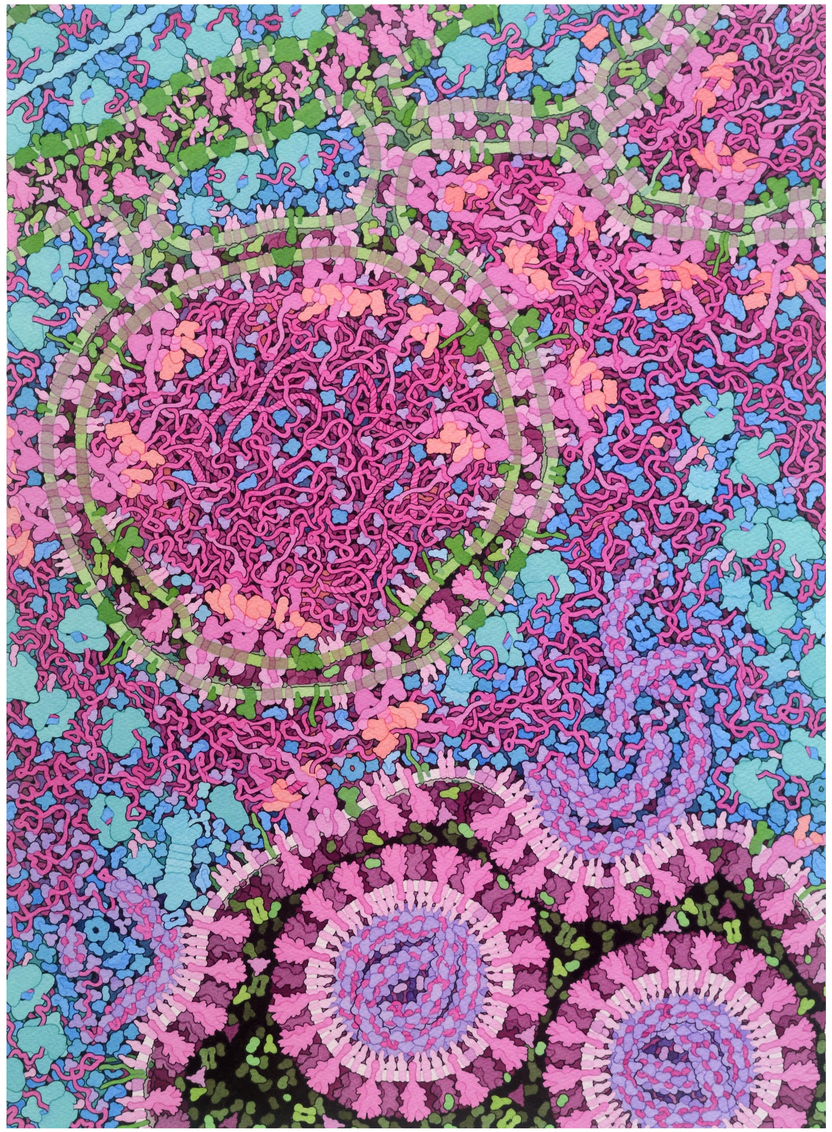
Large Language Models and Generative AI for Life Science

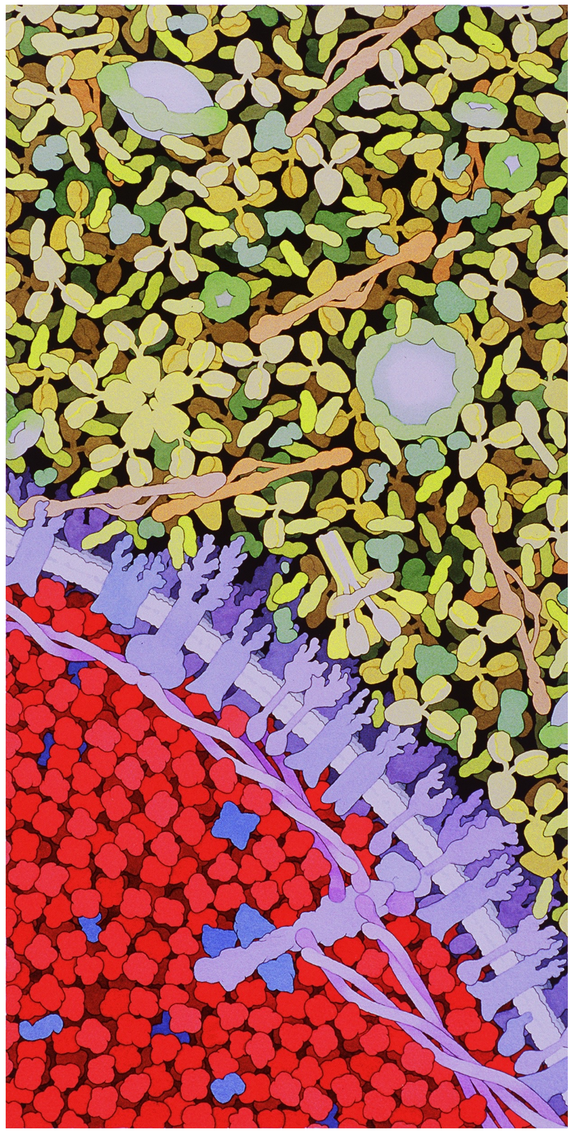
Christian Dallago, Senior Solution Architect for Healthcare and Life Sciences

Molecular Landscapes
David S. Goodsell

HIV Vaccine, 2022
David S. Goodsell, RCSB Protein Data Bank and Scripps Research. doi: 10.2210/rcsb_pdb/goodsell-gallery-043



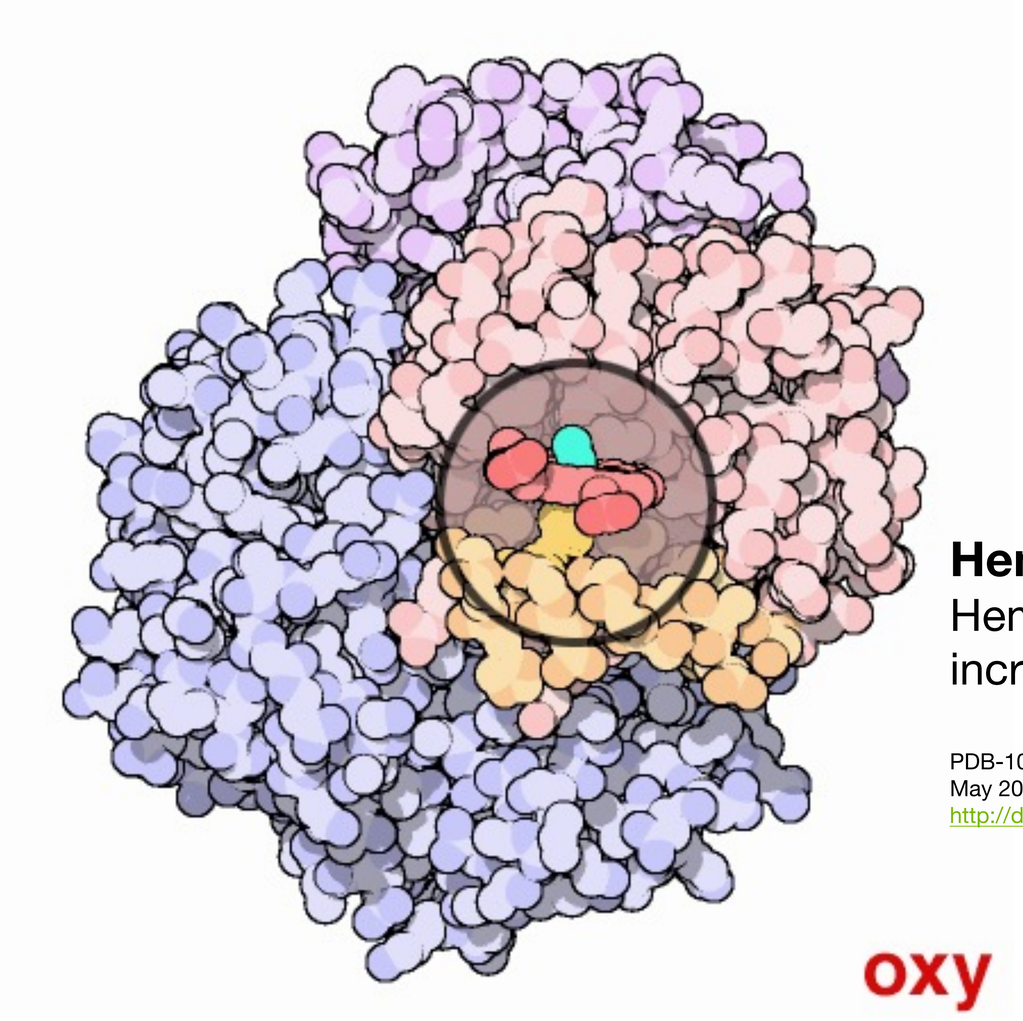
Coronavirus Life Cycle, 2020
David S. Goodsell, RCSB Protein Data Bank; doi: 10.2210/rcsb_pdb/goodsell-gallery-023



Blood, 2000
Illustration by David S. Goodsell, The Scripps Research Institute. doi: 10.2210/rcsb_pdb/goodsell-gallery-003

Proteins perform most functions necessary for life

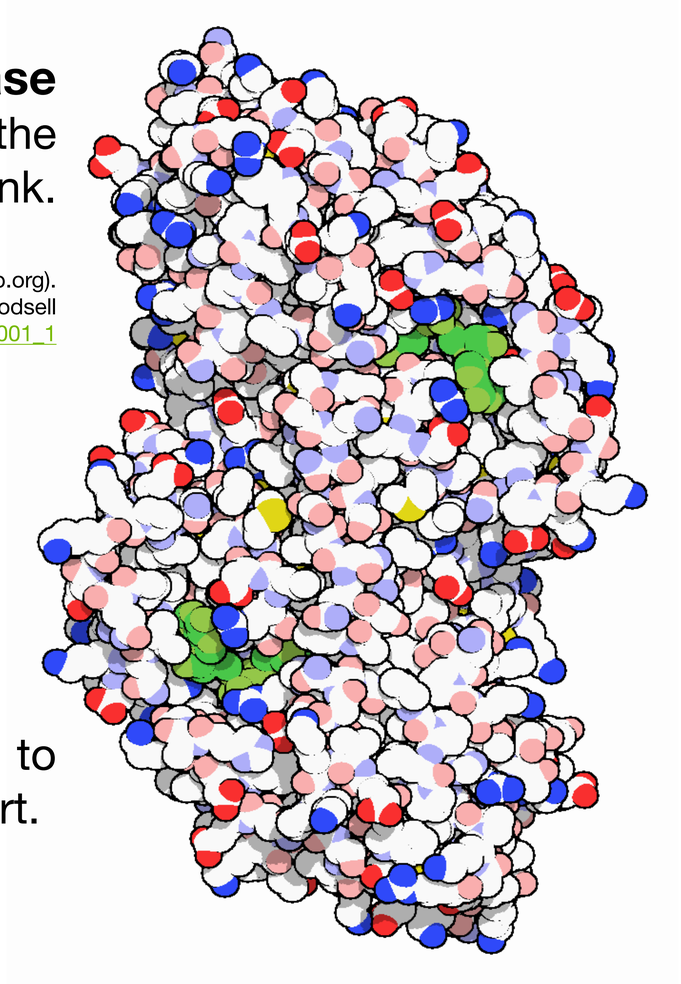
Their 3D shapes and the changes in shape determine their functions



Alcohol Dehydrogenase

Alcohol dehydrogenase detoxifies the ethanol we drink.

PDB-101 (PDB101.rcsb.org).
January 2001, David Goodsell
http://doi.org/10.2210/rcsb_pdb/mom_2001_1



Hemoglobin

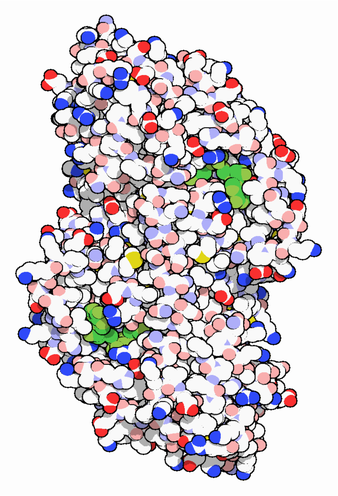
Hemoglobin uses a change in shape to increase the efficiency of oxygen transport.

PDB-101 (PDB101.rcsb.org).
May 2003, Shuchismita Dutta, David Goodsell
http://doi.org/10.2210/rcsb_pdb/mom_2003_5

In biology (and chemistry), a string of letters can identify a 3D object

Protein sequences are uniquely linked to their structures

`MALLHSARVLSG...`



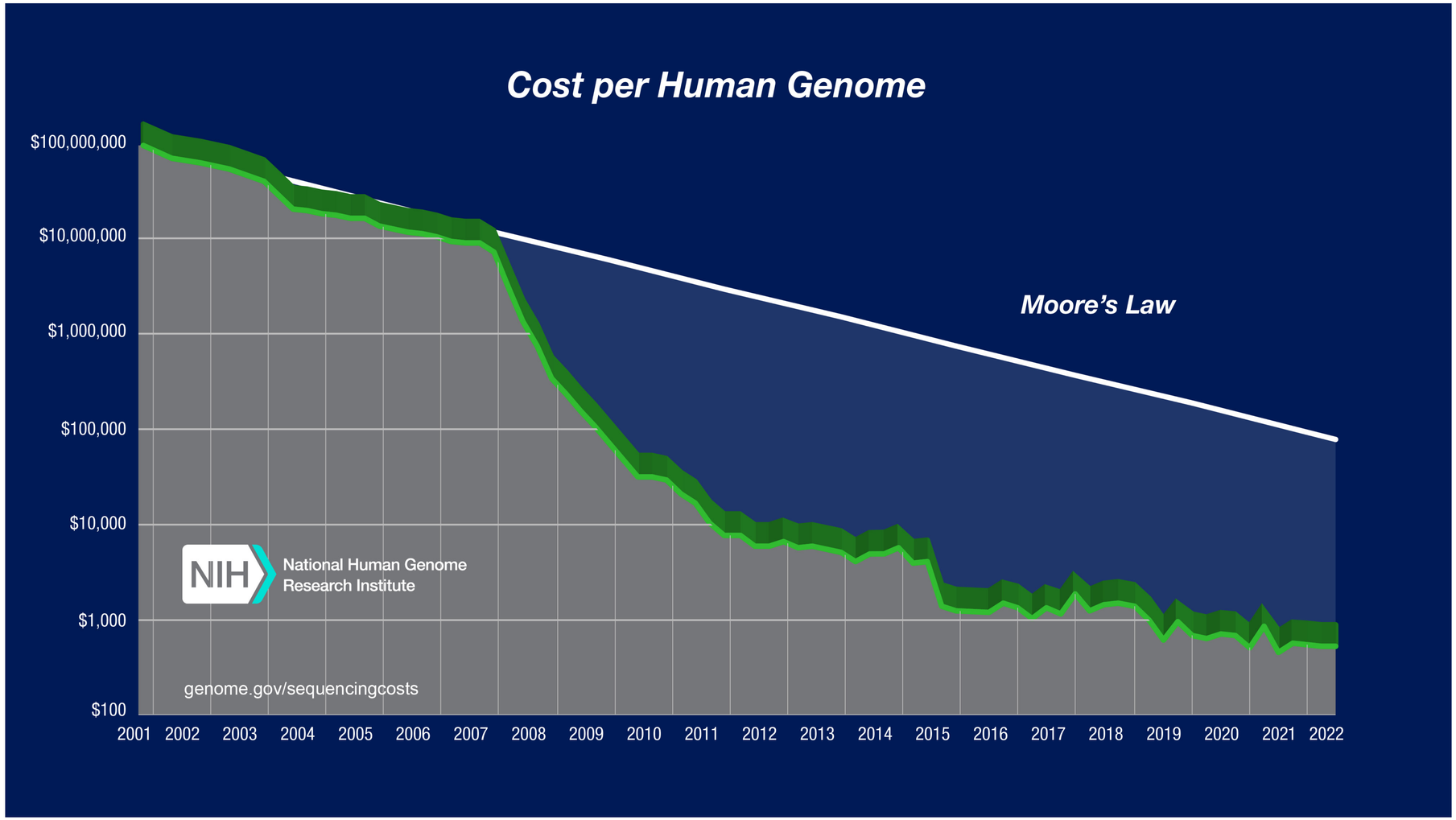
PDB-101 (PDB101.rcsb.org).
January 2001, David Goodsell
http://doi.org/10.2210/rcsb_pdb/mom_2001_1

`Red apple`

CC0 by PixBay
https://pixabay.com/de/illustrations/apple-apfelbaum-fr%C3%BCchte-obst-essen-1986660/

Sequencing has become chaper and fast

Vast amounts of primary biological data are recorded daily

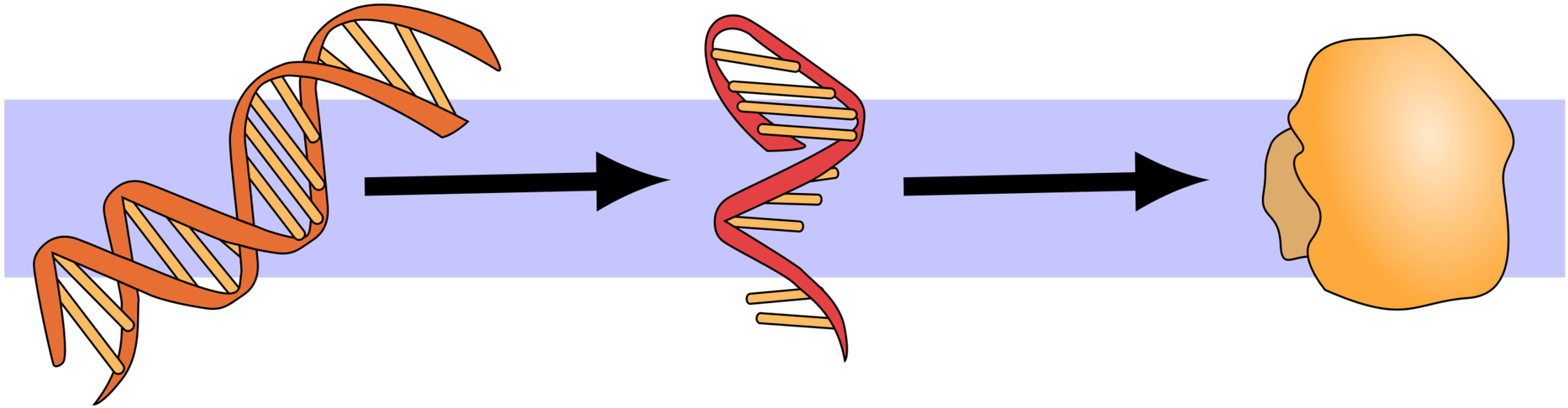


Cost per genome 2022, NHGRI
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

The central dogma of molecular biology meets technology

From easy to sequence to molecular workhorses

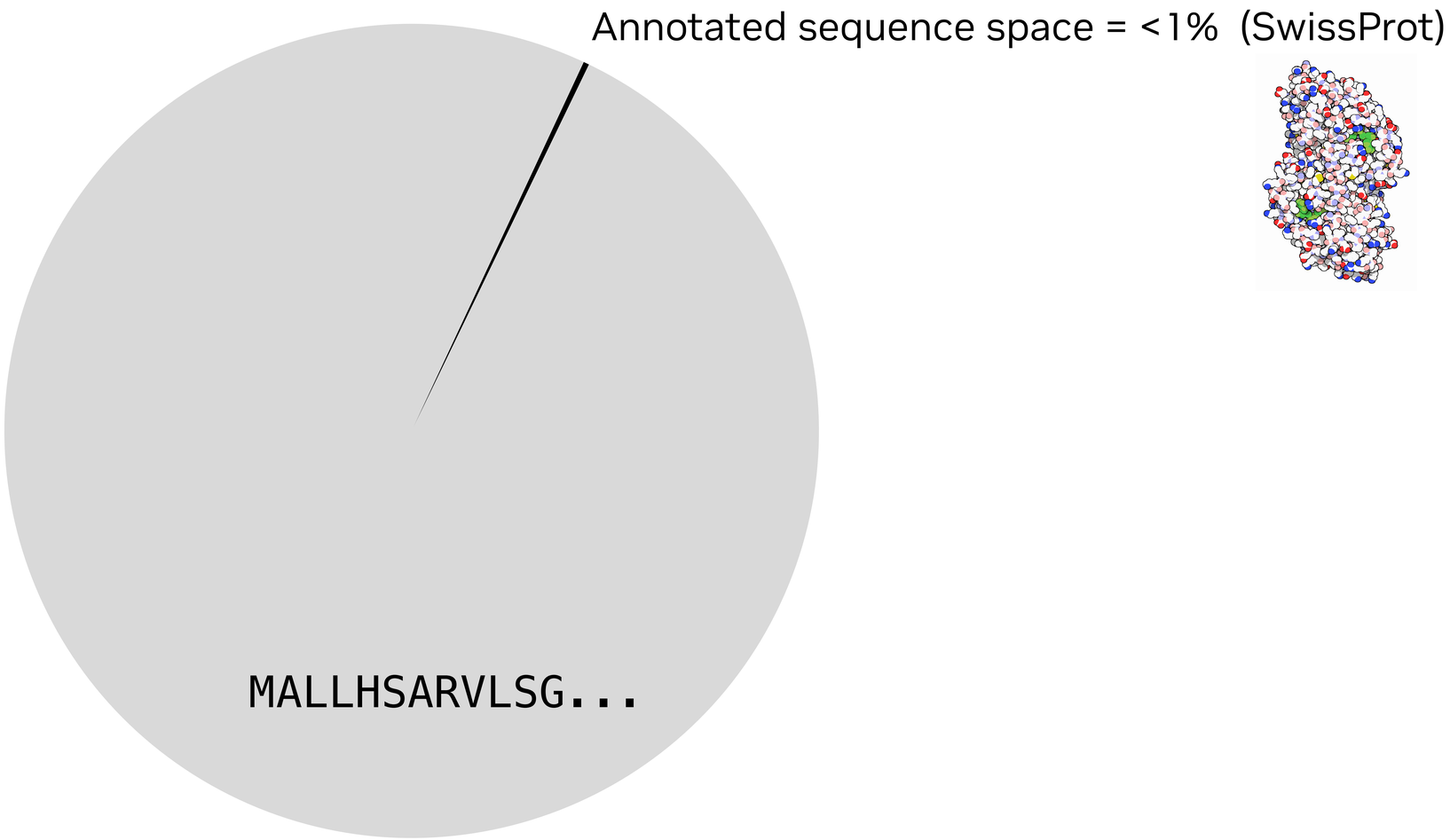
DNA	RNA	Protein



We can sequence en masse	We can recruit cells to express	Workhorse of the cell

There’s a gap between what we can sequence and what we know

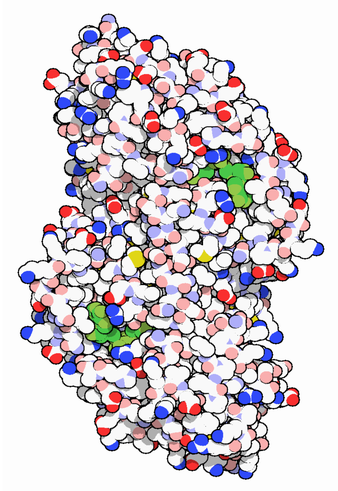
We have vast libraries of genomes, but we don’t yet know how to interpret them



Predictions drove 30 years of scientific innovation

The search for computational oracles to tell us what sequences *mean*

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG...



