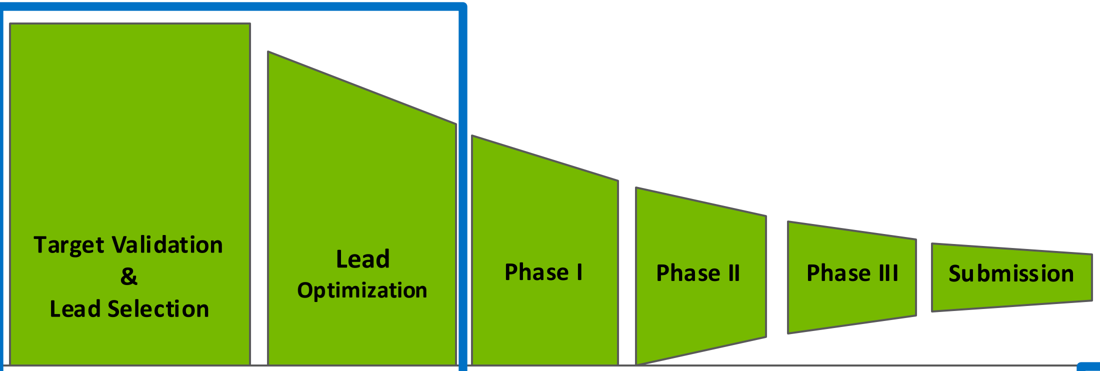
NVIDIA.

Scientific Discovery: From the Lab Bench to the GPU

Michelle L. Gill, PhD; Tech Lead and R&D Manager, NVIDIA

PyDataNYC | 3rd November, 2023

Motivation: Drug Development is a Long and Expensive Process



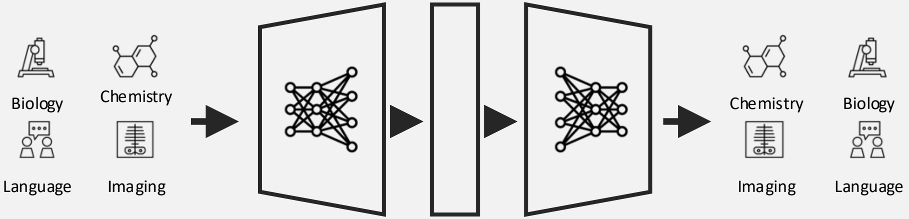
	Target Validation & Lead Selection	Lead Optimization	Phase I	Phase II	Phase III	Submission	Total
Cost **Time**	$1.1B 4.5Yr	$230M 1.0Yr	$225M 1.5Yr	$500M 2.5Yr	$500M 2.5Yr	$75M 1.5Yr	**$2.8B** **>10Yr**

$2.8B and >10 Years to Bring a Drug to Market

Original source: *Developability assessment as an early de-risking tool for biopharmaceutical development*, J. Zurdo, 2013, DOI: 10.4155/pbp.13.3

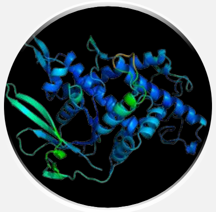
Language Models are Revolutionizing Discovery

- Information from biomedical literature
- Prediction of chemical reactions
- Biomolecular property prediction
- Structure prediction and docking

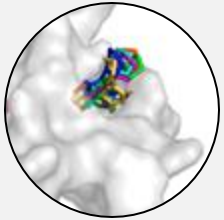


BIOMEDICAL NLP
Learn all of PubMed

GENERATIVE CHEMISTRY
Novel Drug Candidates

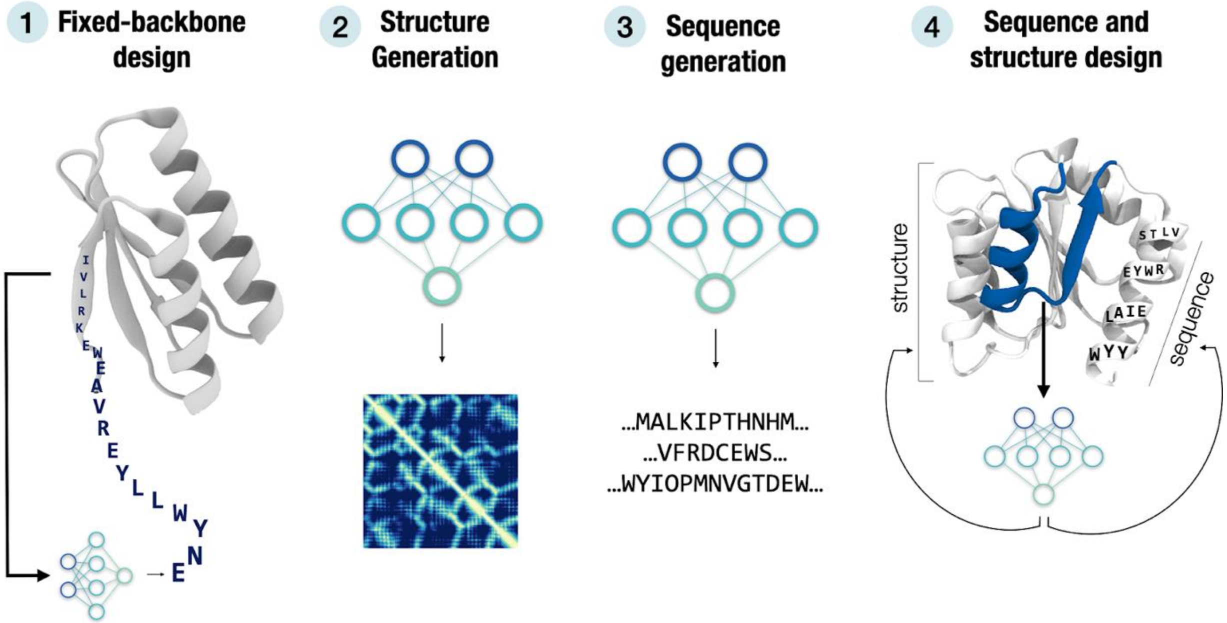


PROTEIN STRUCTURE
Predict 3D Structures



VIRTUAL SCREENING
Docking and Pose Prediction

From Sequence to 3D and Back Again



Ferruz, N. *et al.* *bioRxiv* 2022.08.31.505981 (2022)

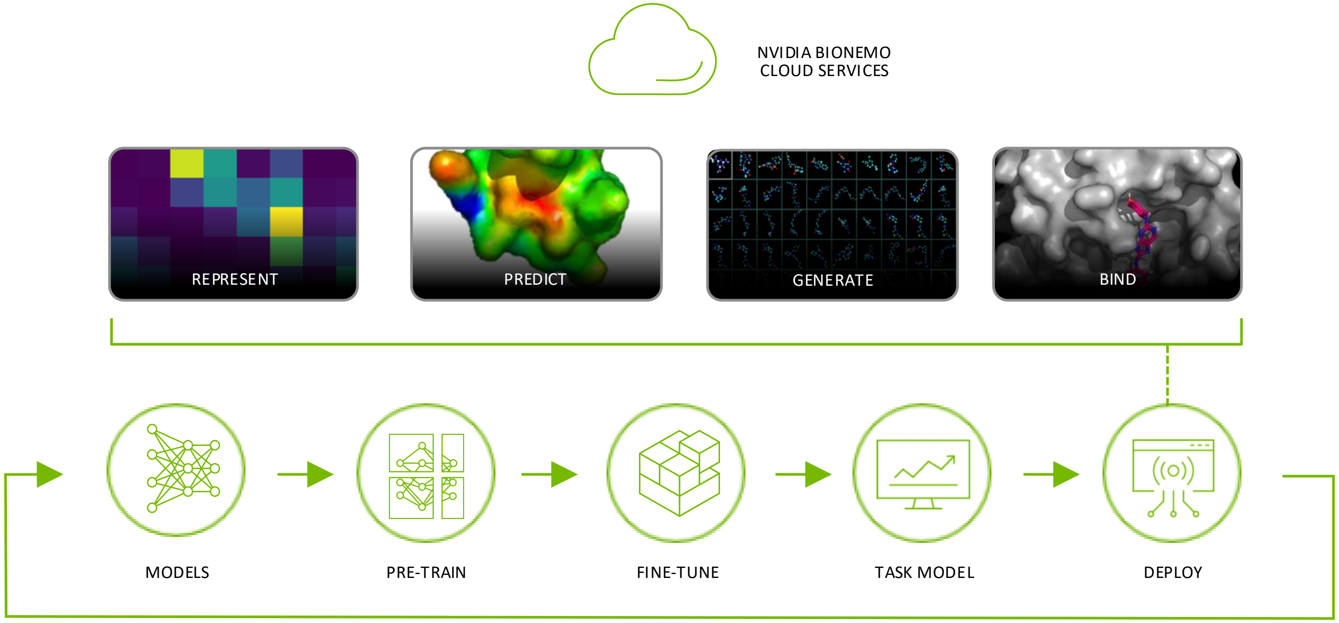
Outline

- Overview of BioNeMo: Inference Service and Training Framework
- MolMIM: Development of a Small Molecule Foundation Model for Generation
- Career Progression and Lessons from the Field

BioNeMo Overview: Inference Service and Framework

NVIDIA BioNeMo

AI Tools, Frameworks, and Applications for Drug Discovery



NVIDIA BIONEMO
CLOUD SERVICES

REPRESENT

PREDICT

GENERATE

BIND

MODELS

PRE-TRAIN

FINE-TUNE

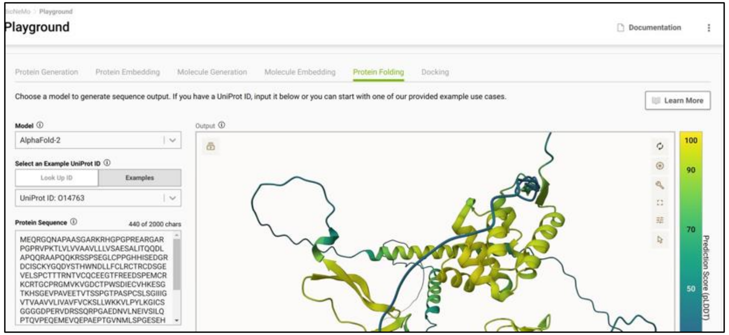
TASK MODEL

DEPLOY

NVIDIA BIONEMO
FRAMEWORK

Multiple Interfaces to a BioNeMo Model in the Inference Service

Interactive UI and Jupyter Workflows



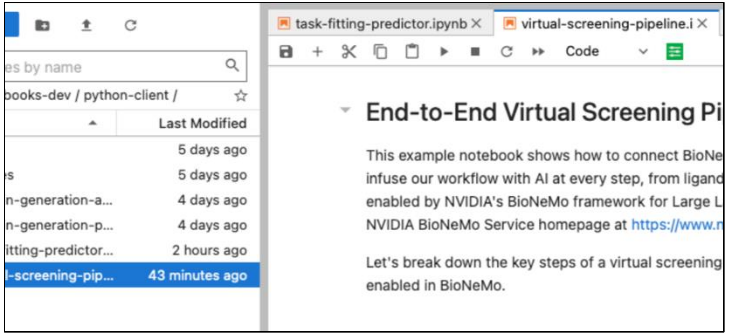
Choose a model to generate sequence output. If you have a UniProt ID, input it below or you can start with one of our provided example use cases.

Model: AlphaFold-2

Select an Example UniProt ID: UniProt ID: O14763

Protein Sequence — 440 of 2000 chars

```
MEQRGQNAPAASGARKRHGPGPREARGAR
PGPRVPKTLVLVVAAVLLLVSAESALITQQDL
APQQRAAPQQKRSSPSEGLCPPGHHISEDGR
DCISCKYGQDYSTHWNDLLFCLRCTRCDSGE
VELSPCTTTRNTVCQCEEGTFREEDSPEMCR
KCRTGCPRGMVKVGDCTPWSDIECVHKESG
TKHSGEVPAVEETVTSSPGTPASPCSLSGIIIG
VTVAAVVLIVAVFVCKSLLWKKVLPYLKGICS
GGGGDPERVDRSSQRPGAEDNVLNEIVSILQ
PTQVPEQEMEVQEPAEPTGVNMLSPGESEH
```



es by name
books-dev / python-client /
Last Modified
task-fitting-predictor.ipynb
virtual-screening-pipeline.i
Code

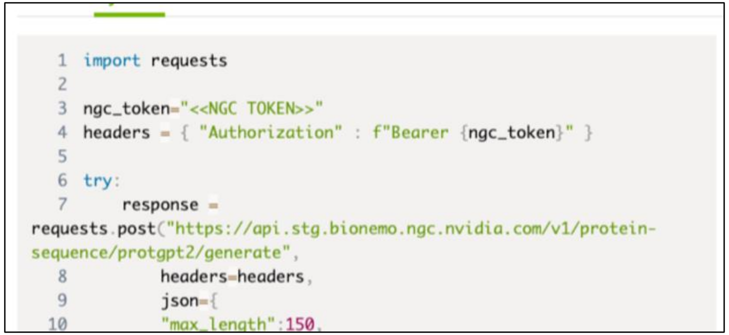
	Last Modified
	5 days ago
s	5 days ago
n-generation-a...	4 days ago
n-generation-p...	4 days ago
itting-predictor...	2 hours ago
l-screening-pip...	43 minutes ago

End-to-End Virtual Screening Pi

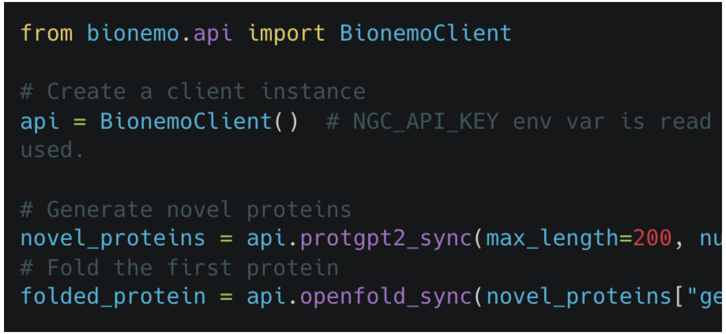
This example notebook shows how to connect BioNe
infuse our workflow with AI at every step, from ligand
enabled by NVIDIA's BioNeMo framework for Large L
NVIDIA BioNeMo Service homepage at https://www.n

Let's break down the key steps of a virtual screening
enabled in BioNeMo.

