



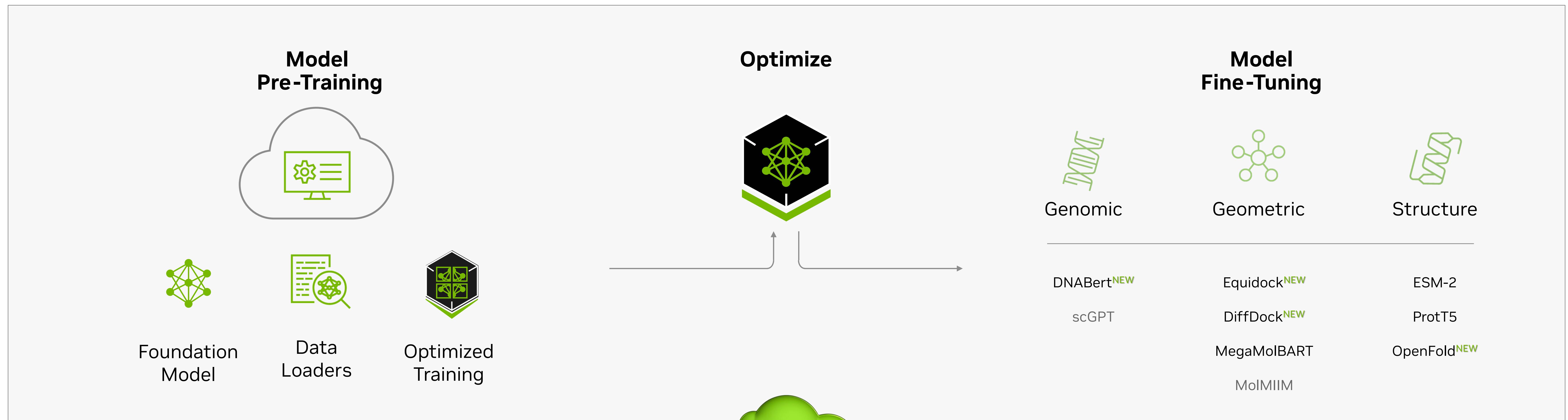
NVIDIA BioNeMo Hands-on Workshop

Colleen Ruan, PhD.,
Senior Deep Learning Solution Architect, Healthcare
2024/6/5

Introduction to BioNeMo Framework

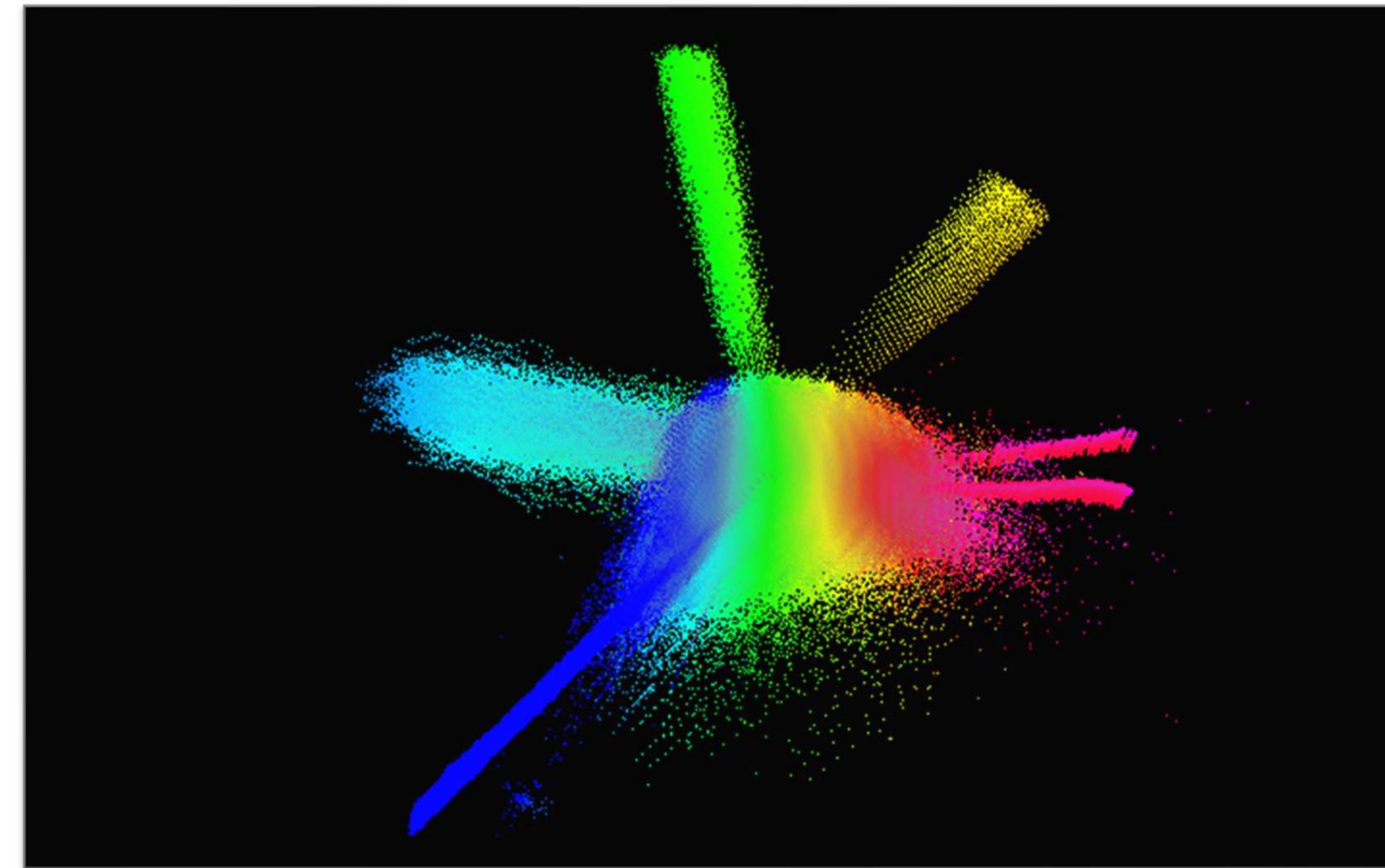
NVIDIA BioNeMo Framework

Enables Data Scientists and Researchers Train on DNA, Protein, Chemistry Data

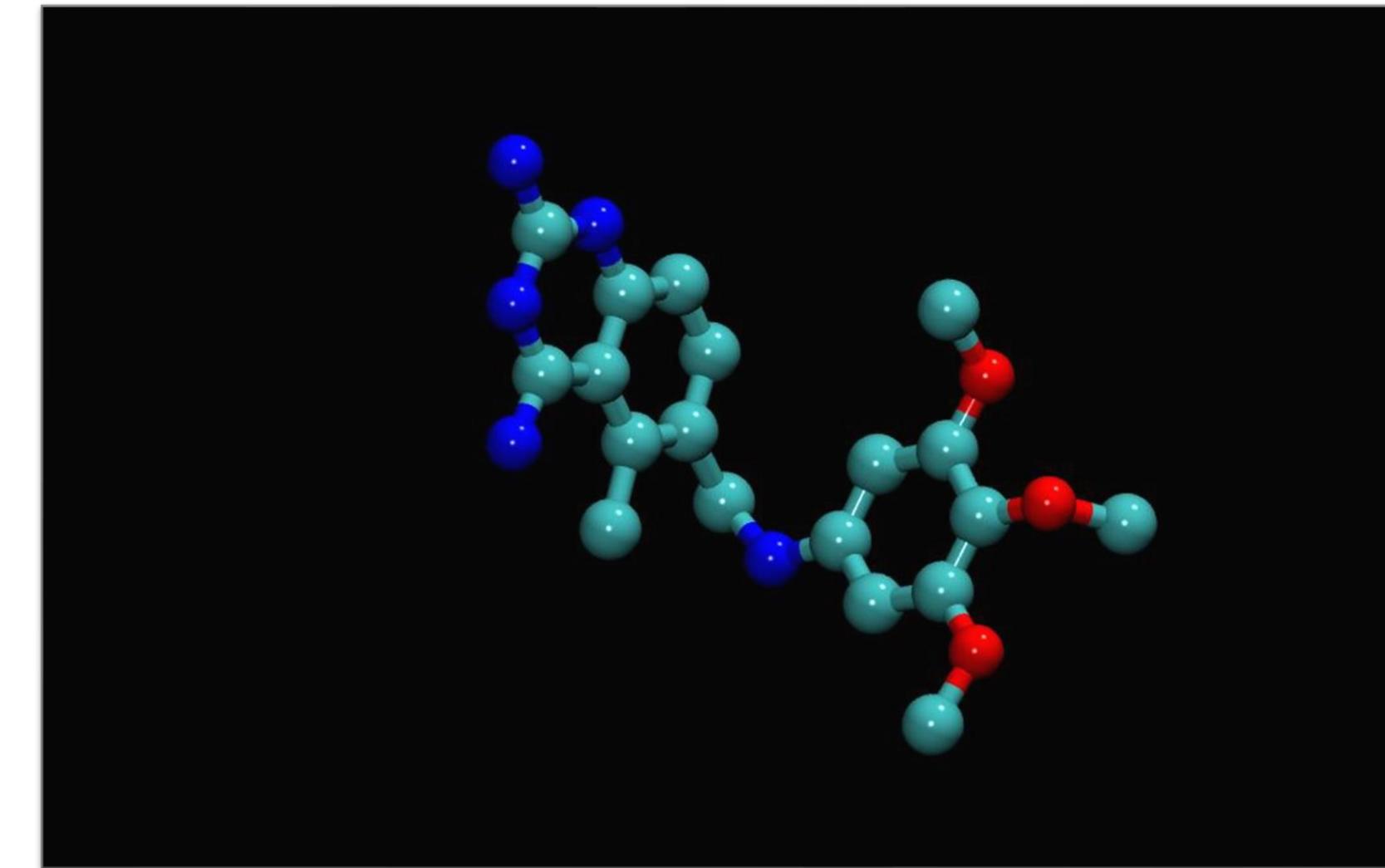


BioNeMo Framework Supports Optimized Biomolecular Models

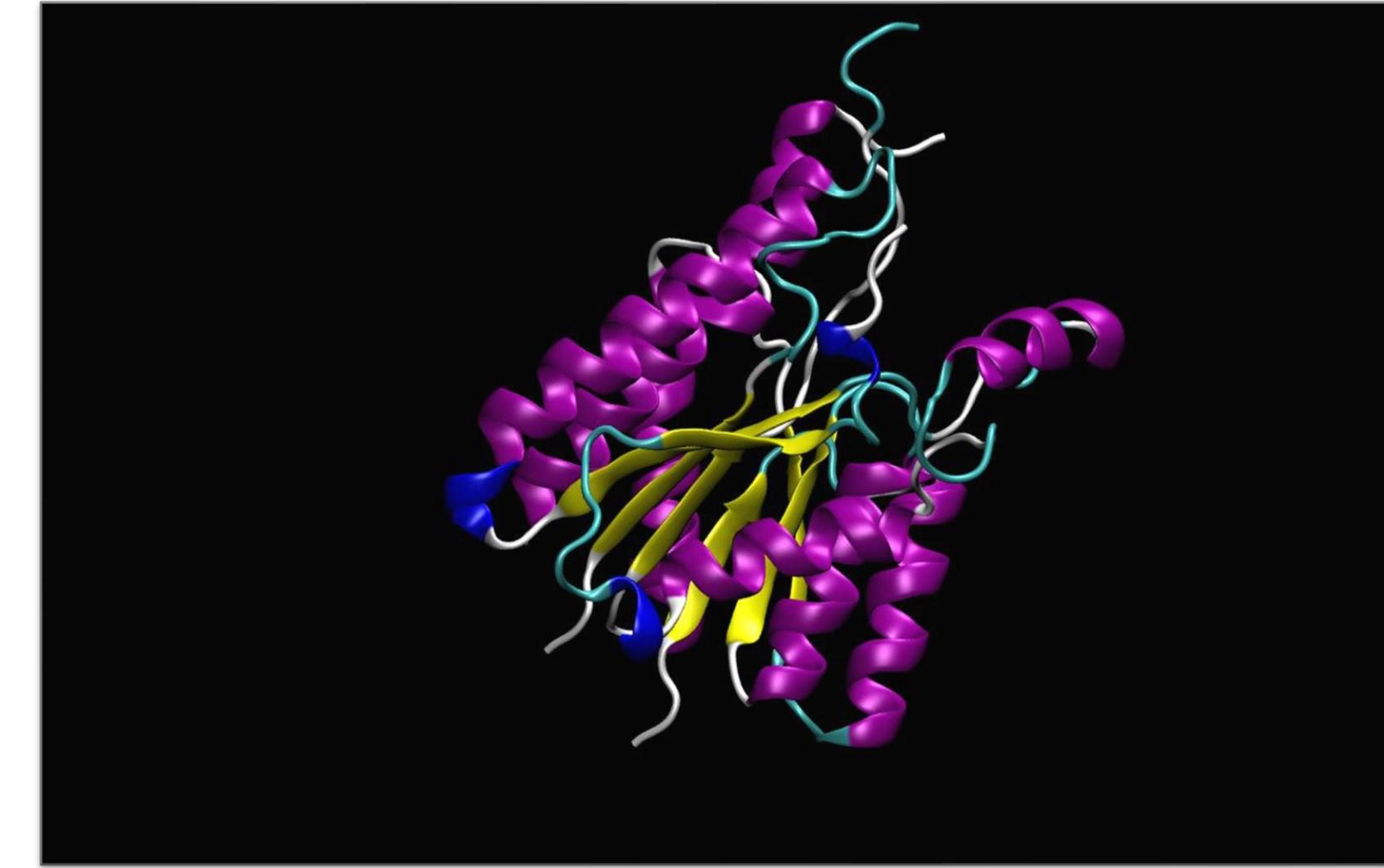
Proteins | Small Molecules | Genomics



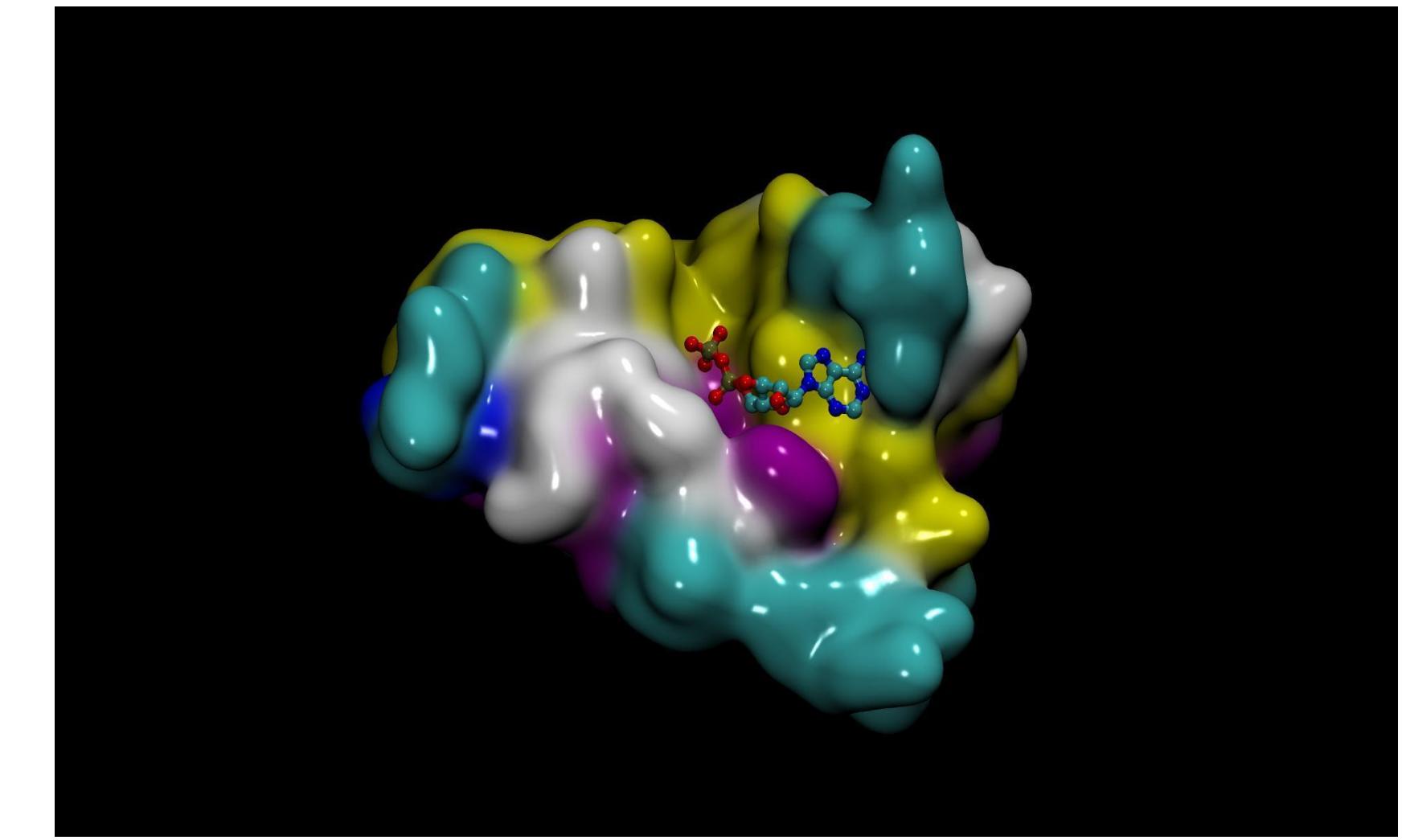
ESM-1 | ESM-2
Protein LLMs



MegaMolBART
Generative Chemistry Model



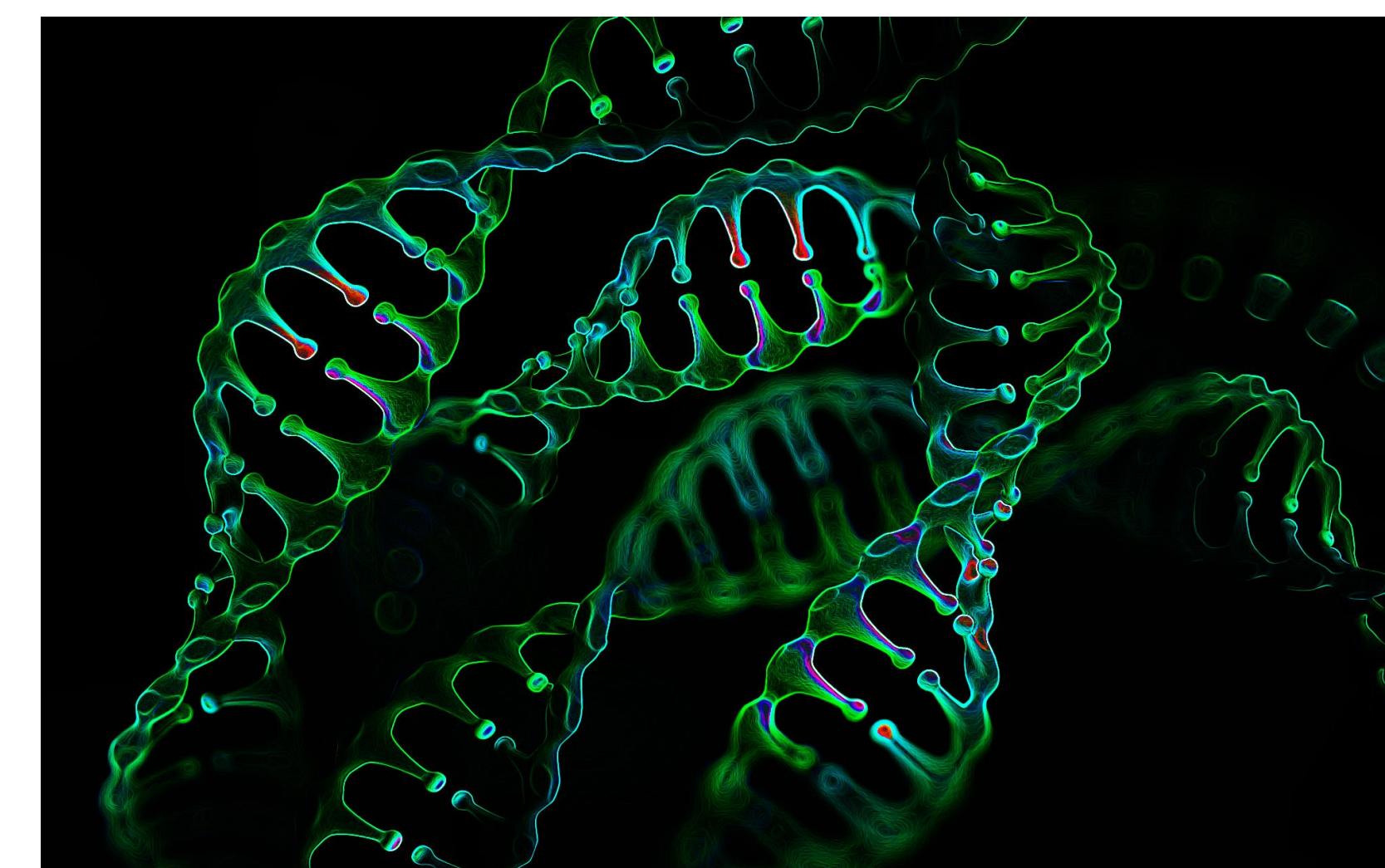
ProtT5
Protein Sequence Generation



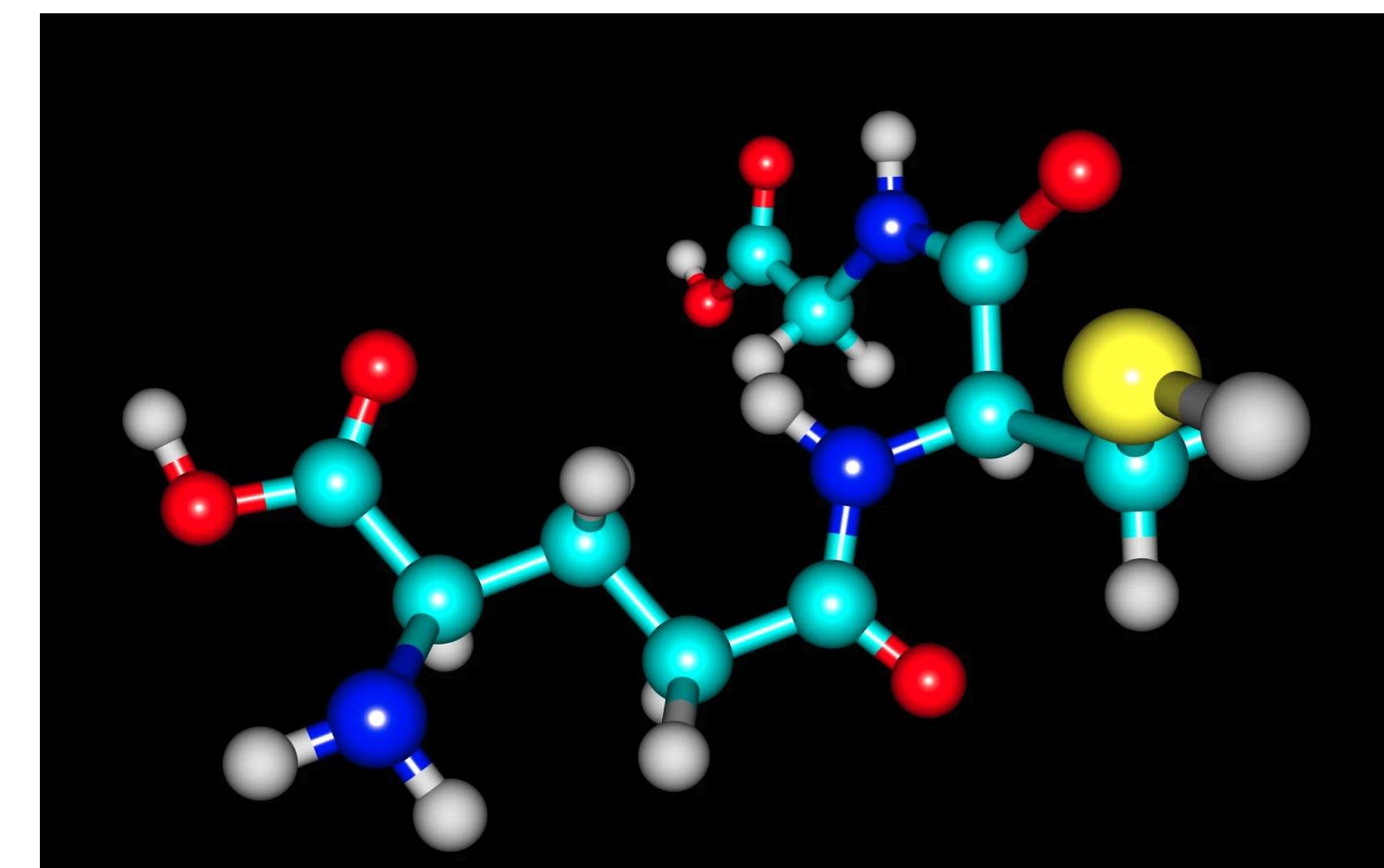
NEW: DiffDock | EquiDock
Docking Prediction



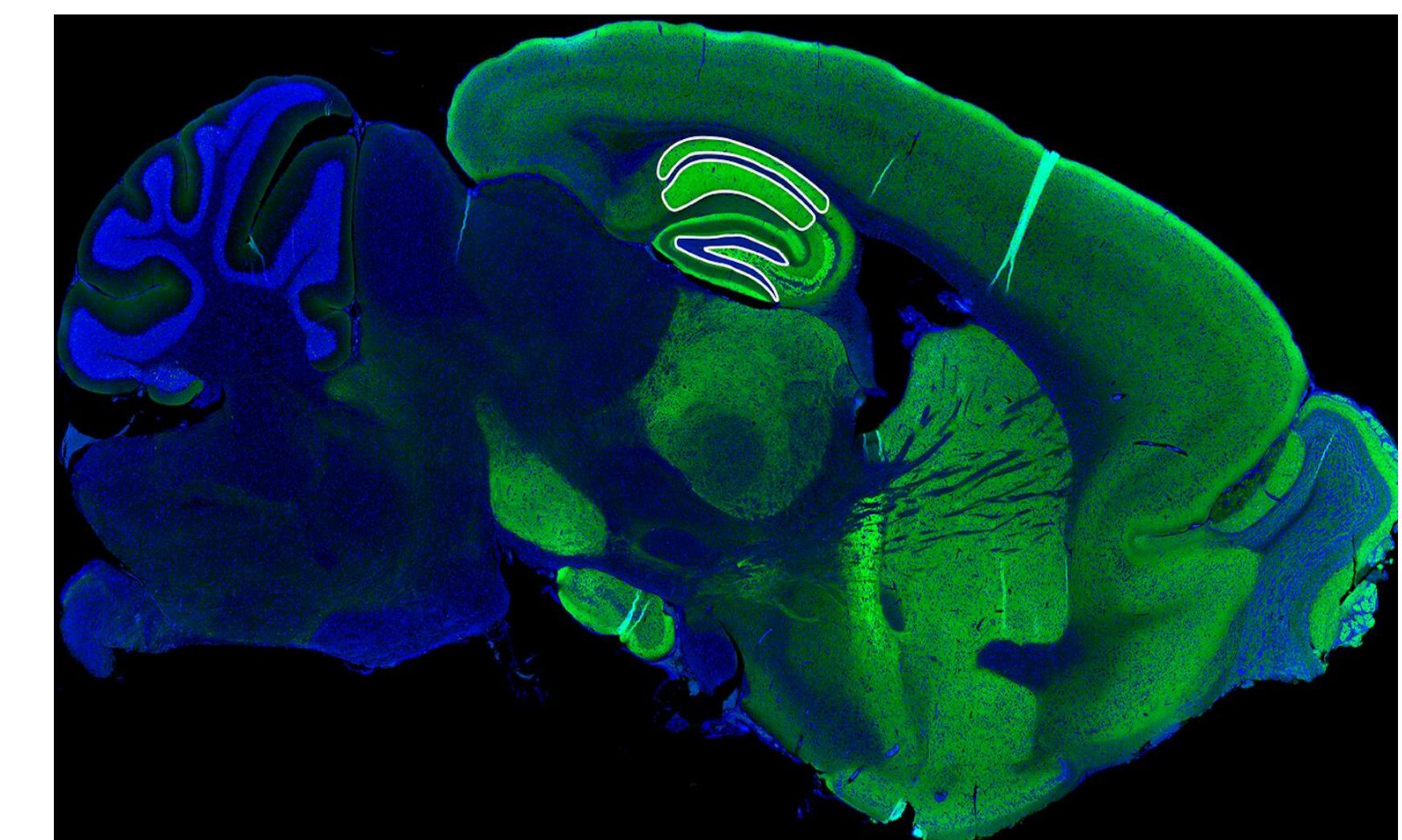
NEW: OpenFold
3D Protein Structure Prediction



NEW: DNABERT
DNA Sequence Model



NEW: MolMIM
Molecular Generation