PDB-101 (PDB101.rcsb.org).
January 2001, David Goodsell
http://doi.org/10.2210/rcsb_pdb/mom_2001_1

How proteins were predicted until 2019

Extract features from alignment and use ML

1 `> MALLHSAR` **Start from a sequence of interest**

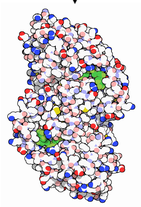
2 **Align to a database of protein sequences**

```
>  MALLHSAR
P2 MA---SAR
P3 MARNDSYP
.  ........
.  ........
.  ........
```

3 **Feature extraction**

- Position Specific Scoring Matrix (PSSM)
- Direct Couplings Analys (DCA)

4 AI Model



Find out about the history here (credit: Sergey Ovchinnikov): https://www.solab.org/resources/coevolution-structure-prediction

How proteins were predicted until 2019

Extract features from alignment and use ML

Can we learn this with AI?

1 `> MALLHSAR` **Start from a sequence of interest**

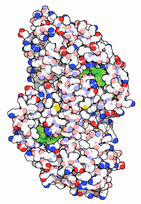
2 **Align to a database of protein sequences**

```
> MALLHSAR
P2 MA---SAR
P3 MARNDSYP
.  ........
.  ........
.  ........
```

3 **Feature extraction**

- Position Specific Scoring Matrix (PSSM)
- Direct Couplings Analys (DCA)

4 AI Model



Find out about the history here (credit: Sergey Ovchinnikov): https://www.solab.org/resources/coevolution-structure-prediction

Large language models introduce a new way to learn text

Masked language modelling, powered by large compute and larg models delivers excellent performance

The more you buy, the more you save!

ProtT5 is a transformer LLM trained on millions of protein sequences

Objective: Given a corrupted sequence, predict the residues (letters) that were corrupted

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEA

Modeling aspects of the language of life through transfer-learning protein sequences.
Heinzinger M., Elnaggar A., Wang Y., Dallago C., Nechaev D., Matthes F., & Rost B.

ProtT5 is a transformer LLM trained on millions of protein sequences

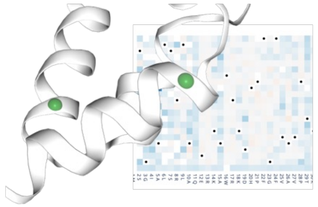
Result: ProtT5 has an internal representation of a protein sequence

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEA

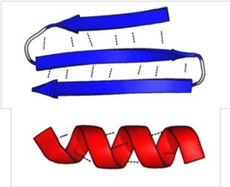
Modeling aspects of the language of life through transfer-learning protein sequences.
Heinzinger M., Elnaggar A., Wang Y., Dallago C., Nechaev D., Matthes F., & Rost B.

ProtT5 Protein LLM widely applicable

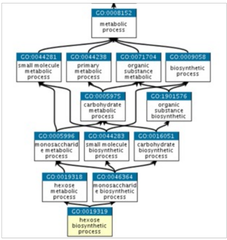
Protein Function ◀····· ·····▶ Protein Structure



Embeddings from protein language models predict conservation and variant effects
Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Michael Bernhofer, Kyra Erckert, Burkhard Rost
2021 —> Human genetics

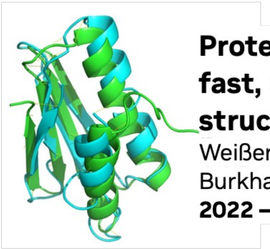


ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, Burkhard Rost **2019 —> IEEE Trans. on Patt. An. And Mach. Int.**

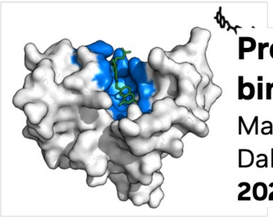


Embeddings from deep learning transfer GO annotations beyond homology.
Littmann, Maria, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost.
2021 —> In Review

TMbed: transmembrane proteins predicted through language model embeddings.
Bernhofer, Michael, and Burkhard Rost
2022 —> BMC bioinformatics

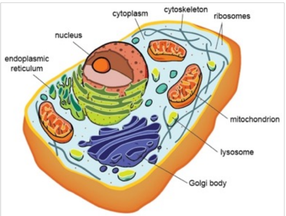


Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction.
Weißenow, Konstantin, Michael Heinzinger, and Burkhard Rost.
2022 —> Structure



Protein embeddings and deep learning predict binding residues for various ligand classes
Maria Littmann, Michael Heinzinger, Christian Dallago, Konstantin Weissenow & Burkhard Rost
2021 —> Scientific reports

Contrastive learning on protein embeddings enlightens midnight zone.
Heinzinger, Michael, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost.
2022 —>NAR genomics and bioinformatics



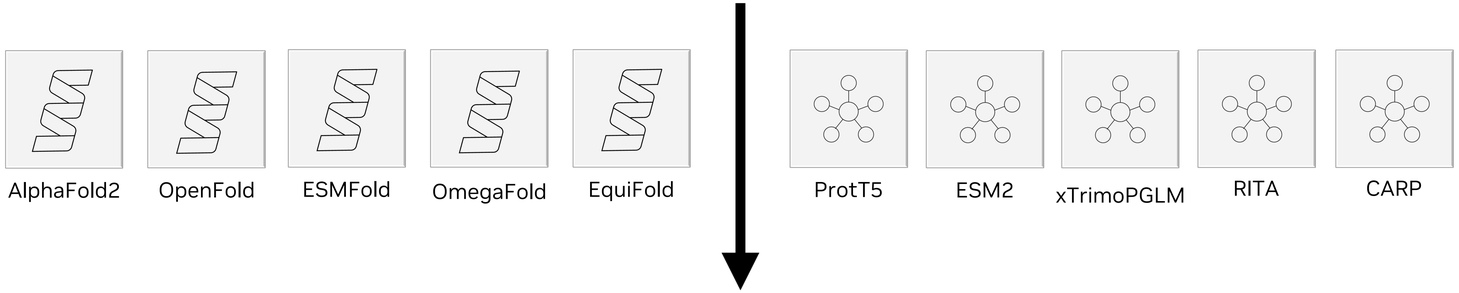
Light Attention Predicts Protein Location from the Language of Life
Hannes Stärk, Christian Dallago, Michael Heinzinger, Burkhard Rost
2021 —> Bioinformatics Advances

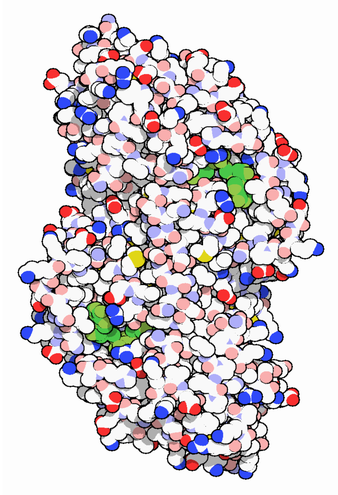
SETH predicts nuances of residue disorder from protein embeddings.
Ilzhoefer, Dagmar, Michael Heinzinger, and Burkhard Rost.
2022 —> In Review

Predictions drove 30 years of scientific innovation

The search for computational oracles to tell us what sequences *mean*

MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKMNLGVG...





Find a collection of 3D folding tools & protein language models at https://github.com/biolists/folding_tools

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

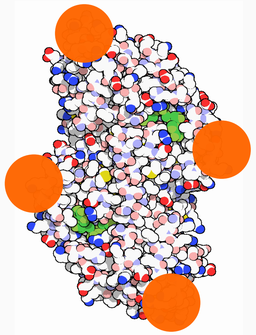
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

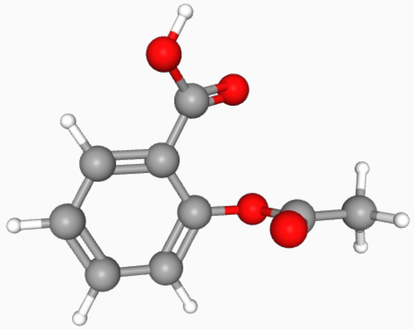
+

CC(=O)OC1=CC=CC=C1C(
=O)O

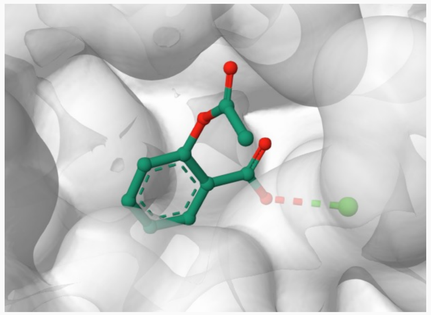
Protein structures



Chemical structures



Bound structures



The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

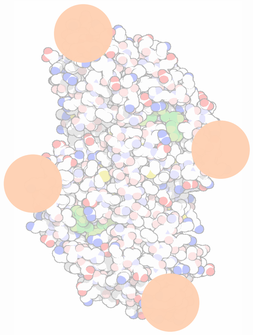
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

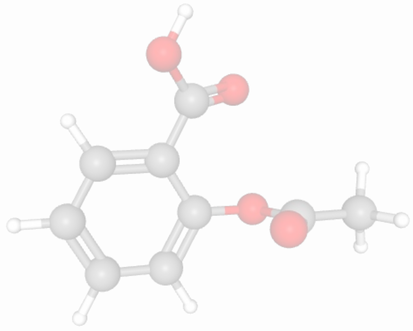
+

CC(=O)OC1=CC=CC=C1C(
=O)O

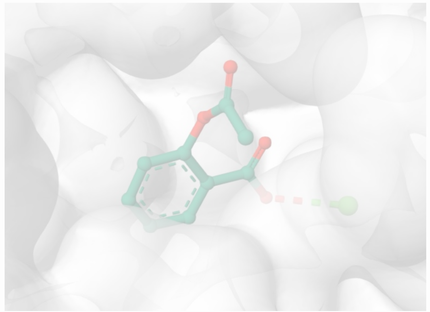
Protein structures



Chemical structures



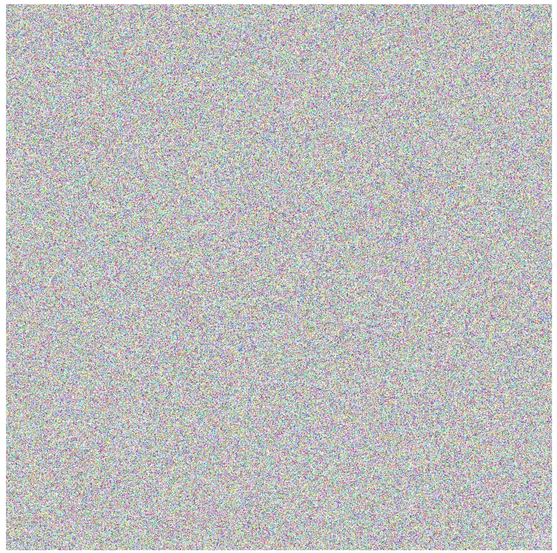
Bound structures



Diffusion models are making waves

These models train to create order from noise

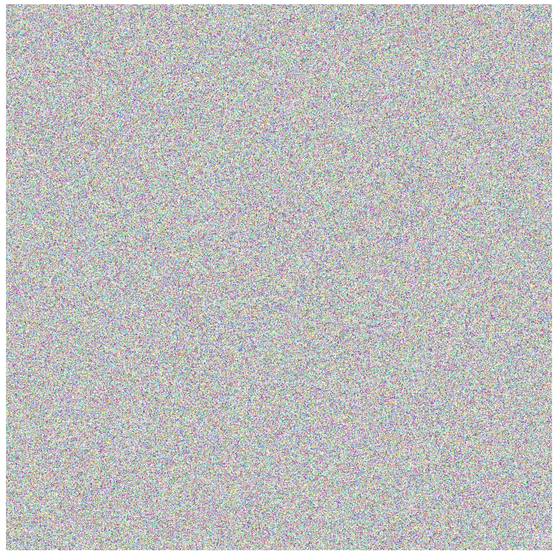
Noise



Diffusion models are making waves

These models train to create order from noise

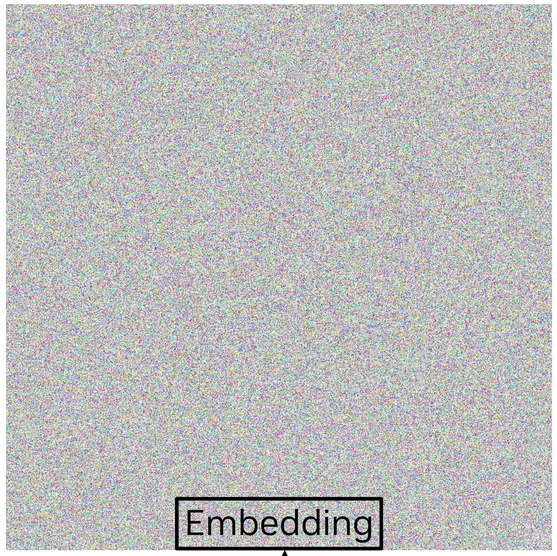
Noise → **Order**



Diffusion models are making waves

These models train to create order from noise

Noise

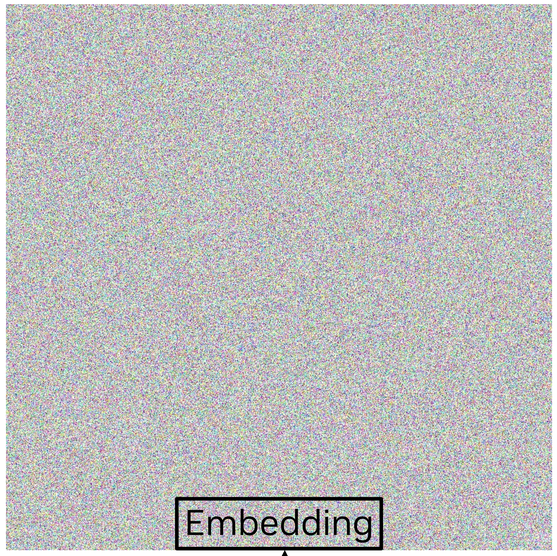


“A wave in a tumultuous sea”

Diffusion models are making waves

These models train to create order from noise

Noise → **Order**



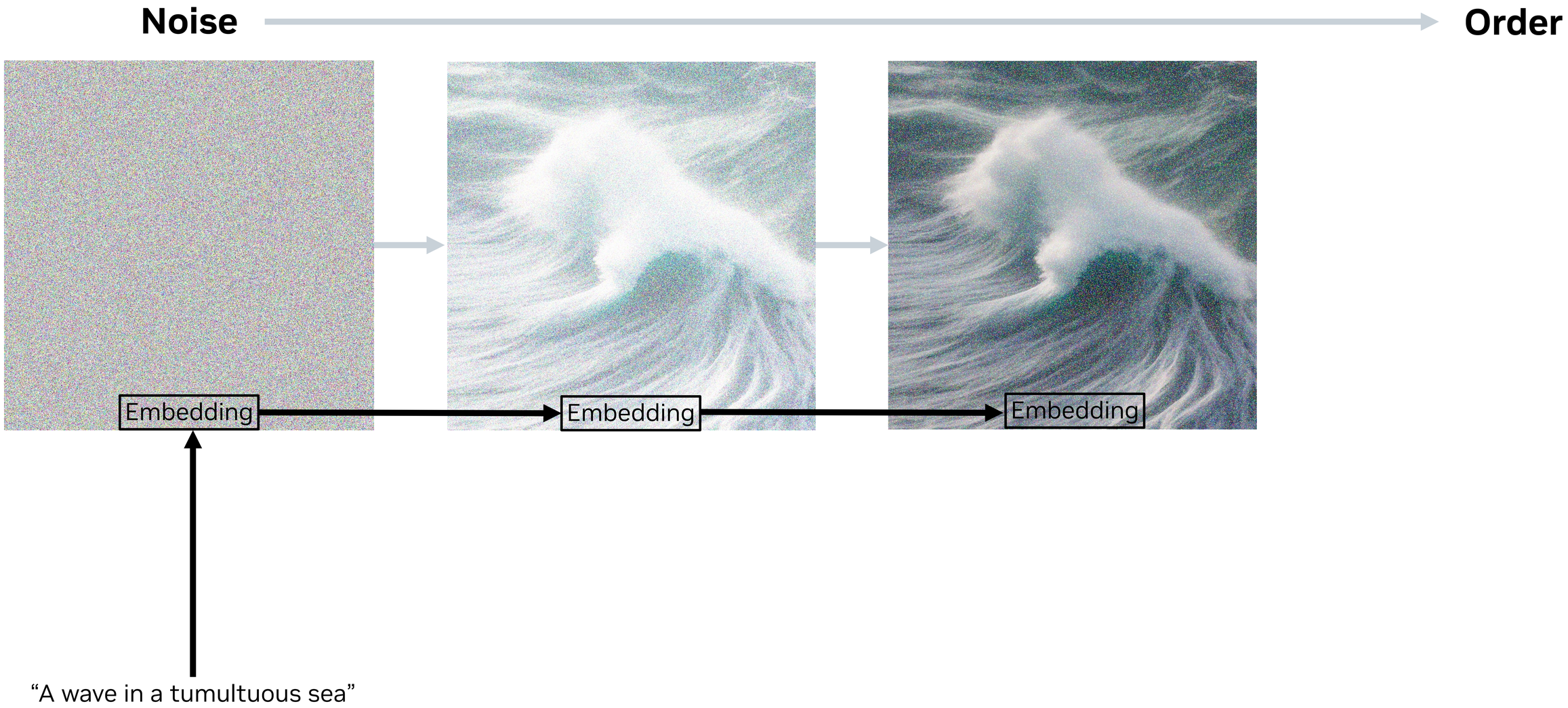
"A wave in a tumultuous sea"

