



Atlas AI:

Unlocking the Power of GenAI for Drug Discovery

Daniel Ferrante, PhD

dferrante@deloitte.com

March 18, 2024



Mapping Data to Knowledge

SFL Scientific a Deloitte Business M2D2 Team

BY THE NUMBERS

100%

U.S.-Based

80%

have Ph.Ds.

100%

have a graduate degree

MISSION & VISION

Reimagining the entire drug discovery process from first principles with a **Scientific Data-Centric & AI-First approach**.

We customize our innovative drug discovery solutions to optimally fit your data, processes, and workflows, providing seamless integration and accelerated time-to-insights.

With **expertise across AI/ML, MLOps, & data engineering as well as deep, hands-on expertise in drug discovery**, we are uniquely qualified to develop and deploy these customized solutions to improve bench scientist workflows.

DRUG DISCOVERY LEADERSHIP



Daniel Ferrante, PhD



Annabel Romero, PhD



Chris Hayduk, MSc



Michael Koetting, PhD

SOLUTIONS & CAPABILITIES



AI & Machine Learning

Daniel Ferrante
Chris Hayduk
Michael Koetting



R&D Strategy

Daniel Ferrante
Annabel Romero



DataOps & AIOps

Daniel Ferrante
Chris Hayduk



Drug Discovery

Annabel Romero
Michael Koetting

NVIDIA and DGX Cloud

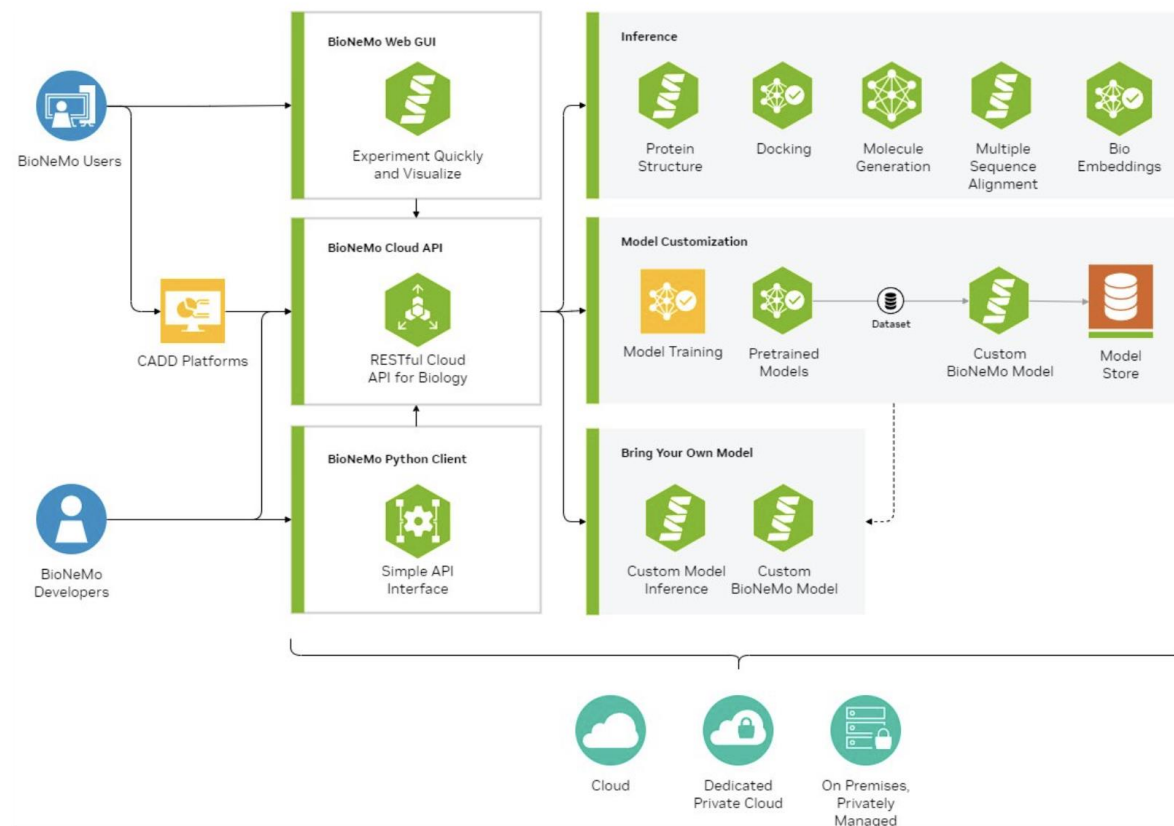
ENABLING CUTTING EDGE TECHNOLOGY IN DRUG DISCOVERY & DEVELOPMENT

➤ SFL Scientific, a Deloitte Business:

- **Started** NVIDIA's AI and DS partnership track back in Summer **2017**.
- **Using DGX Cloud** for *months*:
 - Creating our **ChemLLMs**.
 - Creating our **retraining** pipeline for BioNeMo pLMs with **QLoRA**.
 - Creating pipelines for BioNeMo's **MolMIM** and **Oracles** using our **ChemLLM embeddings**.

BioNeMo Cloud Architecture

High Level Architecture



Partnerships: Atlas AI and NVIDIA

With the AI capabilities of NVIDIA paired with SFL Scientific, a Deloitte Business' unrivaled domain and industry expertise, we are setting the standard in the market on how to imagine a new customer experience and unparalleled operational efficiency.

- **Business and Strategy Expertise:**

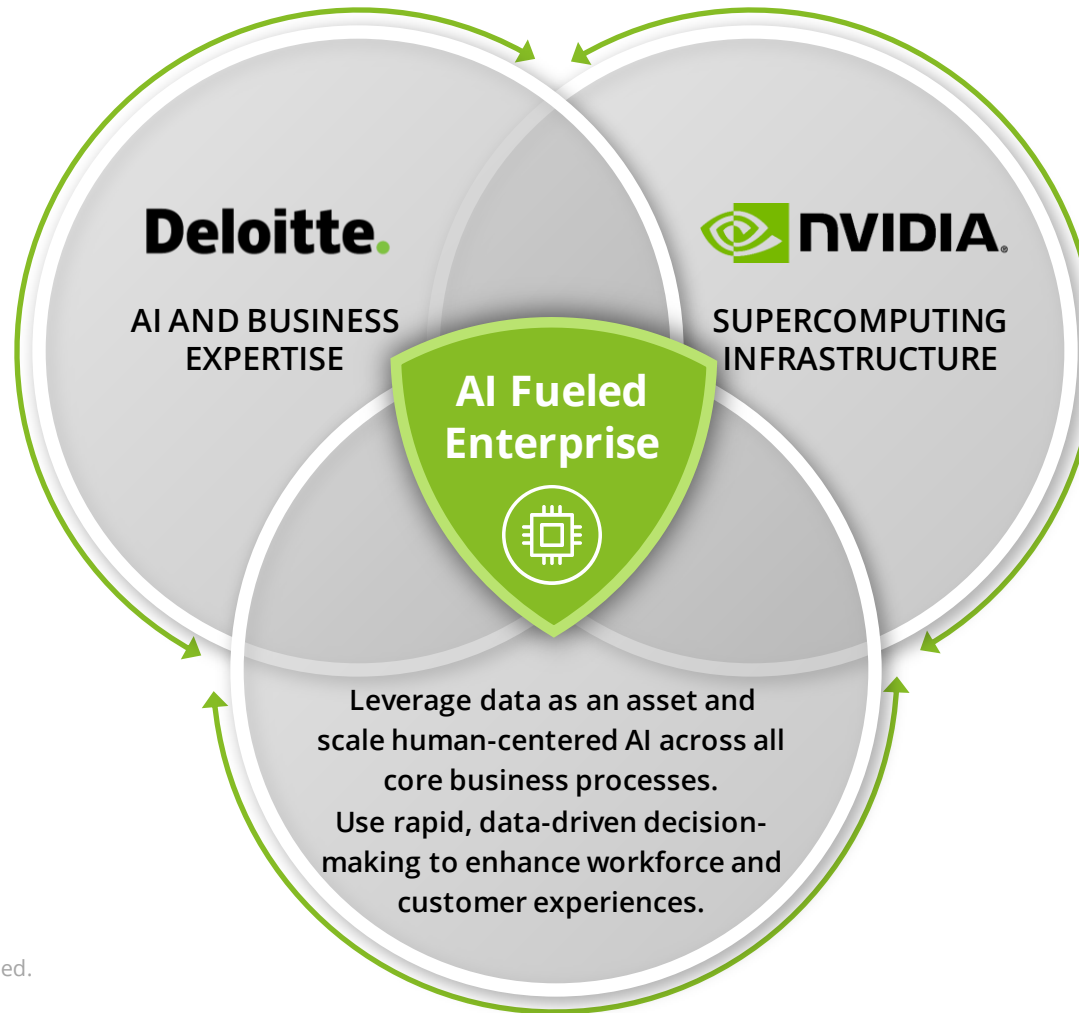
Sector and domain expertise with **an account and insight-led strategy** that drives large scale business transformation.

- **AI Talent and Scale:**

Breadth and depth of industry expertise in the application of AI and data science. **2,000+ data scientist practitioners with 250+ NVIDIA GPU-certified** and growing

- **Intellectual Property**

The **Deloitte Center for AI Computing** underpinned by Trustworthy AI Framework and assets across domain, industry, ML Models



- **Platform:**

AI Hardware and Supercomputing infrastructure

- **Expertise:**

GPU, Omniverse, DGX and NVAI Enterprise. NVIDIA is the leader in full stack AI solutions from systems, cards, AI software, and AI skills.

- **Solution Accelerators:**

Pre-trained **ML Models and Scripts** and a vast Industry Solutions Catalogue

- **Market position:**

Leading market solutions for today's emerging challenges. **Strong AI brand power** to align with longstanding leadership of Deloitte.

Unmet Need: Data & Metadata Harmonization

Data integration guarantees the **maintenance of corporate knowledge** as well as an **organization across functional teams**.

