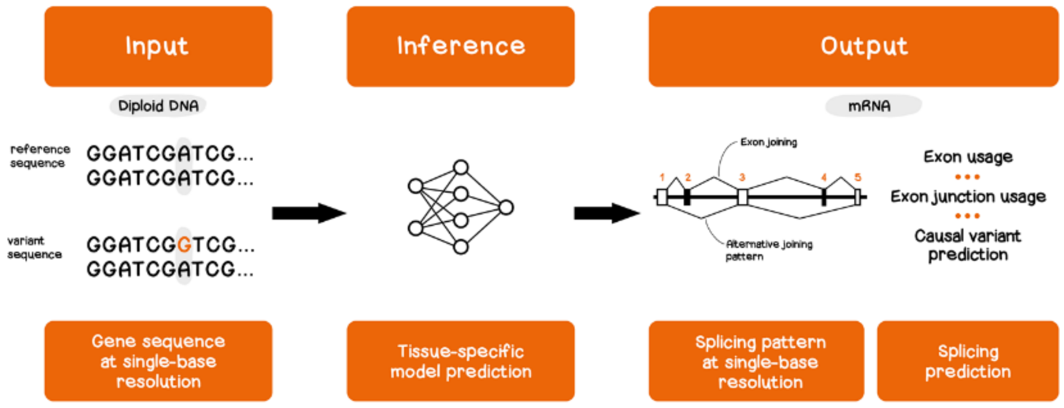
This motivated the development of our ExonNet model – a neural network ensemble trained on the task of predicting the (alternatively) spliced RNA outputs from the DNA sequence of a gene while accounting for how the tissue-specific cellular context affects RNA processing. As the tissue-specific cellular context influences this process, we trained a suite of tissue-specific models. The principal training datasets for ExonNet, including the public Genotype-Tissue Expression (GTEx) dataset, enable mapping from an individual human donor's diploid DNA sequence (whole-genome sequence data) to a tissue-specific RNA expression readout (RNA-seq) in the same donor. This allows the model to learn the tissue-specific effects of DNA variants. Running this model twice, once with a reference sequence and once with an alternative sequence, allows us to predict the effect of new variants on splicing patterns.

ExonNet's detailed splicing outputs at single-base resolution, including explicit splice-site usage and relative usage of alternative splicing junctions, combined with broad tissue specificity separates it from previous strong deep learning contributions in the literature (SpliceAI [3], Pangolin [4] and the recently published Borzoi [5] and BigRNA [6]. An additional machine learning model uses ExonNet predictions to output causal probabilities of variants, turning it into a highly accurate variant effect predictor (VEP) (AUPRC of 0.91 on a balanced causal/non-causal dataset of high confidence sQTLs; 10× higher recall at 0.9 precision than Ensemble's VEP for splicing). ExonNet greatly increases the positive predictive values of variant-to-gene mapping methods when benchmarked with OMIM and PharmaProjects as a source of truth labels.

Source: https://gsk.ai/blogs/gsk-s-aiml-genomic-sequence-models-for-interpreting-variant-effects/

GSK’s AIML Genomic Sequence Models (Continued)

ExonNet has been applied to predict cell-type-specific splicing effects for hundreds of millions of common and rare variants at GSK, informing target evaluations and fine mapping of disease-causing variants (identifying causal variants hidden amongst groups of statistically entangled contender variants), as well as our oncology and ASO development programs. In particular, an ability to predict the effects of oligonucleotides that modulate splicing and other RNA maturation processes will accelerate the development of oligonucleotide drugs for specific targets.



Source: https://gsk.ai/blogs/gsk-s-aiml-genomic-sequence-models-for-interpreting-variant-effects/

Seneca: providing missing directionality data for drug targets through analysis of variant effects

Early in the target identification process, we frequently face a ‘directionality’ problem. We have succeeded in associating a target (gene product) with a disease, but we do not know whether we should be seeking to increase or decrease the abundance or activity of the target to alleviate the disease. Perhaps surprisingly, this information can be missing in up to 90% of targets under scrutiny. This creates a bottleneck and introduces doubt over whether a target is even tractable under available drug modalities, or if it will lead to a potentially lengthy and expensive scientific dead end.

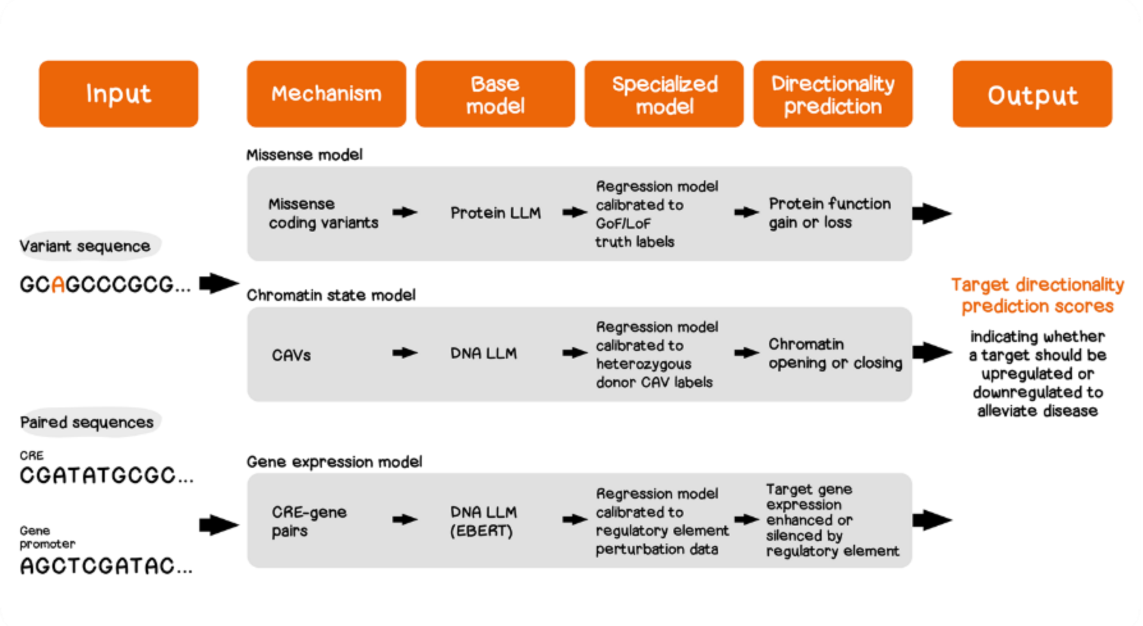
Seneca tackles the directionality question by using a suite of binary predictors that, when used together, tile the space of variant effect mechanisms and handle both coding variants (directly affecting protein sequences) and non-coding variants (acting through long-range regulatory mechanisms that up- or down-regulate a gene’s expression).

This has led to the development of three first-in-class proprietary models, operating across coding and non-coding mechanisms:

1. **Missense model**: Predicts whether an amino acid change to a protein sequence arising due to a ‘missense’ variant leads to a gain or loss in protein function.
2. **Chromatin state model**: Predicts whether a variant in a gene regulatory element leads to a local opening (activation) or closing (inactivation) of DNA.
3. **Gene expression model**: Predicts, for a regulatory element–gene pair of sequences, whether the regulatory element acts to increase or decrease the expression of the gene.

GSK’s AIML Genomic Sequence Models (Continued)

At GSK, Seneca has halved the number of early targets with missing directionality evidence, de-risking targets and the probability of clinical failure.



Input
Mechanism
Base model
Specialized model
Directionality prediction
Output
Missense model
Variant sequence
GCAGCCCGCG...
Missense coding variants
Protein LLM
Regression model calibrated to GoF/LoF truth labels
Protein function gain or loss
Chromatin state model
CAVs
DNA LLM
Regression model calibrated to heterozygous donor CAV labels
Chromatin opening or closing
Target directionality prediction scores
indicating whether a target should be upregulated or downregulated to alleviate disease
Paired sequences
CRE
CGATATGCGC...
Gene promoter
AGCTCGATAC...
Gene expression model
CRE-gene pairs
DNA LLM (EBERT)
Regression model calibrated to regulatory element perturbation data
Target gene expression enhanced or silenced by regulatory element

Fig. 2. Seneca encompasses three separate models. Predictions can be used individually or combined to predict the desired directionality of effect for a potential target. CRE, cis-regulatory element; CAV, chromatin-altering variants; GoF, gain of function; LoF, loss of function.

Source: https://gsk.ai/blogs/gsk-s-aiml-genomic-sequence-models-for-interpreting-variant-effects/

Future of genomic sequence models at GSK

Both ExonNet and Seneca are highly beneficial to our work at GSK. The ability to predict the effects of variants within molecular sequences, also accounting for long-range dependencies, means we can better understand biology and design better drugs.

In each case, the products fill critical gaps in the genetic and biological evidence for target–disease pairs that are infeasible or impossible to fill by experimental approaches. Strong evidence has shown that drug targets with genetic validation are twice as likely to succeed.

Beyond splicing and directionality, there are many more untapped opportunities where we can leverage AIML sequence models to close critical gaps in our understanding of biological sequence properties and activities underlying disease, and to design sequence-based therapeutics.

References

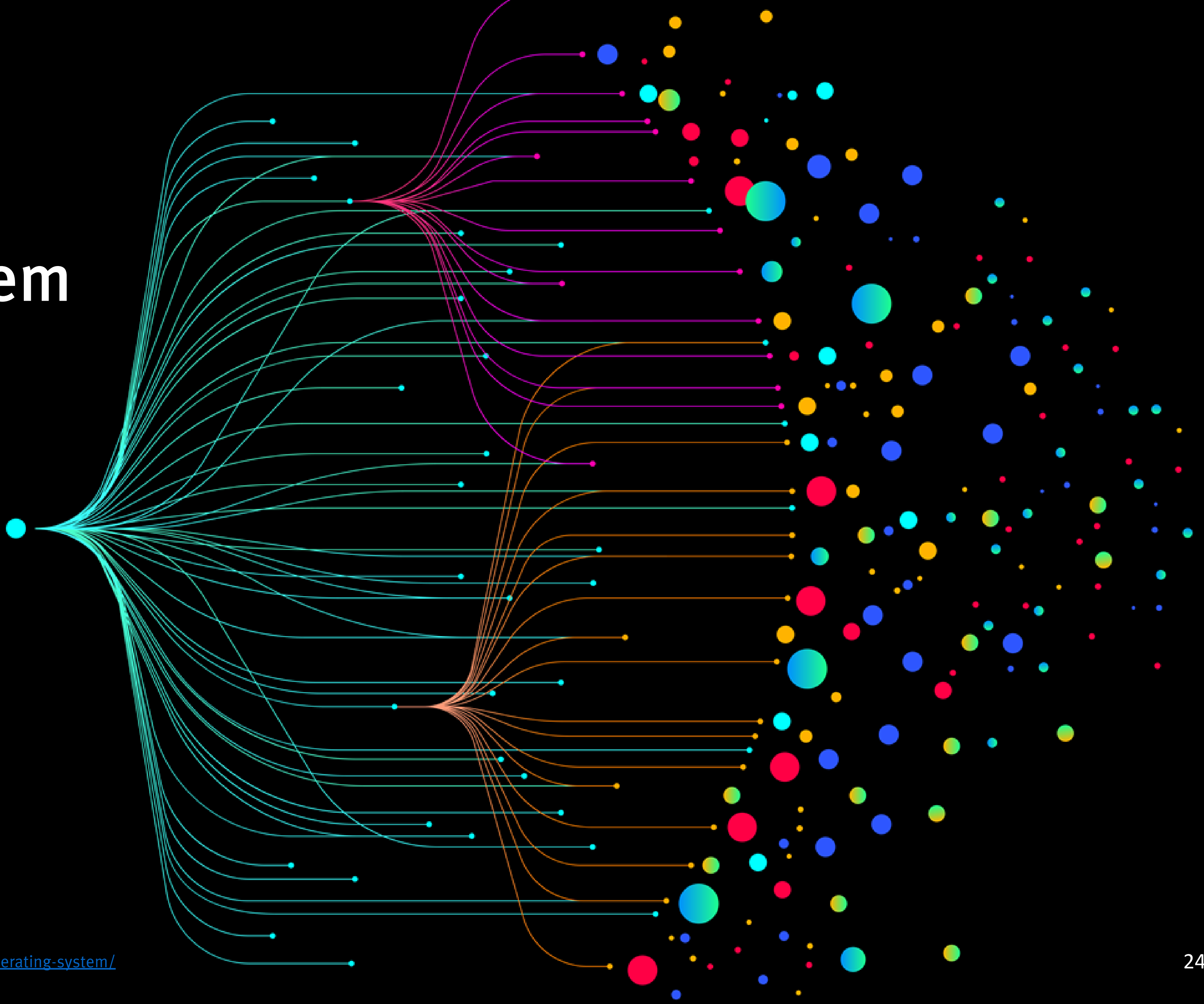
1. Claussnitzerm M. et al. A brief history of human disease genetics. Nature 577, 179–189 (2020)
2. Phuong, M. & Hutter, M. Formal algorithms for transformers. bioRxiv https://doi.org/10.48550/arXiv.2207.09238 (2022).
3. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535–584 (2019).
4. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. 23, 103 (2022).
5. Linder, J. et al. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv https://doi.org/10.1101/2023.08.30.555582 (2023).
6. Celaj, A. et al. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. bioRxiv https://doi.org/10.1101/2023.09.20.558508 (2023).

JulesOS: GSK's Agent-based Operating System

JulesOS has a conversational interface, allowing users to explore complex research questions without understanding the details of GSK's data ecosystem, and without the need to code.

This means GSK staff can find answers for any questions they may have – automatically discovering and extracting relevant data from structured and unstructured sources (such as databases and literature), analyzing and reasoning about that data, and synthesizing the resulting insights into an actionable conclusion.

JulesOS provides a community of agents, from specialized AIML models to more general code creation and analysis tools. This gives everyone at GSK a team of virtual assistants. Together with JulesOS, everyone at GSK can get ahead of disease.

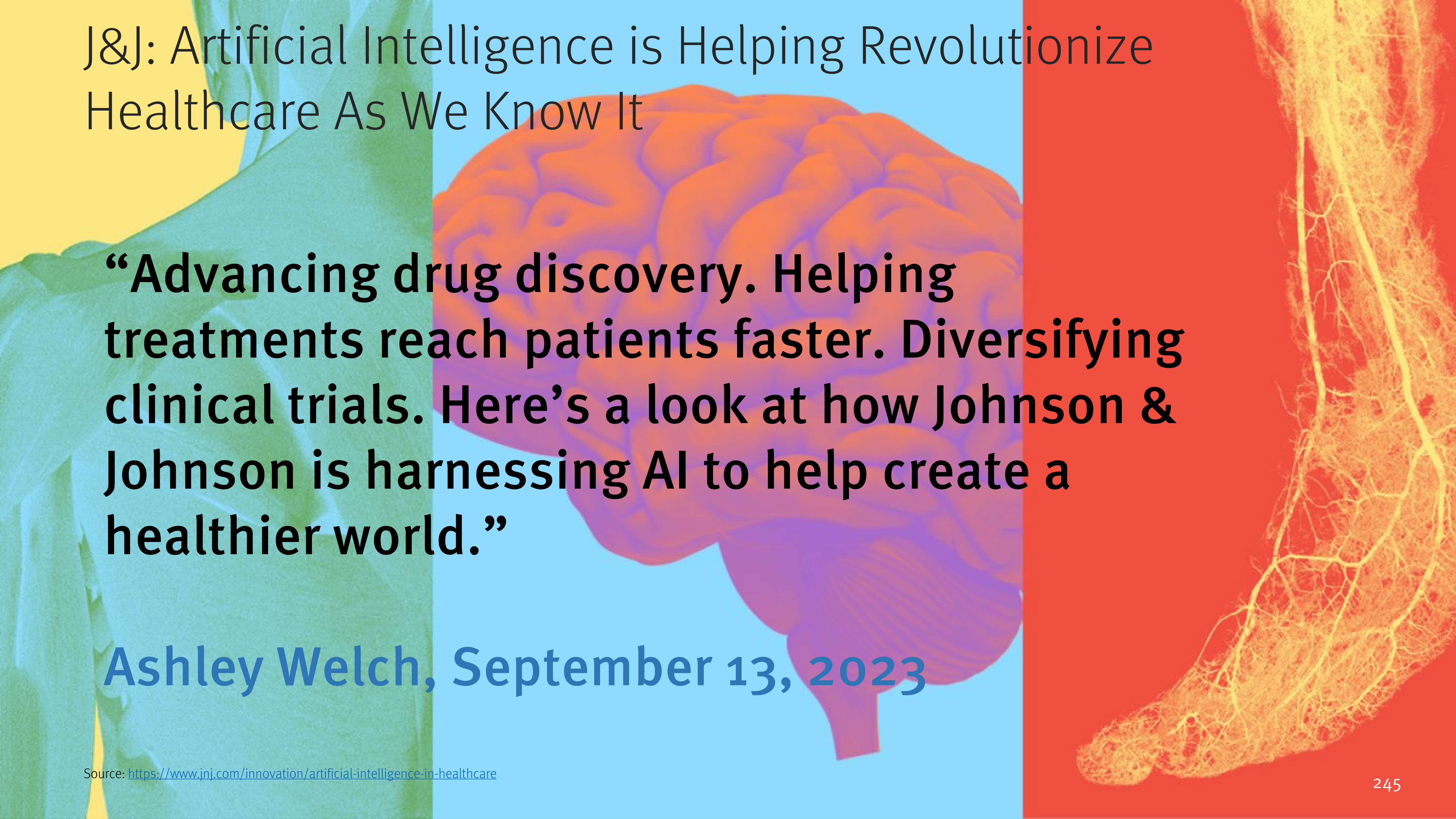


Source: https://gsk.ai/blogs/julesos-gsk-s-agent-based-operating-system/

Johnson&Johnson

Work on Artificial Intelligence and Machine Learning

J&J: Artificial Intelligence is Helping Revolutionize Healthcare As We Know It



“Advancing drug discovery. Helping treatments reach patients faster. Diversifying clinical trials. Here’s a look at how Johnson & Johnson is harnessing AI to help create a healthier world.”

Ashley Welch, September 13, 2023

Source: https://www.jnj.com/innovation/artificial-intelligence-in-healthcare

J&J: AI Revolutionizing Healthcare (continued)

In a time when chronic diseases are on the rise and people are living longer than ever before, novel solutions for better patient care are urgently needed. In healthcare's next chapter, a new type of technology will play a bigger role than ever before.

Enter artificial intelligence, or AI. Rooted in the simulation of human intelligence by computer systems and machines, AI has the potential to transform how humans learn, work and interact with one another in every aspect of life.

It's also primed to revolutionize healthcare.