Diffusion models are making waves

These models train to create order from noise

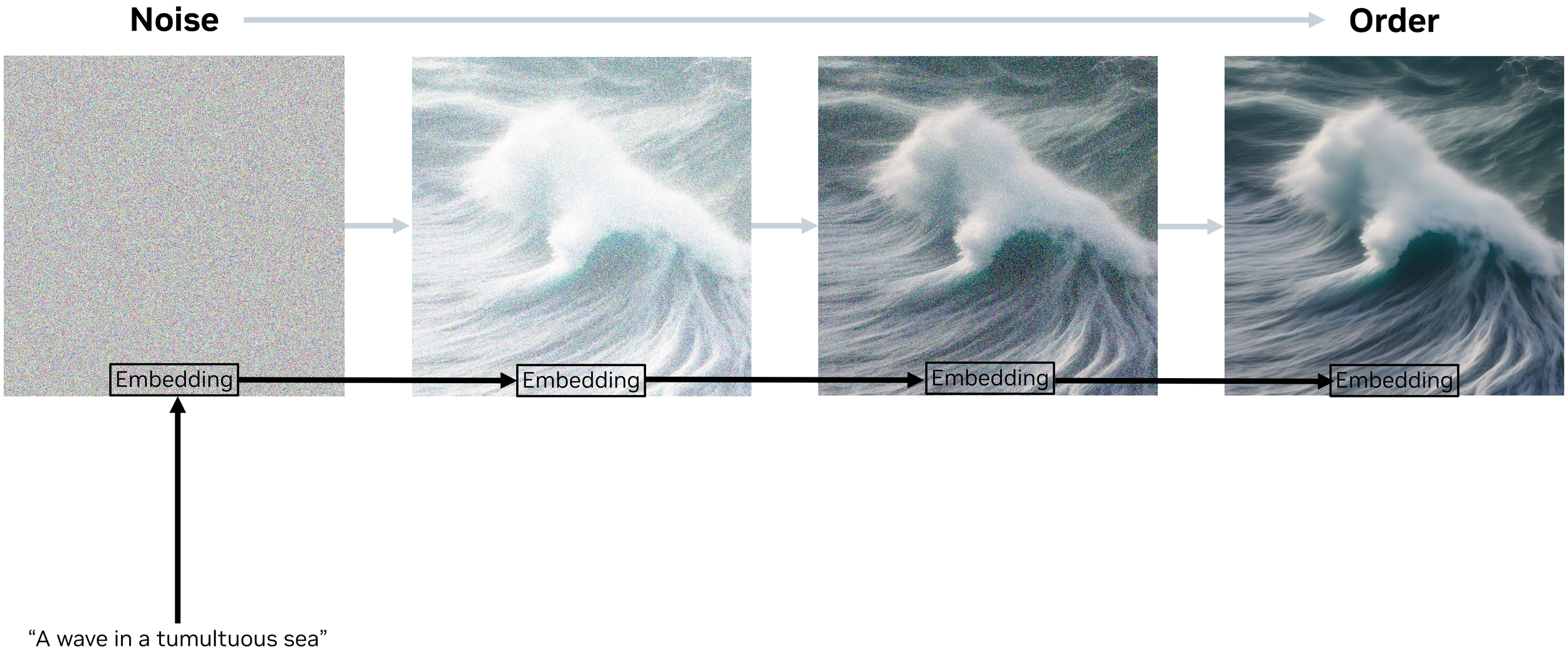
Diffusion models are making waves

These models train to create order from noise



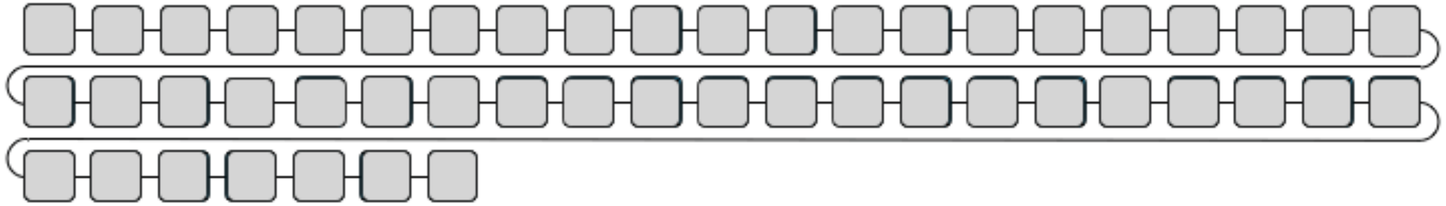
Diffusion models are making waves

These models train to create order from noise



EvoDiff: diffusion models to generate protein sequences

We can design entirely new protein sequences…

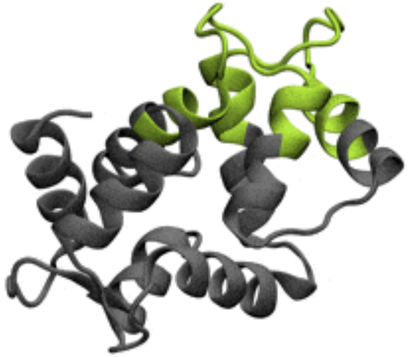


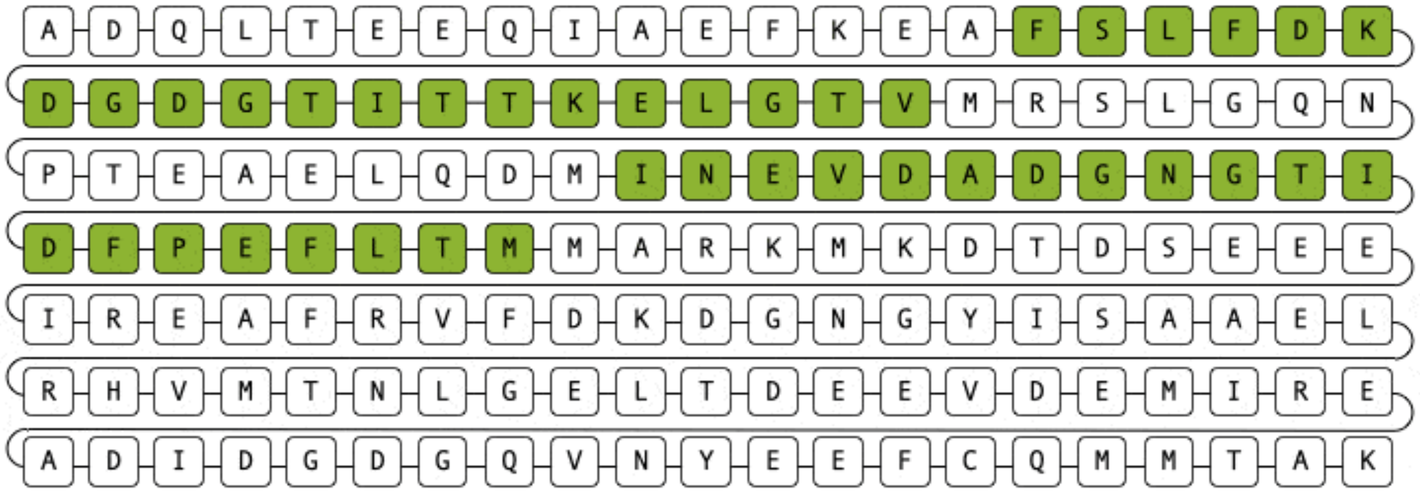
Protein generation with evolutionary diffusion: sequence is all you need. Alamdari S., Thakkar N., et al & Yang K.K.

© Microsoft; Alamdari S., Thakkar N., et al & Yang K.K., with permission from Yang K.K.

EvoDiff: diffusion models to generate protein sequences

Or condition the generation to known motifs!





Protein generation with evolutionary diffusion: sequence is all you need. Alamdari S., Thakkar N., et al & Yang K.K.

© Microsoft; Alamdari S., Thakkar N., et al & Yang K.K., with permission from Yang K.K.

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

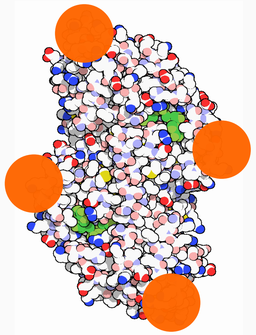
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

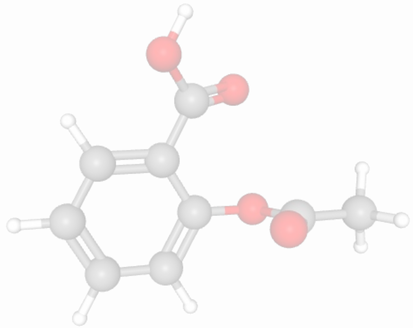
+

CC(=O)OC1=CC=CC=C1C(
=O)O

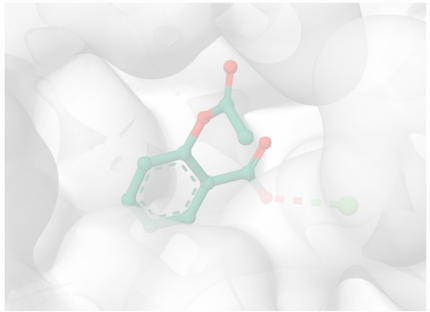
Protein structures



Chemical structures

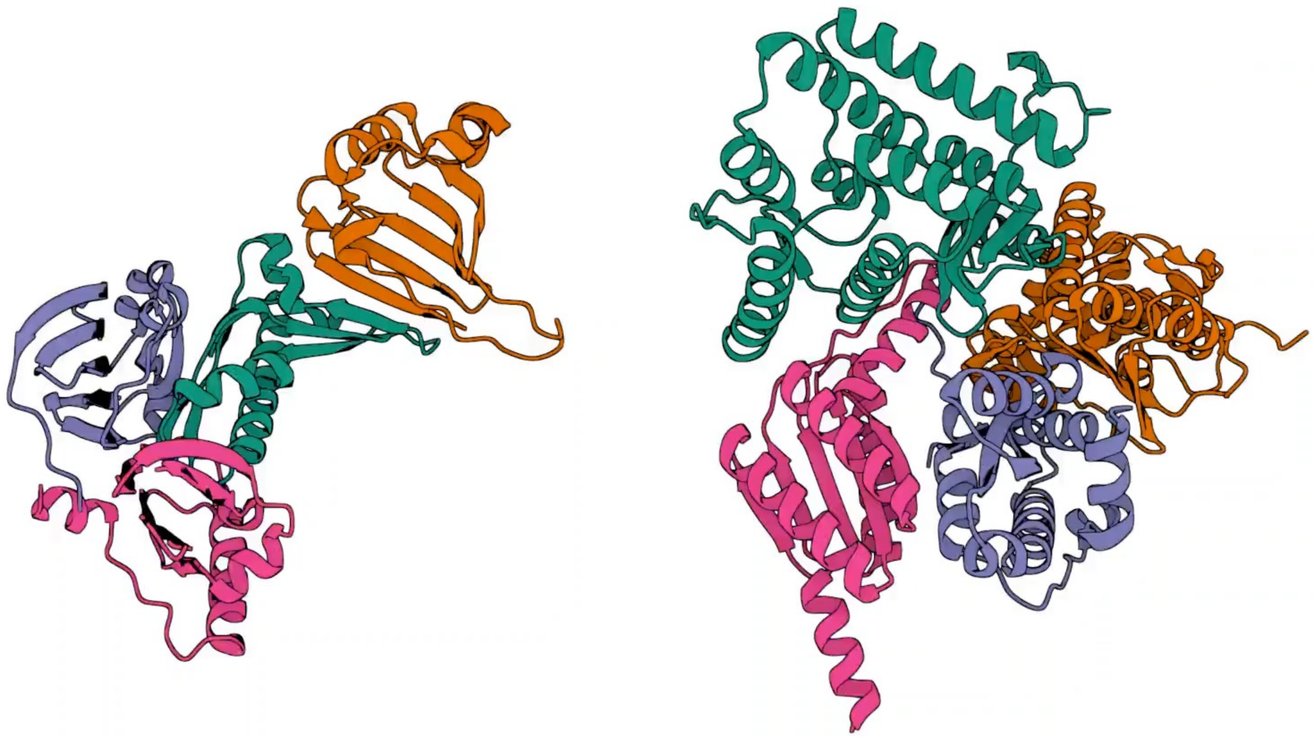


Bound structures



Chroma: diffusion model to generate protein structures

Generate 3D structures directly and tune them to desired shapes and geometries



Illuminating protein space with a programmable generative model. Ingraham J., Baranov M., et al. & Grigoryan G.

© Generate Biomedicines; Ingraham J., Baranov M., et al. & Grigoryan G., with permission from Grigoryan G.

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

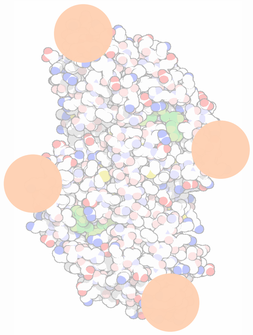
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

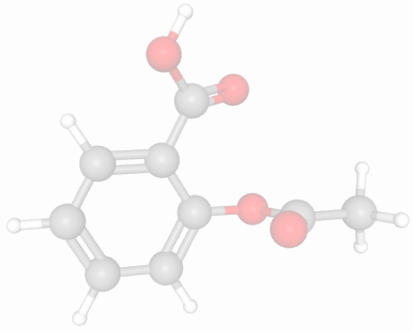
+

CC(=O)OC1=CC=CC=C1C(
=O)O

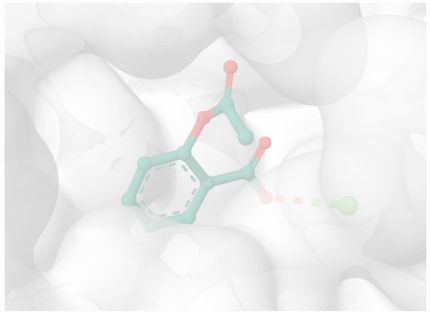
Protein structures



Chemical structures

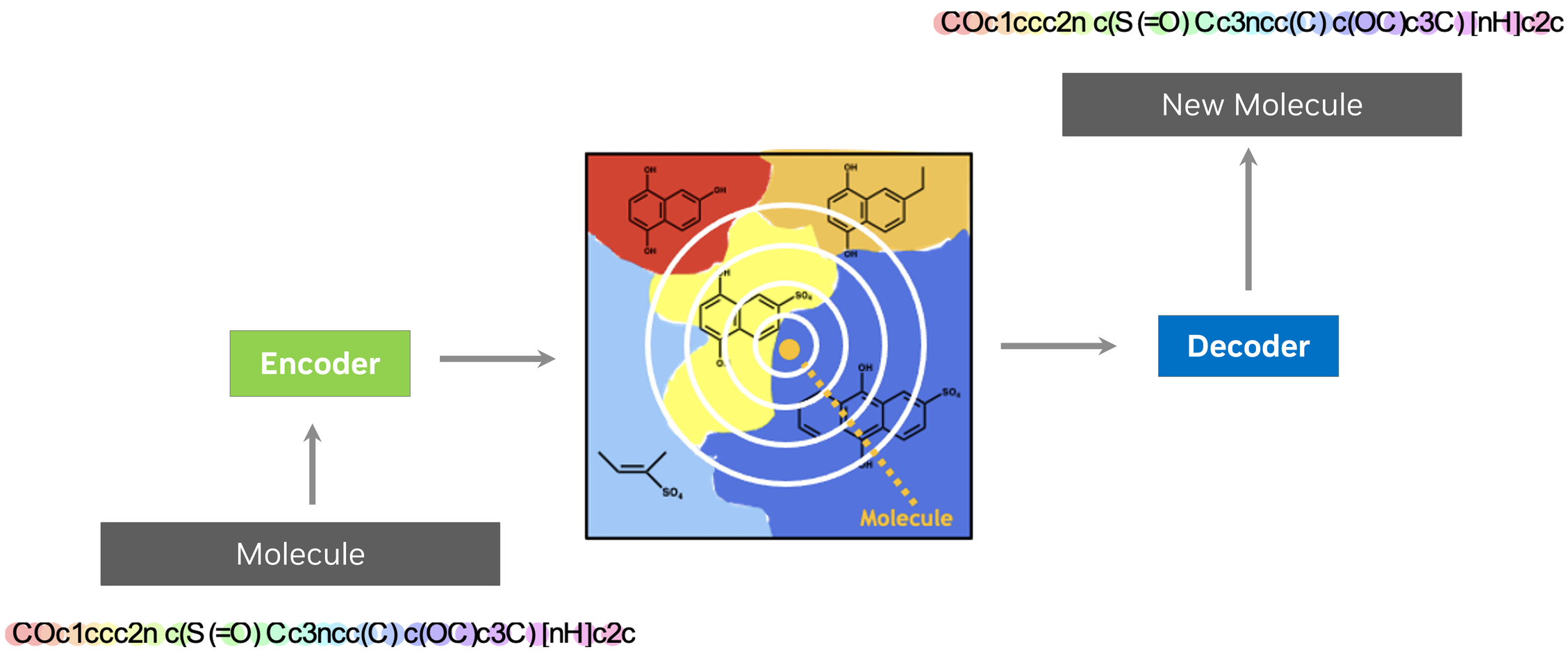


Bound structures



MegaMolBART: encoder/decoder LLM to sample SMILES

The generation in the LLM case happens via gaussian noise added to the embedding of a known sequence



MolMIM: LLM encoder + VAE and LLM decoder to generate smiles

Changing the interface between encoder and decoder through a VAE allows more diverse generation

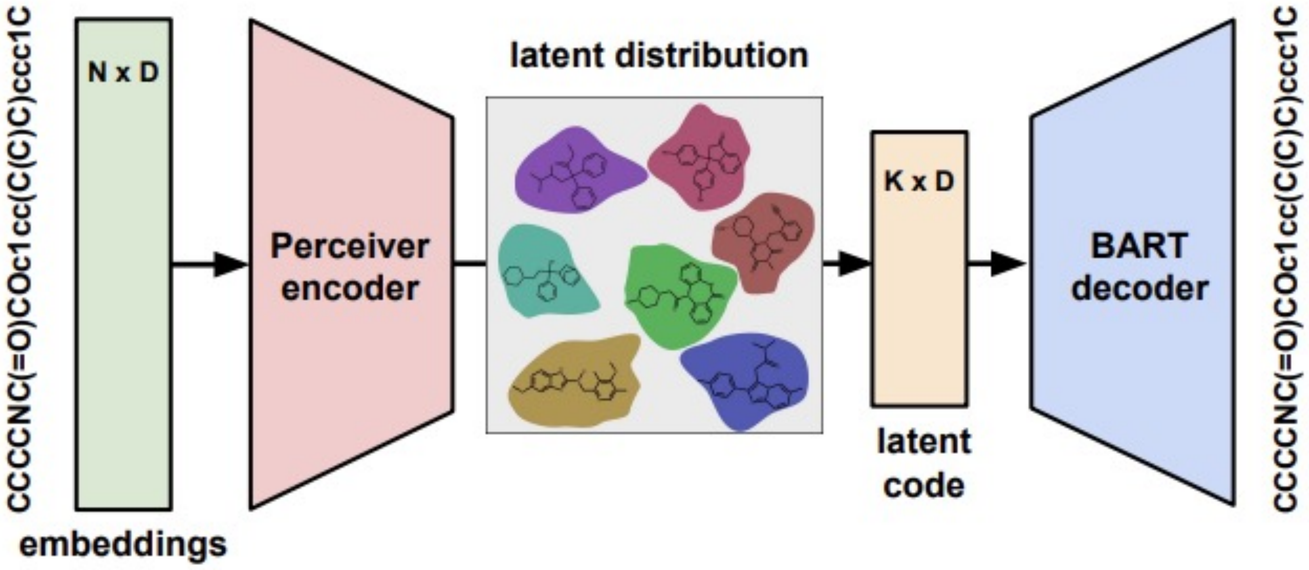


Figure 1: MolMIM is a probabilistic encoder-decoder model with a fixed-size representation that learns a latent distribution clustered around similar molecules.

