# Scientific Discovery: From the GPU to the Lab Bench

Michelle L. Gill, PhD; Applied Research Manager, Life Sciences, NVIDIA

May 11th, 2024

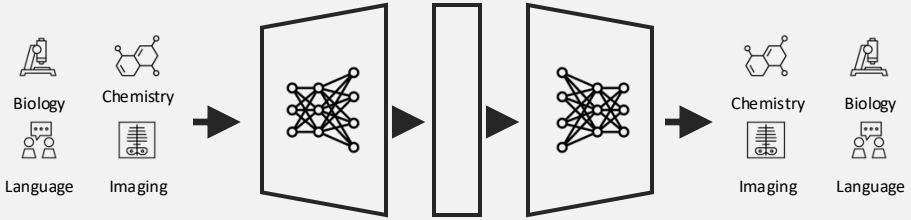
# Outline

- Foundation model development for science -- small molecules, proteins, and genomics

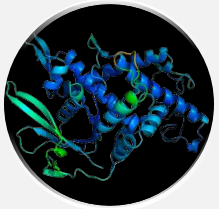
- Advice for scientists in the age of artificial intelligence

# Language Models in Scientific Discove

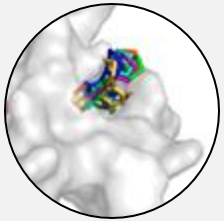
- Information from biomedical literature
- Protein structure prediction and ligand docking
- Prediction of chemical reactions
- Biomolecular property prediction




BIOMEDICAL NLP
Learn all of PubMed



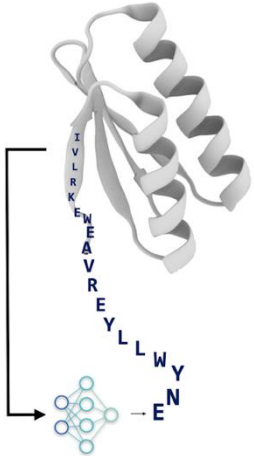
PROTEIN STRUCTURE
Predict 3D Structures



VIRTUAL SCREENING
Docking and Pose Prediction

# From Sequence to 3D and Back Again

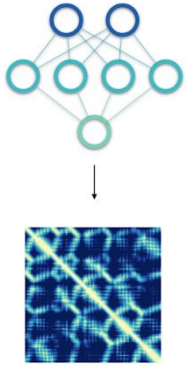
## 1 Fixed-backbone design




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

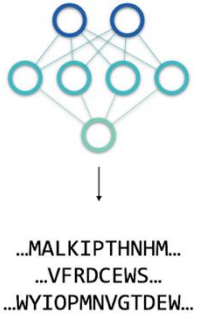
## 2 Structure Generation



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

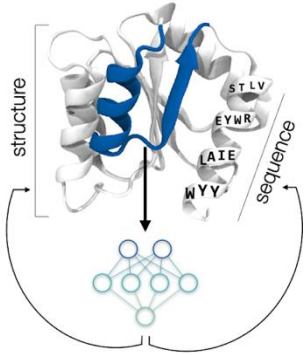



Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022)

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

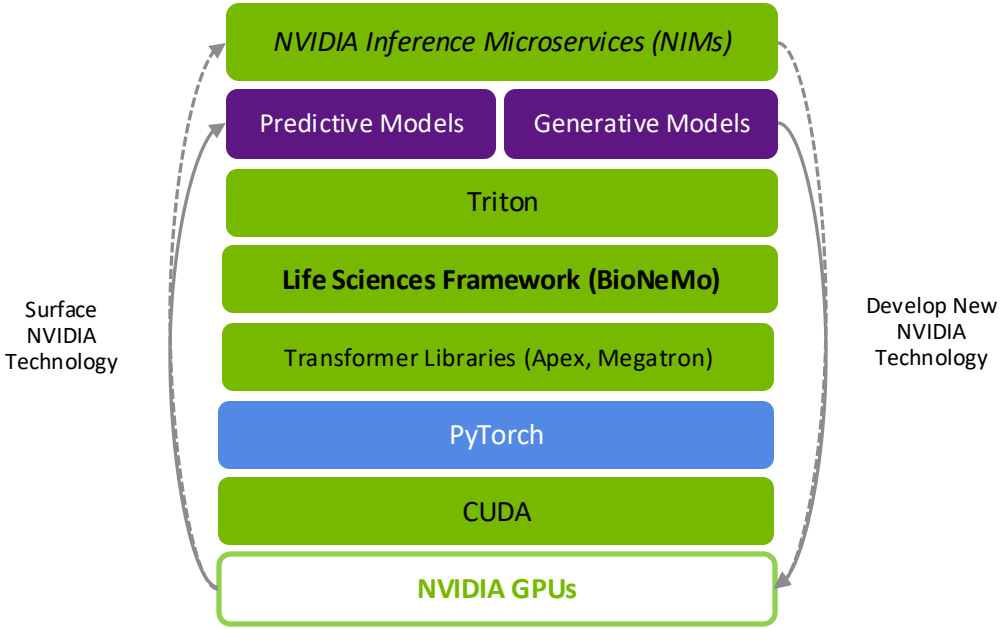
Jin, W., Wohlwend, J., Barzilay, R. & Jaakkola, T. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. *arXiv [q-bio.BM]* (2021)

Ferruz, N. *et al.* From sequence to function through structure: deep learning for protein design. *bioRxiv* 2022.08.31.505981 (2022) doi:10.1101/2022.08.31.505981

