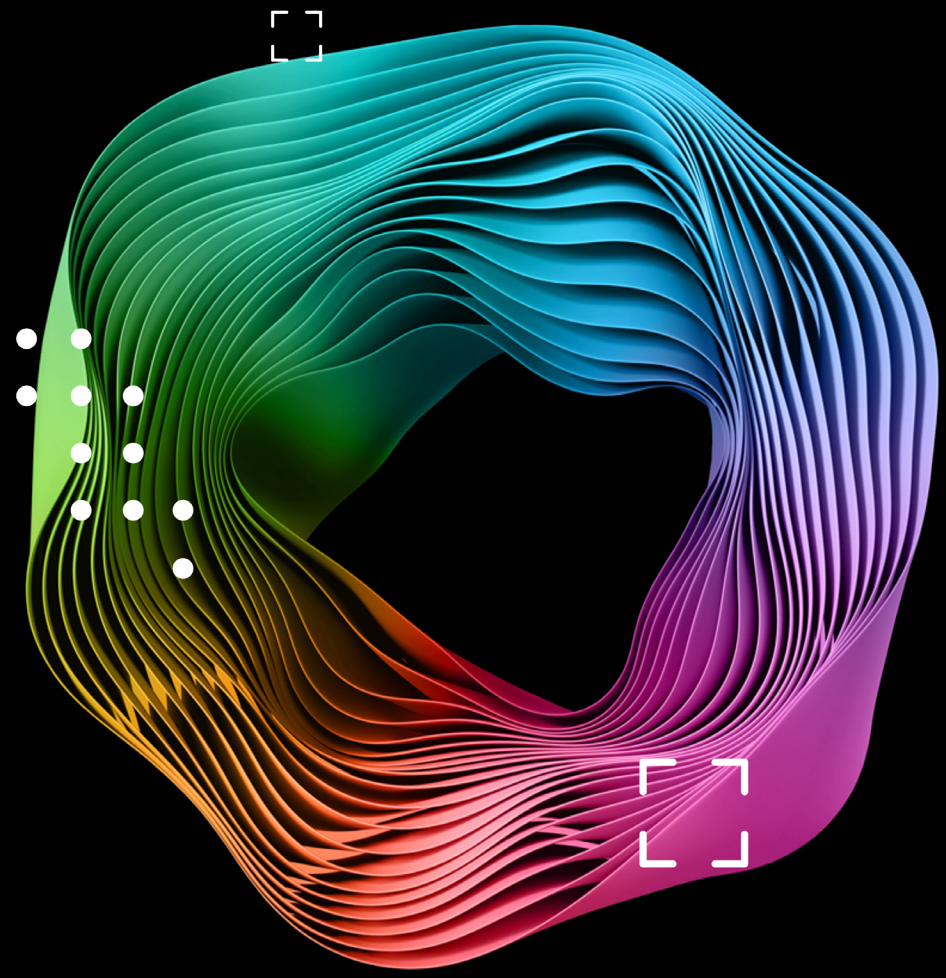
Deloitte. | Google Cloud



A new perspective
on data harmonization, connectivity, and AI models for drug discovery

SECTION 1

Current challenges in drug discovery

Significant challenges persist in contemporary drug discovery despite advances in technology and methodologies. The process from drug discovery to market availability remains protracted, typically spanning over a decade and involving substantial financial investments, often around $2 billion per drug[1]. This extended timeline and high-cost stem from various factors, including rigorous regulatory requirements, lengthy clinical trials, and the complex nature of biological systems.

Moreover, within research laboratories, redundancy in experimental efforts and data silos obstruct efficient collaboration and hinder knowledge sharing. This redundancy often arises due to the fragmented nature of research, where different teams may unknowingly investigate similar pathways or targets. Data silos further exacerbate this issue, impeding the integration and analysis of valuable information across departments or institutions.