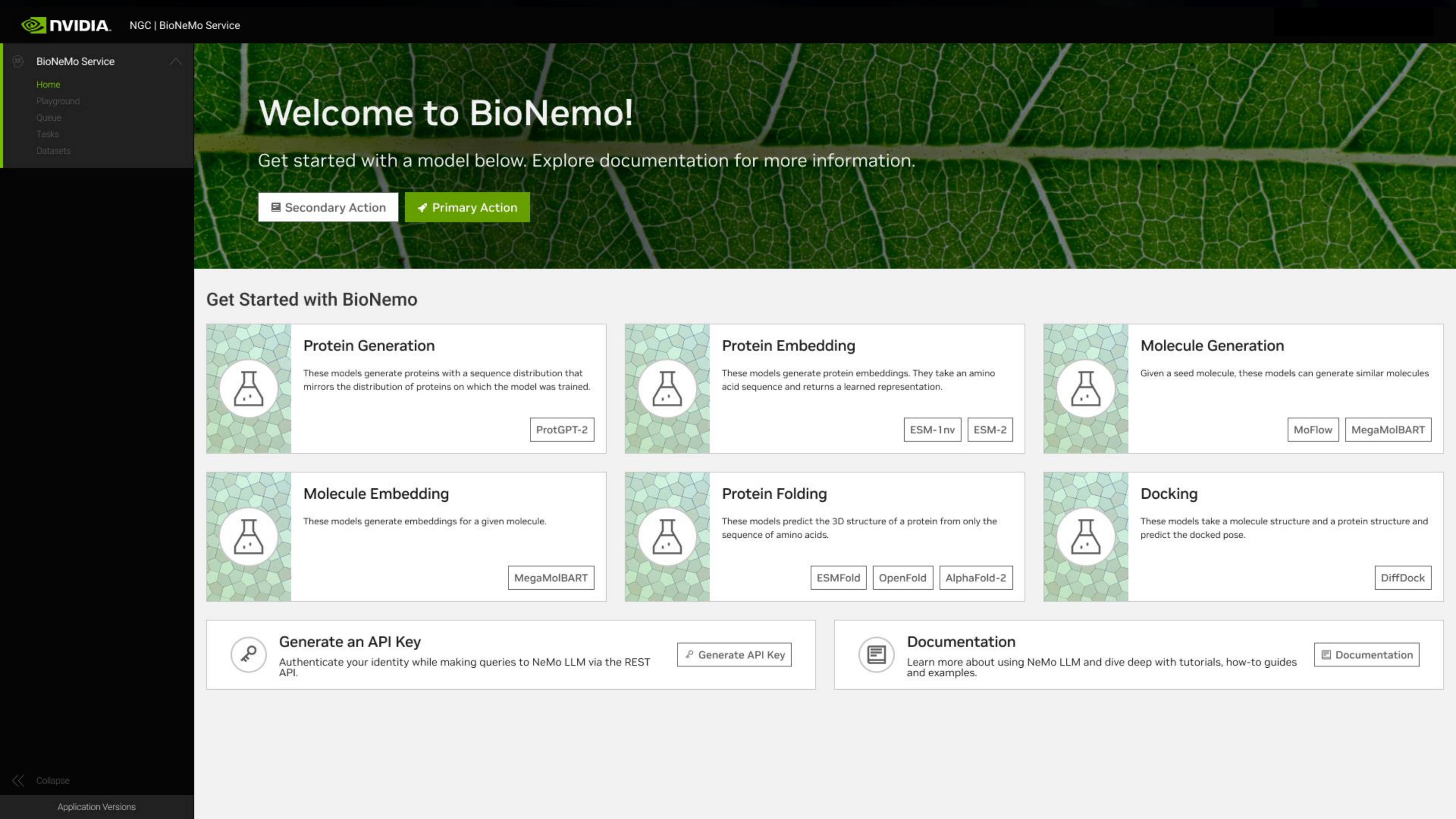
API and Python Client



```
import requests

ngc_token="<<NGC TOKEN>>"
headers = { "Authorization" : f"Bearer {ngc_token}" }

try:
      response =
requests.post("https://api.stg.bionemo.ngc.nvidia.com/v1/protein-
sequence/protgpt2/generate",
            headers=headers,
            json={
            "max_length":150,
```



```
from bionemo.api import BionemoClient

# Create a client instance
api = BionemoClient()  # NGC_API_KEY env var is read
used.

# Generate novel proteins
novel_proteins = api.protgpt2_sync(max_length=200, nu
# Fold the first protein
folded_protein = api.openfold_sync(novel_proteins["ge
```



Welcome to BioNemo!

Get started with a model below. Explore documentation for more information.

Get Started with BioNemo

Protein Generation

These models generate proteins with a sequence distribution that mirrors the distribution of proteins on which the model was trained.

Protein Embedding

These models generate protein embeddings. They take an amino acid sequence and returns a learned representation.

Molecule Generation

Given a seed molecule, these models can generate similar molecules

Molecule Embedding

These models generate embeddings for a given molecule.

Protein Folding

These models predict the 3D structure of a protein from only the sequence of amino acids.

Docking

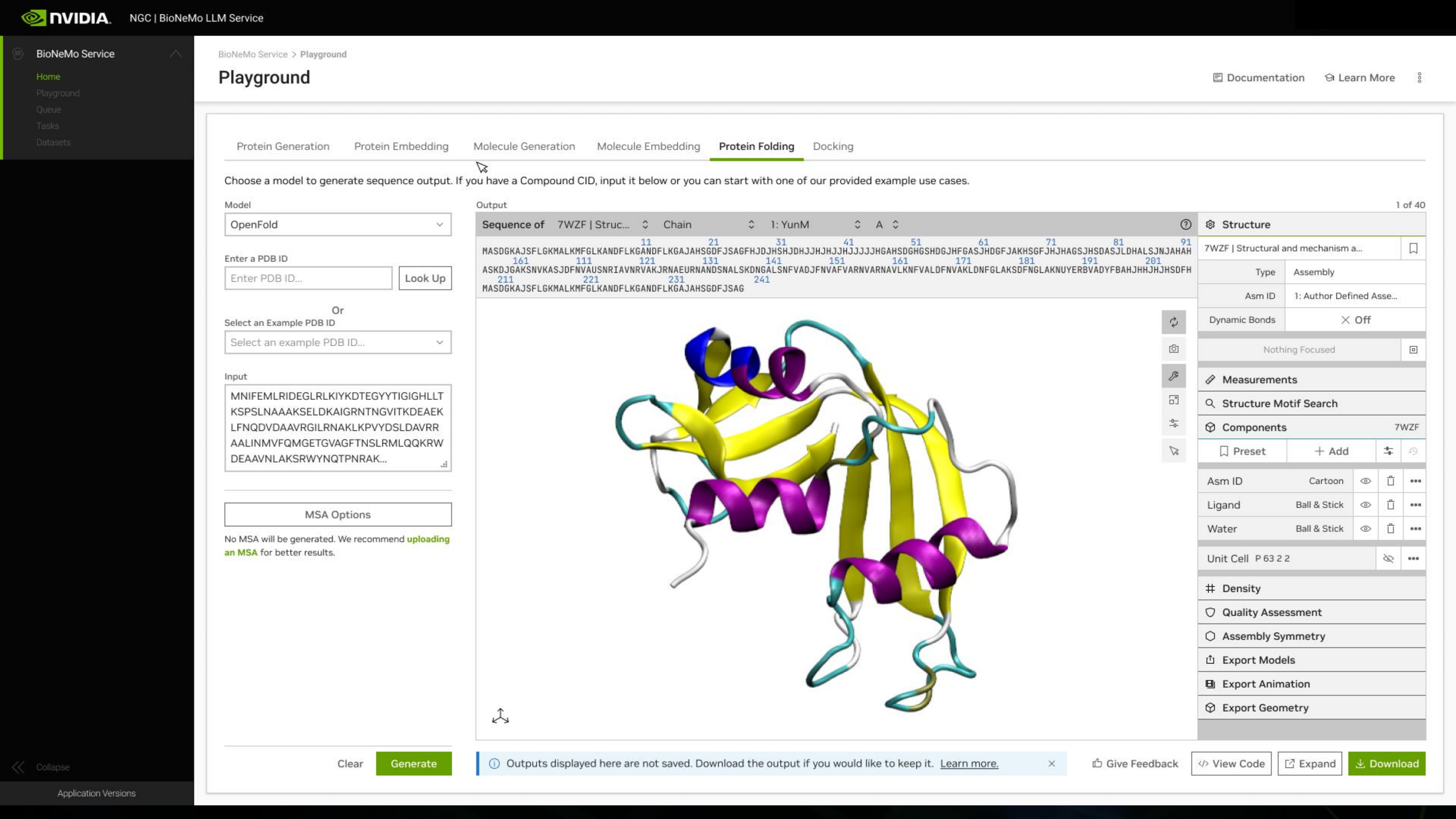
These models take a molecule structure and a protein structure and predict the docked pose.

Generate an API Key

Authenticate your identity while making queries to NeMo LLM via the REST API.

Documentation

Learn more about using NeMo LLM and dive deep with tutorials, how-to guides and examples.



BioNeMo Service > Playground

Playground

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

Model: OpenFold

Enter a PDB ID: Enter PDB ID... [Look Up]

Or

Select an Example PDB ID: Select an example PDB ID...

Input:

```
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLT
KSPSLNAAAKSELDKAIGRNTNGVITKDEAEK
LFNQDVDAAVRGILRNAKLKPVYDSLDAVRR
AALINMVFQMGETGVAGFTNSLRMLQQKRW
DEAAVNLAKSRWYNQTPNRAK...
```

No MSA will be generated. We recommend uploading an MSA for better results.

Output

1 of 40

Sequence of 7WZF | Struc... Chain 1: YunM A

```
                              11         21         31         41         51         61         71         81         91
MASDGKAJSFLGKMALKMFGLKANDFLKGANDFLKGAJAHSGDFJSAGFHJDJHSHJDHJJHJHJJHJJJJJHGAHSDGHGSHDGJHFGASJHDGFJAKHSGFJHJHAGSJHSDASJLDHALSJNJAHAH
     161        111        121        131        141        151        161        171        181        191        201
ASKDJGAKSNVKASJDFNVAUSNRIAVNRVAKJRNAEURNANDSNALSKDNGALSNFVADJFNVAFVARNVARNAVLKNFVALDFNVAKLDNFGLAKSDFNGLAKNUYERBVADYFBAHJHHJHJHSDFH
   211        221        231        241
MASDGKAJSFLGKMALKMFGLKANDFLKGANDFLKGAJAHSGDFJSAG
```

Structure

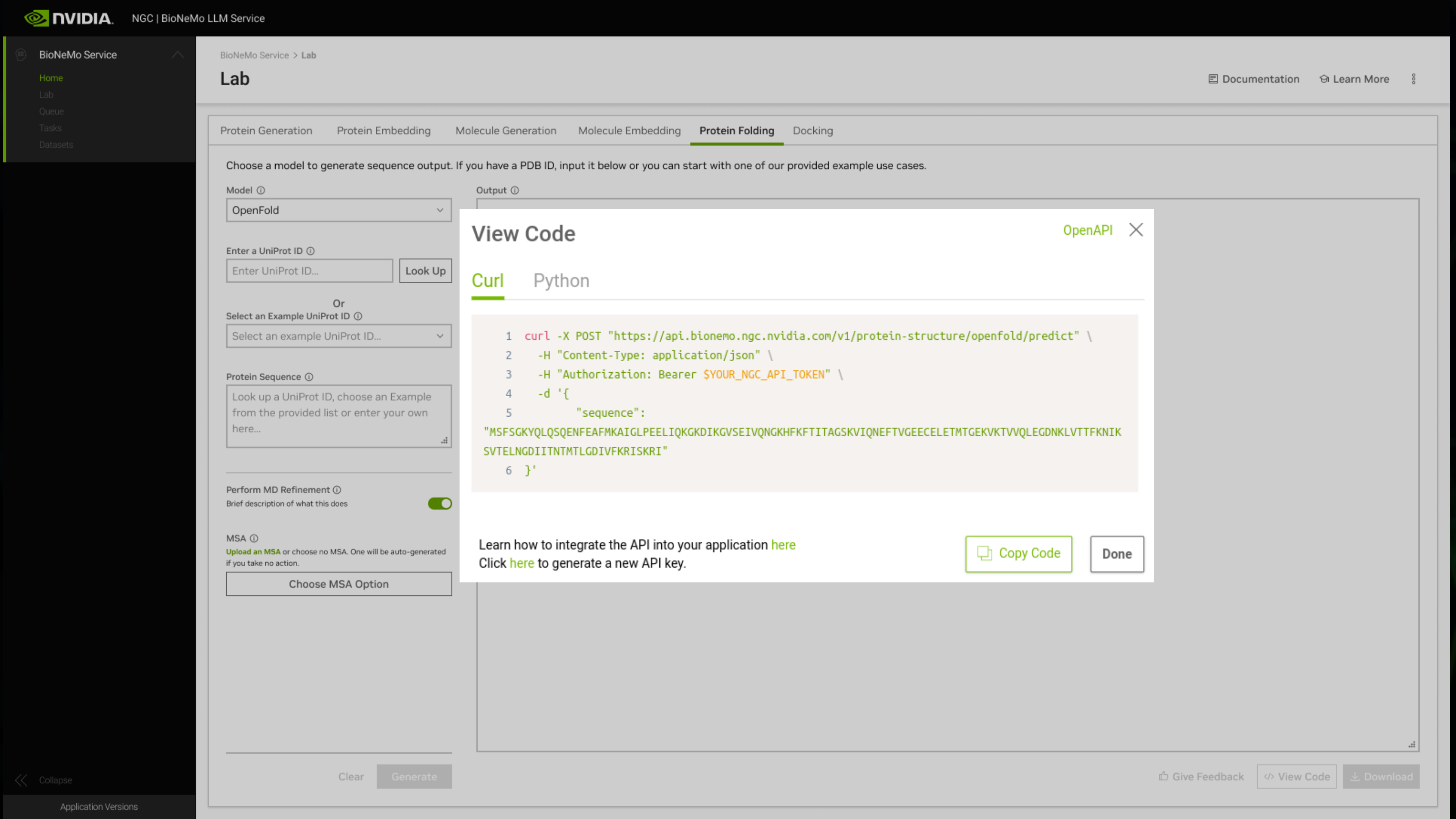
7WZF | Structural and mechanism a...

Type	Assembly
Asm ID	1: Author Defined Asse...
Dynamic Bonds	Off

Nothing Focused

Asm ID	Cartoon
Ligand	Ball & Stick
Water	Ball & Stick
Unit Cell P 63 2 2	

Outputs displayed here are not saved. Download the output if you would like to keep it. Learn more.



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...

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...

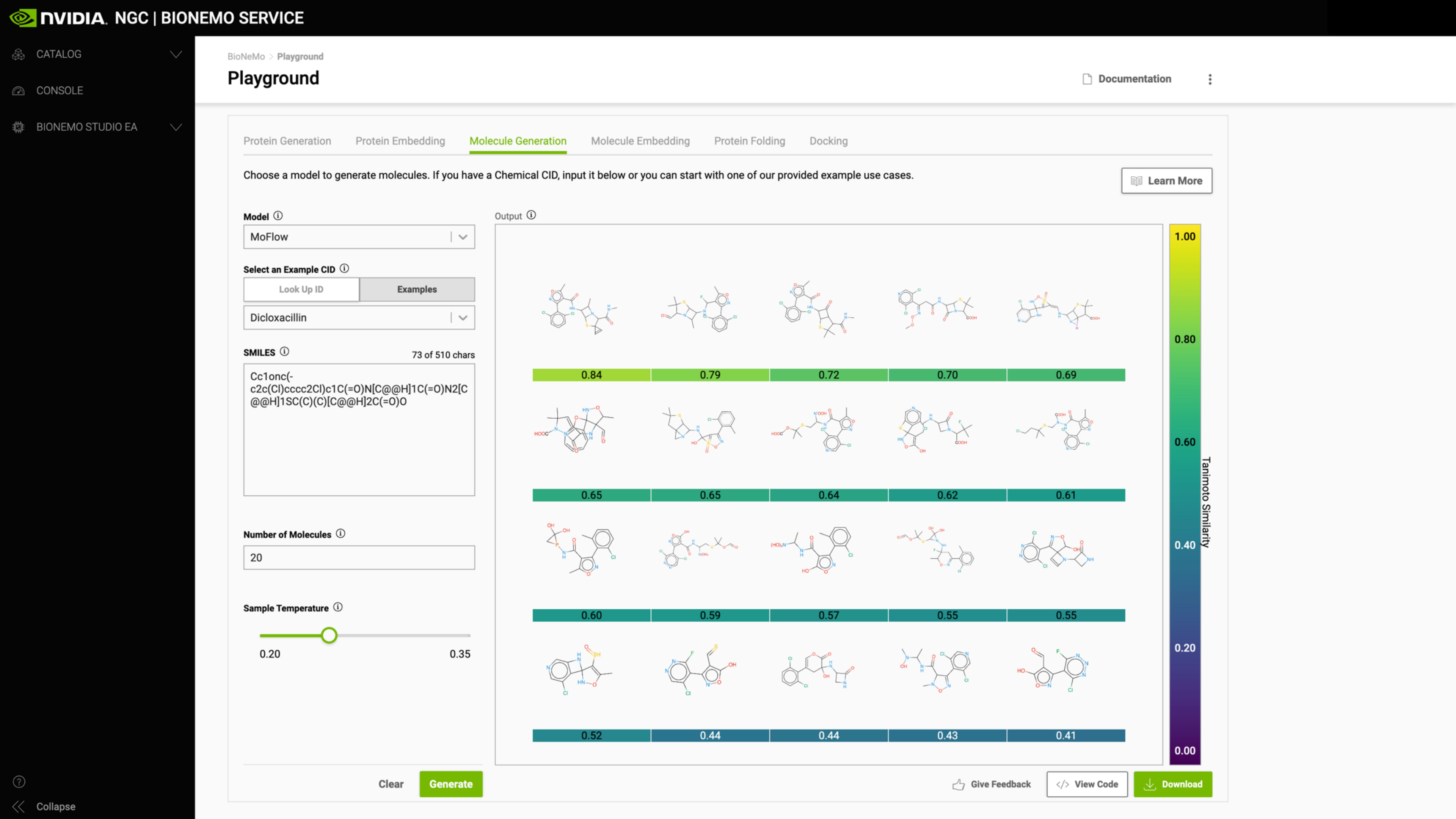
Perform MD Refinement
Brief description of what this does