"The rapid growth in available healthcare-related data in recent years allows us to ask bigger questions," says Jeff Headd, Vice President, Commercial Data Science, Janssen North America Business Technology. "Using the latest innovations in AI and machine learning (ML), we are able to quickly analyze these vast datasets (including electronic medical records, lab results or even medical imaging like X-rays, MRIs and CT scans), uncover new insights and then drive actions with real potential to improve patient outcomes."

The promise that AI holds is why Johnson & Johnson is actively using the technology in different ways, from speeding up the process of discovering new medicines to helping surgeons analyze the results of procedures. It's also why, during this year's South by Southwest conference, the company hosted a panel about AI's role in transforming healthcare.

"There's a deep demand for solutions in the healthcare space," says Shan Jegatheeswaran, Global Head of MedTech Digital, Johnson & Johnson, who spoke on the panel. "But it's important to remember that the most sophisticated thing in the clinical workflow is still the human brain. The role of AI is to augment a human decision or action in a way that improves speed, quality or both."

Source: https://www.jnj.com/innovation/artificial-intelligence-in-healthcare

J&J: AI Revolutionizing Healthcare (continued)

Driving Drug Discovery

Traditionally, discovering and developing new drugs to treat disease is a long and complex undertaking, but AI is primed to help accelerate this process.

To develop medicines, researchers need to understand what biological and genetic variations cause diseases to develop. By applying AI to anonymized medical datasets, such as electronic health records or lab results, scientists can fill in missing information as to what causes those diseases.

AI is also enabling researchers to develop more targeted medicines, driving progress toward precision medicine. For example, in oncology an AI algorithm can be applied to digitized images of biopsies to help identify subtle differences between tumors, pointing to the presence of genetic mutations in a subset of patients. Researchers can use these findings to develop medicines specifically designed for that subset of patients. Those same algorithms that can help identify genetic mutations could then be used to find these patients in the real world to facilitate clinical trial recruitment and clinical decision-making.

“Drug discovery is an extremely challenging process with only a small percentage of lead compounds moving into clinical trials and an even smaller percentage becoming approved medicines,” says Chris Moy, Scientific Director, Oncology, Data Science & Digital Health, R&D, Janssen. “AI is not only helping us identify the right targets for complex diseases, but it’s also helping us design fit-for-purpose molecules to treat diseases and optimize them to provide targeted treatment to the disease while also reducing the impact of side effects.”

Together, these applications of AI will help researchers place the most promising candidate drugs into clinical development, with the ultimate goal of improving the probability of successfully bringing a drug to market and rapidly getting new treatments to patients who need them the most.

Source: https://www.jnj.com/innovation/artificial-intelligence-in-healthcare

J&J: AI Revolutionizing Healthcare (continued)

Enabling more targeted clinical trial recruitment

One of the biggest challenges when it comes to running clinical trials is quickly and efficiently recruiting and enrolling patients that meet the selection criteria. Adopting AI technology into the process may help solve this problem.

Example: At Johnson & Johnson, researchers are applying AI and ML algorithms to large anonymized datasets to identify and locate clinical research sites with patients who could potentially benefit from medicines being studied. The clinical trial operations team can then work to determine the likelihood of enrolling the newly identified sites into their trials.

“Historically, many clinical trials have largely taken place at major academic medical centers, but we know that not all patients have access to these centers,” says Nicole Turner, Senior Director of Global Development, Data Science & Digital Health, R&D, Janssen. “Our goal is to leverage the power of AI to bring trials to more patients, rather than waiting for patients to come to us.”

Data and AI are also helping researchers diversify clinical trials, as advanced analytics are finding locations and healthcare institutions where diverse patients are more likely to be treated. Researchers can then prioritize recruiting eligible patients from those study sites into clinical trials. This is critical, given the importance of ensuring medicines are studied in diverse patient populations representative of those impacted by diseases.

> “
> Our goal is to leverage the power of AI to bring trials to more patients, rather than waiting for patients to come to us.”
>
> Nicole Turner, Senior Director of Global Development, Data Science & Digital Health, R&D, Janssen

Source: https://www.jnj.com/innovation/artificial-intelligence-in-healthcare

J&J Hired Thousands of Data Scientists. Will The Strategy Pay Off?

The 137-year-old pharmaceutical and medical-device company is taking a new direction in its efforts to discover drugs

Peter Loftus, *Wall Street Journal*, November 30, 2023

Johnson & Johnson is making one of the biggest bets in the healthcare industry on using data science and artificial intelligence to bolster its work.

The 137-year-old pharmaceutical and medical-device company has hired 6,000 data scientists and digital specialists in recent years, and spent hundreds of millions of dollars on their work, such as using machines to scour massive health-record datasets. Last year the company opened a state-of-the-art research site near San Francisco that houses advanced data science.

Some early efforts focus on diagnostics, like an algorithm that analyzes heart tests to spot a deadly type of high blood pressure much sooner than humans can, and voice-recognition technology to analyze speech for early signs of Alzheimer's disease. There's a virtual-reality goggle set to help train surgeons on procedures like knee replacements.

The long game, though, is a goal that has seen a lot of hype but less concrete proof that it will become a reality: using AI for drug discovery.

Startup biotechs are in the early days of human testing of AI-discovered drugs. Google this year introduced cloud-based AI tools to assist drugmakers in finding new treatments. But it could still be years before an AI-discovered drug is approved for sale by regulators.

Some pharmaceutical leaders have expressed skepticism that AI could ever discover new drugs any better than humans can.

J&J says it has an edge: a massive database called med. AI that it can sift for patterns to help speed up drug development. The info includes "real-world data"—anonymized information collected from everyday patient visits to doctors and hospitals—and years of clinical-trial results.

Source: https://www.wsj.com/tech/biotech/johnson-johnson-hiring-data-scientists-ai-ccbf2c07

J&J Data Science Story (continued)

“AI and data science are going to be the heart of how we are transforming and innovating,” says Najat Khan, chief data science officer and global head of strategy and operations for J&J’s pharmaceutical research unit. “The amount of data is increasing, the algorithms are getting better, the computers are getting better.”

J&J says it has already used machine learning to help design an experimental cancer drug that is scheduled to start human testing next year.

A few things make J&J’s effort different, Khan says. Its data-science workers are tightly integrated into the company’s strategic decisions on drug research. The company’s massive datasets—med. AI has more than three petabytes of information—are made available to tens of thousands of employees. And it has hired people who aren’t just data scientists but also have skills in chemistry, biology or drug development.

Khan, who has a doctorate in organic chemistry, joined J&J in 2018 after working for Boston Consulting Group advising drugmakers on research and development strategies. She was tapped to spearhead the use of data science in the pharmaceutical R&D operation, and also works alongside the scientists.

Analysts consider J&J to be one of the most active large drugmakers in its commitment to AI. Market intelligence firm CB Insights recently ranked it third of 50 companies in its Pharma AI Readiness Index, which tracks companies’ patent applications, investments, dealmaking and other efforts related to AI.

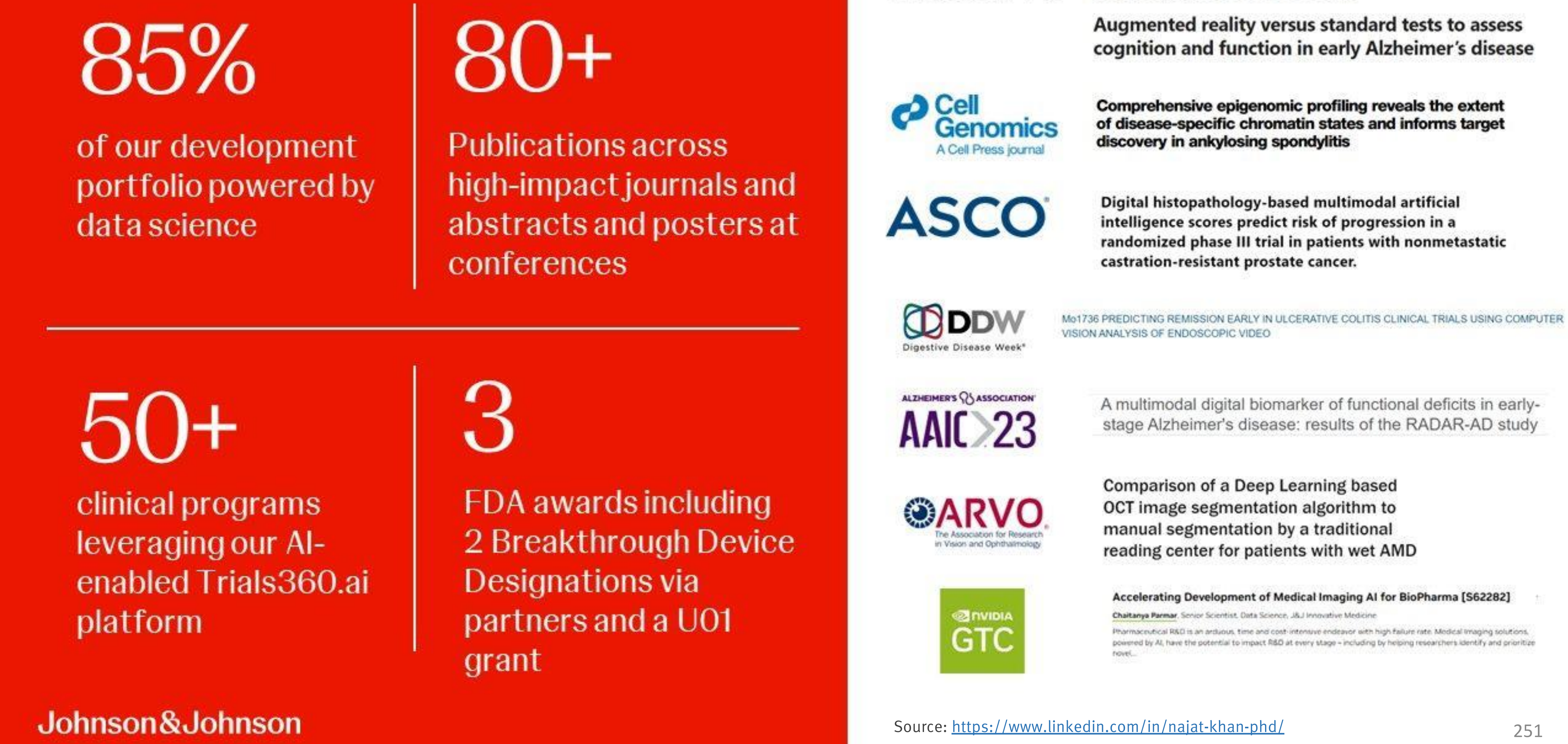
J&J’s sprawling business—more than 130,000 employees and $80 billion in annual global sales—has had data-based projects for years, but the company’s leaders began to take a more coordinated approach about a decade ago, and ramped up investments around four years ago.

Today, most of the company’s drug-development projects incorporate some aspects of data science, up from just a handful five years ago. Its San Francisco-area research site in Brisbane, Calif., places data-science projects alongside R&D focused on finding treatments for retinal and infectious diseases. Many of J&J’s data workers are spread across multiple company locations including in the U.S., China and Belgium.

One hallmark of J&J’s approach is collaborations—more than 50 external partnerships with data-science startups and others. “They seem to be making more investments in other companies, startups and initiatives, in more of a venture-feeling sort of way than some of the other life-sciences firms,” says Daniel Faggella, CEO and head of research at Emerj Artificial Intelligence Research, a Boston firm that conducts market research on corporate use of AI.

Source: https://www.wsj.com/tech/biotech/johnson-johnson-hiring-data-scientists-ai-ccbf2c07

J&J Accomplishments in Data Science, Apr 2023 to Mar 2024



85%

of our development portfolio powered by data science

80+

Publications across high-impact journals and abstracts and posters at conferences

50+

clinical programs leveraging our AI-enabled Trials360.ai platform

3

FDA awards including 2 Breakthrough Device Designations via partners and a U01 grant

Johnson&Johnson

nature

Plasma proteomic associations with genetics and health in the UK Biobank

Augmented reality versus standard tests to assess cognition and function in early Alzheimer's disease

Cell Genomics A Cell Press journal

Comprehensive epigenomic profiling reveals the extent of disease-specific chromatin states and informs target discovery in ankylosing spondylitis

ASCO

Digital histopathology-based multimodal artificial intelligence scores predict risk of progression in a randomized phase III trial in patients with nonmetastatic castration-resistant prostate cancer.

DDW Digestive Disease Week®

Mo1736 PREDICTING REMISSION EARLY IN ULCERATIVE COLITIS CLINICAL TRIALS USING COMPUTER VISION ANALYSIS OF ENDOSCOPIC VIDEO

ALZHEIMER'S ASSOCIATION AAIC 23

A multimodal digital biomarker of functional deficits in early-stage Alzheimer's disease: results of the RADAR-AD study

ARVO The Association for Research in Vision and Ophthalmology

Comparison of a Deep Learning based OCT image segmentation algorithm to manual segmentation by a traditional reading center for patients with wet AMD

NVIDIA GTC

Accelerating Development of Medical Imaging AI for BioPharma [S62282]

Chaitanya Parmar, Senior Scientist, Data Science, J&J Innovative Medicine

Pharmaceutical R&D is an arduous, time and cost-intensive endeavor with high failure rate. Medical imaging solutions, powered by AI, have the potential to impact R&D at every stage – including by helping researchers identify and prioritize novel...

Source: https://www.linkedin.com/in/najat-khan-phd/

Lilly

Work on Artificial Intelligence and Machine Learning

The CEO of Pharma Giant Eli Lilly Shares 3 Ways AI Could Transform His Industry

***Business Insider*, June 15, 2023 (excerpt)**

A handful of biotech companies are testing AI-developed drugs in people. Meanwhile, digital-health companies, providers, and insurers are grappling with how to use technologies including ChatGPT to speed up tasks such as assessing patients and completing medical notes, while still maintaining the safety and privacy of their patients.

According to David Ricks, the CEO of the pharma giant Eli Lilly, the technology has the potential to upend the industry. Eli Lilly is developing dozens of drugs through clinical trials and expects to bring in more than $30 billion in revenue this year.

Ricks told *Insider* that AI is "one of the most exciting technological moves" he's seen in a long time.

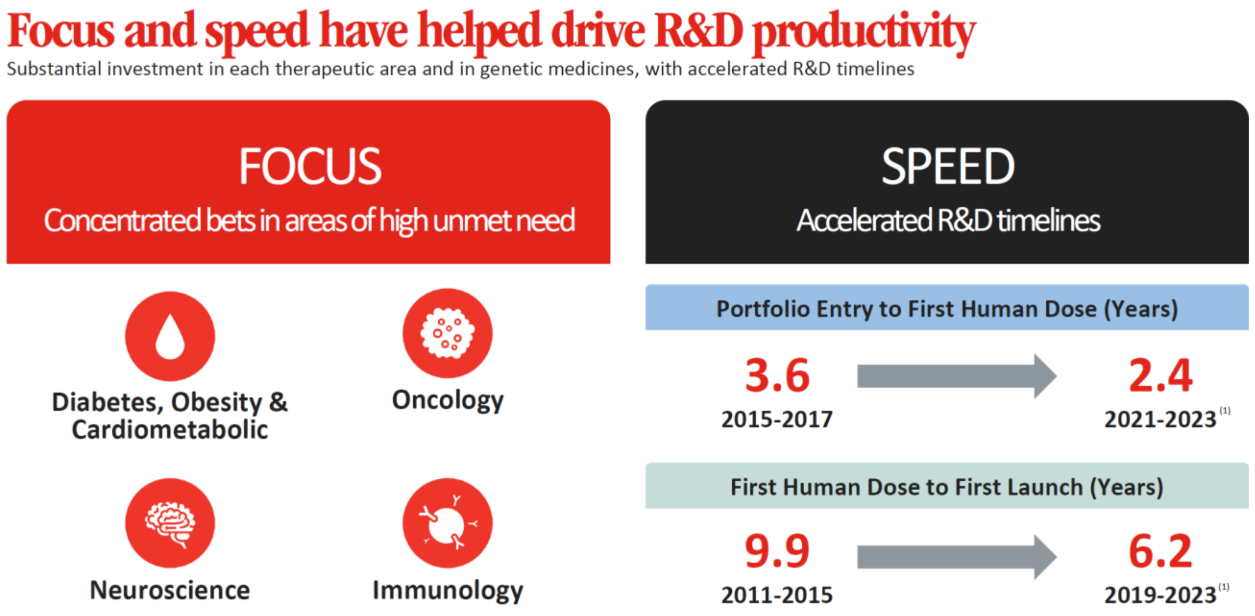
A spokesperson for the company said that **Lilly is investing in artificial intelligence and machine learning in areas including drug discovery, natural-language generation, robotic-process automation, and chatbots.**

The goal is to grow what Lilly calls its "digital worker-equivalent workforce," a concept that the company says helps quantify the hours saved by using technology instead of human labor.