# What is a Foundation Model?

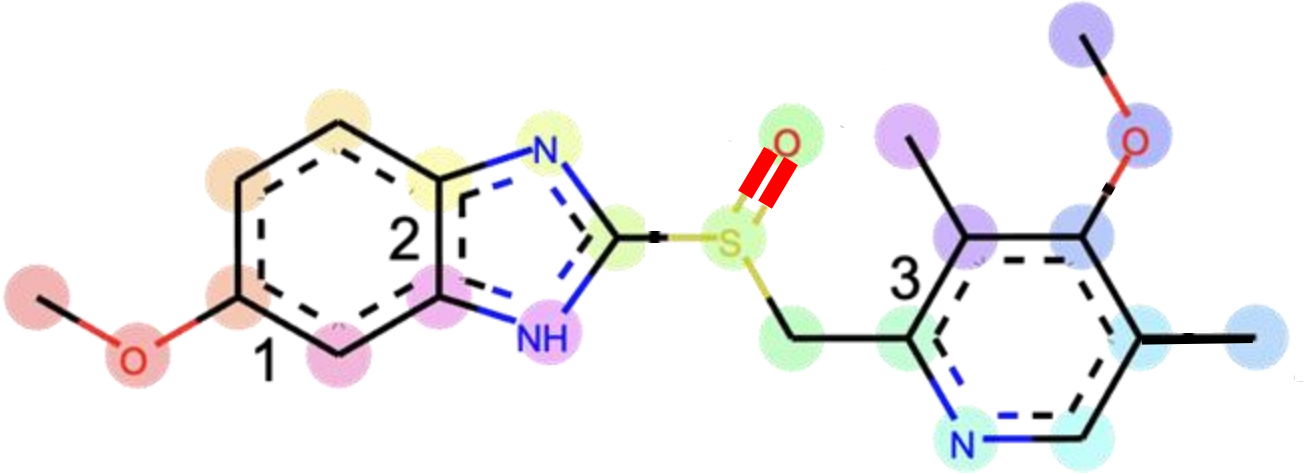
- **Large scale (pre-)training** – models are trained on vast amounts of data, often multiple topics and modalities
- **Generality** – capable of performing many different functions
- **Adaptability and fine tuning** -- general purpose models can be specialized for desired task
- **Accessibility** – pre-trained models serve as a starting point for researchers to build upon
- **Emergence** – very large models can develop capabilities beyond those that they were trained to perform

# NVIDIA Generative AI Life Sciences Software Stack



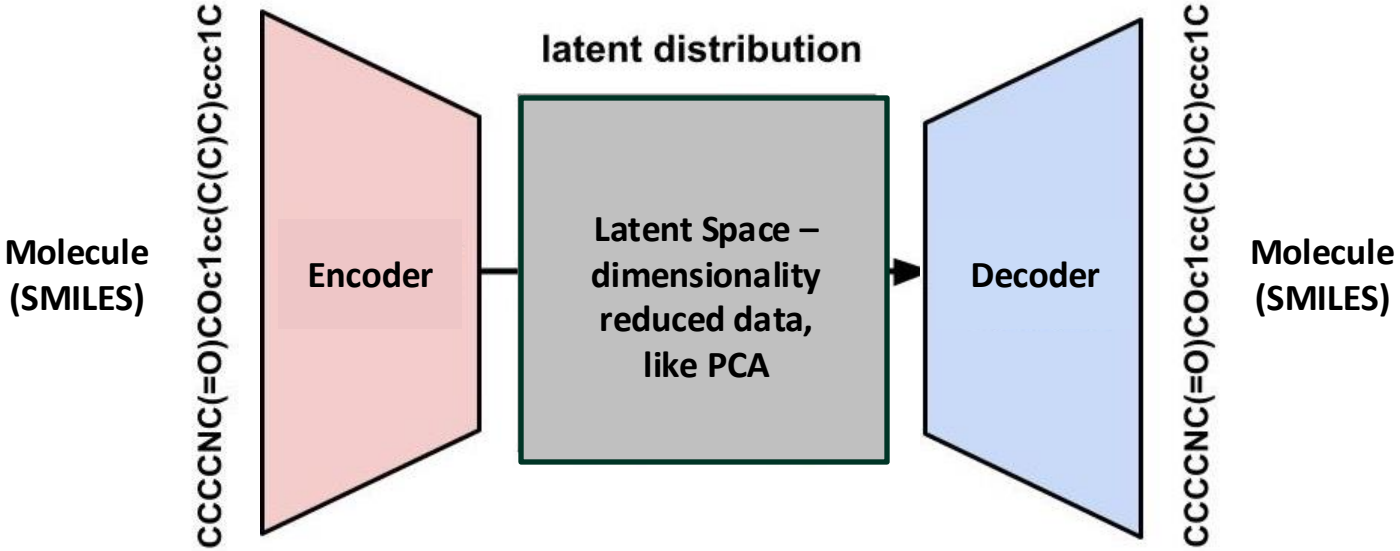

- Surface new technology from NVIDIA hardware and software; and feedback domain specific advancements to improve them
- GPU-accelerated life sciences frameworks, e.g. BioNeMo, depend on CUDA and accelerated deep learning libraries
- NVIDIA deployment libraries and (soon) microservices bring accelerated model inference and APIs to researchers and developers

