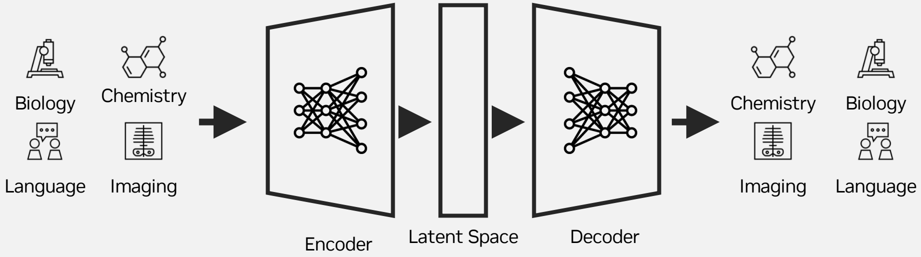
NVIDIA BioNeMo: A Framework and Service for Generative AI in Drug Discovery

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

6th RSC-BMCS / RSC-CICAG AI in Chemistry | 5th September, 2023

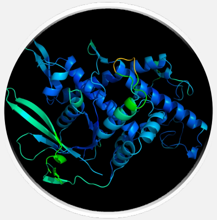
Language Models are Revolutionizing Discovery

- Information from biomedical literature
    - Named entity and relationship extraction
- Reaction prediction
    - Reaction and retrosynthesis prediction
    - Molecular optimization
- Property prediction
    - Sequence level
    - “Token” level (amino acid, motif, SMILES)
- Structure prediction and docking
    - Secondary structure analysis
    - Protein representation for model inputs

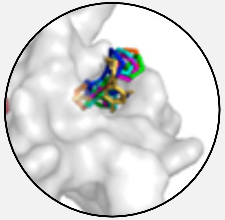


BIOMEDICAL NLP
Learn all of PubMed

GENERATIVE CHEMISTRY
Novel Drug Candidates



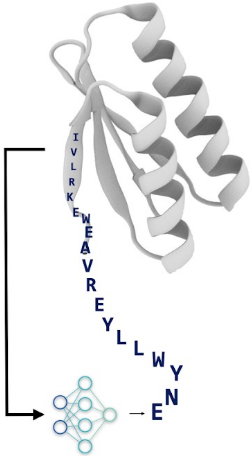
PROTEIN STRUCTURE
Predict 3D Structures



VIRTUAL SCREENING
Docking and Pose Prediction

From Sequence to 3D and Back Again

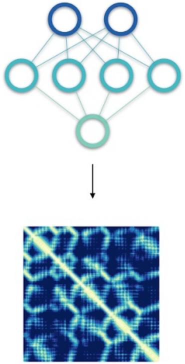
1 Fixed-backbone design



Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., III & Anandkumar, A. Dynamic-Backbone Protein-Ligand Structure Prediction with Multiscale Generative Diffusion Models. *arXiv [q-bio.QM]* (2022)

Verkuil, R. *et al.* Language models generalize beyond natural proteins. *bioRxiv* 2022.12.21.521521 (2022) doi:10.1101/2022.12.21.521521

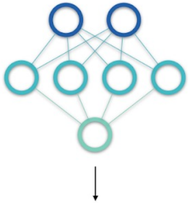
2 Structure Generation



Jing, B. *et al.* EigenFold: Generative protein structure prediction with diffusion models. *arXiv [q-bio.BM]* (2023)

Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* 1–4 (2023) doi:10.1038/s41592-022-01760-4

3 Sequence generation



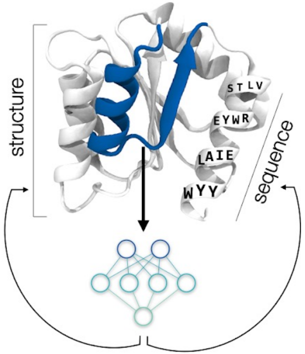
...MALKIPTHNHM...
...VFRDCEWS...
...WYIOPMNVGTDEW...

Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022)

Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. *arXiv [cs.LG]* (2022)

Munsamy, G., Lindner, S., Lorenz, P. & Ferruz, N. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes.

4 Sequence and structure design



Lisanza, S. L. *et al.* Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. *bioRxiv* 2023.05.08.539766 (2023) doi:10.1101/2023.05.08.539766

Jin, W., Wohlwend, J., Barzilay, R. & Jaakkola, T. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. *arXiv [q-bio.BM]* (2021)

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

Perspective on BioNeMo

Outline

- Overview of BioNeMo: Inference Service and Training Framework
- MolMIM: Development of a Small Molecule Foundation Model for Generation
- DiffDock Optimization: From Research to Enterprise Quality Software

Outline

- Overview of BioNeMo: Inference Service and Training Framework
- MolMIM: Development of a Small Molecule Foundation Model for Generation
- DiffDock Optimization: From Research to Enterprise Quality Software

Outline

- Overview of BioNeMo: Inference Service and Training Framework
- MolMIM: Development of a Small Molecule Foundation Model for Generation
- DiffDock Optimization: From Research to Enterprise Quality Software

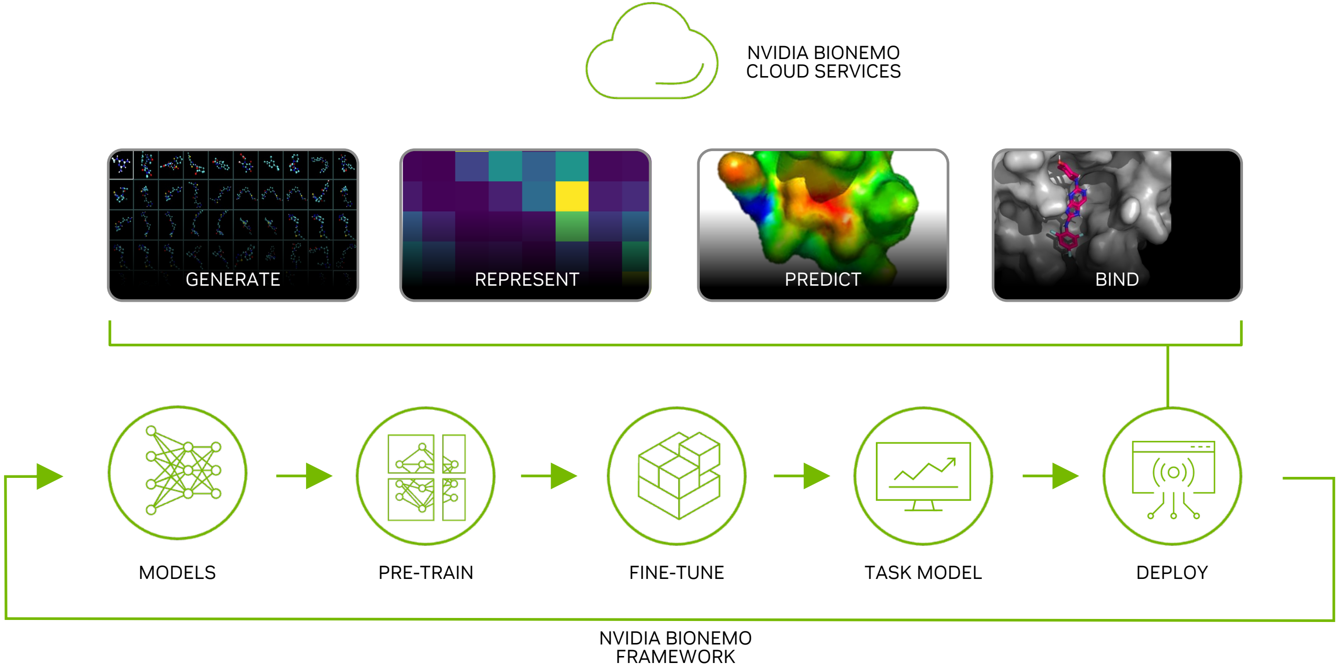
Outline

- Overview of BioNeMo: Inference Service and Training Framework
- MolMIM: Development of a Small Molecule Foundation Model for Generation
- DiffDock Optimization: From Research to Enterprise Quality Software

BioNeMo Overview: Inference Service and Framework

NVIDIA BioNeMo

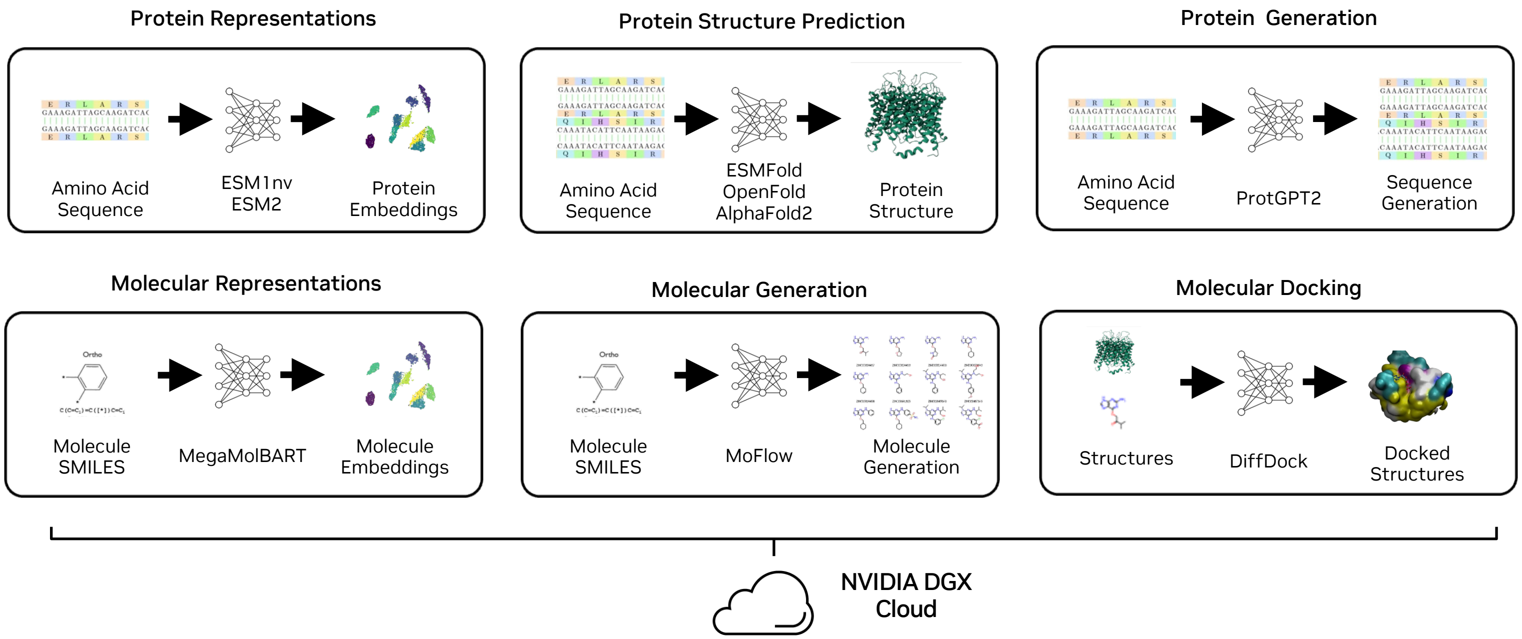
AI Tools, Frameworks, and Applications for Drug Discovery