The drug discovery landscape may also face the challenge of diminishing "low-hanging fruit," where the most readily exploitable drug targets have already been identified and pursued. This scarcity necessitates exploration into more complex biological mechanisms and novel therapeutic modalities, demanding interdisciplinary collaboration, increased data interoperability, and innovative research approaches to overcome.

Historically, there were two approaches to drug discovery: target-based drug discovery (TDD), which is hypothesis-driven (gene or protein-based), and phenotypic drug discovery (PDD), which is function-based and animal and clinical data-driven. Nowadays, scientists use a mix of both approaches to find and optimize an optimal drug.

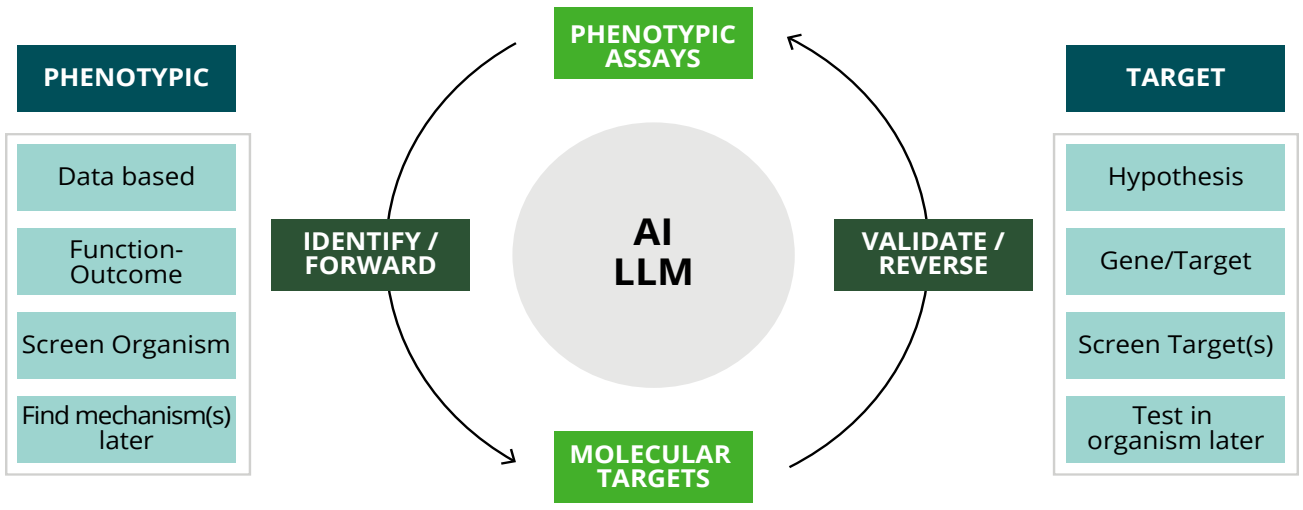
However, to successfully integrate experiments from either a target or phenotypic point of view, they require advanced technologies, such as artificial intelligence (AI) and large language models (LLMs). These technologies can enhance both data and hypothesis-driven research by providing data harmonization and connectivity. One of the many advantages of AI and LLMs is that they can help reduce the required experiments to a minimum.

This paper demonstrates how, in collaboration with Google Cloud, Deloitte can provide capabilities that allow the integration of TDD and PDD approaches. Integrating target and phenotypic approaches allows for harmonization and novel connectivity across multi-modal data using LLMs and knowledge graphs (KG). This implementation can contribute to generating a feedback loop between PDD experimental data and TDD screening and lead optimization experiments that can then be connected to PDD target deconvolution.

The integration of data facilitated by LLMs is also fed into multiple customized scientific pipelines.