# SMILES: a Natural Language Representation of Small Molecules

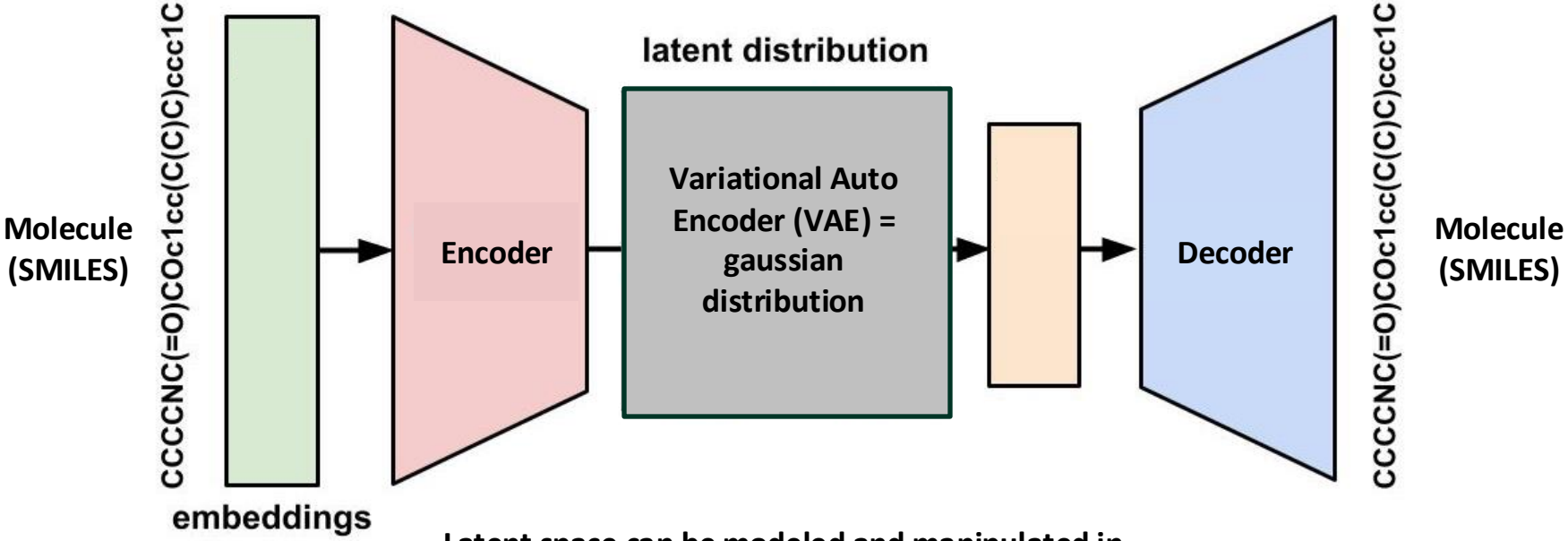



COc1ccc2n c(S (=O) Cc3ncc(C) c(OC)c3C) [nH]c2c

# Anatomy of an Auto Encoder Model



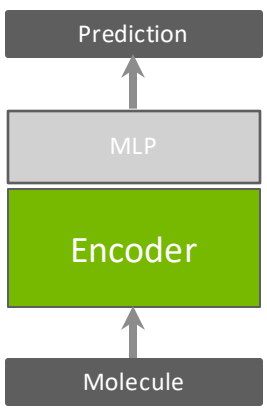
# Deep Learning Models as Lego Blocks

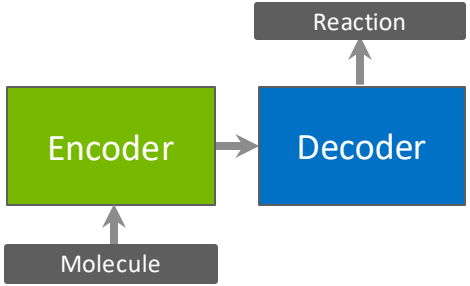



Latent space can be modeled and manipulated in many ways

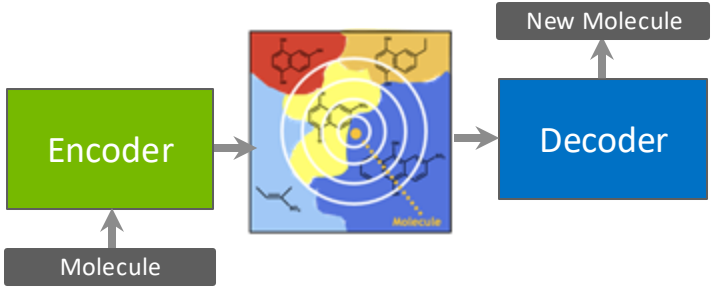
# Objectives of a Cheminformatics Foundation Model

## Representation and Translation



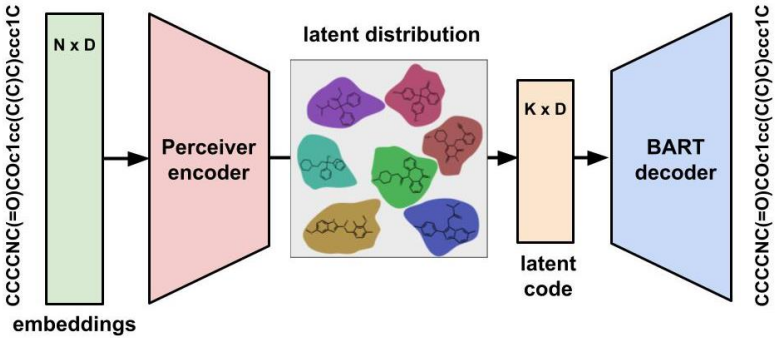

## Generation



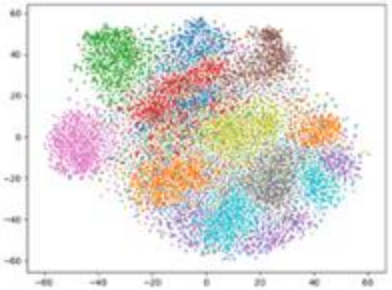

Cheminformatics foundation models can be applied to a wide range of predictive tasks (physical chemical properties, retrosynthesis) and the generation of novel molecules

# A Clustered Latent Space with Mutual Information Machine (MIM)

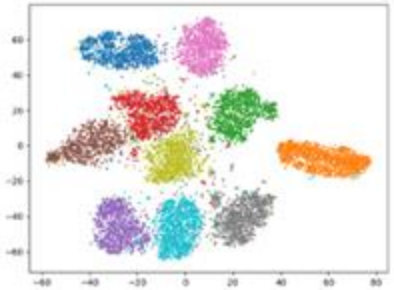
- A variational autoencoder (VAE) loss smooths the latent space resulting in blurring
- MIM loss results in a clustered space




VAE



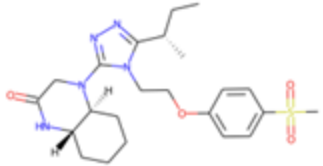
MIM



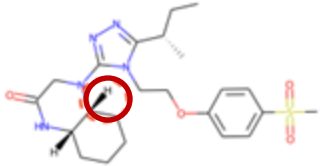
Danny Reidenbach, Micha Livne, Rajesh Illango