absci · ALCHEMAB THERAPEUTICS · AMGEN · astellas · AstraZeneca · BROAD INSTITUTE · Deloitte. · EVOZYNE NATURAL MACHINES · Flagship Pioneering · Innophore

Insilico Medicine · InstaDeep™ · MAYO CLINIC · Meta · Mila · OpenFold Democratizing AI for Biology · Relation · ROSTLAB. · VYASA

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

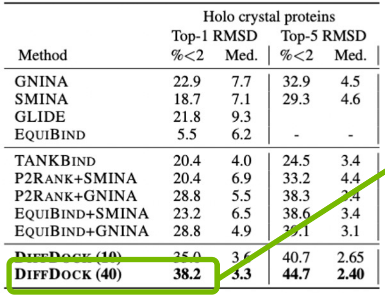
Molecular Docking

Structures → DiffDock → Docked Structures

NVIDIA DGX Cloud

Life Cycle of a BioNeMo Model in the Inference Service

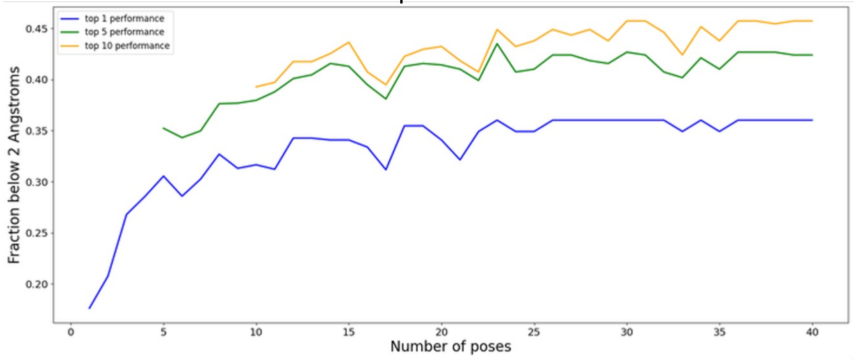
- Model checkpoints are accelerated using a variety of NVIDIA tools – standard tricks to custom CUDA kernels
- All quantitative and qualitative results are reproduced
- For DiffDock, the RMSD metrics were reproduced under a variety of different conditions



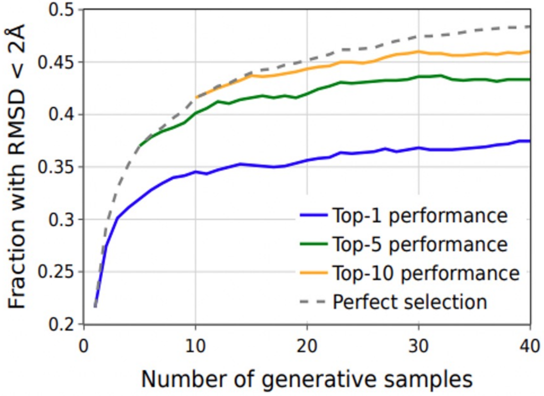
	Holo crystal proteins			
	Top-1 RMSD		Top-5 RMSD	
Method	%<2	Med.	%<2	Med.
GNINA	22.9	7.7	32.9	4.5
SMINA	18.7	7.1	29.3	4.6
GLIDE	21.8	9.3		
EQUIBIND	5.5	6.2	-	-
TANKBIND	20.4	4.0	24.5	3.4
P2RANK+SMINA	20.4	6.9	33.2	4.4
P2RANK+GNINA	28.8	5.5	38.3	3.4
EQUIBIND+SMINA	23.2	6.5	38.6	3.4
EQUIBIND+GNINA	28.8	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

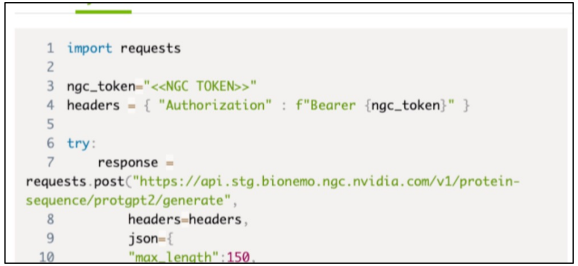


Publication



Life Cycle of a BioNeMo Model in the Inference Service

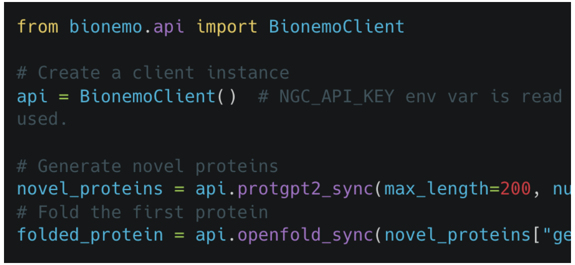
- API and Python interface developed



```
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,
```

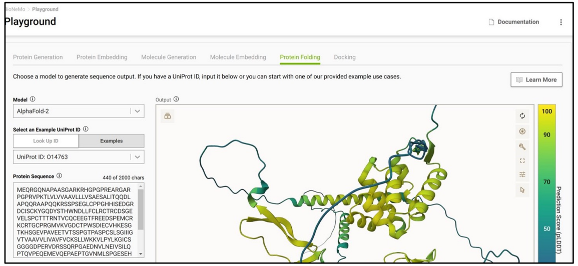


```
from bionemo.api import BionemoClient

# 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
```

- Interactive UI and example Jupyter notebooks



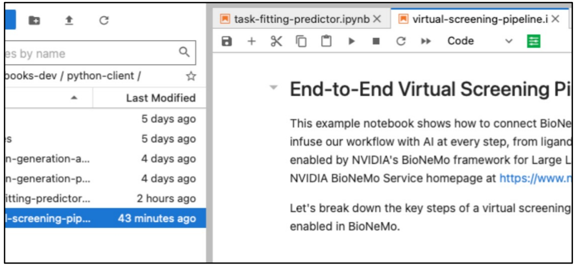
Choose a model to generate sequence output. If you have a UniProt ID, input it below or you can start with one of our provided example use cases.