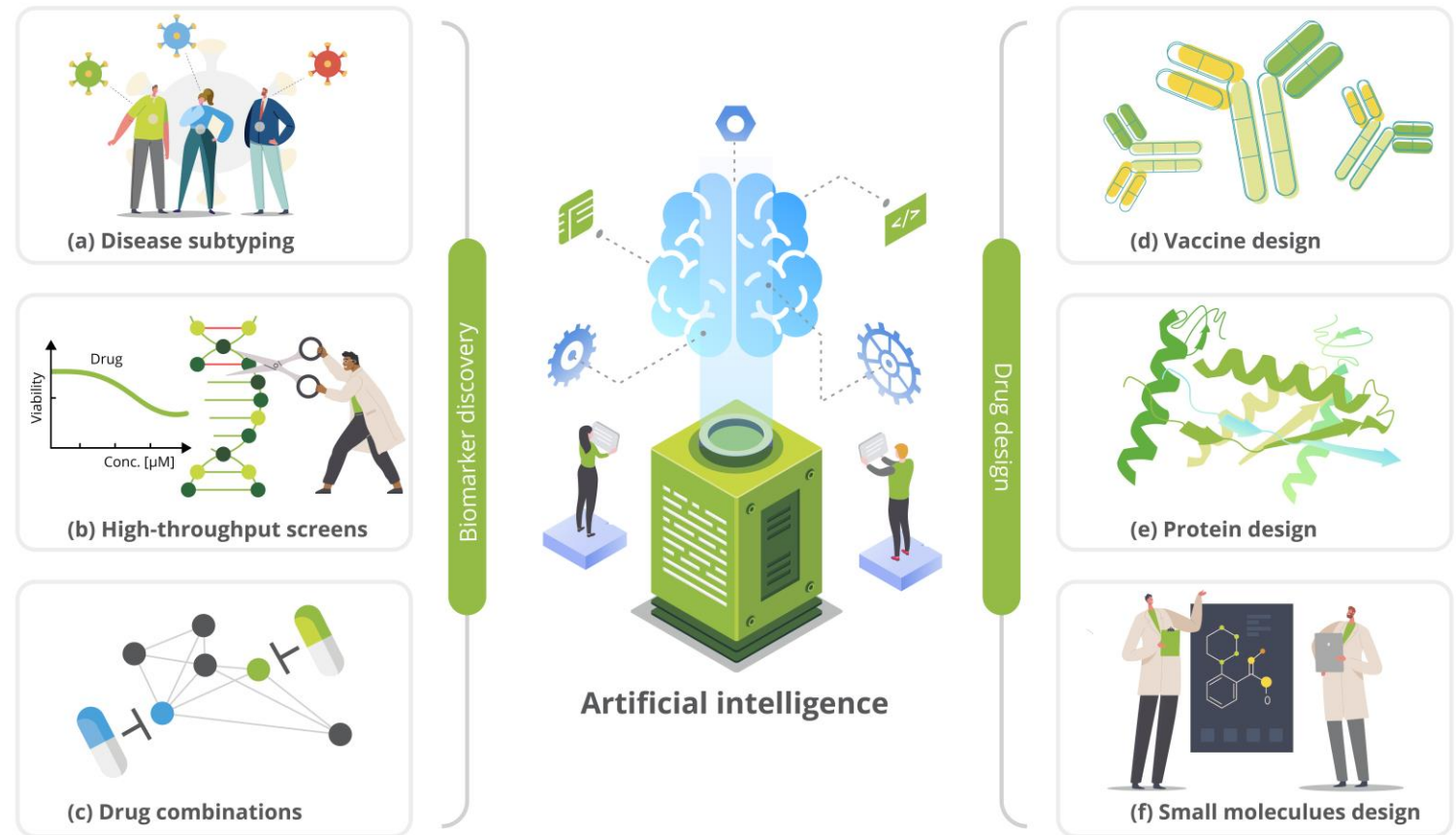
Challenge (Data is splintered): /data and experiments are designed to answer one or a few questions. If the experiment fails, most of the time the result is not considered further.

Drug Discovery Data is an Interconnected Network (multimodal):

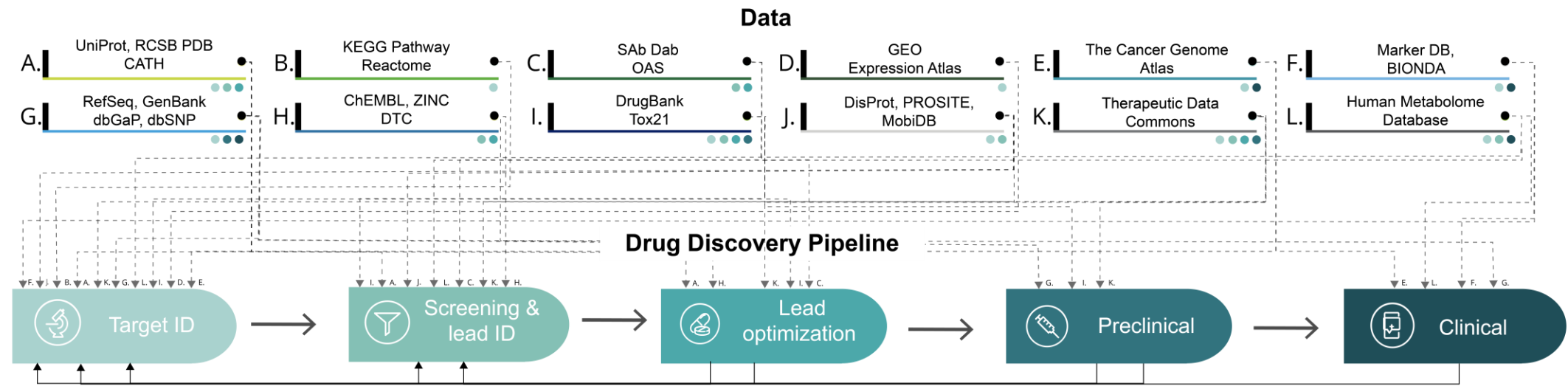
Implementation of a data aggregation tool to sift through the experimental disconnected data to generate a general purpose, centralized, and organized knowledge graph.

Data integration based on applications:

Data/Experiments considered for a single application or department can be interconnected with other departments for further acceleration of drug discovery research.



Data harmonization to accelerate AI-based Drug Discovery



Scientific Pipeline Opportunity

— ● A. ● J.

DATA: UniProt, RCSB PDB, CATH, DisProt, PROSITE, MobiDB
SOLUTION: STRUCTURAL BIOLOGY
TOOL: Protein Structure Prediction
PPI Hotspot Prediction
Protein Knowledge Graph

— ● B.

DATA: KEGG Pathway Database, Reactome
SOLUTION: STRUCTURAL BIOLOGY & MULTIOMICS
TOOL: PPI Hotspot Prediction
Pathway Knowledge Graph

— ● C.

DATA: OAS, SAbDab
SOLUTION: IMMUNOTHERAPEUTICS
TOOL: Antibody Structure Prediction
Antibody Knowledge Graph

— ● D. ● G. ● L. ● E. ● F.

DATA: RefSeq, GenBank, dbGaP, dbSNP, GEO, Exp. Atlas, HMdb, TCGA, MarkerDB, BIONDA
SOLUTION: MULTIOMICS
TOOL: Gene disease association, biomarker KG
Expression, Metabolite KG

— ● H. ● I. ● K.

DATA: ChEMBL, ZINC, DTC, DrugBank, Tox21, TDC
SOLUTION: CHEMOINFORMATICS
TOOL: Binding, ADMET/QSAR, HTS prediction
Chemical properties and effects

— ● A.

DATA: UniProt, RCSB PDB, CATH
SOLUTION: STRUCTURAL BIOLOGY
TOOL: Protein Structure Prediction
PPI Hotspot Prediction
Protein Knowledge Graph

— ● C.

DATA: OAS, SAbDab
SOLUTION: IMMUNOTHERAPEUTICS
TOOL: Antibody Structure Prediction
Antibody Knowledge Graph

— ● I.

DATA: DrugBank, Tox21
SOLUTION: CHEMOINFORMATICS
TOOL: Efficacy and Safety - Toxicity prediction

— ● E.

DATA: The Cancer Genome Atlas
SOLUTION: MULTIOMICS & IMAGING
TOOL: Pathology and gene biomarker association KG

— ● F.

DATA: MarkerDB, BIONDA
SOLUTION: MULTIOMICS
TOOL: Biomarker discovery and characterization KG

— ● L.

DATA: Human Metabolite Database
SOLUTION: MULTIOMICS
TOOL: Metabolite KG

— ● G.

DATA: GenBank
SOLUTION: MULTIOMICS
TOOL: Gene - Protein - Disease KG

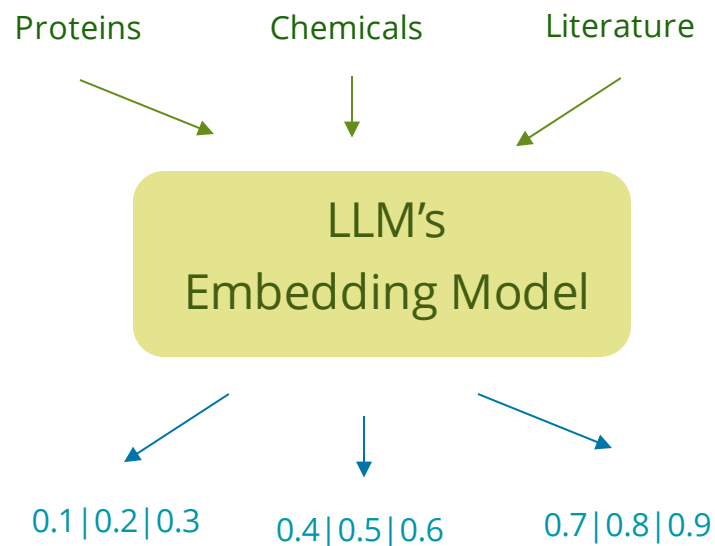
— ● K.