# MolMIM – Sampling Distance Can Be Tuned for Similarity

**Small Perturbations**

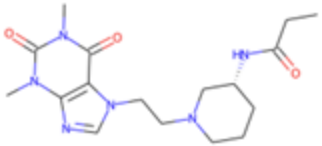


Seed Molecule

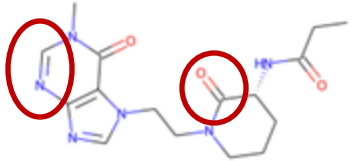


Sampled Molecule

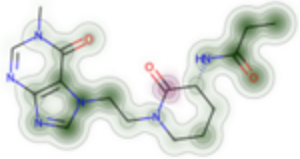
**Larger Perturbations**



Seed Molecule

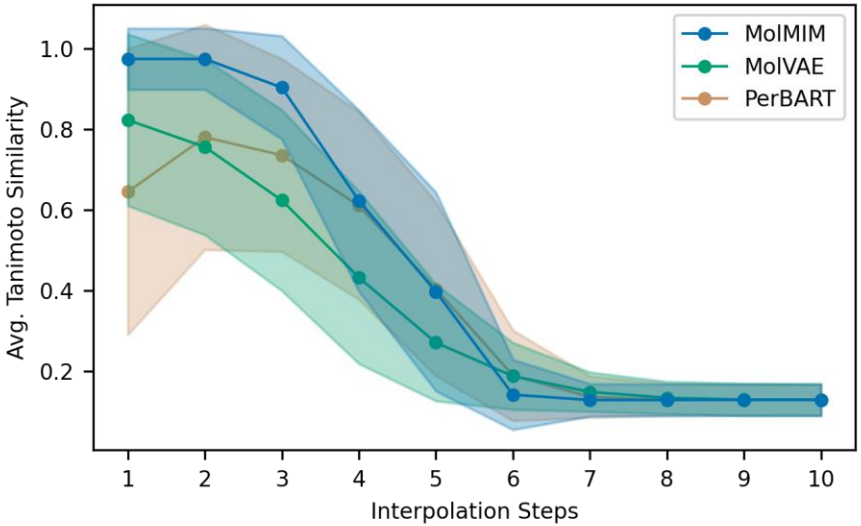


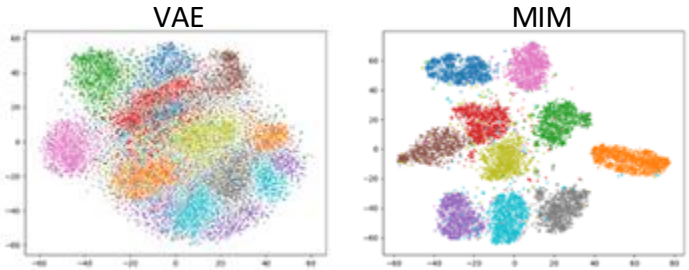
Sampled Molecule



Similarity Map

# Probing Latent Structure by Molecule Interpolation



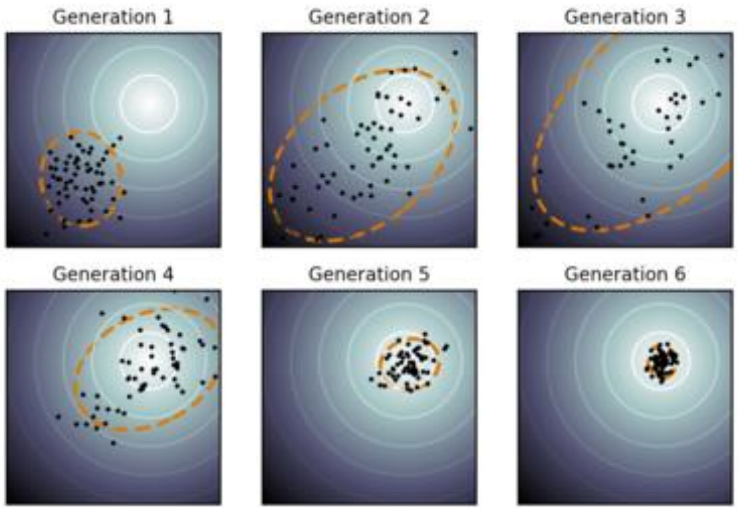




- Pairwise interpolations between 1,000 molecules performed at ten evenly spaced steps
- Similarity between starting molecule and each interpolated molecule calculated
- Molecules sampled from baseline models (PerBART, MolVAE) have reduced similarity at start and high variance at early interpolation steps
- MolMIM molecules are similar to each other and have smallest variance at initial steps

Danny Reidenbach, Micha Livne, Rajesh Illango

# 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).

# Multi-Objective Property Optimization

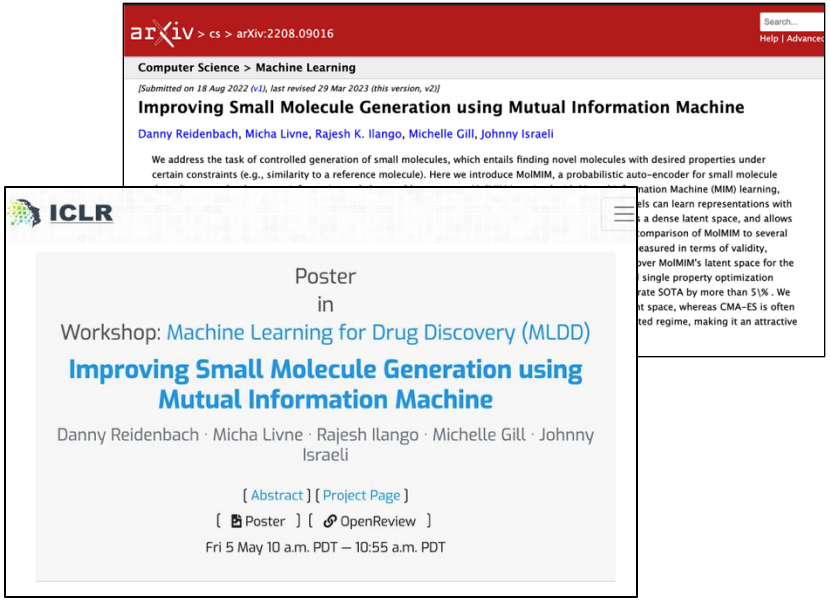
- Performed multi-objective molecule optimization to jointly optimize two molecular properties (QED and SA), and binding to two targets (JNK3 and GSK4β).
- Objective was to maximize success, novelty, and diversity metrics.
- Optimization methods:
  - *Random*: subset of randomly selected molecules
  - *Approximate*: subset of molecules that partially satisfy optimization criteria
  - *Exemplar*: subset of molecules that satisfy all criteria
- MolMIM is competitive for success and diversity -- novelty has since been improved considerably

