RAYCA™

WHITE PAPER

AI-Powered Drug Discovery

Rayca Precision Drug Discovery Suite

De-risking drug targets, saves billions and accelerates timelines by decades

Powered by NVIDIA BioNeMo

Protein Libraries

Create extensive protein libraries at large scale.

Property Predictors

Refine protein libraries using embeddings for precise property predictions.

Molecule Generation

Craft small molecules with tailored properties.

3D Folding Structure Prediction

Visualize and predict the 3D structures of billions of proteins.

Pose Estimations

Run campaigns for ligand-to-small-molecule pose estimations.

By leveraging NVIDIA BioNeMo, our platform empowers scientists to design proteins, predict 3D folding structures and properties, and expedite drug candidate discovery and validation.

Overview

This integrated computational solution for drug discovery, hosts a diverse array of ten (10) LLMs accessible through Rayca Precision Drug Discovery Suite and its APIs/CLI interfaces. This platform simplifies workflows, allowing researchers to concentrate on adapting AI models without navigating complex infrastructure configurations. This suite marks a paradigm shift in drug discovery, slashing the traditional timelines by adopting the power of advanced AI models.

Integrative Approach

Traditional drug development timelines often exceed a decade. Rayca Precision's Drug Discovery Suite seeks to revolutionize this process using cutting-edge large language models (LLMs) capable of understanding the intricacies of biological and chemical text. By leveraging NVIDIA DGX Cloud, our platform empowers scientists to design proteins, predict 3D structures, and expedite drug candidate discovery.

Generative AI Models

Protein Structure Prediction

AlphaFold-2 | ESMFold | OpenFold

Protein Property Predictions

ESM-1 | ESM-2 | ESM-1nv

Protein Generation

ProtGPT2

Small Molecule Generation

MegaMolBART | MoFlow

Molecular Docking

DiffDock

Protein Structure Prediction

Models Architecture Overview

AlphaFold-2	*Deep learning with JAX workflow*
OpenFold	*Faithful reproduction based on PyTorch*
ESMFold	*Transformer-based, ultrafast model*

Gen-AI Engines:

AlphaFold-2 | ESMFold | OpenFold

The challenge lies in the vast number of possible conformations a protein can adopt, influenced by its amino acid sequence. Experimental methods for structure determination, such as X-ray crystallography and cryo-electron microscopy, are resource-intensive and time-consuming. Computational methods driven by generative AI models offer a promising avenue to accelerate this process.

Predicting the 3D structure of proteins from their primary amino acid sequences is a complex and critical task in drug discovery. The folds and interactions of proteins play a pivotal role in their functions.

AlphaFold-2: Breaking Ground

Achieved near-experimental accuracy for predicted protein 3D structures. Utilizes deep learning to predict the relationship between amino acid sequences and 3D structures. Significant milestone achieved at CASP14.

Model Description

Model Type	*DeepMind's AlphaFold 2 utilizes deep learning with a JAX workflow.*
Training Data	*Diverse protein sequence datasets including CASP14 dataset.*
Functionality	▪ *Predict the 3D structures of proteins with remarkable accuracy.* ▪ *Achieved near-experimental accuracy during the CASP14 competition, showcasing its proficiency in transforming amino acid sequences into detailed 3D models.* ▪ *Eliminating the need for manual intervention or supplementary information.*

Powered by NVIDIA. BioNeMo

OpenFold: PyTorch-Based Reproduction

Replicates AlphaFold 2's accuracy in predicting 3D protein structures. Significantly accelerated by 6x in NVIDIA's BioNeMo. Trainable, allowing variants for specialized research.

Model Description

Model Type	*OpenFold is a faithful reproduction based on PyTorch, inspired by AlphaFold 2.*
Training Data	*Trained on millions of protein sequences*
Functionality	▪ *A faithful reproduction of AlphaFold 2, leveraging the PyTorch framework for flexibility and ease of adaptation.* ▪ *Accelerates protein structure prediction by 6x.* ▪ *Reduces analysis times, allowing to analyze larger datasets, iterate more rapidly.*

ESMFold: Ultrafast 3D Structure Prediction

Achieves ultrafast 3D protein structure prediction. Utilizes embeddings without requiring many homologous sequences. Single NVIDIA GPU prediction time significantly faster than AlphaFold-2

Model Description

Model Type	*Meta's ESMFold is a transformer-based, ultrafast model.*
Training Data	*Trained on millions of protein sequences without multiple sequence alignment (MSA).*
Functionality	▪ *Making predictions for a protein with 384 residues in just 14.2 seconds on a single NVIDIA GPU.* ▪ *Enabling end-to-end, single-sequence structure prediction without requiring multiple sequence alignments (MSA).*

Powered by NVIDIA. BioNeMo

Protein Property Predictions: Bridging the Knowledge Gap

Protein property predictions aim to unveil characteristics such as subcellular location, thermostability, and water solubility. The challenge lies in decoding the intricate relationships between amino acids that give rise to these properties.