DATA: Therapeutic Data Commons
SOLUTION: IMMUNOTHERAPEUTICS
TOOL: Antibody affinity prediction

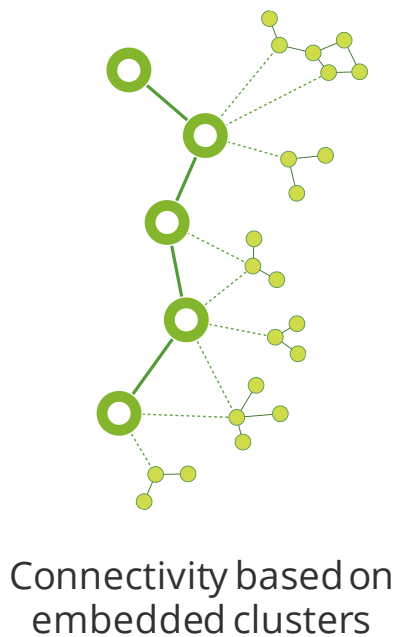
Generative AI and Knowledge Representation & Reasoning (KRR)

Complexity Science: Bringing together LLM's, Knowledge Representation and Networks

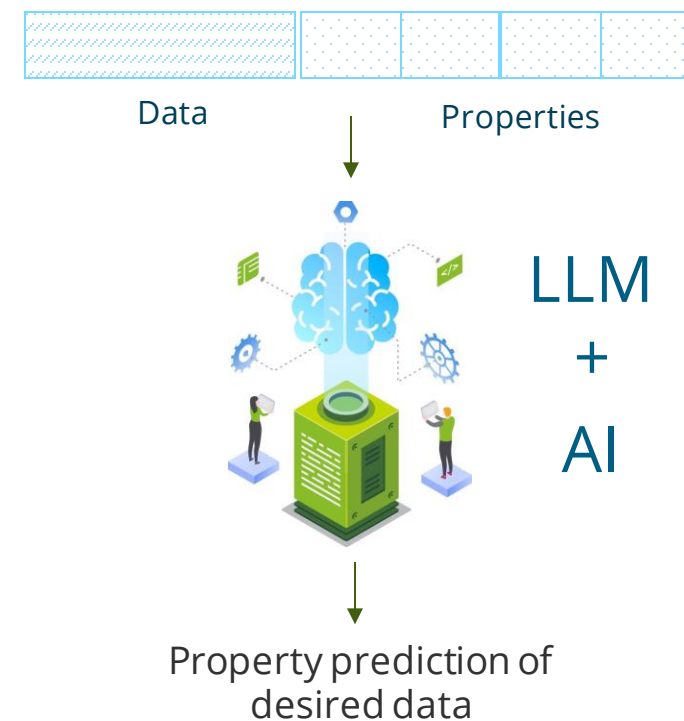
Data ingestion



Knowledge Graph Integration



Model Fine-tuning



Atlas AI | Empowering Scientists with Data-Centric AI

Atlas AI is an innovative, AI-powered web application that aims to empower scientists to generate & validate high-quality hypotheses quickly through three key pillars:

1. Enhanced Data Connectivity with LLM Integration

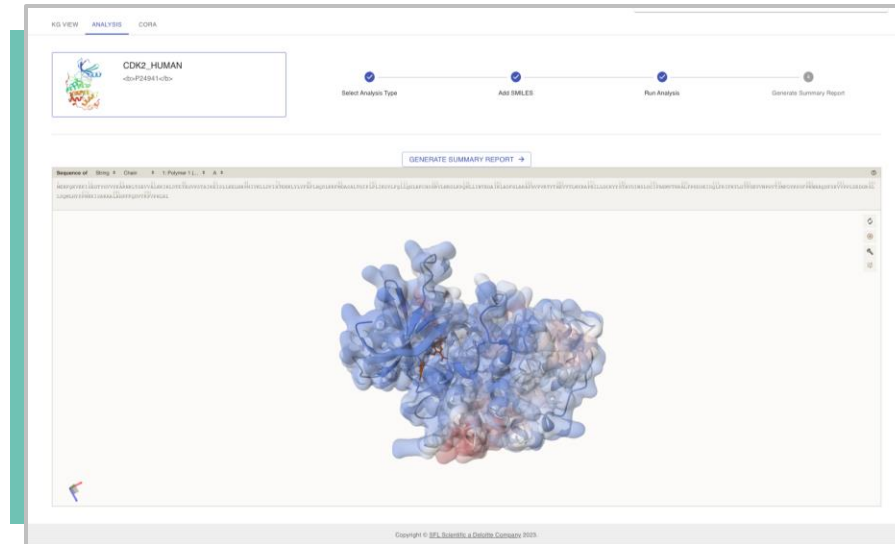
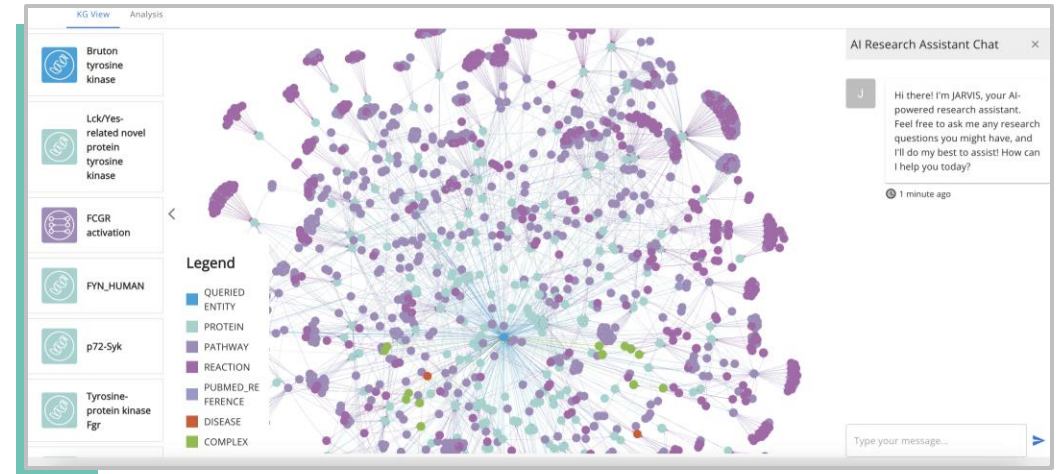
Atlas AI's Knowledge Graph **captures connections between dozens of different data sources and generates new hypotheses using LLMs.**

2. State-of-the-Art Generative AI

Assist researchers in quickly simulating scientific experimental pipelines, allowing rapid validation of initial hypotheses.

3. Intuitive User Interface

Atlas AI is designed from a scientist-first perspective. We expose our data, language models, and **AI-powered scientific pipelines in a no-code interface.**

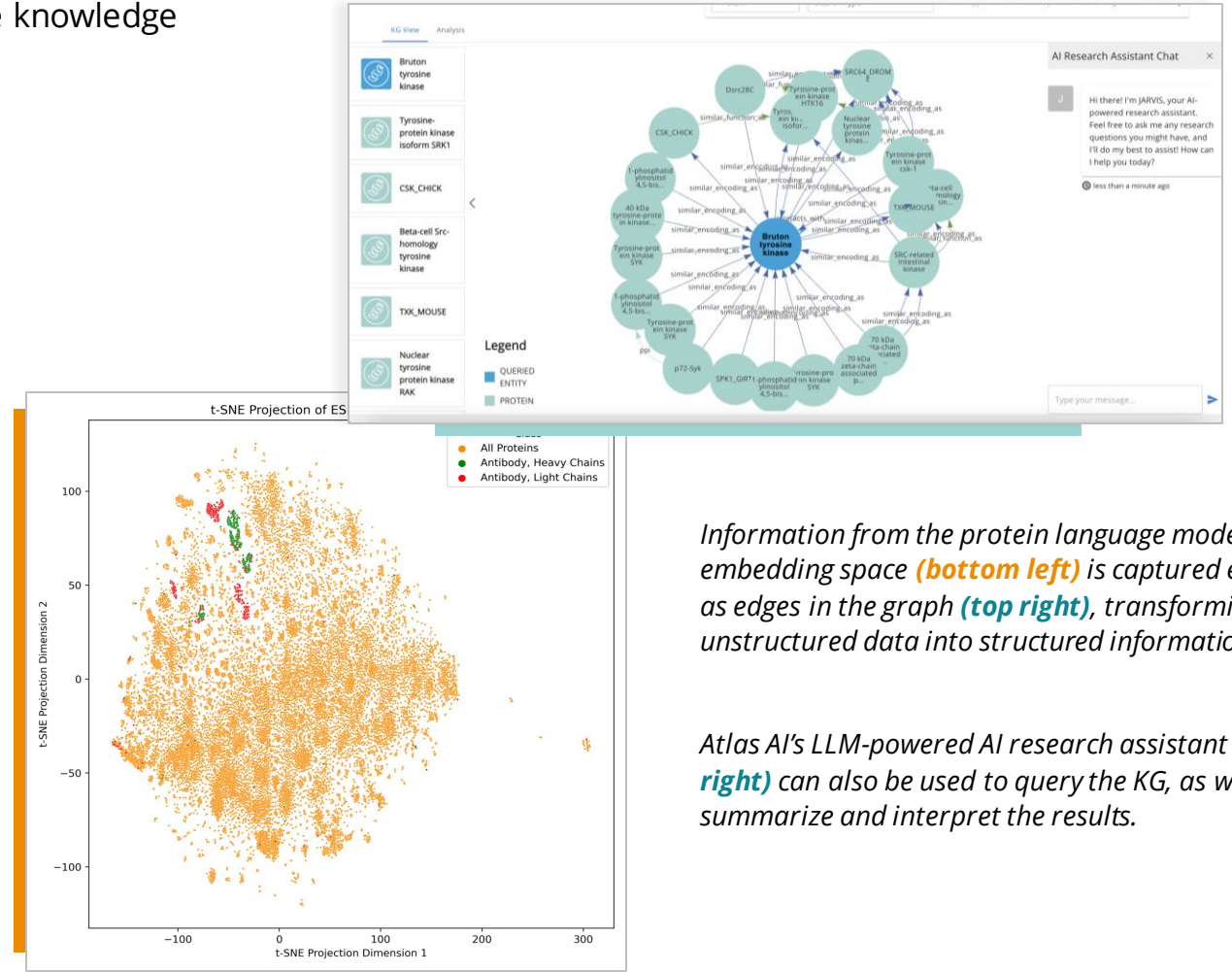


Atlas AI | Combining Knowledge Graphs and Generative AI

Connecting multimodal data to enrich information & create knowledge

Key features of Atlas AI's Knowledge Graph (KG) + Large Language Models (LLMs):

- ✓ Atlas AI's knowledge graph provides **real-world grounding for the LLM**, avoiding hallucinations
- ✓ Non-technical users unfamiliar with database query languages can **use natural language to interact with their data**
- ✓ Atlas AI's knowledge graph provides **deeper levels of connections and relationships** between datapoints, surfacing **unexpected insights** across domains
- ✓ Atlas AI's LLM can interpret user messages and **call other tools and APIs to help answer questions.**



Information from the protein language model embedding space (**bottom left**) is captured explicitly as edges in the graph (**top right**), transforming unstructured data into structured information.

Atlas AI's LLM-powered AI research assistant (**top right**) can also be used to query the KG, as well as summarize and interpret the results.