| 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 |

Danny Reidenbach, Micha Livne, Rajesh Illango

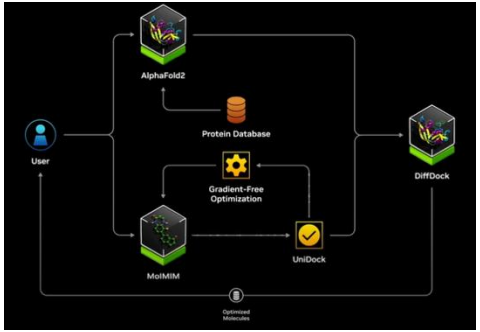
Results above solid bar as in B. Chen, X. Fu, R. Barzilay, T. Jaakkola, ArXiv (2021)
QED, SA, JNK3, and GSK4β oracles from Therapeutic Data Commons

# MolMIM: Applied Research to Productization

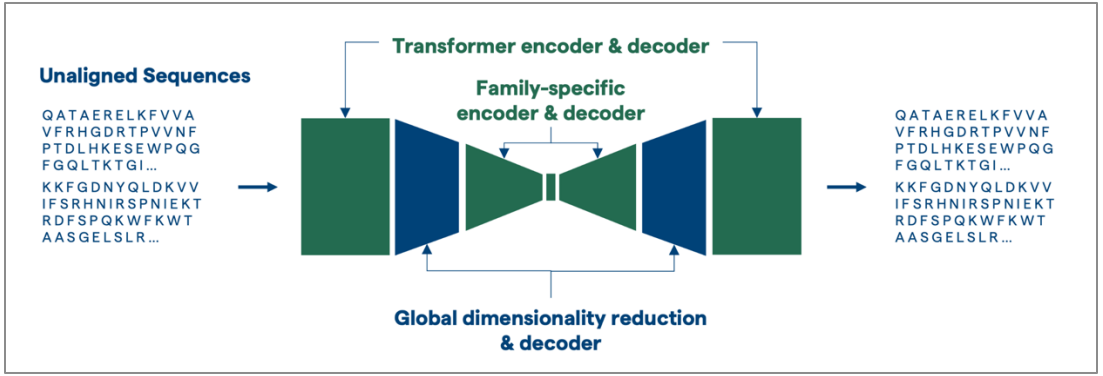



- MolMIM and controlled generation is hallmark feature of BioNeMo NIMs
- Model released on BioNeMo framework and accelerated inference workflows for controlled generation available soon on NIMs
- *On-going work:*
  - Improving encoder representations to make MolMIM well-rounded foundation model
  - Development of more comprehensive benchmarks

MolMIM Featured in Jensen's 2024 GTC Keynote:



# Improving Enzyme Function with Protein Language Models (I)

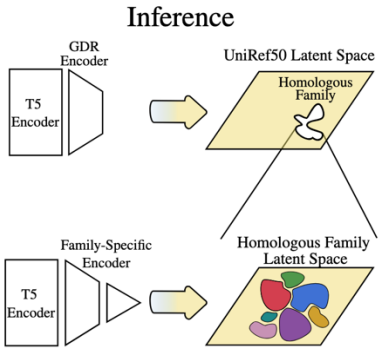


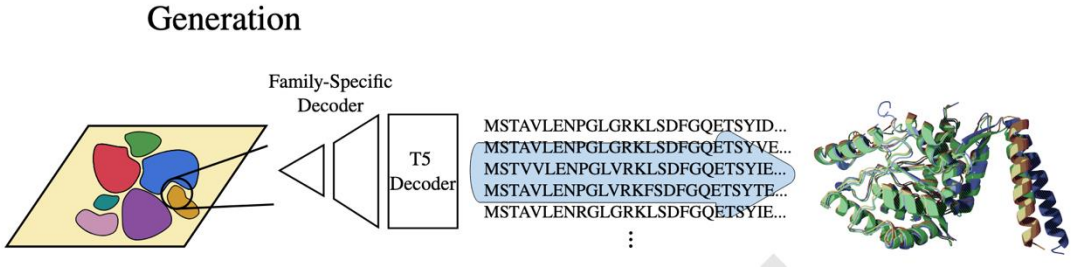

ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design

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.