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

View Code

```
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.



Playground

Choose a model to generate molecules. If you have a Chemical CID, input it below or you can start with one of our provided example use cases.

Model: MoFlow

Select an Example CID: Dicloxacillin

SMILES 73 of 510 chars

```
Cc1onc(-c2c(Cl)cccc2Cl)c1C(=O)N[C@@H]1C(=O)N2[C@@H]1SC(C)(C)[C@@H]2C(=O)O
```

Number of Molecules: 20

Sample Temperature: 0.20 – 0.35

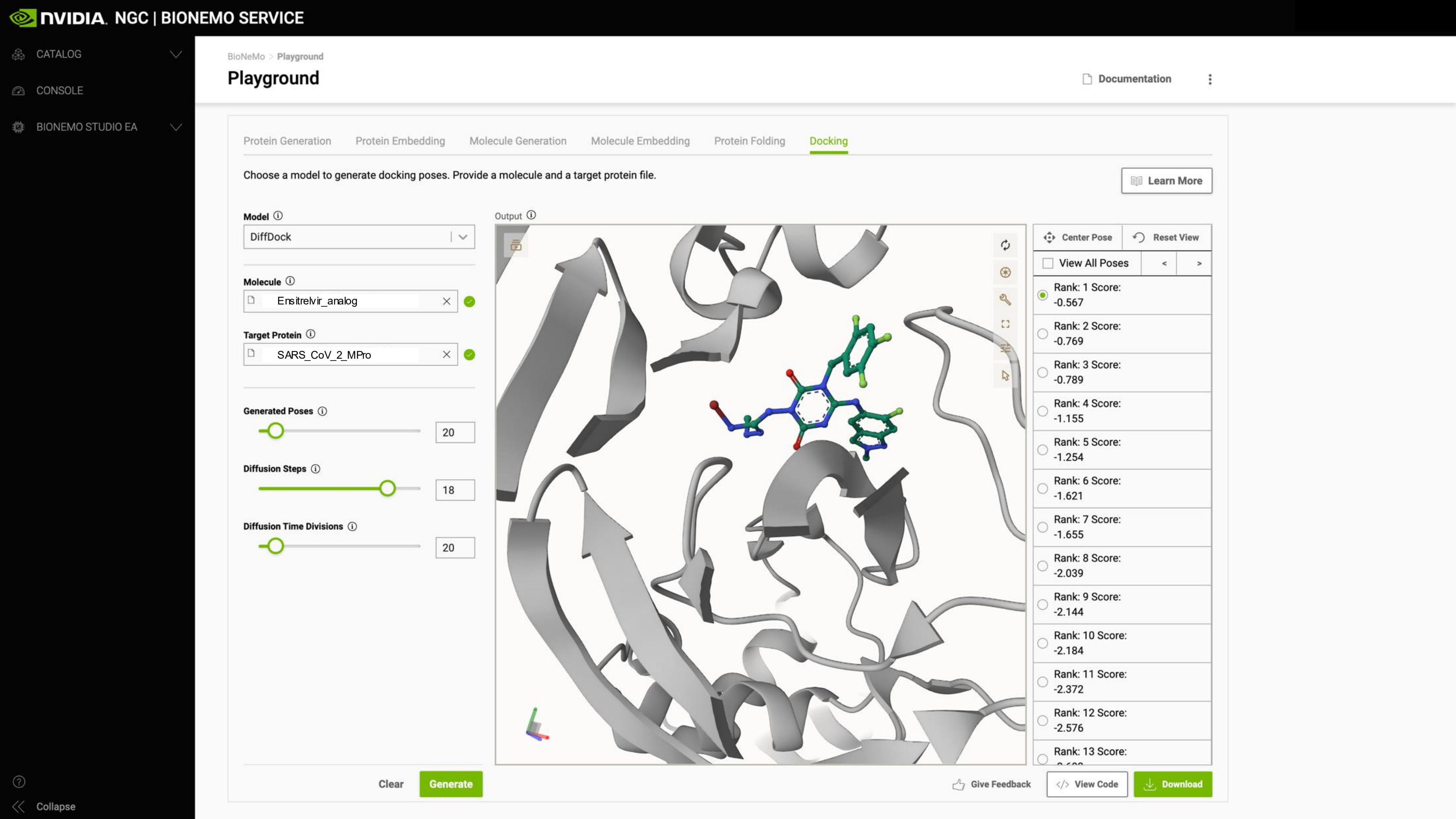
Output

0.84 0.79 0.72 0.70 0.69

0.65 0.65 0.64 0.62 0.61

0.60 0.59 0.57 0.55 0.55

0.52 0.44 0.44 0.43 0.41



Playground

Choose a model to generate docking poses. Provide a molecule and a target protein file.

Model: DiffDock

Molecule: Ensitrelvir_analog

Target Protein: SARS_CoV_2_MPro

Generated Poses: 20

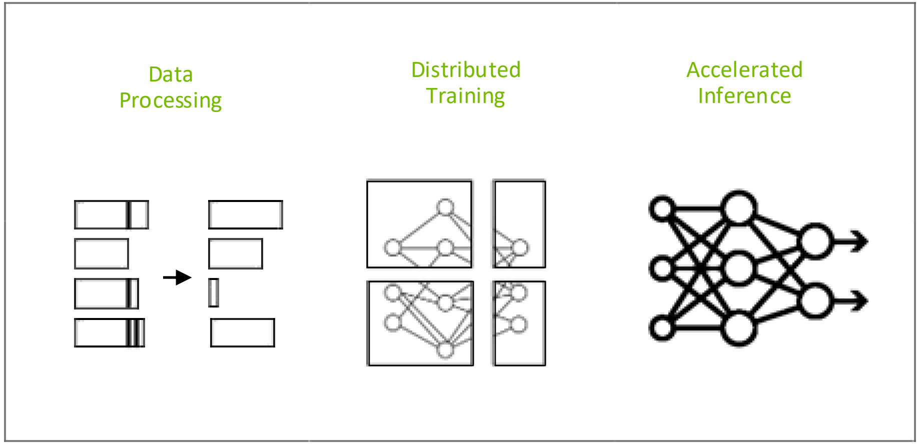
Diffusion Steps: 18

Diffusion Time Divisions: 20

□ View All Poses < >

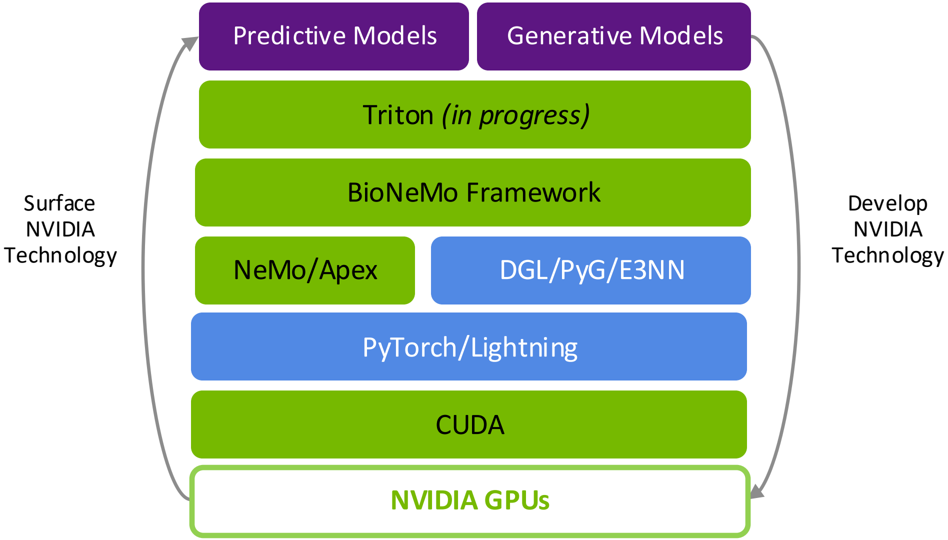
● Rank: 1 Score: -0.567
○ Rank: 2 Score: -0.769
○ Rank: 3 Score: -0.789
○ Rank: 4 Score: -1.155
○ Rank: 5 Score: -1.254
○ Rank: 6 Score: -1.621
○ Rank: 7 Score: -1.655
○ Rank: 8 Score: -2.039
○ Rank: 9 Score: -2.144
○ Rank: 10 Score: -2.184
○ Rank: 11 Score: -2.372
○ Rank: 12 Score: -2.576
○ Rank: 13 Score:

BioNeMo Framework Overview



- Includes dataset processing, training, fine tuning, and example downstream tasks
- Support for multi-GPU and multi-node training
- Data parallelism, and three types of model parallelism
- Currently three LLM models for cheminformatics and protein applications – more models and model types coming soon

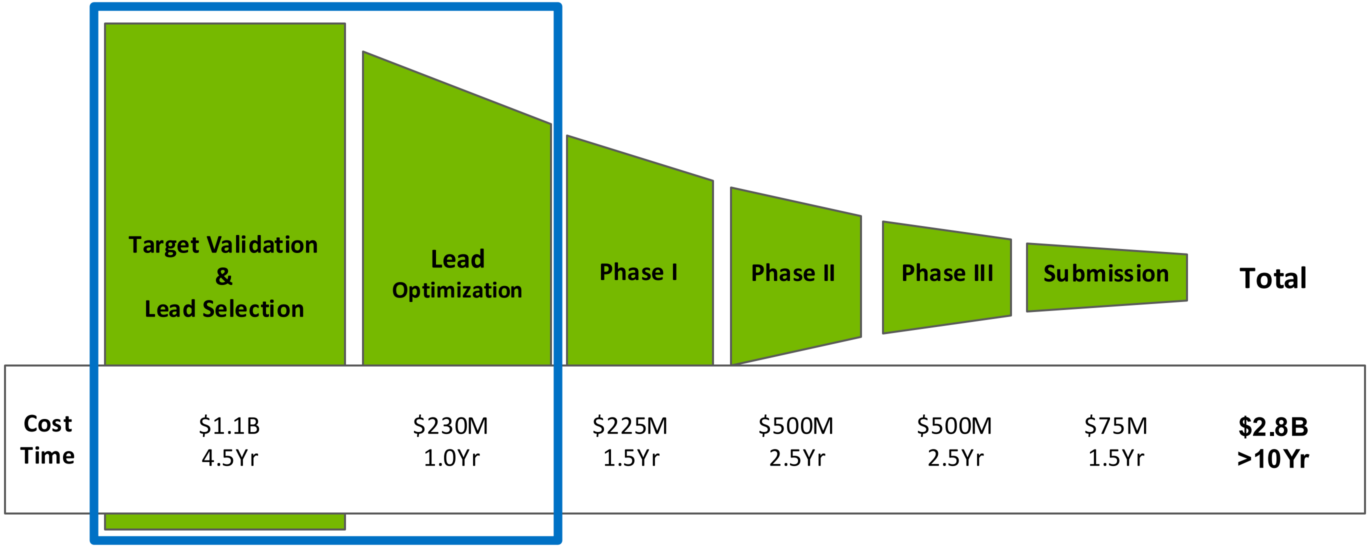
BioNeMo Framework Technology Stack



- Based on NVIDIA NeMo, which is a library for development and training of LLMs
- Automated deployment with Triton is in progress
- Surface and develop new software and hardware technology

MolMIM: Development of a Small Molecule Foundation Model for Generation

Motivation: Drug Development is a Long and Expensive Process



	Target Validation & Lead Selection	Lead Optimization	Phase I	Phase II	Phase III	Submission	Total
Cost **Time**	$1.1B 4.5Yr	$230M 1.0Yr	$225M 1.5Yr	$500M 2.5Yr	$500M 2.5Yr	$75M 1.5Yr	**$2.8B** **>10Yr**

$2.8B and >10 Years to Bring a Drug to Market

Original source: *Developability assessment as an early de-risking tool for biopharmaceutical development*, J. Zurdo, 2013, DOI: 10.4155/pbp.13.3

Lead Discovery: Three Years for Design-Make-Test-Analyze Cycle

Hit Compound

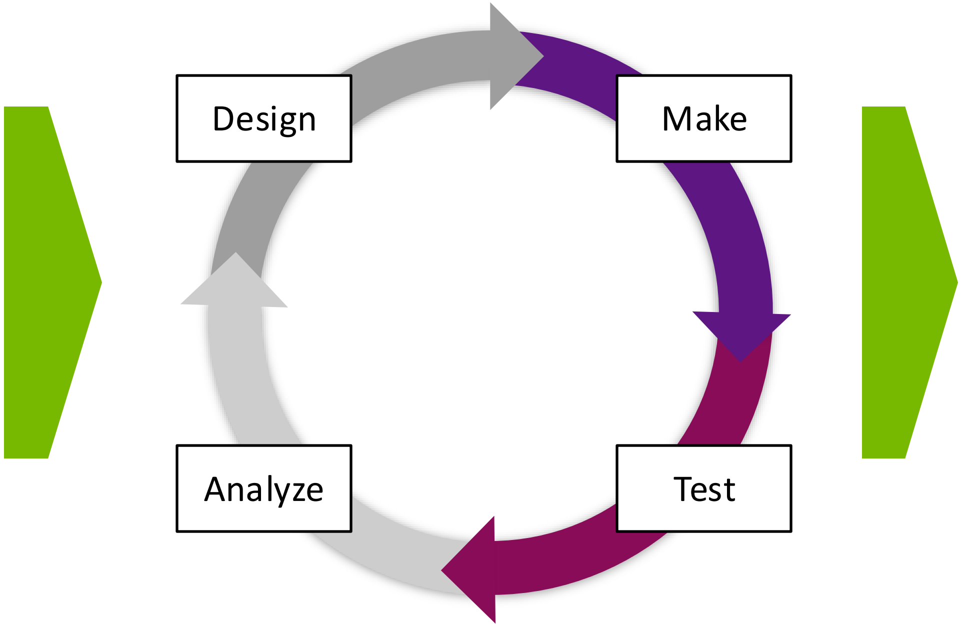
Known or experimentally determined

Weakly active

Target unselective

Toxicity risk

Low metabolic stability



Candidate Drug

Highly potent

Effective for *in vivo* models

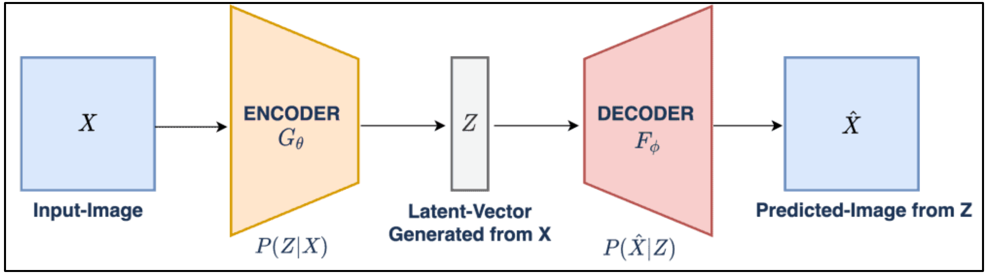
Metabolically stable

No toxicity issues

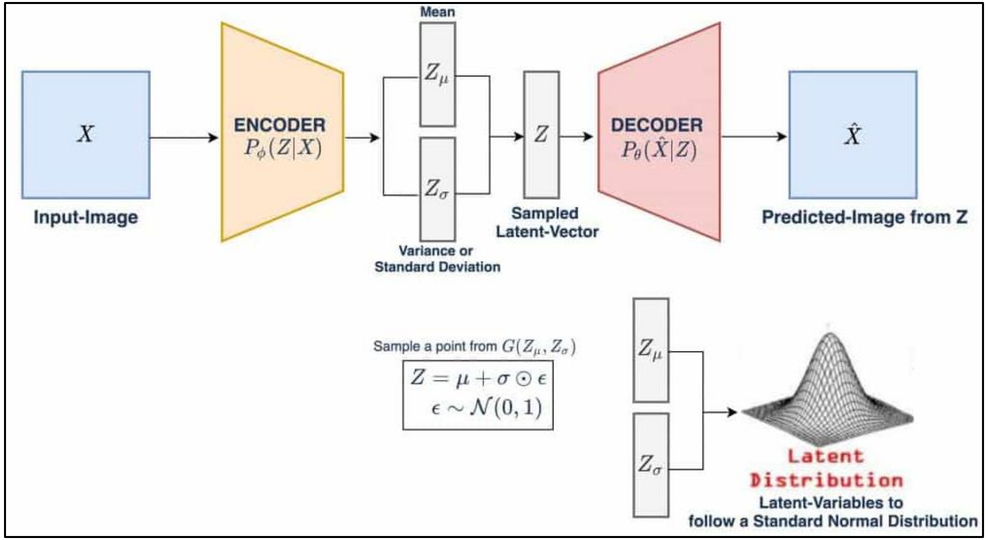
Multiple of DMTA cycles at 4-6 weeks/cycle
Transition between multiple labs

Autoencoder Models in a Nutshell

Autoencoder



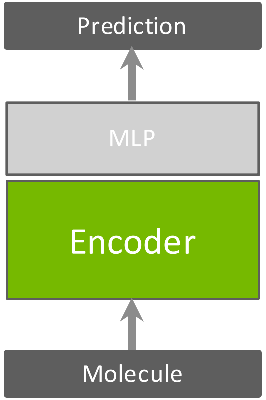
Variational Autoencoder (VAE)



Also works with sequences -- seq2seq models

Cheminformatics Foundation Model Objectives

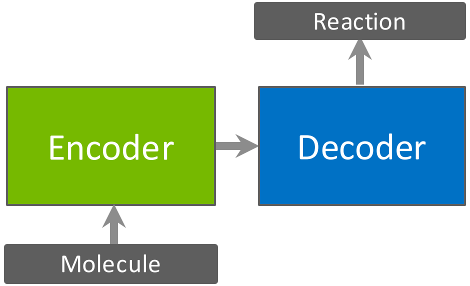
Representation and Translation


Prediction

MLP

Encoder

Molecule

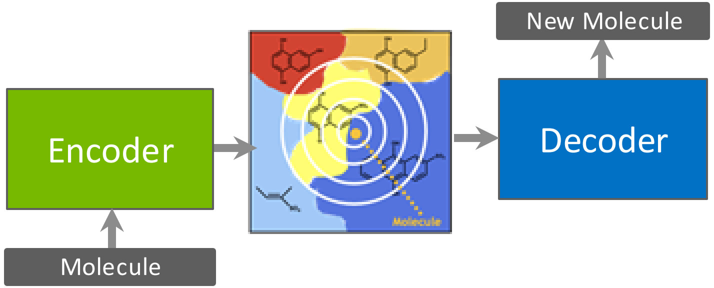

Reaction

Encoder

Decoder

Molecule

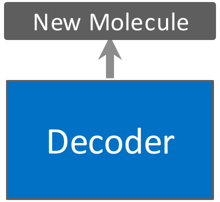
Generation


New Molecule

Encoder

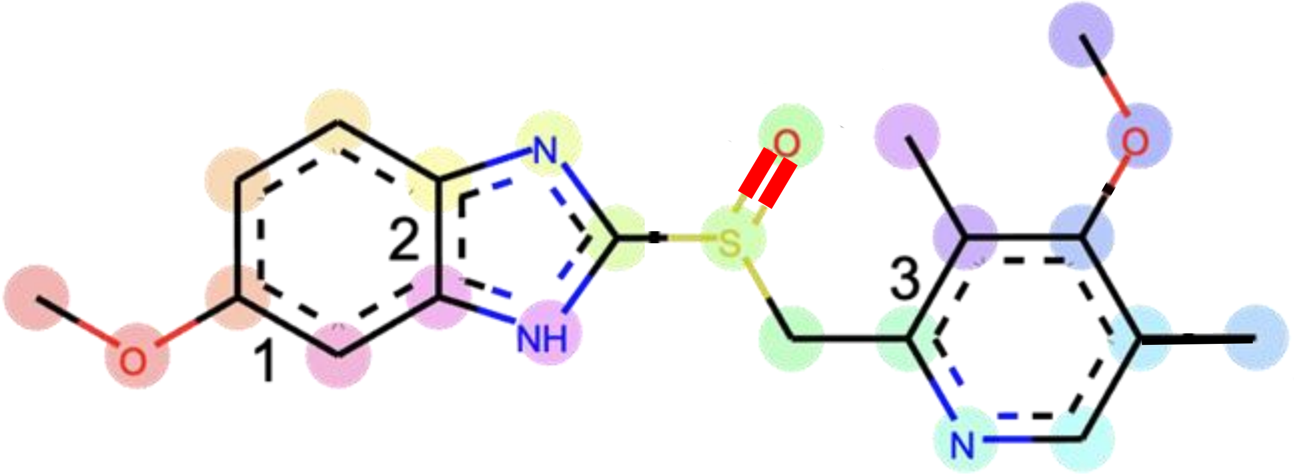
Decoder

Molecule


New Molecule

Decoder

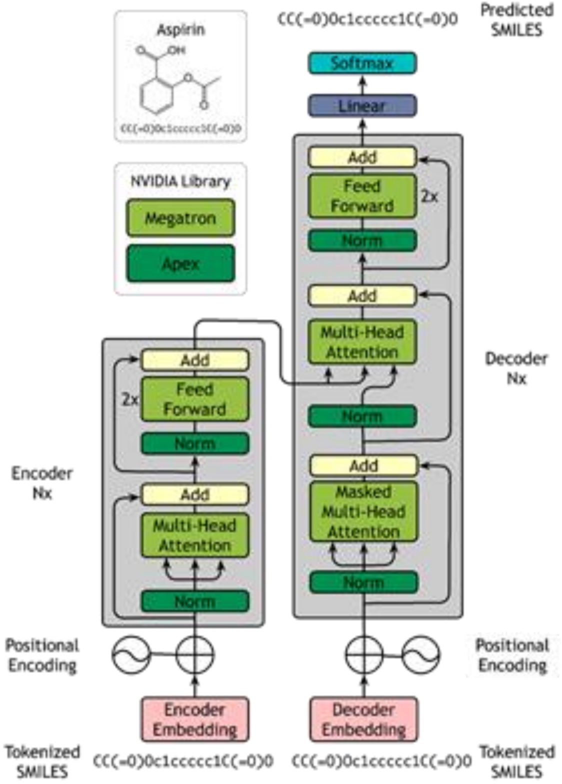
SMILES: a Natural Language Representation of Small Molecules



COc1ccc2nc(S(=O)Cc3ncc(C)c(OC)c3C)[nH]c2c

MegaMolBART Molecule Representations

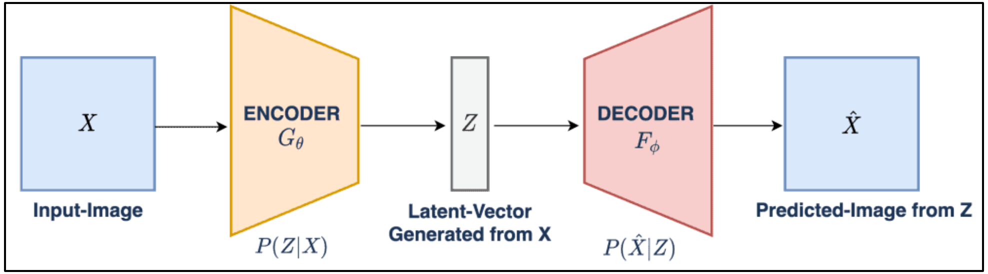
- MegaMolBART is a sequence-to-sequence developed in collaboration with AstraZeneca
- Based on BART NLP model
- Trained on 1.5B small molecules in SMILES format