> ***By creating a loop across data and scientific workflows, connectivity between TDD and PDD is now possible without the considerable expense and time of multiple and redundant experiments (Figure 1).***

Figure 1 *Technologies such as AI and LLMs incorporate two complementing views of drug discovery: target and phenotypic focus, providing harmonization and connectivity among data and hypothesis-driven research.*



SECTION 2

What is needed in 21st century drug discovery

Mathematical modeling and machine learning (ML) have made significant progress in areas as diverse as recommender systems, computer vision, and natural language processing, solving many of the most challenging benchmarks published in these fields. However, in the context of biology, machine learning progress needs to catch up. AlphaFold2's release in 2021 represented a significant step forward in AI applied to biology, setting a new standard for protein folding prediction. Nevertheless, this and newer models only address a particular corner of what is needed to accelerate the drug discovery process significantly.

In response to this discrepancy, Deloitte's Atlas AI for drug discovery emerges as a pioneering approach. It adopts an integrated, systemic perspective on drug discovery based on the complexity of science. It weaves together three foundational pillars: interconnected data modalities, AI-based scientific pipelines, and large language models.

This integration facilitates a synergistic feedback loop where enhanced data interconnectivity bolsters AI model training, generating superior synthetic data. Thus, enhanced data interconnectivity enriches the pool and utility of interconnected data, allowing large language models to propose novel and actionable hypotheses.

Establishing this feedback mechanism catalyzes a pharmaceutical research and development (R&D) flywheel effect. ***Each research cycle incrementally advances the individual components of data, AI models, and LLMs and accelerates the overall pace of discovery, promising a more rapid and efficient pathway toward innovative drug development (see Sidebars 1 and 2).***

Sidebar 1 - Knowledge Graph to Scientific Pipelines Feedback Loop Example

Small molecule optimization using MolMIM

Atlas AI exploits its rich data connectivity to inform AI experiments, providing strong starting points that increase the speed to convergence and improve the quality of the final desired molecule. One use case accelerated by this feedback loop is molecular optimization. Researchers can specify a desired set of properties for a compound and use the knowledge graph to select starting compounds that have similar properties. These starting compounds can then be optimized using NVIDIA's MolMIM model to arrive at a fully customized molecule that satisfies the specified parameters (such as binding affinity, solubility, and more).

Sidebar 2 – Large Language Model to Knowledge Graph Feedback Loop Example

Creating data connectivity using LLMs

Atlas AI uses modality-specific language models to create edges in the knowledge graph that capture similarities and differences between entities. One use case for this capability is determining structural and functional relationships between proteins using only their amino acid sequences. For example, bet v1-a, is one of the most studied allergens, its structure determines the shape of many known allergens. However, the relationship across structures can only be known once a structure is resolved. Using LLMs with diverse information such as amino acid sequence and function. With this method we are able to relate distant proteins such as thebaine synthase with a very similar structure but only 15% identity in amino acid sequence.

Complexity in science and biology

> ***"Every object that biology studies is a system of systems."***
>
> — Francois Jacob, 1965 winner of the Nobel Prize in Medicine[2].