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
```



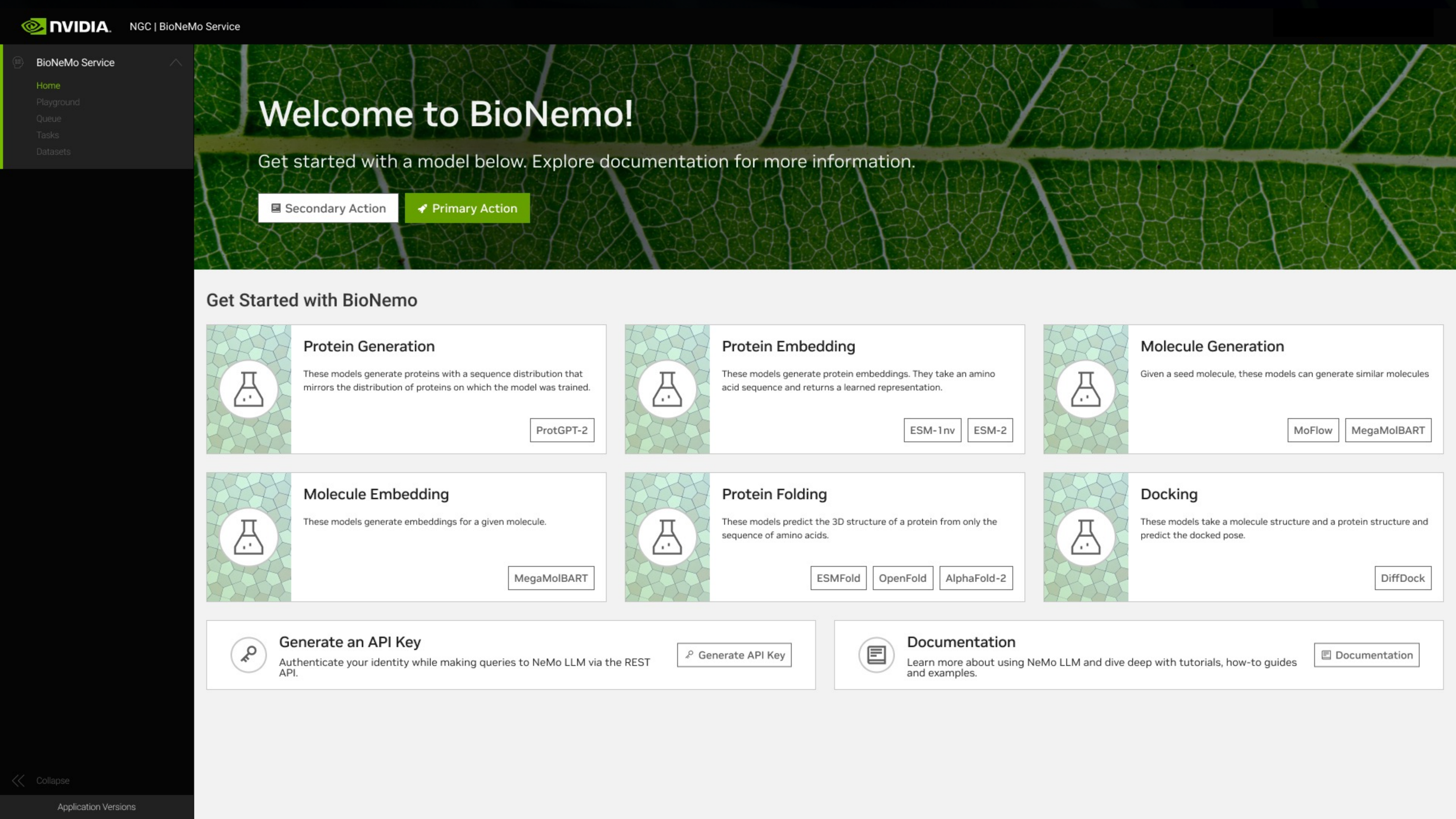
ooks-dev / python-client /

	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.



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

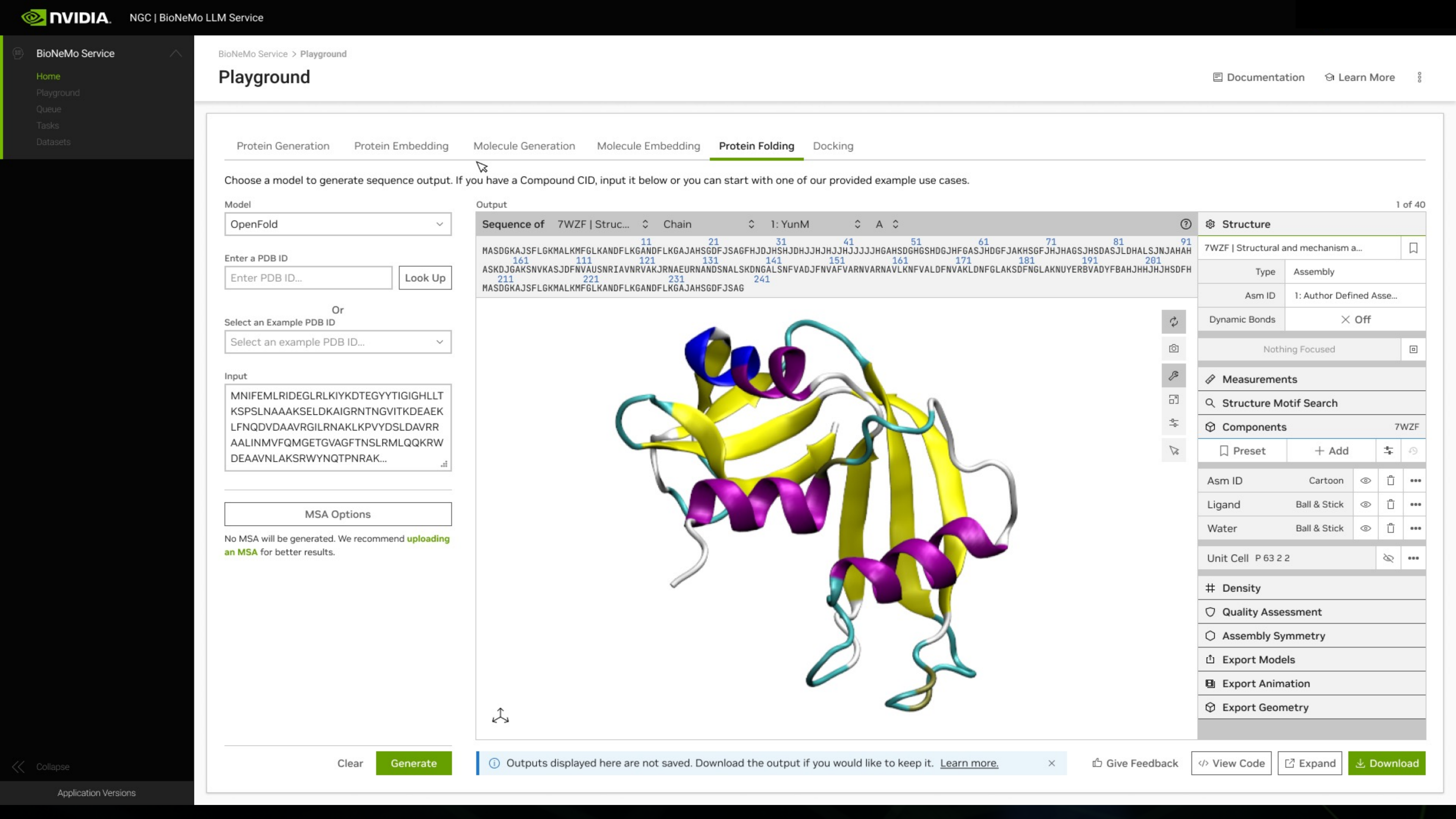
These models take a molecule structure and a protein structure and predict the docked pose.

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.



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:

MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAK...

MSA Options

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

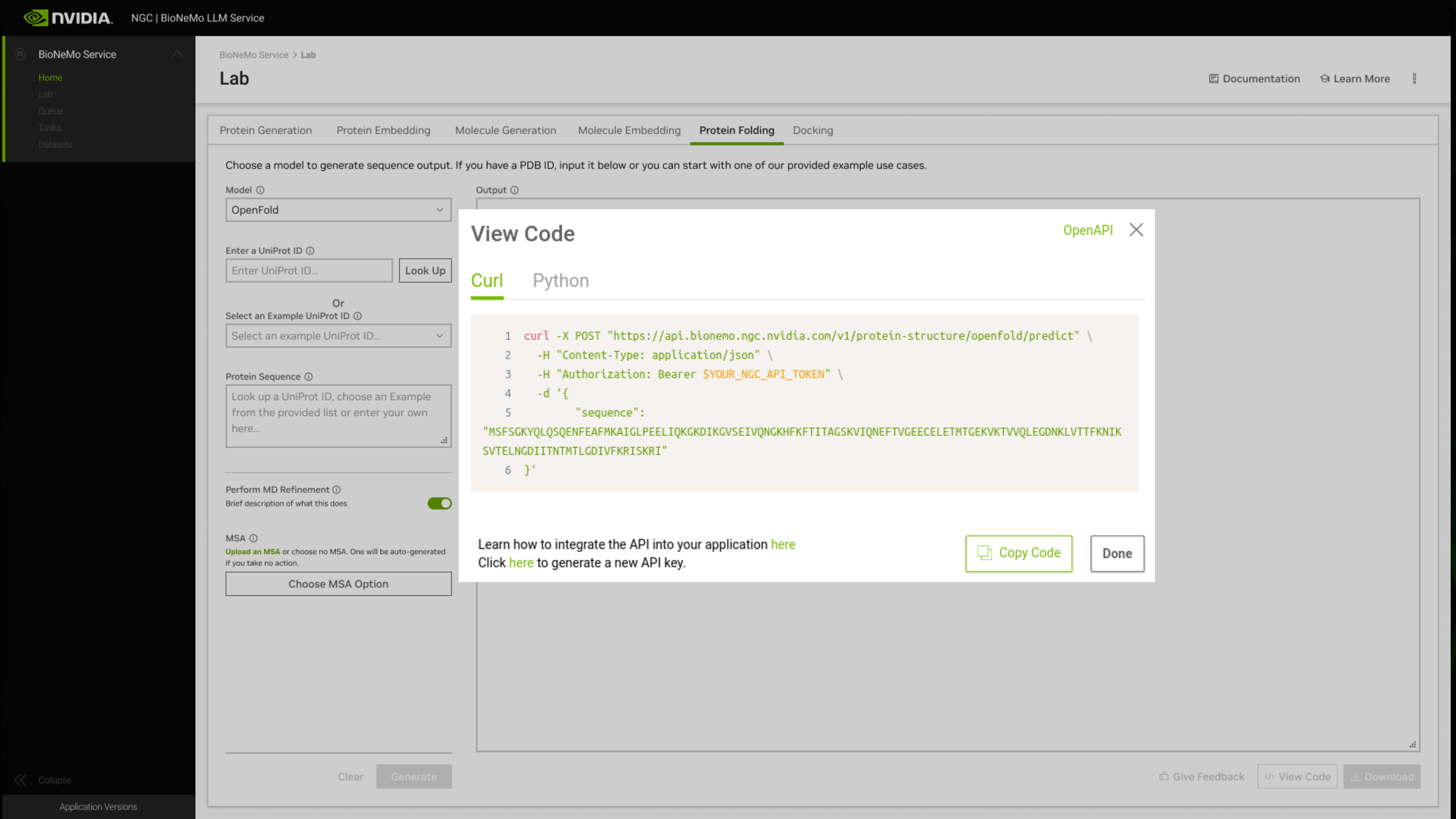
7WZF | Structural and mechanism a...

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.



Lab

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...

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

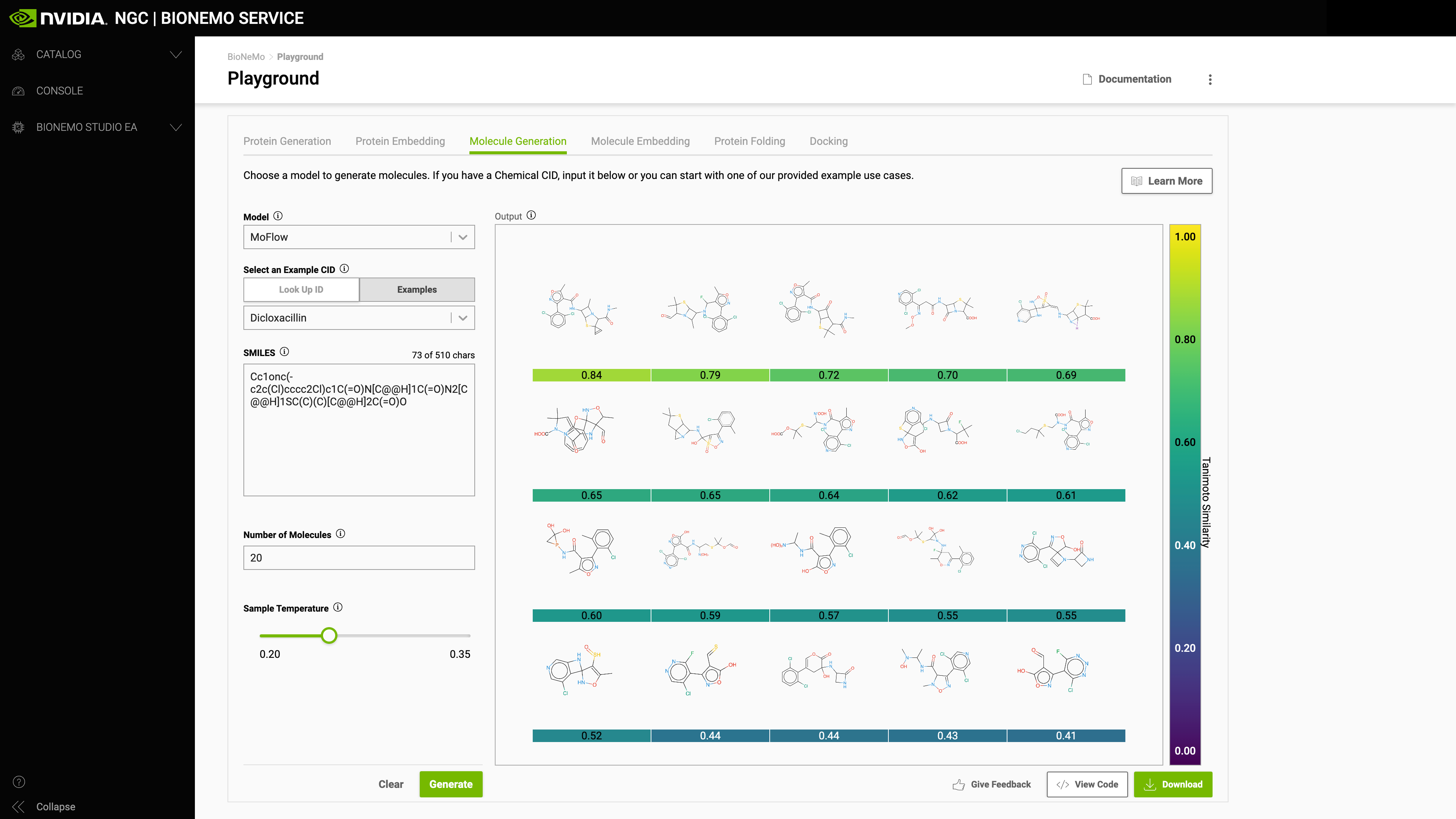
MSA
Upload an MSA or choose no MSA. One will be auto-generated if you take no action.

Output

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":
"MSFSGKYQLQSQENFEAFMKAIGLPEELIQKGKDIKGVSEIVQNGKHFKFTITAGSKVIQNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTTFKNIKSVTELNGDIITNTMTLGDIVFKRISKRI"
}'
```