Image from Lilly booth at BIO

Source: https://ca.style.yahoo.com/ceo-pharma-giant-eli-lilly-193606225.html

Lilly R&D Productivity To Date Based on Speed and Focus: Not on AI



Focus and speed have helped drive R&D productivity

Substantial investment in each therapeutic area and in genetic medicines, with accelerated R&D timelines

FOCUS

Concentrated bets in areas of high unmet need

- Diabetes, Obesity & Cardiometabolic
- Oncology
- Neuroscience
- Immunology

SPEED

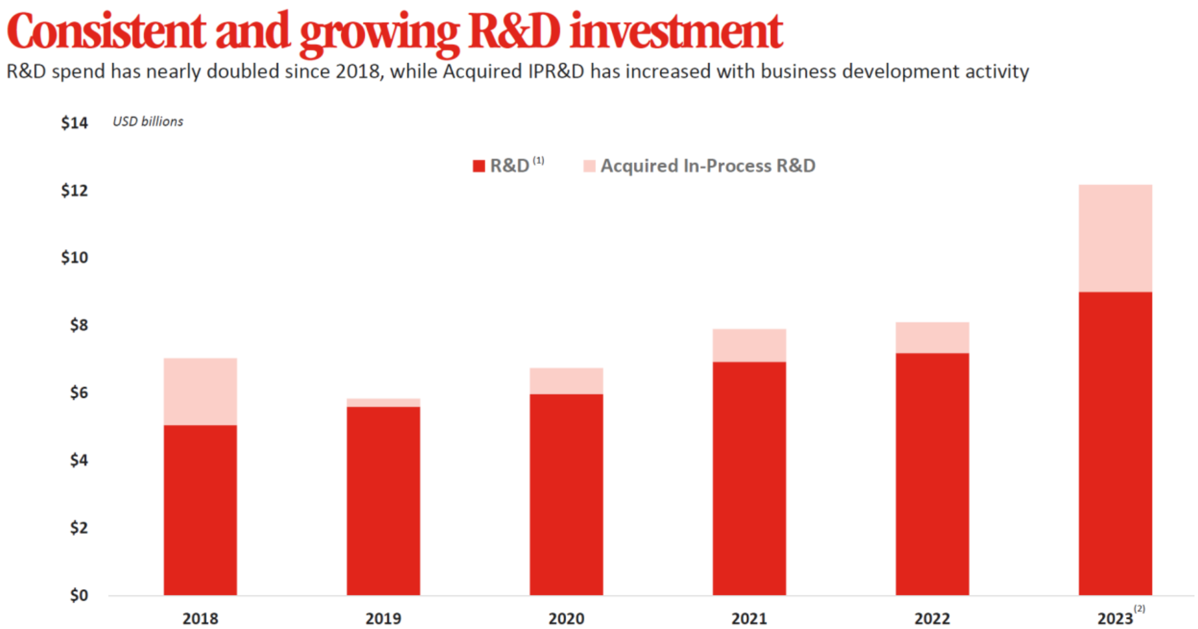
Accelerated R&D timelines

Portfolio Entry to First Human Dose (Years)

3.6 (2015-2017) → 2.4 (2021-2023[1])

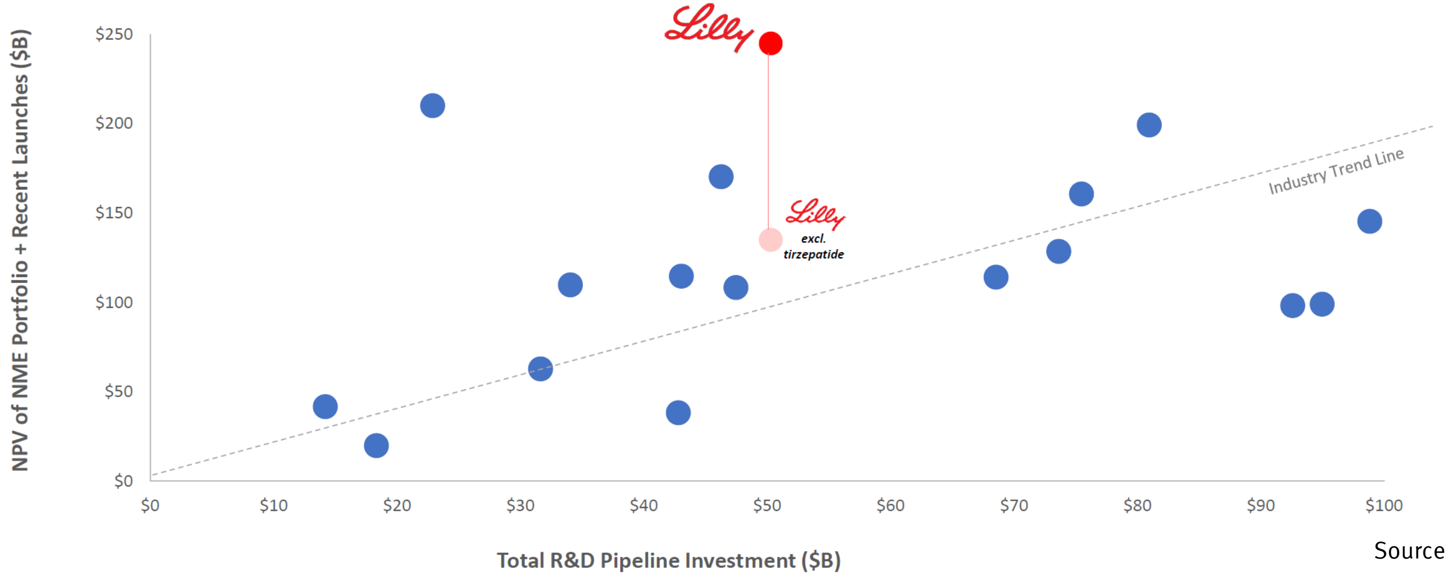
First Human Dose to First Launch (Years)

9.9 (2011-2015) → 6.2 (2019-2023[1])



Consistent and growing R&D investment

R&D spend has nearly doubled since 2018, while Acquired IPR&D has increased with business development activity



Slides from David Ricks Investor Presentation

JP Morgan Healthcare Conference

January 9, 2024

Source: https://investor.lilly.com/static-files/c700e00d-2474-42b0-9a92-6a539961e312 novo nordisk®

Work on Artificial Intelligence and Machine Learning

Novo Nordisk has a foundation to leverage data and AI

Internal and external setup

SELECT

Best-in-class clinical data

within cardiometabolic space

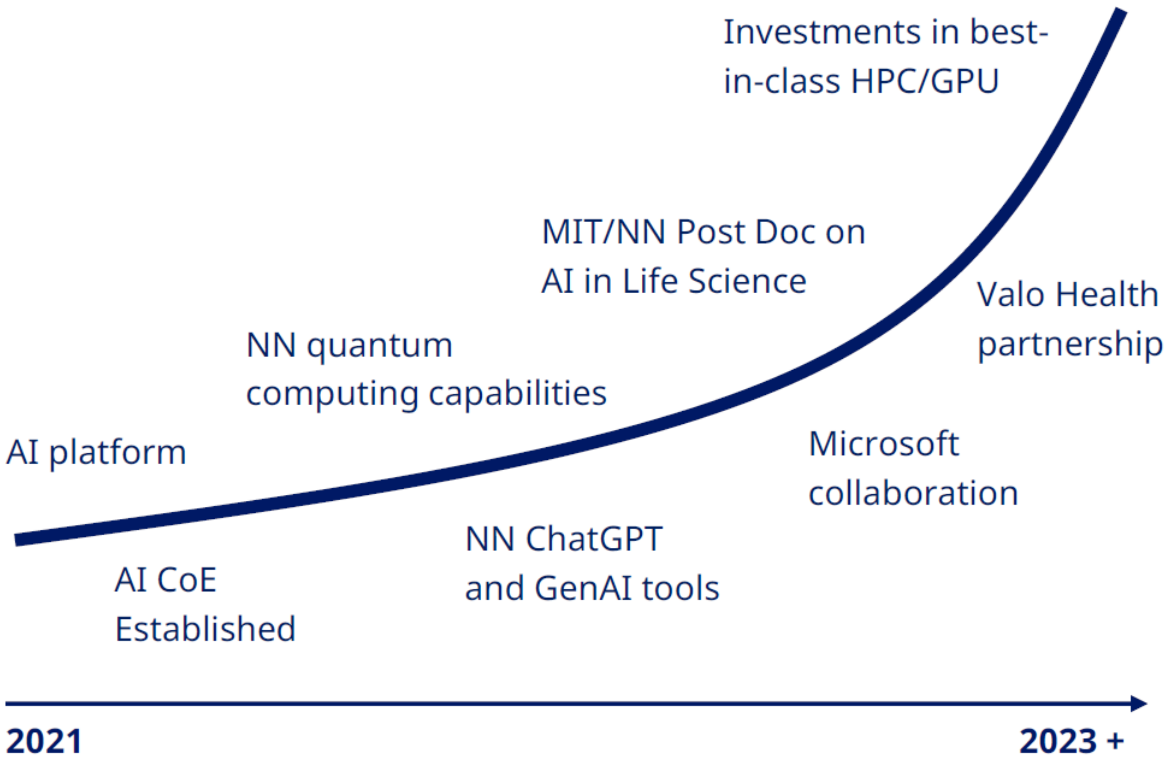
11

Digital Hubs globally

30+

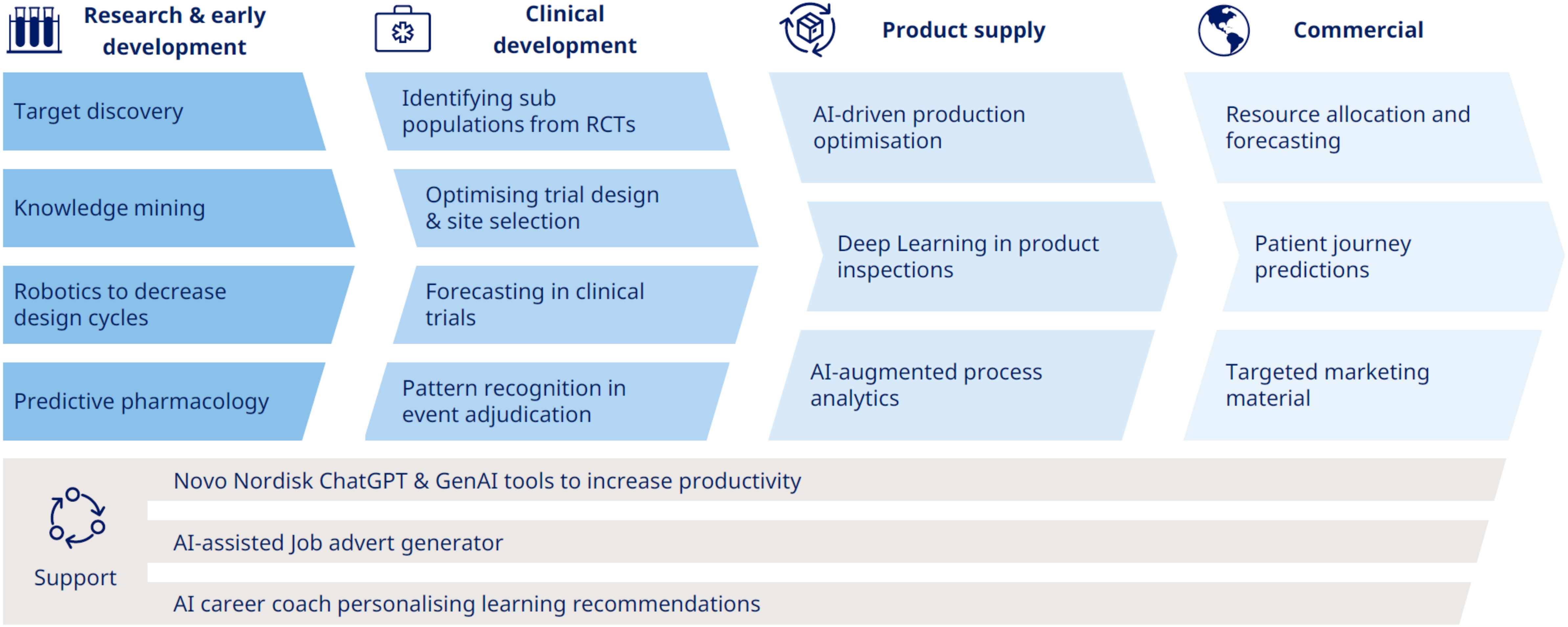
AI-enabling partnerships

Artificial Intelligence is not new to Novo Nordisk


Investments in best-in-class HPC/GPU
MIT/NN Post Doc on AI in Life Science
Valo Health partnership
NN quantum computing capabilities
AI platform
Microsoft collaboration
NN ChatGPT and GenAI tools
AI CoE Established
2021
2023 +

Source: Novo Nordisk Capital Markets Day, March 8, 2024 (https://www.novonordisk.com/investors/capital-markets-day-2024.html)

AI brings innovation across the Novo Nordisk value chain



Research & early development

- Target discovery
- Knowledge mining
- Robotics to decrease design cycles
- Predictive pharmacology

Clinical development

- Identifying sub populations from RCTs
- Optimising trial design & site selection
- Forecasting in clinical trials
- Pattern recognition in event adjudication

Product supply

- AI-driven production optimisation
- Deep Learning in product inspections
- AI-augmented process analytics

Commercial

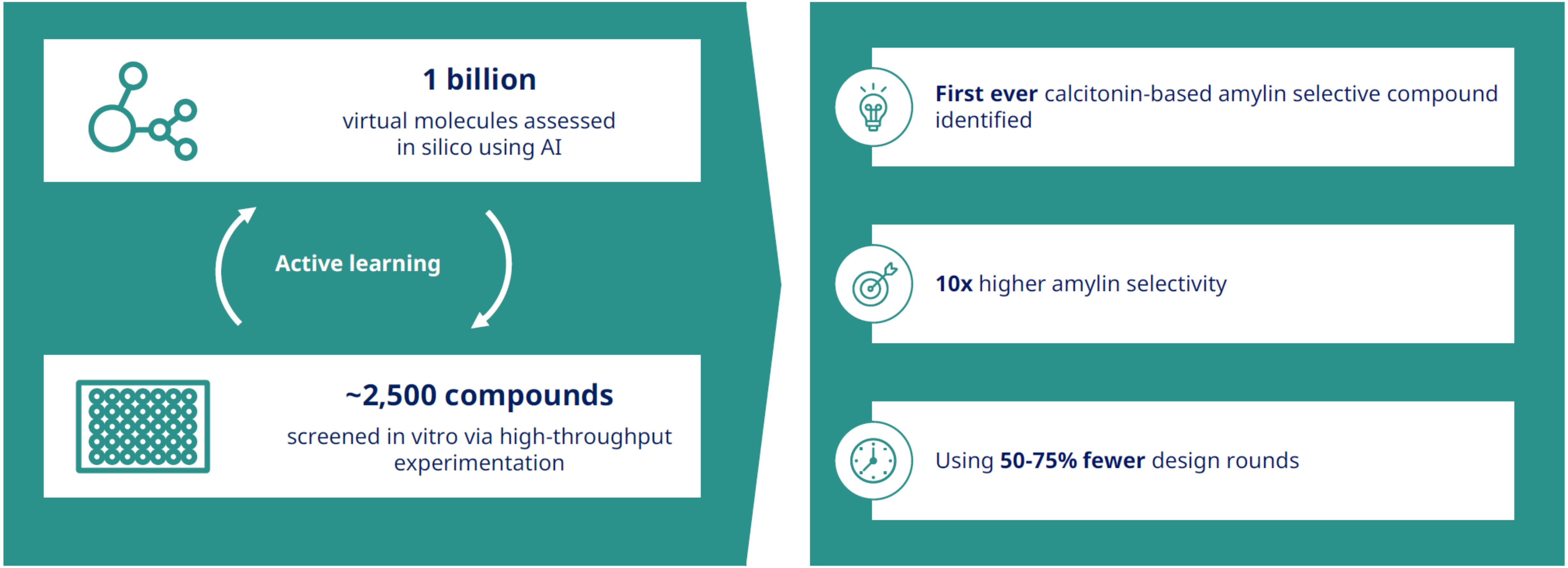
- Resource allocation and forecasting
- Patient journey predictions
- Targeted marketing material

Support

- Novo Nordisk ChatGPT & GenAI tools to increase productivity
- AI-assisted Job advert generator
- AI career coach personalising learning recommendations

Source: Novo Nordisk Capital Markets Day, March 8, 2024 (https://www.novonordisk.com/investors/capital-markets-day-2024.html)

Combining AI with high-throughput experimentation yields first-ever new compound class for amylin selection



Inputs and process

1 billion

virtual molecules assessed in silico using AI

Active learning

~2,500 compounds

screened in vitro via high-throughput experimentation

Realised value

First ever calcitonin-based amylin selective compound identified

10x higher amylin selectivity

Using **50-75% fewer** design rounds

Source: Novo Nordisk Capital Markets Day, March 8, 2024 (https://www.novonordisk.com/investors/capital-markets-day-2024.html)

REGENERON

Work on Artificial Intelligence and Machine Learning

Regeneron R&D Philosophy And AI/ML

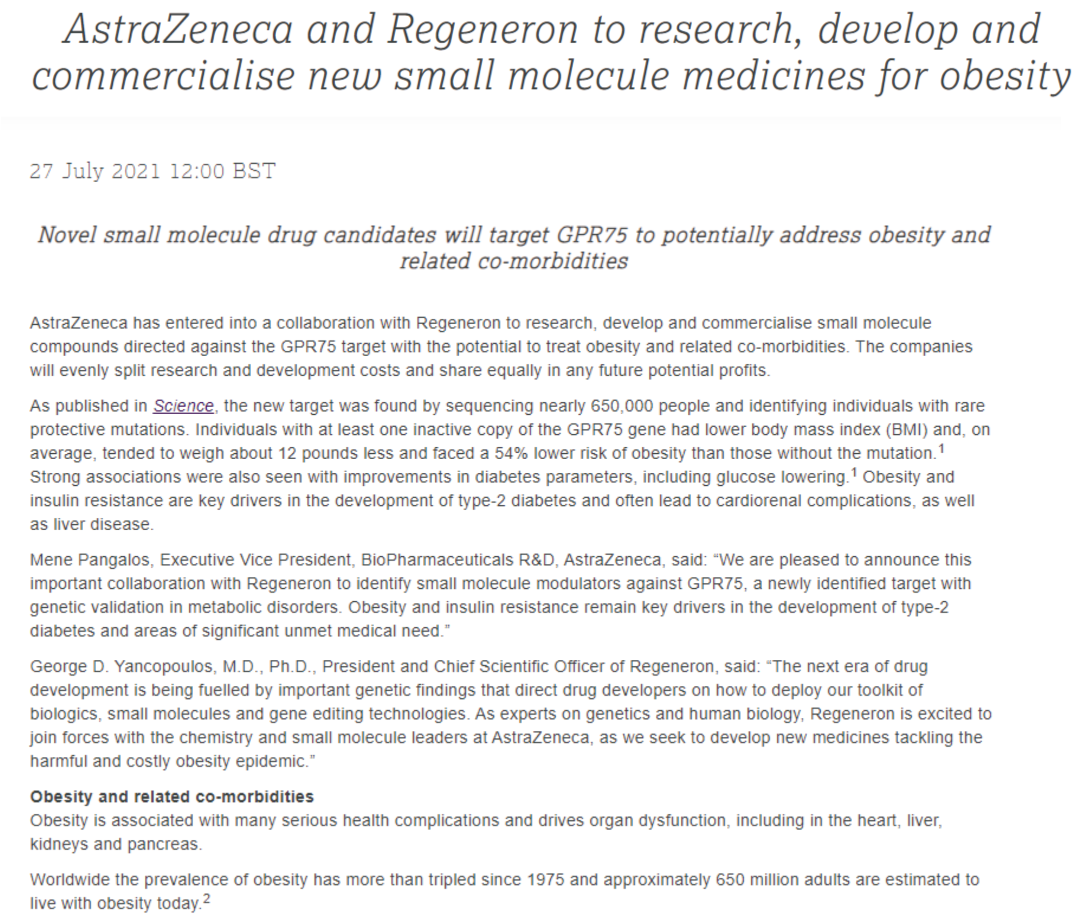
At a time when many pharmas are making big investments in AI and machine learning, Regeneron has been conspicuously relaxed about the topic.

We had the chance to ask a representative of Regeneron about their attitude towards AI at an industry conference in January 2024.

The response was “we don’t think AI is that important.” The speaker explained that “If we can find a novel target for a disease, our scientists have generally been effective at finding a drug.”

It’s important to note that Regeneron has spent 30 years building a biologics generation platform that works from target to disease and then drug design.