Atlas AI | Knowledge Graph by the numbers

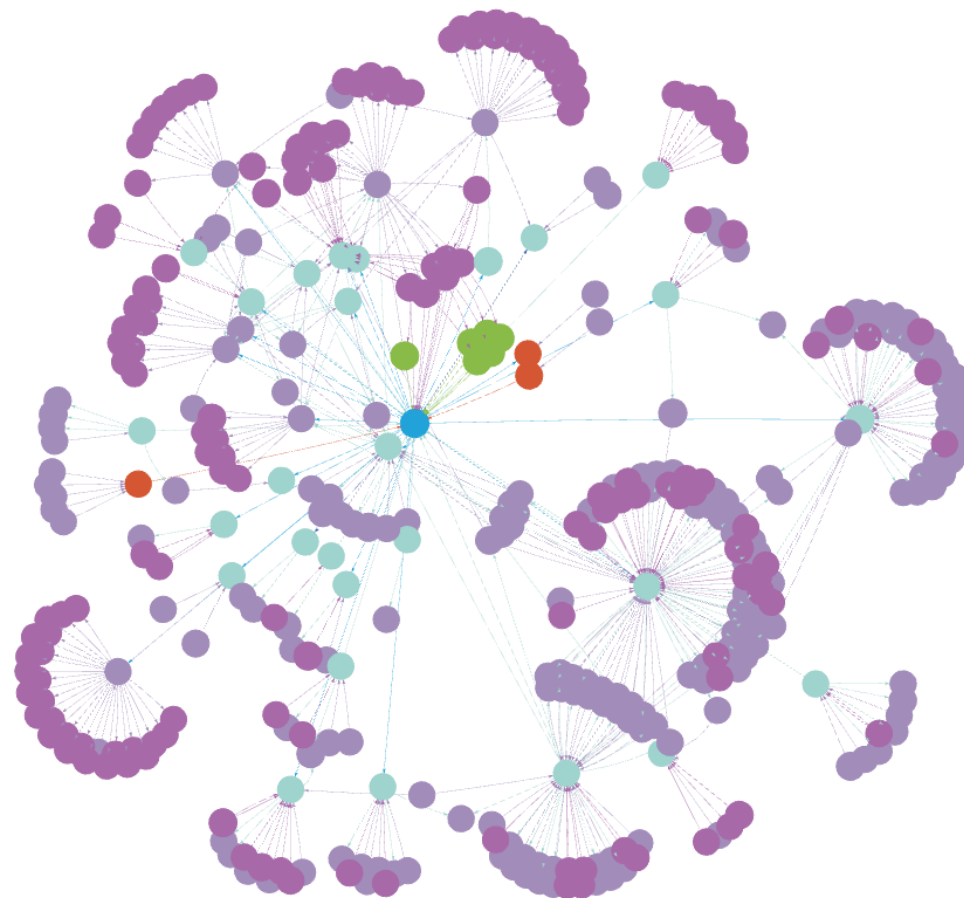
SUMMARY STATISTICS

on current version of knowledge graph:

11,229,138 total nodes and **96,355,621 relationships**, comprised of:

- **794,847** proteins
 - including 102,402 variants, 89,754 mutations, and 885 antibodies
- **8151** functional protein domains
- **431,029** detailed descriptions of protein function
- **7,630,236** chemicals/drug compounds
- **17,372** diseases and ailments
- **89,767** biological reactions across 22,020 biological pathways
- **1,386,431** drug-target pairs with experimental binding data
- **338,376** total PubMed citations across 121,267 unique PubMed references
- **2,356,105** distinct patents connected to compounds in the knowledge graph
- Links to **142,469** experimentally-determined protein structures

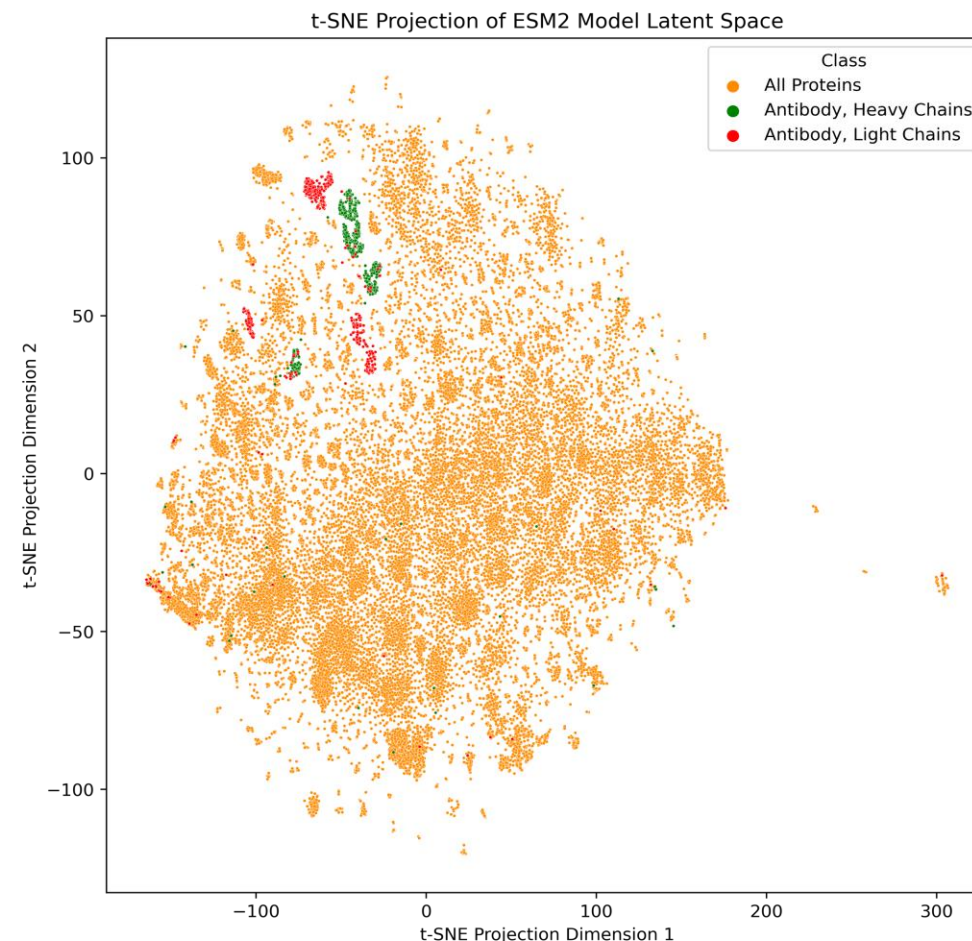
All the above is easily searchable within seconds!



Atlas AI | NVIDIA BioNeMo Model Finetuning

Deloitte is currently leveraging the advance BioNeMo finetuning framework to create a series of state-of-the-art models designed to improve understanding of antibody structure and function.

1. **Antibody Sequence Embedding** – ESM-2 clusters the light chains and heavy chains of antibodies well (see image), but this suggests that it may have trouble distinguishing fine-grained details of the sequences. We will finetune ESM-2 on antibody sequence data to improve separation.
2. **Antibody Structure Prediction** – using the improved antibody embeddings, we will finetune ESMFold to improve downstream structure prediction of antibodies
3. **Protein Function Prediction Model** – we will finetune ESM-2 with a classifier network attached to predict a protein's biological function using its sequence alone



ESM-2 clusters antibody sequences well, but the tight clustering of antibody heavy chains suggests that it may struggle to distinguish differences in functionality and structure

Improving ESM-2 & ESMFold for antibody prediction

ESM-2 encodes significant protein information but lacks focus on key aspects of antibody structure

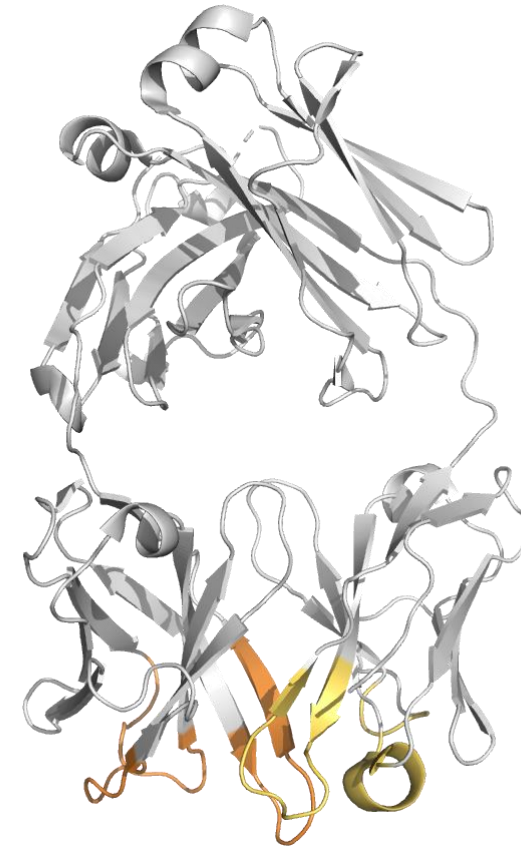
Antibody sequences have large conserved portions consistent across antibodies, with **small changes in variable regions known as CDRs determining the vast majority of the difference** in functionality.

Since CDRs make up only a small percentage (about 15%) of amino acids in the antibody's heavy chain, **ESM-2 cannot distinguish these large functional differences**, as it applies equal weight to all amino acids in a sequence.

Deloitte has worked in conjunction with the NVIDIA BioNeMo and BCP teams to address these shortcomings in ESM-2 and **develop state-of-the-art training pipelines.**

IMPROVEMENT STEPS

- ✓ **Completed:** Curate a dataset of millions of paired antibody sequences from the Observed Antibody Space
- ✓ **Completed:** Finetune ESM-2 to predict amino acids in the paired antibody sequences as well as to predict which amino acids belong to each CDR region.
- ✓ **In progress:** Use the newly-finetuned ESM-2 as the new base of ESMFold. Finetune ESMFold with its new base for antibody structure prediction



CDR loops (colored in the image) determine antibody binding but represent a small percentage of the overall structure, making accurate prediction difficult. Deloitte has trained ESM-2 to automatically label these critical regions in novel antibodies.