COMING SOON: Single Cell BERT
Single Cell Expression Model

NVIDIA Training Times

Pre-Training from Scratch

Model	LLM Arch	Model Size (M Params)	Dataset	Training Time (Days) on 20 DGX	Experiments / Month
ESM	BERT	670	270M Proteins	7.5	4
ProtT5	T5	198	46M Proteins	7	4
MegaMolBART	BART	45	1.54B Compounds	1.5	20
GPT-3	GPT	5000	500B Tokens	5	6

Example Workloads:

- **Infrastructure**

- 20 DGX A100 Nodes
- Base Command Platform Software

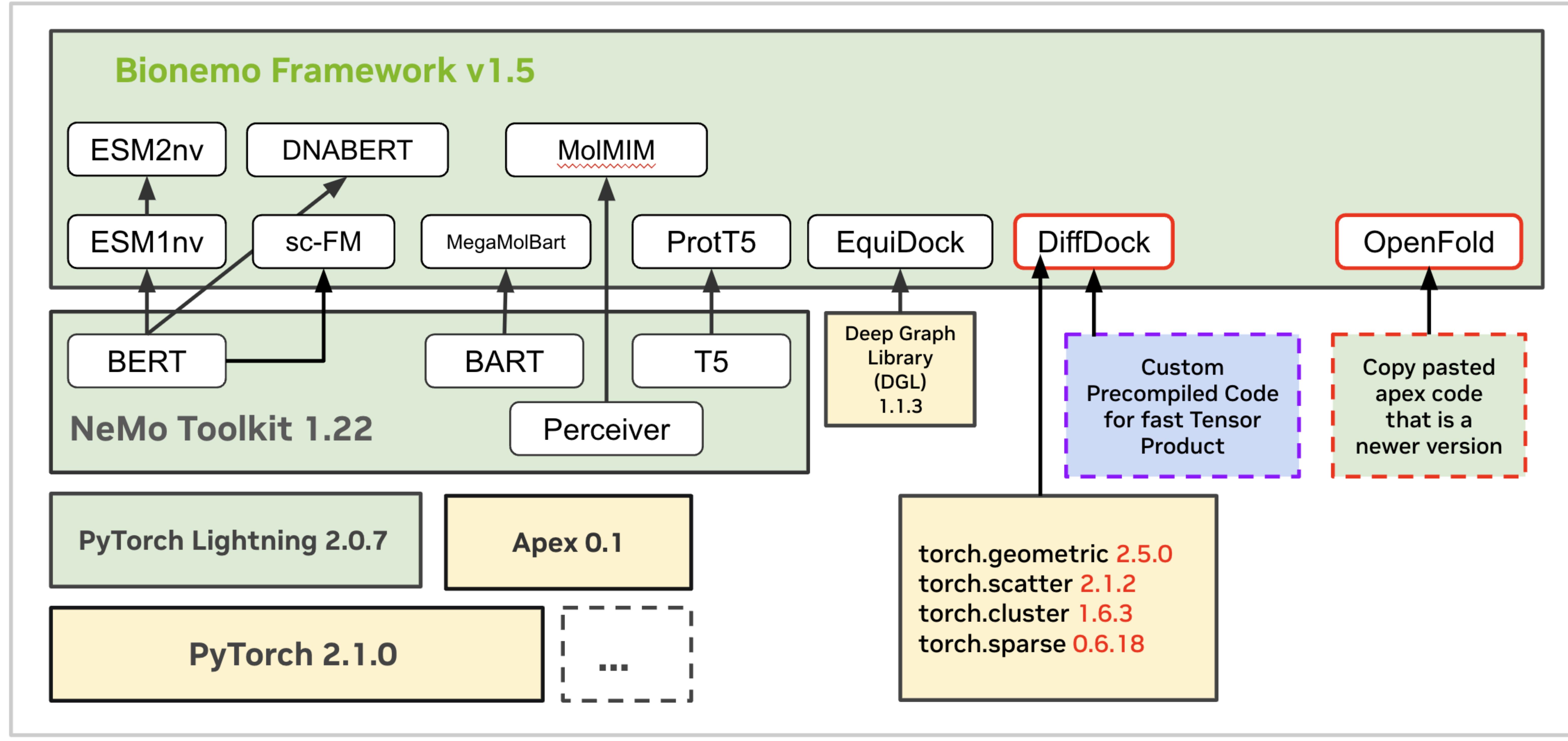
- **Software**

- NVIDIA AI Enterprise for
 - BioNeMo
 - NemoMegatron
 - Rapids

DGX Cloud Sample Sizing Guide

Analysis for Building Generative AI Models

BIONEMO FW STACK FOR V1.5 -- INSIDE CONTAINER



Legend

Python Class/Module	Python code	Compiled binaries	Compiled/Python Mix	Modified in this release (any red)	Temporary dependency that will need refactoring - tech debt (dashed)
---------------------	-------------	-------------------	---------------------	------------------------------------	--