Improving Small Molecule Generation using Mutual Information Machine.
Danny Reidenbach, Micha Livne, Rajesh K. Ilango, Michelle Gill, Johnny Israeli

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

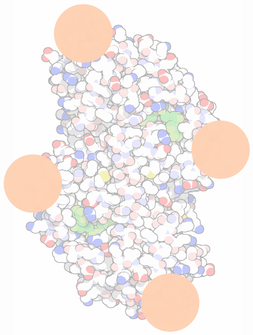
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

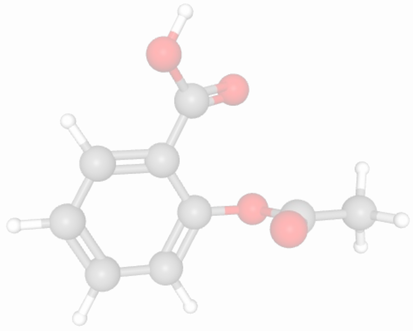
+

CC(=O)OC1=CC=CC=C1C(
=O)O

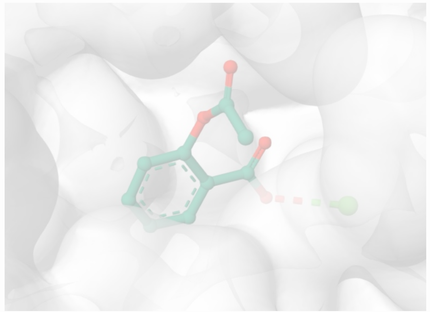
Protein structures



Chemical structures

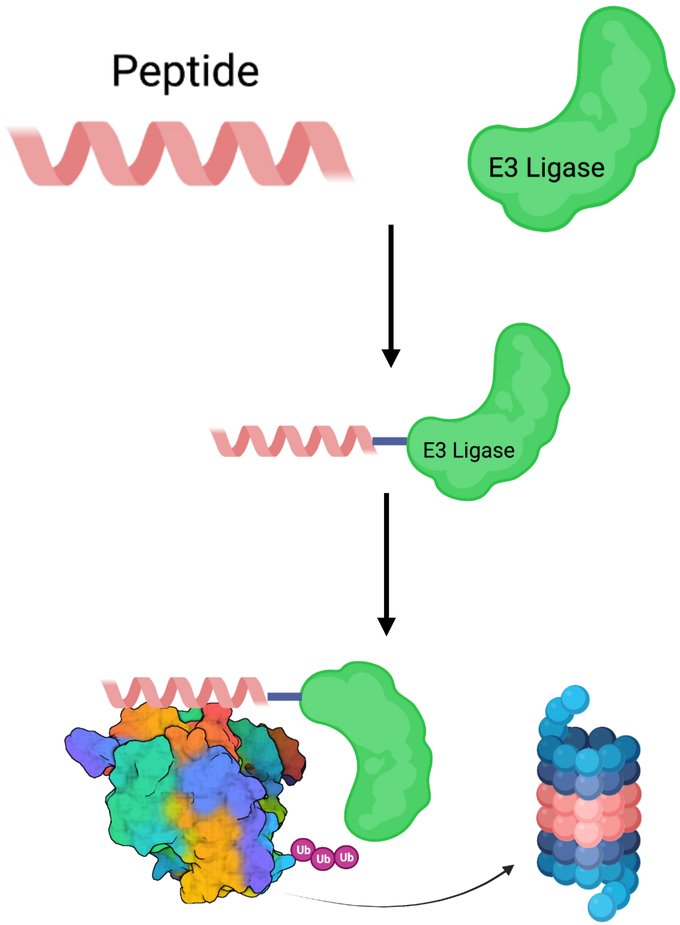


Bound structures



PepPrCLIP: generate peptide sequences that bind to proteins

Peptides are promising biological therapeutic candidates

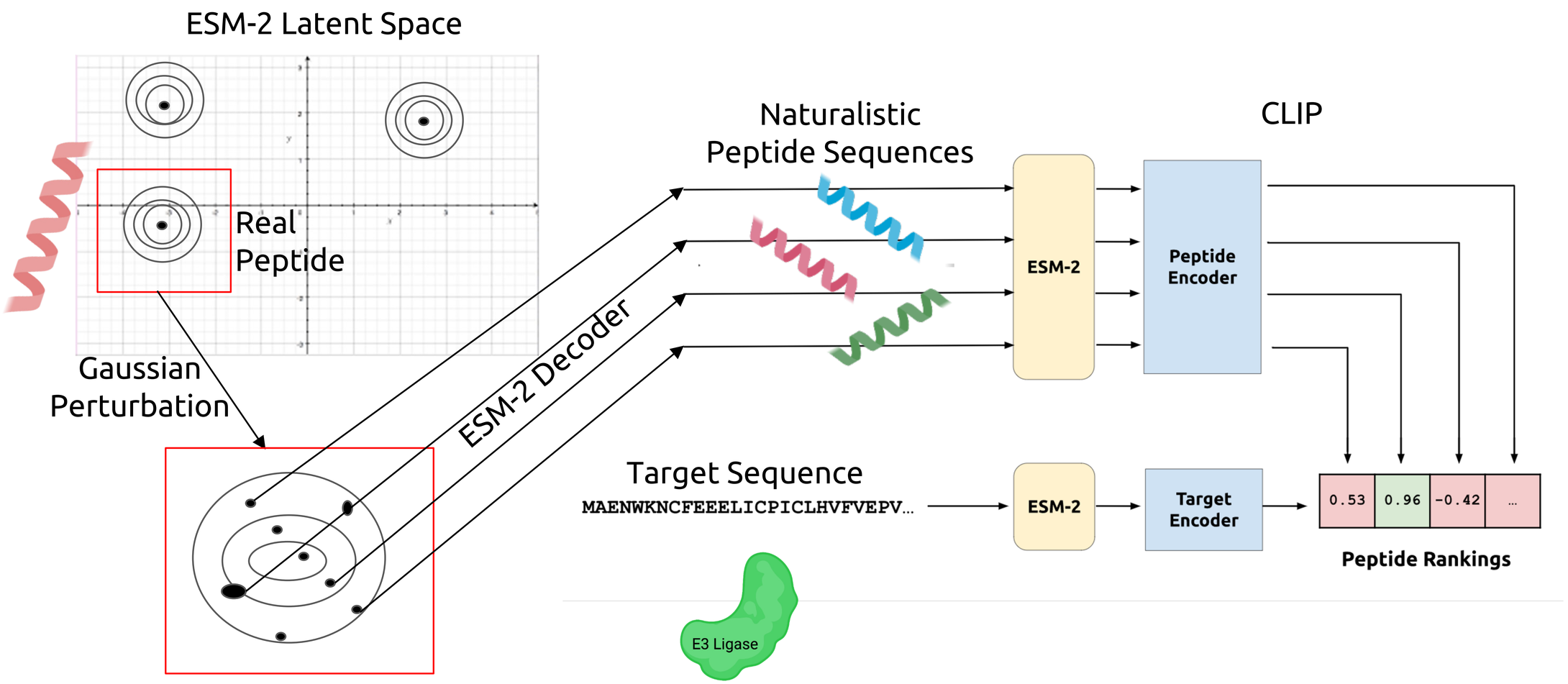


De Novo Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone. Bhat S., Palepu K., et al & Chatterjee P.

© Bhat S., Palepu K., et al & Chatterjee P., with permission from Chatterjee P.

PepPrCLIP: generate peptide sequences that bind to proteins

Peptides are promising biological therapeutic candidates



De Novo Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone. Bhat S., Palepu K., et al & Chatterjee P.

© Bhat S., Palepu K., et al & Chatterjee P., with permission from Chatterjee P.

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

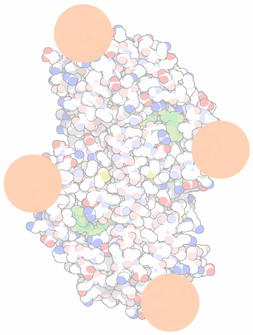
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…

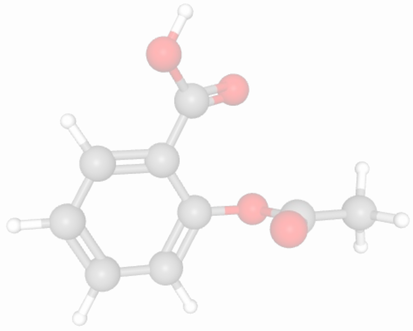
+

CC(=O)OC1=CC=CC=C1C(
=O)O

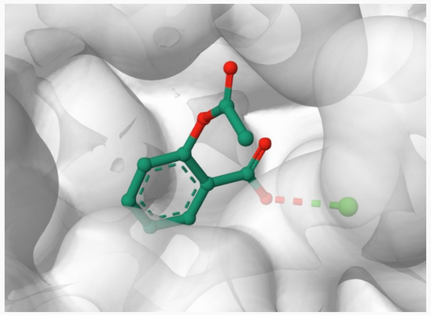
Protein structures



Chemical structures

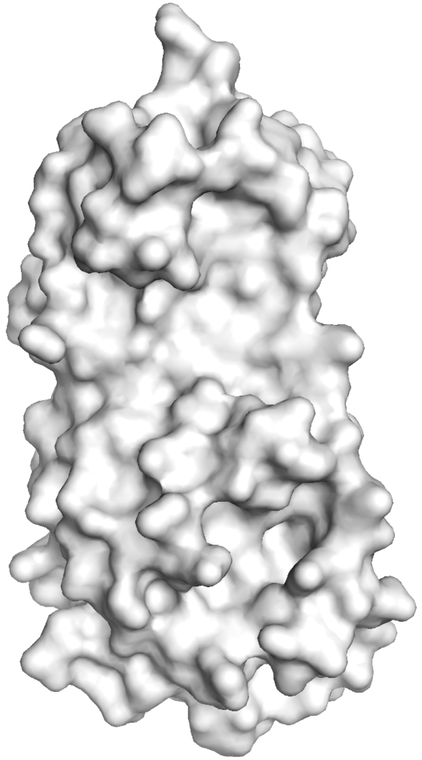


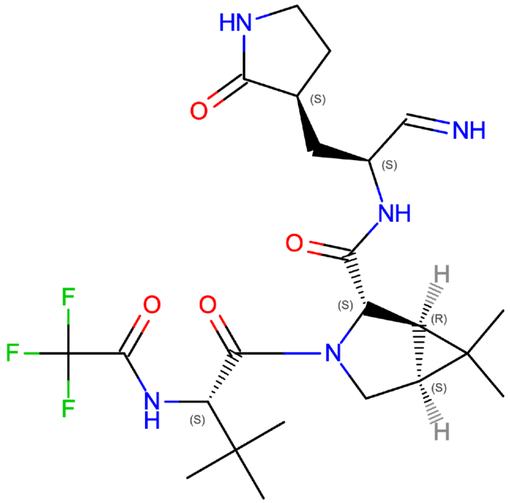
Bound structures

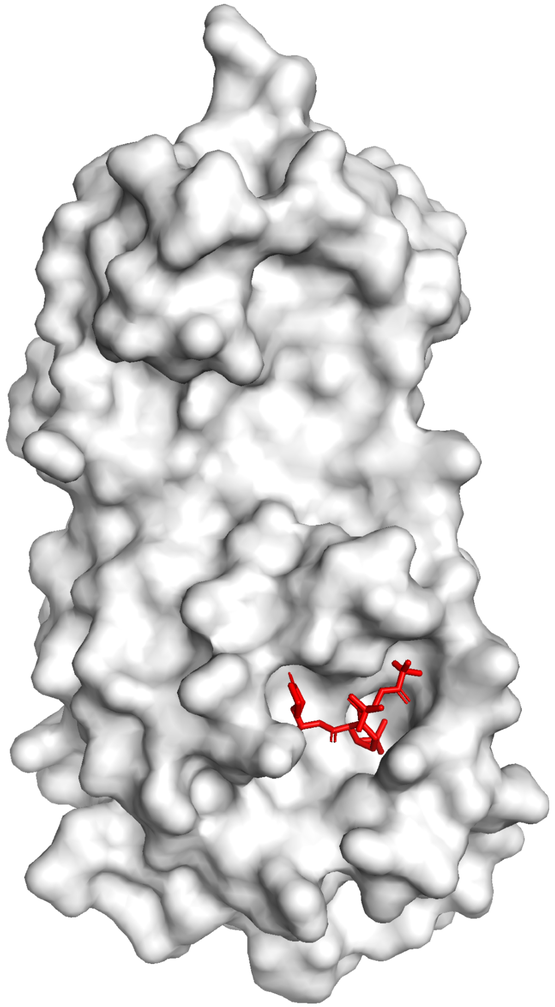


DiffDock: diffusion model for molecular docking

DiffDock predicts the binding pose of small molecules on protein surfaces





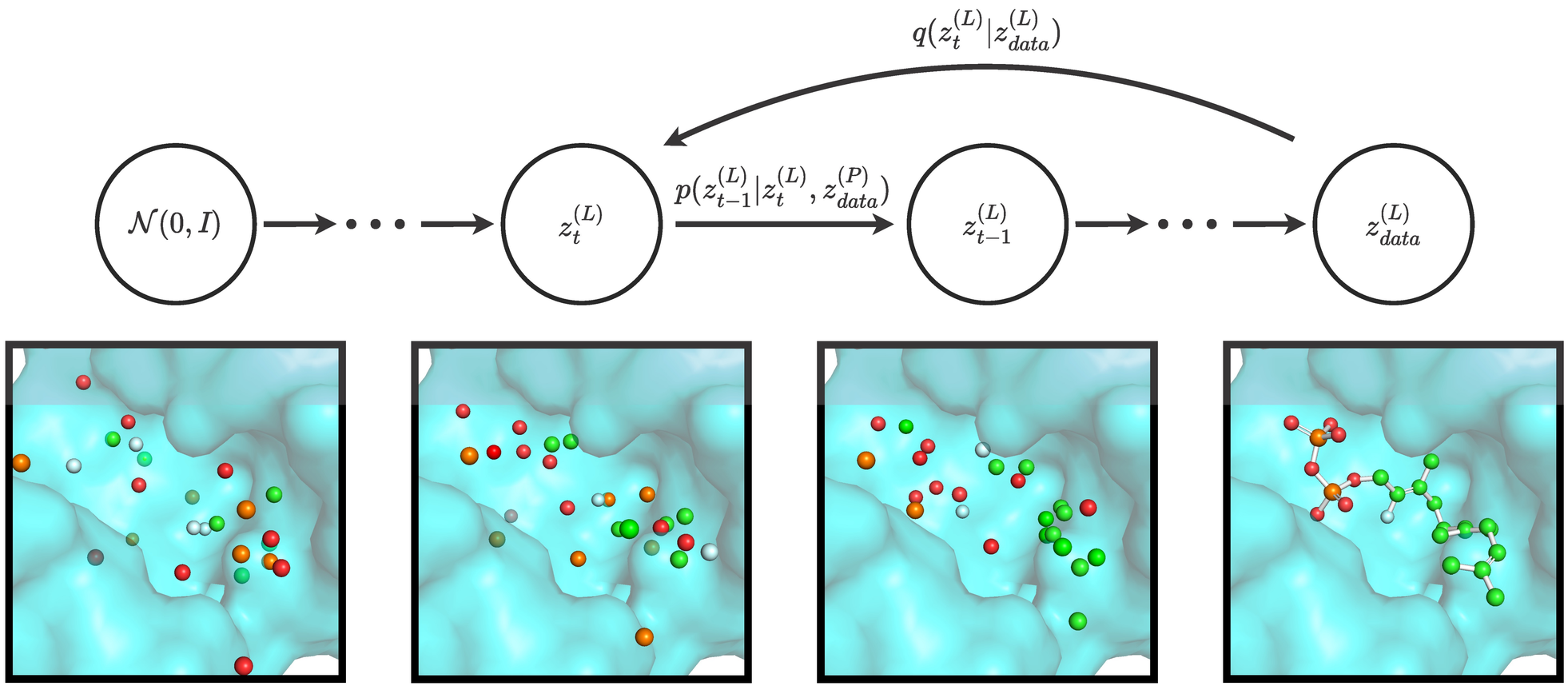


Input: protein structure + molecule

DiffDock Diffusion Steps, Twists, and Turns for Molecular Docking. Corso G., Stärk H., Jing B., Barzilay R., Jaakkola T.
Adapted from public domain material. © Corso G., Stärk H., Jing B., Barzilay R., Jaakkola T., with permission from Stärk H.

DiffSBDD: generating small molecules binding protein pockets

DiffSBDD generates 3D structures of molecules that bind protein 3D structures



Structure-based Drug Design with Equivariant Diffusion Models. Schneuing A., Du Y., et al. & Correia B.

© Schneuing A., Du Y., et al. & Correia B. With permission from Schneuing A.

The next 30 years of scientific innovation is programmable biology

The quest to generate biology and chemestry that meets our needs

Protein sequences

MALLHSARVASGVASAFHPG
LAKGASARASSWWAHVEAMP
PDPILGVTEASKRDTNSKKM
NLGVG…

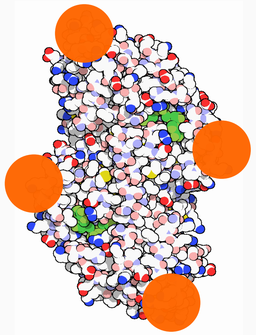
Chemical sequences

CC(=O)OC1=CC=CC=C1C(
=O)O

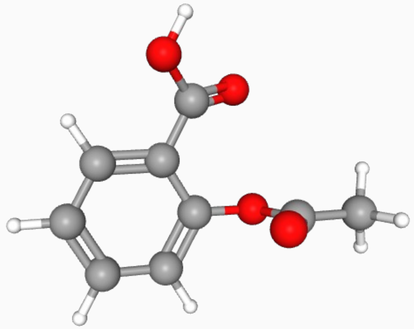
Bound sequences

MALLHSARVASGVASAFHPG
LAKGA…
+
CC(=O)OC1=CC=CC=C1C(
=O)O

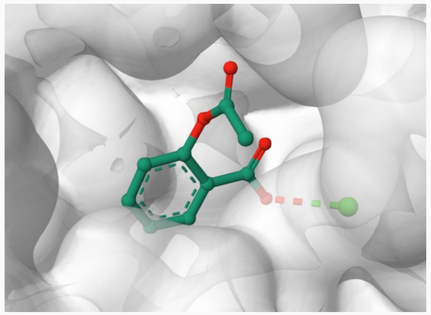
Protein structures



Chemical structures



Bound structures



Thanks for help

- Kevin K. Yang – Microsoft Research New England
- Bruno Correia & Arne Schneuing – EPFL
- Yuanqi Du – Cornell
- Chatterjee Pranam – Duke
- Burkhard Rost & Michael Heinzinger – TUM
- John Ingraham & Grigoryan Grigoryan – Gnerate Biomedicines
- Hannes Stärk – MIT
- Charles Harris – U. of Cambridge
- … many more from the community!

Links to many of the papers mentioned in this presentation at github.com/biolists

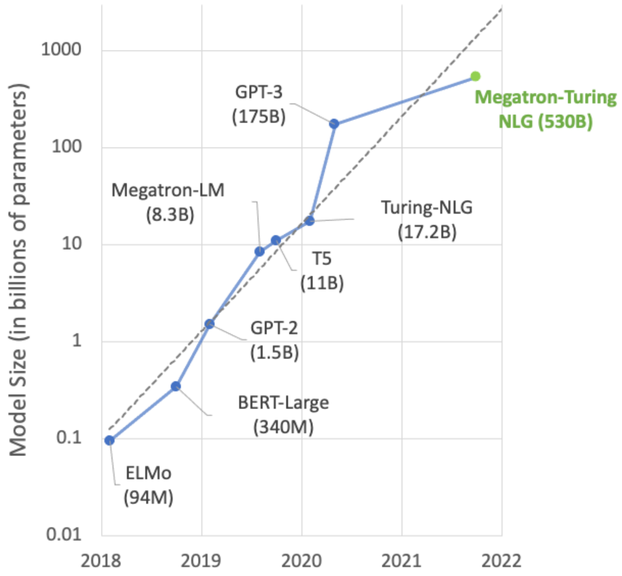
Reach out via X (@sacdallago) or LinkedIn: linkedin.com/in/sacdallago

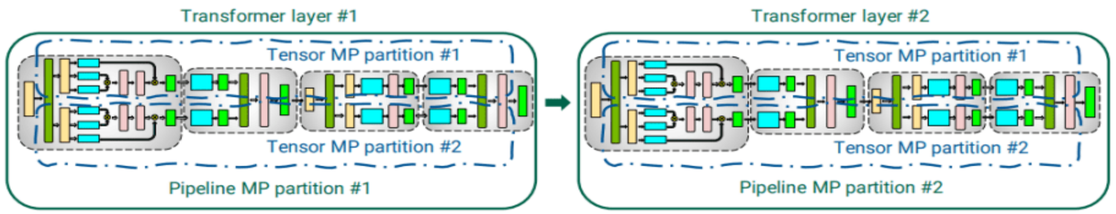
Stay tuned for BioNeMo!