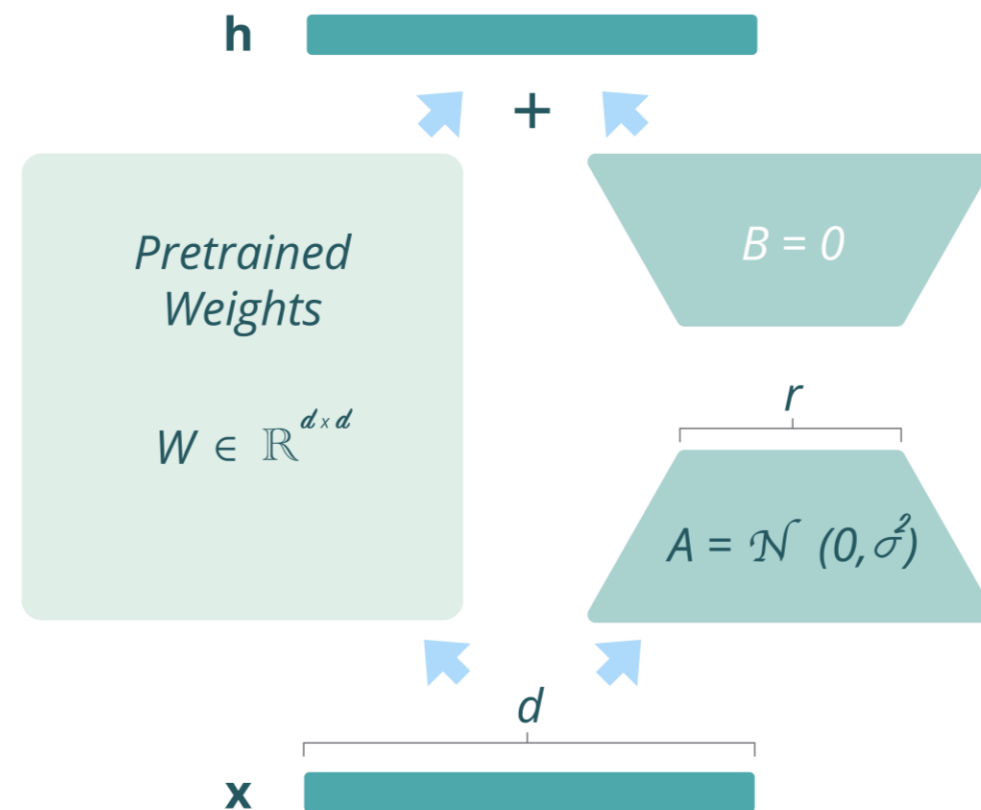
Finetuning Protein Language & Structure Models on Nvidia BCP

With the assistance of the NVIDIA BCP and BioNeMo teams, **Deloitte has become the first team** to implement a scalable, distributed finetuning framework for **protein language models using the QLoRA training paradigm**.

QLoRA training allows for memory efficient finetuning of large (10+ B parameter) protein language models, allowing research groups to cheaply and efficiently:

- Improve amino acid prediction for targeted protein groups
- Adapt ESM-2 for novel downstream tasks such as gene ontology label prediction or CDR region identification

In addition, Deloitte is working with the NVIDIA BCP and BioNeMo teams to become **the first group with to successfully finetune ESMFold for improved antibody structure prediction**.



QLoRA adds low rank matrices (A and B in the image above) to approximate updates to the frozen weight matrix with a smaller memory footprint.

ESM-2 Antibody Model Finetuning

Deloitte has leveraged the advanced BioNeMo finetuning framework on NVIDIA BCP hardware to create an **improved ESM-2 model for antibody encoding and understanding**.

Deloitte's ESM-2 antibody model:

- **Reduces errors when predicting amino acids by an order of magnitude** when compared to the original ESM-2 model
- Provides **automated data labeling** for antibody sequences by predicting CDR regions

The Deloitte team aims to use the NVIDIA BioNeMo and BCP frameworks to train **a state-of-the-art antibody-specific structure prediction model based on ESMFold, utilizing the representation learned by our custom-developed ESM-2 antibody model**.

ESM-2 Antibody Finetuning Results

Validation Loss After 8 Epochs of QLoRA Fine-Tuning of ESM2 (15B) on Paired OAS Antibody Data		Model	
		ESM2 + CDR Predictor	Control: ESM2 before fine-tuning
Validation Mask Type	CDR-Only Mask	0.04533	0.3774
	Full Mask	0.05692	0.4054

Our finetuned ESM-2 model is able to significantly reduce errors when predicting amino acids in antibody sequences when compared to the base ESM-2 model.

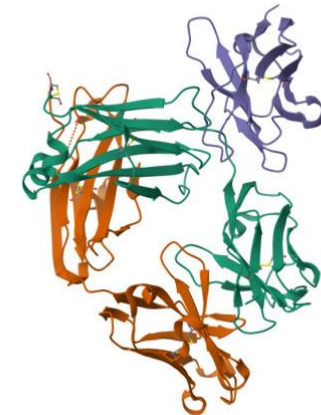
Beyond Self-Supervised Learning & Masked Language Modeling

- **ESMFold provides a major speed advantage** over state-of-the-art multiple sequence alignment (MSA)-based models such as AlphaFold2 by using the ESM-2 protein language model to obviate the costly MSA step.
- A key component of ESM-2 is its **self-supervised training** setup, in which it is tasked with **predicting missing amino acids in a protein sequence**. This gives ESM-2 a strong representational capacity but, critically, **ignores the detailed attributes that have been experimentally derived** and verified by researchers.
- **Deloitte's seeks to be the first team to successfully repurpose the ESMFold model**, using our **finetuned ESM-2 model for antibodies** to combine masked language modeling with experimentally-determined labels.
 - We hypothesize that training the ESM-2 encoder to predict CDR regions within the antibody sequence forces the model to learn a **richer feature space for therapeutic antibody development than could be achieved by masked language modeling alone**
- By forcing the model to encode this extra information regarding the CDR regions within the model's hidden state, we believe we will gain a **substantial advantage on downstream tasks** that are highly dependent on the CDR region:
 - Antibody structure prediction
 - Binding to a therapeutic target or antigen

EVQLLESGGGLVQPGG...



Sequence
To
Structure



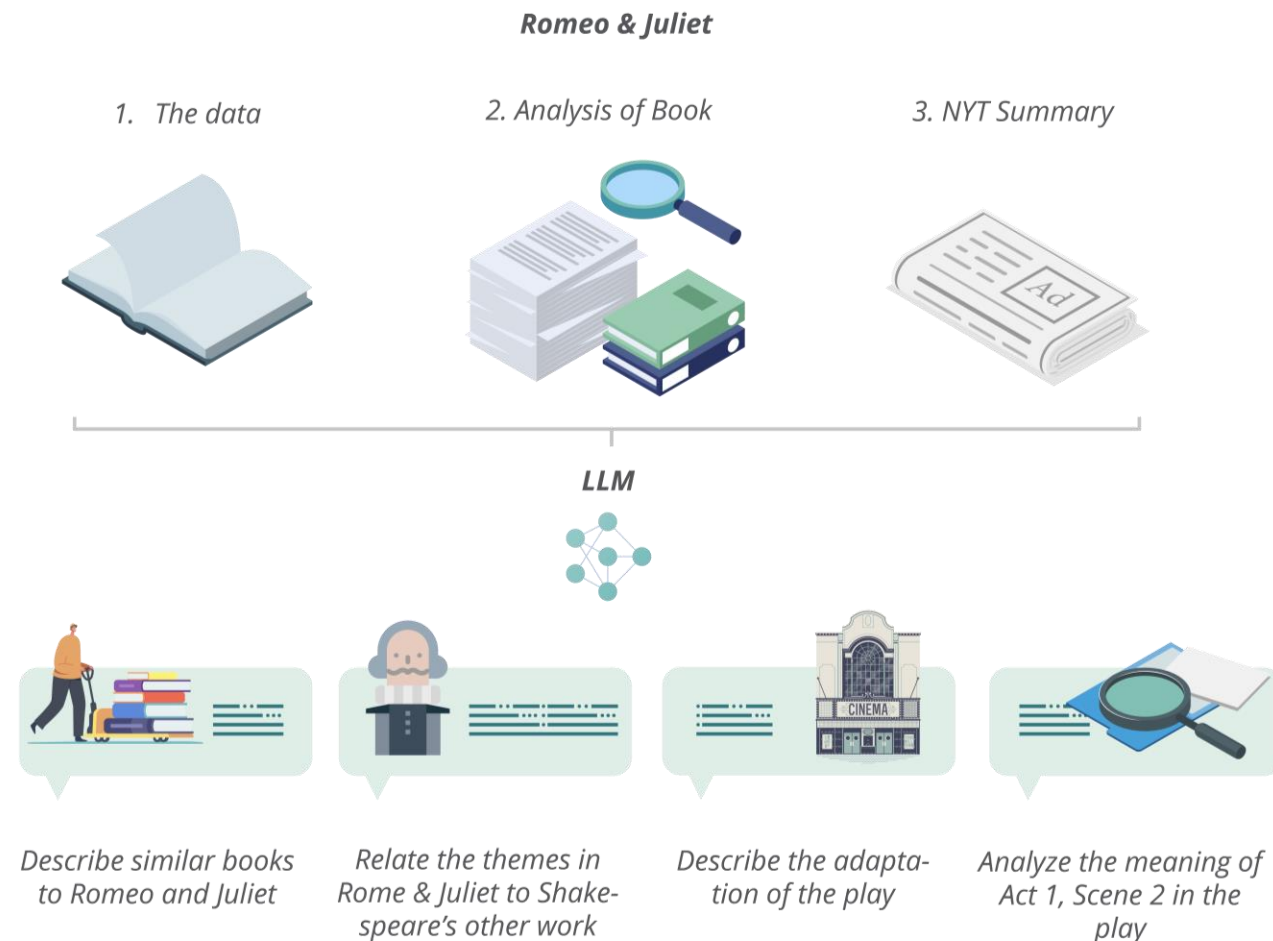
With our customized ESMFold model, Deloitte will enable scientists to move directly from antibody sequence to antibody structure, yielding greater prediction accuracy than the pretrained ESMFold and much higher throughput than MSA-based methods (e.g., AlphaFold2).