- Useful for representation and sequence translation tasks
- Not well suited for generation tasks -- lacks an organized and uniformly shaped latent space



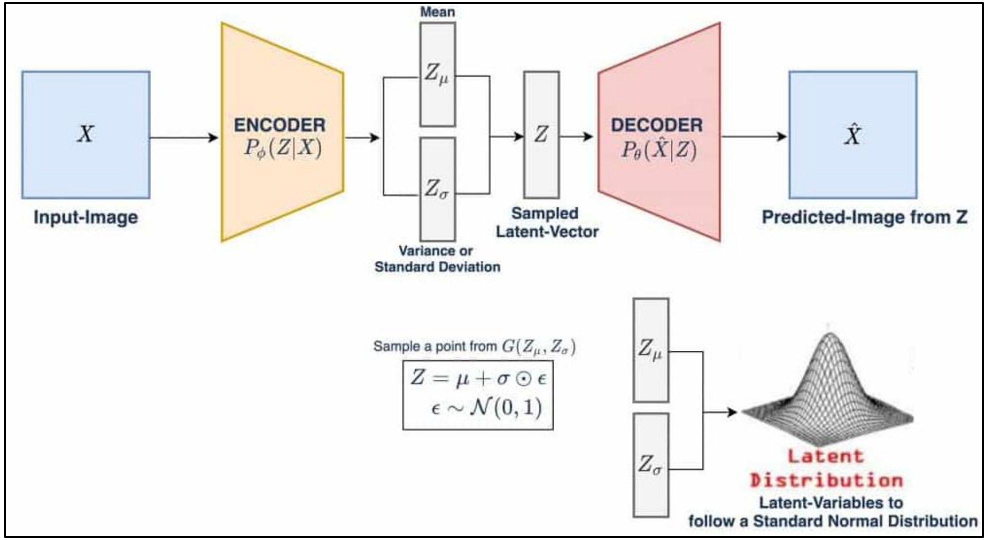
Chemformer publication: Irwin, R., *et al*, Mach. Learn.: Sci. Technol. 3 (2022).

Autoencoder Models in a Nutshell

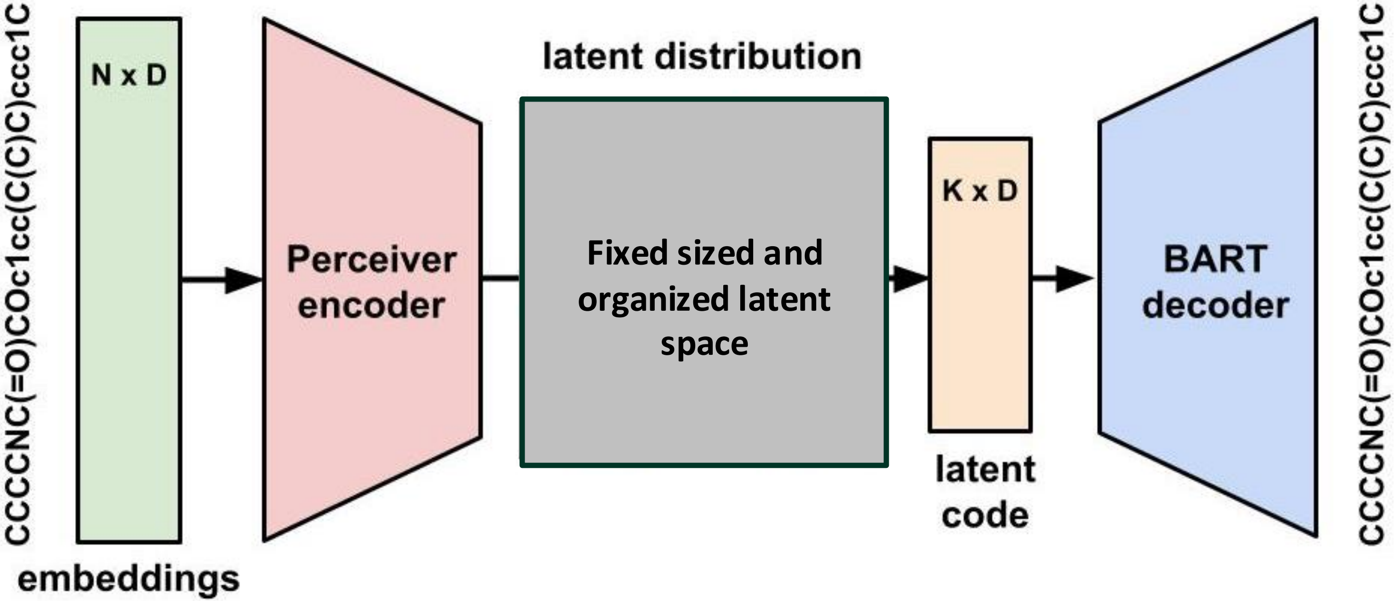
Autoencoder



Variational Autoencoder (VAE)



Development of MolMIM for Molecule Generation



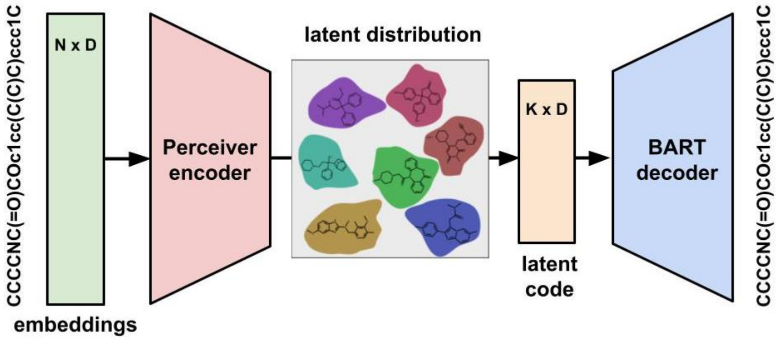
CCCCNC(=O)COc1cc(C(C)C)ccc1C
N x D
embeddings
Perceiver encoder
latent distribution
Fixed sized and organized latent space
K x D
latent code
BART decoder
CCCCNC(=O)COc1cc(C(C)C)ccc1C

A. Jaegle, *et al.*, ArXiv (2021).

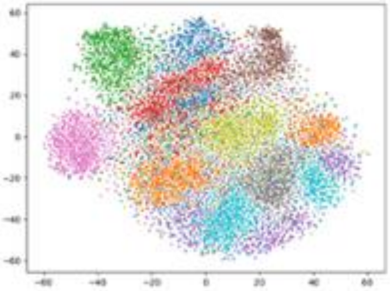
A Clustered Latent Space with Mutual Information Machine

- Mutual information machine (MIM) has a loss function that maximizes mutual information and minimizes marginal entropy

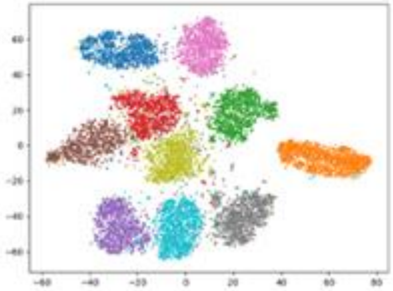
- MIM loss results in a clustered space while variational autoencoder (VAE) loss smooths the latent space resulting in blurring



VAE



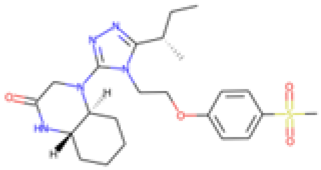
MIM



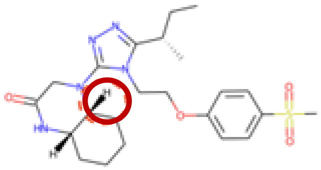
M. Livne, K. Swersky, D. J. Fleet, ArXiv (2019).

MolMIM – Sampling Distance Can Be Tuned for Similarity

Small Perturbations

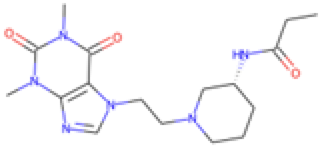


Seed
Molecule

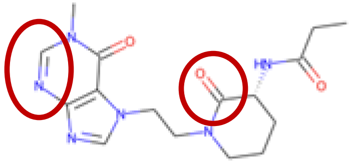


Sampled
Molecule

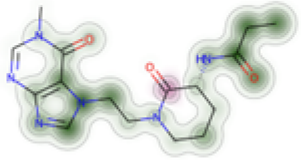
Larger Perturbations



Seed
Molecule



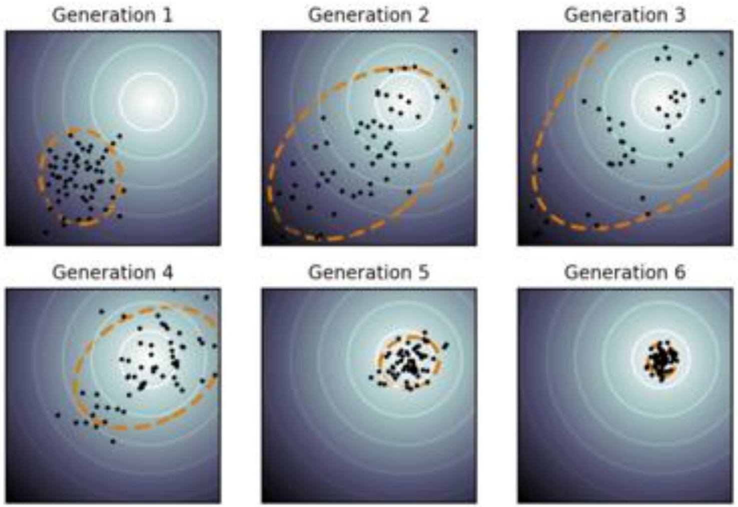
Sampled
Molecule



Similarity
Map

Measuring the Controllability of MolMIM Generation

- **Hypothesis:** having a structured latent space will improve performance of property guided optimization
- Chose covariance matrix adaptation (CMA-ES), which is a zeroth order optimization method
- CMA-ES is non-parametric and uses only a single scoring function per sample



N. Hansen, A. Ostermeier, *Evol. Comput.* 9, 159–195 (2001).

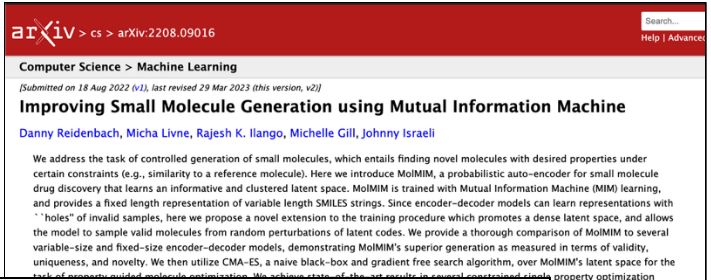
Multi-Objective Property Optimization

- Performed multi-objective optimization to jointly optimize two molecule properties (QED, SA) and binding to two protein targets (JNK3, GSK4β)
- Novelty is proportion of molecules with similarity metric (0.0 – 1.0) less than ≤ 0.4 relative to any other molecule
- Diversity is average similarity across all compounds
- MolMIM is competitive for success and diversity, but novelty has room for improvement

Model	QED + SA + JNK3 + GSK4β		
	Success (%)	Novelty (%)	Diversity
RationaleRL	74.8	56.1	0.621
MARS	92.3	82.4	0.719
JANUS	**100**	32.6	**0.821**
FaST	**100**	**100**	0.716
MolMIM (R)	97.5	71.1	0.791
MolMIM (A)	96.6	63.3	0.807
MolMIM (E)	98.3	55.1	0.767
MolMIM (E)[†]	99.2	54.8	0.772

Results above solid bar as in B. Chen, X. Fu, R. Barzilay, T. Jaakkola, ArXiv (2021)
QED, SA, JNK3, and GSK4β oracles from Therapeutic Data Commons

MolMIM: Research to Productization



Computer Science > Machine Learning

[Submitted on 18 Aug 2022 (v1), last revised 29 Mar 2023 (this version, v2)]

Improving Small Molecule Generation using Mutual Information Machine

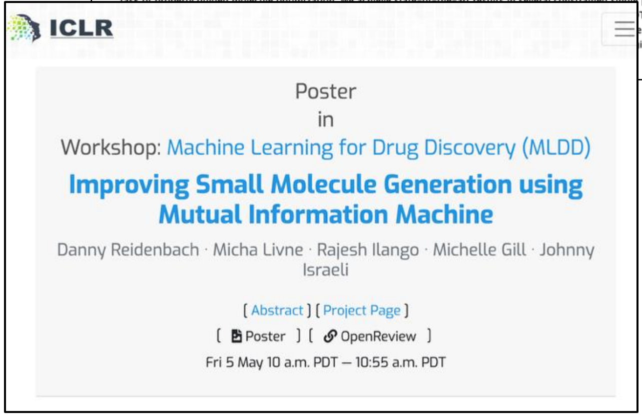
Danny Reidenbach, Micha Livne, Rajesh K. Ilango, Michelle Gill, Johnny Israeli

We address the task of controlled generation of small molecules, which entails finding novel molecules with desired properties under certain constraints (e.g., similarity to a reference molecule). Here we introduce MolMIM, a probabilistic auto-encoder for small molecule drug discovery that learns an informative and clustered latent space. MolMIM is trained with Mutual Information Machine (MIM) learning, and provides a fixed length representation of variable length SMILES strings. Since encoder-decoder models can learn representations with ``holes" of invalid samples, here we propose a novel extension to the training procedure which promotes a dense latent space, and allows the model to sample valid molecules from random perturbations of latent codes. We provide a thorough comparison of MolMIM to several variable-size and fixed-size encoder-decoder models, demonstrating MolMIM's superior generation as measured in terms of validity, uniqueness, and novelty. We then utilize CMA-ES, a naive black-box and gradient free search algorithm, over MolMIM's latent space for the task of property guided molecule optimization. We achieve state-of-the-art results in several constrained single property optimization