It’s worth noting that occasionally Regeneron might make a genetic discovery that calls for a small molecule approach. In the case of its GPR75 discovery, Regeneron partnered up with a company that has excellent capabilities in this area:



AstraZeneca and Regeneron to research, develop and commercialise new small molecule medicines for obesity

27 July 2021 12:00 BST

Novel small molecule drug candidates will target GPR75 to potentially address obesity and related co-morbidities

AstraZeneca has entered into a collaboration with Regeneron to research, develop and commercialise small molecule compounds directed against the GPR75 target with the potential to treat obesity and related co-morbidities. The companies will evenly split research and development costs and share equally in any future potential profits.

As published in *Science*, the new target was found by sequencing nearly 650,000 people and identifying individuals with rare protective mutations. Individuals with at least one inactive copy of the GPR75 gene had lower body mass index (BMI) and, on average, tended to weigh about 12 pounds less and faced a 54% lower risk of obesity than those without the mutation.[1] Strong associations were also seen with improvements in diabetes parameters, including glucose lowering.[1] Obesity and insulin resistance are key drivers in the development of type-2 diabetes and often lead to cardiorenal complications, as well as liver disease.

Mene Pangalos, Executive Vice President, BioPharmaceuticals R&D, AstraZeneca, said: “We are pleased to announce this important collaboration with Regeneron to identify small molecule modulators against GPR75, a newly identified target with genetic validation in metabolic disorders. Obesity and insulin resistance remain key drivers in the development of type-2 diabetes and areas of significant unmet medical need.”

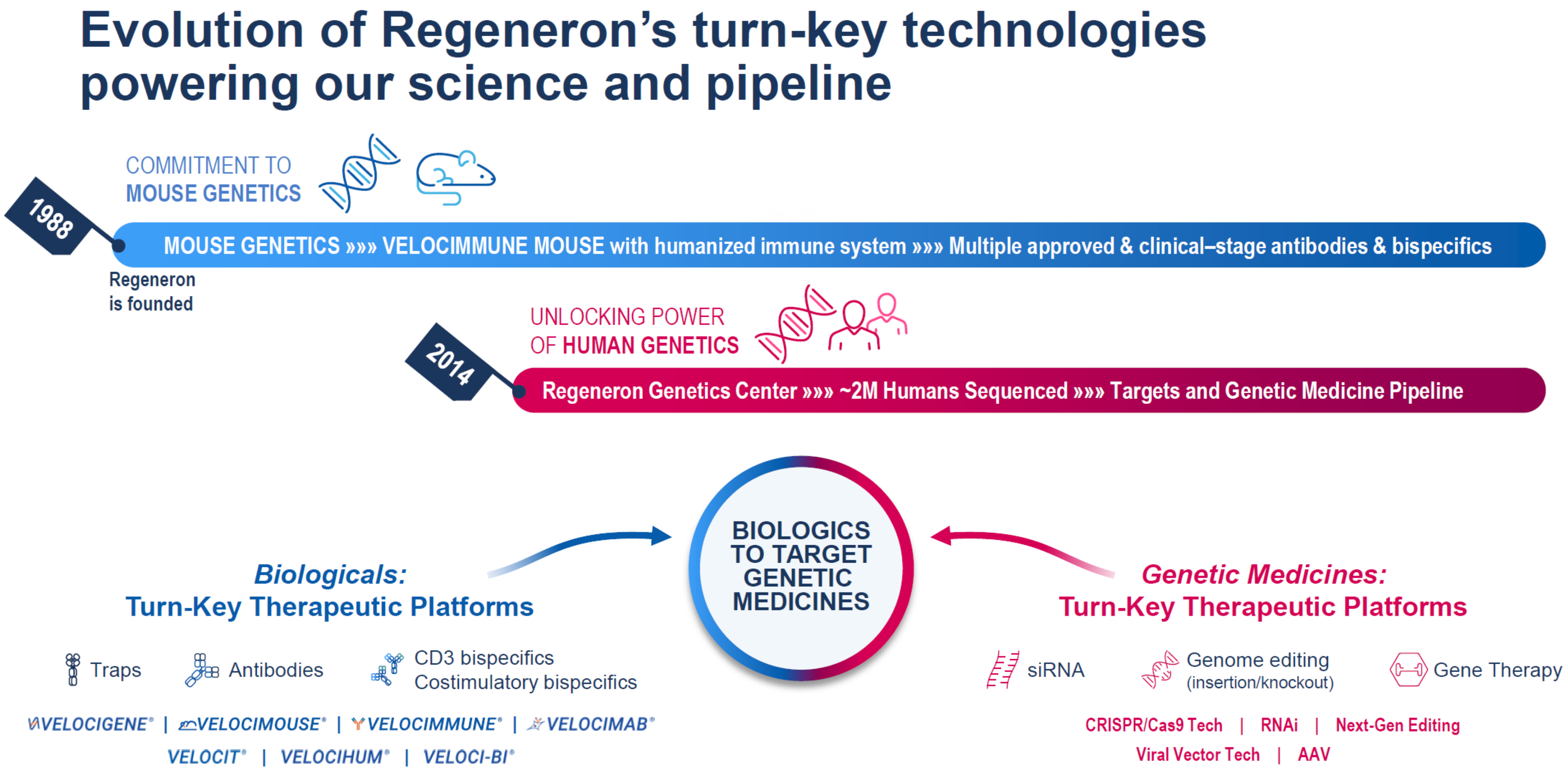
George D. Yancopoulos, M.D., Ph.D., President and Chief Scientific Officer of Regeneron, said: “The next era of drug development is being fuelled by important genetic findings that direct drug developers on how to deploy our toolkit of biologics, small molecules and gene editing technologies. As experts on genetics and human biology, Regeneron is excited to join forces with the chemistry and small molecule leaders at AstraZeneca, as we seek to develop new medicines tackling the harmful and costly obesity epidemic.”

Obesity and related co-morbidities
Obesity is associated with many serious health complications and drives organ dysfunction, including in the heart, liver, kidneys and pancreas.

Worldwide the prevalence of obesity has more than tripled since 1975 and approximately 650 million adults are estimated to live with obesity today.[2]

Source: https://www.astrazeneca.com/media-centre/press-releases/2021/astrazeneca-and-regeneron-to-research-develop-and-commercialise-new-small-molecule-medicines-for-obesity.html#

Regeneron R&D Platform Based on Genetic Target ID and Relevant Therapeutic Platforms



Evolution of Regeneron’s turn-key technologies powering our science and pipeline

MOUSE GENETICS »»» VELOCIMMUNE MOUSE with humanized immune system »»» Multiple approved & clinical–stage antibodies & bispecifics

Regeneron Genetics Center »»» ~2M Humans Sequenced »»» Targets and Genetic Medicine Pipeline

BIOLOGICS TO TARGET GENETIC MEDICINES

Biologicals:
Turn-Key Therapeutic Platforms

Traps | Antibodies | CD3 bispecifics, Costimulatory bispecifics

VELOCIGENE® | VELOCIMOUSE® | VELOCIMMUNE® | VELOCIMAB®
VELOCIT® | VELOCIHUM® | VELOCI-BI®

Genetic Medicines:
Turn-Key Therapeutic Platforms

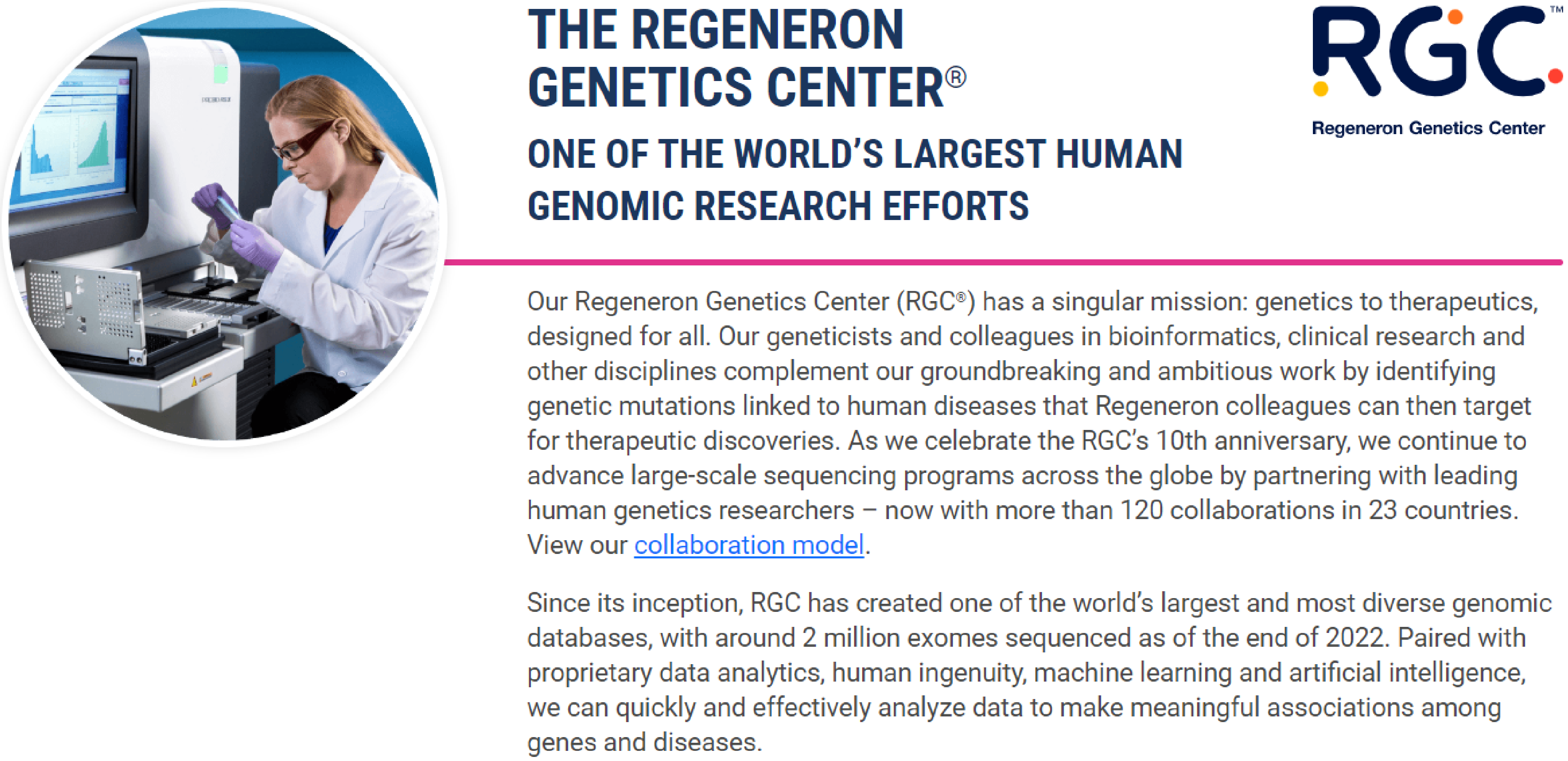
siRNA | Genome editing (insertion/knockout) | Gene Therapy

CRISPR/Cas9 Tech | RNAi | Next-Gen Editing
Viral Vector Tech | AAV

Source: Regeneron Investor Presentation, January 2024

Regeneron is Focused on Target Identification Using Genetic Data

This excerpt from Regeneron’s website makes it clear that Regeneron is not opposed to using AI/ML to study associations between genetic polymorphisms and disease.



THE REGENERON GENETICS CENTER®

ONE OF THE WORLD’S LARGEST HUMAN GENOMIC RESEARCH EFFORTS

Our Regeneron Genetics Center (RGC®) has a singular mission: genetics to therapeutics, designed for all. Our geneticists and colleagues in bioinformatics, clinical research and other disciplines complement our groundbreaking and ambitious work by identifying genetic mutations linked to human diseases that Regeneron colleagues can then target for therapeutic discoveries. As we celebrate the RGC’s 10th anniversary, we continue to advance large-scale sequencing programs across the globe by partnering with leading human genetics researchers – now with more than 120 collaborations in 23 countries. View our collaboration model.

Since its inception, RGC has created one of the world’s largest and most diverse genomic databases, with around 2 million exomes sequenced as of the end of 2022. Paired with proprietary data analytics, human ingenuity, machine learning and artificial intelligence, we can quickly and effectively analyze data to make meaningful associations among genes and diseases.

Source: https://yearinreview.regeneron.com/early-research-and-technology

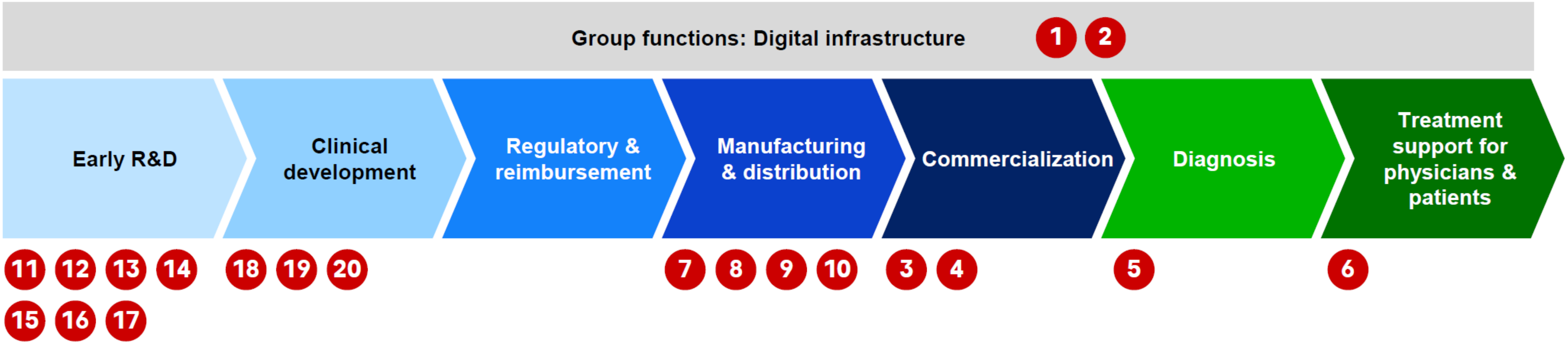
Work on Artificial Intelligence and Machine Learning

Roche is Partnering Heavily to Accelerate AI Adoption

NVIDIA	First-of-its-kind collaboration to accelerate drug discovery using generative AI
Prescient Design A Genentech Accelerator	ML to implement "lab in a loop" for protein drug discovery
Recursion.	AI platform leveraging large-scale, multi-omic perturbation data to generate maps of human biology
Genesis Therapeutics	Neural network algorithms to discover small molecules for challenging targets
sysnav HEALTH CARE	Develop digital endpoints as regulatory approved standards of outcome measurement, and design of next generation wearable technology
Elekta Kaiku	Digital tools for real-time symptom management by patients and HCPs to improve patient support and provide personalized cancer care
ShapeTX	RNA editing capsids platform to generate tissue-specific adeno-associated viruses for gene therapy
DYNO THERAPEUTICS	AI-powered CapsidMap technology to engineer AAV capsids for targeted gene therapy

Non-exhaustive and illustrative overview of deals and partnerships signed over recent years; ML=machine learning

Roche is Implementing AI / Digitalization Across the Enterprise



Group functions: Digital infrastructure

Early R&D

Clinical development

Regulatory & reimbursement

Manufacturing & distribution

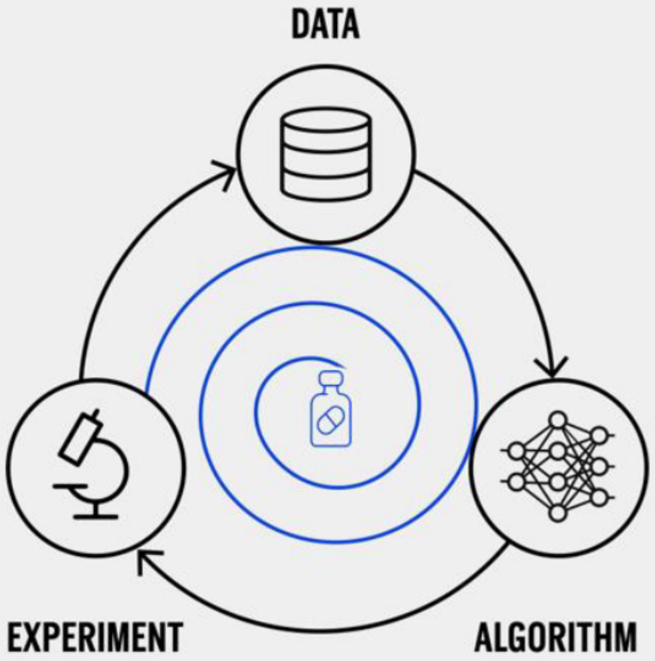
Commercialization

Diagnosis

Treatment support for physicians & patients

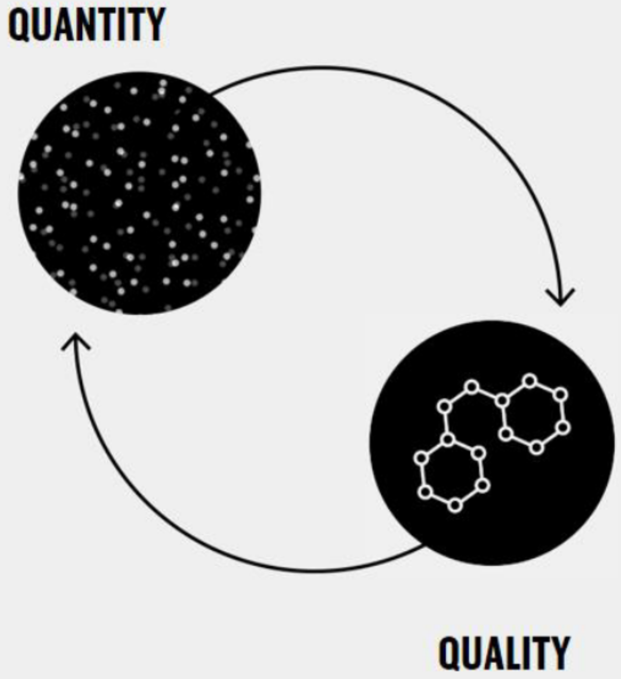
1. GALILEO: Global generative AI & LLM strategy
2. ASPIRE: Creating the digital process backbone
3. nafivy Integrator
4. navify Operational Excellence Applications
5. navify Algorithm Suite
6. Remote patient management solutions
7. Demand forcasting in Diagnostics
8. Supply chain risk management in Diagnostics
9. Manufacturing yield improvements using AI
10. Transportation network optimization
11. SCimilarity: Reverse cell search
12. Massively parallel, high content pooled screens for function
13. Improving small molecule PTS
14. EquiFold: Antibody structure prediction
15. HLApollo transformer-based model: Personalized neoantigen vaccines
16. LLM pillars in research & development
17. Using AI to increase efficiency of small molecule early R&D
18. Leveraging RWD to optimize patient eligibility criteria
19. Utilizing digital biomarkers to support clinical drug development
20. Applying a deep learning algorithm in ophthalmology