Multi-Scale Data & The Future Trajectory of Drug Discovery AI

Large language models (LLMs) based on natural language have a significant advantage – language data is self-referential. That is, within a text dataset, there will be pieces of text that refer to the same thing at different levels of abstraction. Examples include:

- **The text itself:** The text of a famous book or play, like *Romeo & Juliet*
- **A higher-scale representation:** Reviews or overview text that summarize the play
- **A classification of the text:** Text that states the genre of *Romeo & Juliet*
- **A lower-scale representation:** Detailed essays that spend pages analyzing a few lines from *Romeo & Juliet*

Since LLMs are trained on mountains of self-referential data like the above, they are able to learn multiple scales of representation for a wide range of concepts providing them with strong generalization and an approximation of human reasoning. However, **this is not the case for drug discovery language models.**



Multi-Scale Data & The Future Trajectory of Drug Discovery AI

BIOLOGICAL DATA IS INHERENTLY MULTISCALE

Unlike text, where, if we ingest enough of it, we see nearly every "scale" of thought, **it is impossible to move from one scale to another in biology by only looking at a single data modality.**

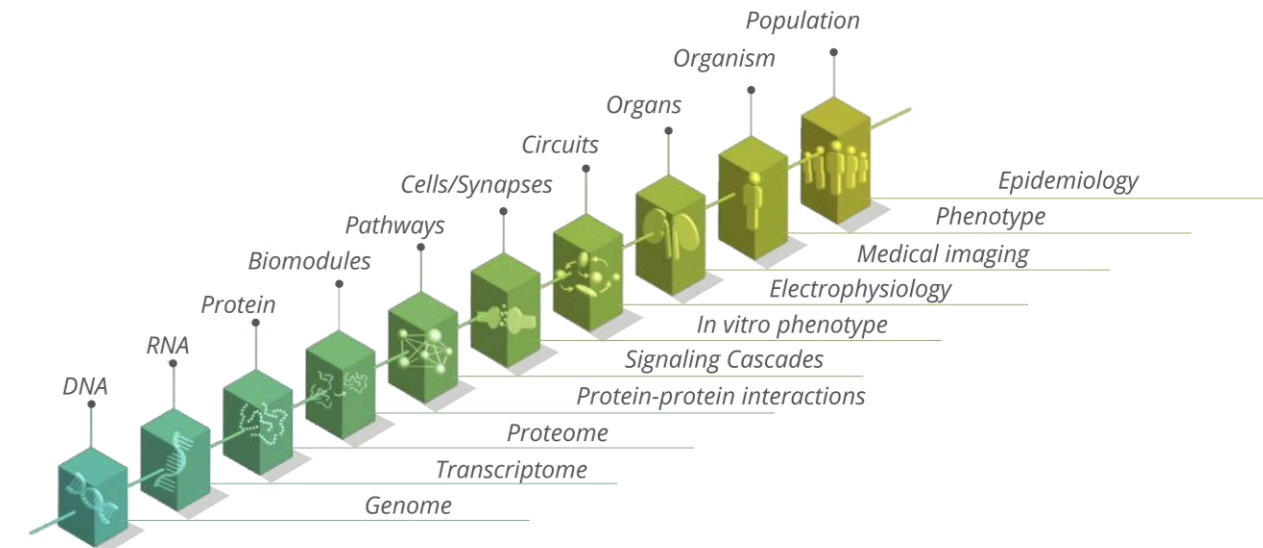
Addressing this problem requires a fundamental shift in how we approach training drug discovery language models.

Deloitte's - NVIDIA BioNeMo and BCP team take the first steps to bring multi-scale data to the realm of drug discovery:

- Our ESM-2 antibody learns not just the amino acid sequence, but also how to label the amino acids' presence in the CDR region
- A second ESM-2 model we are developing predicts amino acids in a sequence and **global** labels of the protein's overall function

Extending these approaches, with the computing power and frameworks provided by NVIDIA, will lead to drug discovery language models that are far more robust and encode a much richer understanding of biology.

Multiple scales of biology



The multiscale nature of biological data. DNA & RNA code for proteins, proteins bind to form cells, cells are organized into tissues, etc.

Ensuring exclusivity of the drug market using GenAI

Use case: AI optimization of a drug soon to lose patent exclusivity

BACKGROUND

Revlimid (Lenalidomide) is a cancer drug for treatment of multiple myeloma, owned by Bristol Myers Squibb.

Revlimid netted \$12.1 billion annually in sales as of 2020. The upcoming **loss of US patent exclusivity** (2025-2026) will likely lead to the **loss of much of this market**.

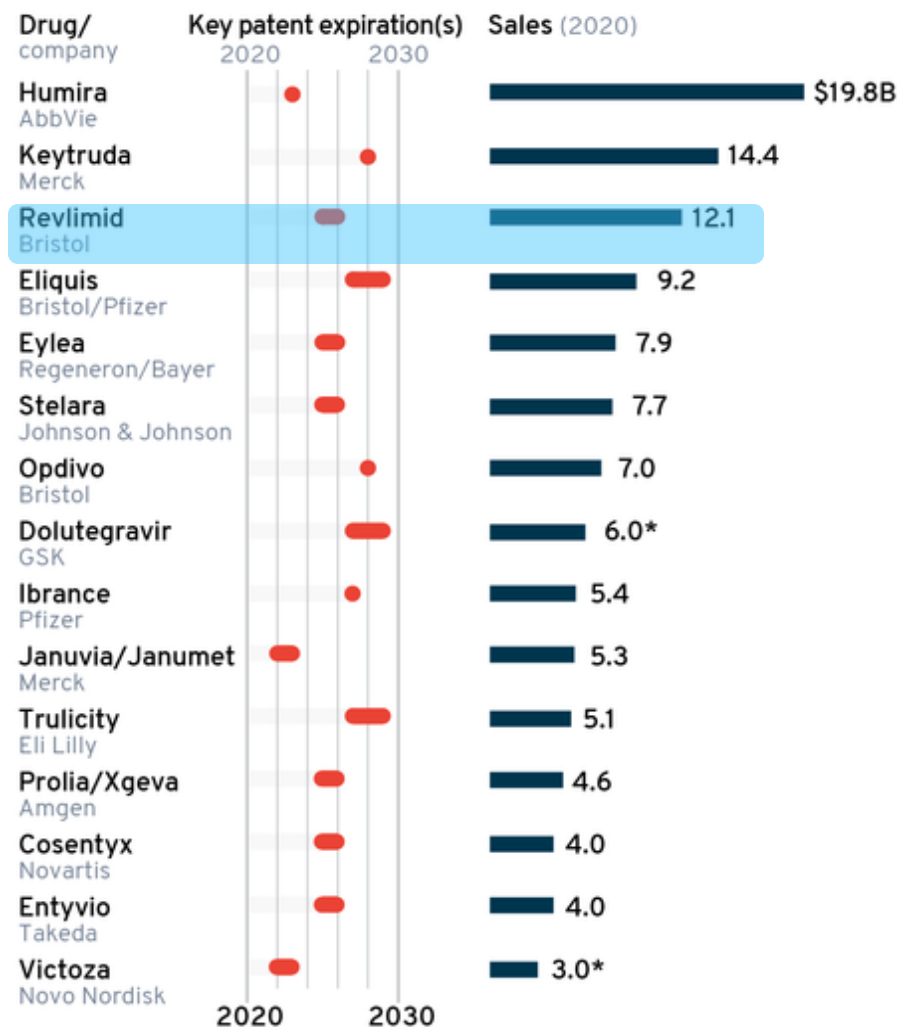
GENAI MOLECULE OPTIMIZATION

To recapture much of this market while also improving patient outcomes, we can design *in silico* a **better, patentable version** of the molecule that binds its target protein (cereblon) more potently while also exhibiting higher oral bioavailability:

- Improved drug-target binding
- Enhanced solubility
- Greater intestinal permeability
- Longer half-life *in vivo*

Major drugs set to lose patents in next decade

The 15 top selling drugs facing expirations pulled in more than \$100 billion in sales last year.



*Estimated

Source: Moody's and company filings

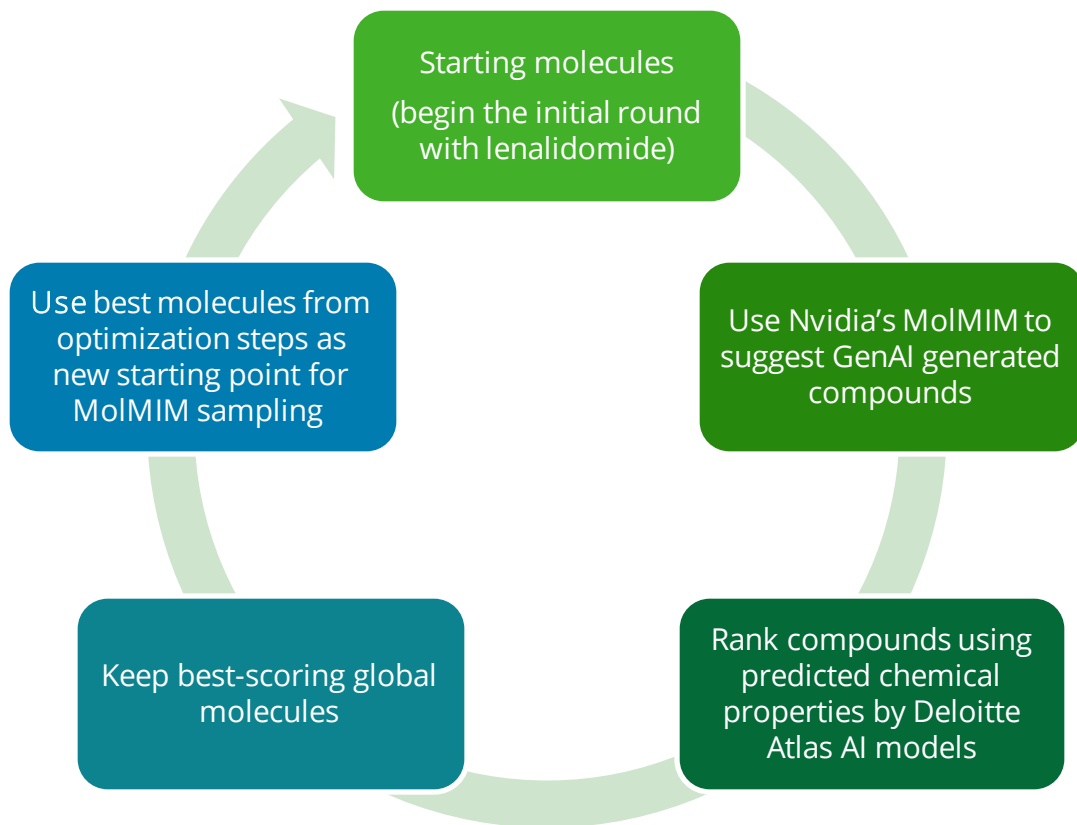
By Randy Leonard



<https://www.fiercepharma.com/special-report/top-15-blockbuster-patent-expirations-coming-decade>