DNABERT

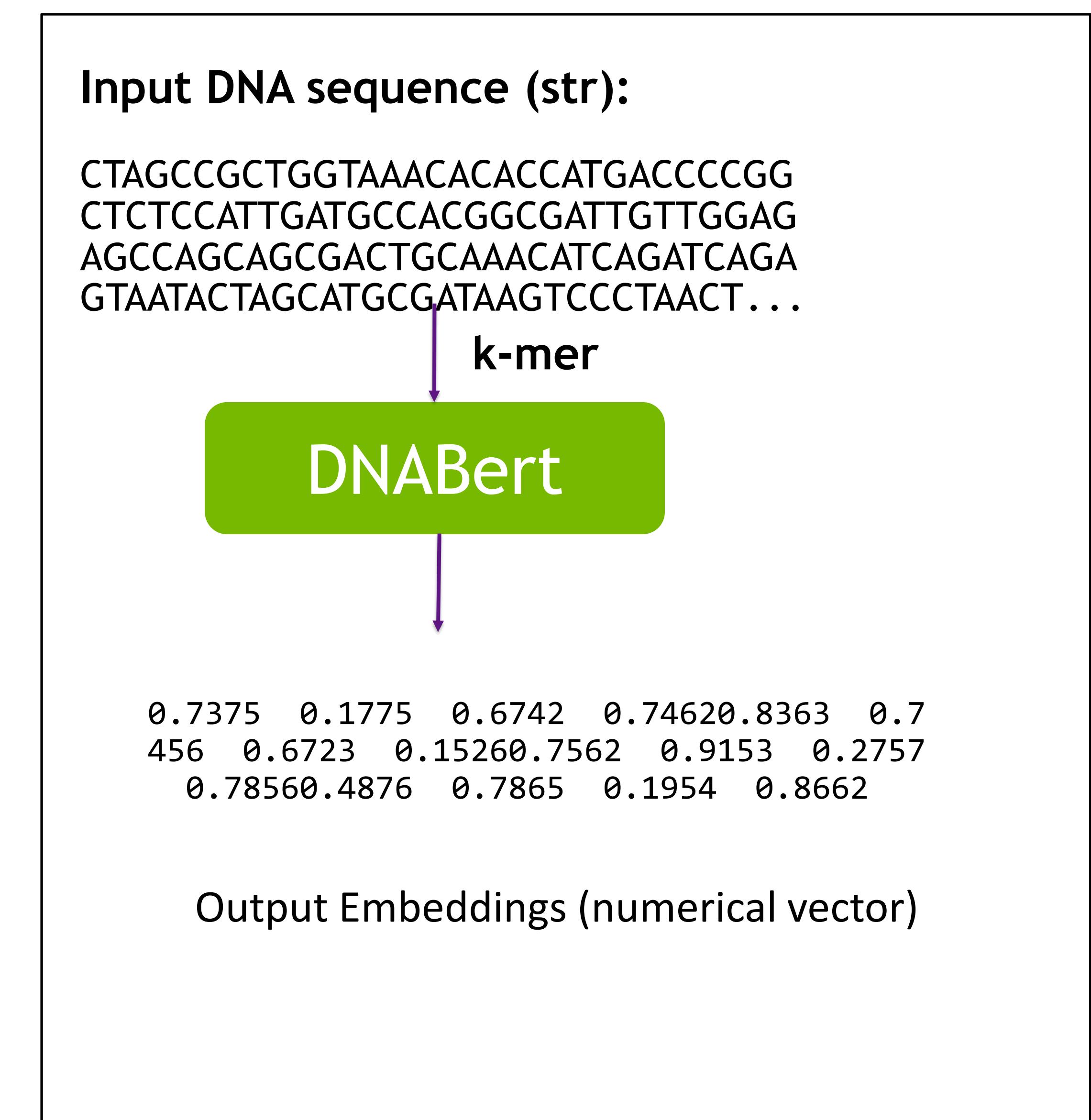
- **Usage:** embedding, prediction of transcription factor binding sites, prediction of splice sites, prediction of functional genetic variants
- **Type of model:** BERT model trained on DNA sequences. The primary use case of DNABERT is for end-users to train from scratch on their own data, rather than using pre-trained checkpoints.

- **In scope deliverables**

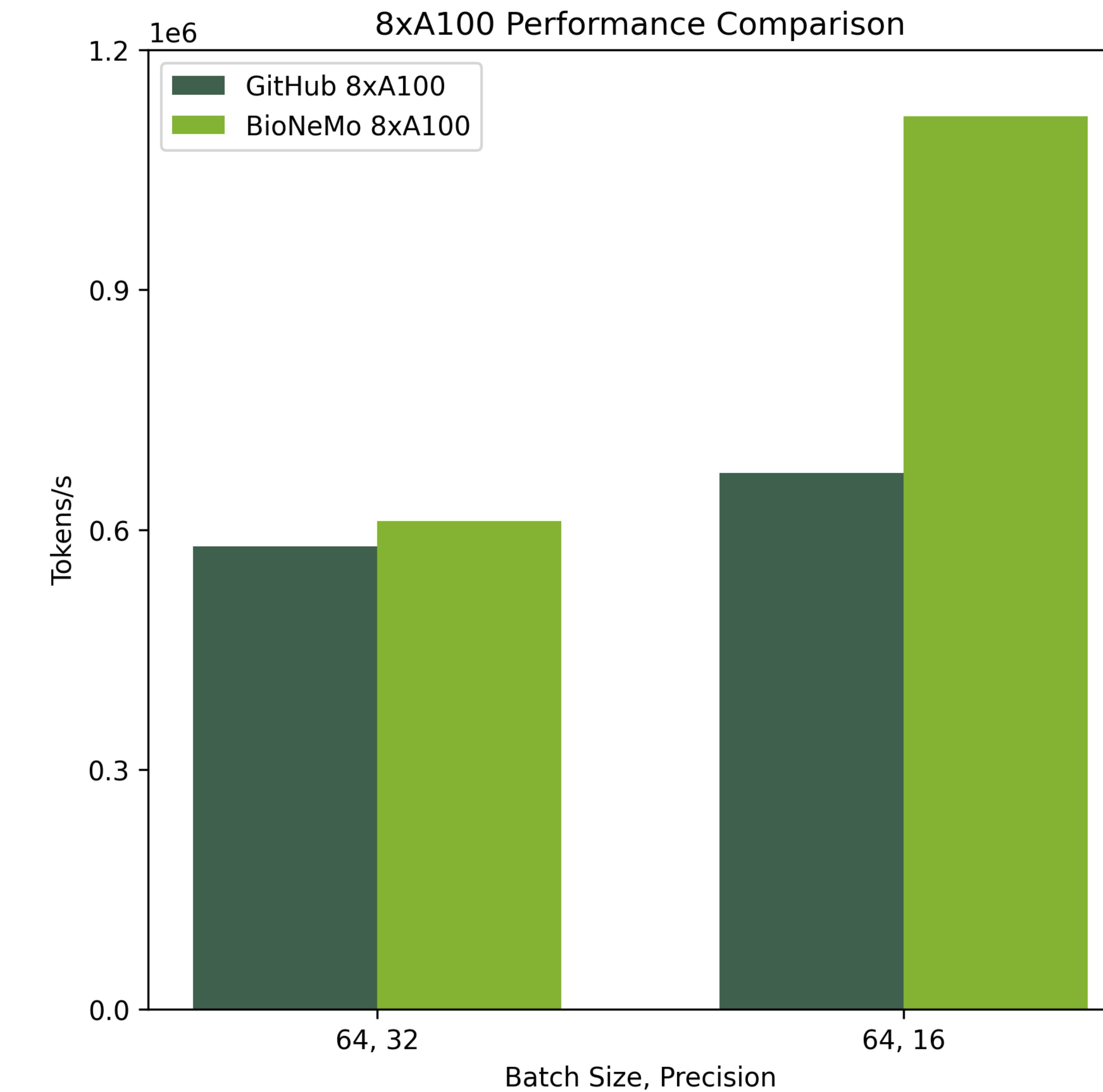
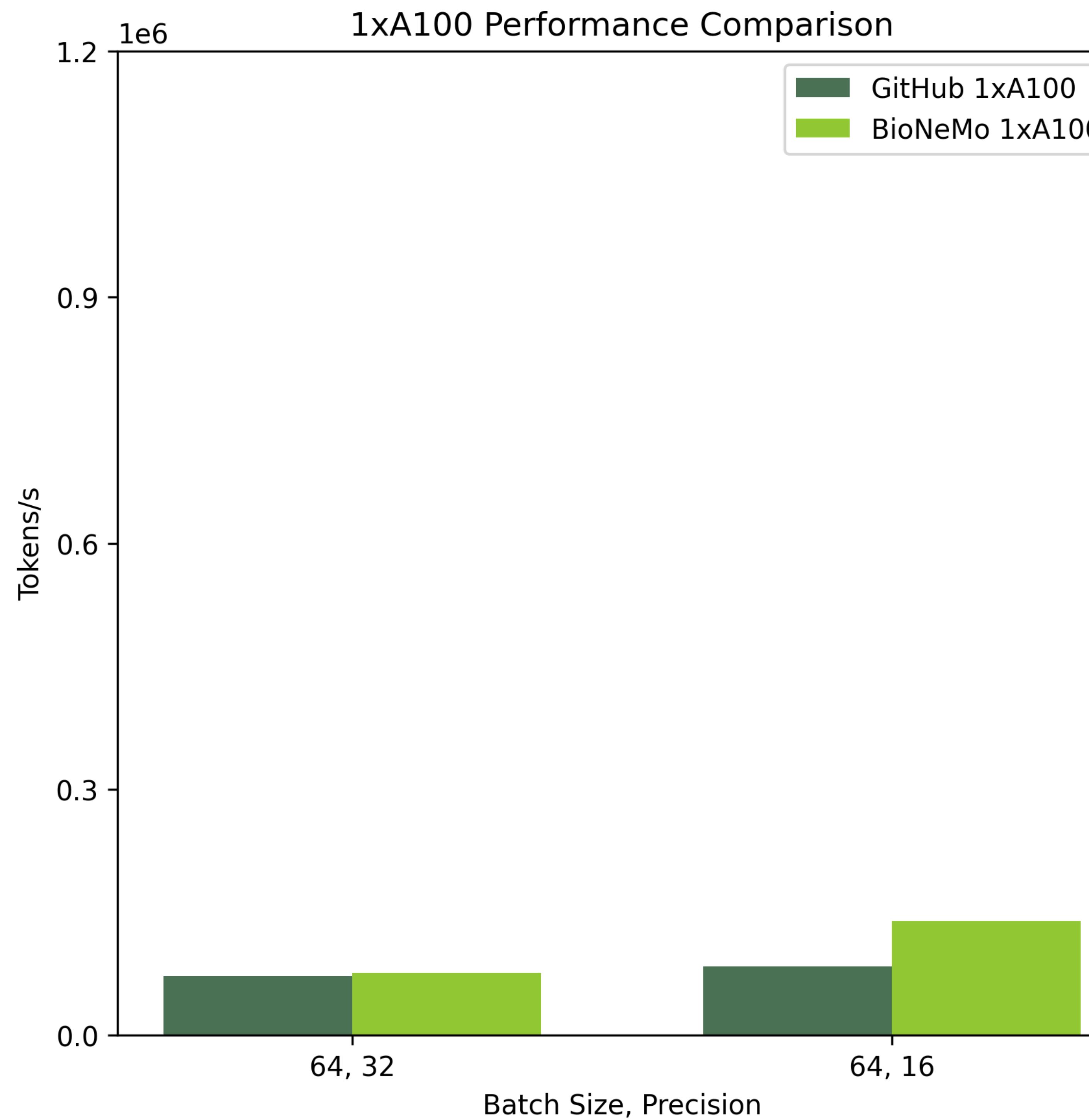
- DNABert training code.
 - Based on NeMo BERT, with tensor and pipeline parallelism support.
 - Downstream task support: Splice Site Prediction
- Pretrained model checkpoint:
 - Checkpoint was trained from scratch, 86M parameters, not expected to match paper accuracy, because this is not relevant for typical DNABert users.

Training statistics

- Dataset: Homo sapiens genome assembly GRCh38.p13 - NCBI - NLM (nih.gov)
3.2B nucleotides of the human genome sequence.
- Input: Plain Text of nucleotide sequences (A, C, G, T).
- Output: Text predictions in the form of dense numerical embeddings.