Lab in a Loop, integrated



Full stack, across all aspects of R&D; up to "self drive"

Scale & resolution



Maximize benefit of large size: proprietary legacy data and data generation capacity

Leading partnerships and acquisitions

Partnership
Accelerated computing

NVIDIA

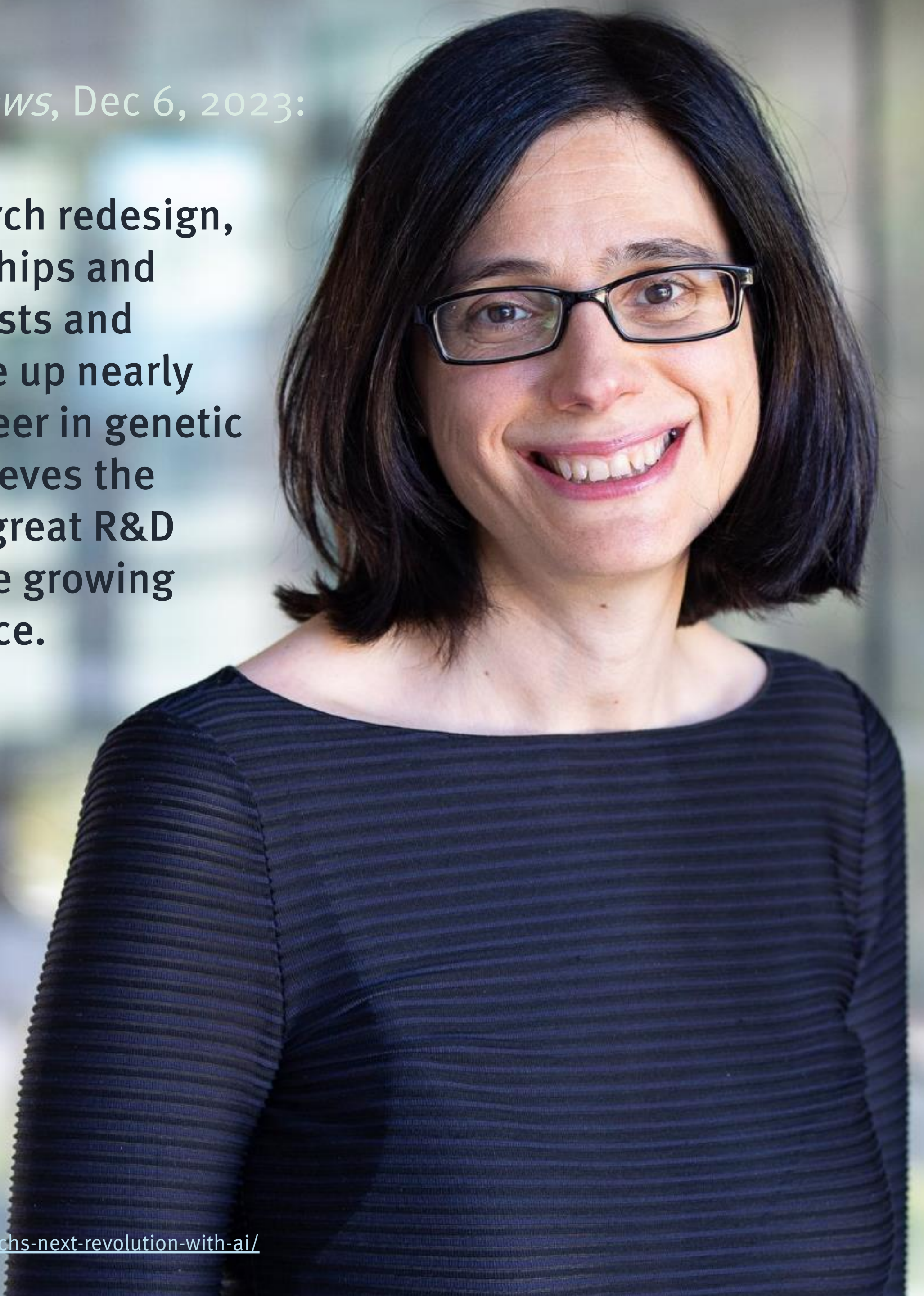
Acquisition

Prescient Design
A Genentech Accelerator

Top capabilities in house, to ensure full loop; Partnership around unique data generation and hardware

Aviv Regev

EVP Genentech Research and Early Development



Andrew Dunn, *Endpoints News*, Dec 6, 2023:

“Regev is leading the research redesign, which focuses as much on chips and compute power as on chemists and compounds. Genentech rose up nearly half a century ago as a pioneer in genetic engineering, and Regev believes the company can lead the next great R&D revolution by harnessing the growing power of artificial intelligence.

Virtually every pharma giant has dipped their toes into AI. But Genentech has plunged in since hiring Regev, poaching academic luminaries, creating labs, and earlier this year establishing a 400-employee computational sciences unit.”

Source: https://endpts.com/aviv-regev-leads-genentechs-next-revolution-with-ai/

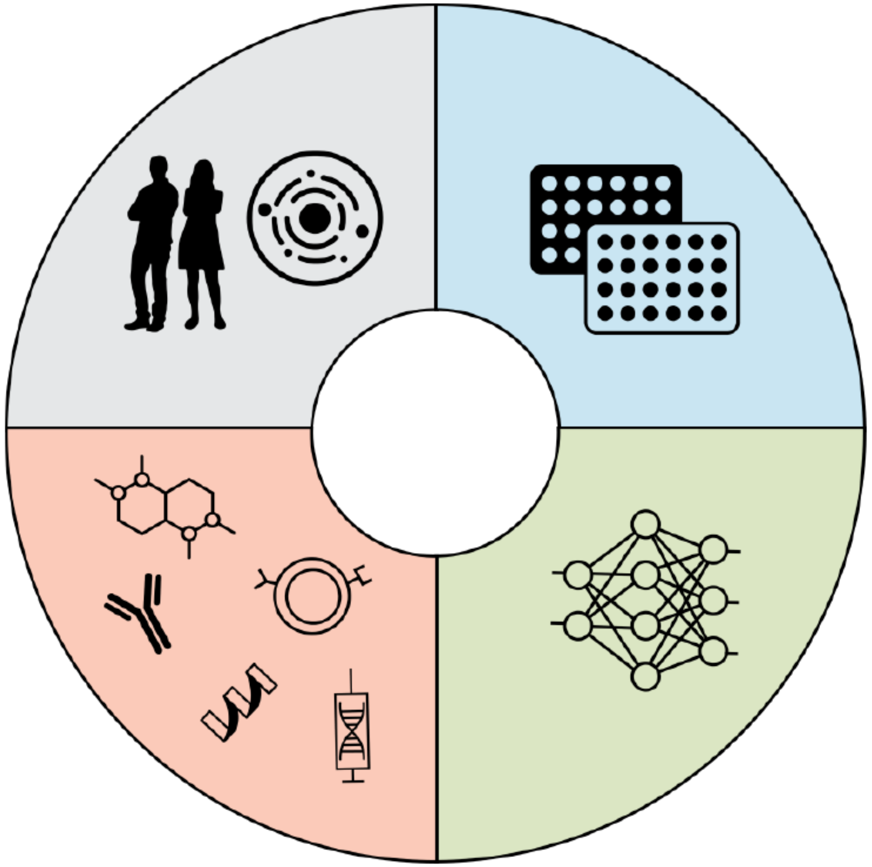
Aviv Regev Talk: Four Multiplicative Levers to Amplify R&D

Human biology

Study disease processes directly in patients or in human derived models

Therapeutic modalities

Expand existing and introduce new modalities to be able to tackle unprecedented targets, with better efficacy and safety



High resolution & massive scale

Do deep and high resolution experiments at extreme scale for only a marginally added cost

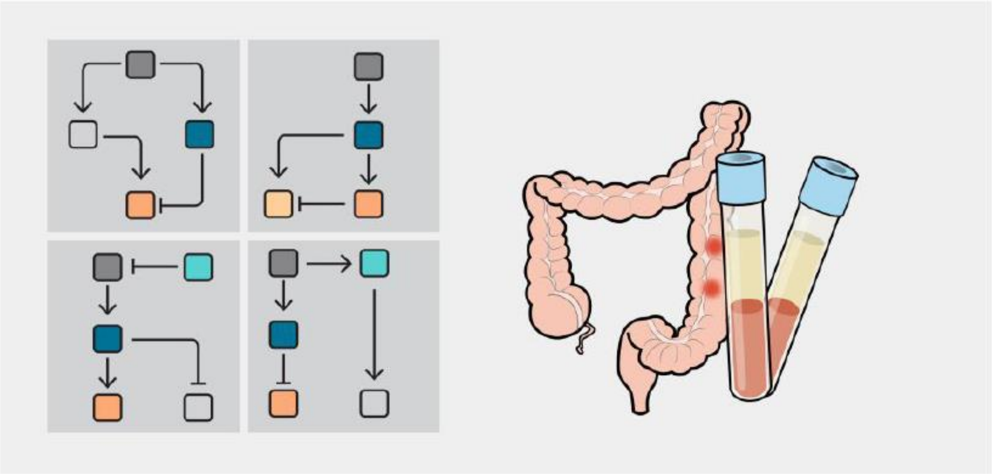
Artificial intelligence / machine learning (AI/ML)

Leverage ML and other advanced computation to make discoveries we would not make otherwise, predict outcomes better, and increase capacity and speed

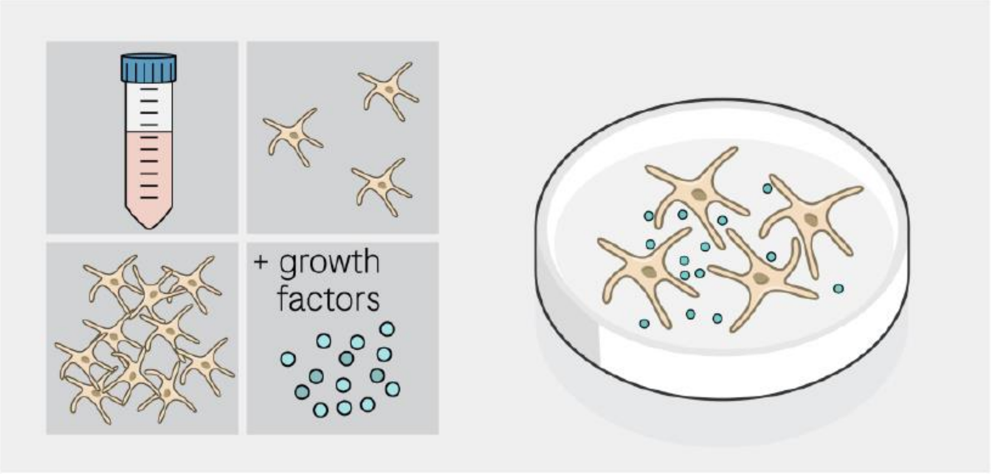
Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 49.

Genentech: Finding Disease Cells and Targets Essential Across R&D

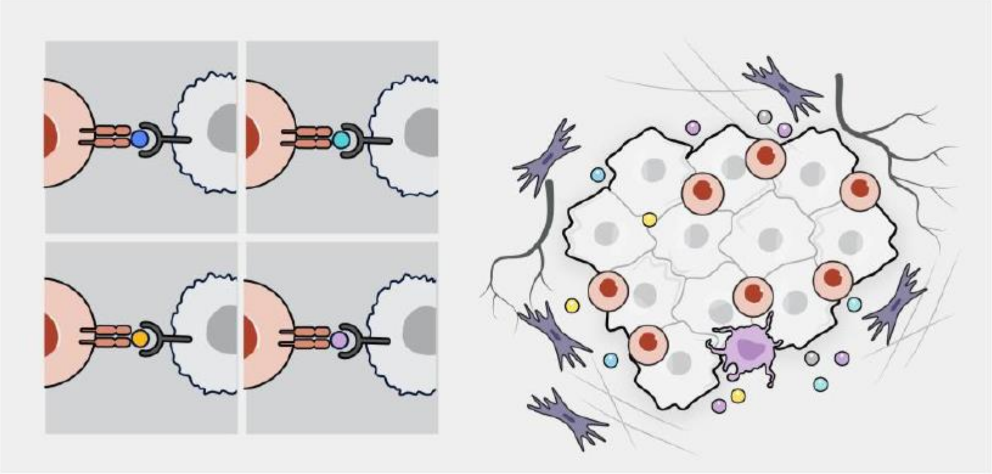
Primary and additional disease indications



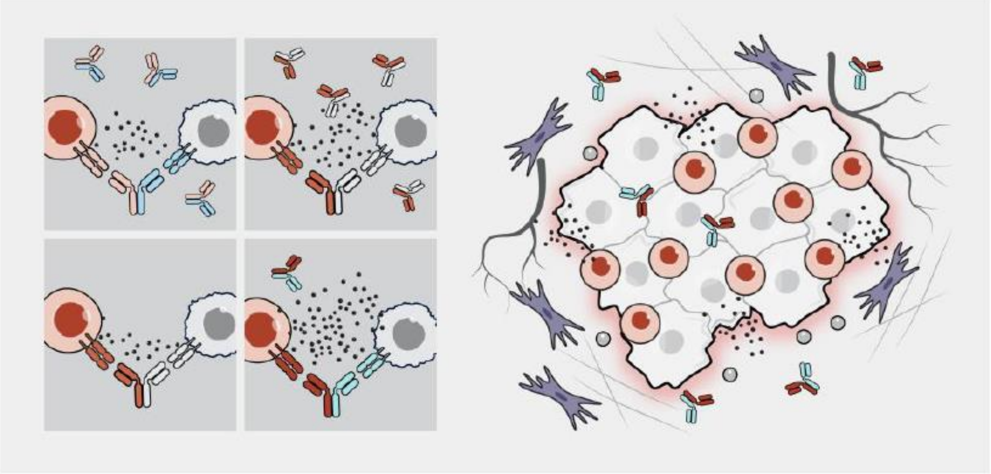
Growth conditions for screens



Bi/multi specifics targets

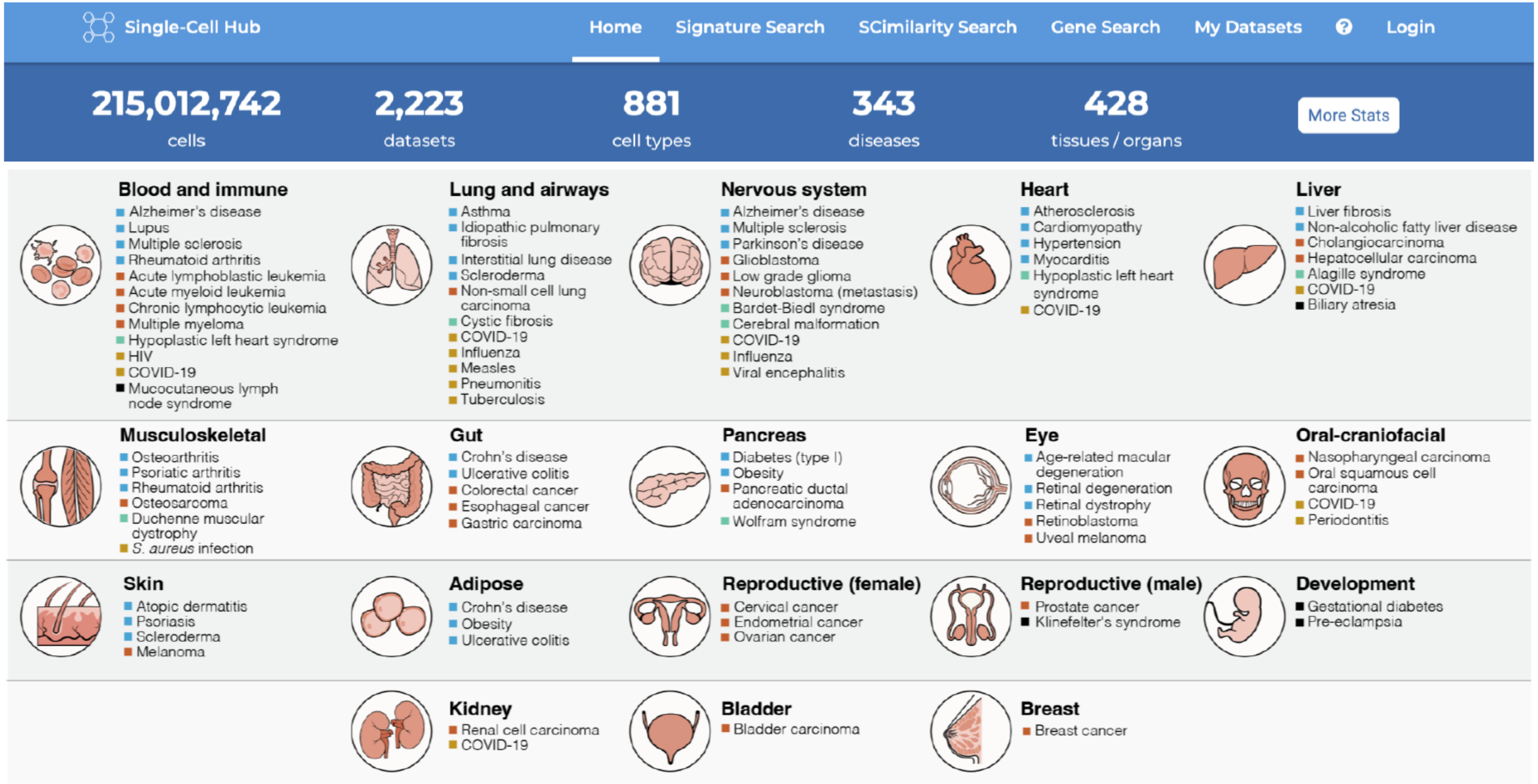


On-target toxicities



Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 53

Genentech: SCHub Has >200M Cells From >2.2K Studies



215,012,742 cells

2,223 datasets

881 cell types

343 diseases

428 tissues / organs

Blood and immune

- 🟦 Alzheimer's disease
- 🟦 Lupus
- 🟦 Multiple sclerosis
- 🟦 Rheumatoid arthritis
- 🟧 Acute lymphoblastic leukemia
- 🟧 Acute myeloid leukemia
- 🟧 Chronic lymphocytic leukemia
- 🟧 Multiple myeloma
- 🟩 Hypoplastic left heart syndrome
- 🟨 HIV
- 🟨 COVID-19
- ⬛ Mucocutaneous lymph node syndrome

Lung and airways

- 🟦 Asthma
- 🟦 Idiopathic pulmonary fibrosis
- 🟦 Interstitial lung disease
- 🟦 Scleroderma
- 🟧 Non-small cell lung carcinoma
- 🟩 Cystic fibrosis
- 🟨 COVID-19
- 🟨 Influenza
- 🟨 Measles
- 🟨 Pneumonitis
- 🟨 Tuberculosis