Gen-AI Engines:

ESM-1 | ESM-2 | ESM-1nv

Understanding the properties of proteins is crucial for deciphering their roles in cellular processes and identifying potential drug targets.

ESM-1 and ESM-2: Inspired by Evolution

Drawing inspiration from the BERT architecture with the ability to predict evolutionary-scale protein properties and support diverse downstream tasks positions them as valuable assets in deciphering the complexities of protein behavior in the drug discovery landscape.

Model Description

Model Type	*ESM-1 and ESM-2 are large language models inspired by the BERT architecture.*
Training Data	*Millions of protein sequences including UniRef database*
Functionality	▪ *Predict evolutionary-scale protein properties.* ▪ *Leverage masked language modeling to capture amino acid patterns and dependencies.* ▪ *Support downstream tasks like 3D structure prediction and variant effect prediction.*

Powered by NVIDIA. BioNeMo

ESM-1nv: Re-trainable Evolutionary Modeling

ESM-1nv, crafted by Meta AI, is a potent Large Language Model (LLM) built on the BERT architecture. Trained on vast protein datasets, it employs Masked Language Modeling (MLM) to unravel amino acid sequence intricacies, offering insights into protein structure. Its uniqueness lies in adaptability—re-trainable, extensible, and fine-tunable on powerful infrastructures like NVIDIA DGX Cloud.

Model Description

Model Type	*ESM-1nv is a faithful reproduction of Meta's ESM-1b, based on the BERT architecture.*
Training Data	*Trained on millions of protein sequences from UniProt database*
Functionality	▪ *Optimized for re-training on large compute infrastructures like DGX Cloud.* ▪ *Predicts multiple protein properties from amino acid sequences.* ▪ *Offers adaptability and extensibility for specialized research.*

ESM-1, ESM-2, and ESM-1nv, utilize advanced architecture and diverse training data to provide accurate predictions for crucial protein properties. Bridging the gap between amino acid sequences and functional characteristics, these models enhance the understanding of protein behavior in drug discovery processes.

In Rayca Precision's Drug Discovery Suite, ESM-1nv swiftly predicts varied protein properties, from subcellular location to thermostability, leveraging its optimized embeddings in BioNeMo Service for seamless integration into downstream tasks, accelerating drug discovery workflows.

Powered by NVIDIA BioNeMo

Protein Generation

AI Models ProtGPT2 MoFlow MegaMolBART

ProtGPT2, developed at ISMB and the Universität of Bayreuth, Germany, stands as a significant player in de novo protein sequence generation. This Large Language Model (LLM) is based on the GPT2 transformer architecture, featuring 36 layers with 738M parameters. Trained using a causal modeling objective, ProtGPT2 predicts the next token (oligomer) in the sequence, effectively learning an internal representation of proteins and the language they follow.

ProtGPT2: De Novo Protein Sequences

ProtGPT2, with 738 million parameters, excels in de novo protein sequence generation. Trained on UniRef50, it identifies unique structures, properties, and functions, valuable for exploring novel protein sequences in drug discovery.

Model Description

Model Type	*GPT2 transformer-based architecture, featuring 36 layers with 738M parameters.*
Training Data	*UniRef50 protein space database*
Functionality	▪ *De novo protein sequence generation* ▪ *Identifying unique protein structures, properties, and functions* ▪ *Effective in scenarios with limited training data* ▪ *Valuable for exploring uncharted territories in protein sequence design*

The integration of ProtGPT2 and ESM-1nv within the Drug Discovery Suite represents a leap forward in addressing the challenge of creating de novo protein sequences with sophistication and efficiency.

Powered by NVIDIA. BioNeMo

MoFlow: Transformative Molecule Generation

As a flow-based generative model, MoFlow excels in its ability to craft molecular graphs with precision, leveraging its unique approach to invertible mappings and producing molecular structures that align with real-world chemical principles.