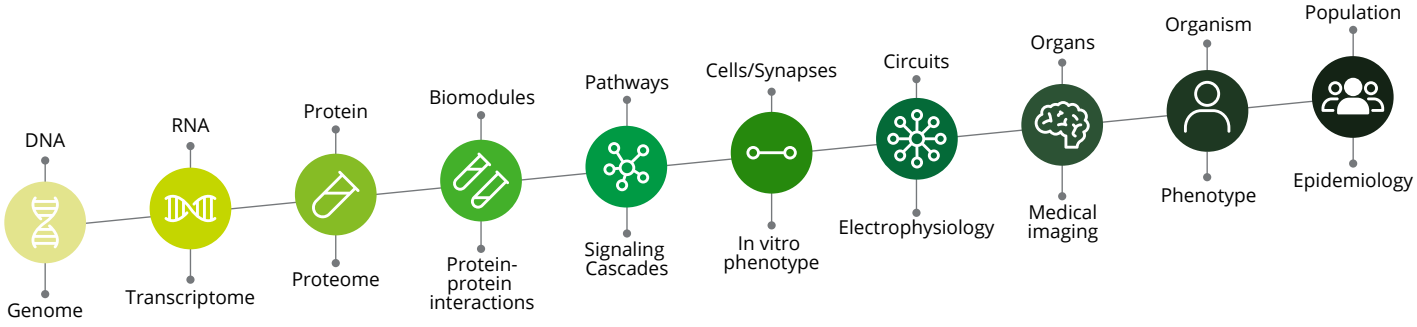
Biology, at its core, is a complex system; complex systems are made of components that interact with each other and the environment in multiple ways. These interactions make it difficult to study the components and isolation, making it impossible to infer the system's behavior as a whole by examining only its constituent elements (e.g., trying to infer brain network activity by studying a single neuron). These interactions give rise to emergence, where the properties of a complex system are generally very different from those of its constituent parts (e.g., a protein might have a different behavior alone rather than in a complex in a particular function in the cell). The notion of scale and the various types of modeling needed to address it are critical to understanding why biology is difficult to predict. It requires new models, data, and modes of thinking when moving from the lower scale of constituent parts to the higher scale of the system.

Biology is characterized by complexity that gives rise to multiple scales (see Figure 2). There is an emergence that occurs when complexes of proteins interact in the body, giving rise to higher-scale systems. **As a result, if we want to make accurate predictions across the multiple scales of biology, we require:**

1. A data collection process that captures data at multiple scales, e.g., gene and amino acid sequences, protein structures, interaction pathway involvement, etc.
2. A data model that allows us to relate relevant data at multiple scales, e.g., given a protein/ target, and shows us which pathways it is involved in, the cell types it is expressed in, etc.
3. AI models that can process data at multiple scales simultaneously.

Figure 2 *In biology, there are different scales that we can study, beginning from DNA and amino acid sequences at the protein level and going all the way to cells, organisms, and entire populations. With growing amounts of information across multi-omics data (genome, transcriptome, etc.) and in different scales, the complexity of information is not unidirectional.*

Multiple scales of biology



By integrating the insights from our exploration into the complexities of biological systems and the necessity of understanding these systems across multiple scales, we arrive at a sophisticated approach to harnessing AI for biology.

First, implementing a knowledge graph serves as a foundational data layer, enabling the seamless connection between scales—from individual proteins to complex cellular pathways and beyond—and aggregating labels within each scale.

This structure can effectively organize the vast, multidimensional datasets inherent to biological research and not only aids in navigating complexity but can also enhance our capacity to identify meaningful relationships within biological data.

Furthermore, by using **LLMs for semantic data labeling** (see Sidebar 2) within this knowledge graph, we can significantly improve the quantity and quality of labels across and within scales. ***This augmentation can increase the likelihood of accurately identifying the critical variables necessary to predict emergent features of biological systems, thereby addressing one of the primary challenges in modeling biological complexity (see Sidebars 1 and 3).***

Sidebar 3 - Scientific Pipelines to Knowledge Graph Feedback Loop

AlphaFold and hotspot identification for protein structure and surface prediction

Atlas AI's AI scientific pipelines generate predictions that can be fed back into the knowledge graph as new data points. For instance, our protein prediction pipeline incorporates AlphaFold and geometric deep learning methods to predict 3D structures and their relevant surfaces to identify druggable hotspots for targets that might have not been experimentally resolved yet. These AI-generated structures and surfaces are then ingested into the knowledge graph and can be easily queried or used as the basis for new experiments.

Finally, the application of **graph neural networks** (GNNs) capitalizes on the enriched labeling provided by LLMs, employing these labels to navigate the multi-scale and multimodal data embedded within the knowledge graph. By considering the local structure of the graph and the interconnected data it contains, GNNs can consider data that spans multiple scales of complexity and make nuanced predictions across those scales, thereby unlocking new potentials in our understanding and modeling of biological phenomena.