Nervous system

- 🟦 Alzheimer's disease
- 🟦 Multiple sclerosis
- 🟦 Parkinson's disease
- 🟧 Glioblastoma
- 🟧 Low grade glioma
- 🟧 Neuroblastoma (metastasis)
- 🟩 Bardet-Biedl syndrome
- 🟩 Cerebral malformation
- 🟨 COVID-19
- 🟨 Influenza
- 🟨 Viral encephalitis

Heart

- 🟦 Atherosclerosis
- 🟦 Cardiomyopathy
- 🟦 Hypertension
- 🟦 Myocarditis
- 🟩 Hypoplastic left heart syndrome
- 🟨 COVID-19

Liver

- 🟦 Liver fibrosis
- 🟦 Non-alcoholic fatty liver disease
- 🟧 Cholangiocarcinoma
- 🟧 Hepatocellular carcinoma
- 🟩 Alagille syndrome
- 🟨 COVID-19
- ⬛ Biliary atresia

Musculoskeletal

- 🟦 Osteoarthritis
- 🟦 Psoriatic arthritis
- 🟦 Rheumatoid arthritis
- 🟧 Osteosarcoma
- 🟩 Duchenne muscular dystrophy
- 🟨 *S. aureus* infection

Gut

- 🟦 Crohn's disease
- 🟦 Ulcerative colitis
- 🟧 Colorectal cancer
- 🟧 Esophageal cancer
- 🟧 Gastric carcinoma

Pancreas

- 🟦 Diabetes (type I)
- 🟦 Obesity
- 🟧 Pancreatic ductal adenocarcinoma
- 🟩 Wolfram syndrome

Eye

- 🟦 Age-related macular degeneration
- 🟦 Retinal degeneration
- 🟦 Retinal dystrophy
- 🟧 Retinoblastoma
- 🟧 Uveal melanoma

Oral-craniofacial

- 🟧 Nasopharyngeal carcinoma
- 🟧 Oral squamous cell carcinoma
- 🟨 COVID-19
- 🟨 Periodontitis

Skin

- 🟦 Atopic dermatitis
- 🟦 Psoriasis
- 🟦 Scleroderma
- 🟧 Melanoma

Adipose

- 🟦 Crohn's disease
- 🟦 Obesity
- 🟦 Ulcerative colitis

Reproductive (female)

- 🟧 Cervical cancer
- 🟧 Endometrial cancer
- 🟧 Ovarian cancer

Reproductive (male)

- 🟧 Prostate cancer
- ⬛ Klinefelter's syndrome

Development

- ⬛ Gestational diabetes
- ⬛ Pre-eclampsia

Kidney

- 🟧 Renal cell carcinoma
- 🟨 COVID-19

Bladder

- 🟧 Bladder carcinoma

Breast

- 🟧 Breast cancer

Rood et al. *Nature Medicine* 2022

🟦 **Common complex disease** 🟧 **Cancer** 🟩 **Rare disease** 🟨 **Infectious disease** ⬛ **Other**

Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 54

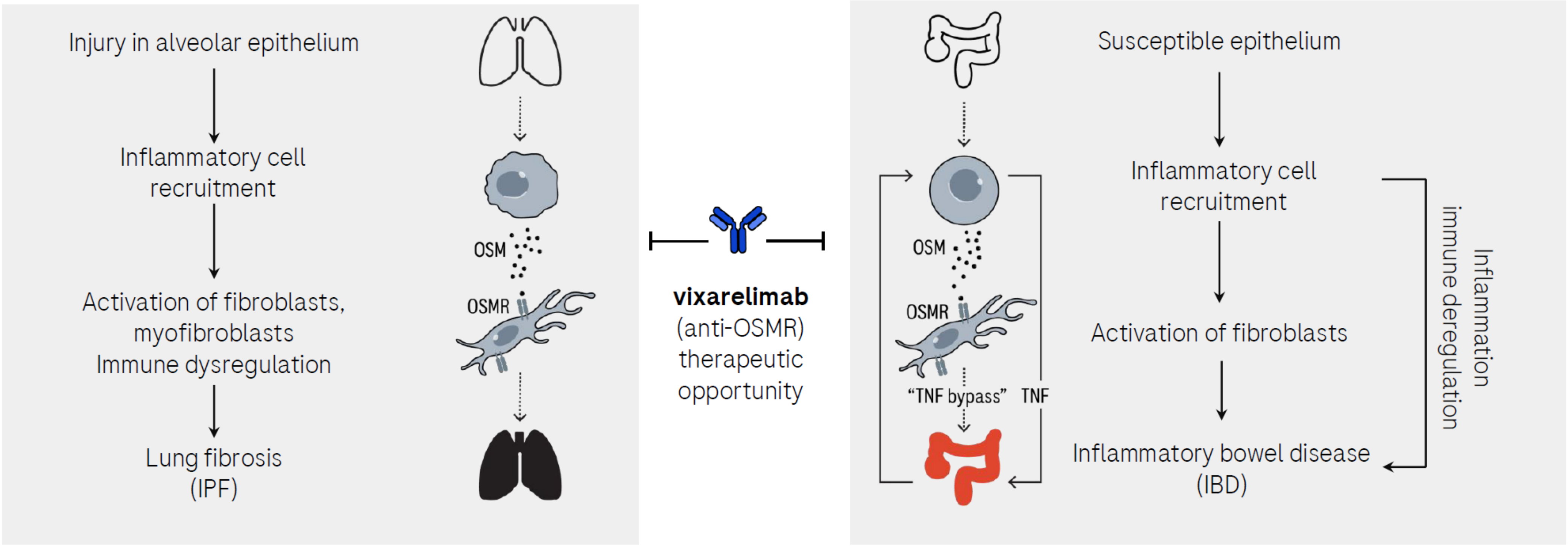
Portfolio Impact: Target OSM/OSMR Pathway Implicated in Both IPF and IBD Through Different Cells

Original goal: Antagonizing OSMR to prevent progressive lung fibrosis

Hypothesis: Blockade of OSMR on myofibroblast will stall progressive fibrosis

Additional goal: Antagonizing OSMR to prevent IBD

Hypothesis: Targeting OSMR in IBD to block inflammatory pathways that drive disease



Injury in alveolar epithelium

Inflammatory cell recruitment

Activation of fibroblasts, myofibroblasts Immune dysregulation

Lung fibrosis (IPF)

vixarelimab (anti-OSMR) therapeutic opportunity

Susceptible epithelium

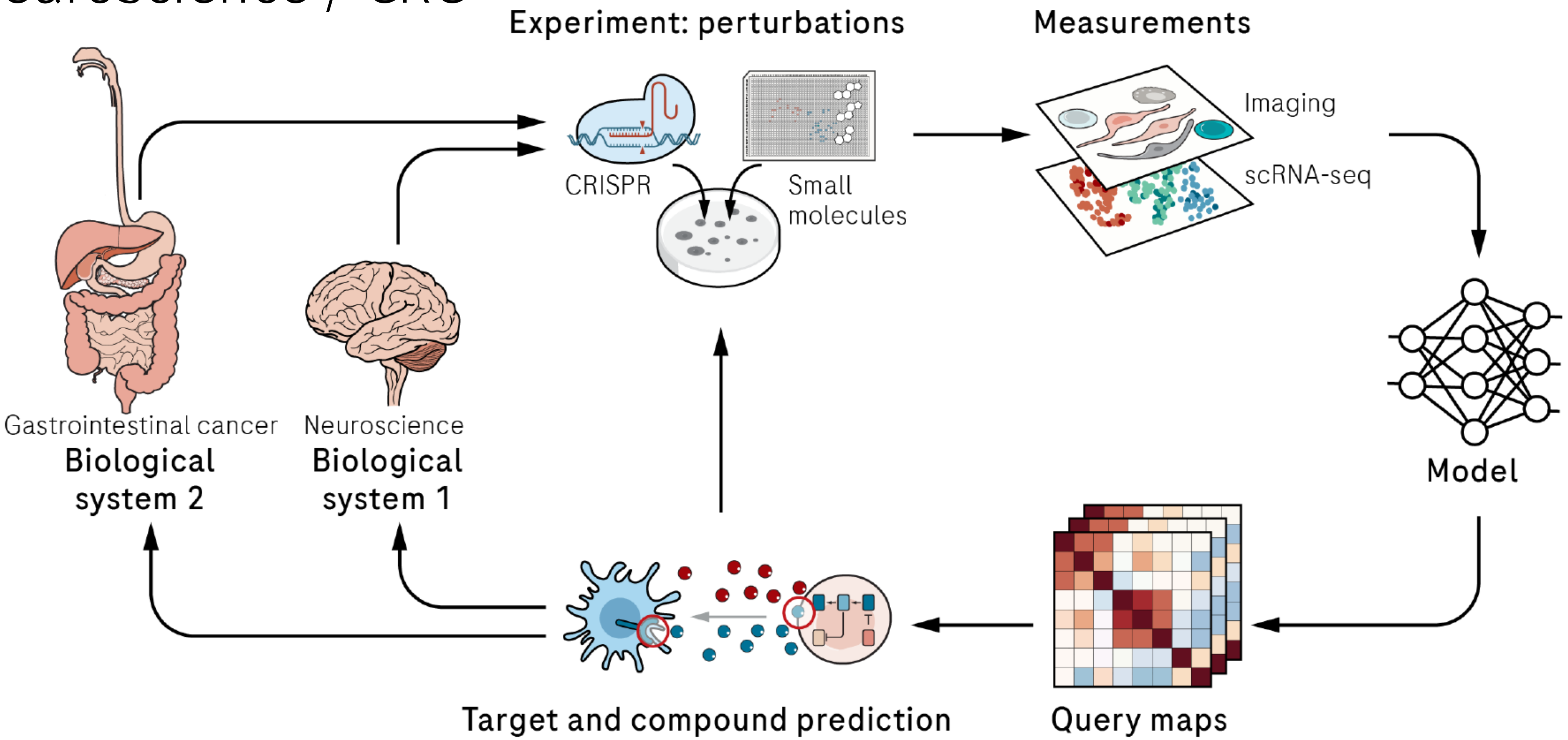
Inflammatory cell recruitment

Activation of fibroblasts

Inflammatory bowel disease (IBD)

Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 58

Portfolio Impact: Lab in a Loop for Target and Compound Prediction in Neuroscience / CRC

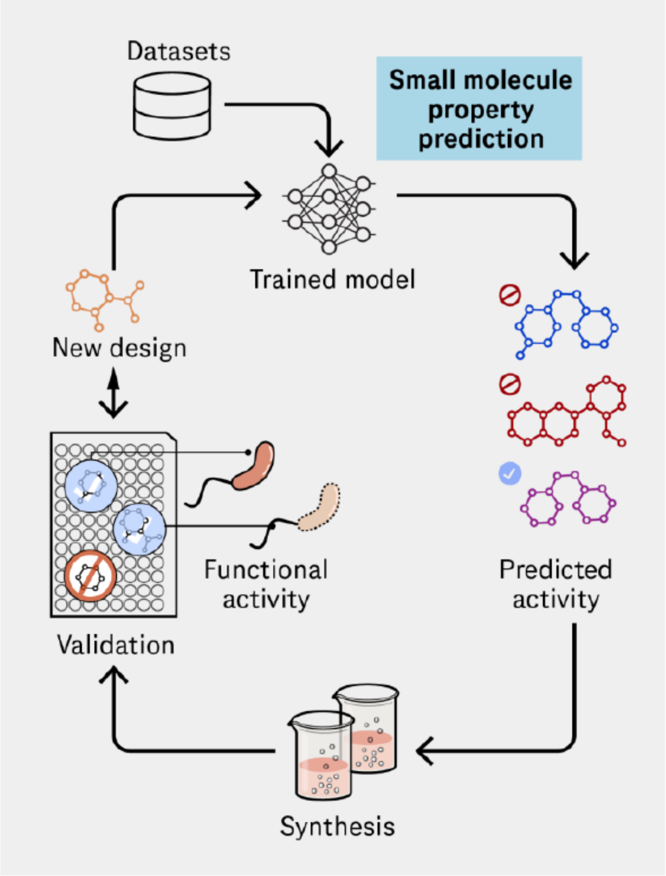


In collaboration with Recursion

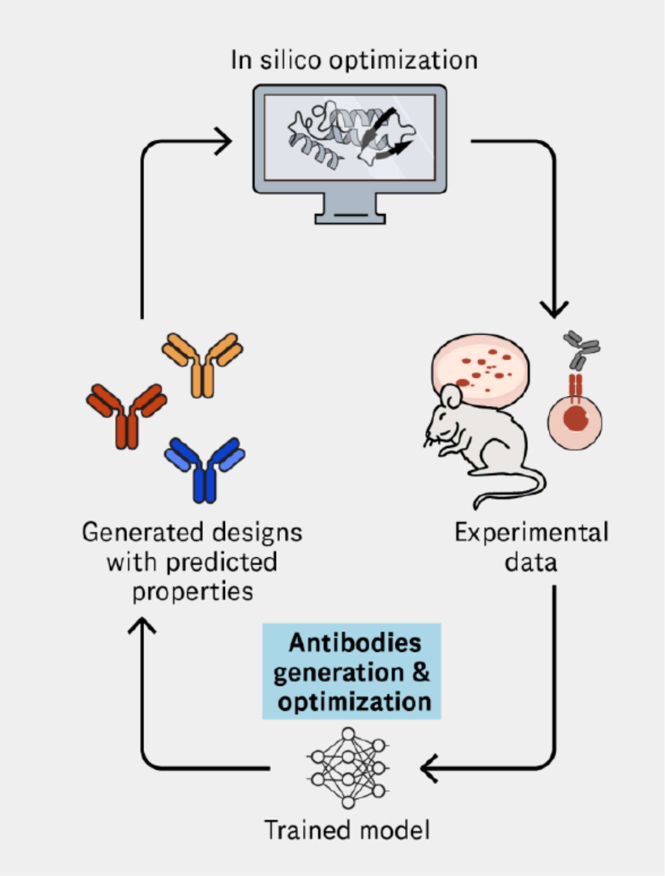
Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 53

Genentech: Improving Drug Discovery PTS with AI

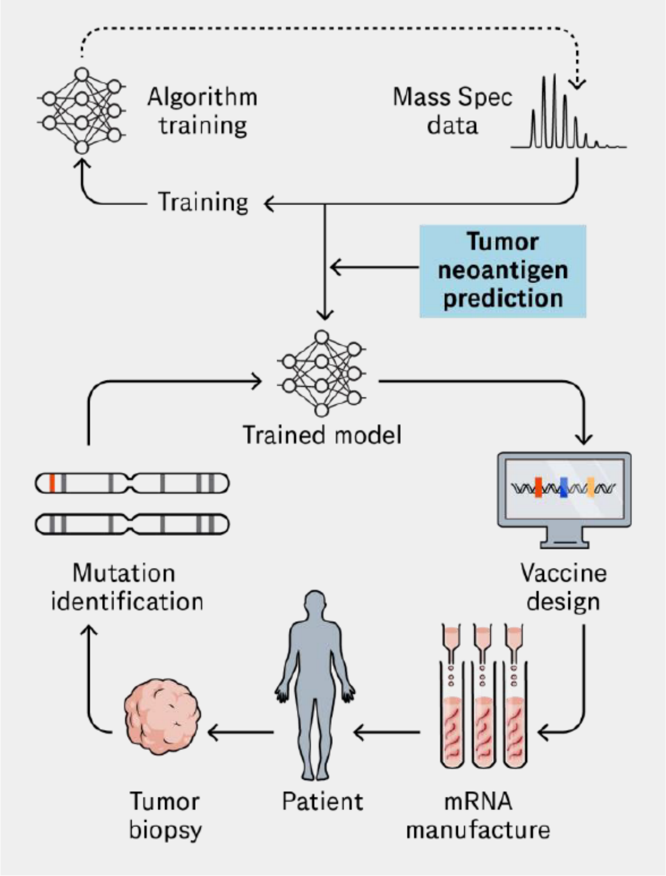
Small molecules
property prediction



Antibodies
optimization and generation


In silico optimization
Generated designs with predicted properties
Experimental data
Antibodies generation & optimization
Trained model

Tumor neoantigen
prediction

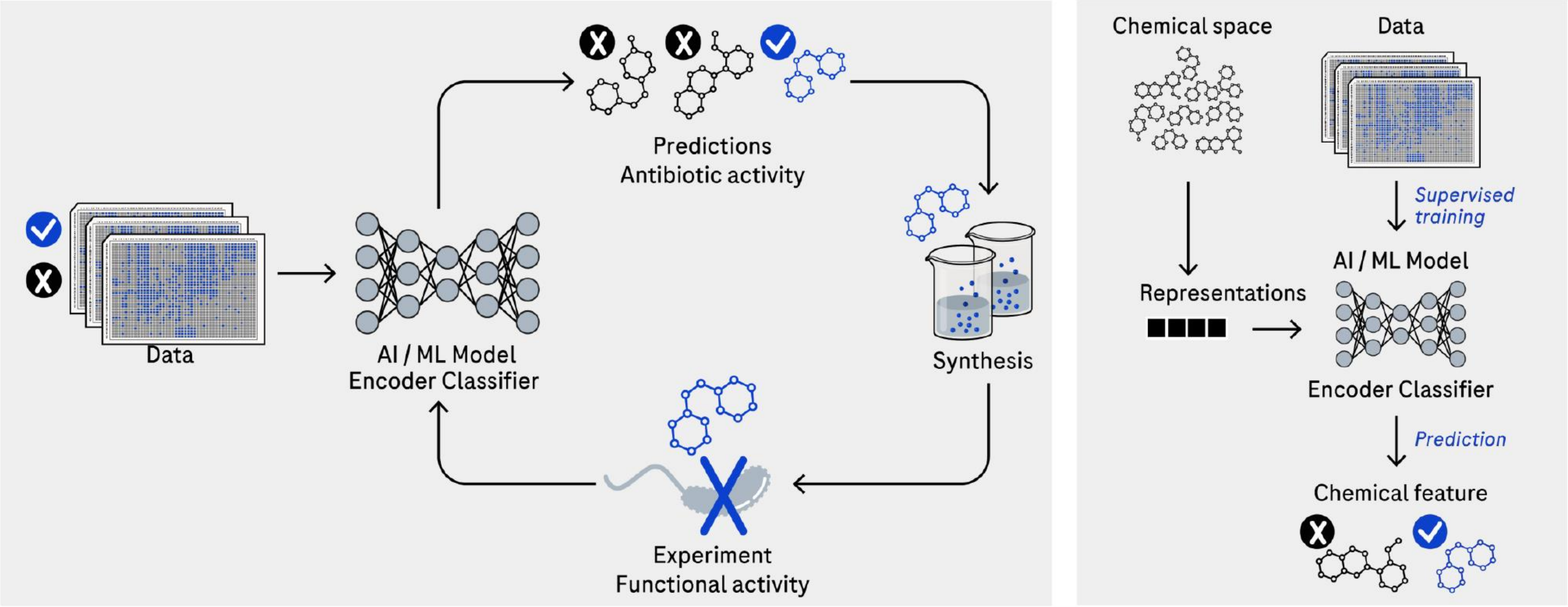


New ML/AI algorithms for small molecule, antibodies, vaccine design, etc.