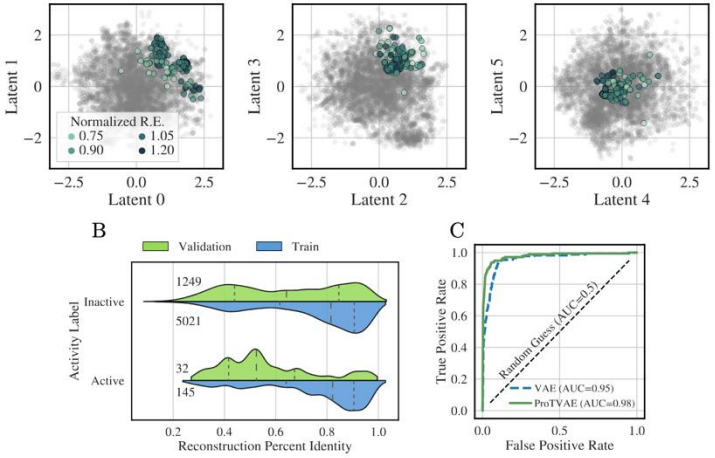




Maria Korshunova, Micha Livne

E. Sevgen, *et al.*, bioRxiv, doi:10.1101/2023.01.23.525232.
*Under review*

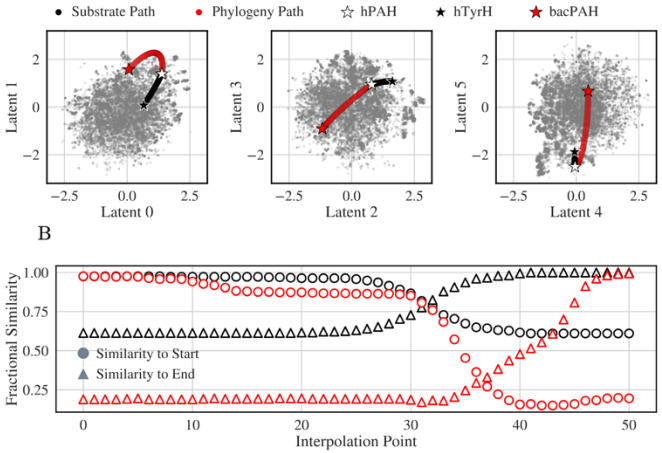
# Improving Enzyme Function with Protein Language Models (II)

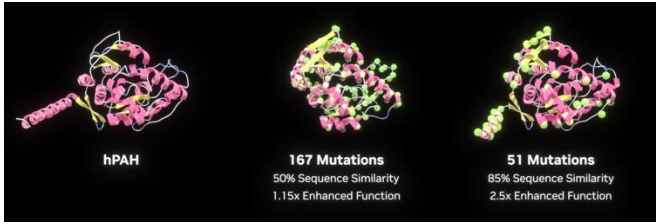
## Src Homology 3 (SH3)




*In vivo* assay that measures incorporation of designed SH3 constructs in *S. cerevisiae* by relative enrichment of sequencing counts

## Phenylalanine Hydroxylase (hPAH)




Maria Korshunova, Micha Livne

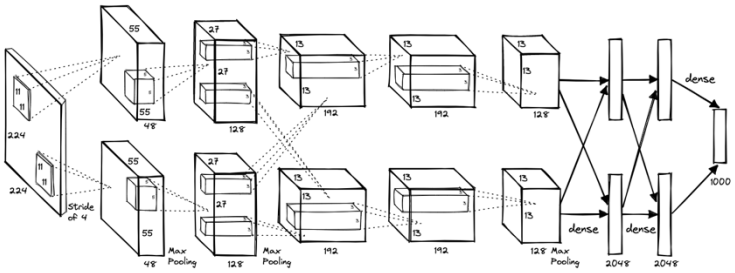
E. Sevgen, *et al.*, bioRxiv, doi:10.1101/2023.01.23.525232.
*Under review*

# Outline

- Foundation model development for science -- small molecules, proteins, and genomics
- Advice for scientists in the age of artificial intelligence

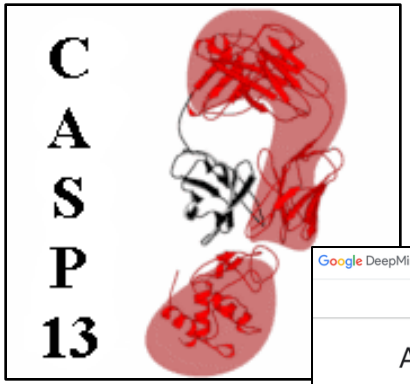
# Don’t Miss the Forest Through the (Chemis)Trees

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.

AlphaFold Won CASP13 in 2018




Google DeepMind About Technologies Impact Discover

Overview Blog The Podcast Visualising AI

RESEARCH

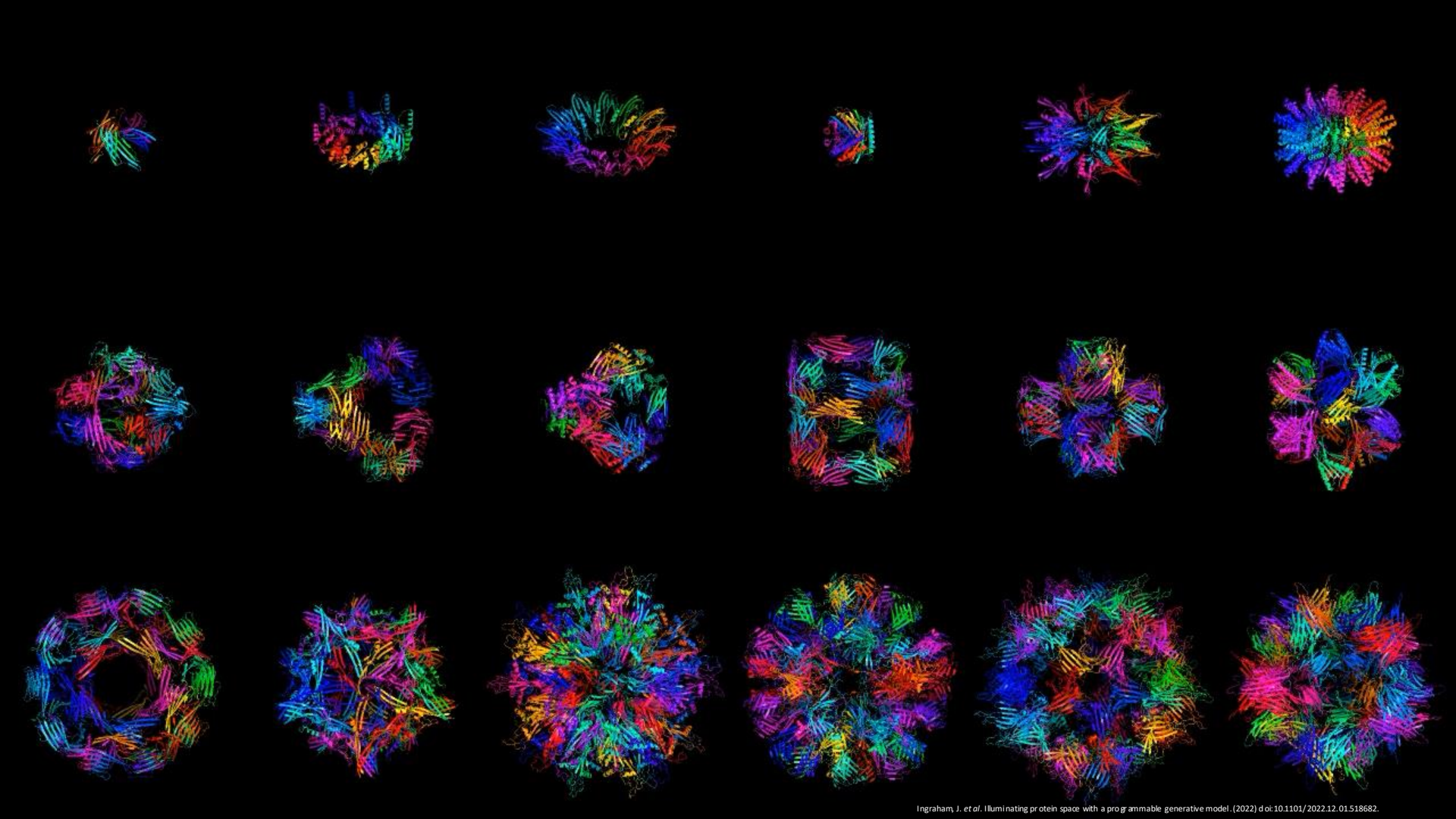
AlphaFold: a solution to a 50-year-old grand challenge in biology