This integrative approach, combining knowledge graphs, LLMs, and GNNs, represents a pivotal advancement in our ability to model the intricate systems of biology, promising to accelerate progress in fields such as drug discovery by providing a complete understanding of biological processes at all scales (Sidebar 4).

Sidebar 4 – Knowledge Graph and Scientific Pipelines to LLM Feedback Loop

LLM finetuning

Atlas AI uses the data contained in the knowledge graph to finetune its LLMs, providing improved hypotheses and a richer similarity-encoding space. Datasets in the KG have been leveraged for a variety of use cases, including: (1) fine-tuning a protein language model to focus on Complementarity-determining regions (CDRs) in monoclonal antibodies, resulting in improved antibody structure prediction, and (2) training a state-of-the-art chemical language model that can identify pharmaceutically-relevant compounds and predict diverse properties of those compounds (e.g. ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity).

Integrating AI into the wet lab

The ultimate test of any predictive model's utility lies in its verification through empirical evidence. In biology, this necessitates a close integration between AI-driven predictions and wet lab experimentation. The development of automated data ingestion pipelines for experimental data, alongside the capability to link this data with both public and internal datasets, is crucial.

To this end, Atlas AI is designed to extract insights from internal datasets and ingests experimental data and electronic notebooks, playing a pivotal role in bridging the gap between AI predictions and experimental validation. This integration can improve causal inference capabilities in our models and grounds AI predictions in empirical reality.

> *By closely aligning AI-driven insights with wet lab experimentation, we can accelerate the pace of discovery and enhance the reliability of predictions across the complex landscape of biological research.*

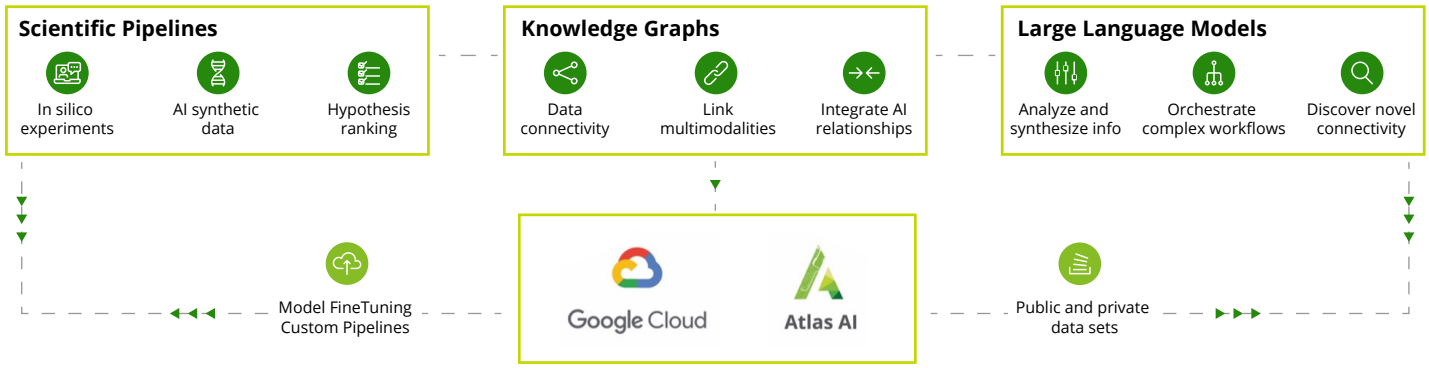
The Atlas AI™ scientific feedback loop

Atlas AI creates a knowledge graph data layer that integrates historical experimental data internal to an organization with existing public datasets, providing multiscale data connectivity. The knowledge graph is combined with AI models and LLMs, forming an innovative feedback loop structure designed to improve research and development (R&D) continuously.