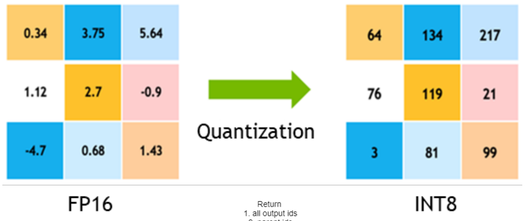
Leveraging BioNeMo for Biomolecular Generative AI

Christian Dallago, Senior Solution Architect for Healthcare and Life Sciences

Training LLMs is perceived as a prohibitively difficult problem







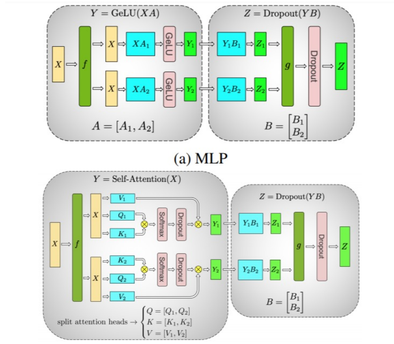
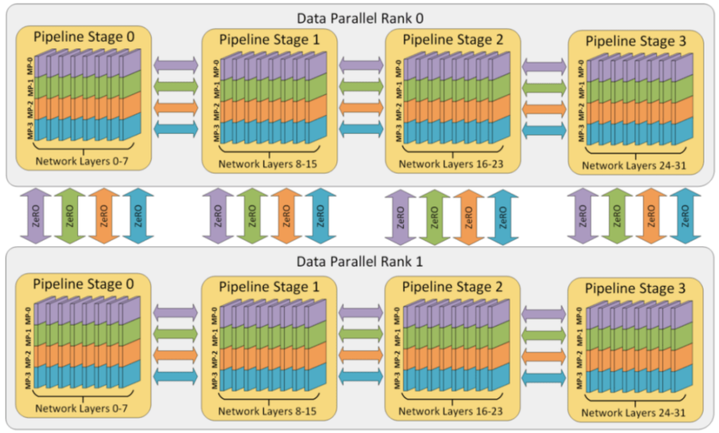
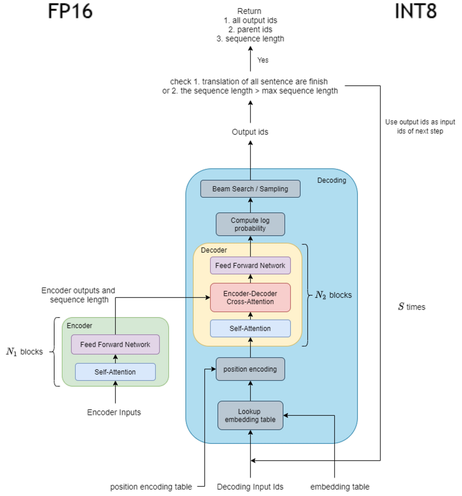
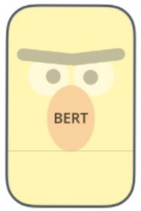


Figure 3. Blocks of Transformer with Model Parallelism. f and g are conjugate. f is an identity operator in the forward pass and all reduce in the backward pass while g is an all reduce in the forward pass and identity in the backward pass.







GPT3

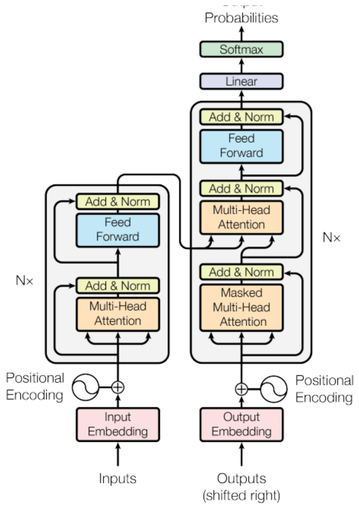


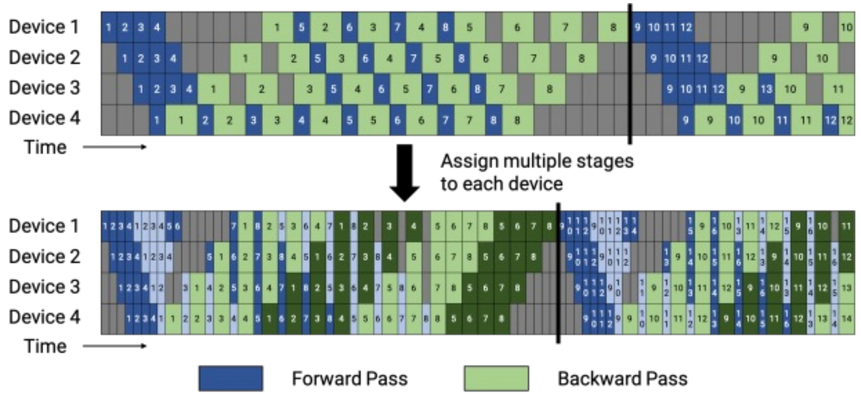
Figure 1: The Transformer - model architecture.

$$AGG(\{h_u^{(l-1)}, \forall u \in N_m(v)\}) = \frac{1}{|N_m(v)|} \sum_{u \in N_m(v)} h_u^{(l-1)}$$

$$\alpha_m = \frac{\exp(e_m)}{\sum_{m=1}^{M} \exp(e_m)}$$

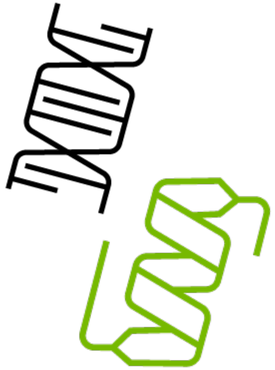
$$h_v^{(l)} = \sum_{m=1}^{M} \alpha_m \cdot h_v^{(l)[m]}$$

$$h_v^{(l)[m]} = W^{(l)[m]} \cdot \mathrm{CONCAT}\left(W_d^{(l)[m]} \cdot h_v^{(l-1)}, W_s^{(l)[m]} \cdot AGG(\{h_u^{(l-1)}, \forall u \in N_m(v)\})\right)$$



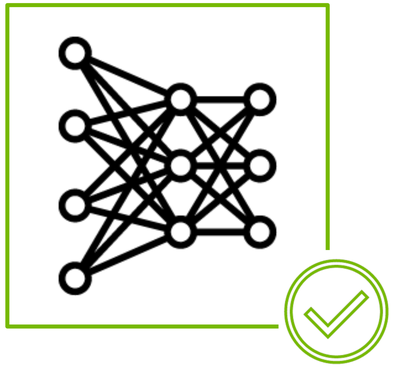
BioNeMo Training Service

Biomolecular LLM training on DGX Cloud



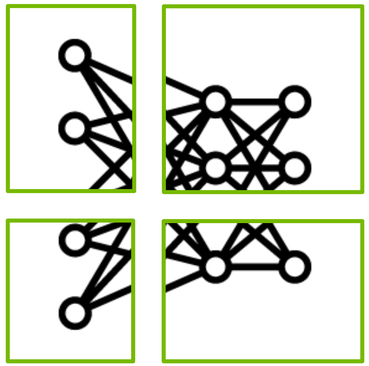
Domain Specific

Built in support for pre-processing biology and chemistry data



Pre-Trained Models

Out-of-the-box architectures and validated checkpoints for state-of-the-art models



Seamless Scaling

Distributed training recipes for model-parallel multi-GPU training

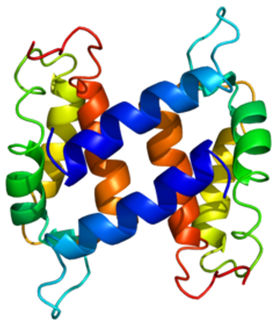
Full Stack Optimization

Optimized for NVIDIA's DGX Cloud to maximize throughput and reduce cost

Protein and Chemical Language Model Training Paradigm

Pretraining Enables Learned Representations For Downstream Tasks, But How Good Are They?

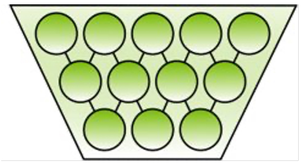
SEQUENCE



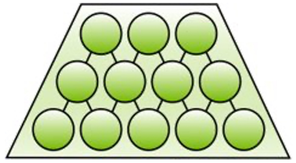
```
ATIKVTKANTLFKIAAY
PMMVAKKRIILRAMKRI
ILRIKVTKANTLFKIIL
RIKVTNTLFKIILRIKV
TNTMVAKKRITNKLISL
ALALIRMKTLILRIKVT
KAVTKANTLFKIILRIK
VGGKANTLFKIILRIKV
```

PRE-TRAINING OF PROTEIN LM

ATIKVTKANTLFKIAAY...



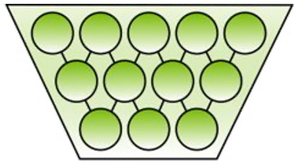
Embedding



ATIKVTKANTLFKIAAY...

EMBEDDINGS EXTRACTION

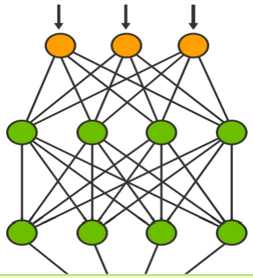
MVAKKRITNGGKLISLALA...



6.3	3.	1.9	-8	4.3	6.	1.7	0.9	0.8	8.5	2.9
1.1	4.1	7.	-3.	13.	7.7	7.8	3.8	0.4	7.6	2.8
3.6	4.9	1.2	1.7	8.3	1.1	2.8	0.5	4.4	8.8	10.
2.	8.	7.8	2.8	1.3	1.5	5.7	8.2	5.4	1.3	0.5
…	…	…	…	…	…	…	…	…	…	…
1.6	5.	5.7	-6.	6.8	2.6	4.8	3.7	0.8	6.1	7.9
3.1	9.5	4.8	3.9	-3	2.7	11.	1.2	1.1	3.6	2.7
1.3	1.7	4.4	4.7	4.5	2.2	7.	1.9	2.8	1.2	0.3

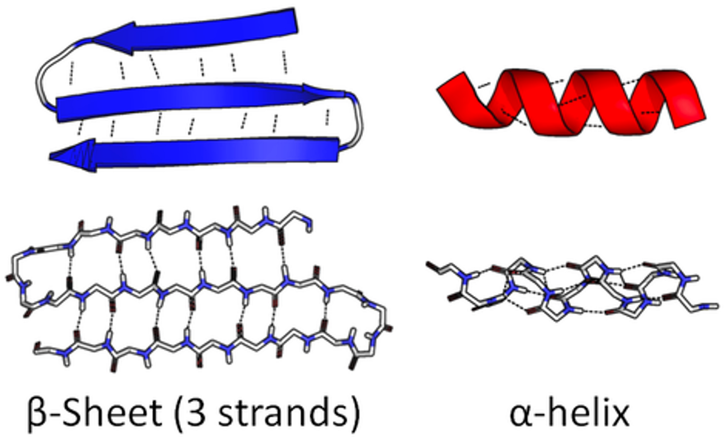
SECONDARY STRUCTURE PREDICTION

6.3	3.	1.9	-8	4.3	6.	1.7	0.9	0.8	8.5	2.9
1.1	4.1	7.	-3.	13.	7.7	7.8	3.8	0.4	7.6	2.8
3.6	4.9	1.2	1.7	8.3	1.1	2.8	0.5	4.4	8.8	10.
2.	8.	7.8	2.8	1.3	1.5	5.7	8.2	5.4	1.3	0.5
…	…	…	…	…	…	…	…	…	…	…
1.6	5.	5.7	-6.	6.8	2.6	4.8	3.7	0.8	6.1	7.9
3.1	9.5	4.8	3.9	-3	2.7	11.	1.2	1.1	3.6	2.7
1.3	1.7	4.4	4.7	4.5	2.2	7.	1.9	2.8	1.2	0.3

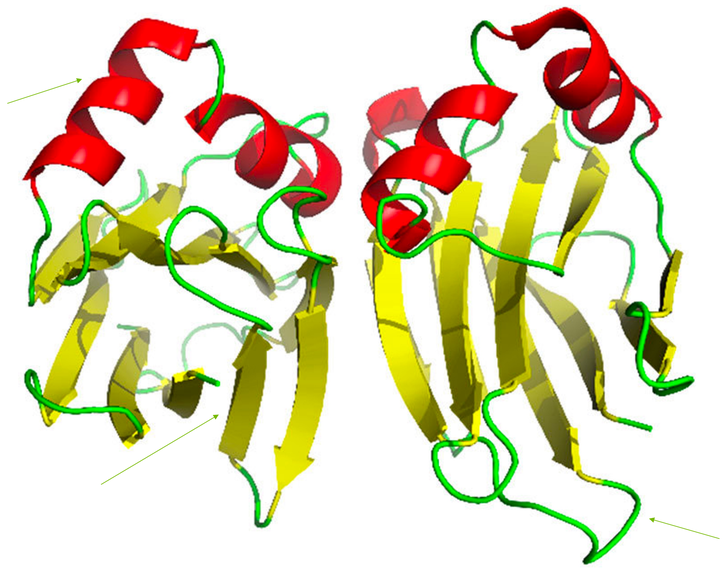


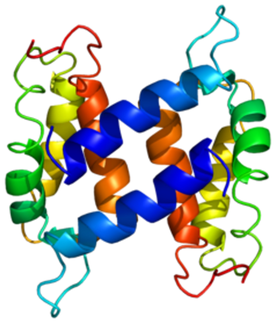
CEEEECCEEEEEECCCCCC…

Protein Secondary Structure



- Secondary structure is a protein's local spatial conformation
- 3 main states – strand (E), helix (H) and coil (C)
- Can be classified with more granularity into 8 states
- Each residue belongs to some structural state
- Secondary structure prediction is classification problem



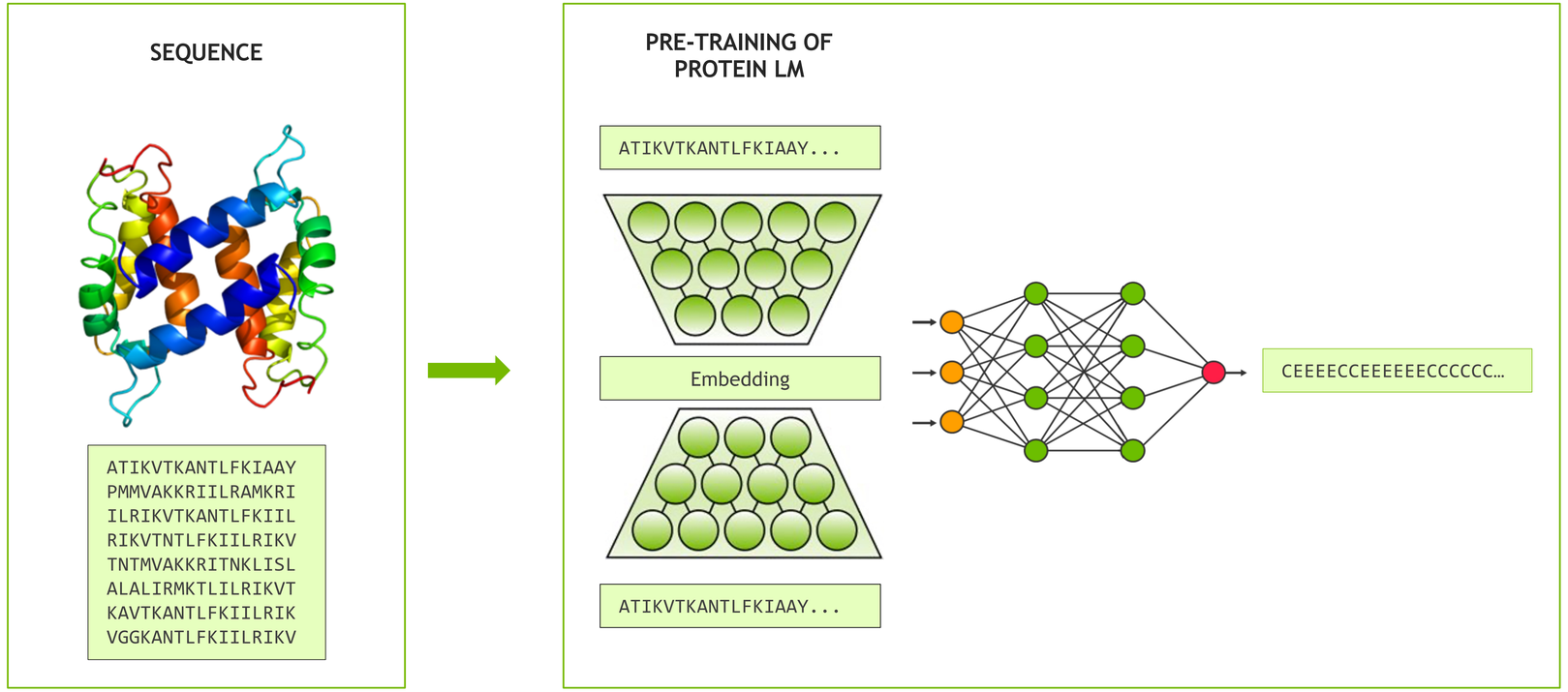


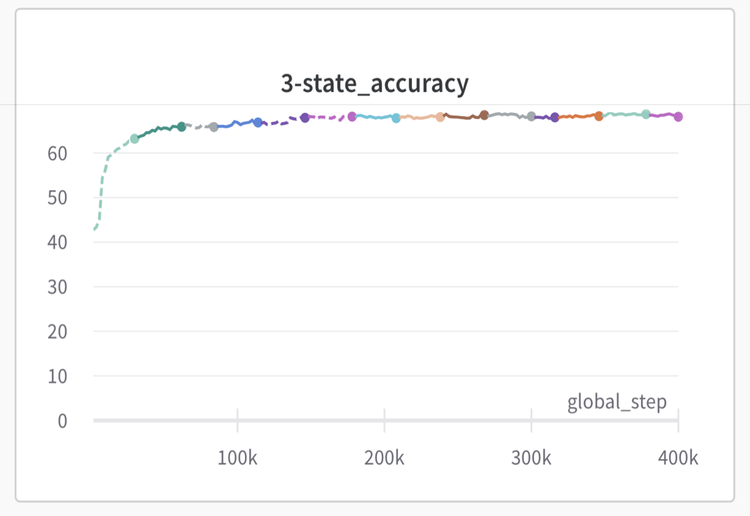
```
ATIKVTKANTLFKIAAY
PMMVAKKRIILRAMKRI
ILRIKVTKANTLFKIIL
RIKVTNTLFKIILRIKV
TNTMVAKKRITNKLISL
ALALIRMKTLILRIKVT
KAVTKANTLFKIILRIK
VGGKANTLFKIILRIKV
```

```
CEEEECCEEEEEECCCC
CCHHHHHHHHHHCCCCC
CCCEECCCCEEEEEECC
CCEEEECHHHHHCCCCC
CCCCHHHHCCCCCCCCC
CCCCCCCCCCCCCCCCC
CCCCCCEEECCCCCCEE
EECCCCCCHHHHHCC
```

Validation-In-The-Loop During Protein LM Pre-Training

Secondary structure prediction as a validation step

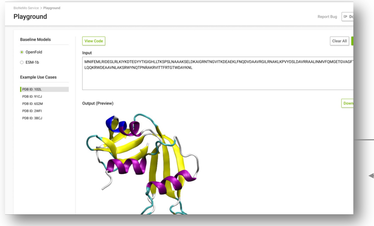




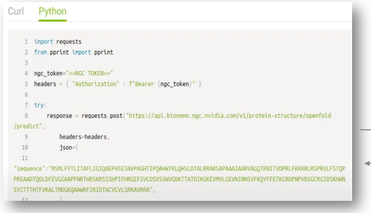
BioNeMo

For Generative AI in Biology - Supporting 9 SOTA Small Molecule & Protein AI Models at HPC Scale

BioNeMo



WEB



API

Drug Discovery Applications

- Protein Structure Prediction
- Biomolecule Representation
- Biomolecule Generation
- Molecular Docking

BioNeMo Models

- AlphaFold2
- OpenFold
- ESMFold
- ESM1
- ESM2
- ProtGPT2
- MegaMolBART
- MoFlow
- DiffDock

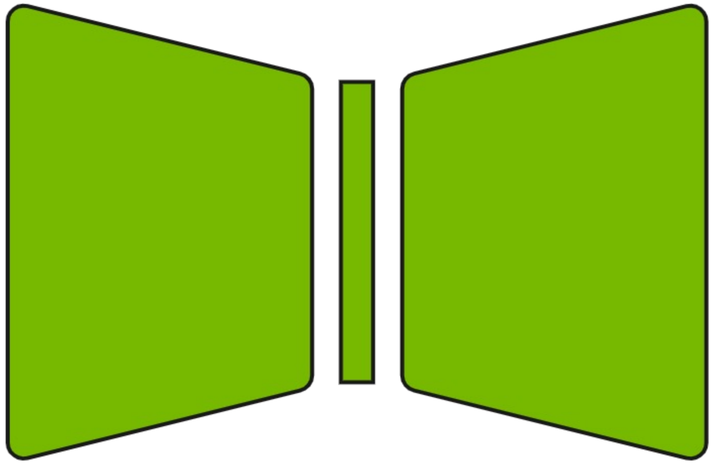
NVIDIA DGX Cloud

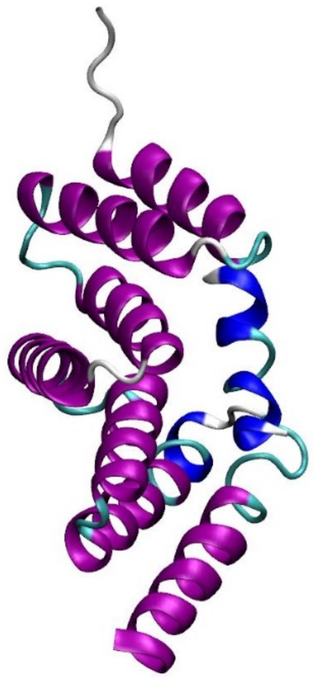
Training, fine-tuning, inferencing and deploying