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

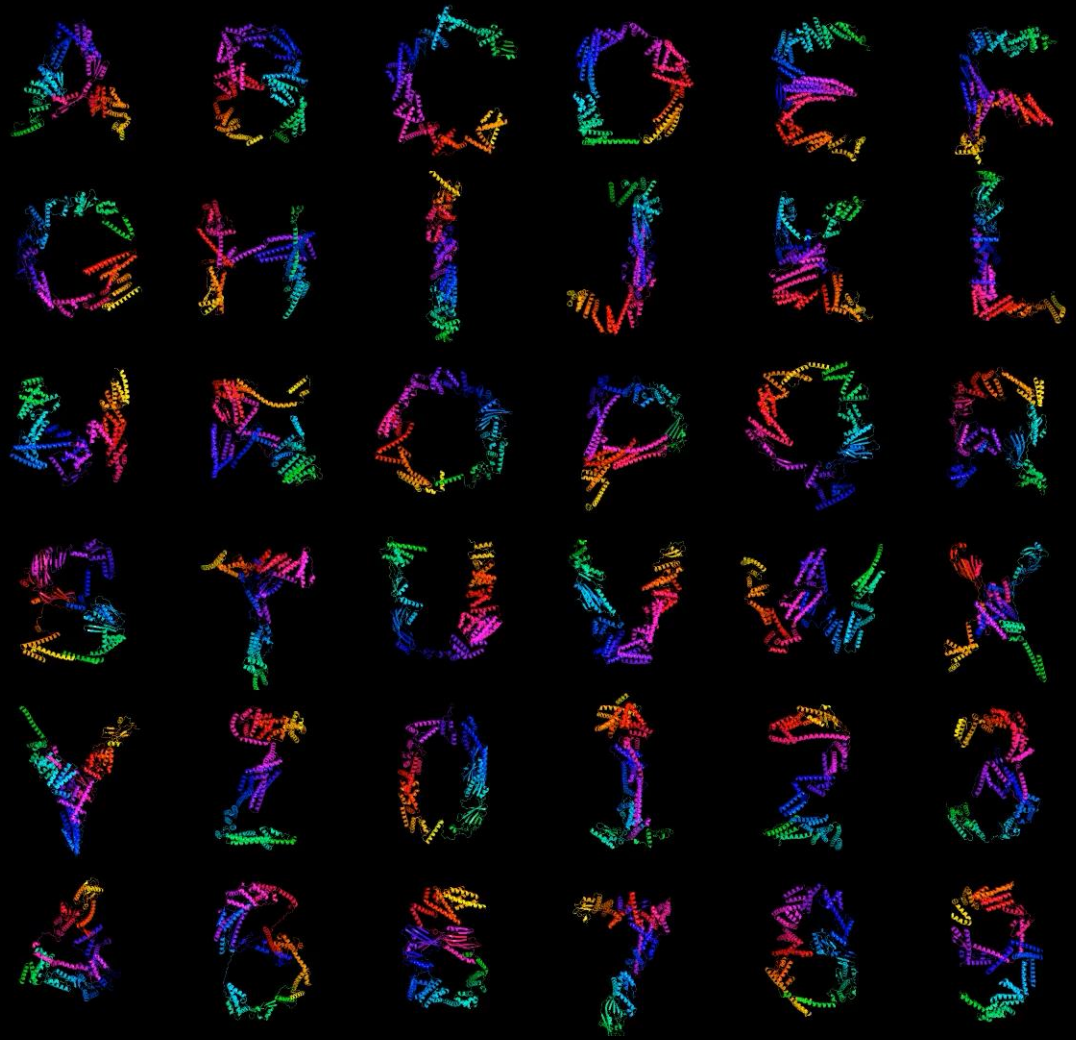
Dec 14, 2022 — Two years later, **AlphaFold** still **dominates** the competition. Deepmind itself did not participate in this round, but **AlphaFold** has been open ...