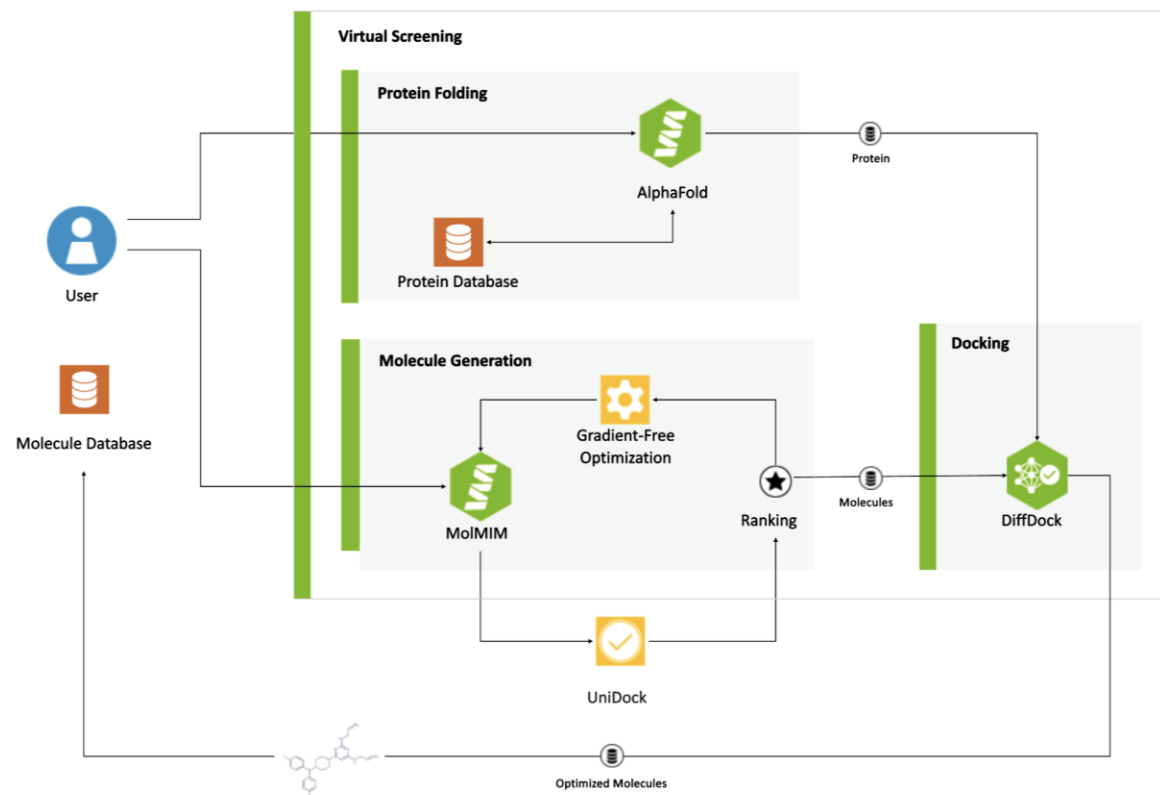
Gradient-Free Optimization Path

Universally applicable optimization cycle moves progressively toward molecules with improved properties without requiring any prior knowledge about the optimum.



Virtual Screening

uServices: MolMIM, AlphaFold-2, NeuralPLexer



Optimization Criteria

Selected scoring objectives and pharmaceutical relevance

ORACLE FUNCTION

- Specify an objective scoring function to determine optimized molecule(s) comprised of any predictable metrics specific to a given compound

CRITERIA CHOSEN:

1. **Binding strength (K_d):** Increased binding strength to the target of interest.
2. **Aqueous solubility and intestinal permeability:** Improving solubility and permeability increases bioavailability via oral delivery allowing for higher doses.
3. **Half-life:** Increasing the half-life of the drug compound *in vivo* will enable it to persist in the patient's bloodstream longer times.
4. **Molecular weight:** Jointly attempting to maximize molecular weight while simultaneously maximizing solubility/permeability for molecules with desirable physical properties.
5. **Drug-likeness:** Using QED score for drug-likeness and Deloitte's Pharmaceutical Compound Classifier to score for drug-likeness.
6. **hERG Cardiotoxicity:** Use publicly available hERG cardiotoxicity predictor models to identify and filter out any compounds likely to block hERG channels and cause cardiotoxicity.
7. **PAINS / Brenk et al. substructure filters:** Filtering out compounds containing substructures identified by either of these lists can avoid false positives in terms of drug-target binding and reduce the likelihood of negative clinical outcomes.

Optimization

ORACLE FUNCTION / OPTIMIZATION STRATEGY

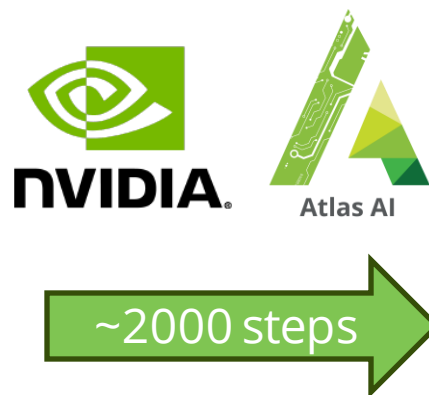
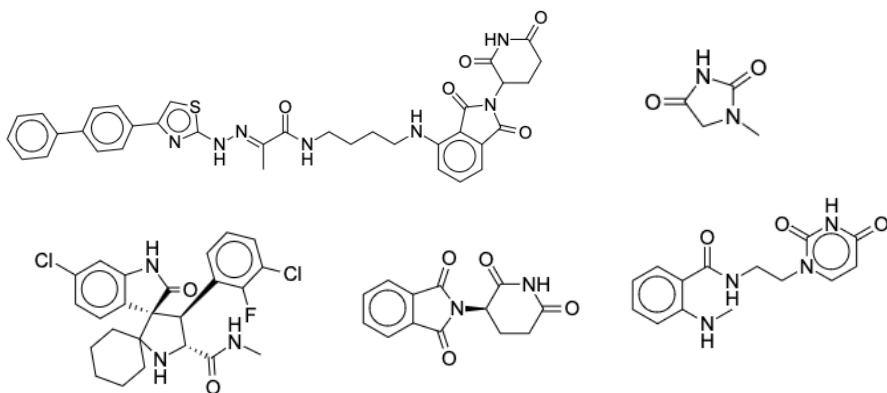
Starting compounds were identified using Atlas AI, returning all compounds tested to bind protein target cereblon (CRBN)

Score = (1.0 * **Binding**) + (-0.25 * **Log-solubility**) + (-0.25 * **Log-permeability**) + (-0.1 * **Log-half-life**) + (-0.75 * **MW**) + (-0.4 * **QED**) + (100 if **PAINS/Brenk**) + (5 x **hERG-blocker probability**) + (-2.0 * **Pharma-like compound probability**)

RESULTS

18 starting compounds

Best starting score: -0.862

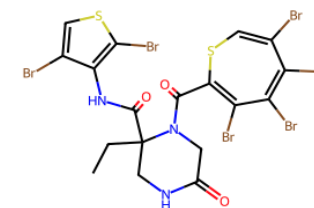


Optimized Examples

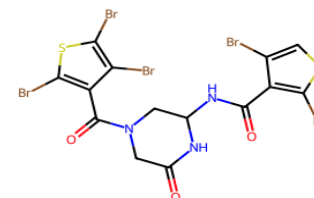
Score

Compound

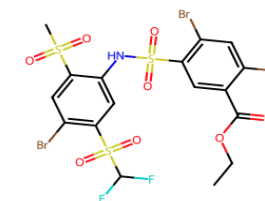
-2.293



-2.199

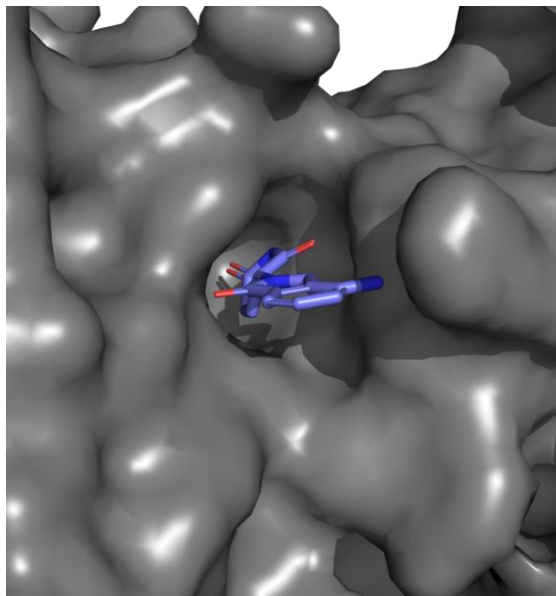


-2.027

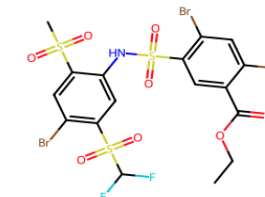
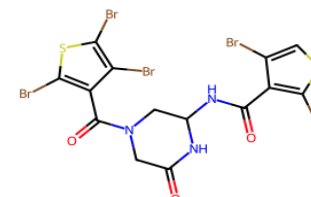
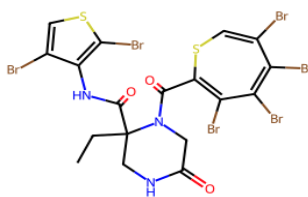
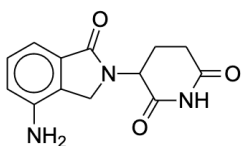
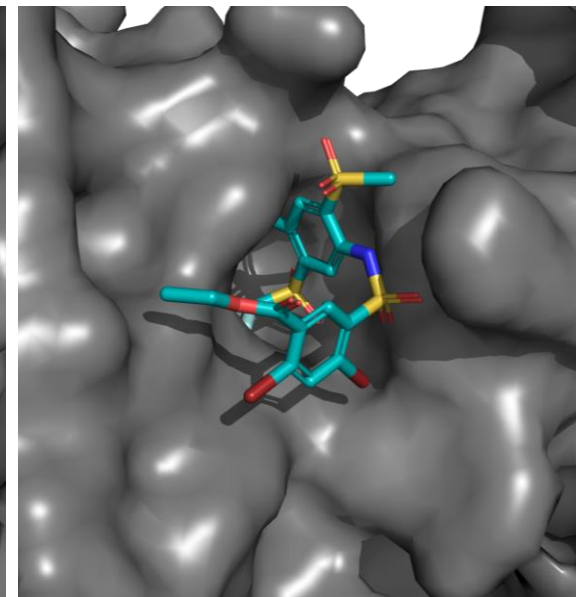
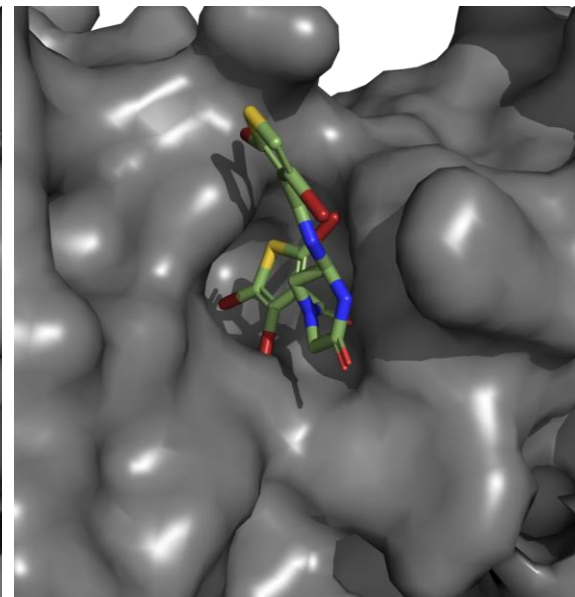
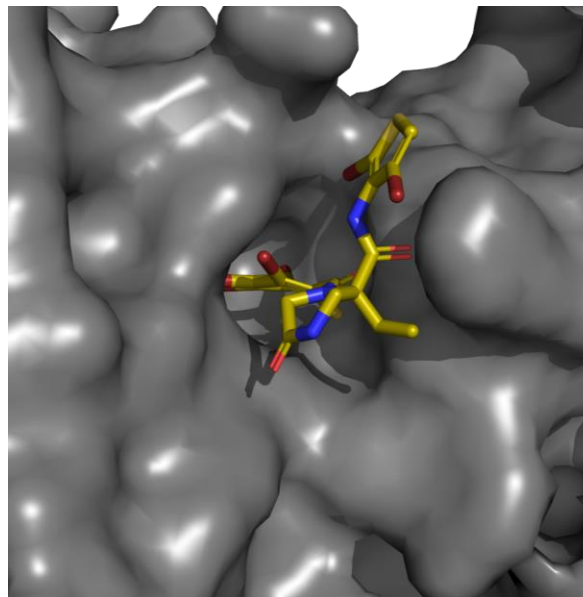