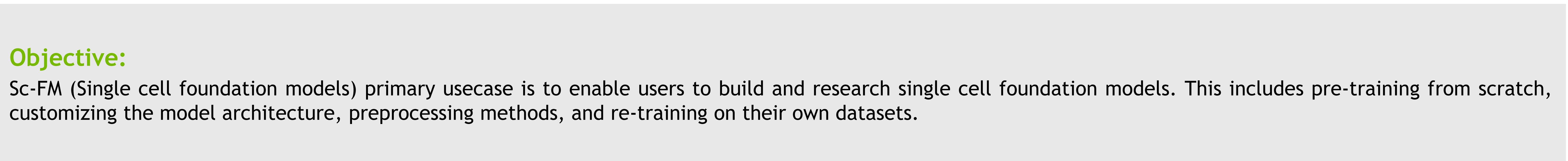


DNABERT

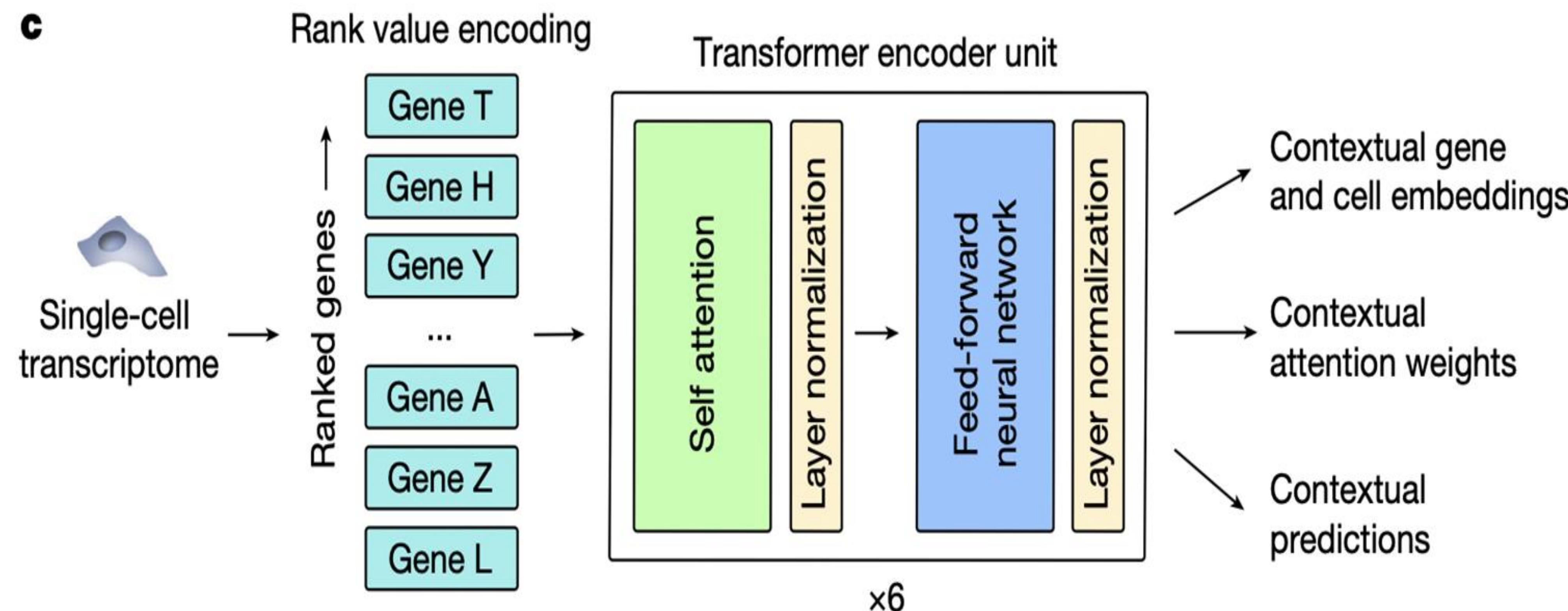


GENEFORMER

Enabling state of the art foundation models for single cell RNA-seq in BioNeMo



- **Architecture** - Geneformer-style BERT pre-training pipeline with restricted context length (1k genes)
 - Tensor, Pipeline Model parallelism, scale to 1B parameters.

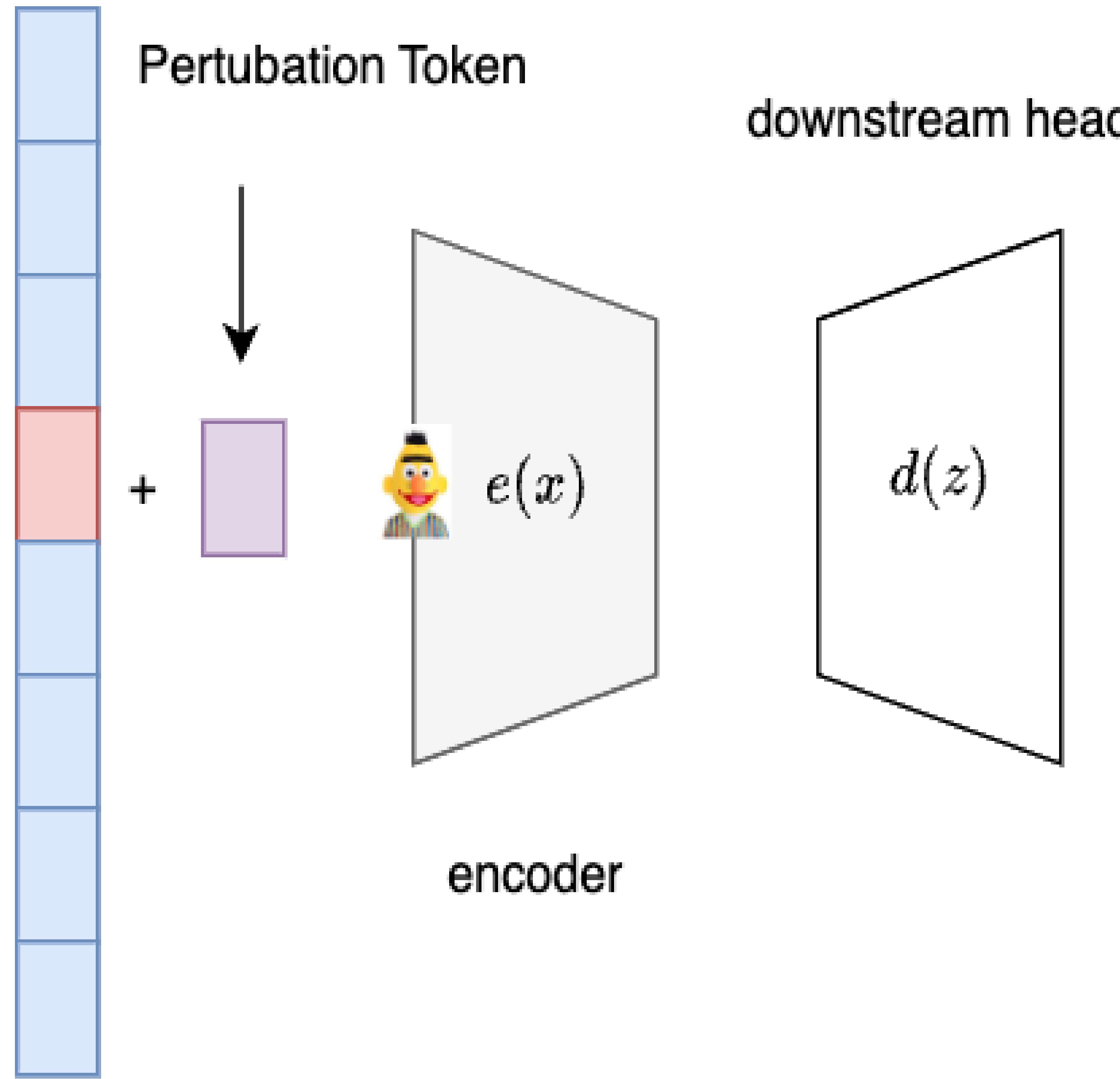


- This method uses a simple normalization technique based on the rank of gene expression.
- Limitation: It hides all of the resolution involved in gene expression behind ranks.

Geneformer

Downstream task

- Downstream tasks - Fine-tuning and zero-shot example using PERTURB-seq

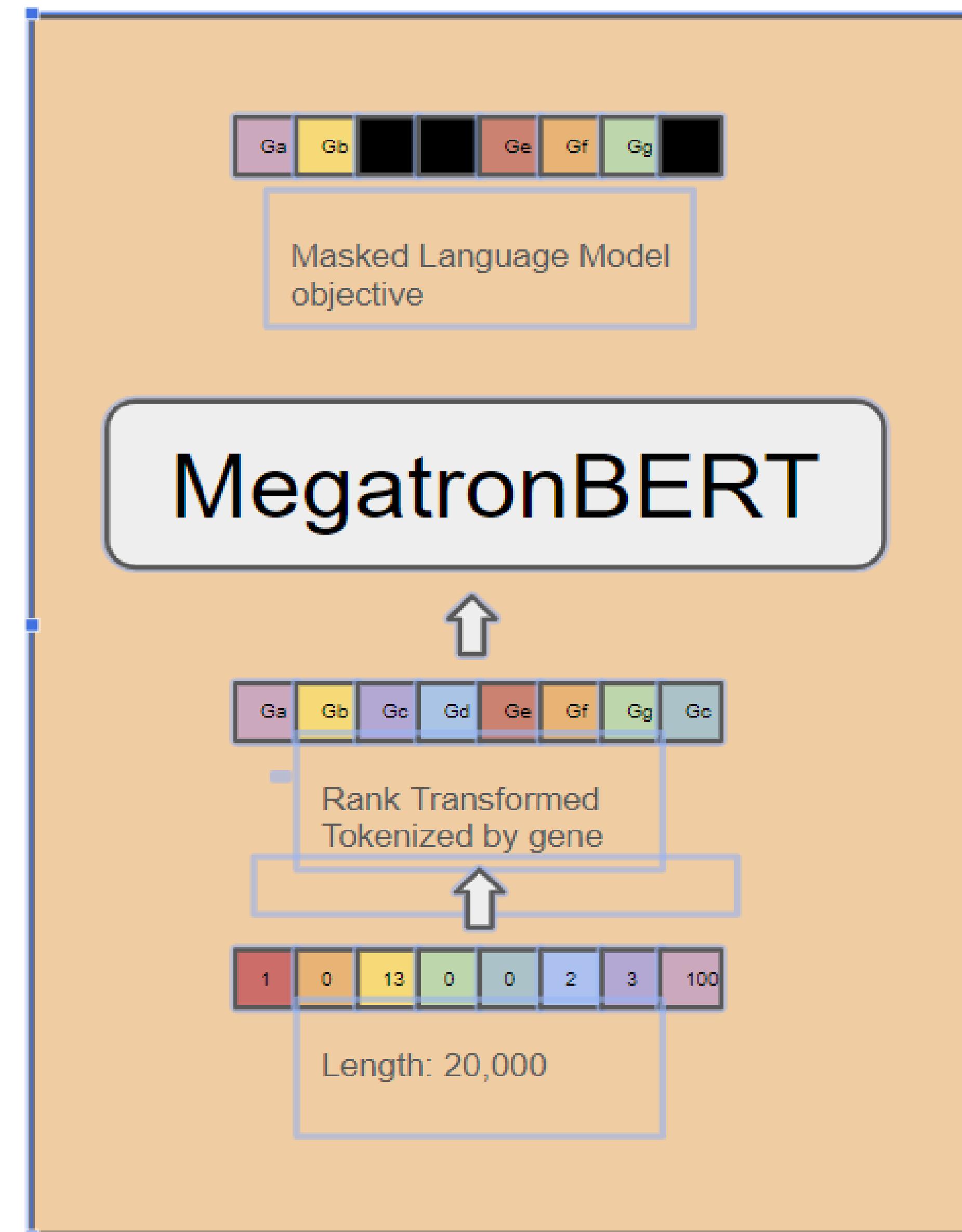


Requires modification to the input

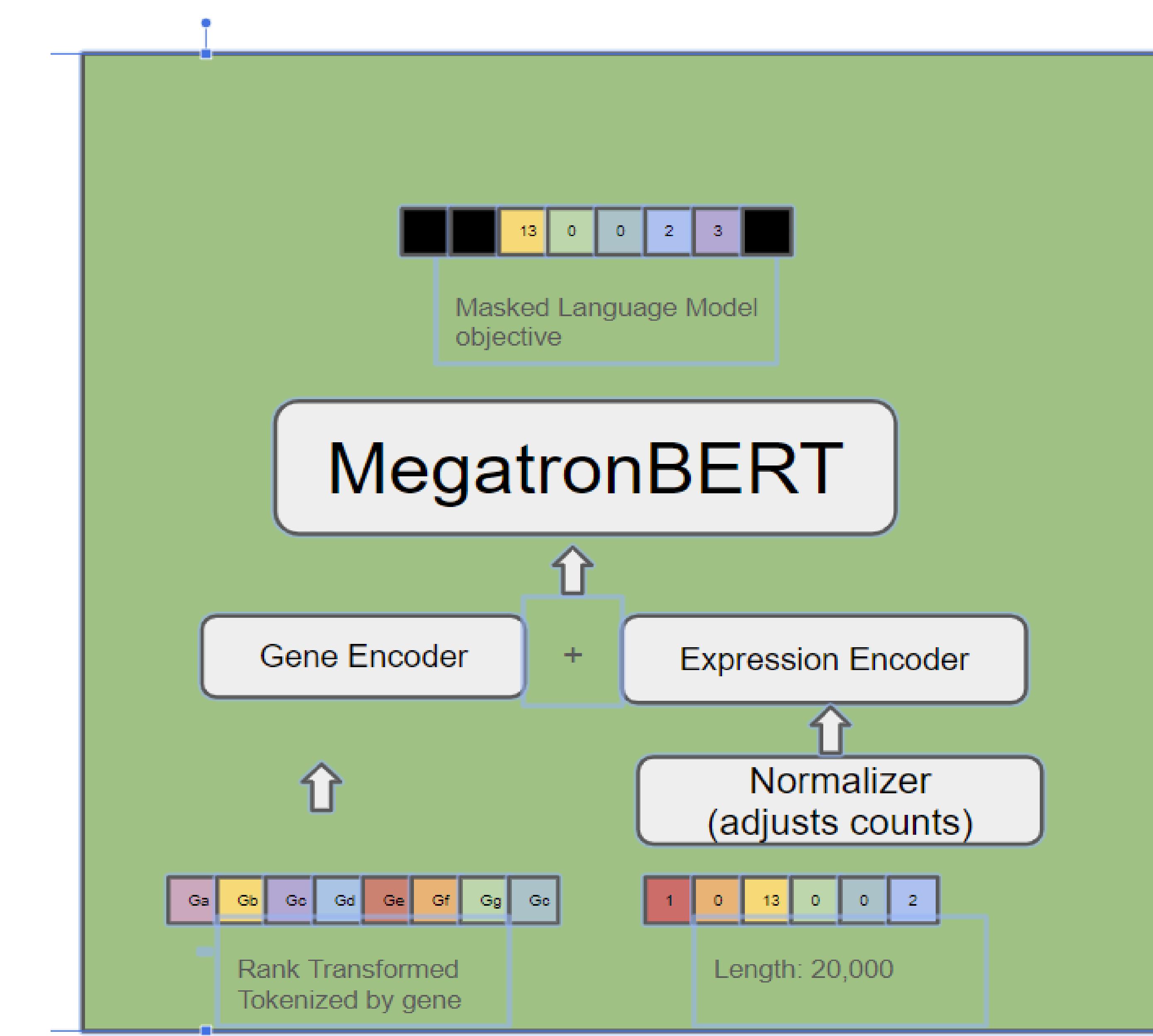
Option 1: use tokentype during fine tuning (BioNemo)

Option 2: add perturbed gene embedding @ output of the encoder

Geneformer



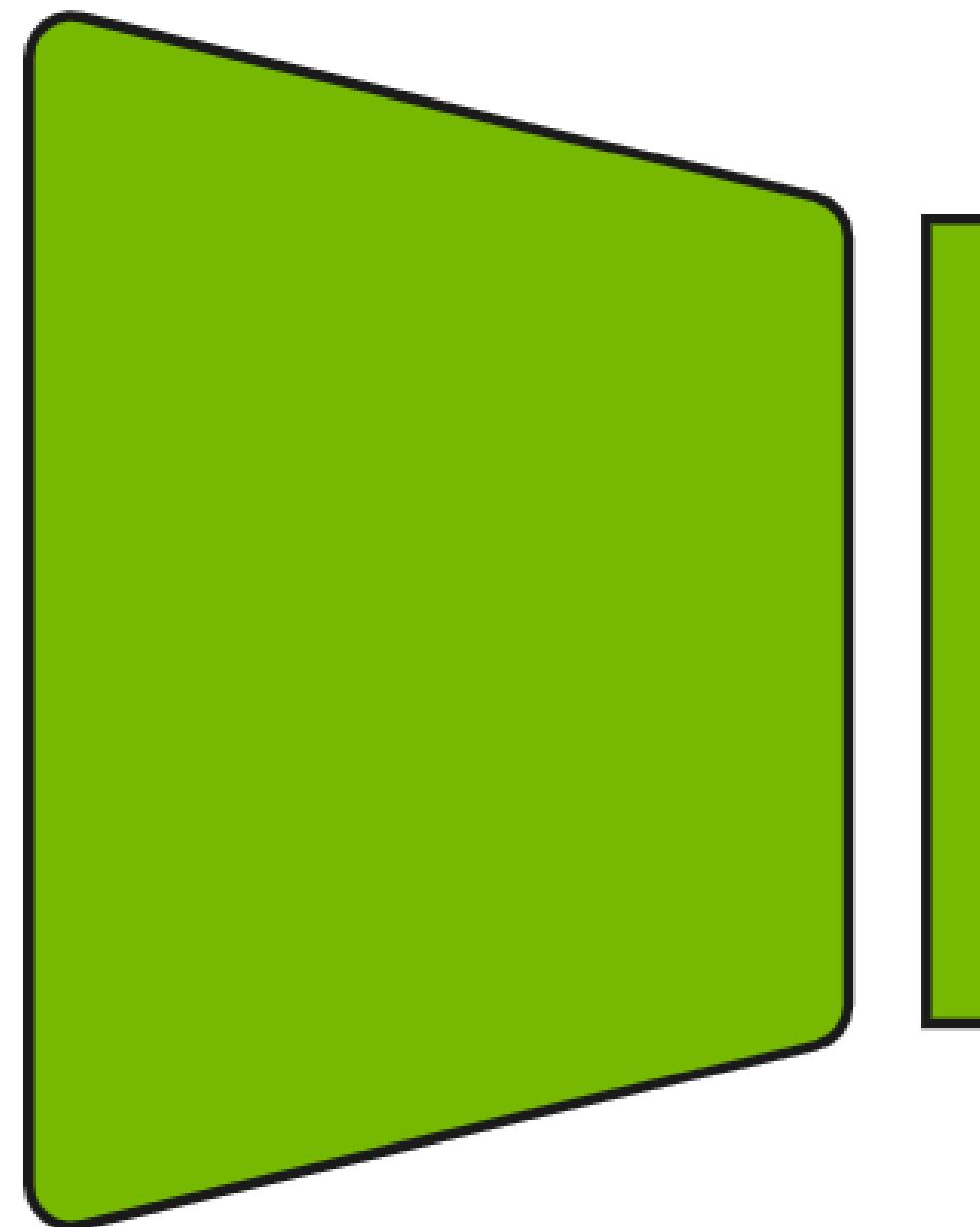
The model known as “geneformer”



An example, users will want to use more than one encoder for single cell foundation models, which is not supported by BioNemo yet.