TA by more than 5\% . We

e, whereas CMA-ES is often

ime, making it an attractive



Poster
in
Workshop: Machine Learning for Drug Discovery (MLDD)

Improving Small Molecule Generation using Mutual Information Machine

Danny Reidenbach · Micha Livne · Rajesh Ilango · Michelle Gill · Johnny Israeli

[Abstract] [Project Page]
[Poster] [OpenReview]

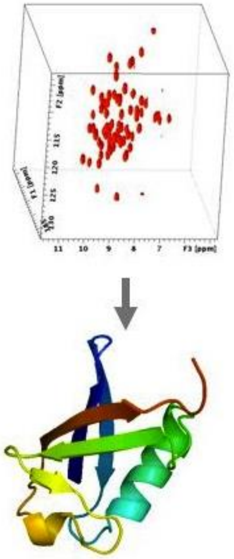
Fri 5 May 10 a.m. PDT — 10:55 a.m. PDT

- Integration of MolMIM model into BioNeMo inference service
- Productionize model architecture and training framework
- Accelerated inference
- Improving encoder representations

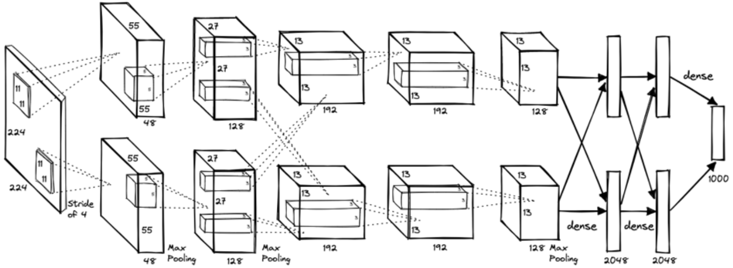
“How I Got Here” and Lessons Learned Along the Way

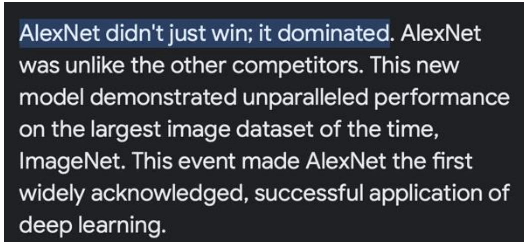
From Structural Biologist to Data Scientist

Postdoctoral Research: Enzyme Dynamics by NMR Spectroscopy



AlexNet Won ImageNet Challenge in 2012





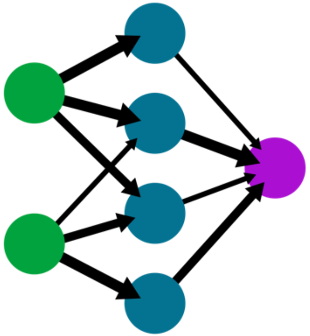
AlexNet didn't just win; it dominated. AlexNet was unlike the other competitors. This new model demonstrated unparalleled performance on the largest image dataset of the time, ImageNet. This event made AlexNet the first widely acknowledged, successful application of deep learning.

Don't miss the bigger picture: Machine learning will have an impact on every industry.

From Structural Biologist to Data Scientist

coursera

edX



PyData

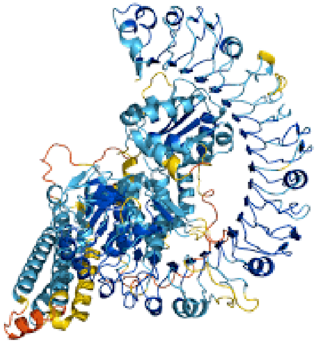
NYC 2013

NVIDIA

A Deep Learning Model Became the World’s Best Protein Structure Predictor

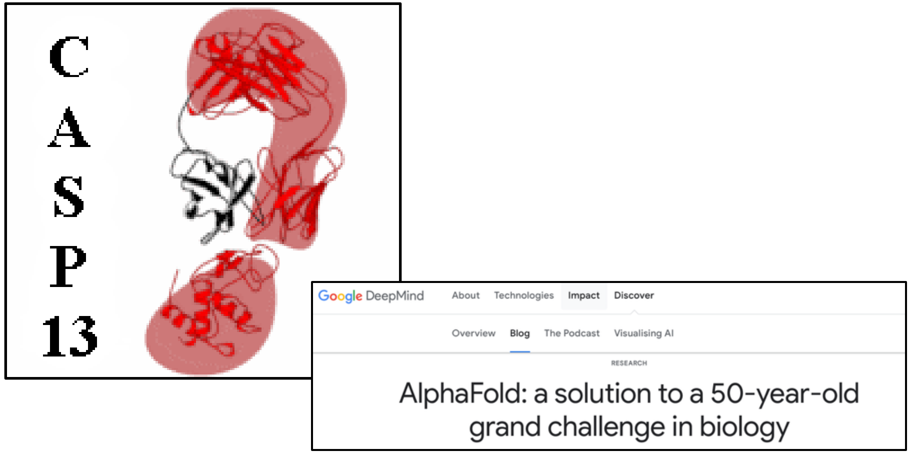
DeepMind

...MALKIPTHNHM...
...VFRDCEWS...
...WYIOPMNVGTDEW...



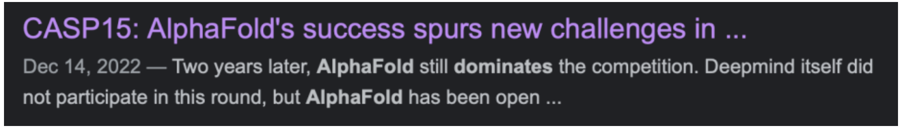
Sequence

Structure



RESEARCH

AlphaFold: a solution to a 50-year-old grand challenge in biology

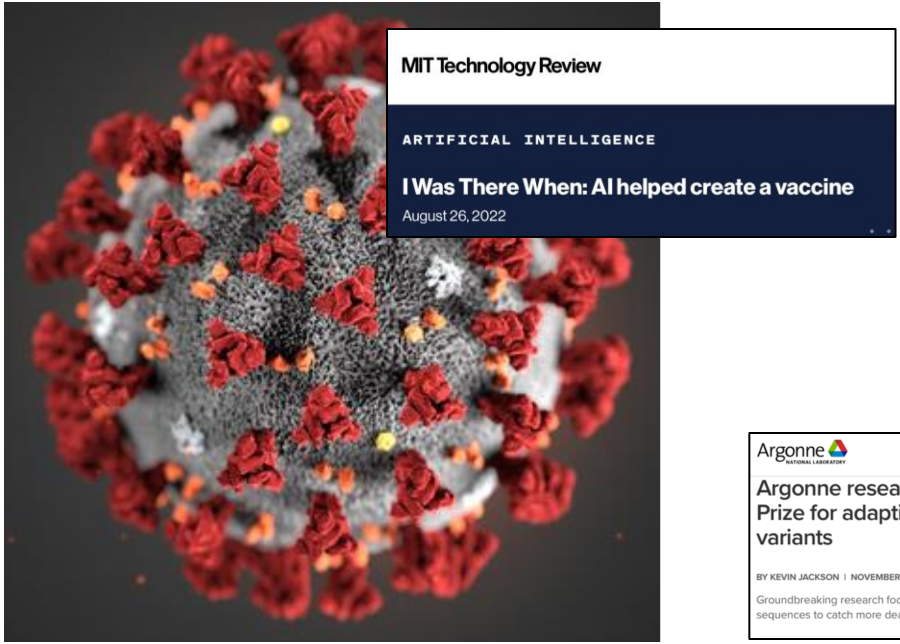


CASP15: AlphaFold's success spurs new challenges in ...

Dec 14, 2022 — Two years later, **AlphaFold** still **dominates** the competition. Deepmind itself did not participate in this round, but **AlphaFold** has been open ...

AlphaFold won the Critical Assessment of Protein Structure Prediction (CASP13) Competition in 2018 ... and has done so every year since

AI and the Race for a COVID-19 Vaccine

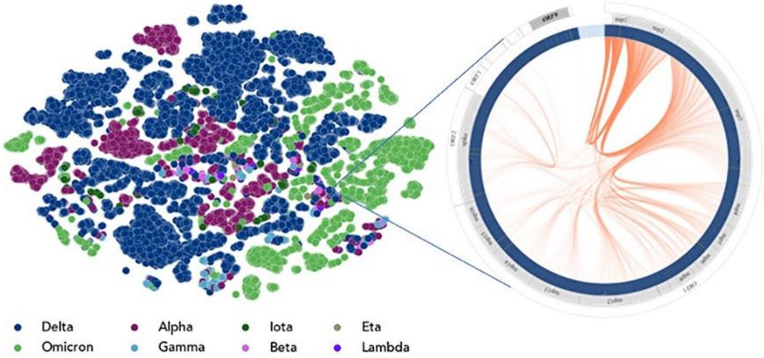


MIT Technology Review

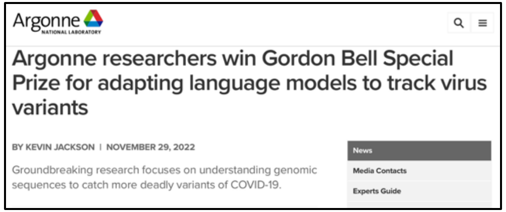
ARTIFICIAL INTELLIGENCE

I Was There When: AI helped create a vaccine

August 26, 2022



Genome-scale language models (GenSLMs) discover distinct evolutionary patterns in SARS-CoV-2



Argonne NATIONAL LABORATORY

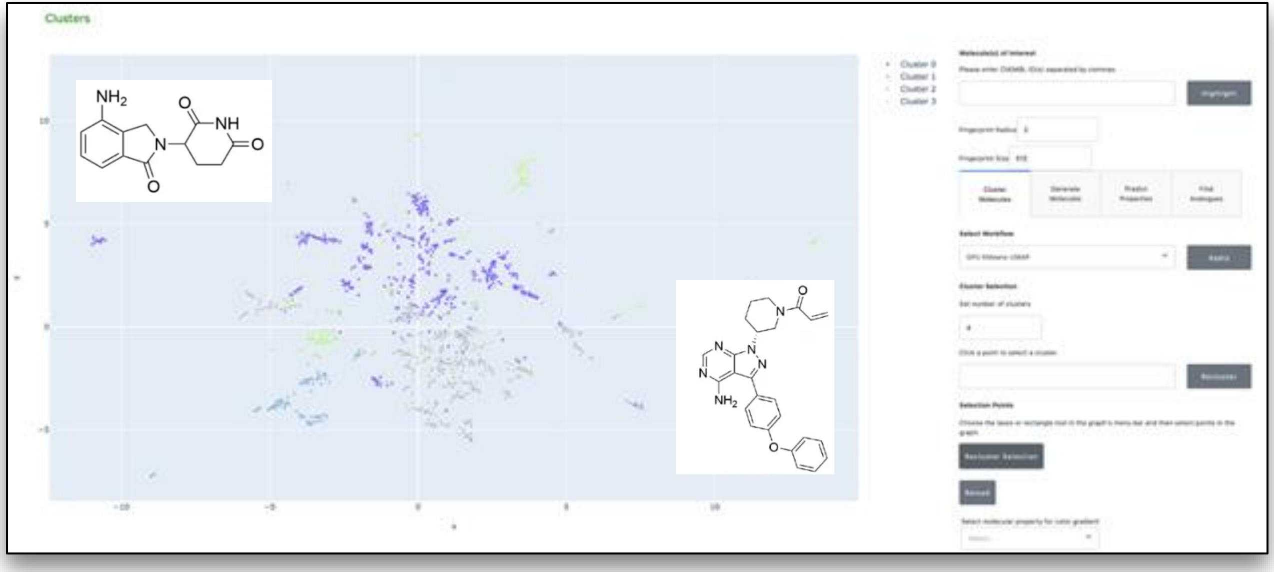
Argonne researchers win Gordon Bell Special Prize for adapting language models to track virus variants

BY KEVIN JACKSON | NOVEMBER 29, 2022

Groundbreaking research focuses on understanding genomic sequences to catch more deadly variants of COVID-19.

Zvyagin, M. *et al.* Biorxiv 2022.10.10.511571 (2022) doi:10.1101/2022.10.10.511571.

First Effort: Interface for Clustering and Visualization of Small Molecules



Molecule(s) of Interest

Please enter ChEMBL ID(s) separated by commas

Fingerprint Radius: 2

Fingerprint Size: 512

Select Workflow

GPU KMeans-UMAP

Cluster Selection

Set number of clusters

4

Click a point to select a cluster.

Selection Points

Choose the lasso or rectangle tool in the graph's menu bar and then select points in the graph.

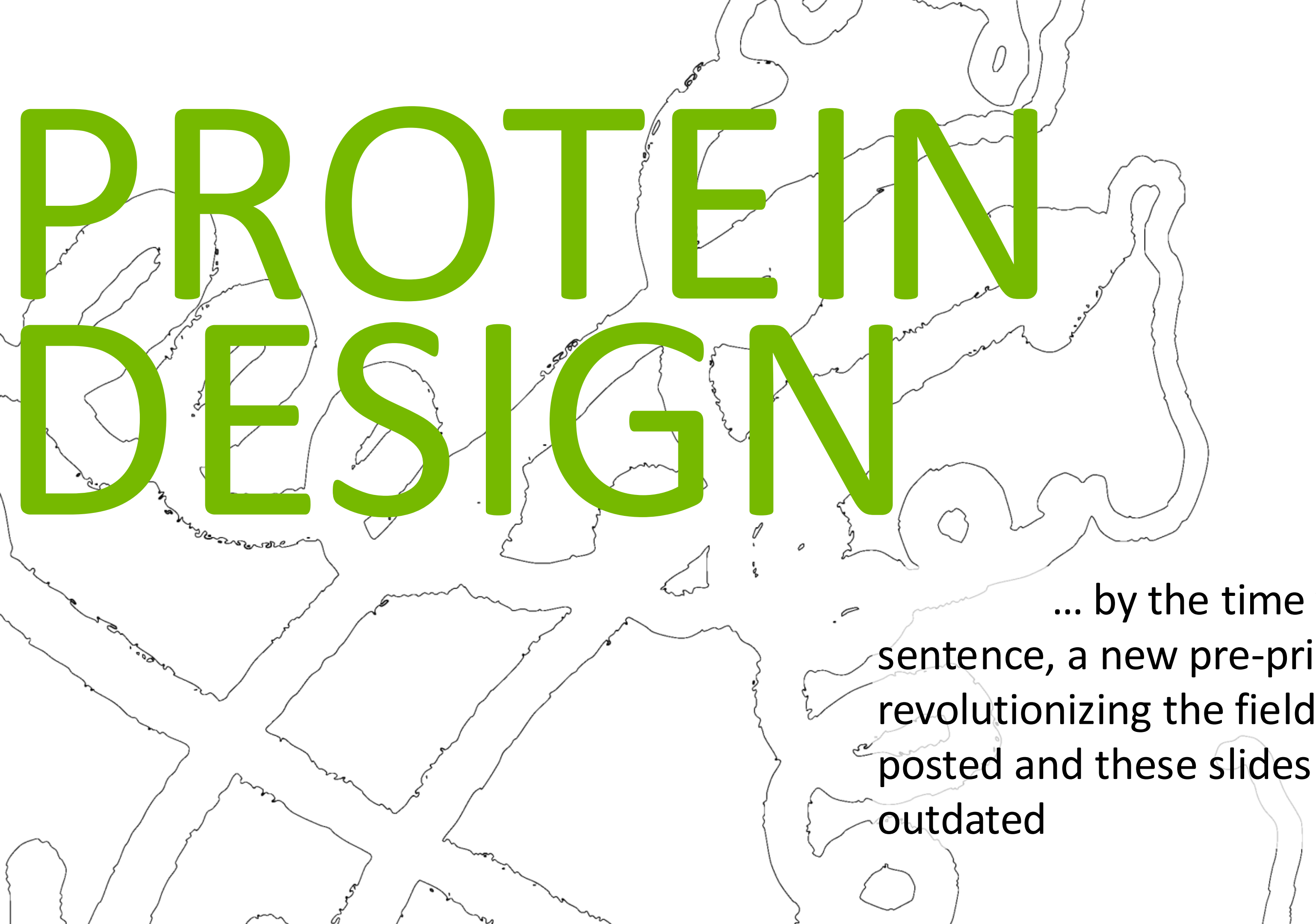