Model Description

Model Type	*An Invertible Flow Model for Generating Molecular Graphs*
Training Data	*UniRef50 protein space database*
Functionality	▪ *Learns invertible mappings between molecular graphs and latent representations* ▪ *Incorporates graph convolutions for capturing structural nuances* ▪ *Achieves high accuracy and efficiency in molecule generation.* ▪ *By conforming to bond-valence constraints, the model produces molecular structures that align with real-world chemical principles.*

MegaMolBART: Transformer-Based Generative Chemistry

It excels in molecular design for drug discovery. A collaborative effort by AstraZeneca and NVIDIA, uses SMILES notation, ensuring compatibility with diverse databases. Pretrained on ZINC-15 with 1.45 billion molecules

Model Description

Model Type	*Seq2Seq Transformer BART*
Training Data	*Billions of SMILES from the ZINC15 database*
Functionality	▪ *employs SMILES notation for chemical structure representation, ensuring compatibility with various databases.* ▪ *optimized for molecular design tasks* ▪ *Focused on generating molecules for tasks such as molecular optimization, ensuring practical utility in drug discovery workflows.*

Molecular Docking

The AI Model DiffDock

Challenges in Molecular Docking

Molecular docking, a crucial step in drug discovery, involves predicting the binding structure of a small molecule ligand to a protein. Accurate predictions facilitate the design of targeted drugs. However, this process presents challenges, including the need for fast inference times, confidence estimates, and high selective accuracy.

DiffDock: A Diffusion Generative AI Model

DiffDock, developed by MIT's Jameel Clinic, is a state-of-the-art diffusion generative AI model tailored for molecular docking or pose prediction. It outperforms previous methods, achieving a remarkable 38% top-1 prediction with RMSD<2A on the PDBBind blind docking benchmark. This surpasses both search-based methods like SMINA and GLIDE, as well as recent deep learning methods like EquiBind and TANKBind.

Model Description

Model Type	*A diffusion generative AI model designed for molecular docking predictions*
Training Data	*DiffDock was trained on the complexes from PDBBind, which is a large collection of protein-ligand structures collected from PDB*
Functionality	▪ *Predicts binding structure of small molecules ligands to proteins* ▪ *DiffDock provides confidence estimates with high selective accuracy* ▪ *Swift inference times, making it practical for handling large datasets and supporting iterative drug design*

DiffDock's fast inference times, confidence estimates, and benchmark performance position it as a pivotal tool for large datasets and supporting iterative drug design.

Integrating Rayca Precision Bioinformatics Suite with Drug Discovery Platform

Unifying Data Insights: A Holistic Approach

- Identification of Alternative Splicing Events
- Mass Spectrometry Data Integration
- Functional Implications of Alternative Splicing
- Detection of Fusion Transcripts

Identification of Alternative Splicing Events

Utilization of RNAseq data for the identification of alternative splicing events, revealing the diversity of transcript isoforms arising from a single gene.

- Prioritization of alternative splicing events for downstream functional and structural analyses.
- Tailoring therapeutics based on the specific isoforms expressed, considering functional differences.
- Integration with structure prediction models for predicting 3D structures of proteins encoded by alternatively spliced transcripts.

Functional Implications of Alternative Splicing

Exploitation of RNAseq data to understand the functional consequences of alternative splicing, recognizing alterations that lead to the inclusion or exclusion of specific exons.

- Identification of proteins with altered functional domains for prioritization as drug targets.
- Integration with structural prediction models for understanding the impact on protein structure.
- Customization of therapeutics targeting specific functional isoforms.

Powered by NVIDIA. BioNeMo

Detection of Fusion Transcripts

Leveraging RNAseq data to detect fusion transcripts arising from genetic rearrangements, providing insights into potential oncogenic drivers and therapeutic targets.

- Prioritization of fusion transcripts for further investigation
- Integration with structural prediction models for 3D structure analysis
- Exploration of fusion-driven pathways for drug development

Mass Spectrometry Data Integration

Integration of mass spectrometry data into the Drug Discovery Suite, offering detailed insights into the proteome, including protein abundance, post-translational modifications, and variations.

- Prioritization of druggable proteins based on abundance and modifications
- Validation of predicted protein structures through experimental evidence
- Dynamic profiling for personalized medicine by understanding changes over time

From alternative splicing revelations to fusion transcript discoveries and precise proteomic profiling, our holistic approach accelerates precision medicine, unraveling the complexities of disease for innovative therapeutic breakthroughs

Powered by NVIDIA. BioNeMo