ESM-1nv

MHHHHHSSGLVPRGSGMKGTAAKFERQH
MDSPDLGTDDDKAMADIGSENLYFQSMSK
IFVNPSAIRAGMADLEMAEETVDLINRNIE
DNQAHLQGEPIEVDSLPEIDIENLYFQGMES
DKIVFKVNNQLVSVKPEVIVDQEYKYPAI
QDHTKPSITLGKAPDLNKAYKSILSGMNAA
KLDPDDVCSYLAAMELFEGVCPEDWTSYG
IMIARKGDKITPATLVDIKRTDIEGNWALT
GGQDLTRDPTVAEHASLVGLLFFSRVEHLY
SAIRVGTVVTAYEDC5GLVSFTGFIKQIN



0.7375	0.1775	0.6742	0.7462
0.8363	0.7456	0.6723	0.1526
0.7562	0.9153	0.2757	0.7856
0.4876	0.7865	0.1954	0.8662



- Downstream task models
- Physical properties
- Sequence-level tasks
- Residue-level tasks
- Clustering

RAPIDS

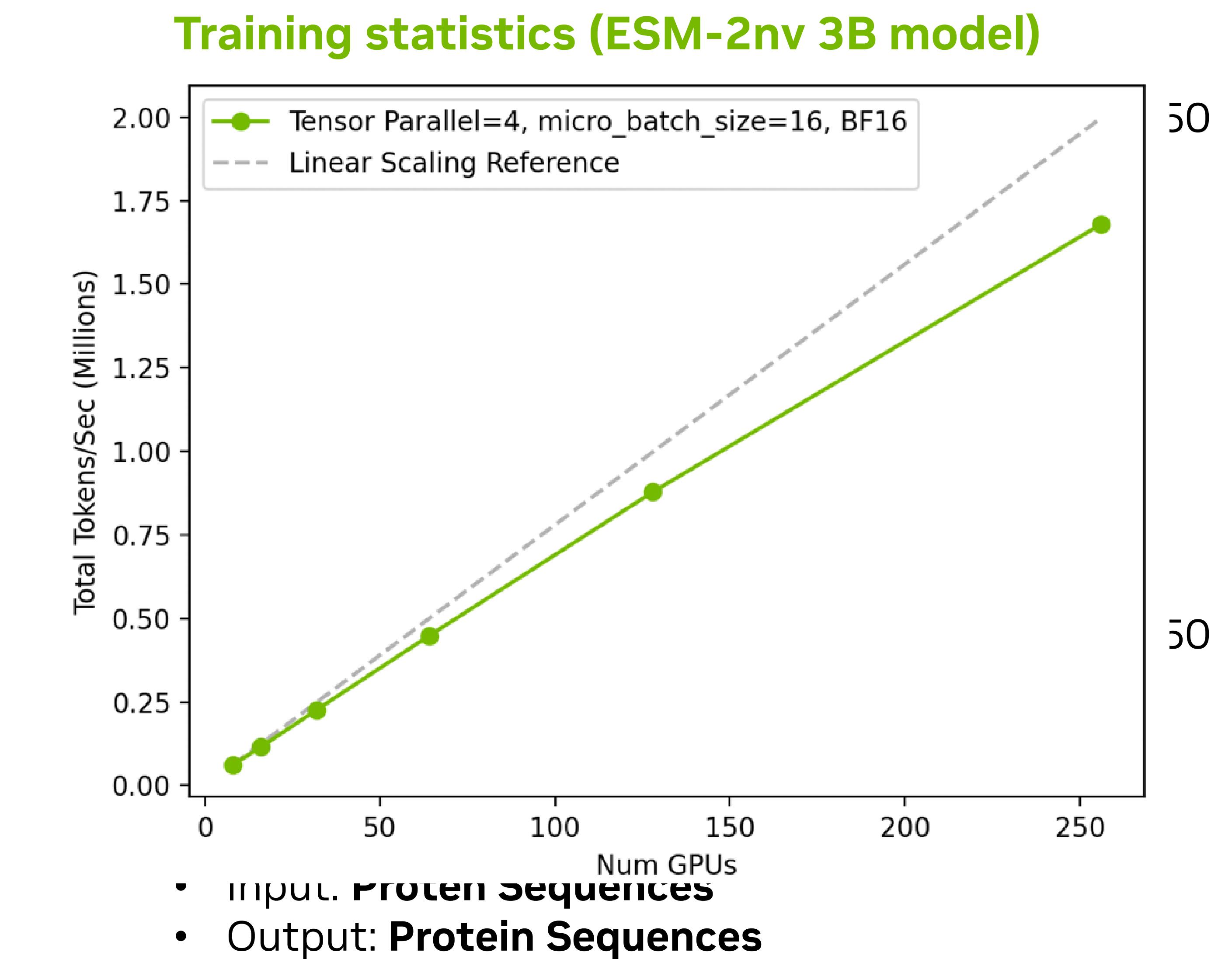
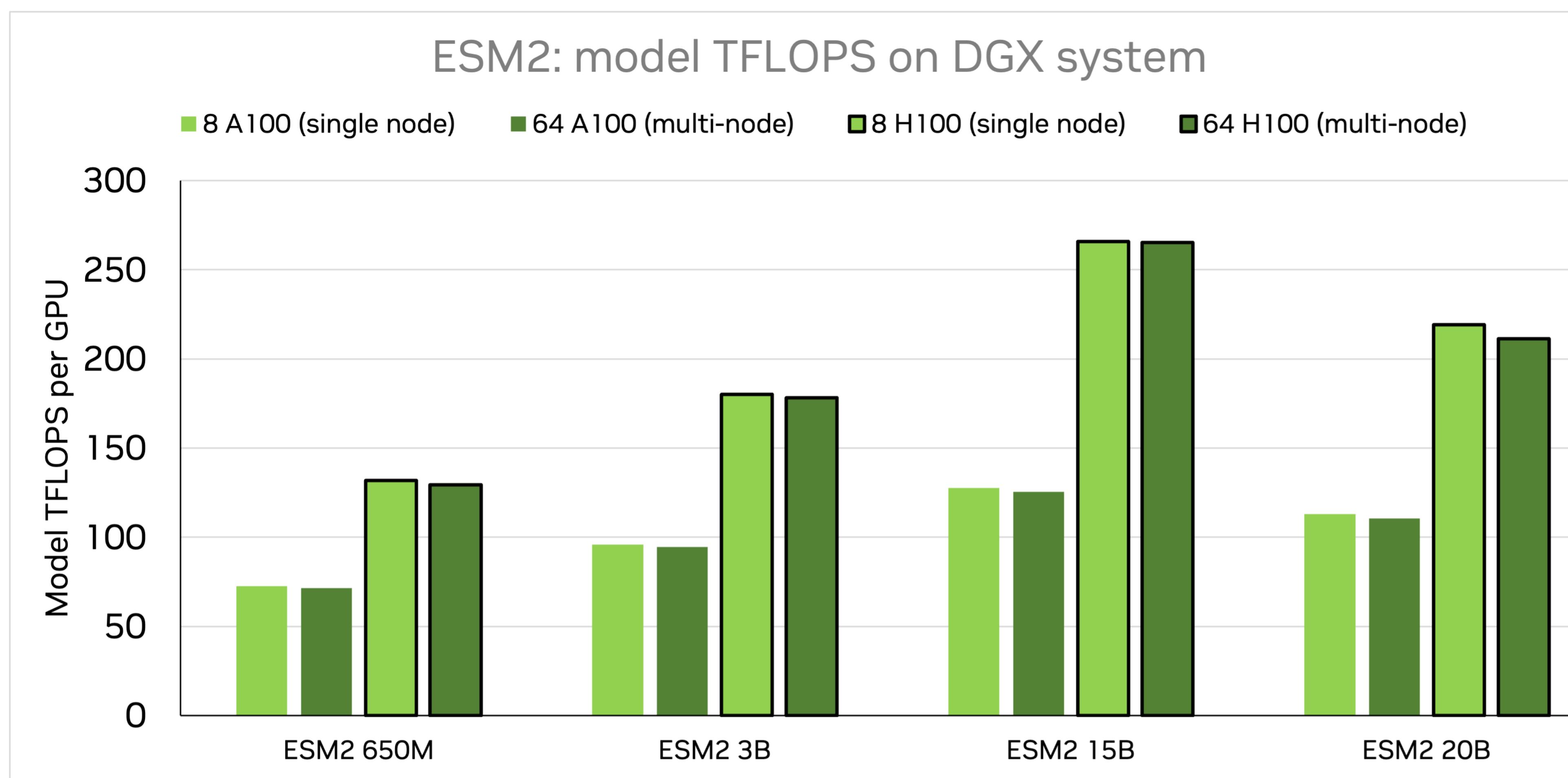
PyTorch

Training statistics

- Trained on **46 M** protein sequences from UniRef50
- **44 M** parameters
- Hidden dim size: **768**
- Maximum input length: **512**
- Maximum output length: **512**
- Input: **Protein Sequences**
- Output: **Protein Sequences**

ESM-2nv

- **Usage:** embedding, predict structure, function and other protein properties
- **Type of model:** BERT and is based on ESM-2 model published by Meta AI
- **Data:** Unlike ESM-2 pre-training data, the curated pre-training dataset provided with ESM-2nv release contains hits for de novo proteins, since sequences in UniRef100, UniRef90, and UniRef50 with high sequence similarity to a non-public 81 de novo proteins are not filtered.



ESM2 LORA

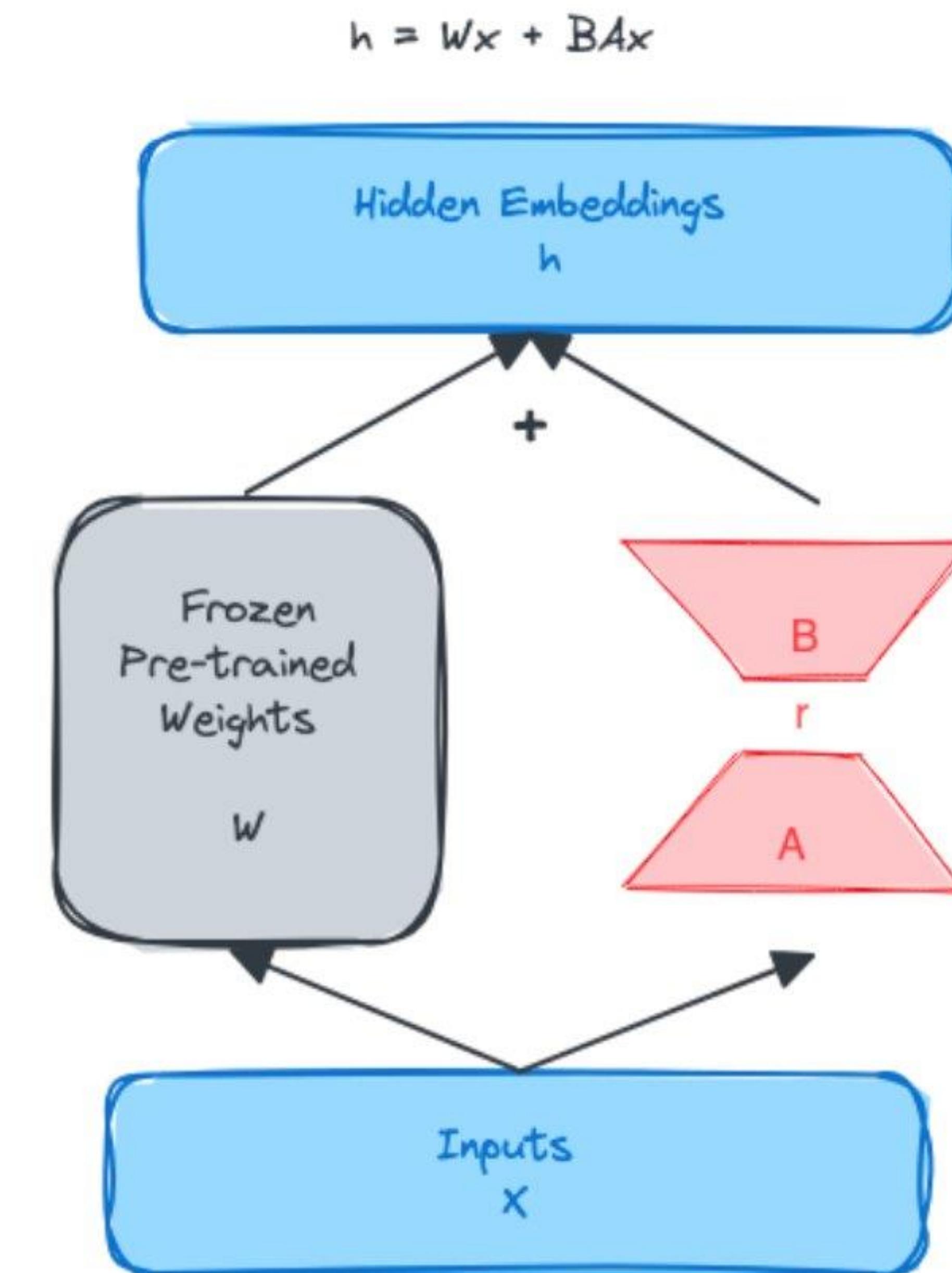
Enabling parameter-efficient finetuning technique for ESM2

Objective:

LoRA addresses some of the drawbacks of fine-tuning by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture. The key advantage of this decomposition is that it significantly reduces the number of trainable parameters.