Select molecular property for color gradient

Select...

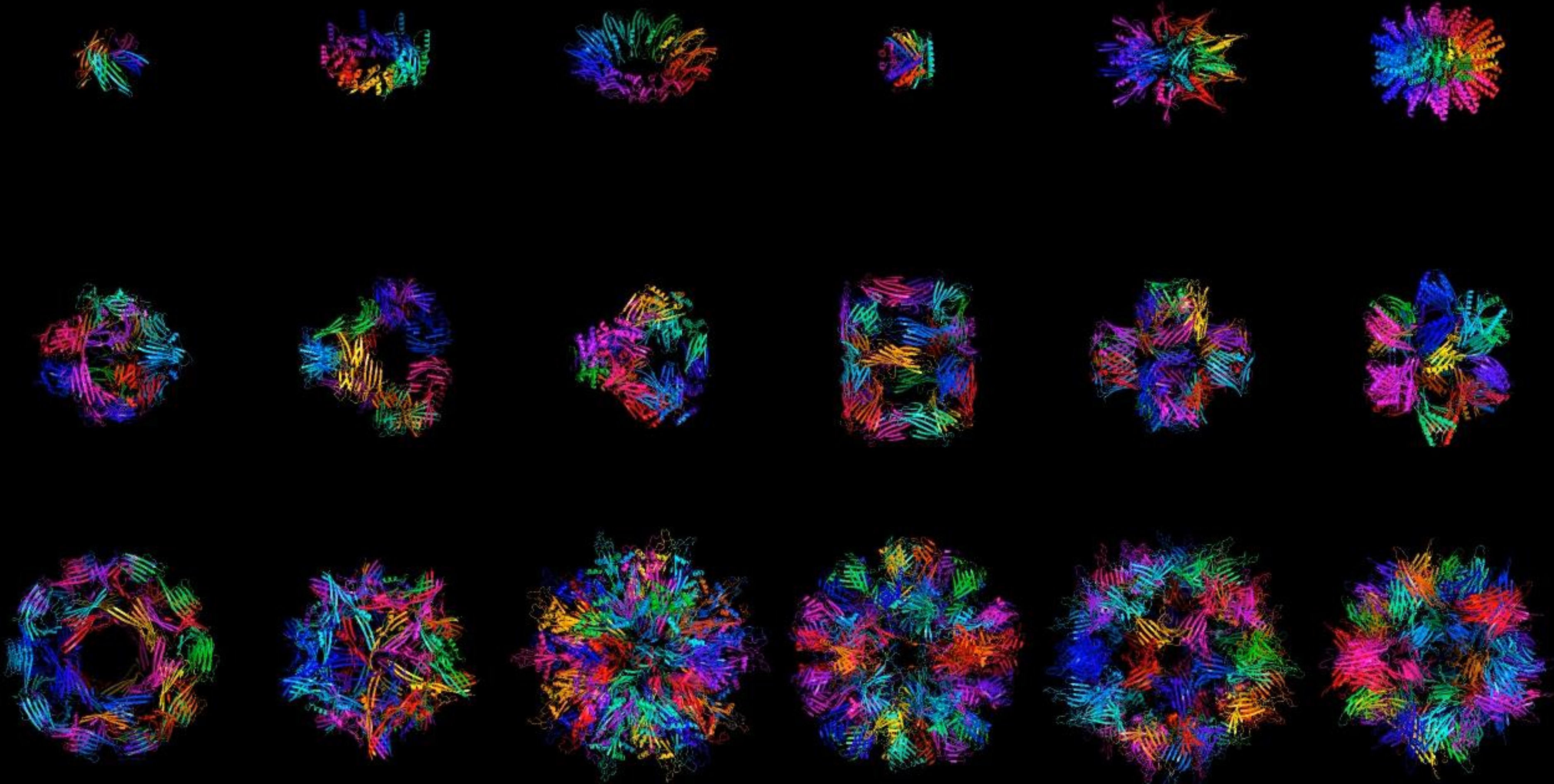
dask RAPIDS plotly | Dash

Deep learning is high risk. Ensure the project will succeed if deep learning fails.

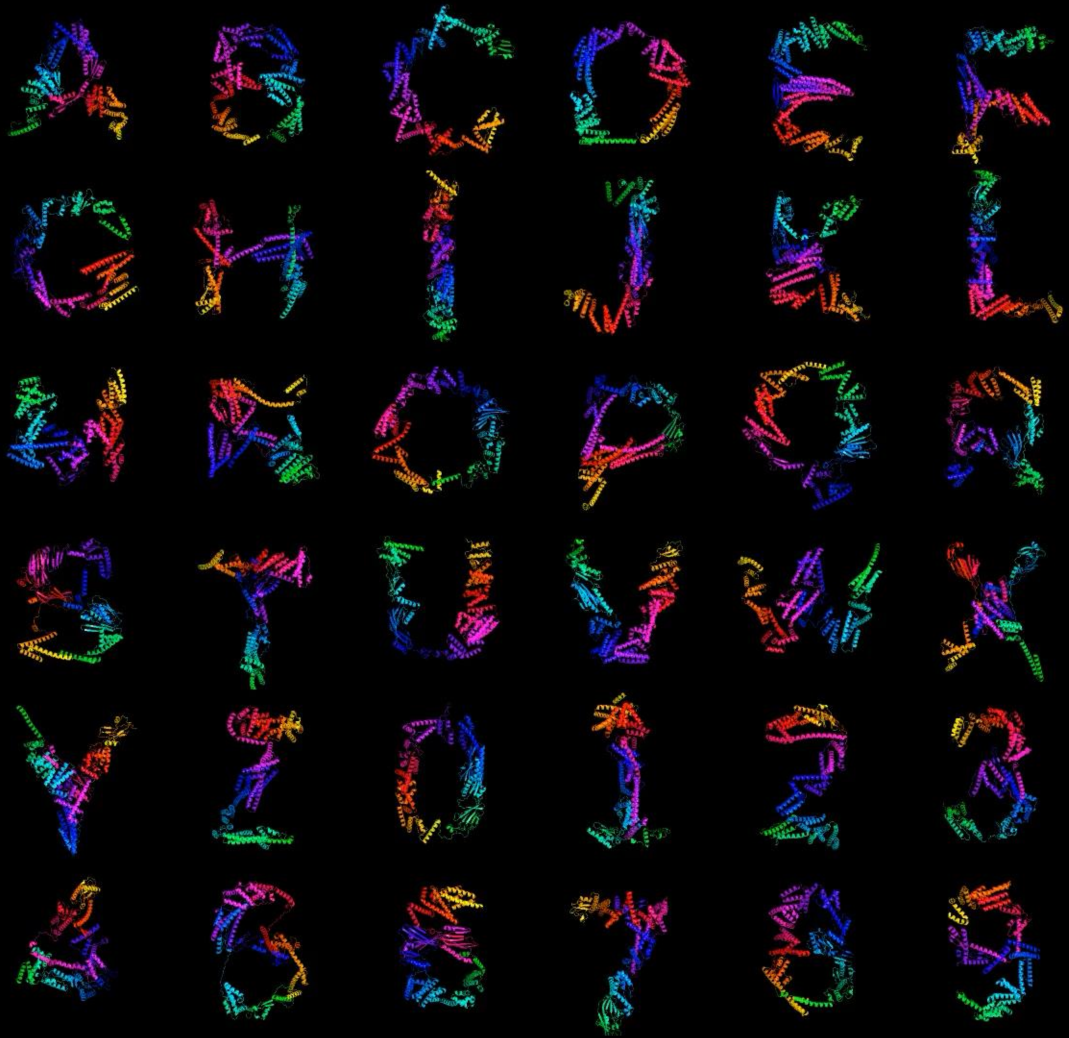
PROTEIN DESIGN



… by the time you’ve read this sentence, a new pre-print revolutionizing the field has been posted and these slides are totally outdated

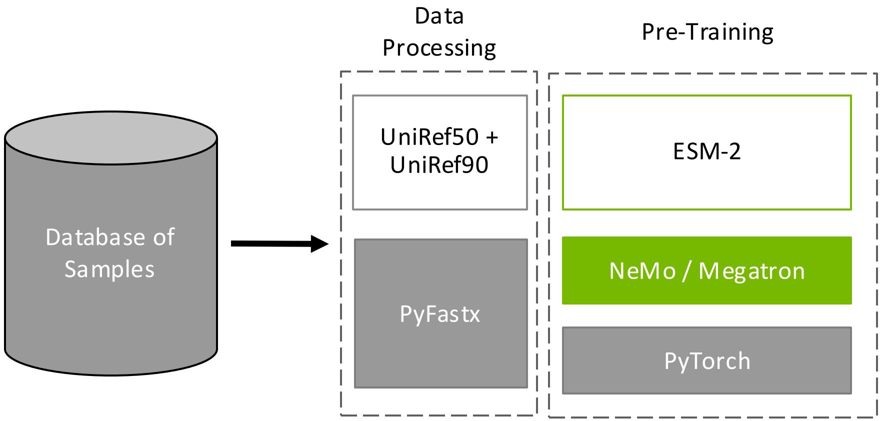


Ingraham, J. *et al*. Illuminating protein space with a programmable generative model.(2022) doi:10.1101/2022.12.01.518682.



Ingraham, J. *et al*. Illuminating protein space with a programmable generative model. (2022) doi:10.1101/2022.12.01.518682.

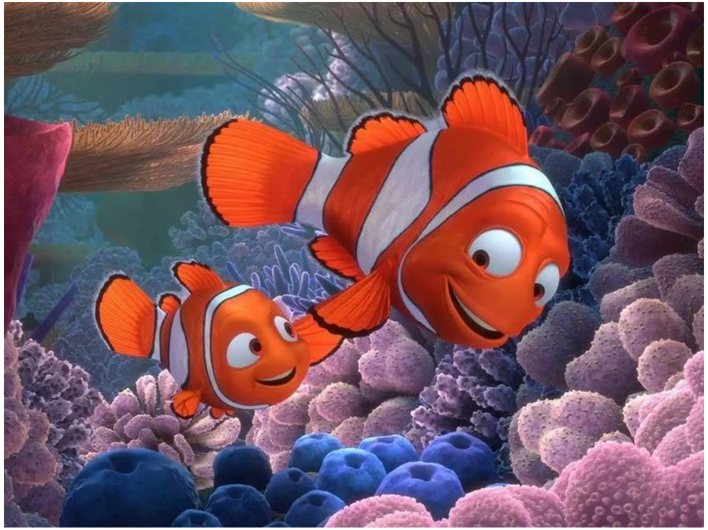
Developing Deep Learning Models at Scale



Model Size (Param)	Training Time (Days)	
	512 x V100s	512 x A100s
650M	8	???
3B	30	

Successes from calculated risks provide justification for growing a team.

Rapid Team Growth and Adventures in Management



Two Engineers

< Two Years

Over Thirty Engineers

Deep learning is hard, but growing and managing a team is the most challenging problem.

Conclusions

- BioNeMo is a framework and inference service for developing, training, deploying, and using deep learning models and tools for drug discovery
- MolMIM is a cheminformatics language model trained on SMILES with a structured latent space for molecule design
- Careers are long compared to the pace of machine learning advancement
- Capitalize on new opportunities and enjoy the ride!

BioNeMo Inference Service early access : https://www.nvidia.com/bionemo
BioNeMo Framework general access coming next week!

The BioNeMo Team

Johnny Israeli

Alireza Moradzadeh

Arkadiusz Nowaczynski

Camir Ricketts

Danny Reidenbach

Dejun Lin

Dorota Toczydlowska

Emine Kucukbenli

Eric Dawson

Farhad Ramezanghorbani

Gagan Kaushik

George Armstrong

Guoqing Zhou

Han-Yi Chou

Jasleen Grewal

Kevin Boyd

Maria Korshunova

Mario Geiger

Marta Stepniewska-Dziubinska

Micha Livne

Neha Tadimeti

Ohad Mosafi

Pablo Ribalta

Rajesh Ilango

Sara Rabhi

Simon Chu

Srimukh Vecchan

Steven Kothen-Hill

Tomasz Grzegorzek

Timur Rvachov

Yuxing Peng

Zachary McClure

Thank You!

Questions:

Fireside Chat

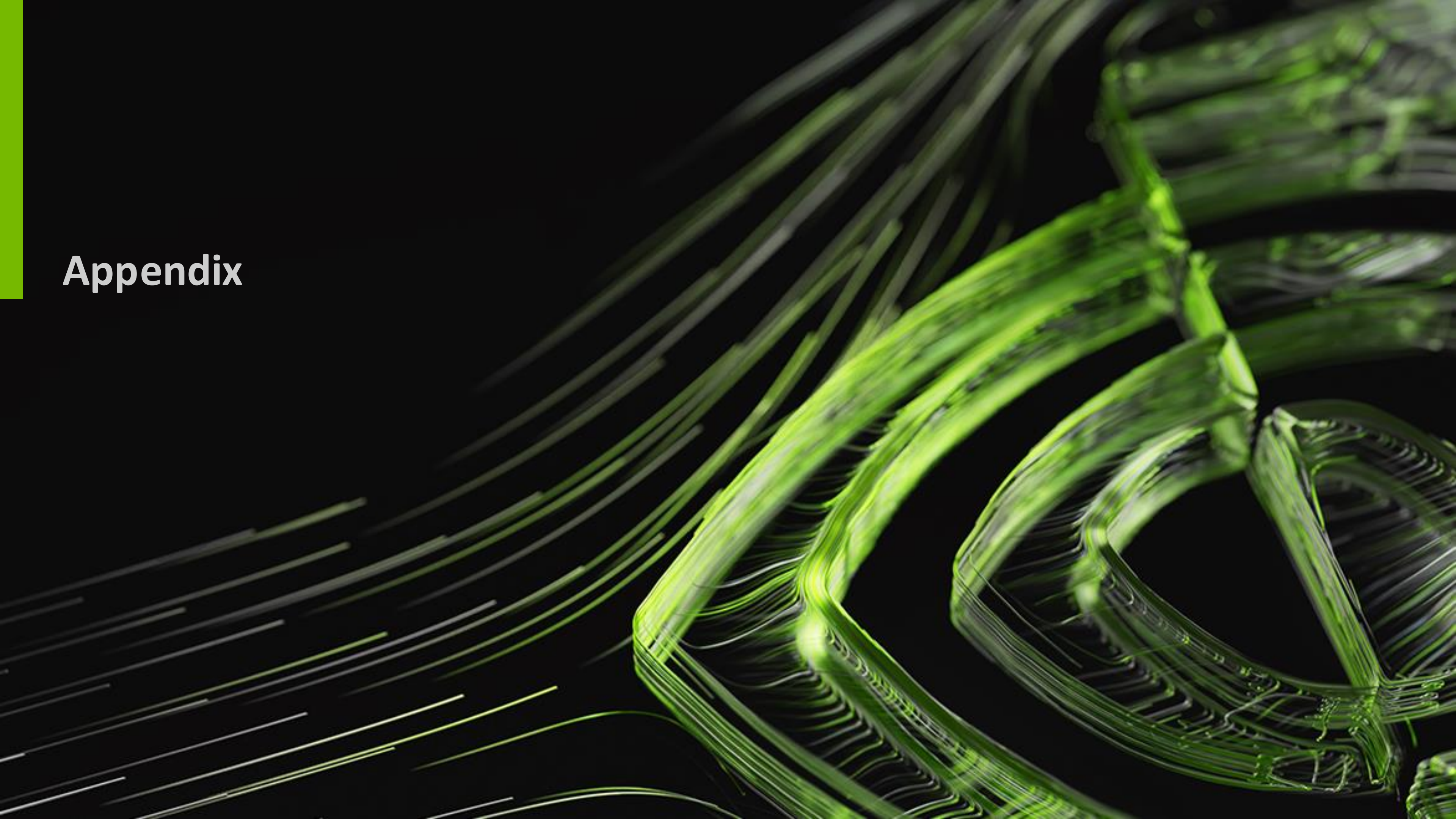
10:15 – 10:55am

Central Park East

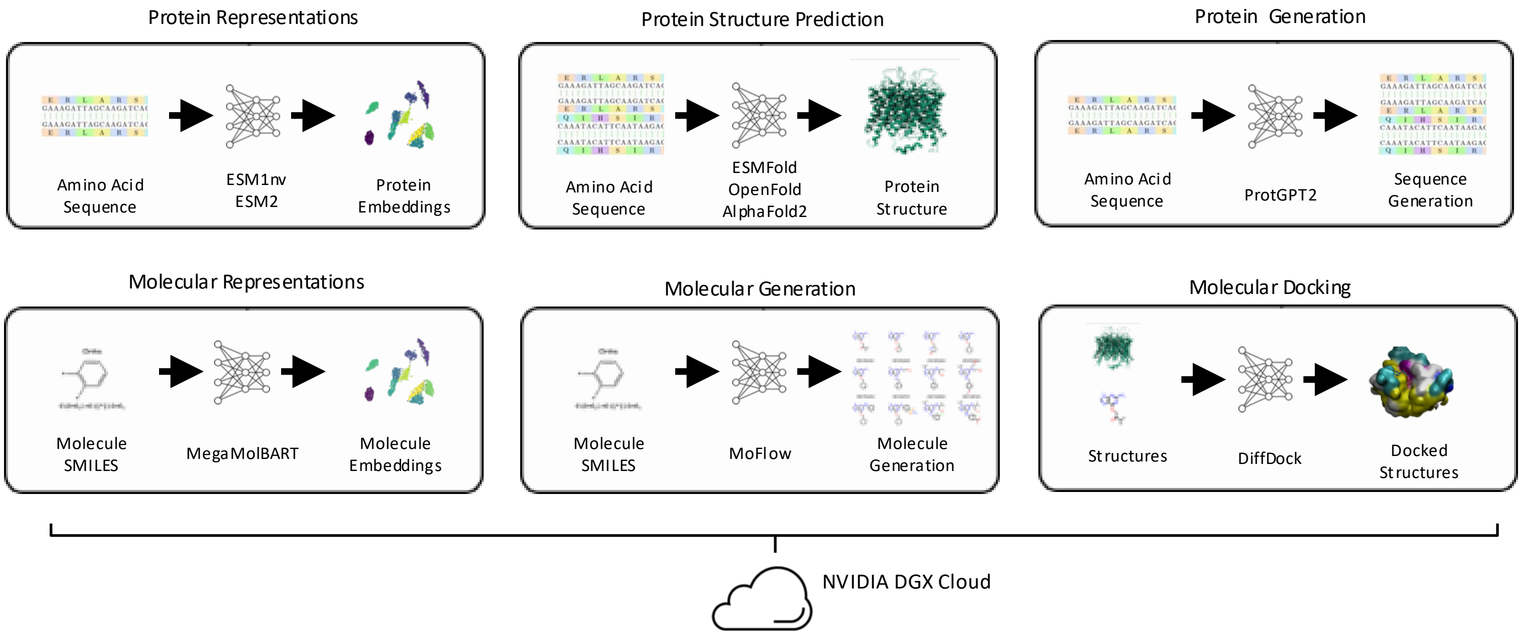
✉ mgill@nvidia.com

michellelynngill.com

Appendix



Nine Models in Inference Service for Drug Discovery Applications



Protein Representations

Amino Acid Sequence → ESM1nv ESM2 → Protein Embeddings

Protein Structure Prediction

Amino Acid Sequence → ESMFold OpenFold AlphaFold2 → Protein Structure

Protein Generation

Amino Acid Sequence → ProtGPT2 → Sequence Generation

Molecular Representations

Molecule SMILES → MegaMolBART → Molecule Embeddings

Molecular Generation

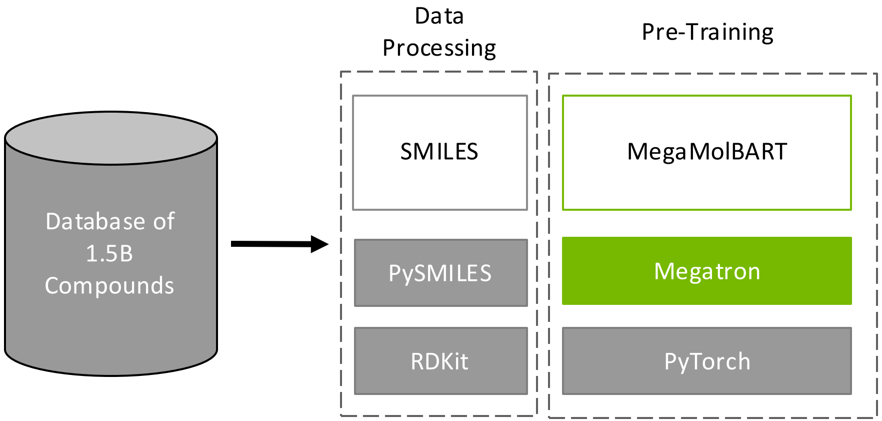
Molecule SMILES → MoFlow → Molecule Generation

Molecular Docking

Structures → DiffDock → Docked Structures

NVIDIA DGX Cloud

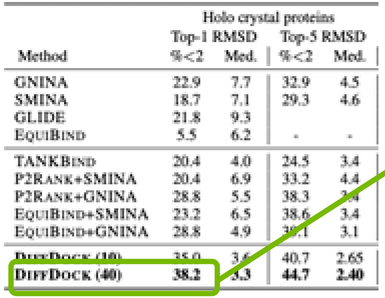
Deep Learning at Scale



Attention Heads	Layers	Hidden Size	Feed Forward	Parameters
8	4	256	1024	10M
8	6	512	2048	45M
16	8	1024	4096	230M

Life Cycle of a BioNeMo Model in the Inference Service

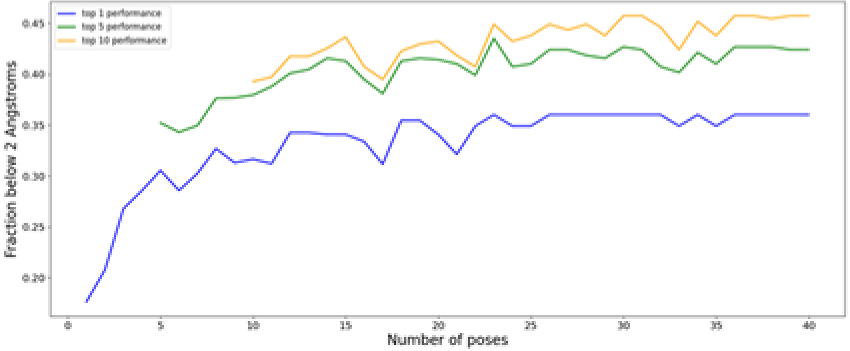
- Model checkpoints are accelerated using a variety of NVIDIA tools – standard tricks to custom CUDA kernels
- All quantitative and qualitative results are reproduced
- For DiffDock, the RMSD metrics were reproduced under a variety of different conditions



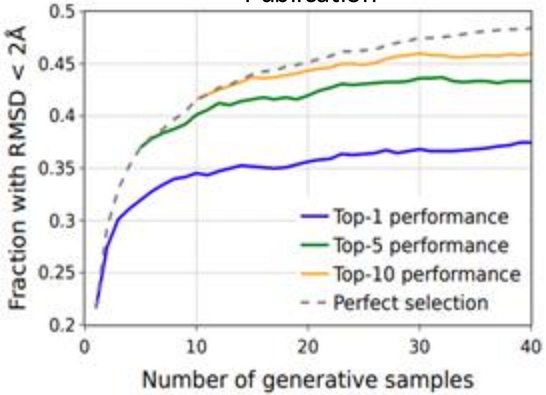
	Holo crystal proteins			
	Top-1 RMSD		Top-5 RMSD	
Method	%<2	Med.	%<2	Med.
GNINA	22.9	7.7	32.9	4.5
SMINA	18.7	7.1	29.3	4.6
GLIDE	21.8	9.3		
EQUIBIND	5.5	6.2	-	-
TANKBIND	20.4	4.0	24.5	3.4
P2RANK+SMINA	20.4	6.9	33.2	4.4
P2RANK+GNINA	28.8	5.5	38.3	3.4
EQUIBIND+SMINA	23.2	6.5	38.6	3.4
EQUIBIND+GNINA	28.8	4.9	39.1	3.1
DIFFDOCK (10)	35.0	3.6	40.7	2.65