The Atlas AI scientific feedback loop aims to make drug discovery more efficient over time through reversing the trend of increasing costs and timelines in the drug discovery process. The feedback loop consists of six key steps (Figure 3):

1. Data ingestion of historical experiments for connectivity and to fine-tune models.
2. AI models are fine-tuned based on historical data
3. Results from AI models and LLMs' hypotheses inform new wet lab experiments.
4. Wet lab experiments are performed, and data is collected.
5. Experimental data and electronic notebooks are fed back into the knowledge graph as new evidence.
6. The newly ingested experimental data then improves the AI models' predictive performance and LLMs' hypotheses.

Figure 3 *The Atlas AI scientific feedback loop integrates public and private data, informs of new hypotheses using LLMs, and ingests experimental evidence to improve the performance of AI models.*



SECTION 3

Science & Technology in tandem: building blocks for transformation

Atlas AI, on the foundation of Google Cloud, is transforming drug discovery by harnessing the power of cutting-edge AI and secure, scalable infrastructure. This powerful combination can empower researchers with the tools they need to accelerate breakthroughs (Figure 4).

Google Cloud's end-to-end Generative AI (GenAI) ecosystem

Vertex AI: As an end-to-end enterprise-ready predictive and GenAI platform, Vertex AI equips Atlas AI with access to advanced large language models like Gemini. These powerful LLMs empower researchers to harmonize complex data sets, unlock profound scientific insights, and explore groundbreaking hypotheses that could lead to new drug discoveries.

Streamlined model development: Vertex AI's ML operations (MLOps) tools streamline collaboration and model development. Features like model evaluation, orchestrated workflows, comprehensive model management, efficient feature handling, and robust model monitoring are all housed within a single, unified platform, fostering a seamless and efficient research process.

Google Cloud's comprehensive toolkit for unveiling new possibilities in drug discovery

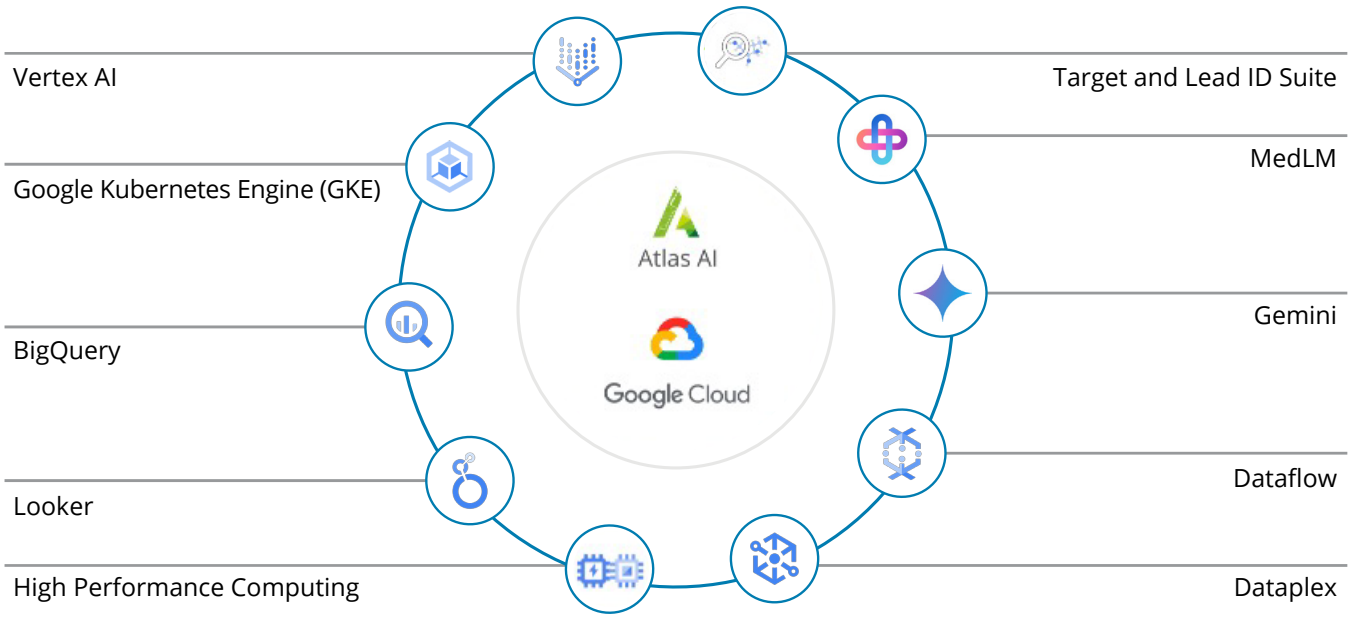
Target and Lead Identification Suite: This suite provides researchers with essential tools like AlphaFold, allowing for precise protein structure analysis and mutation prediction, which are crucial steps in identifying promising drug targets.