# 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

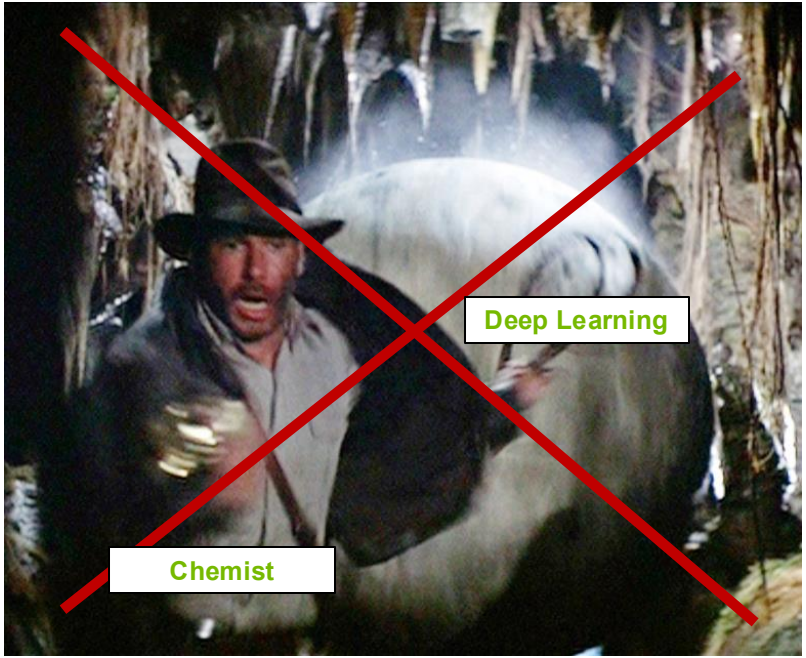


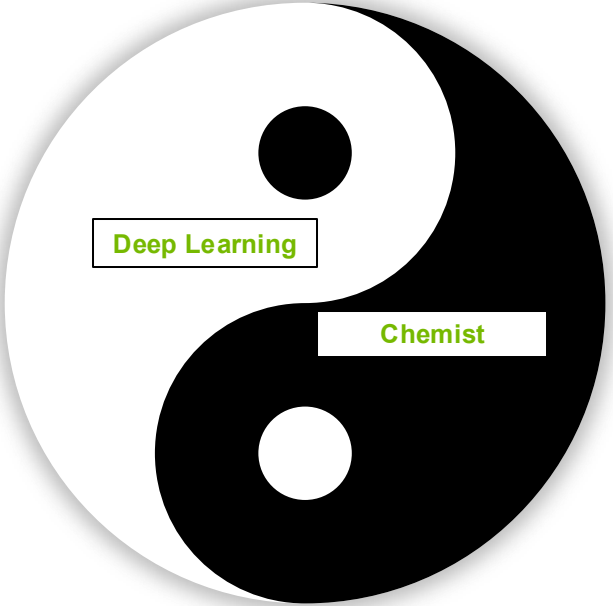
Ingraham, J. *et al*. Illuminating protein space with a programmable generative model. (2022) doi:10.1101/2022.12.01.518682.



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

# Chemistry and Deep Learning are Complementary



# Conclusions

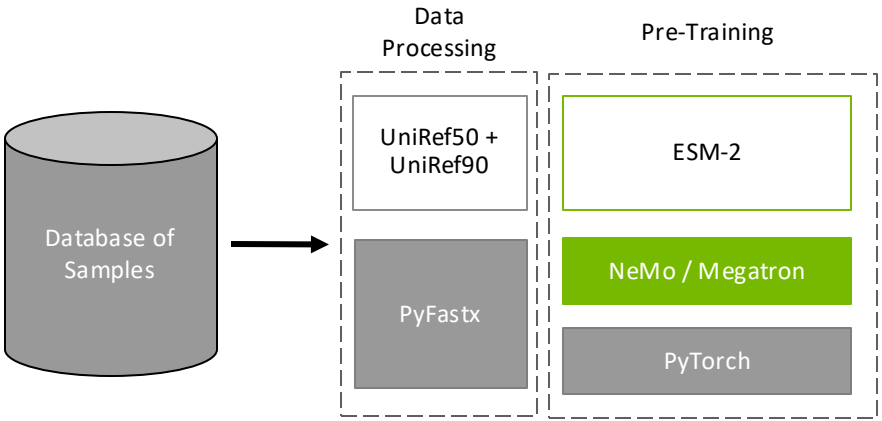
- MolMIM is a cheminformatics model for generation and design of small molecule therapeutics
- Novel enzymes with improved functionality can be designed by protein language models, like ProT-VAE
- Deep learning is a powerful tool for scientific discovery (not a panacea)

# Thank You!

mgill@nvidia.com

# Appendix

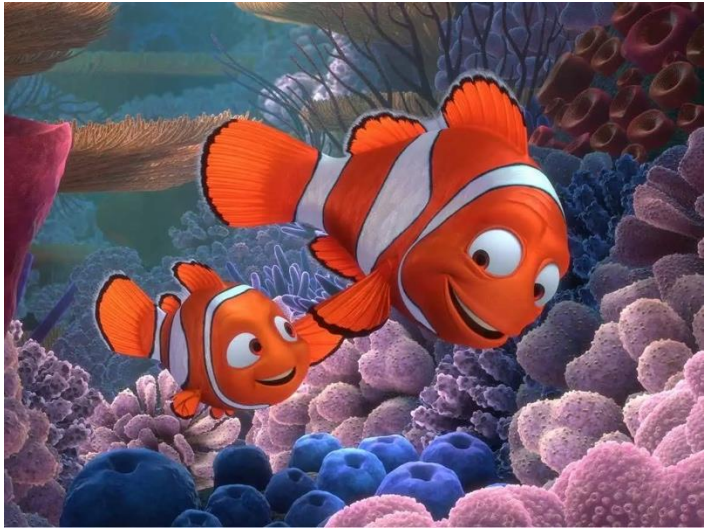
# Developing Deep Learning Models at Scale




| 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.**

# The BioFoundation Model and BioNeMo Teams

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 Veccham

Steven Kothen-Hill

Tomasz Grzegorzek

Timur Rvachov

Yuxing Peng

Zachary McClure