DIFFDOCK (40)	**38.2**	**3.3**	**44.7**	**2.40**

NV Trial #1	38.0
NV Trial #2	35.0
NV Trial #3	38.6
NV Trial #4	39.1
NV Trial #5	38.6

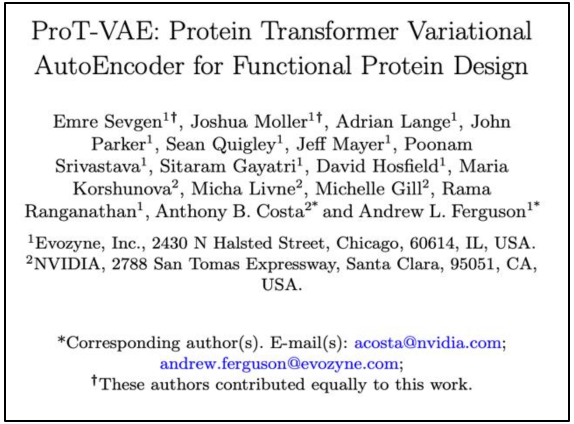
Reproduction



Publication



Proteins Generated from Evozyne's ProT-VAE Models

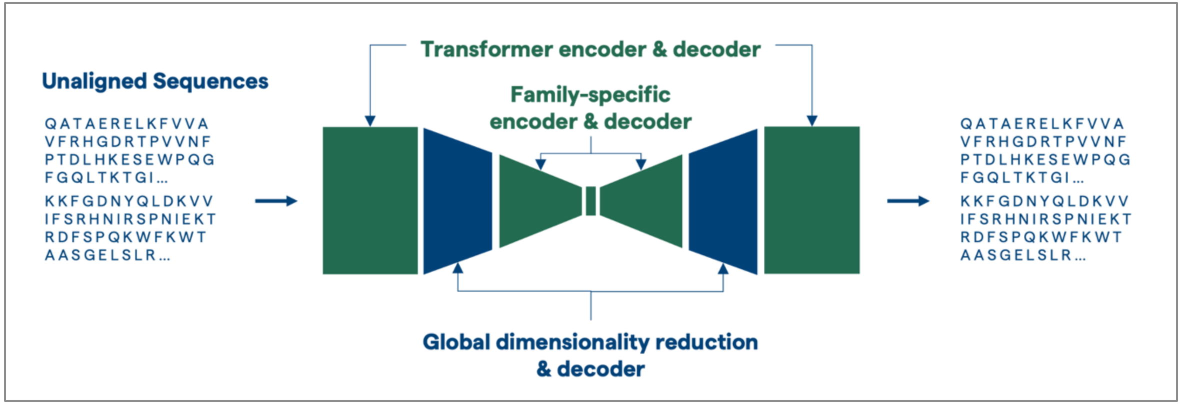


ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design

Emre Sevgen[1†], Joshua Moller[1†], Adrian Lange[1], John Parker[1], Sean Quigley[1], Jeff Mayer[1], Poonam Srivastava[1], Sitaram Gayatri[1], David Hosfield[1], Maria Korshunova[2], Micha Livne[2], Michelle Gill[2], Rama Ranganathan[1], Anthony B. Costa[2*] and Andrew L. Ferguson[1*]

[1]Evozyne, Inc., 2430 N Halsted Street, Chicago, 60614, IL, USA.
[2]NVIDIA, 2788 San Tomas Expressway, Santa Clara, 95051, CA, USA.

*Corresponding author(s). E-mail(s): acosta@nvidia.com; andrew.ferguson@evozyne.com;
†These authors contributed equally to this work.



Unaligned Sequences

QATAERELKFVVA
VFRHGDRTPVVNF
PTDLHKESEWPQG
FGQLTKTGI...

KKFGDNYQLDKVV
IFSRHNIRSPNIEKT
RDFSPQKWFKWT
AASGELSLR...

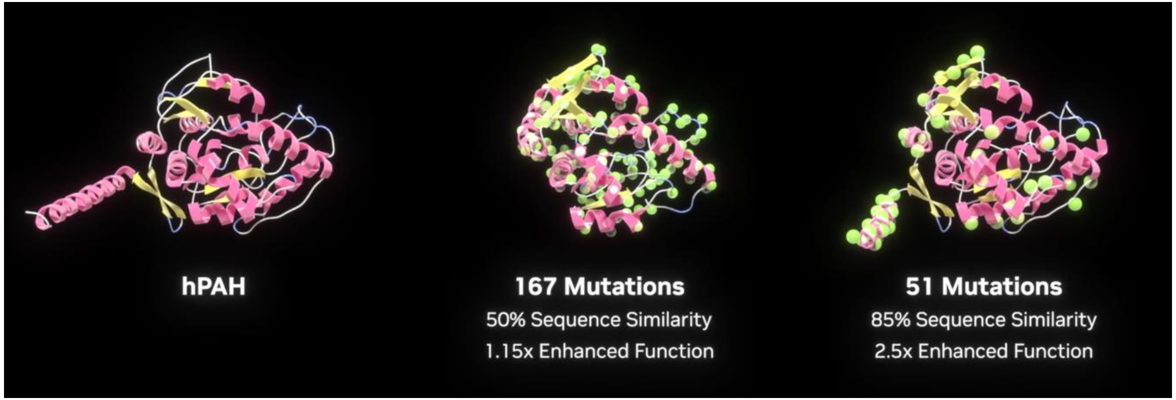
Transformer encoder & decoder

Family-specific encoder & decoder

Global dimensionality reduction & decoder

QATAERELKFVVA
VFRHGDRTPVVNF
PTDLHKESEWPQG
FGQLTKTGI...

KKFGDNYQLDKVV
IFSRHNIRSPNIEKT
RDFSPQKWFKWT
AASGELSLR...

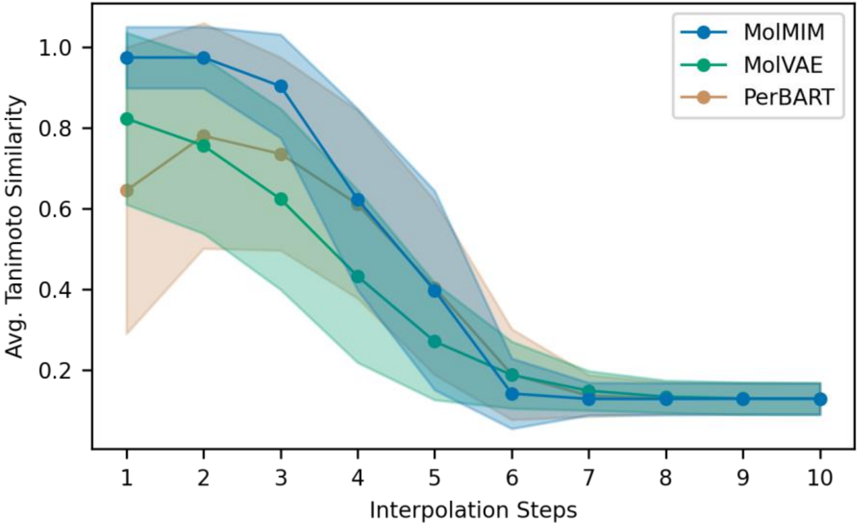


hPAH

167 Mutations
50% Sequence Similarity
1.15x Enhanced Function

51 Mutations
85% Sequence Similarity
2.5x Enhanced Function

Probing Latent Structure by Molecule Interpolation



- Pairwise interpolations performed at ten evenly spaced steps for 1,000 ZINC15 molecules
- Average Tanimoto similarity to first molecule was calculated at each step
- Molecules sampled from Perceiver BART and MolVAE have reduced similarity to start and a large degree of variability at early interpolation steps
- Molecules sampled from MolMIM are similar and have the smallest variance at early steps

MolMIM – Performance on Seed Based Molecule Sampling

- Randomly sampled ten molecules for each of 20k molecules from test split
- Effective novelty is percentage of molecules that are valid, unique, not identical to seed, and novel
- Sampling radius empirically determined to maximize effective novelty
- CDDD used as baseline model – trained with molecular property loss
- Perceiver BART sampling speed improved relative to MegaMolBART
- MolVAE and MolMIM show significant improvements in validity and effective novelty

Model	Latent Dim	Validity (%)	Uniqueness (%)	Novelty (%)	Effective Novelty (%)	Test Runtime
MegaMolBART	Variable	75.0	84.8	94.2	51.1	8.7 hours
Perceiver BART	2048	71.8	94.9	94.6	59.1	38 min
MolVAE	2048	95.7	**100.0**	98.1	93.9	64 min
MolMIM	512	**98.7**	**100.0**	95.5	**94.2**	30 min
CDDD	512	84.5	98.9	**99.5**	82.2	12 hours[†]

[†] CDDD decoding speed limited by batch size.

R. Winter, *et. al.*, Chemical Science. 10, 1692–1701 (2019).

Single Property Optimization with CMA-ES

Model	QED (%)	Penalized logP	
	δ ≥ 0.4	δ ≥ 0.4	δ ≥ 0.6
AtomG2G	73.6	-	-
HeirG2G	76.9	-	-
DESMILES	77.8	-	-
QMO	92.8	7.71 ± 5.65	3.73 ± 2.85
MolGrow	-	8.34 ± 6.85	4.06 ± 5.61
GraphAF	-	8.21 ± 6.51	4.98 ± 6.49
GraphDF	-	9.19 ± 6.43	4.51 ± 5.80
CDGS	-	9.56 ± 6.33	5.10 ± 5.80
FaST	-	18.09 ± 8.72	8.98 ± 6.31
MolMIM	**94.6**	**28.45 ± 54.67**	**7.60 ± 23.62**
MolMIM		9.44 ± 4.12†	4.57 ± 3.87†

- Performed optimization of QED or penalized logP with query budget of 50,000 oracle calls per input molecule
- Success is % of molecules with QED ≥ 0.9 or penalized logP improvement while maintaining Tanimoto similarity δ ≥ {0.4, 0.6}
- MolMIM achieves the highest QED and logP success rates
- Penalized logP results impacted by known exploit where identical functional groups are repeatedly added

Results above solid bar as in B. Chen, X. Fu, R. Barzilay, T. Jaakkola, ArXiv (2021) and S. C. Hoffman, *et al*, Nat Mach Intell. 4, 21–31 (2022)