PTS=probability of technical success

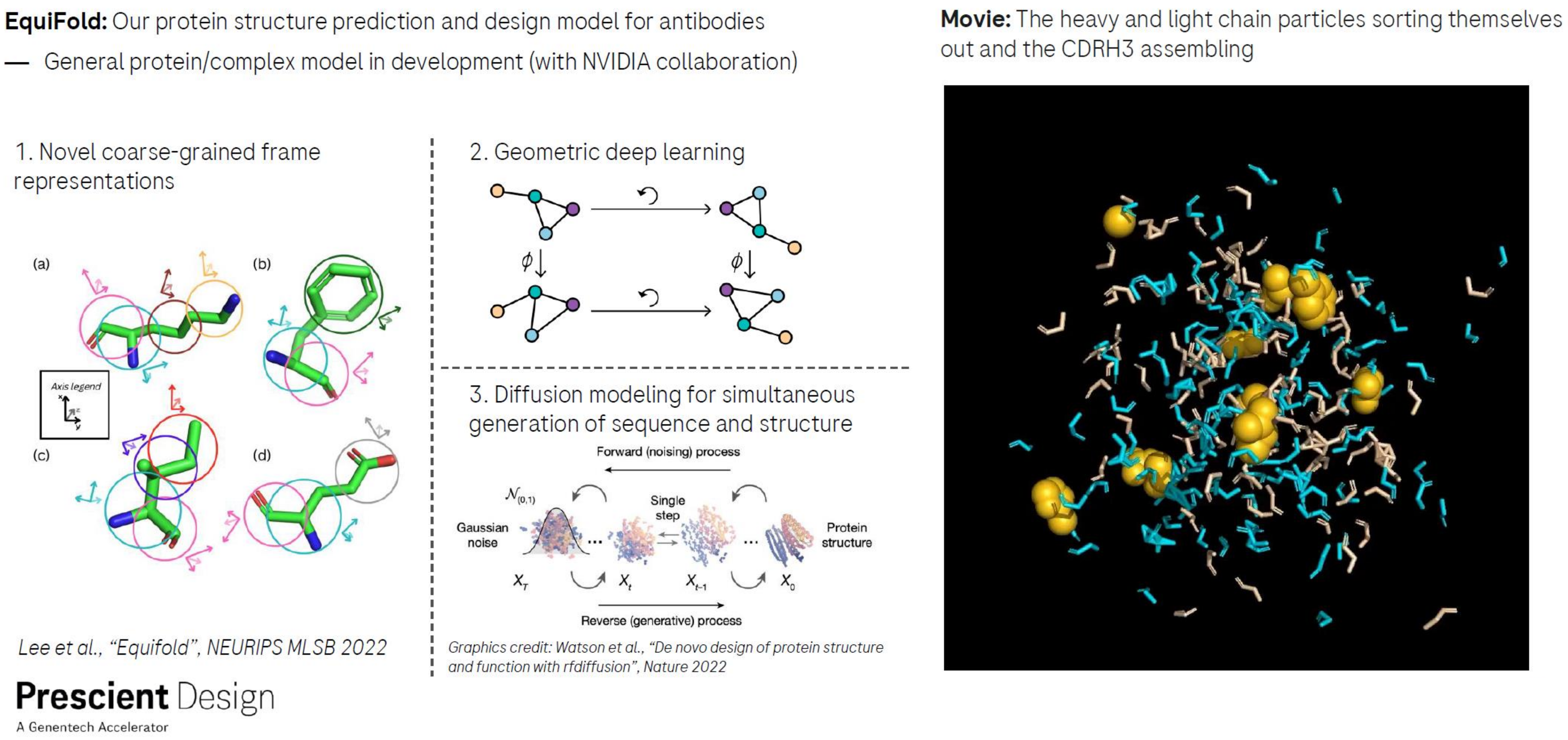
Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 68

Genentech: Lab in a Loop for Small Molecule Property Prediction



Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 69

Genentech: EquiFold Diffusion: Fast Antibody Structure Prediction



EquiFold: Our protein structure prediction and design model for antibodies

— General protein/complex model in development (with NVIDIA collaboration)

1. Novel coarse-grained frame representations

Lee et al., "Equifold", NEURIPS MLSB 2022

Prescient Design

A Genentech Accelerator

2. Geometric deep learning

3. Diffusion modeling for simultaneous generation of sequence and structure

Graphics credit: Watson et al., "De novo design of protein structure and function with rfdiffusion", Nature 2022

Movie: The heavy and light chain particles sorting themselves out and the CDRH3 assembling

Source: https://assets.roche.com/f/176343/x/e60b81765d/20231129_digi-day.pdf, p. 73

sanofi

Work on Artificial Intelligence and Machine Learning

> “Our ambition is to become the first pharma company powered by artificial intelligence at scale, giving our people tools and technologies that focus on insights and allow them to make better everyday decisions. The use of artificial intelligence and data science already support our teams’ efforts in areas such as accelerating drug discovery, enhanced clinical trial design, and improving manufacturing and supply of medicines and vaccines.”

Paul Hudson

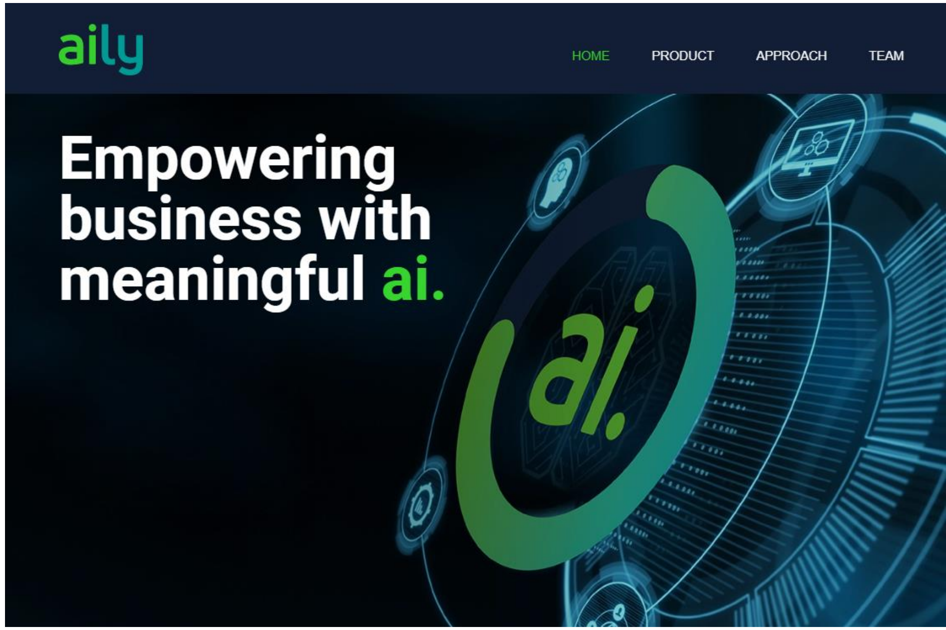
Chief Executive Officer
Sanofi

Source: https://www.sanofi.com/en/media-room/press-releases/2023/2023-06-13-12-00-00-2687072

Press Release: Sanofi "All In" on Artificial Intelligence and Data Science to Speed Breakthroughs for Patients

Paris, June 13, 2023. Sanofi takes the next step in its company-wide digital transformation and rolls-out *plai* at scale. *plai*, Sanofi's industry-leading app developed with artificial intelligence (AI) platform company Aily Labs, delivers real-time, reactive data interactions and gives an unprecedented 360° view across all Sanofi activities. The app aggregates available company internal data across functions and harnesses the power of AI to provide timely insights and personalized "*what if*" scenarios to support thousands of Sanofi team's decision makers to take informed decisions in a simple and modern digital user experience.

plai is an essential enabler in the company-wide digital transformation and data democratization journey. AI-powered tools help Sanofi teams make better and faster data-driven decisions, hence boosting productivity



Source: https://www.sanofi.com/en/media-room/press-releases/2023/2023-06-13-12-00-00-2687072

OWKIN

Thomas Clozel, CEO of Owkin

Frank Nestlé, SVP R&D, Sanofi

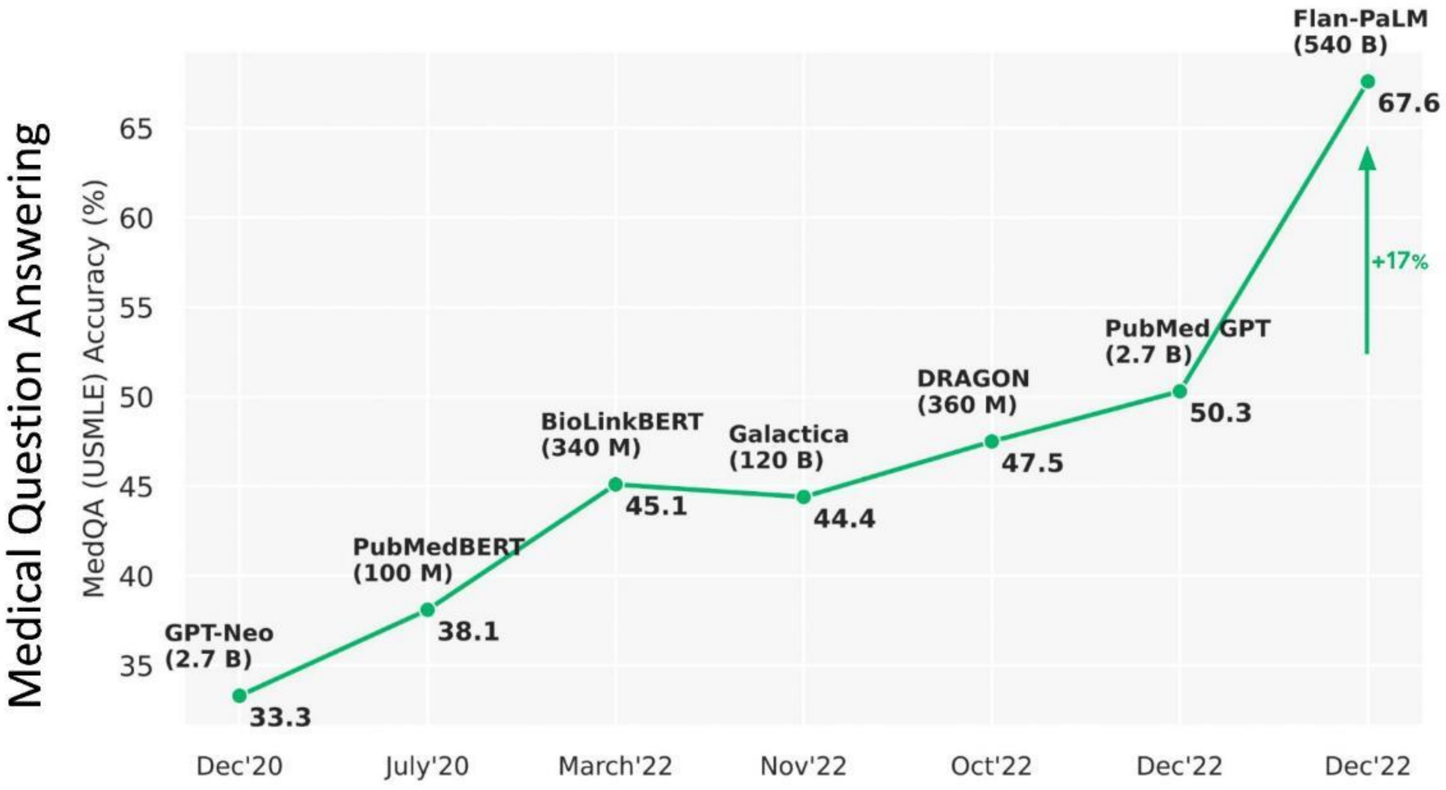
In 2021, Sanofi announced its partnership with artificial intelligence (AI) bio-tech company Owkin, joining the wide collection of companies embracing this rapidly developing technology. Since then, AI has seen growth in almost every industry, but according to Thomas Clozel, CEO of Owkin, we have yet to unlock its greatest potential in drug development, specifically in the area of precision medicine.

AI and Healthcare

ORGANS
"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Can ChatGPT Beat Doctors yet?

PaLM Model Got 67% of Questions right on United States Medical Licensing Examination (USMLE). A good doctor can get 90% of questions right.



See:
https://www.advisory.com/daily-briefing/2023/01/23/ai-exam

For medical doctor reactions to this finding see:

https://www.sciencemediacentre.org/expert-reaction-to-study-on-chatgpt-almost-passing-the-us-medical-licensing-exam/

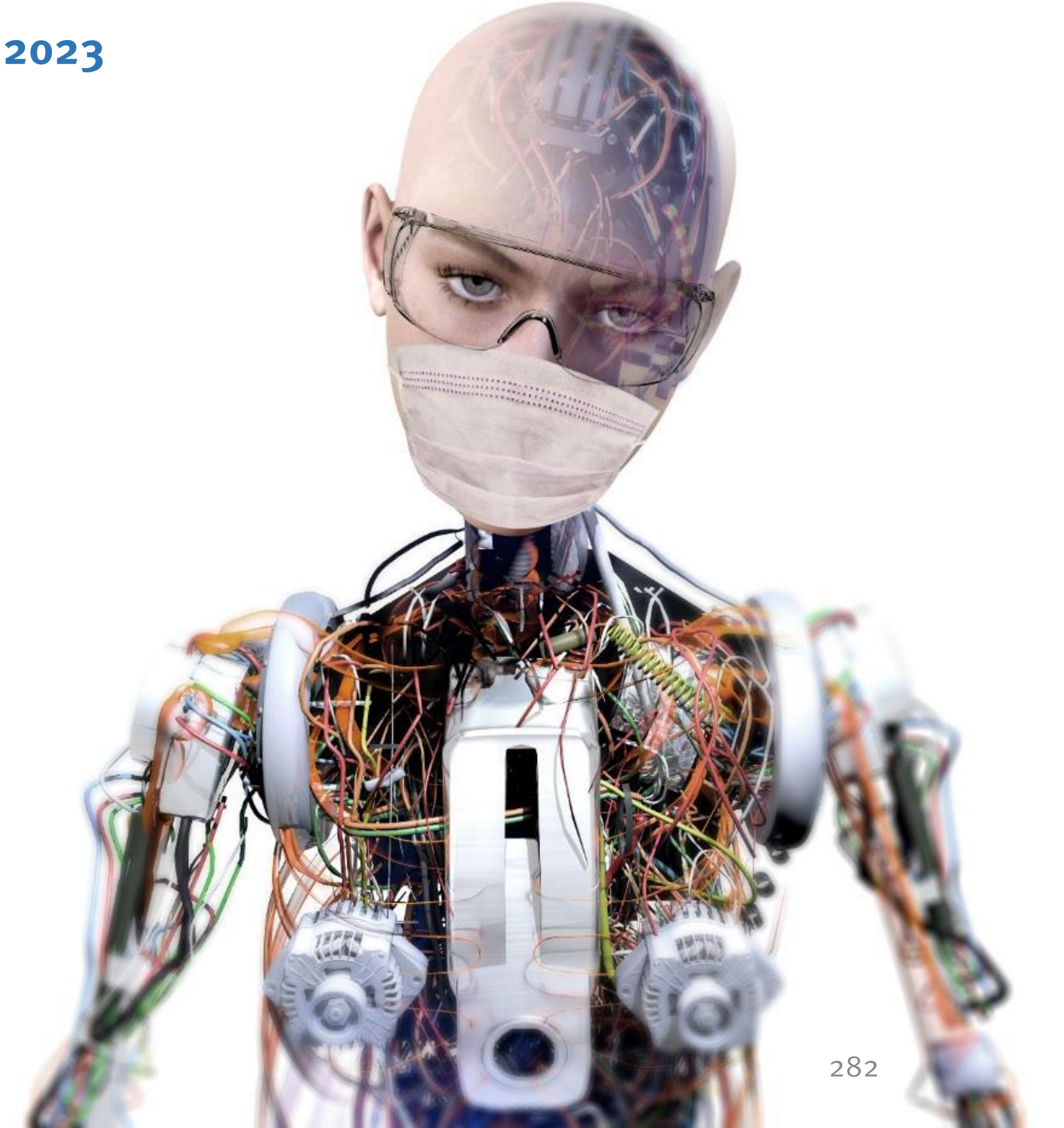
Source: https://erictopol.substack.com/p/when-md-is-a-machine-doctor

Can Machines Replace Doctors?

Eric Topol, “When MD is a Machine Doctor,” Substack, Jan 15, 2023

“It’s very early for LLMs/generative AI/foundation models in medicine, but I hope you can see from this overview that there has been substantial progress in answering medical questions—that AI is starting to pass the tests that approach the level of doctors, and it’s no longer just about image interpretation, but starting to incorporate medical reasoning skills. That doesn’t have anything to do with licensing machines to practice medicine, but it’s a reflection that a force is in the works to help clinicians and patients process their multimodal health data for various purposes. The key concept here is augment; I can’t emphasize enough that machine doctors won’t replace clinicians. Ironically, it’s about technology enhancing the quintessential humanity in medicine.”