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

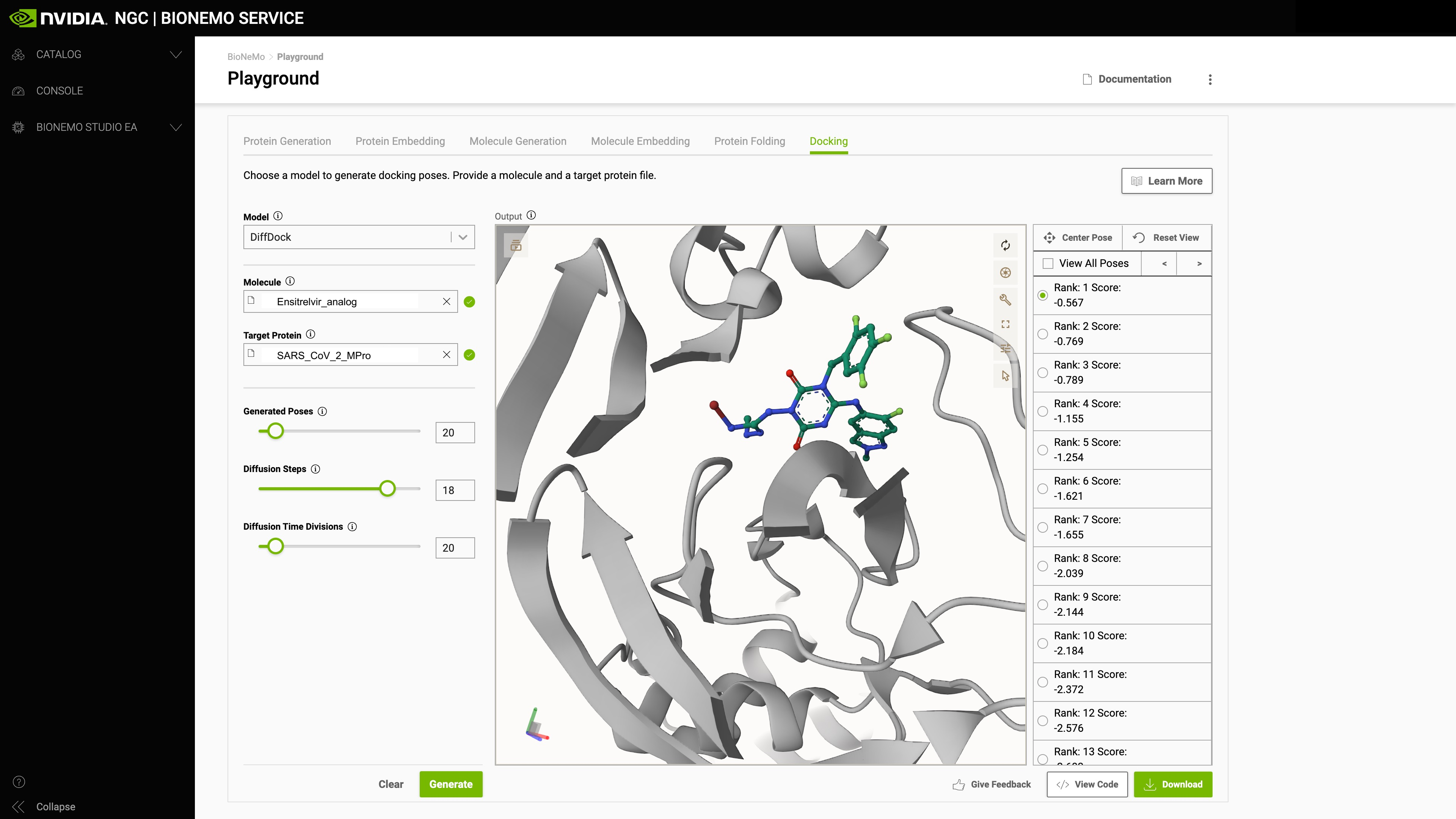
Cc1onc(-c2c(Cl)cccc2Cl)c1C(=O)N[C@@H]1C(=O)N2[C@@H]1SC(C)(C)[C@@H]2C(=O)O

Number of Molecules ⓘ

20

Sample Temperature ⓘ

0.20 0.35



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

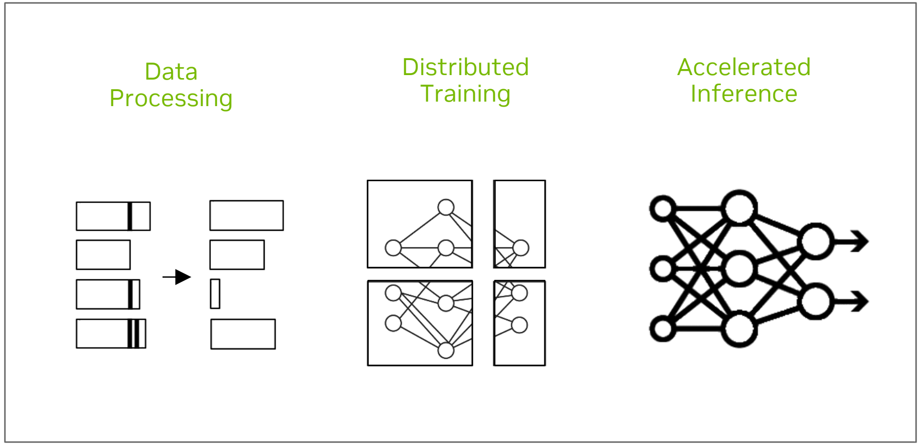
Diffusion Steps: 18

Diffusion Time Divisions: 20

□ View All Poses < >

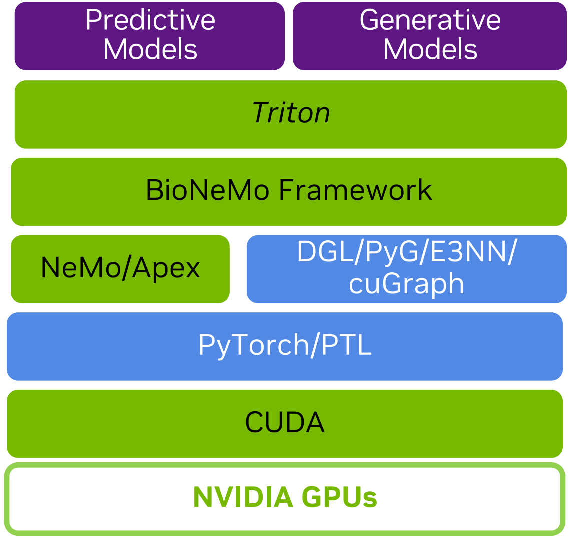
● Rank: 1 Score: -0.567
○ Rank: 2 Score: -0.769
○ Rank: 3 Score: -0.789
○ Rank: 4 Score: -1.155
○ Rank: 5 Score: -1.254
○ Rank: 6 Score: -1.621
○ Rank: 7 Score: -1.655
○ Rank: 8 Score: -2.039
○ Rank: 9 Score: -2.144
○ Rank: 10 Score: -2.184
○ Rank: 11 Score: -2.372
○ Rank: 12 Score: -2.576
○ Rank: 13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