Generative AI Powerhouse: Atlas AI seamlessly integrates with NVIDIA's BioNeMo Platform, a state-of-the-art platform offering large-scale model training and essential tools designed specifically for drug discovery, such as DiffDock, MolMIM, and MoFlow.

Unleashing the power of high-performance computing: Direct integration with Google's elastic high-performance computing infrastructure removes limitations. Researchers can easily tackle large-scale scientific calculations like high-throughput virtual screenings, molecular dynamics, and lead optimization techniques.

Data-driven insights: BigQuery and Looker provide the foundation for enterprise-grade data warehousing, secure data sharing, and insightful business intelligence analytics. This allows researchers to make informed decisions based on a comprehensive understanding of their data.

Figure 4 *Google Cloud provides a foundation layer to support Atlas AI, providing access to diverse tools and capabilities.*



SECTION 4

The path forward, connecting with your needs

Atlas AI harnesses a unique perspective to solve current needs in drug discovery, using AI to connect both hypothesis and data-driven analysis pipelines, weaving together public and private experimental data that can be customized for the client.

To start a project, our team can walk you through common use cases in the drug discovery space or fully customize it to your needs. Deloitte in collaboration with Google can provide multi-modal strategy advisory to integrate research data using Large Language Models and a series of diverse AI models according to the use case needed.

Our team will work with you to help optimize your data analysis to its full potential, utilizing the power of AI, critical thinking, and scientific methodologies to drive AI-driven hypothesis generation and discovery.

Authors

Daniel Ferrante, PhD
Managing Director and RnD Innovation Leader
Deloitte Consulting LLP

Annabel Romero Hernandez, PhD
AI Leader for Drug Discovery
Deloitte Consulting LLP

Punit Sahni
Cloud Sales & GTM Leader
Deloitte Consulting LLP

Vincent Beltrani, PhD
Enterprise Customer Engineer:
Healthcare, Life sciences and Quantum AI
Google Cloud

Chris Hayduk, MSc
Lead Machine Learning Engineer
Deloitte Consulting LLP

Parastou Eslami, PhD
Specialist Leader for AI in Healthcare and Life Sciences
Deloitte Consulting LLP

Michael Koetting, PhD
Lead Data Scientist
Deloitte Consulting LLP

Paul Prohodski
Specialist Leader, AI and Data Strategy
Deloitte Consulting LLP

Acknowledgements

The authors would like to thank Deepak Kannangala, Peter Hunt, Amber DeSimone and Steve Peters for their inputs and dedicated efforts in bringing this paper to life.

1. Congressional Budget Office, Nonpartisan Analysis to the U.S. Congress. Research and Development in the Pharmaceutical Industry, 2021. (https://www.cbo.gov/publication/57025)

2. M. De Domenico, D. Brockmann, C. Camargo, C. Gershenson, D. Goldsmith, S. Jeschonnek, L. Kay, S. Nichele, J.R. Nicolás, T. Schmickl, M. Stella, J. Brandoff, A.J. Martínez Salinas, H. Sayama. Complexity Explained (2019). DOI 10.17605/OSF.IO/TQGNW