BioNeMo – Protein Structure Prediction

AlphaFold2, OpenFold, ESMFold | Fold Proteins with BYO MSA or MSA-Free

MHHHHHHSSGLVPRGSGMKETAAAKFERQH
MDSPDLGTDDDDKAMADIGSENLYFQSMSK
IFVNPSAIRAGMADLEMAEETVDLINRNIE
DNQAHLQGEPIEVDSLPEDIENLYFQGMES
DKIVFKVNNQLVSVKPEVIVDQYEYKYPAI
QDHTKPSITLGKAPDLNKAYKSILSGMNAA
KLDPDDVCSYLAAAMELFEGVCPEDWTSYG
IMIARKGDKITPATLVDIKRTDIEGNWALT
GGQDLTRDPTVAEHASLVGLLLSLYRLSKI
SGQNTGNYKTNIADRIEQIFETAPFAKIVE
HHTLMTTHKMCANWSTIPNFRFLAGTYDMF
FSRVEHLYSAIRVGTVVTAYEDCSGLVSFT
GFIKQIN





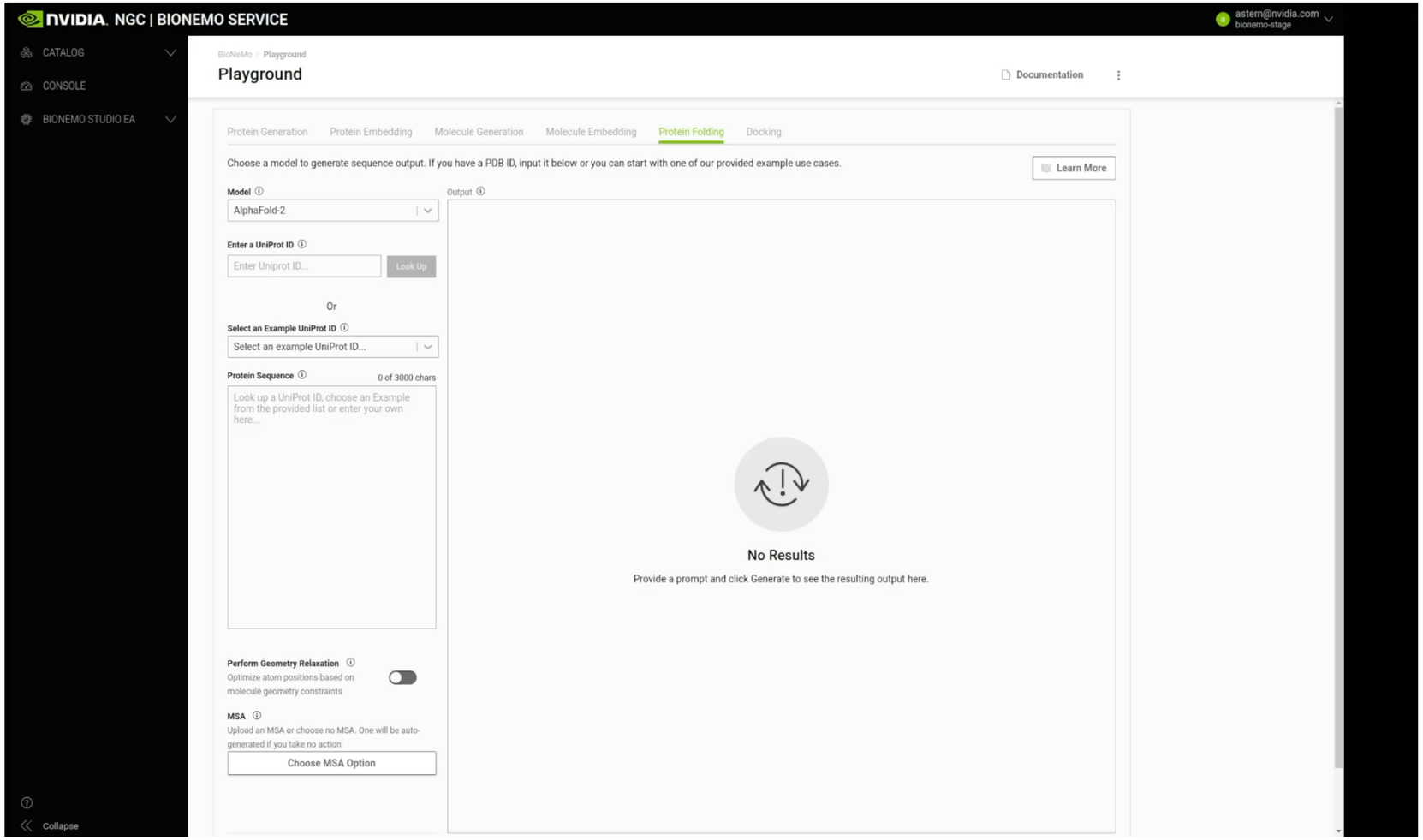
AlphaFold-2: DeepMind; Jumper et al.
RoseTTAFold: UW IPD; Baker et al.
OpenFold: OpenFold Consortium; AlQuraishi et al.
ESMFold: MetaAI; Lin et al.

Models Accessible on an Easy-to-Use Graphic User Interface

Interactive Inference | Visualization | Experimentation

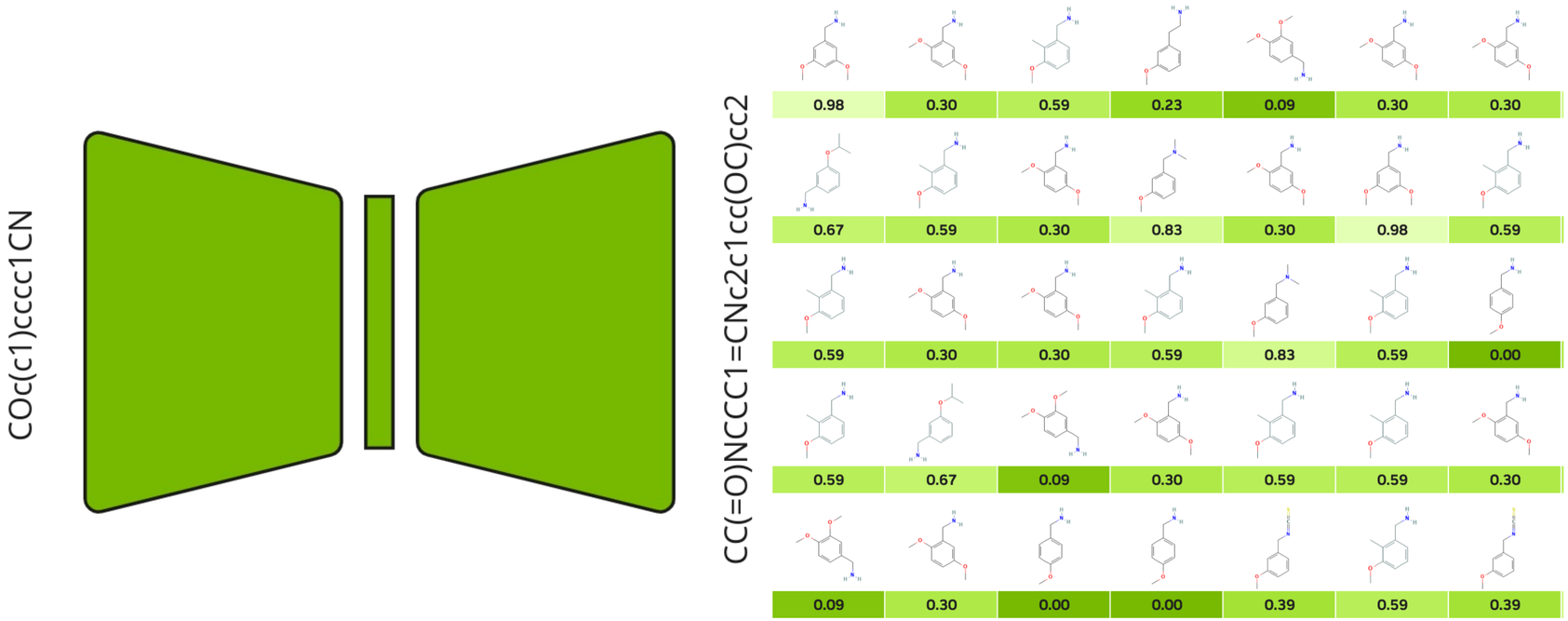
ESM Fold

3D Protein Structure Prediction



BioNeMo – Molecule Generation

MoFlow & MegaMolBART | Generate Molecules with Optimized Properties



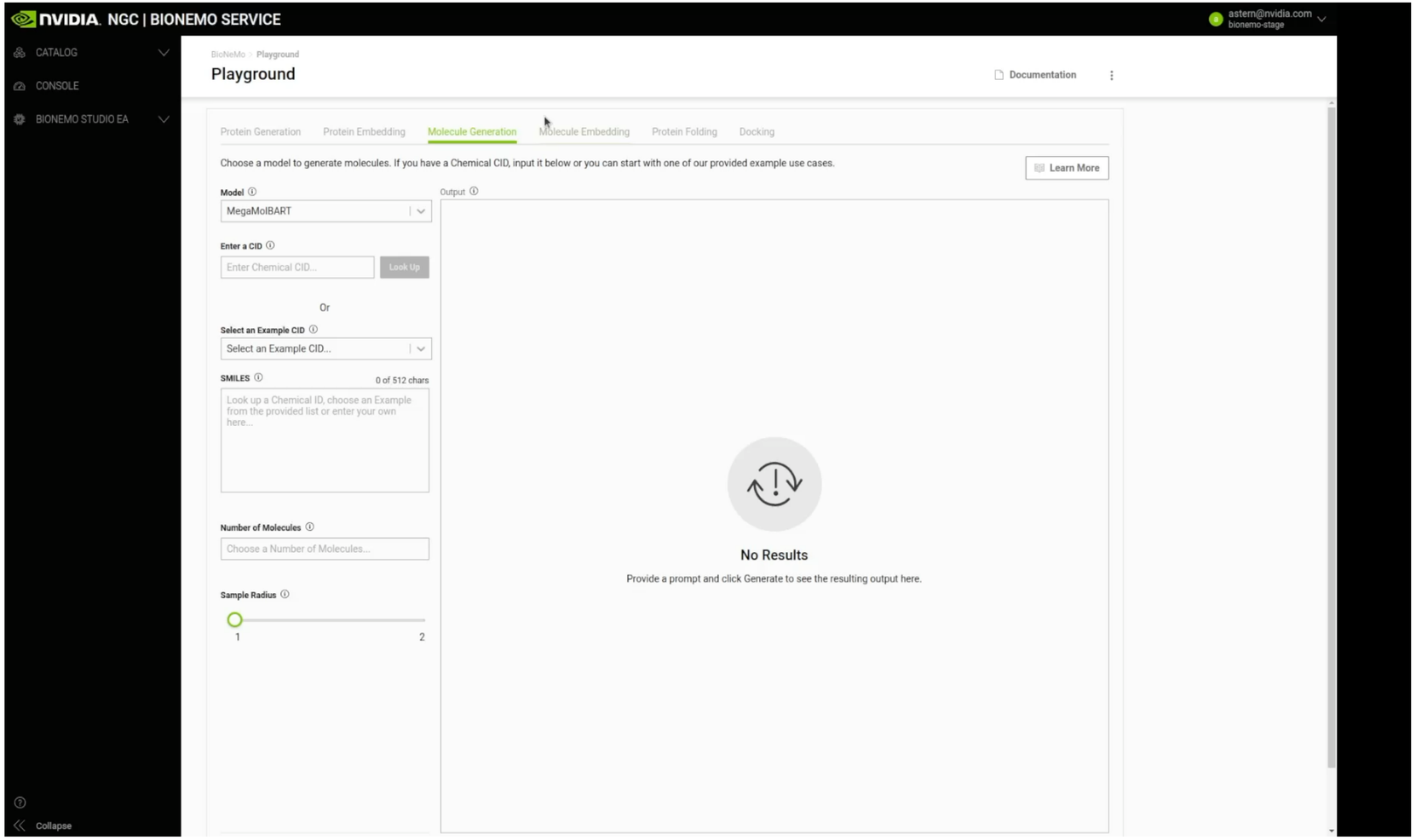
MoFlow: Chengxi Zang and Fei Wang
MegaMolBART: NVIDIA & AstraZeneca; Bjerrum, Irvin, Engkvist, et al.

Models Accessible on an Easy-to-Use Graphic User Interface

Interactive Inference | Visualization | Experimentation

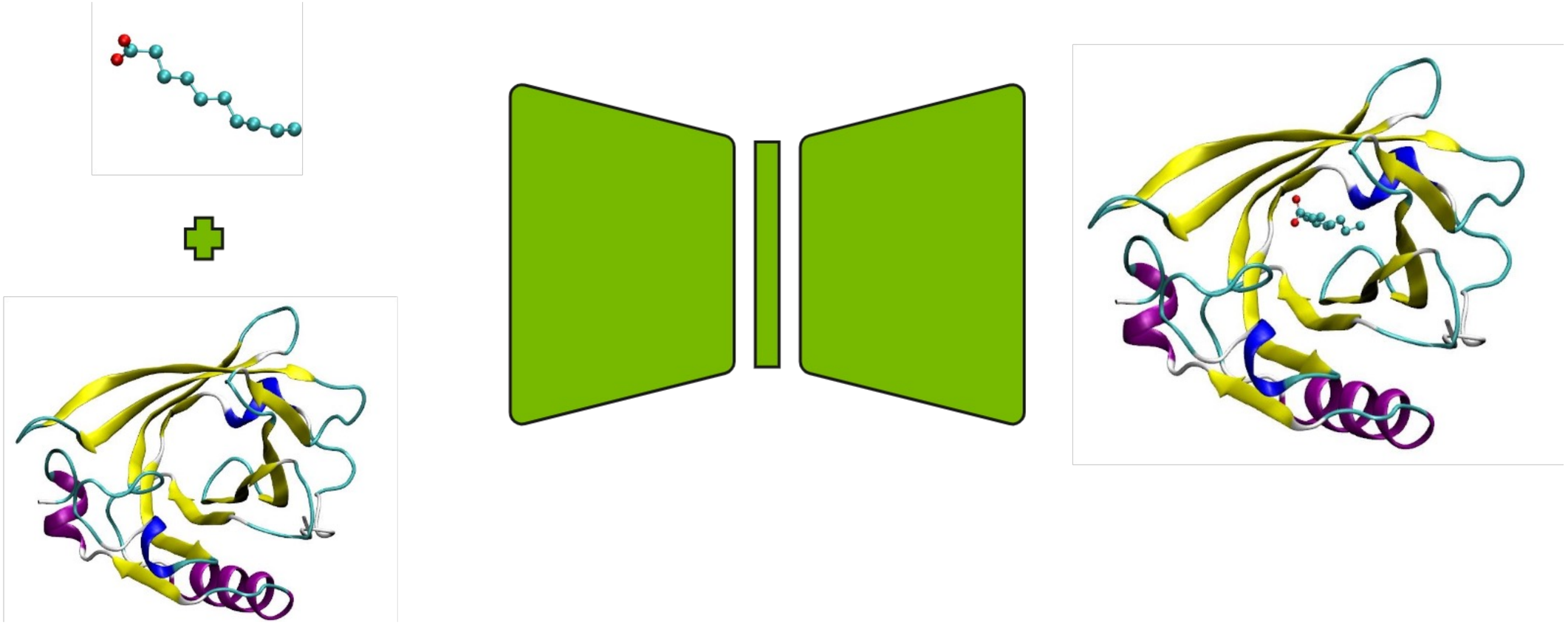
MegaMolBART

Molecular Generation



BioNeMo – Docking Pose Prediction

DiffDock | Pose Prediction Significantly Exceeding State-of-the-Art Methods



38% top-1 success rate (RMSD<2A) on PDB-Bind

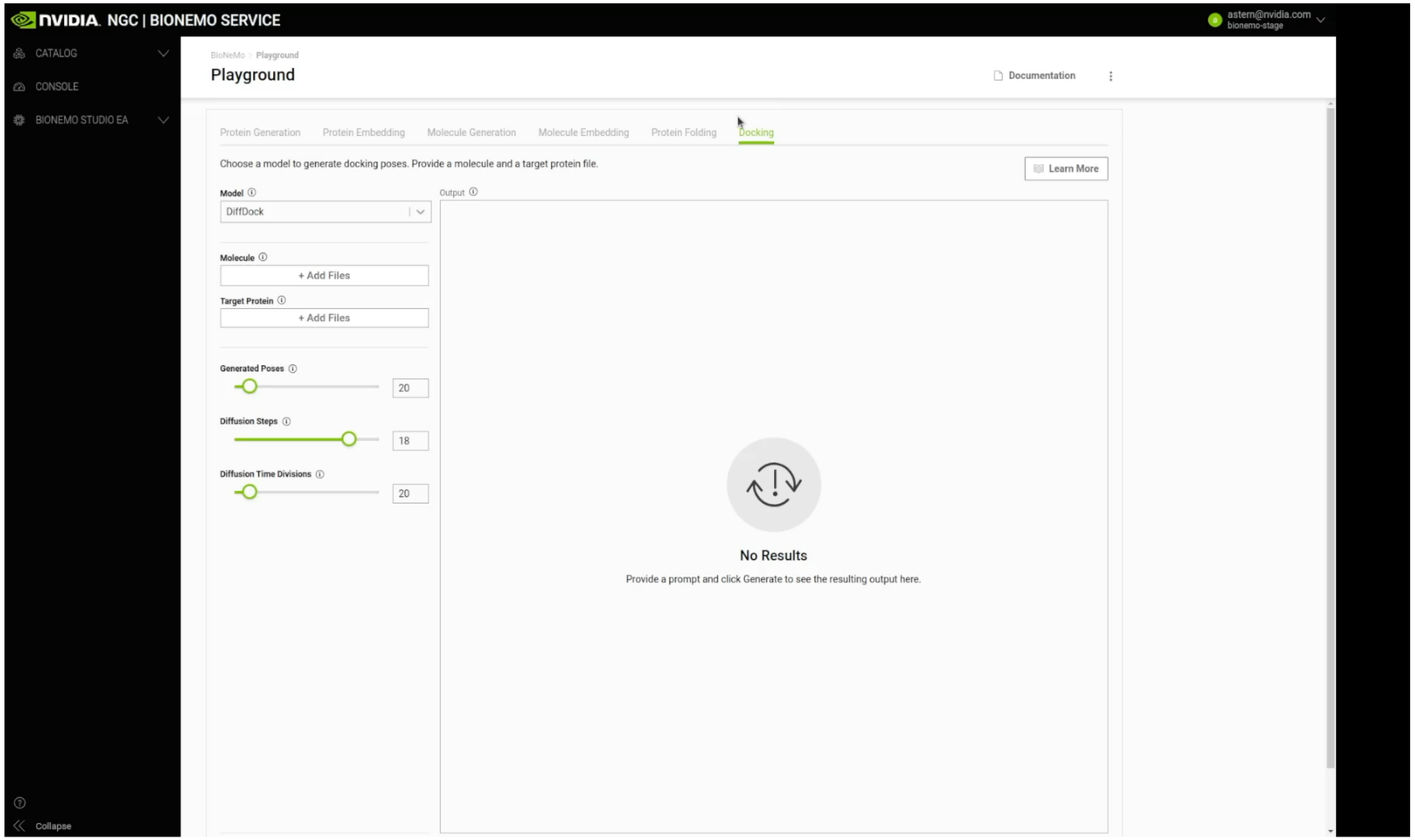
Gabriele Corso, Hannes Stark, Bowen Jing, Regina Barzilay & Tommi J

Models Accessible on an Easy-to-Use Graphic User Interface

Interactive Inference | Visualization | Experimentation

DiffDock

Molecular Docking



NVIDIA NGC | BioNeMo LLM Service

BioNeMo Service
Home
Lab
Queue
Tasks
Datasets

BioNeMo Service > Lab

Lab

Documentation Learn More

Protein Generation | Protein Embedding | Molecule Generation | Molecule Embedding | **Protein Folding** | Docking

Choose a model to generate sequence output. If you have a PDB ID, input it below or you can start with one of our provided example use cases.

Model
OpenFold

Enter a UniProt ID
Enter UniProt ID... Look Up

Or
Select an Example UniProt ID
Select an example UniProt ID...

Protein Sequence
Look up a UniProt ID, choose an Example from the provided list or enter your own here...