BioNeMo Framework Overview



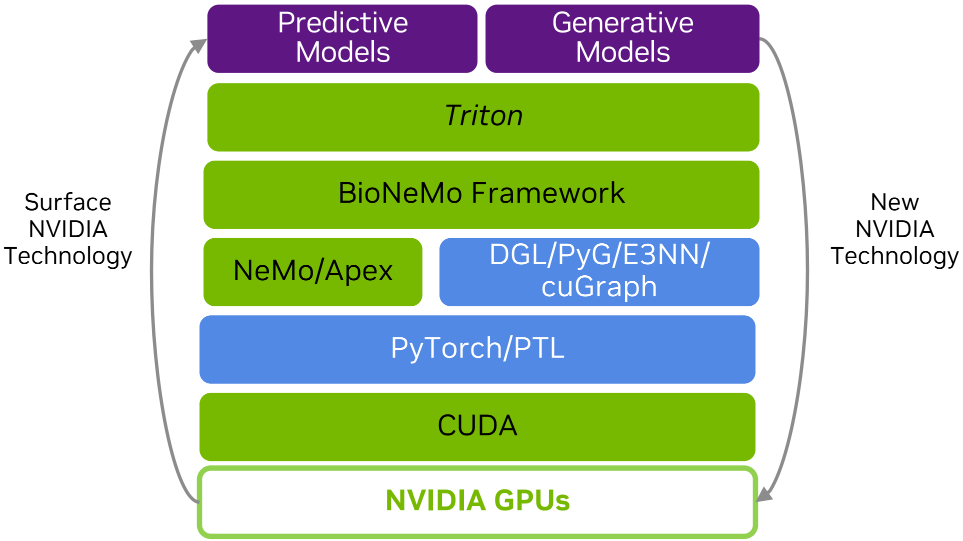
- Includes dataset process, model pre-training optional fine tuning, and example downstream tasks
- gRPC based class for inference and example notebook – automated deployment coming
- **Currently:** three LLM models for cheminformatics and protein applications (MegaMolBART, ESM1, ProtT5)
- Additional models in development
  - *LLM:* ESM-2, nucleic acid models, **MolMIM**
  - *Equivariant:* EquiDock, OpenFold, **DiffDock**

BioNeMo Framework Technology Stack



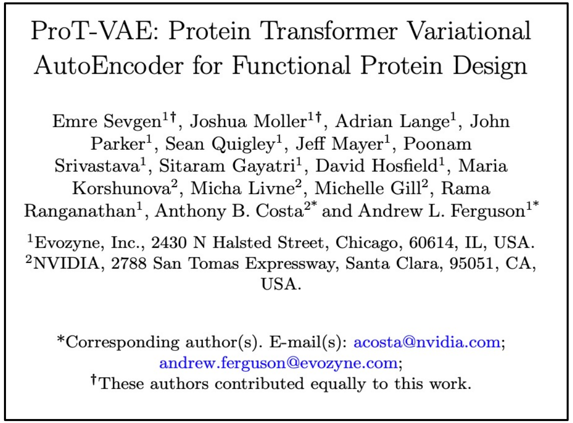
- Based on NVIDIA NeMo, which is a library for development and training of LLMs (as well as text-to-speech, etc.)
  - Provides support for multi-GPU and multi-node training
  - Data parallelism supported
  - Model parallelism supported for all LLMs: tensor parallelism, pipeline parallelism, and sequence parallelism
- Automated deployment with Triton is coming

BioNeMo Framework Technology Stack



- Based on NVIDIA NeMo, which is a library for development and training of LLMs (as well as text-to-speech, etc.)
  - Provides support for multi-GPU and multi-node training
  - Data parallelism supported
  - Model parallelism supported for all LLMs: tensor parallelism, pipeline parallelism, and sequence parallelism
- Automated deployment with Triton is coming
- Surface and develop new software and hardware technology

Proteins Generated from Evozyne's ProT-VAE Models

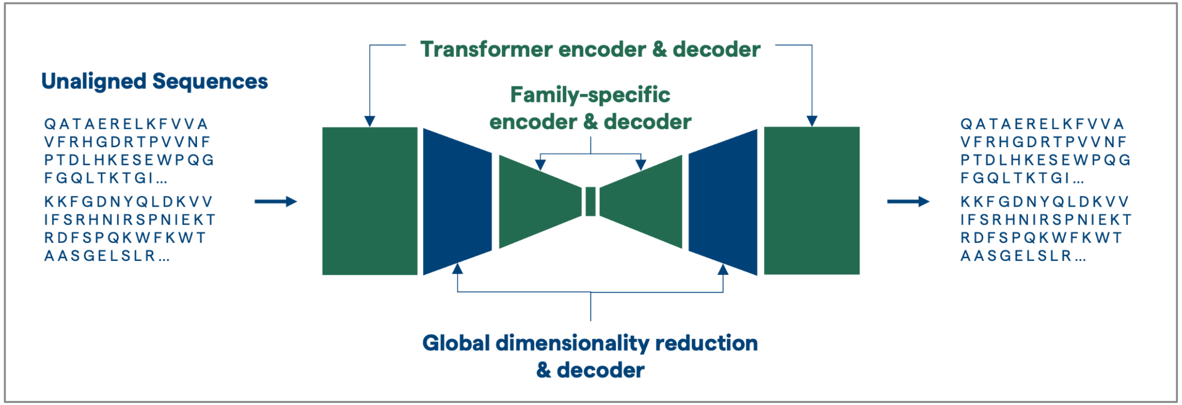


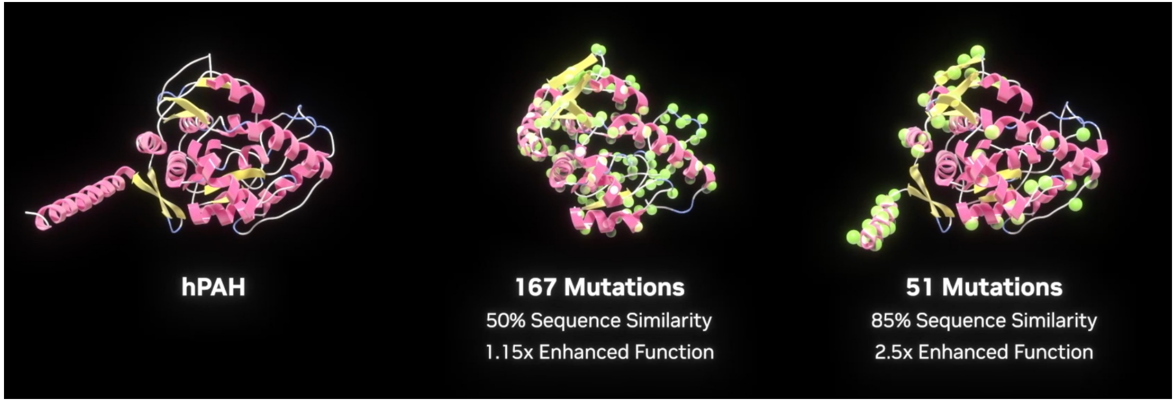
ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design

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*]