QED and logP oracles from Therapeutic Data Commons.
†logP improvement limited to ≤ 20

Single Property Optimization with CMA-ES

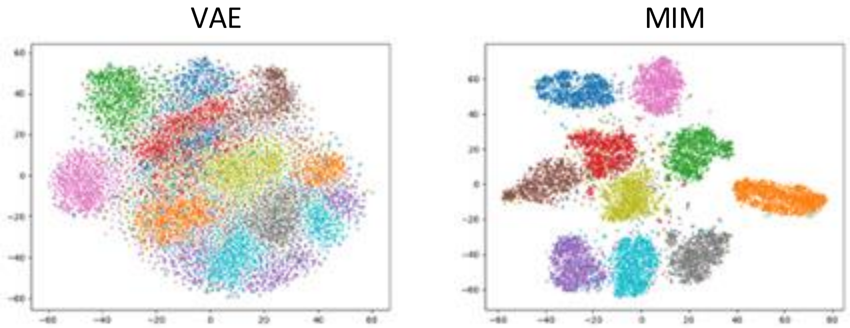
Model	QED (%)	Penalized logP	
	δ ≥ 0.4	δ ≥ 0.4	δ ≥ 0.6
AtomG2G	73.6	-	-
HeirG2G	76.9	-	-
DESMILES	77.8	-	-
QMO	92.8	7.71 ± 5.65	3.73 ± 2.85
MolGrow	-	8.34 ± 6.85	4.06 ± 5.61
GraphAF	-	8.21 ± 6.51	4.98 ± 6.49
GraphDF	-	9.19 ± 6.43	4.51 ± 5.80
CDGS	-	9.56 ± 6.33	5.10 ± 5.80
FaST	-	18.09 ± 8.72	8.98 ± 6.31
MolMIM	**94.6**	**28.45 ± 54.67**	**7.60 ± 23.62**
MolMIM		9.44 ± 4.12†	4.57 ± 3.87†

- Performed optimization of QED or penalized logP with query budget of 50,000 oracle calls per input molecule
- Success is % of molecules with QED ≥ 0.9 or penalized logP improvement while maintaining Tanimoto similarity δ ≥ {0.4, 0.6}
- MolMIM achieves the highest QED and logP success rates
- Penalized logP results impacted by known exploit where identical functional groups are repeatedly added
- Recall: MolMIM trained without chemical properties, activity, or fragment knowledge

Results above solid bar as in B. Chen, X. Fu, R. Barzilay, T. Jaakkola, ArXiv (2021) and S. C. Hoffman, *et al*, Nat Mach Intell. 4, 21–31 (2022)
QED and logP oracles from Therapeutic Data Commons.
†logP improvement limited to ≤ 20

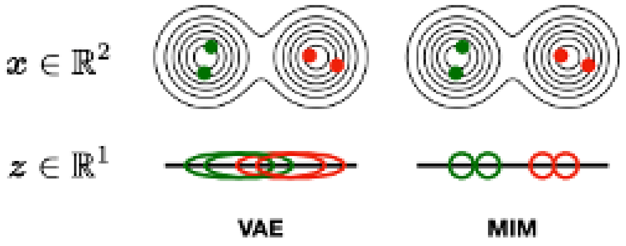
A Clustered Latent Space with Mutual Information Machine

- Same architecture as VAE, but loss maximizes mutual information and minimizes marginal entropy
- MIM results in an informative and clustered latent space


VAE

MIM

$$\begin{aligned}\mathcal{L}_{\text{A-MIM}}(\boldsymbol{\theta}) &= \frac{1}{2}\Big(CE\left(\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z}),\, q_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\right) \\ &\quad + CE\left(\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z}),\, p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\right)\Big) \\ &\geq H_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{x}) + H_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{z}) - I_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{x};\boldsymbol{z})\end{aligned}$$


$\boldsymbol{x} \in \mathbb{R}^2$
$\boldsymbol{z} \in \mathbb{R}^1$
VAE
MIM

M. Livne, K. Swersky, D. J. Fleet, ArXiv (2019).

Model	QED (%)	Penalized logP	
	δ ≥ 0.4	δ ≥ 0.4	δ ≥ 0.6
JT-VAE	8.8	1.03 ± 1.39	0.28 ± 0.79
GCPN	9.4	2.49 ± 1.30	0.79 ± 0.63
MolDQN	-	3.37 ± 1.62	1.86 ± 1.21
MMPA	32.9	-	-
VSeq2Seq	58.5	3.37 ± 1.75	2.33 ± 1.17
VJTNN+GAN	60.6	-	-
VJTNN	-	3.55 ± 1.67	2.33 ± 1.24
MoFlow	-	4.71 ± 4.55	2.10 ± 2.86
GA	-	5.93 ± 1.41	3.44 ± 1.09
AtomG2G	73.6	-	-
HeirG2G	76.9	-	-
DESMILES	77.8	-	-
QMO	92.8	7.71 ± 5.65	3.73 ± 2.85
MolGrow	-	8.34 ± 6.85	4.06 ± 5.61
GraphAF	-	8.21 ± 6.51	4.98 ± 6.49
GraphDF	-	9.19 ± 6.43	4.51 ± 5.80
CDGS	-	9.56 ± 6.33	5.10 ± 5.80
FaST	-	18.09 ± 8.72	8.98 ± 6.31
MolMIM	**94.6**	**28.45 ± 54.67**	**7.60 ± 23.62**
MolMIM		9.44 ± 4.12†	4.57 ± 3.87†

gle Property Optimization

- Repeated QED and penalized logP optimization with query budget of 50,000 oracle calls per input molecule
- Success is % of molecules with QED ≥ 0.9 or penalized logP improvement while maintaining Tanimoto similarity δ ≥ {0.4, 0.6}
- MolMIM achieves the highest QED and logP success rates
- Penalized logP results impacted by known exploit where identical functional groups are repeatedly added
- MolMIM results were repeated with logP improvement limited

Results above solid bar as in B. Chen, X. Fu, R. Barzilay, T. Jaakkola, ArXiv (2021) and S. C. Hoffman, *et al*, Nat Mach Intell. 4, 21–31 (2022).

†logP improvement limited to ≤ 20

Perspective on BioNeMo

- Models have a finite lifespan, the value is in the learnings

- Developing and productizing internal research is useful for driving improvements to the platform

- Scalability and acceleration are differentiating factors

- Surface NVIDIA technologies, and use bottlenecks to drive the development software and hardware improvements