Perform MD Refinement
Brief description of what this does

MSA
Upload an MSA or choose no MSA. One will be auto-generated if you take no action.
Choose MSA Option

Output

View Code

OpenAPI

Curl | Python

```
curl -X POST "https://api.bionemo.ngc.nvidia.com/v1/protein-structure/openfold/predict" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $YOUR_NGC_API_TOKEN" \
  -d '{
        "sequence":
"MSFSGKYQLQSQENFEAFMKAIGLPEELIQKGKDIKGVSEIVQNGKHFKFTITAGSKVIQNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTTFKNIK
SVTELNGDIITNTMTLGDIVFKRISKRI"
}'
```

Learn how to integrate the API into your application here
Click here to generate a new API key.

Copy Code | Done

Clear | Generate

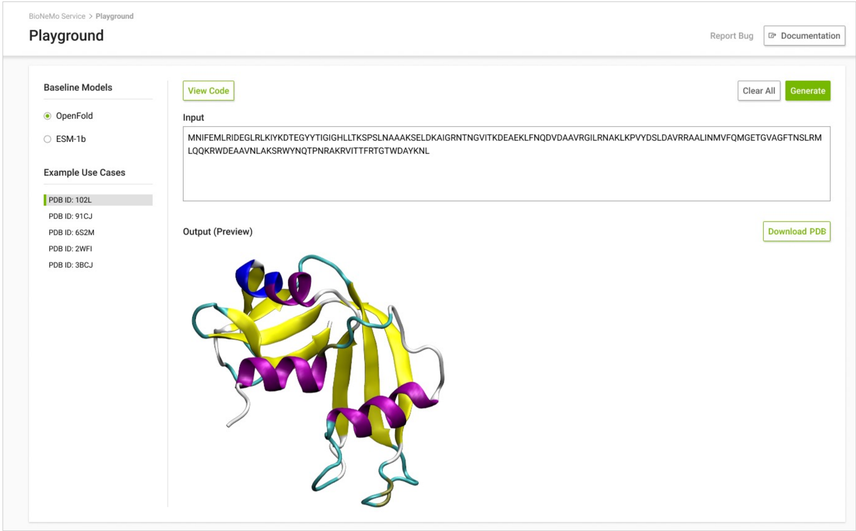
Give Feedback | View Code | Download

Collapse

Application Versions

Building AI Drug Discovery Workflows & Inferencing with BioNeMo Service

Easy Access to State-of-the-Art AI



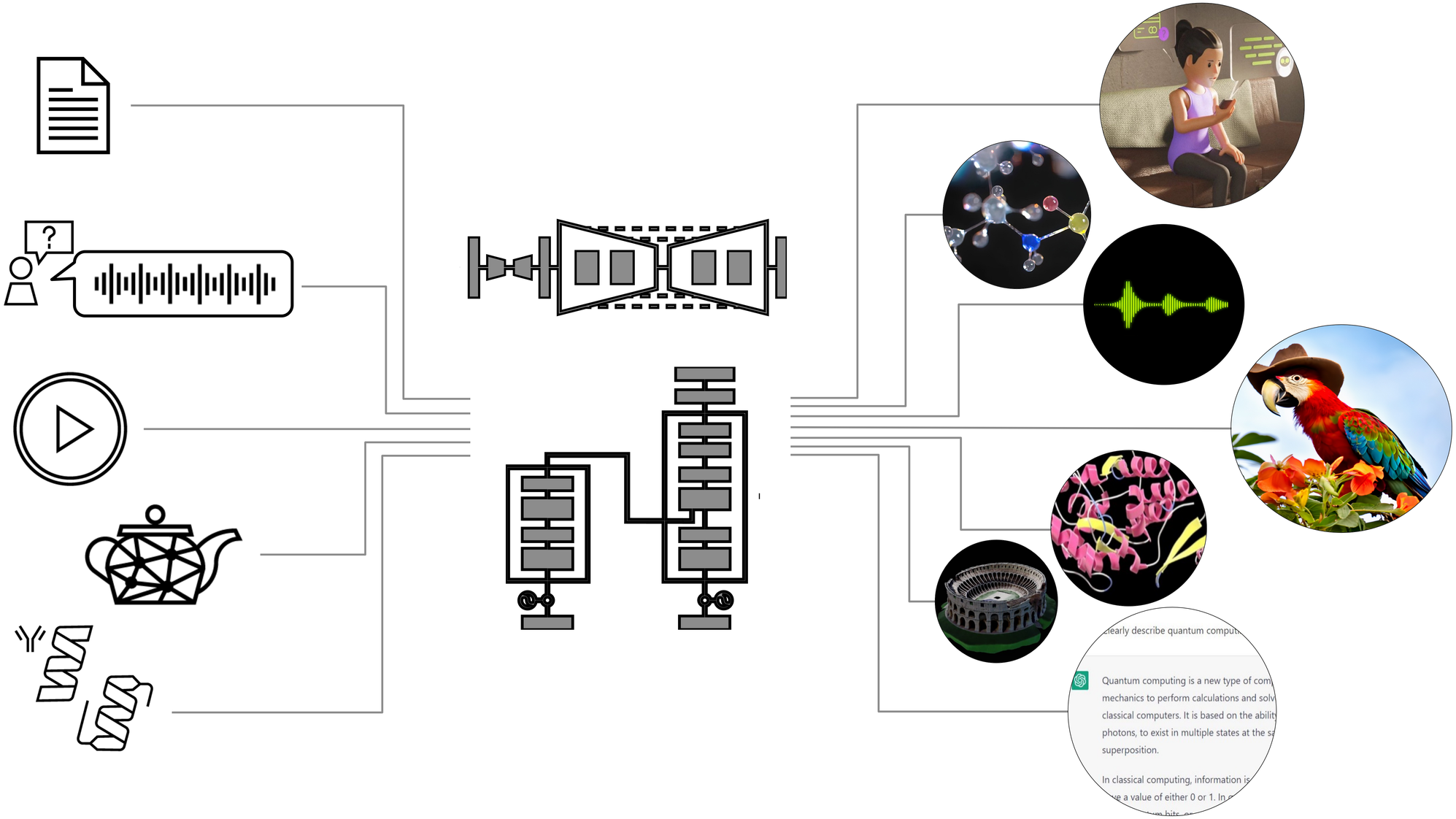
- Instant Access to Generative AI Workflows
  - 3D Protein Structure Prediction
  - Biomolecule Representation
  - Biomolecule Generation of proteins and small molcules
  - Pose Prediction
- State-of-the-Art Model Architectures
  - 3D Protein Structure Prediction: AlphaFold2, OpenFold, ESMFold
  - Protein Generation: ProtGPT2
  - Small Molecule Generation: MegaMolBART, MoFlow
  - Protein Property Predictions: ESM1nv, ESM2
  - Molecular Docking: DiffDock
- Seamless, instantaneous and scalable application integration with fully managed API endpoints
- Easy Web Interface with Interactive 3D Visualization and dataset download

Register for Early Access:
NVIDIA.com/Bionemo

Large Language Models and Generative AI for Life Sciences

Dr. Chelsea Sumner, NALA Healthcare AI Startup Lead

What is Generative AI?



Generative AI is Transforming Business

TEXT GENERATION

Summarization

GPT-3

Marketing Copy

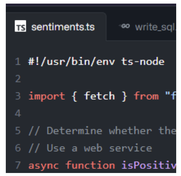
TRANSLATION

Translating Wikipedia

NLLB-200

Real-Time Translation

CODING



Dynamic Code Commenting

CODEX

Function Generation

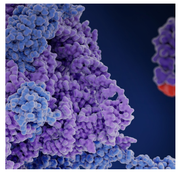
VISUAL CONTENT

Brand Creation

e-Diffi

Gaming Characters

LIFE SCIENCE



Molecular Representations

MegaMolBART

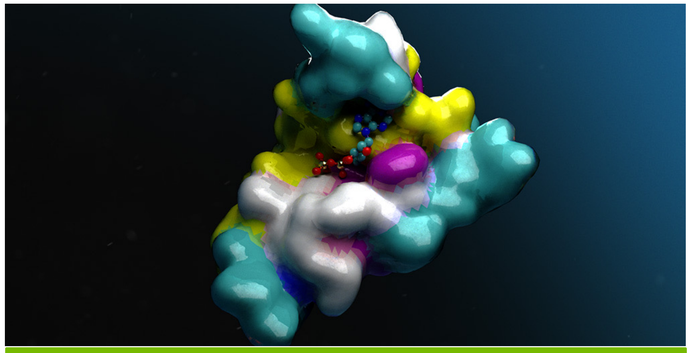
Drug Discovery

Enterprises that adopt next-generation AI like LLMs and Generative AI are **2.6X more likely to increase revenue by 10% or more** but must invest in their AI infrastructure to fully reap the benefits.

-Accenture Research. Breakthrough Innovation: Is your organization equipped for breakthrough innovation? WEF 2023.

NVIDIA Generative AI Platform

NeMo
Language



BioNeMo
Life Sciences

Picasso
Visual Content

NVIDIA AI Enterprise

DGX & DGX Cloud

aws Google Cloud Microsoft Azure ORACLE Cloud Infrastructure

Cloud

DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO

On-Prem

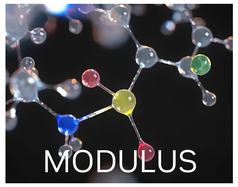
Accelerated Compute Infrastructure

NVIDIA Solutions for Transforming BioPharma R&D

Accelerating workflows from early drug discovery to clinical trials

Applications

Medical Imaging AI Framework

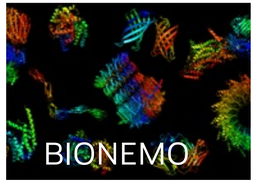


Physics-Informed Neural Networks

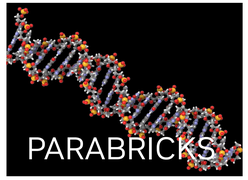
Recommender Systems

Customizable Conversational AI

NLP



AI Drug Discovery for Biomolecular Models



GPU-Accelerated & AI-Based Genomic Analysis Software

Acceleration Libraries

TensorRT	Triton	RAPIDS	CUDA Quantum	FLARE
Deep learning inference	Inference server	GPU accelerated data science library	Accelerated quantum computing research and development	Secure, privacy preserving multi-party collaboration

NVIDIA AI

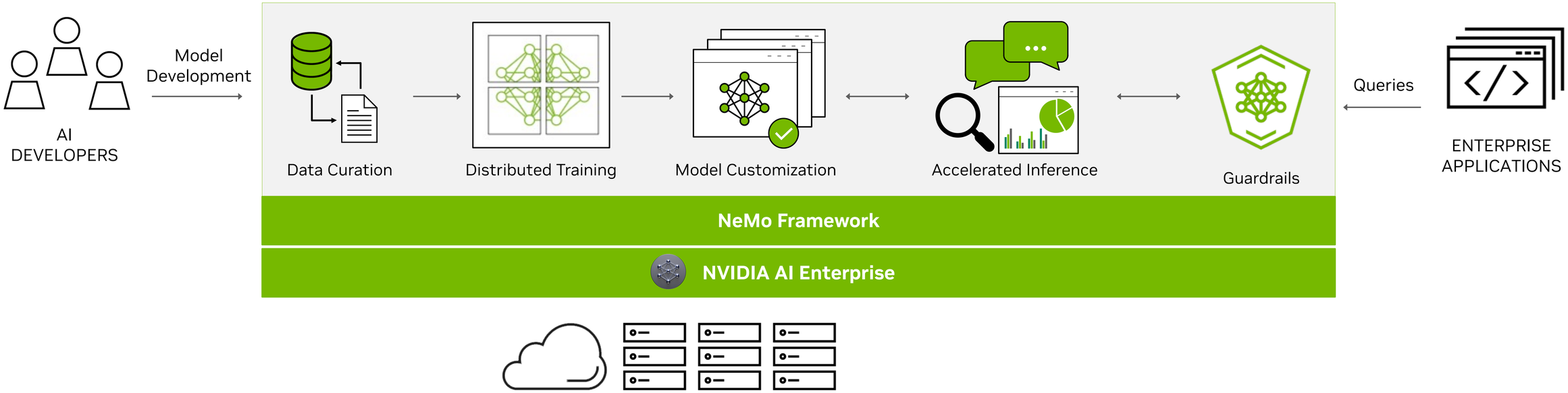
Systems

DGX | HGX	EGX | IGX
H100 | A100 | V100 A40 | A30 | A10	RTX A6000 | T4 | A10 | Jetson

Cloud | Data Center | Edge | Embedded

NVIDIA NeMo for Custom LLMs

End-to-end, cloud-native framework to build, customize and deploy generative AI models

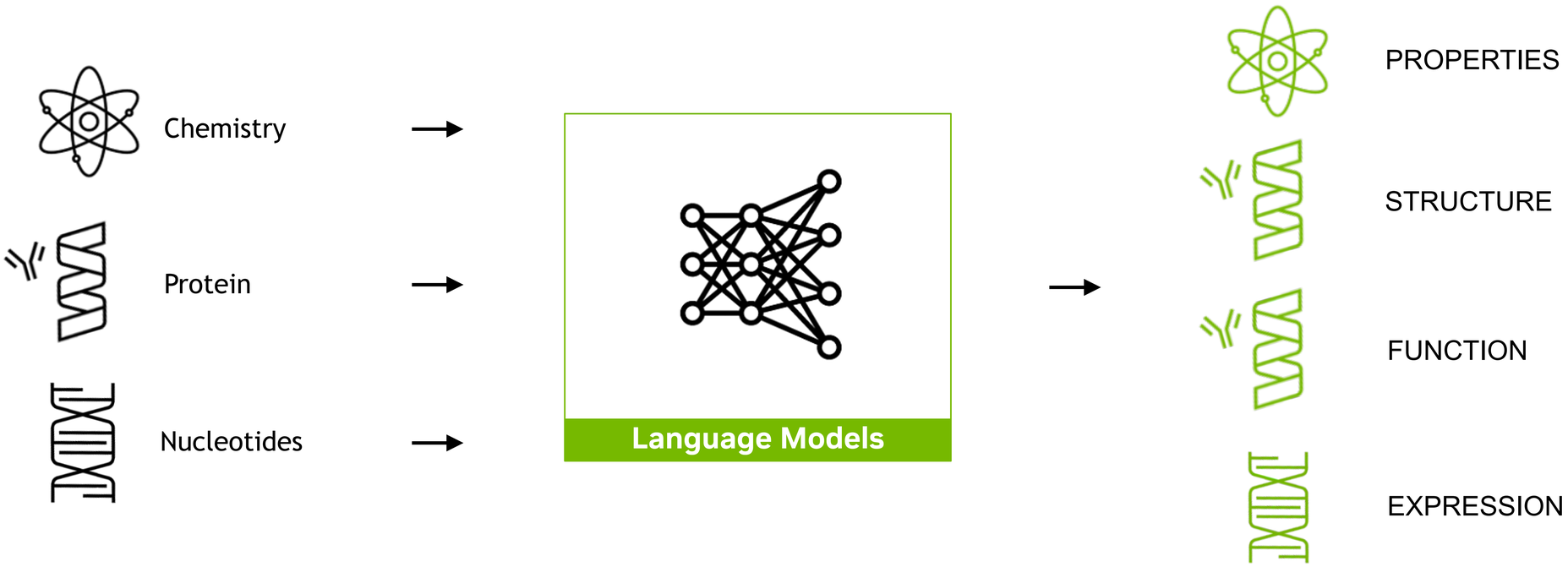


Microsoft Azure | aws | ORACLE | NVIDIA DGX Cloud | DELL Technologies | Hewlett Packard Enterprise | Lenovo | SUPERMICRO

Multi-Modality	Data Curation at Scale	Optimized Training	Model Customization	Deploy at Scale	Guardrails	Support
Build language, image, generative AI models	Extract, deduplicate, filter info from large unstructured data @ scale	Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.	Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi	Run optimized inference at-scale anywhere	Keep applications aligned with safety and security requirements using NeMo Guardrails	NVIDIA AI Enterprise and experts by your side to keep projects on track

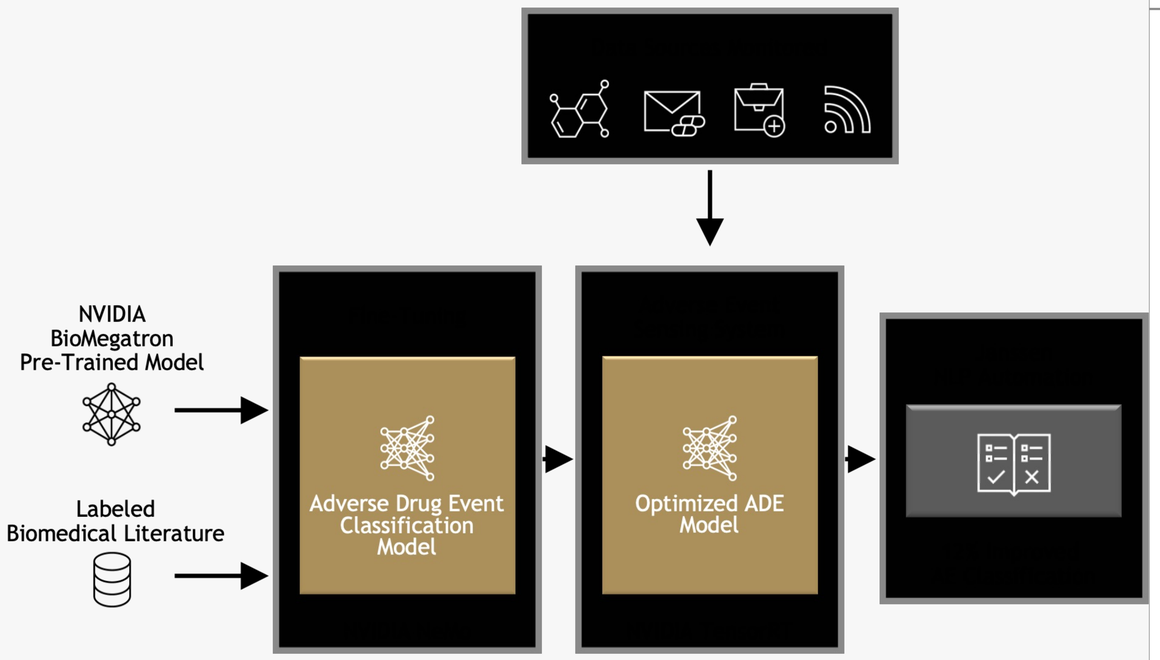
Large Language Models For Digital Biology

Predicting Biomolecular Properties to Accelerate Drug Discovery



Adverse Event Detection

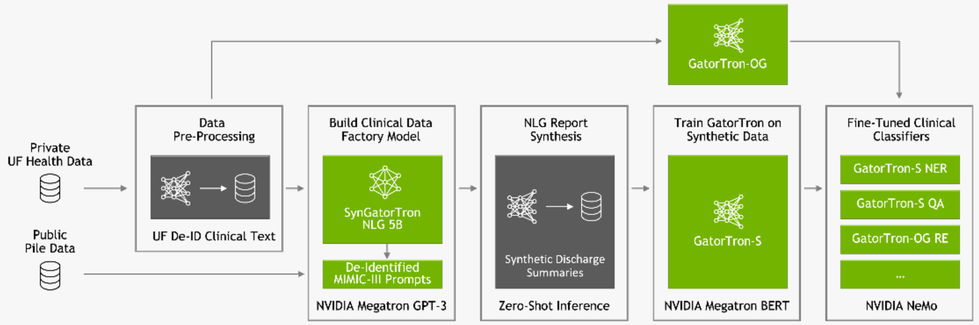
BioMegatron, NeMo, TensorRT



- Customized BioMegatron using NVIDIA NeMo
- Increases accuracy by 12%, to achieve 88% accuracy
- TensorRT improves model throughput by 2x
- Continuously monitor drug safety & efficacy
- Discover unknown side effects

Worlds Largest Clinical Generative NLP Models

GatorTron, SynGatorTron



NER Results		RE Results		QA Results	
Gatortron-S	0.8893	Gatortron-S	0.9601	Gatortron-S	72.75
Gatortron-OG	0.8859	Gatortron-OG	0.9599	Gatortron-OG	71.81

- Clinical Language Generation Model: SynGatorTron NLG 5B
- Used to generate synthetic, de-identified clinical notes & reports
- Trained GatorTron-S with Synthetic Data
- GatorTron-S achieves equivalent SOTA as original GatorTron-OG

SAFETY FOCUSED LARGE LANGUAGE MODEL FOR HEALTHCARE

The World Health Organization predicts a gap of 10 million healthcare workers by 2030- attributed to the current state of healthcare service. To address this issue, Hippocratic AI has developed the industry's first large language model using generative AI, focusing on non-diagnostic, patient-facing applications.

They achieve high performance by utilizing NVIDIA's NEMO framework with A100 GPUs. The LLM is already in use across 10 hospitals in the US and Canada, with plans to further develop its role in patient assistance and hospitals.

NVIDIA.