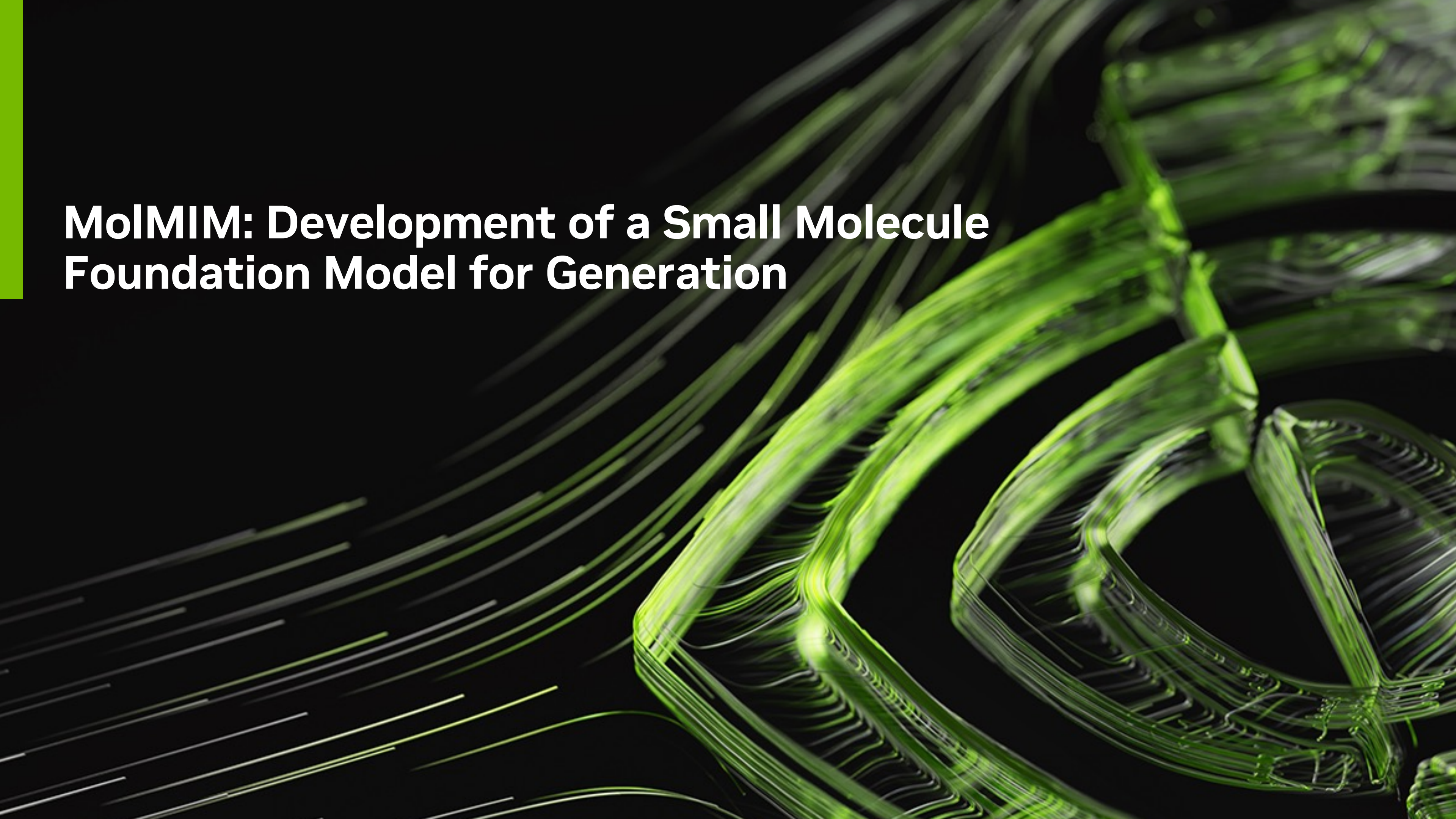
[1]Evozyne, Inc., 2430 N Halsted Street, Chicago, 60614, IL, USA.
[2]NVIDIA, 2788 San Tomas Expressway, Santa Clara, 95051, CA, USA.

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



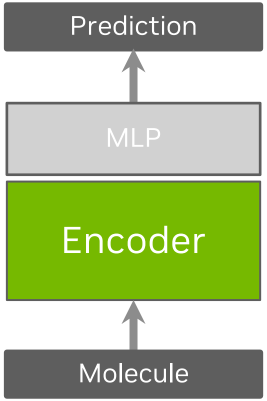


MolMIM: Development of a Small Molecule Foundation Model for Generation



Cheminformatics Foundation Model Objectives

Representation and Translation

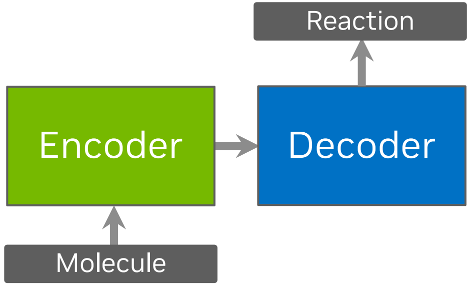


Prediction

MLP

Encoder

Molecule



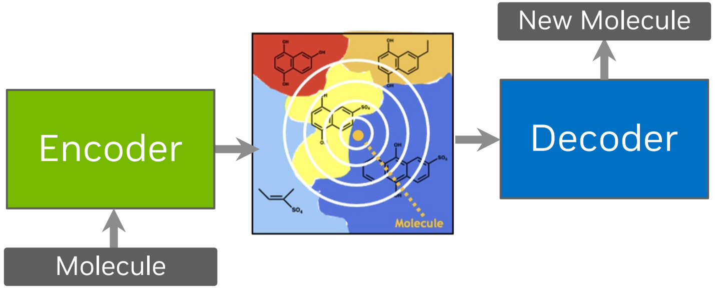
Reaction

Encoder

Decoder

Molecule

Generation

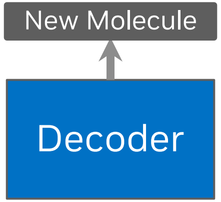


New Molecule

Encoder

Decoder

Molecule

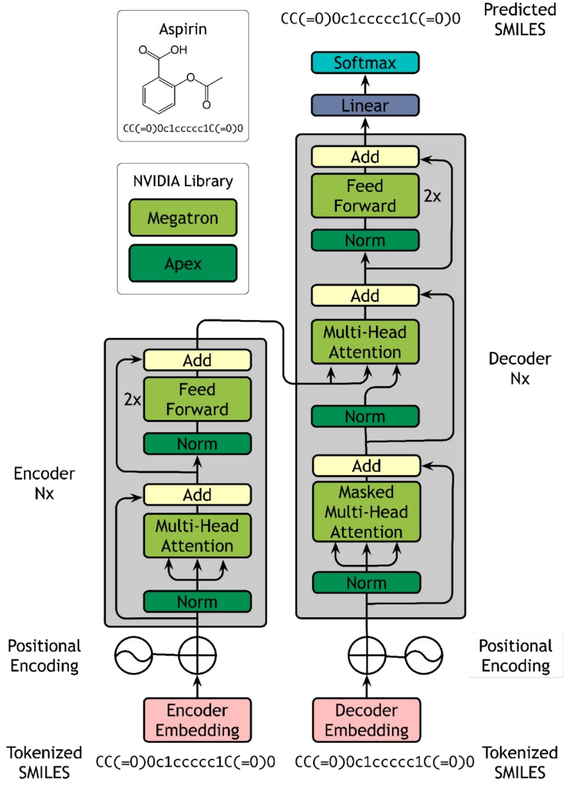


New Molecule

Decoder

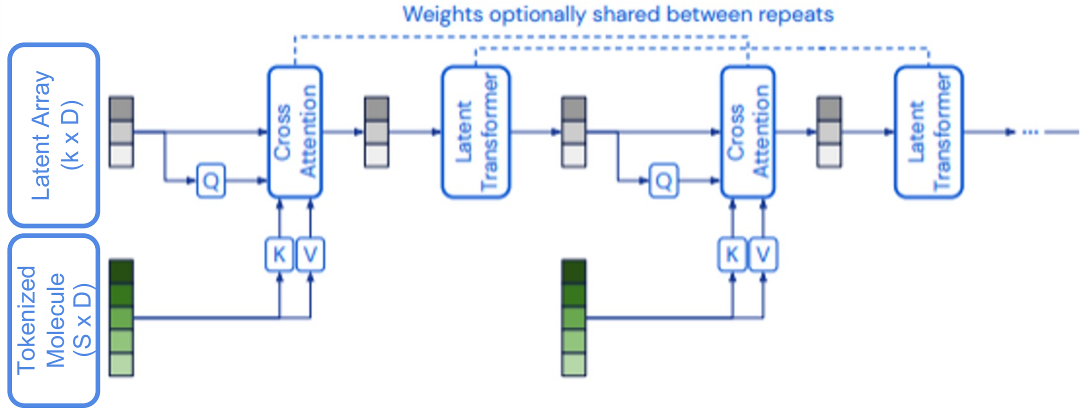
MegaMolBART Molecule Representations

- MegaMolBART developed in collaboration with AstraZeneca, based on published model called Chemformer
- BART model – encoder trained with MLM and autoregressive decoder on 1.5B molecules from ZINC15
- Useful for small molecule representations and sequence translation tasks
- **Challenges with using MegaMolBART for molecule generation:**
- Size of encoder output is variable -- based on number of tokens
- Lacks an organized, smooth latent space



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

Development of a Seq2Seq Model with Fixed Size Latent Dimension



k = Perceiver dimension

- Perceiver encoder utilizes cross-attention to create a fixed size latent space
- Perceiver model has a fixed size representation (k)
- Runtime complexity for the perceiver is O(Sk + k^2), compared to O(S^2) for the transformer
- Perceiver BART was trained on 750M molecules from ZINC15

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

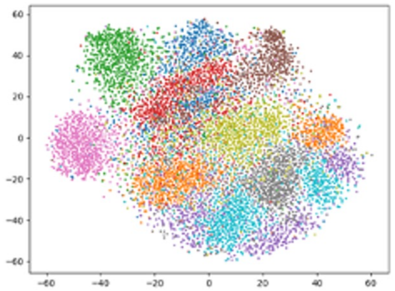
- Utilizes same architecture as VAE

- MIM loss results in a clustered space while KL divergence loss smooths the latent space resulting in blurring

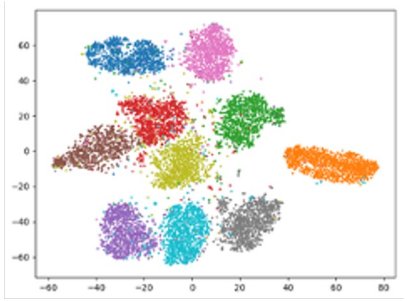
- Important: MIM makes no guarantees about cluster organization

- Developed a MolVAE and MolMIM model and trained both on 750M molecules from ZINC15

VAE



MIM



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

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†

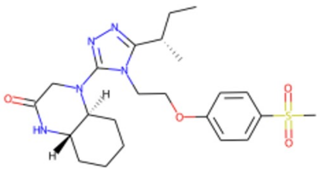
†CDDD decoding speed limited by batch size.