DiffDock Docking Simulations

Lenalidomide



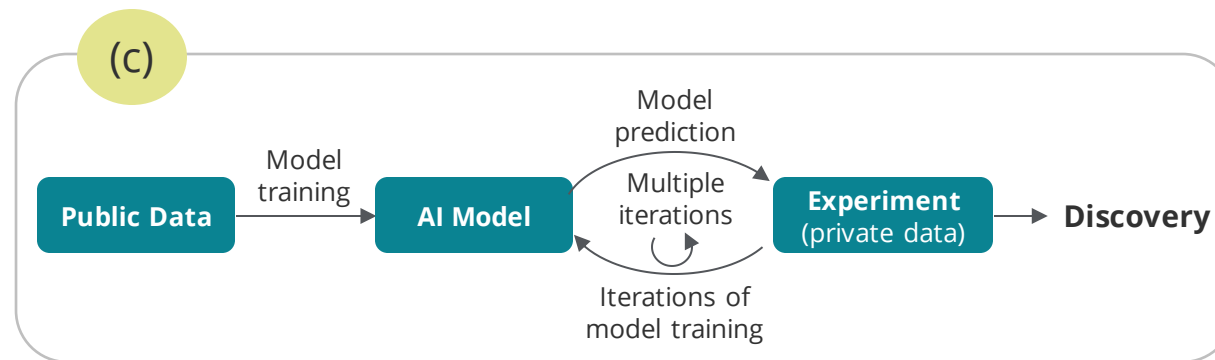
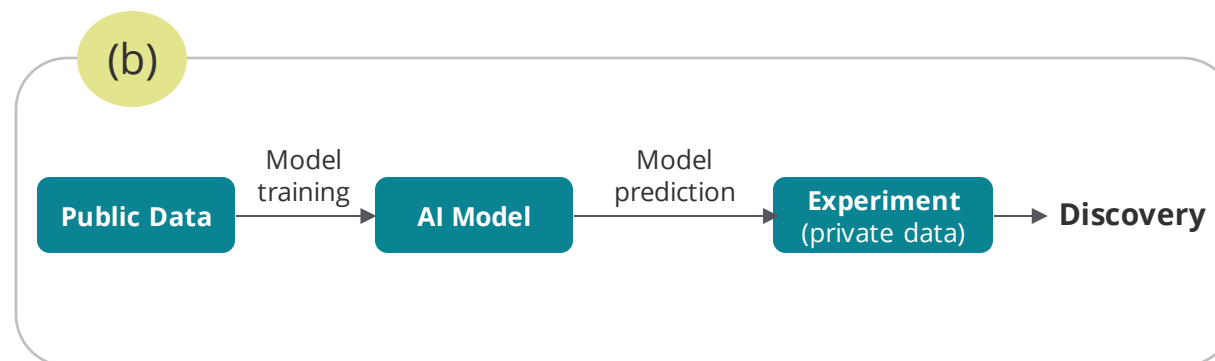
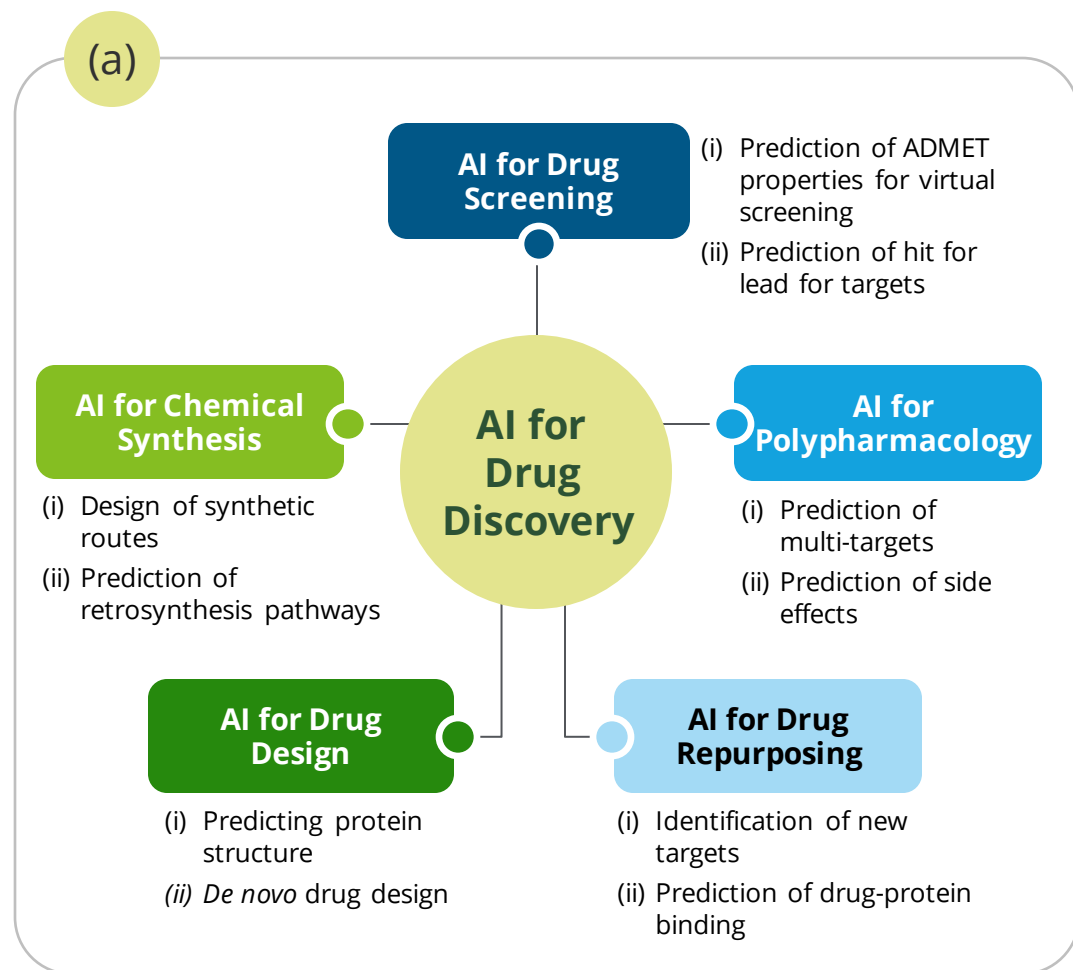
Optimized Examples



- All molecules are predicted to bind to the cereblon protein within the same binding pocket as lenalidomide, suggesting active-site inhibition
- High confidence predictions suggest beneficial geometric and electrostatic interactions between drug and target

Atlas AI | AI/ML Design Patterns

Next Decade's AI-Based Drug Development Features Tight Integration of Data and Computation



Atlas AI | Generative AI and Knowledge Graphs



- Atlas AI's Knowledge Graph can enhance GenAI/LLMs by providing external knowledge for inference and interpretability.
 - Atlas AI complementary unifies GenAI/LLMs and KGs together and simultaneously leverages their advantages.

Atlas AI consists of three general frameworks, namely,

1

KG-enhanced LLMs

which **incorporate KGs during the pre-training and inference phases of LLMs**, or for the purpose of enhancing understanding of the knowledge learned by LLMs.

2

LLM-augmented KGs

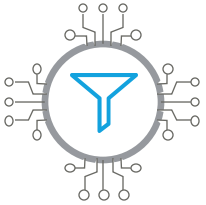
that **leverage LLMs for different KG tasks** such as embedding, completion, construction, graph-to-text generation, and question answering.

3

Synergized LLMs + KGs

in which **LLMs and KGs play equal roles** and work in a mutually beneficial way to enhance both LLMs and KGs for bidirectional reasoning driven by both data and knowledge.

Atlas AI | Semantically Enriched Data



- **Atlas AI addresses important shortcomings of current data science and machine learning solutions** by leveraging "semantic" understanding and reasoning on data in combination with novel tools for data science automation to help with consistent and explainable data augmentation and transformation.
- **Additionally, Atlas AI enables semantics to assist data scientists in a new manner** by helping with challenges related to trust, bias, and explainability in machine learning.
- **Semantic annotation** can also help better explore and organize large data sources.

Atlas AI | Enabling Precision Medicine



Atlas AI integrates **13+ high-quality resources** to describe **~12 million nodes** with **~97 million relationships** representing **over ten major biological scales**, including disease-associated protein perturbations, biological processes and pathways, anatomical and phenotypic scales, and the entire range of approved drugs with their therapeutic action, considerably expanding previous efforts in disease-rooted knowledge graphs.

The Vision of a Lab Powered by Generative AI

Connecting data to create scientific knowledge.

Discover



Hypothesis and idea generation

Our Knowledge graph enhances the hypothesis and reduces the number of variables required to set up an experiment

Experiment



Wet lab

With plenty of data connectivity and predictive information, only a handful of experiments are required.

Analyze



Automated analysis

Every piece of raw data generated in the lab is analyzed by specific workflows and insights are fed into the knowledge graph



Thank you.

As used in this document, 'Deloitte' means Deloitte Consulting LLP, which provides strategy, operations, technology, systems, outsourcing and human capital consulting services; and Deloitte & Touche LLP, which provides audit and risk advisory services. These entities are separate subsidiaries of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Copyright © 2023 Deloitte Development LLC. All rights reserved.