Hippocratic AI
Do No Harm

www.NVIDIA.com/Inception

NVIDIA BioNeMo

Cloud Service for Customizing and Running Generative AI in Drug Discovery

BioNeMo Service

WEB
or
API

BioNeMo Models

AlphaFold2

OpenFold

ESMFold

ESM-2nv

ESM-2

ProtGPT2

MegaMolBART

MoFlow

DiffDock

Drug Discovery Applications

Protein Structure Prediction

Biomolecule Representation

Biomolecule Generation

Molecular Docking

NVIDIA DGX
Cloud

Customizable SOTA Generative AI

Innovate faster and more competitively using proprietary data sets to train and fine-tune drug discovery workflows

Easy and Instant Access to Optimized AI

Eliminate the need for building IT infrastructure, managing open-source software, optimize for throughput

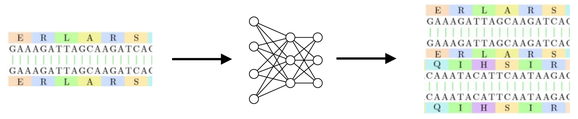
Seamless and Scalable AI Service

Ultimate flexibility in experimenting and building enterprise grade generative AI workflows with GUI & API Endpoints on scalable managed infrastructure

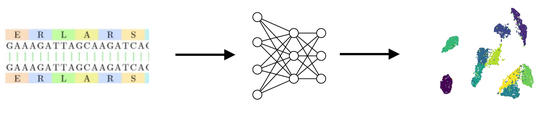
BioNeMo offers 9 Pre-trained Models Optimized for Drug Discovery

Accessible via either web-based UI or API

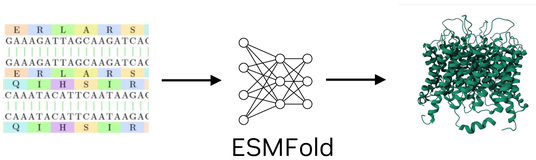
Protein Generation



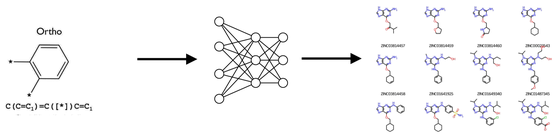
Protein Learned Sequence & Structure



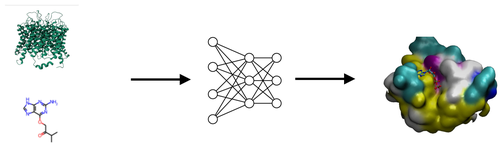
Protein Structure Prediction



Molecular Generation



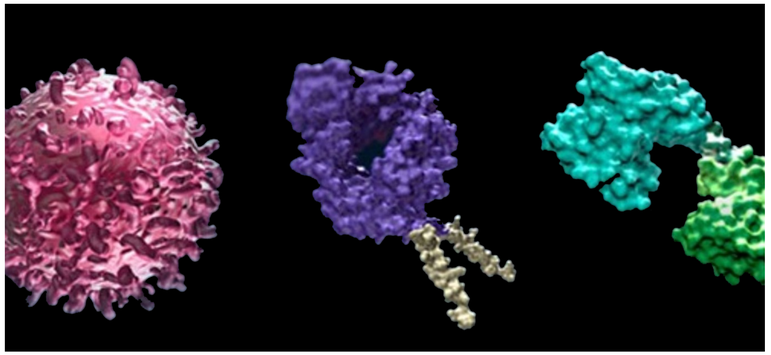
Molecular Docking



NVIDIA DGX Cloud

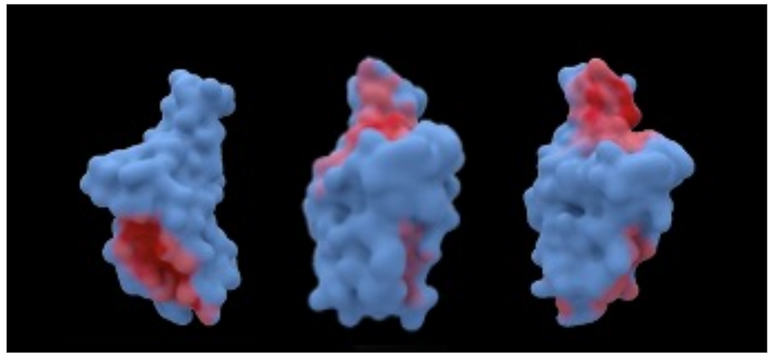
Leaders in Drug Discovery & Development use NVIDIA

Leading pharmaceutical companies, startups, and medical centers



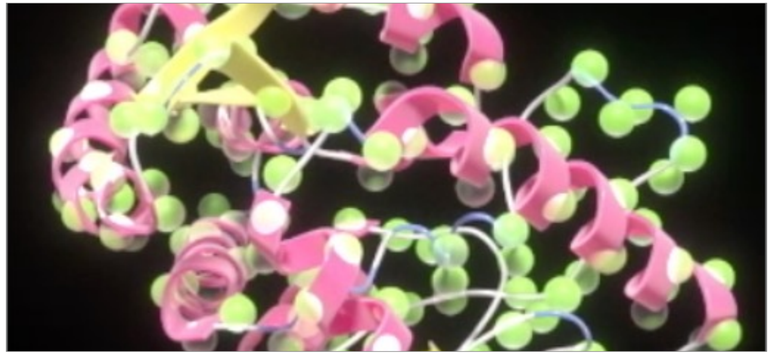
Amgen

5-100x Faster Structure Prediction with BioNeMo
15-100x faster training with proprietary data with BioNeMo on DGX Cloud



Deloitte

Intelligent Structural Biology Pipeline integrating BioNeMo

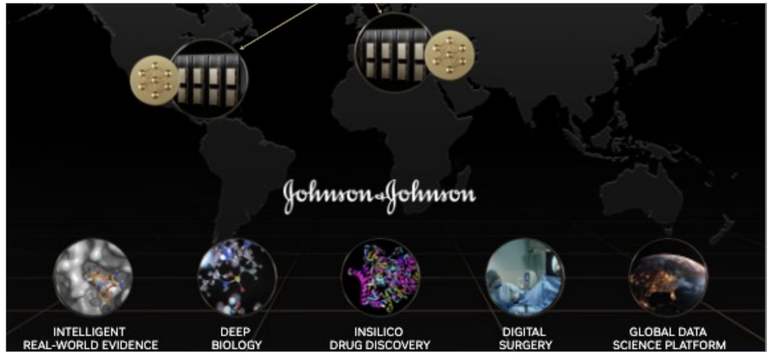


Evozyne

167 Mutations Generated in 1 wk. using BioNeMo
Generated novel proteins for rare disease & climate change

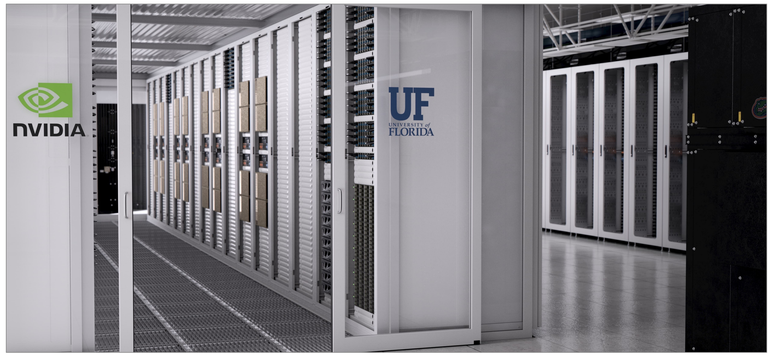
Recursion Pharmaceuticals

AI powered phenotypic screening at scale
DGX SuperPod, BaseCommand



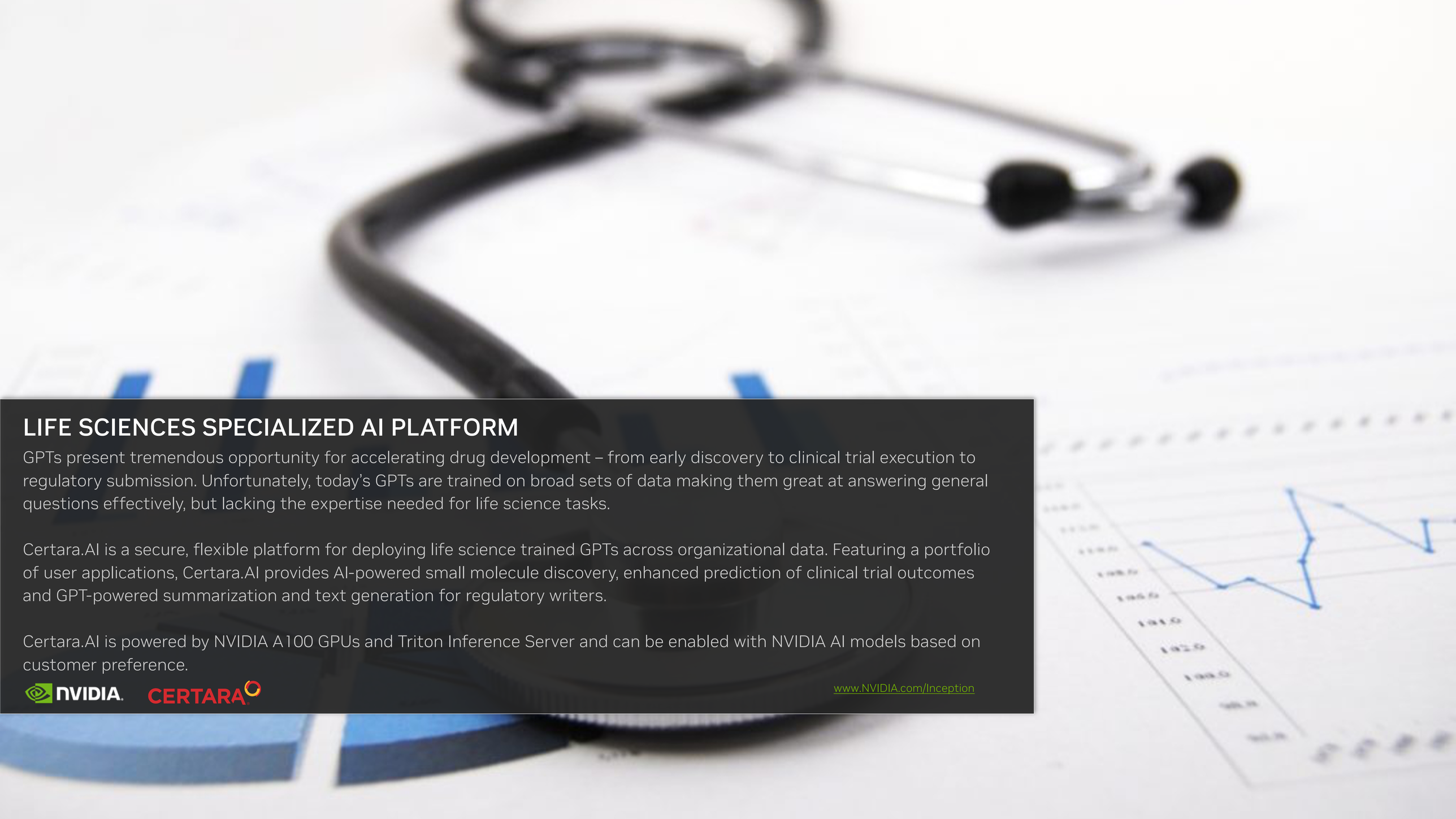
Janssen

Adverse Event Monitoring, Histopathology, CompChem, +
DGX SuperPod, BaseCommand



Univ. Florida Health System

Clinical Language Models, Synthetic Clinical Data Generation
DGX SuperPod



LIFE SCIENCES SPECIALIZED AI PLATFORM

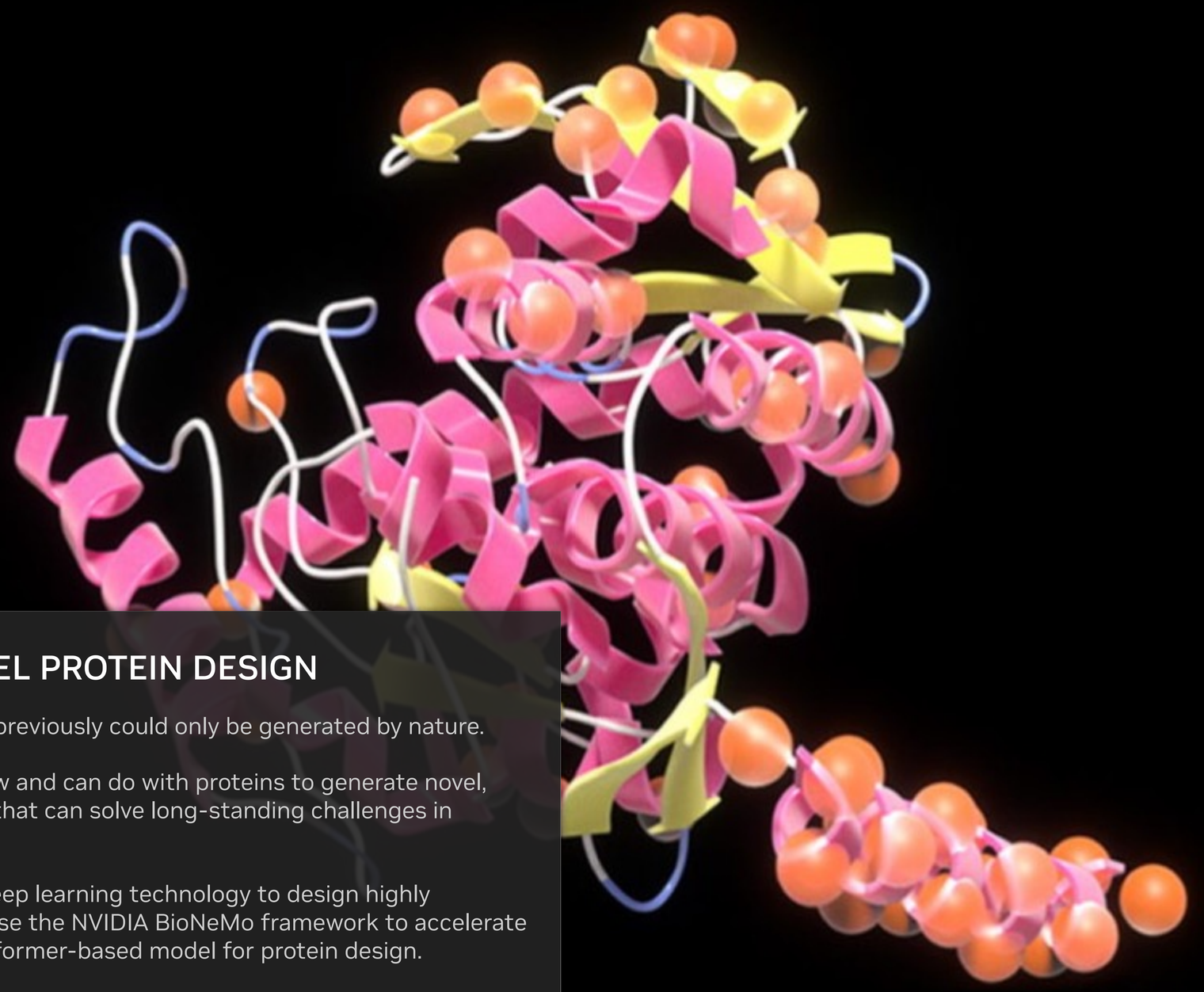
GPTs present tremendous opportunity for accelerating drug development – from early discovery to clinical trial execution to regulatory submission. Unfortunately, today's GPTs are trained on broad sets of data making them great at answering general questions effectively, but lacking the expertise needed for life science tasks.

Certara.AI is a secure, flexible platform for deploying life science trained GPTs across organizational data. Featuring a portfolio of user applications, Certara.AI provides AI-powered small molecule discovery, enhanced prediction of clinical trial outcomes and GPT-powered summarization and text generation for regulatory writers.

Certara.AI is powered by NVIDIA A100 GPUs and Triton Inference Server and can be enabled with NVIDIA AI models based on customer preference.

NVIDIA. CERTARA

www.NVIDIA.com/Inception



GENERATIVE AI FOR NOVEL PROTEIN DESIGN

Proteins make living things work and previously could only be generated by nature.

Evozyne is transforming what we know and can do with proteins to generate novel, adaptive, high-performance proteins that can solve long-standing challenges in therapeutics and sustainability.

Evozyne combines engineering and deep learning technology to design highly functional, synthetic proteins. They use the NVIDIA BioNeMo framework to accelerate the development of ProT-VAE, a transformer-based model for protein design.

NVIDIA EVOZYNE NATURAL MACHINES

www.NVIDIA.com/Inception

"Ease of multi-node training and the ability to use larger batch sizes within DGX Cloud enabled us to achieve our three-month objectives in just four weeks.."

- Chris James Langmead, Director of Digital Biologics Discovery, Amgen

AMGEN

Generative AI Speeds Biologics Drug Discovery

Challenge

Traditional biologics discovery is a costly process, and sparse data make predictions even more challenging.

Amgen wanted to accelerate biologics discovery by using AI models to propose and evaluate designs for candidate drugs.

Required powerful multi-node infrastructure to accelerate training of large protein models with extensive data.

Solution

Trained large language models (LLMs) on Amgen's proprietary data to help predict properties of proteins and develop biologics with enhanced properties.

Leveraged NVIDIA DGX Cloud and BioNeMo for training and fine-tuning of protein LLMs and NVIDIA RAPIDS for faster post-training analysis.

BioNemo on DGX Cloud, a turnkey solution enabled Amgen to get up and running quickly, moving from initial login to training large models in just a few days.

NVIDIA DGX Cloud
AI-training-as-a-service solution

NVIDIA Base Command Platform
for workflow management

NVIDIA AI Enterprise
RAPIDS for data post-processing

NVIDIA BioNeMo
For training and inferencing

20 sec/structure — Faster protein structure prediction

100x — Faster post-training analysis

< 1 month — From onboarding to first pretrained protein LLM

NVIDIA

> *Our process for setting up training jobs went from the hassle of manually pushing code to remote machines and ensuring synchronization to the simplicity of pressing "run" on DGX Cloud.*

— Edward Williams, Machine Learning Engineer, Terray Therapeutics

Need approval

TERRAY

Drug discovery with generative AI delivers the power of hundreds of researchers and infinite possibilities

Challenges

Small molecule drug discovery explores an infinite chemical space, but traditional methods can test only a few dozen to a few hundred compounds per week.

Terray Therapeutics wanted leverage foundation models of chemistry and generative AI to design small molecules using their massive and precise experimental data.

Terray needed a platform that enabled rapid experimentation to support their iterative approach to optimizing these molecules.

Solutions

Terray developed COATI, a pre-trained foundation model for chemistry, trained on a dataset of hundreds of millions of small molecules.

NVIDIA DGX Cloud significantly improved the COATI development process and enabled more efficient experimentation with dedicated multi-node GPUs and on-demand resource scaling.

NVIDIA Base Command Platform enabled Terray to improve infrastructure utilization, with ease of configuring workloads and optimizing models with push-button simplicity.

NVIDIA DGX Cloud
AI-training-as-a-service platform

NVIDIA Base Command Platform
for workflow management

NVIDIA AI Enterprise
PyTorch, CUDA

>4X improved infrastructure utilization over alternate cloud service

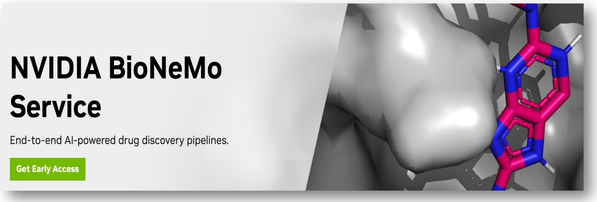
< 1 day to onboard onto DGX Cloud

Reduced training from
a week to a day

NVIDIA

Help and Resources

[NVIDIA Drug Discovery Webpage]



[NVIDIA BioNeMo Webpage with Early Access Form]

[NVIDIA Generative Chemistry LaunchPad Experience]

[Training with BioNeMo on DGX Cloud]

Q&A

Apply to NVIDIA Inception for startups:

NVIDIA.com/startups

Join the NVIDIA Developer community to get access to technical training, technology, AI models and 600+ SDKs:

Developer.nvidia.com/join

Explore More Gen AI/LLM Training:

25% off Workshops for LLM Day registrants*

USE CODE: **TRAIN-LLM**

Instructor-Led Workshops

> Generative AI With Diffusion Models

> Rapid Application Development Using LLMs

> Efficient Large Language Model (LLM) Customizations

Self-Paced Courses

> Generative AI Explained (Free)

> Generative AI With Diffusion Models

View our comprehensive Gen AI/LLM learning path, covering fundamental to advanced topics

> Gen AI/LLM Learning Path

**Offer valid for any of the DLI public workshops scheduled through March 01, 2024.*

NVIDIA