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

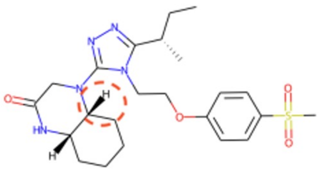
MolMIM – Sampling Distance Can Be Tuned for Similarity

Small Perturbations

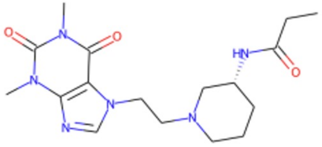
Larger Perturbations



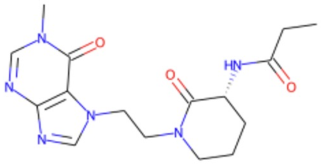
Seed Molecule



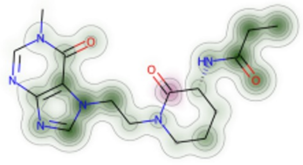
Sampled Molecule



Seed Molecule

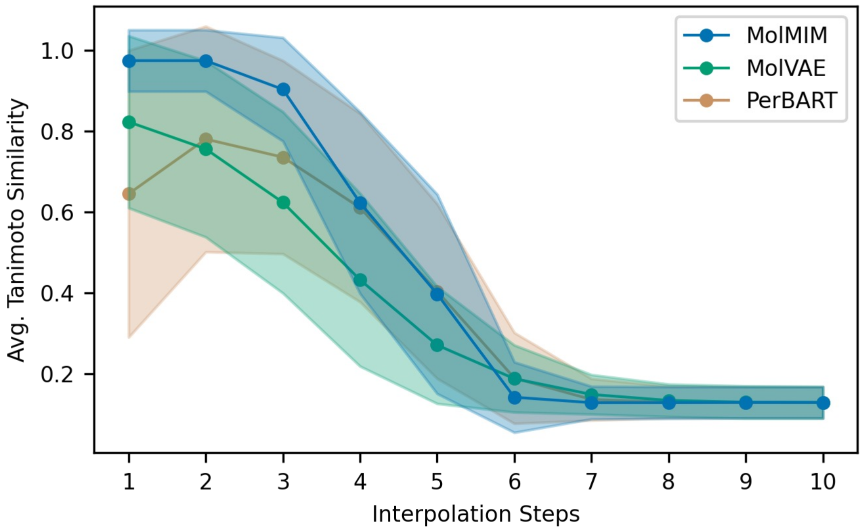


Sampled Molecule



Similarity Map

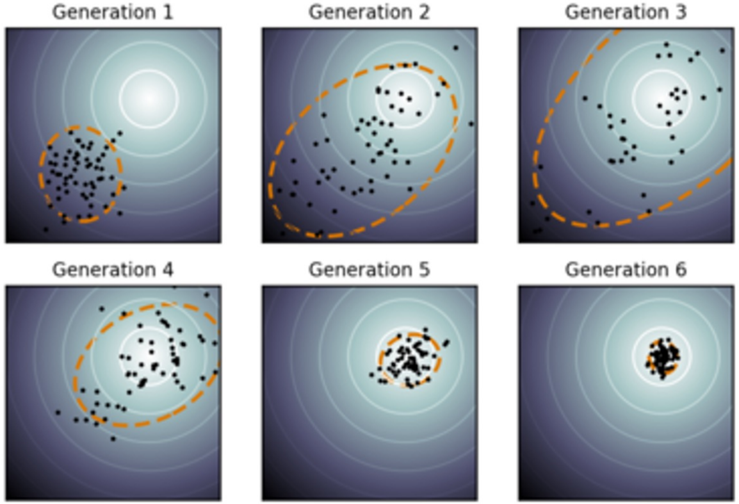
Probing Latent Structure by Molecule Interpolation



- 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

Measuring the Controllability of MolMIM

- **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).

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

Multi-Objective Property Optimization

- Performed multi-objective molecule optimization to jointly optimize QED ≥ 0.6, SA ≤ 4.0, JNK3 ≥ 0.5, GSK4β ≥ 0.5
- Novelty is proportion of molecules with $\delta \leq 0.4$ relative to any molecule in active set
- Diversity is the mean pairwise Tanimoto similarity across all compounds

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

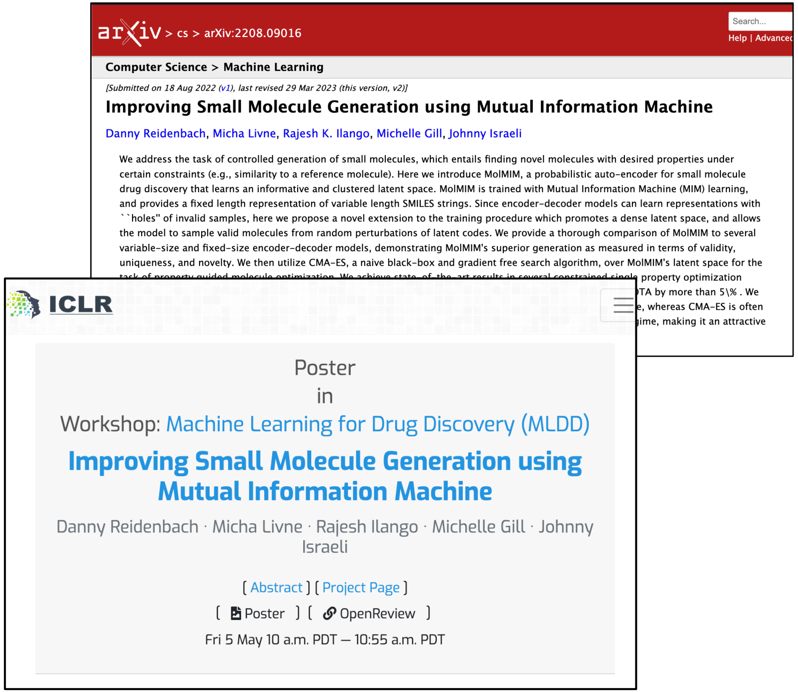
Multi-Objective Property Optimization

- Performed multi-objective molecule optimization to jointly optimize QED ≥ 0.6, SA ≤ 4.0, JNK3 ≥ 0.5, GSK4β ≥ 0.5
- Novelty is proportion of molecules with δ ≤ 0.4 relative to any molecule in active set
- Diversity is the mean pairwise Tanimoto similarity across all compounds
- Optimization types:
  - *Random*: 2,000 ZINC15 test set molecules
  - *Approximate*: 551 molecules that satisfy QED ∊ [0.25, 0.4]; JNK3 and GSK4β ∊ [0.25, 0.35]
  - *Exemplar*: 741 molecules that satisfy success criteria
  - †With Tanimoto similarity ≥ 0.4
- 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

gime, 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
- *Wishlist*: more relevant and comprehensive benchmarks – want to collaborate?

DiffDock Optimization: From Research to Enterprise Quality Software

DiffDock for Diffusion-Based Docking Pose Generation

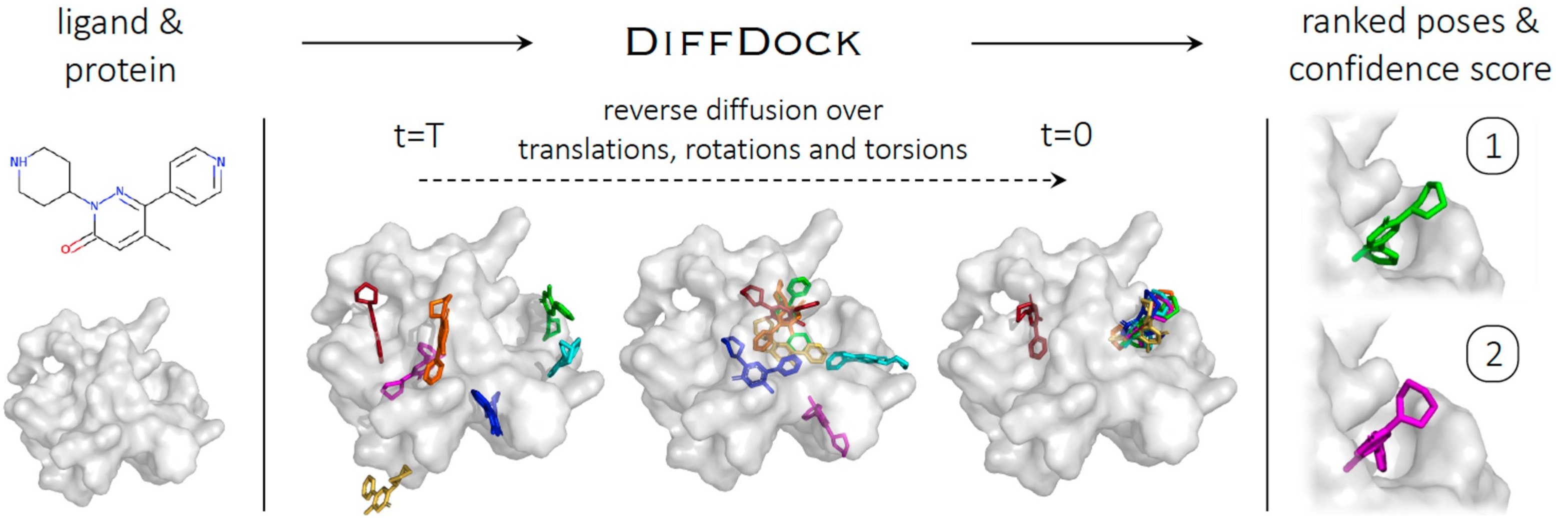
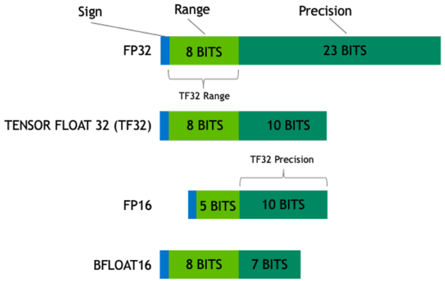
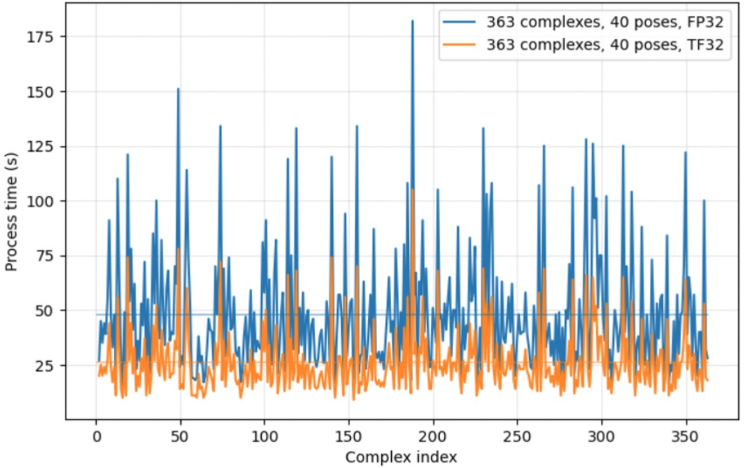


IMAGE: https://github.com/gcorso/DiffDock
G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, Arxiv (2022).