Source: https://erictopol.substack.com/p/when-md-is-a-machine-doctor



Israeli Study Shows that AI-Generated Diagnoses in a Primary Care Setting Generally Get it Right

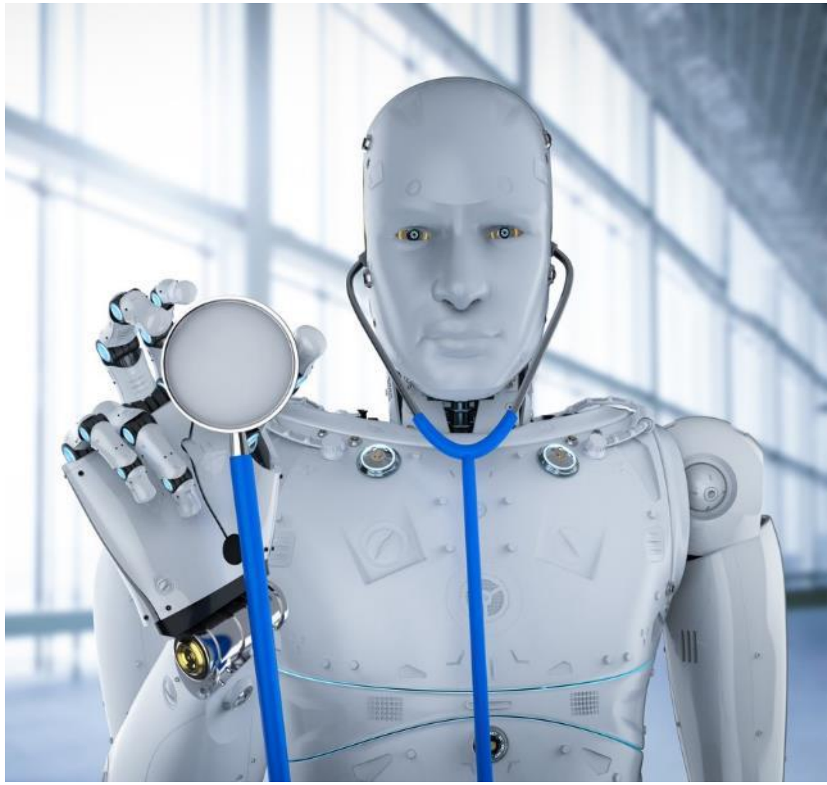
Zeltzer et.al., "Diagnostic Accuracy of Artificial Intelligence in Virtual Primary Care," ***Mayo Clinic Proceedings: Digital Health*****, Sep 20, 2023.**

Objective: To evaluate the diagnostic accuracy of artificial intelligence (AI)-generated clinical diagnoses.

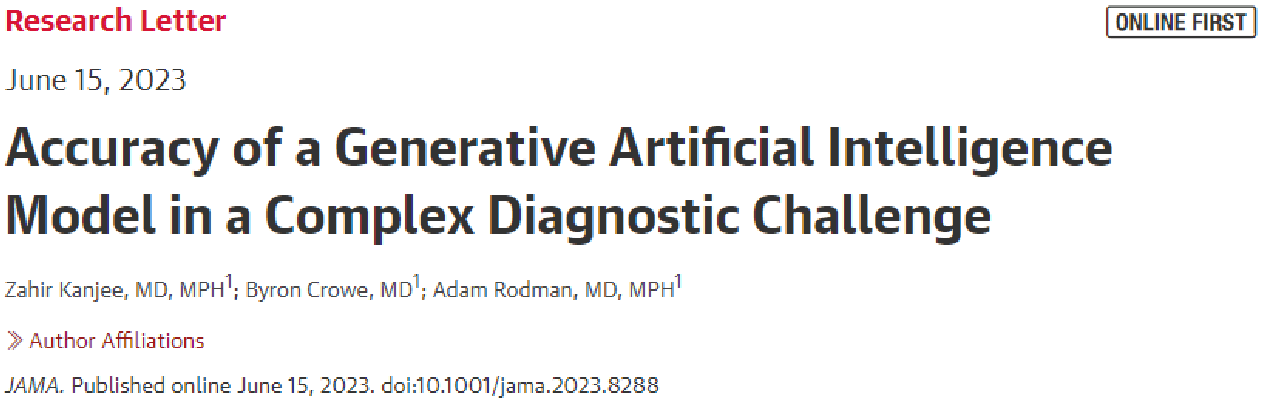
Patients and Methods: A retrospective chart review of 102,059 virtual primary care clinical encounters from October 1, 2022, to January 31, 2023 was conducted. Patients underwent an AI medical interview, after which virtual care providers reviewed the interview summary and AI-provided differential diagnoses, communicated with patients, and finalized diagnoses and treatment plans. Our accuracy measures were agreement between AI diagnoses, virtual care providers, and blind adjudicators. We analyzed AI diagnostic agreement across different diagnoses, presenting symptoms, patient demographic characteristics such as race, and provider levels of experience. We also evaluated model performance improvement with retraining.

Results: Providers selected an AI diagnosis in 84.2% (n = 85,976) of cases and the top-ranked AI diagnosis in 60.9% (n = 62,130) of cases. Agreement rates varied by diagnosis, with greater than or equal to 95% provider agreement with an AI diagnosis for 35 diagnoses (47% of cases, n = 47,679) and greater than or equal to 90% agreement for 57 diagnoses (69% of cases, n = 70,697). The average agreement rate for half of all presenting symptoms was greater than or equal to 90%. Adjusting for case mix, diagnostic accuracy exhibited minimal variation across demographic characteristics. The adjudicators' consensus diagnosis, reached in 58.2% (n = 128) of adjudicated cases was always included in the AI differential diagnosis. Provider experience did not affect agreement, and model retraining increased diagnostic accuracy for retrained conditions from 96.6% to 98.0%.

Conclusion: Our findings show that agreement between AI and provider diagnoses is high in most cases in the setting of this study. The results highlight the potential for AI to enhance primary care disease diagnosis and patient triage, with the capacity to improve over time.



Source: https://www.mcpdigitalhealth.org/article/S2949-7612(23)00070-6/fulltext



Research Letter

ONLINE FIRST

June 15, 2023

Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge

Zahir Kanjee, MD, MPH[1]; Byron Crowe, MD[1]; Adam Rodman, MD, MPH[1]

» Author Affiliations

JAMA. Published online June 15, 2023. doi:10.1001/jama.2023.8288

“We used *New England Journal of Medicine* clinicopathologic conferences. These conferences are challenging medical cases with a final pathological diagnosis that are used for educational purposes; they have been used to evaluate differential diagnosis generators since the 1950s.

We used the first 7 case conferences from 2023 to iteratively develop a standard chat prompt that explained the general conference structure and instructed the model to provide a differential diagnosis ranked by probability. We copied each case published from January 2021 to December 2022, up to but not including the discussant’s initial response and differential diagnosis discussion, and pasted it along with our prompt into the model. We chose recent cases because most of the model’s training data ends in September 2021. Each case, including the cases used to develop the prompt, was run in independent chats to prevent the model applying any “learning” to subsequent cases.

Our prespecified primary outcome was whether the model’s top diagnosis matched the final case diagnosis. Prespecified secondary outcomes were the presence of the final diagnosis in the model’s differential, differential length, and differential quality score using a previously published ordinal 5-point rating system based on accuracy and usefulness (in which a score of 5 is given for a differential including the exact diagnosis and a score of 0 is given when no diagnoses are close). All cases were independently scored by Z.K. and B.C., with disagreements adjudicated by A.R.

Of 80 cases, 10 were excluded (4 were not diagnostic dilemmas; 6 were deleted for length). The 2 primary scorers agreed on 66% of scores (46/70; κ = 0.57 [moderate agreement]). The AI model’s top diagnosis agreed with the final diagnosis in 39% (27/70) of cases. In 64% of cases (45/70), the model included the final diagnosis in its differential (Table). Mean differential length was 9.0 (SD, 1.4) diagnoses. When the AI model provided the correct diagnosis in its differential, the mean rank of the diagnosis was 2.5 (SD, 2.5). The median differential quality score was 5 (IQR, 3-5); the mean was 4.2 (SD, 1.3) (Figure).

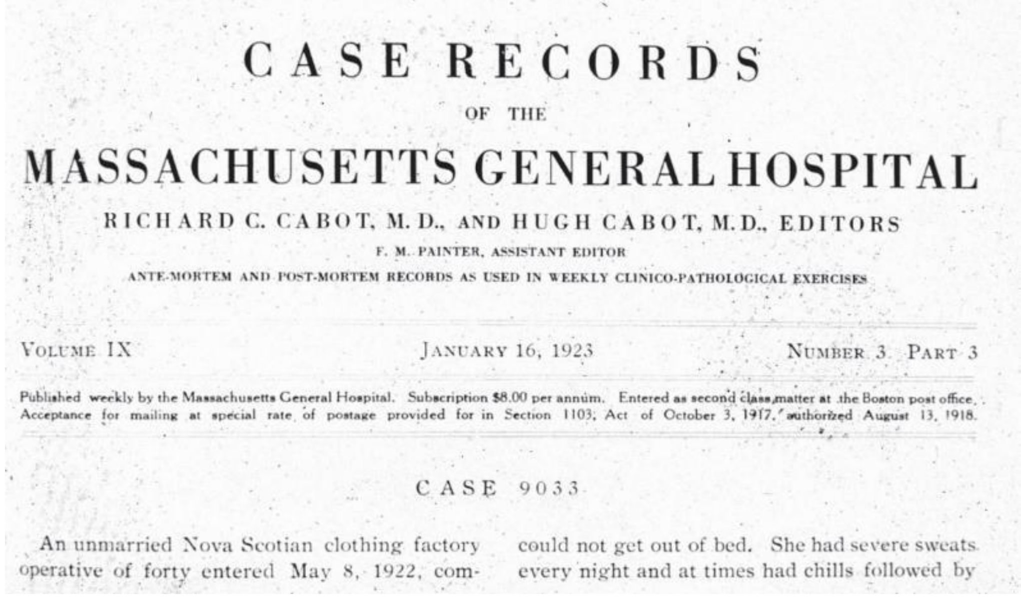
A generative AI model provided the correct diagnosis in its differential in 64% of challenging cases and as its top diagnosis in 39%. The finding compares favorably with existing differential diagnosis generators. A 2022 study evaluating the performance of 2 such models also using *New England Journal of Medicine* clinicopathological case conferences found that they identified the correct diagnosis in 58% to 68% of cases; the measure of quality was a simple dichotomy of useful vs not useful. GPT-4 provided a numerically superior mean differential quality score compared with an earlier version of one of these differential diagnosis generators (4.2 vs 3.8).”

Source: https://jamanetwork.com/journals/jama/fullarticle/2806457

Stump the Medical Expert

A.I., the CPC, and making the diagnosis

By Eric Topol, *Substack*, June 17, 2023



CASE RECORDS

OF THE

MASSACHUSETTS GENERAL HOSPITAL

RICHARD C. CABOT, M. D., AND HUGH CABOT, M.D., EDITORS

F. M. PAINTER, ASSISTANT EDITOR

ANTE-MORTEM AND POST-MORTEM RECORDS AS USED IN WEEKLY CLINICO-PATHOLOGICAL EXERCISES

VOLUME IX JANUARY 16, 1923 NUMBER 3 PART 3

Published weekly by the Massachusetts General Hospital. Subscription $8.00 per annum. Entered as second class matter at the Boston post office. Acceptance for mailing at special rate of postage provided for in Section 1103, Act of October 3, 1917, authorized August 13, 1918.

CASE 9033

An unmarried Nova Scotian clothing factory operative of forty entered May 8, 1922, com-

could not get out of bed. She had severe sweats every night and at times had chills followed by

100 years ago, the first Case Records of the Massachusetts General Hospital was published in the Boston Medical and Surgical Journal, the precursor to the *New England Journal of Medicine*, which has been publishing what are known as clinicopathologic conferences (CPCs) since 1924 on a biweekly basis.

The CPC is a longstanding tradition that continues in many medical centers throughout the world, first introduced in the United States in 1898, undoubtedly influenced by Giovanni Battista Morgagni who published a book of 700 cases with anatomy-clinical correlations in 1761. As they evolved over the years, CPCs were extremely challenging patient cases to stump the medical expert. After presentation of the relevant data, the clinician expert would be asked to provide a differential diagnosis and presumptive final diagnosis, and the actual, definitive diagnosis was established via lab tests, scans, pathology, or autopsy. The CPC educational value is clearcut, but so was there entertainment to see if the noted expert might miss the diagnosis. Of course, there was the expectation that the master clinician—the doctor's doctor— would always get it right. I vividly remember trying to stump UCSF Professor Larry Tierney during my internal medicine residency training, but it was rare that he didn't have the right diagnosis in his differential.

That rarity of expert wrong diagnosis differs substantially from real world medicine. After a classic *Science* 1974 paper about uncertainty, one of its authors, Danny Kahneman, wrote about a study that compared the doctor's diagnosis before death to the autopsy findings. **"Clinicians who were completely certain of the diagnosis antemortem were wrong 40 percent of the time."**

Source: https://erictopol.substack.com/p/stump-the-medical-expert

Stump the Medical Expert: GPT-4 Beats Docs on CPC Diagnoses (Topol Article Continued)

"The New England Journal CPCs have been the benchmark for evaluating medical diagnostic reasoning, as used in the 1959 paper in Science Magazine, emphasizing the role of mathematical techniques and associated use of computers noting: "This method in no way implies that a computer can take over the physician's duties."

To date, the best DDx generator results that have been published are derived from Isabel Health's tool. In the most recent report with Isabel DDx data, two doctors (not medical experts) got 14 of 50 (28%) of the NEJM CPC final diagnoses. As the current report asserted, "GPT-4 provided a numerically superior mean differential quality score compared with an earlier version of one of these differential diagnosis generators (4.2 vs 3.8)."

We need to prospectively assess GPT-4 for its role in facilitating diagnoses. This is a major issue in medicine today: at least 12 million Americans, as out-patients, are misdiagnosed each year. There is real promise for GPT-4 and other large language models (LLMs) to help the accuracy of diagnoses for real world patients, not the esoteric, rare, ultra-challenging NEJM CPC cases. But that has to be proven, and certainly the concern about LLM confabulations is key, potentially leading a physician and patient down a rabbit-hole, towards a major wrong diagnosis and an extensive workup without basis, no less the possibility of an erroneous treatment.

Thee's been much buzz about ChatGPT, GPT 3.5, Med-PaLM, and GPT-4 surpassing the 60% pass threshold for the United States Medical Licensing exam (USMLE) (to reach ~86%). These are fairly contrived comparisons using a subset of representative questions and only those with text, not with visual media. It makes for nice bragging right for LLMs, but we aren't going to be licensing any of them to practice medicine! That's far less relevant than use of these AI tools for promoting accurate diagnoses.

I'm excited about this particular use case of LLMs in the future, especially as they undergo supervised fine tuning for medical knowledge. It clearly needs dedicated, prospective validation work, but ultimately may become a significant support tool for clinicians and patients. **If GPT-4 can perform well with arcane NEJM cases, imagine what might be in store for the common, real world diagnostic dilemmas.**"

Scripps Research

Eric Topol, MD
EVP, Scripps Research

Source: https://erictopol.substack.com/p/stump-the-medical-expert

5 Ways ChatGPT Will Change Healthcare Forever, For Better

Forbes

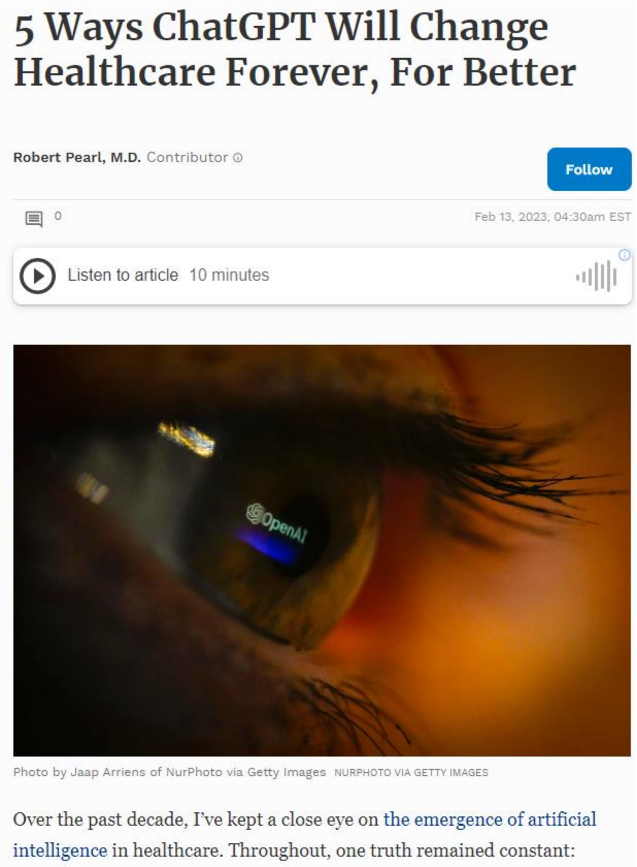
We asked ChatGPT to summarize last week's *Forbes* article on ChatGPT. Here is what we got:

"ChatGPT is a generative AI chatbot from OpenAI that has impressed many by passing graduate-level exams for medical school. With more computing power and data in the future, ChatGPT will possess analytical and problem-solving powers that far exceed current expectations, and revolutionize the medical practice by matching the diagnostic skills of clinicians today.

Generative AI can access terabytes of data in less than a second to predict the next best word or idea in a series of words and concepts. Future generations of generative AI will be able to monitor patients 24/7 and provide ongoing medical expertise, preventing chronic illnesses like heart disease, hypertension, and diabetes, and minimizing their deadly complications.

Machine-learning capabilities will evolve far beyond predicting text, and generative AI can prevent medical errors, which is significant given that nearly 1 in 4 individuals admitted to a hospital will experience harm during their stay. In the future, people won't need to bring a family member with them to protect against deadly mistakes made by humans.

Future generations of AI will help all doctors perform like the best by following a combination of textbooks, journal articles, classroom instruction, and observation of skilled clinicians. The use of generative AI tools will alter medical practice in previously unimaginable ways, becoming an essential part of our lives, and providing around-the-clock medical assistance."



5 Ways ChatGPT Will Change Healthcare Forever, For Better

Robert Pearl, M.D. Contributor

0

Feb 13, 2023, 04:30am EST

Photo by Jaap Arriens of NurPhoto via Getty Images NURPHOTO VIA GETTY IMAGES

Over the past decade, I've kept a close eye on the emergence of artificial intelligence in healthcare. Throughout, one truth remained constant:

To read the actual article go to: https://www.forbes.com/sites/robertpearl/2023/02/13/5-ways-chatgpt-will-change-healthcare-forever-for-better/?sh=161d3bdc7bfc

Disclosure

STIFEL | Healthcare

Stifel collectively refers to Stifel, Nicolaus & Company, Incorporated and other affiliated broker-dealer subsidiaries of Stifel Financial Corp. The information and statistical data contained herein have been obtained from sources that Stifel believes are reliable, but Stifel makes no representation or warranty as to the accuracy or completeness of any such information or data and expressly disclaims any and all liability relating to or resulting from your use of these materials. The information and data contained herein are current only as of the date(s) indicated, and Stifel has no intention, obligation, or duty to update these materials after such date(s). These materials do not constitute an offer to sell or the solicitation of an offer to buy any securities, and Stifel is not soliciting any action based on this material. Stifel may be a market-maker in certain of these securities, and Stifel may have provided investment banking services to certain of the companies listed herein. Stifel and/or its respective officers, directors, employees, and affiliates may at any time hold a long or short position in any of these securities and may from time-to-time purchase or sell such securities. This material was prepared by Stifel Investment Banking and is not the product of the Stifel Research Department. It is not a research report and should not be construed as such. This material may not be distributed without Stifel's prior written consent.

Stifel, Nicolaus & Company, Incorporated | Member SIPC & NYSE | www.stifel.com