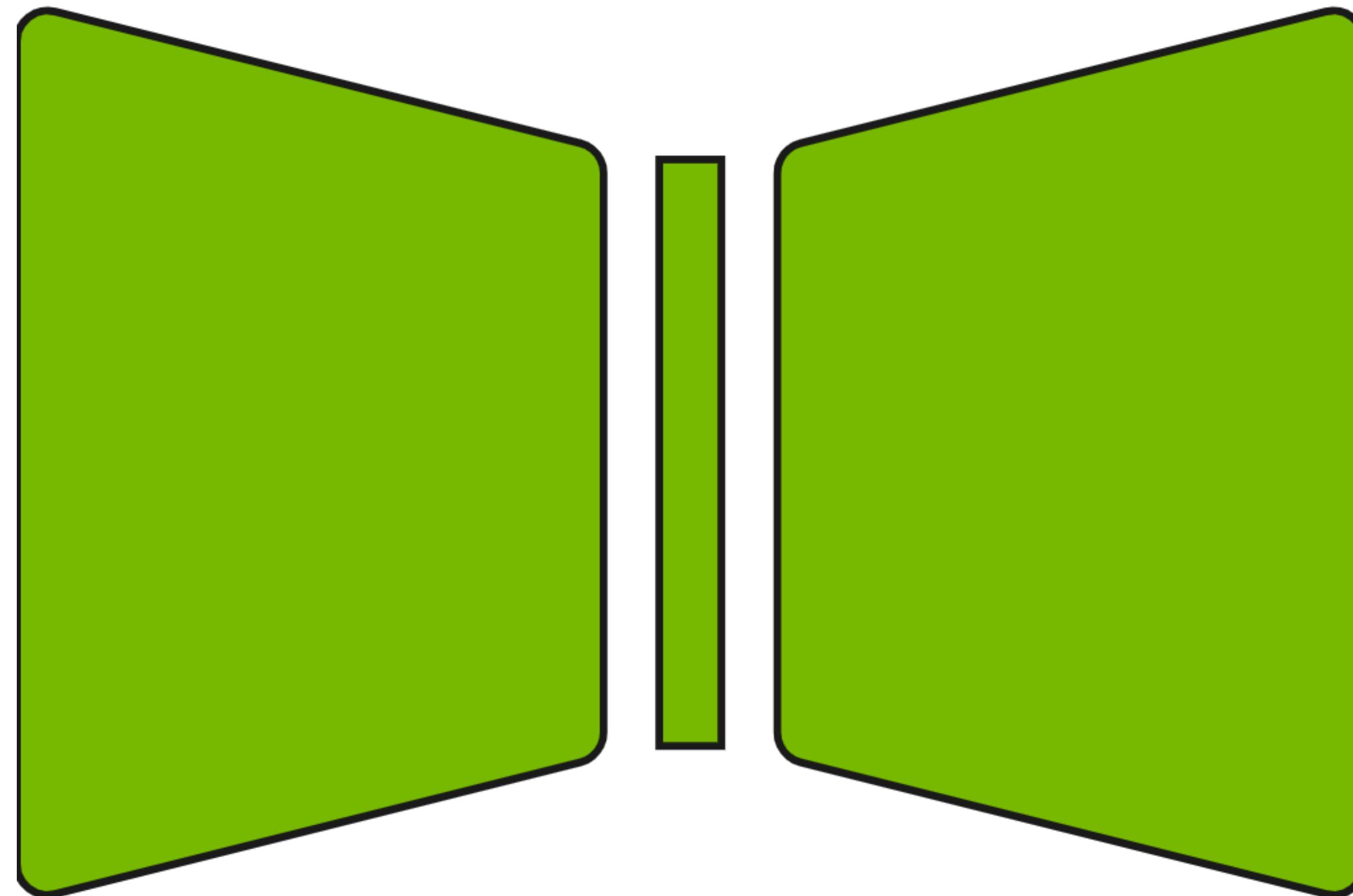
Performance

- Training speed, faster training step time compared to full finetuning (10-25% speedup), 1 node and 8 node A100
- Trainable parameter reduction, ~ 99% reduction compared to full finetuning



ProtT5nv

MHHHHHSSGLVPRGSGMKE TAAAKFERQH
MDSPDLGTDDDKAMADIGSENLYFQSMSK
IFVNPSAIRAGMADLEMAEETVDLINRNIE
DNQAHLQGEPIEVDSLPE DIENLYFQGMES
DKIVFKVNNQLVSVKPEVIVDQEYKYPAI
QDHTKPSITLGKAPDLNKAYKSILSGMNAH
TLMTTHKMCANWSTIPNFRFLAGTYDMFFS
RVEHLYSAIRVGTVWTAYEDCSGLVSFTGF
IKQIN



MHHHHHSSGLVPRGSGMKE TAAAKFERQHMDSPDL
GTDDDKAMADIGSENLYFQSMSKIFVNPSAIRAGM
ADLEMAEETVDLINRNIEEDNQAHLLQGEPIEVDSLPE
DIENLYFQGMESDKIVFKVNNQLVSVKPEVIVD... .

Training statistics

- Trained on **46 M** protein sequences from UniRef50
- **192 M** parameters
- Hidden dim size: **768**
- Maximum input length: **512**
- Maximum output length: **512**
- Input: **Protein Sequences**
- Output: [Protein Sequences, Score]

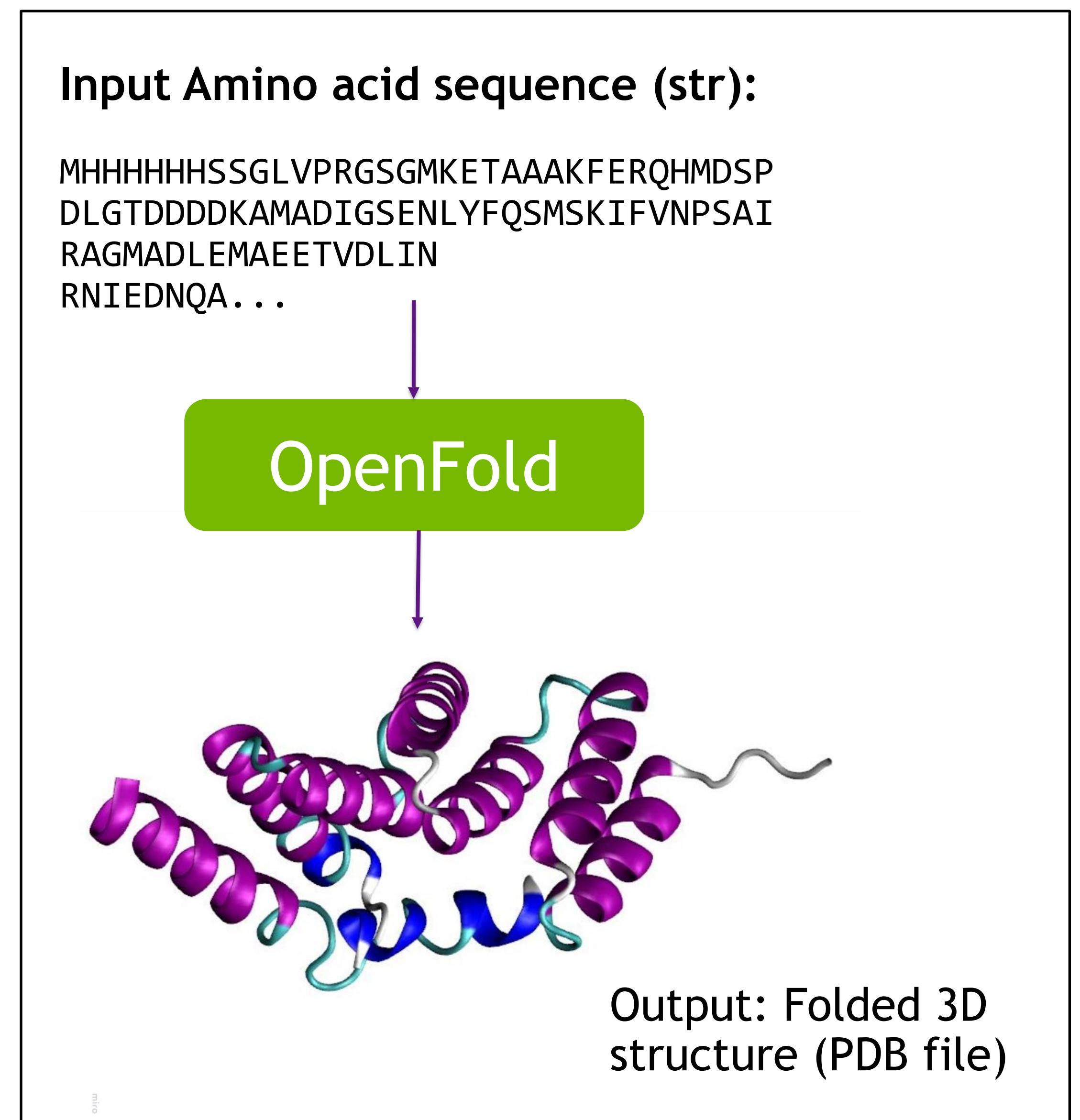
OPENFOLD RELEASE

OpenFold is the open-source version of AlphaFold, which predicts protein 3D structure from amino acid sequences.

- **In scope deliverables**
 - OpenFold training code, with finetuning capability
 - Compatible training data: MSA, PDB, HHR structural templates
 - Two checkpoints:
 - Public converted checkpoint, LDDTCa = 91.5 (min. required LDDTCa 90)
 - Public converted checkpoint + finetunes in the framework, LDDTCa = 91.5 (doesn't get worse)
- **Benchmarks**
 - Training performance, 16 node (128 GPU), A100. Up to 1.7x perf boost (compared to public OpenFold implementation).
 - Accuracy benchmarks for released checkpoints. Document that training from scratch yields LDDTCa = 90.5

Training statistics

- Trained on **200 k** protein structures from PDB-mmCIF dataset and **269 k** sequence alignments from OpenProtein Set
- Input: **Protein Sequences + MSA (optional) + HHR structural template (optional)**
- Output: **PDB + (optional) Confidence Metrics + (optional) Embeddings**



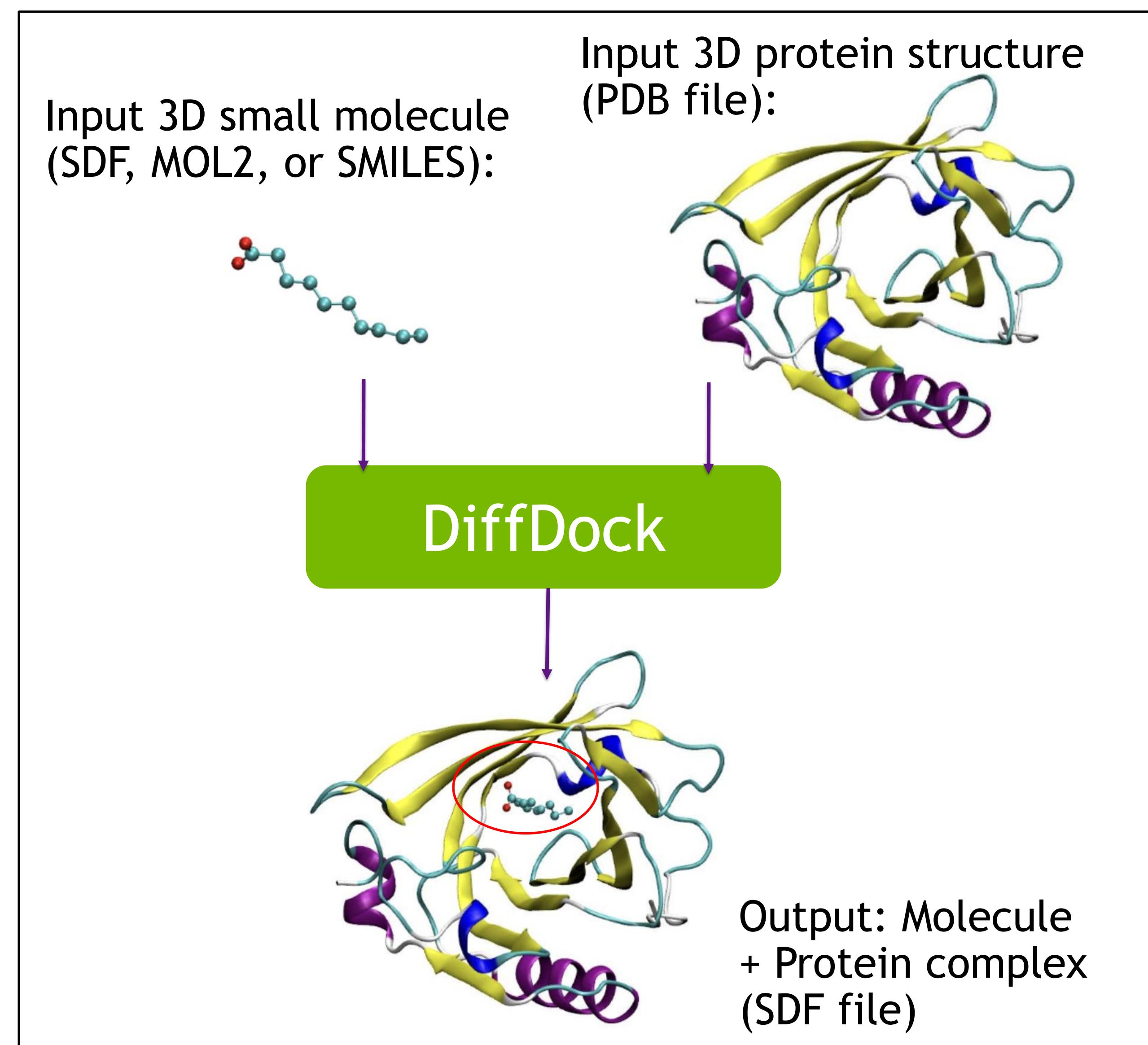
DIFFDOCK RELEASE

DiffDock is an equivariant diffusion model for small molecule – protein docking.

DiffDock consists of two models: the Score and Confidence models.

The Score model: a 6-graph-convolution-layer 3-dimensional equivariant graph neural network with 20M parameters. The Score model is used to generate a series of potential poses for protein-ligand binding by running the reverse diffusion process.

The Confidence model: a 5-graph-convolution-layer 3-dimensional equivariant graph neural network with 5M parameters. The Confidence model is used to rank the generated ligand poses from the Score model.



Training statistics

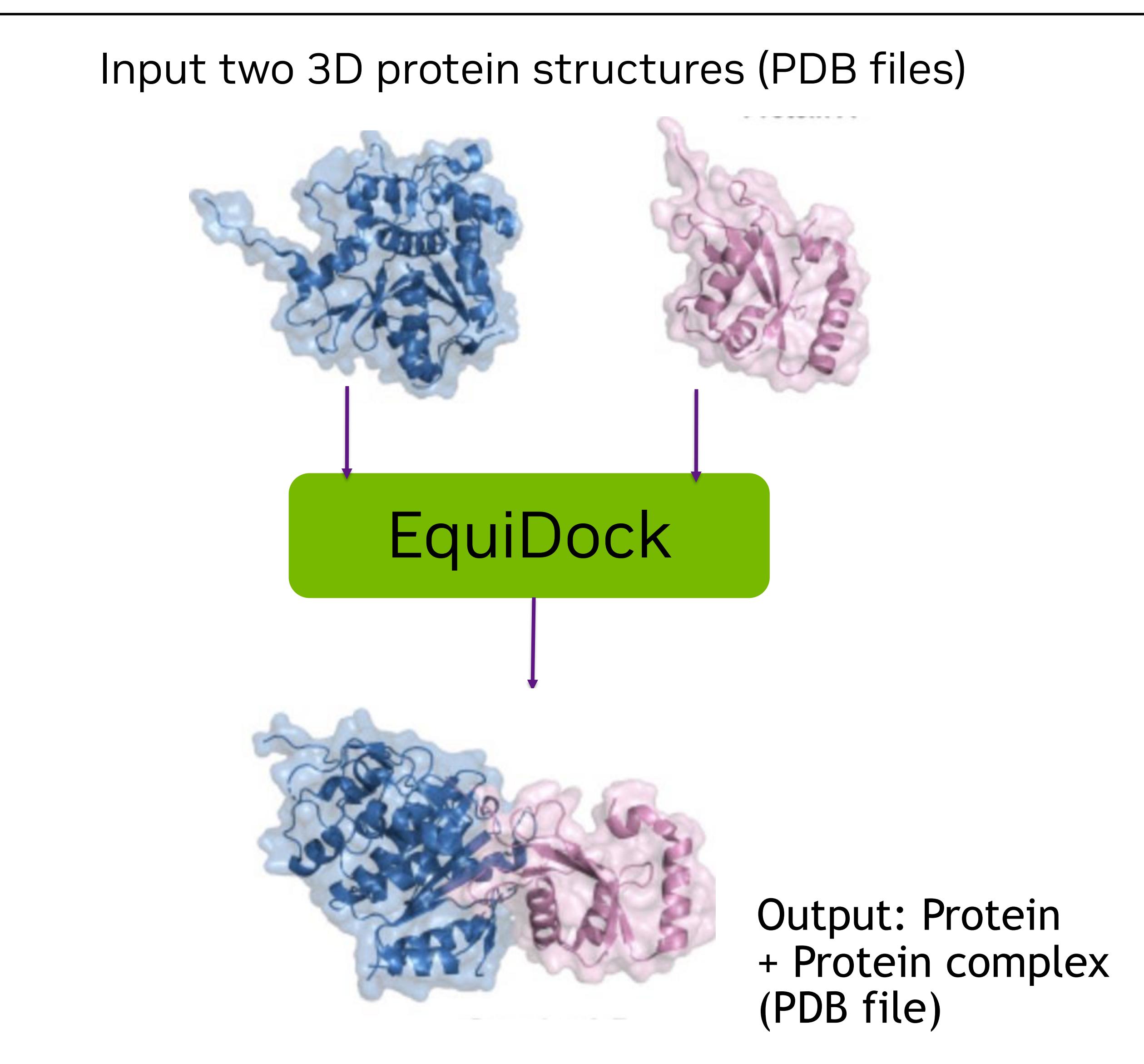
- Evaluated on **428** protein-ligand complexes manually curated using the [PoseBusters benchmark \(PDB\) set](#)
- Input: **Text (PDB, SDF, MOL2, SMILES)**
- Output: **Structural Data Files (SDF) + Confidence Score and the rank based on this score**

EQUIDOCK release

Equidock is an SE(3)-equivariant (GNN) rigid model for prediction of protein-protein complex formation.