GPU Specific Optimization of DiffDock with TF32





- Reducing numerical precision is a common method of accelerating both training and inference, e.g. FP32 → FP16
- However lower precision formats are more susceptible to overflows and can lead to numerical instabilities
- NVIDIA A100 GPUs support a math mode called TensorFloat32 (TF32), which strikes a balance between precision and performance
- Converting DiffDock weights to TF32 required changing one line of code and provided 1.8x speed up of inference, with no impact on benchmarked accuracy
- Similar optimizations are being tested with model training

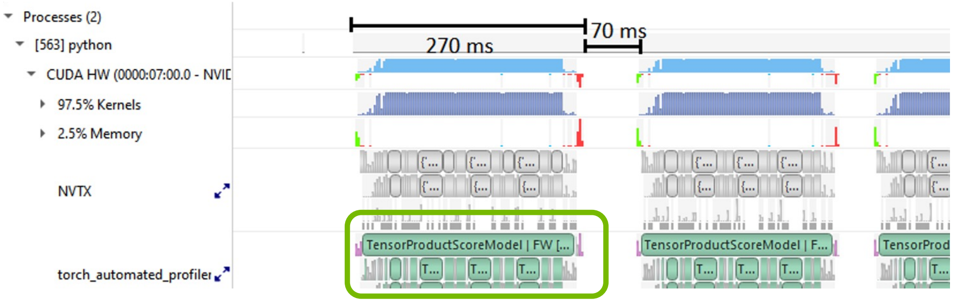
TF32 format: https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format/

Optimization of DiffDock Mathematical Operations

- DiffDock is an equivariant model, data are represented in spherical basis
- One forward pass requires many multiplications involving irreducible representations of a given symmetry group, *e.g.* rigid rotations in 3D
- The tensor product operations are from the e3nn library and comprise a considerable part of computation time (see profile, green circle)
- BioNeMo includes a version of e3nn which has been accelerated with CUDA parallelism
- Profiling reveals other opportunities – data operations and other methods to maximize GPU use

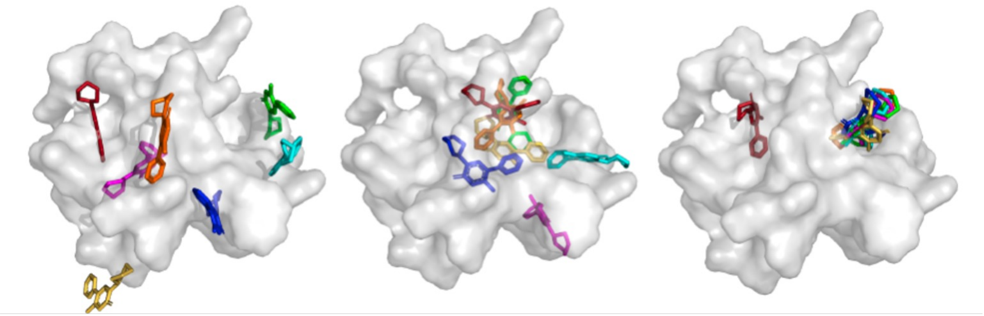
$$\mathbf{h}_a \leftarrow \mathbf{h}_a \bigoplus_{t\in\{\ell,r\}} \mathbf{BN}^{(t_a,t)}\left(\frac{1}{|\mathcal{N}_a^{(t)}|}\boxed{\sum_{b\in\mathcal{N}_a^{(t)}} Y(\hat{r}_{ab}) \otimes_{\psi_{ab}} \mathbf{h}_b}\right)$$

$$\text{with } \psi_{ab} = \Psi^{(t_a,t)}(e_{ab}, \mathbf{h}_a^0, \mathbf{h}_b^0)$$

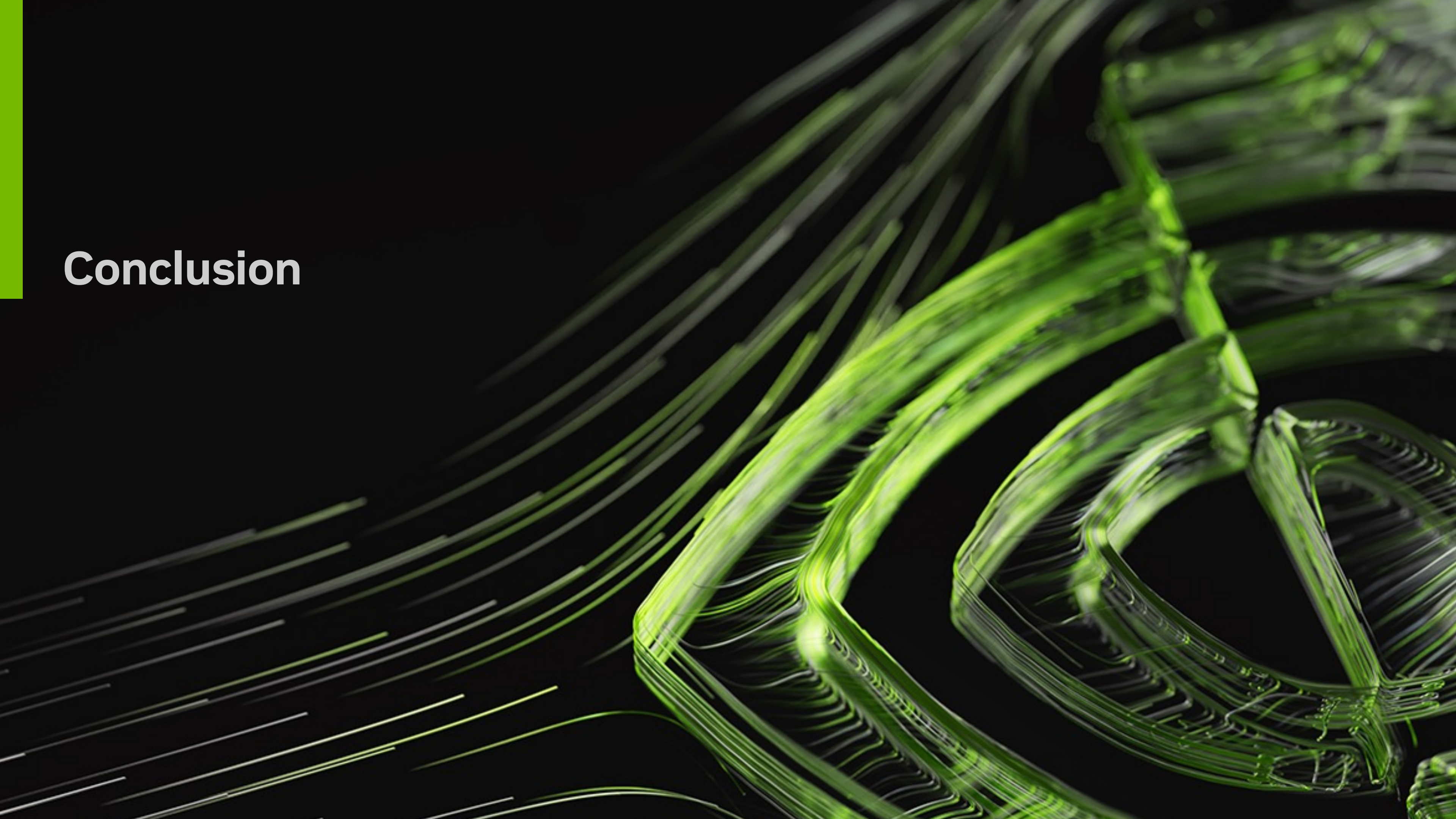


DiffDock: Research to Productization

- MD-assisted refinement of docked poses
- Dataset extension and management
- Drive research and development of accelerate compute functionality for equivariant models



Conclusion



Conclusions

- BioNeMo is a framework and inference service for developing, training, deploying, and using deep learning models and tools for drug discovery
- BioNeMo surfaces NVIDIA hardware and software improvements relevant to life sciences and drives future development
- MolMIM is a cheminformatics model trained on only SMILES with a structured latent space and fixed size embedding for molecule design
- DiffDock acceleration and improvements in numerical stability drive future equivariant model optimizations
- BioNeMo framework open beta coming soon, enroll in service GA here: https://www.nvidia.com/bionemo

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

Hani-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

Steven Kothen-Hill

Tomasz Grzegorzek

Timur Rvachov

Yuxing Peng

Zachary McClure

Contact me: mgill@nvidia.com