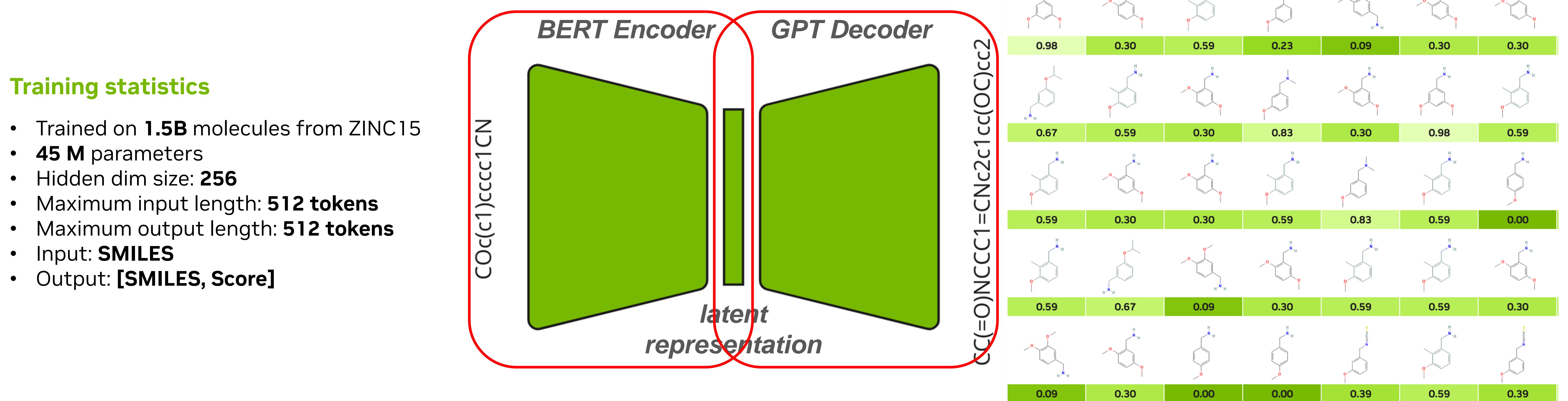
Training statistics

- Dataset:
 - DB5.5 dataset consists of 253 protein structures built by mining the Protein Data Bank (PDB) for pairs of interacting proteins.
 - The Database for Interacting Proteins Structures (DIPS) has 41,876 binary complexes containing bound structures with rigid body docking, while DB5.5 includes unbound protein structures.
 - DB5.5 includes 203/25 training and validation data points, respectively.
 - DIPS includes 39,937/974 training and validation data points respectively.
- Input: **Text (Geometric Protein Structure), Maximum number of residues is 400 per protein and maximum number of atoms is 4000 per protein**
- Output: **Protein Data Bank (PDB)**



MegaMolBART

- **Usage:** embedding, reaction prediction, property prediction, de novo molecular generation
 - **Type of model:** a BART model, seq2seq transformer with layer normalization and GELU activation used throughout
 - **Representation:** variable-size representation for variable length SMILES



MegaMolBART: NVIDIA & AstraZeneca; Bjerrum, Irvin, Engkvist, et al.

<https://ngc.nvidia.com/catalog/models/nvidia:clara:megamolbart>

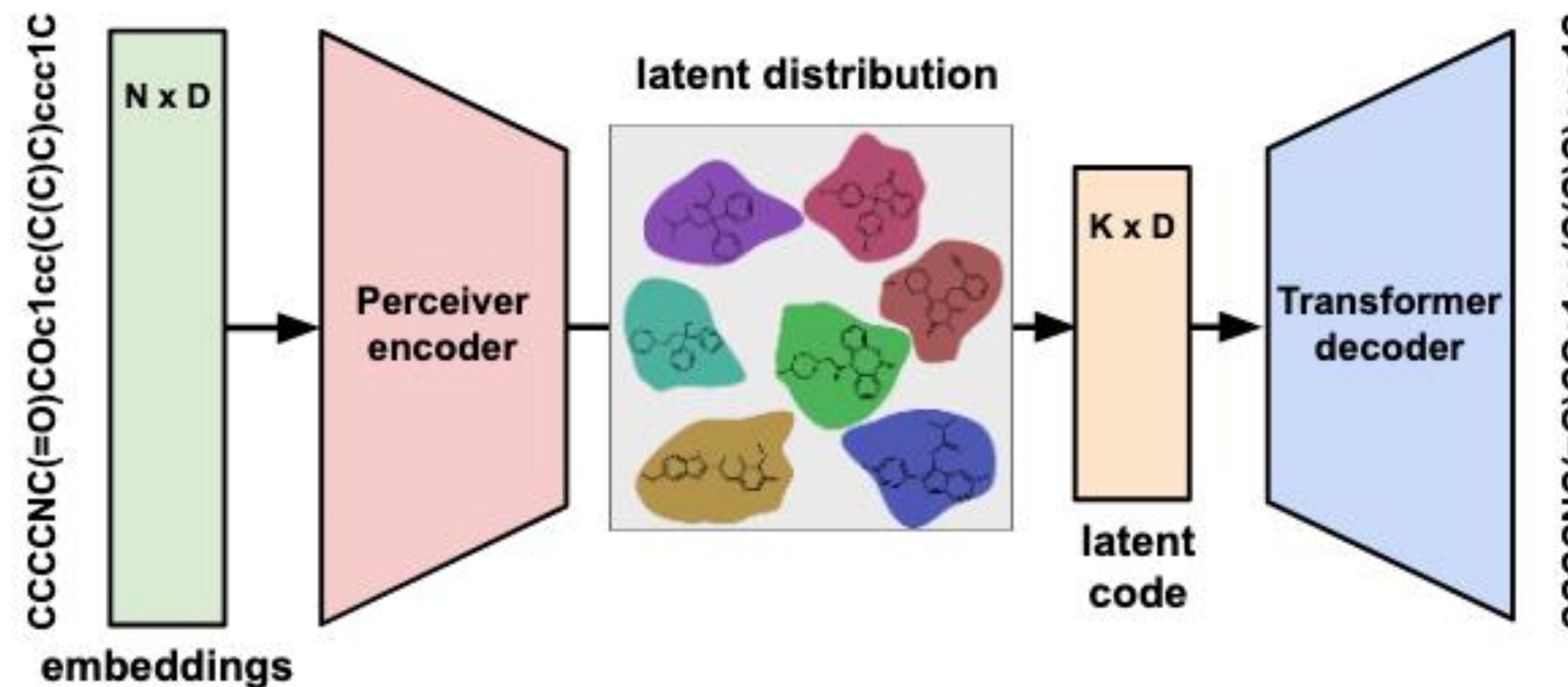


MolMIM

Generative small molecule model developed by NVIDIA

- **Usage:** embedding, reaction prediction, property prediction, molecular optimization, de novo molecular generation
- **Type of model:** auto-encoder trained with mutual information machine learning (MIM)
- **Representation:** fixed-size representation for variable length SMILES

Architecture of MolMIM



Training statistics

- Trained on **730 M** molecules from ZINC15
- **65 M** parameters
- Hidden dim size: **512**
- Maximum input length: **128 tokens**
- Maximum output length: **512 tokens**
- Input: **SMILES**
- Output: **[SMILES, Score]**

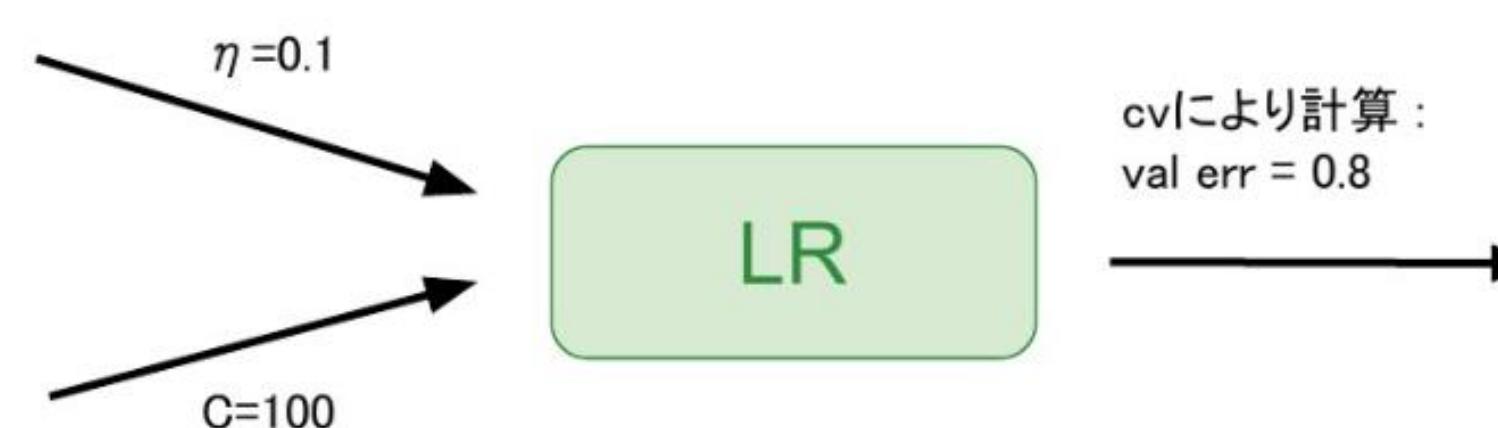
The MIM part of MolMIM refers to the custom Mutual Information Machine (MIM) loss which overcomes some technical shortcomings of variational auto-encoders. MolMIM is uniquely able to sample valid, novel molecules with only slight perturbations of the latent code, due to its unique, dense latent space that clusters chemically similar molecules. These features lead to superior property guided optimization performance.

Hands-on: MolMIM and DNABERT

CMA-ESを使用したMoIMIMによる特性ガイド分子最適化

CMA-ES (Covariance Matrix Adaptation Evolution Strategy) は、進化計算における最も有力なブラックボックス最適化の手法の1つとして知られています。このアルゴリズムは多変量正規分布から解を生成し、その解の評価値を利用して、より良い解を生成するような分布に更新を行う手法です。

ハイパーパラメータ最適化の流れ

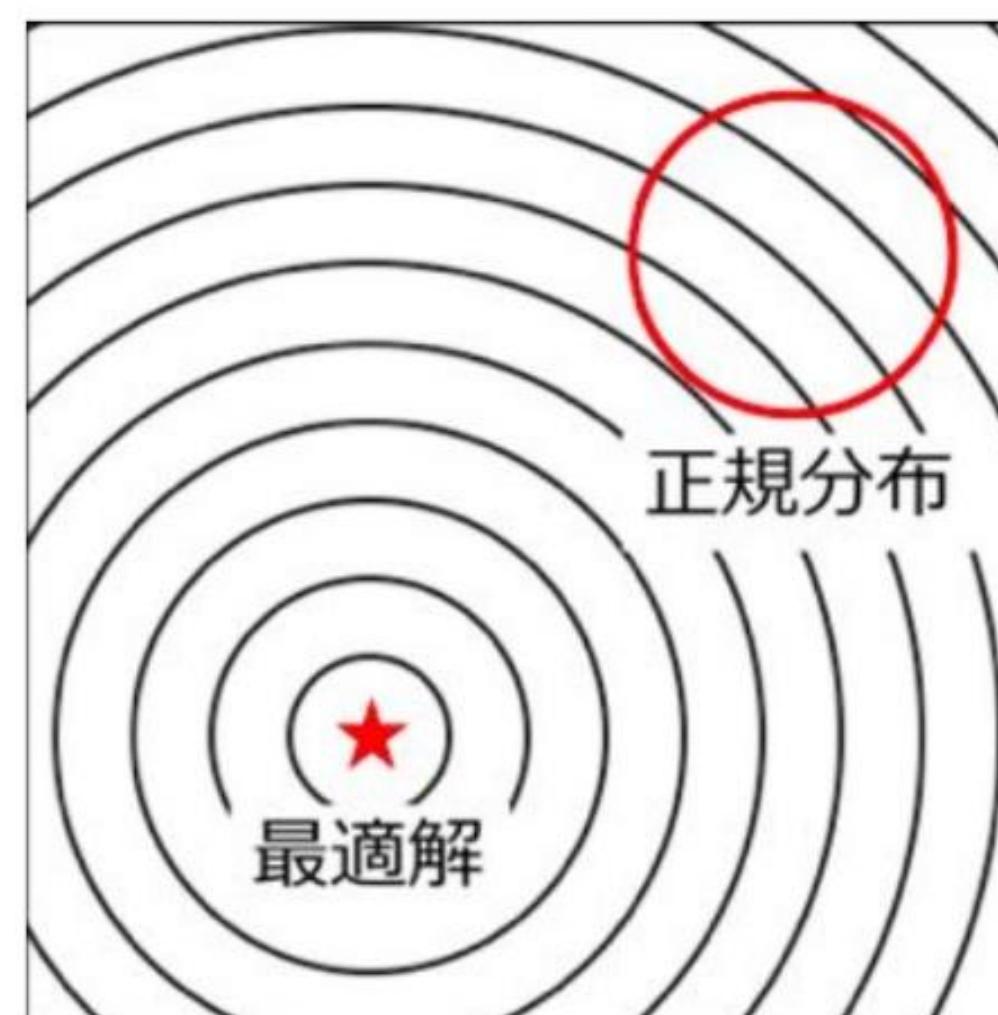


Validationエラーを小さくするように
ハイパーパラメータを決めるという問題



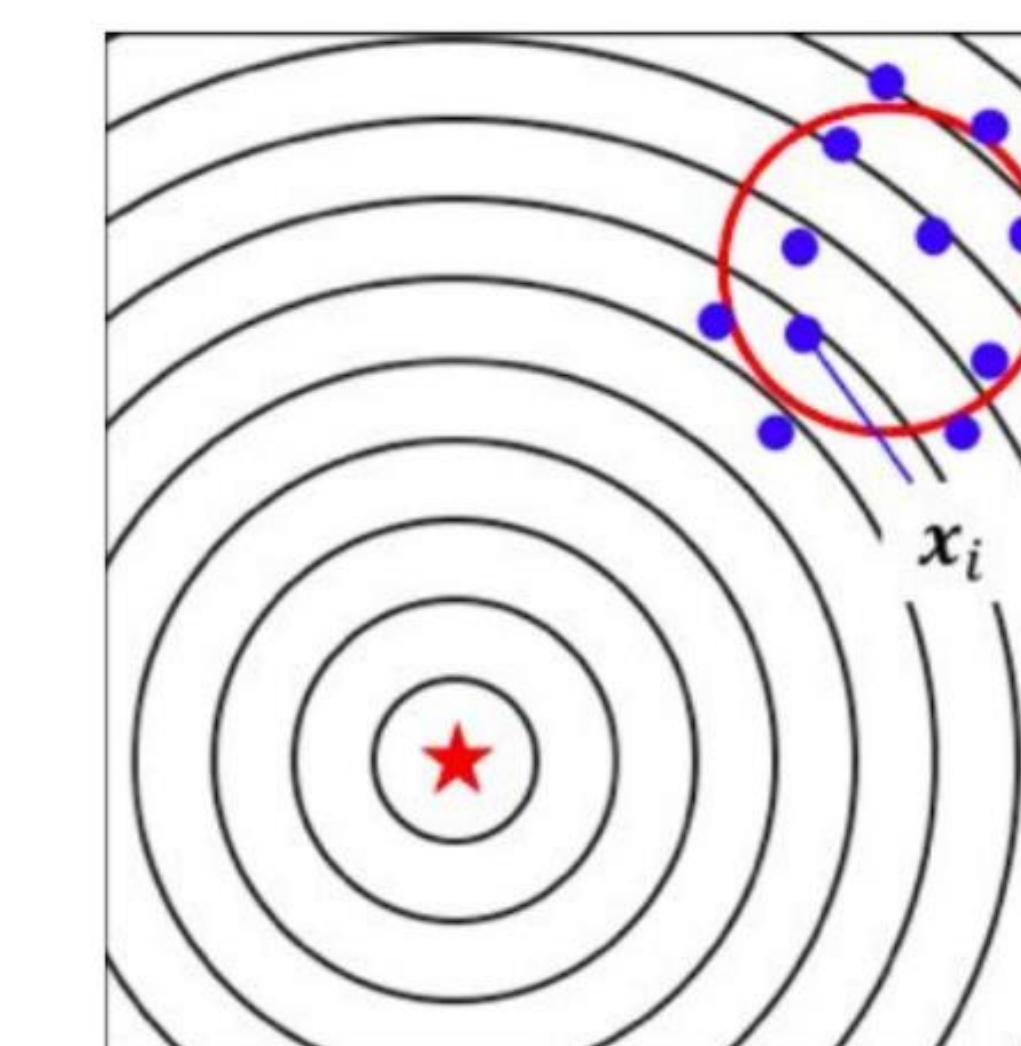
中身がブラックボックスな関数の最適化
= Black-box最適化

CMA-ES



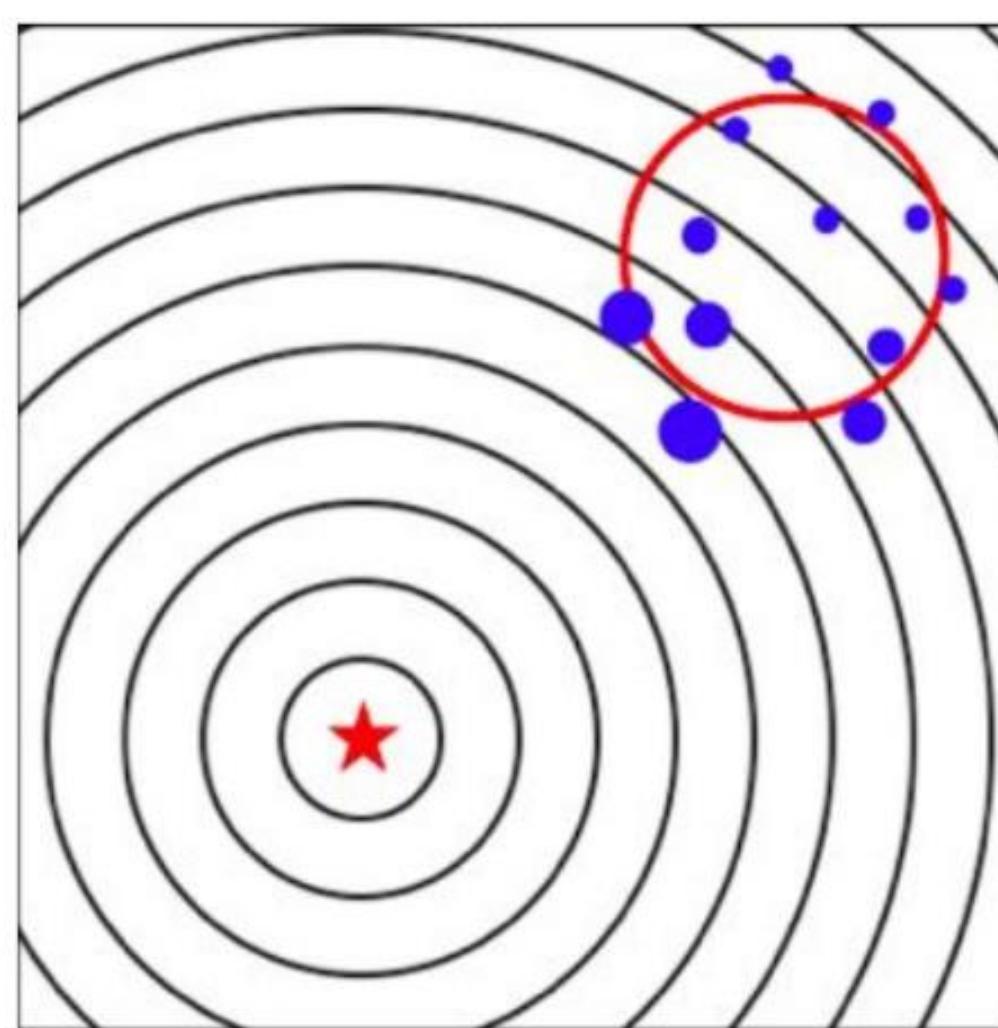
1. 正規分布から解を生成
2. 全ての解を評価して重み付けする
3. 正規分布のパラメータを更新
4. 1.~3.を繰り返す

CMA-ES



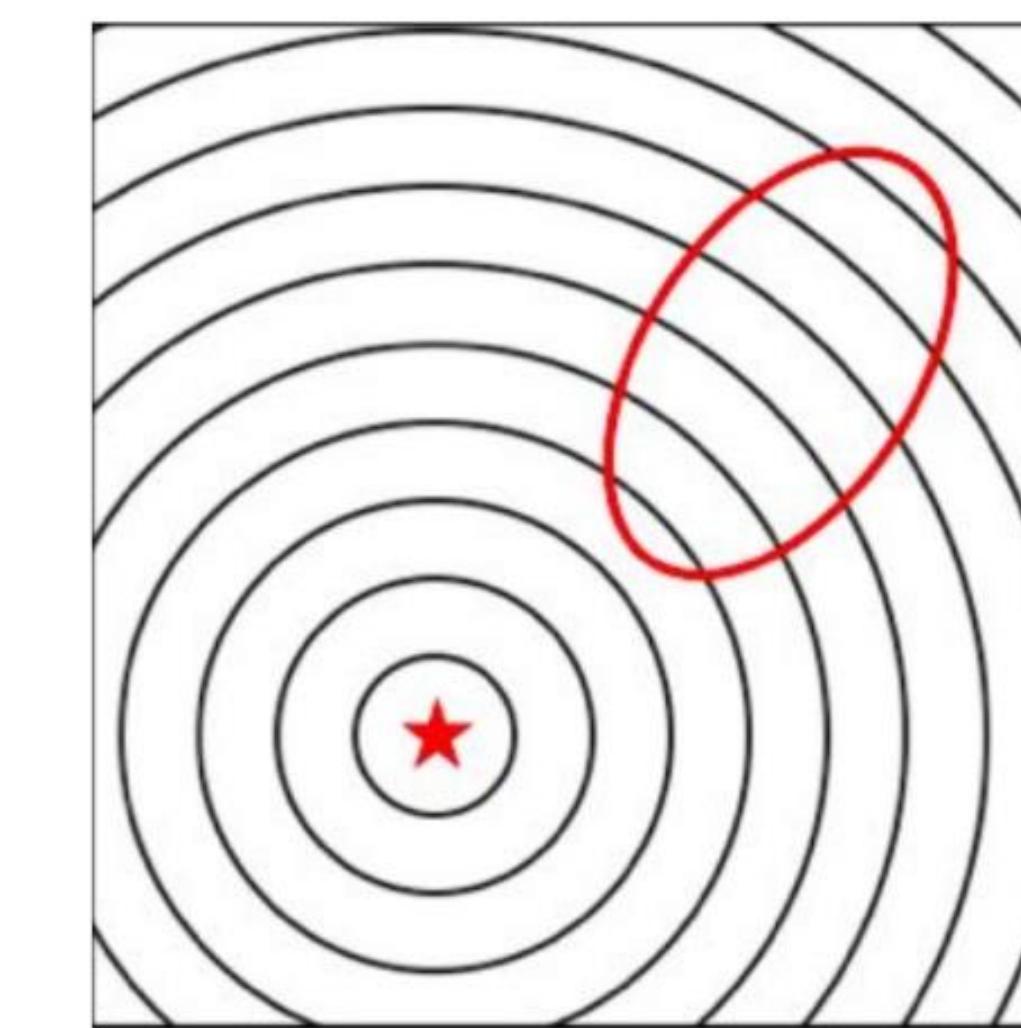
1. 正規分布から解を生成
2. 全ての解を評価して重み付けする
3. 正規分布のパラメータを更新
4. 1.~3.を繰り返す

CMA-ES



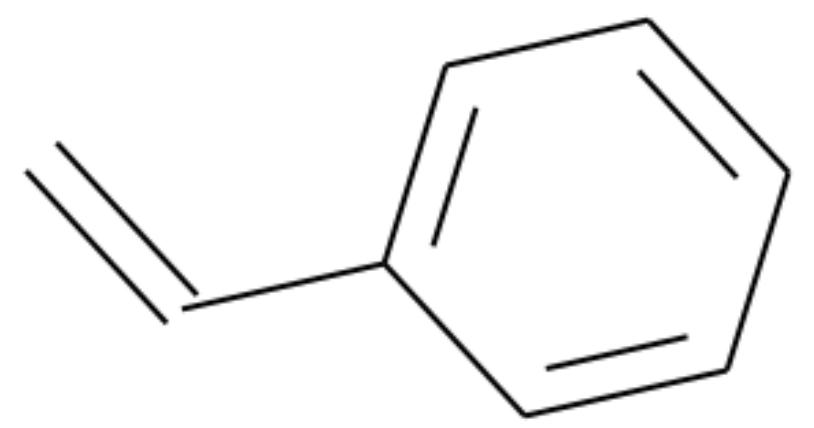
1. 正規分布から解を生成
2. 全ての解を評価して重み付けする
3. 正規分布のパラメータを更新
4. 1.~3.を繰り返す

CMA-ES



1. 正規分布から解を生成
2. 全ての解を評価して重み付けする
3. 正規分布のパラメータを更新
4. 1.~3.を繰り返す

SMILESとCanonical SMILES



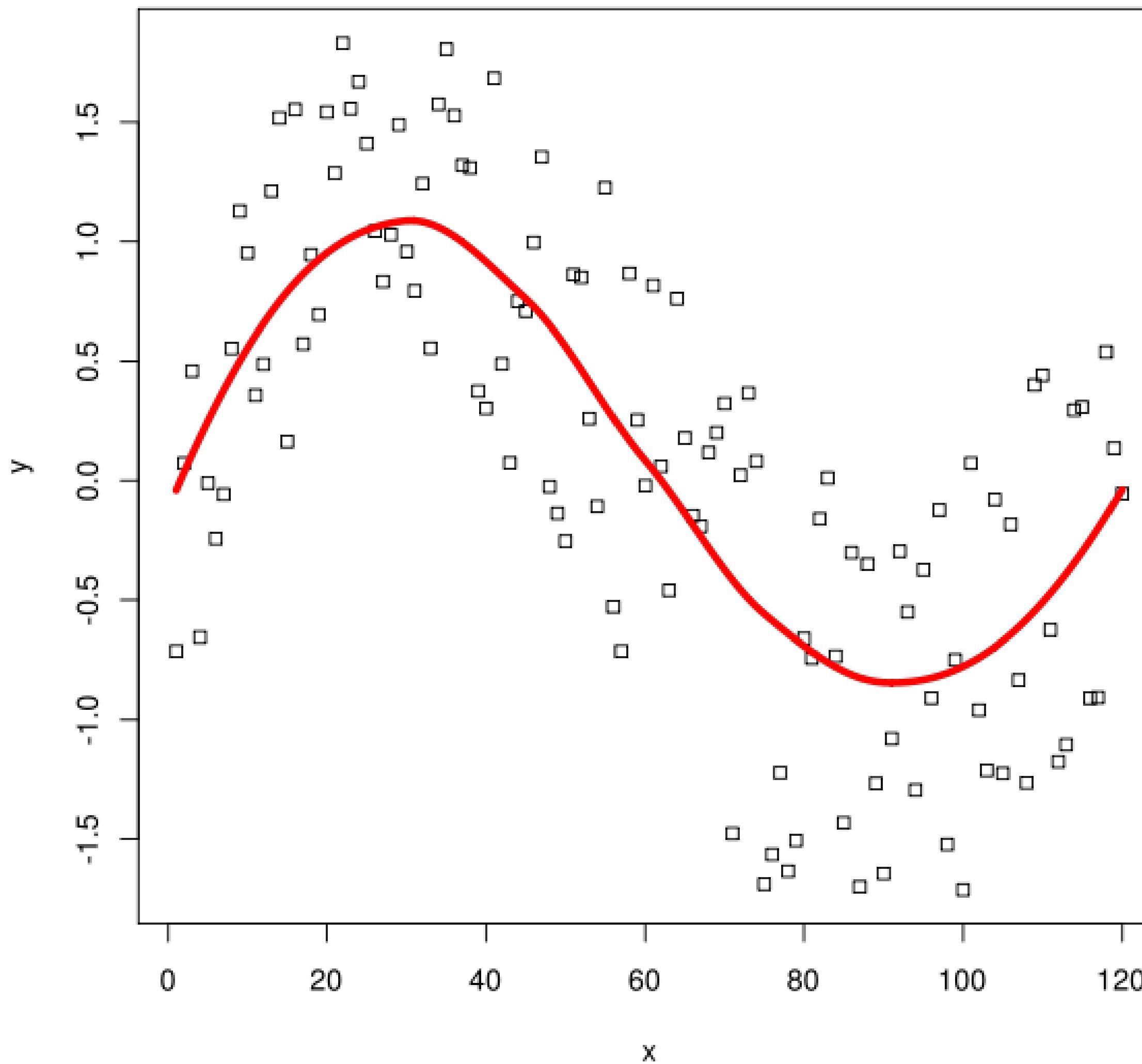
SMILES

- C=CC1=CC=CC=C1
- C1=CC=CC=C1C=C
- C1=C(C=C)C=CC=C1

ルールに従って並び順を決め（正規化 = canonicalize）、それに従って記述しましょう、というのがcanonical SMILESです。

LOWESS Smoothing

LOWESS（局所加重散布図平滑化）、時にはLOESS（局所加重平滑化）とも呼ばれます。回帰分析で使用される一般的なツールです。これは、タイムプロットや散布図にスムーズな線を引き、変数間の関係を見やすくし、傾向を予測するのに役立ちます。



Tanimoto類似度

$$S_{AB} = \frac{c}{a + b - c}$$

- aは分子Aのビット配列で1が立っている数
- bは分子Bのビット配列で1が立っている数
- cは分子AとBで共通に1が立っている数

分子A:

1	0	1	1	1	0	1	1	0	0	1	1
---	---	---	---	---	---	---	---	---	---	---	---

分子B:

0	0	1	1	0	0	1	0	1	0	1	1
---	---	---	---	---	---	---	---	---	---	---	---

$$S_{AB} = \frac{5}{8 + 6 - 5} = 0.56$$

QED

QED (Quantitative Estimate of Drug-likeness) は複数の分子記述子を組み合わせることで薬らしさを評価する方法の1つです。

QEDは771個の経口医薬品をデータセットに用いて、次の8つの記述子を用いてモデル化されました。QEDを用いることで、0（最も薬らしくない）から1（最も薬らしい）に定量化が可能です。

- 分子量 (MW)
- logP (ALOGP)
- 水素結合ドナーの数 (HBDs)
- 水素結合アクセプターの数 (HBAs)
- 極性表面積 (PSA)
- 回転可能結合数 (ROTBs)
- 芳香環の数 (AROMs)
- 忌避構造の数 (ALERTS)