NVIDIA

Scientific Discovery: From the Lab Bench to the GPU

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

April 19th, 2024

Outline

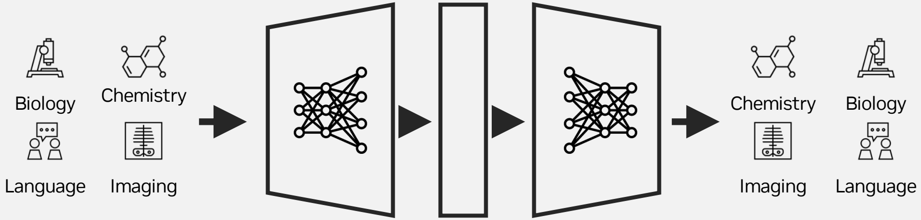
- Foundation model development for science -- small molecules, proteins, and genomics

- What I learned in Andy’s group; and advice for NMR spectroscopists and scientists in the age of AI

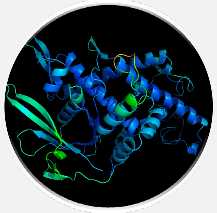
BYRD
FEST

Language Models in Scientific Discovery

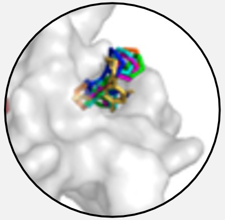
- Information from biomedical literature
- Protein structure prediction and ligand docking
- Prediction of chemical reactions
- Biomolecular property prediction



BIOMEDICAL NLP
Learn all of PubMed



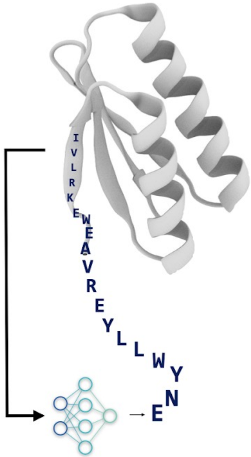
PROTEIN STRUCTURE
Predict 3D Structures



VIRTUAL SCREENING
Docking and Pose Prediction

From Sequence to 3D and Back Again

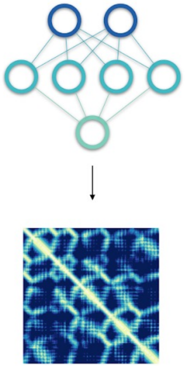
1 Fixed-backbone design



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

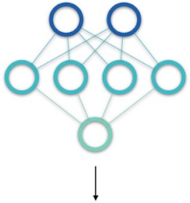
2 Structure Generation



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



...MALKIPTHNHM...

...VFRDCEWS...

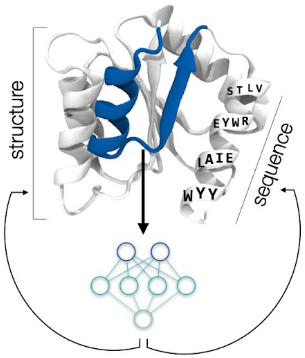
...WYIOPMNVGTDEW...

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

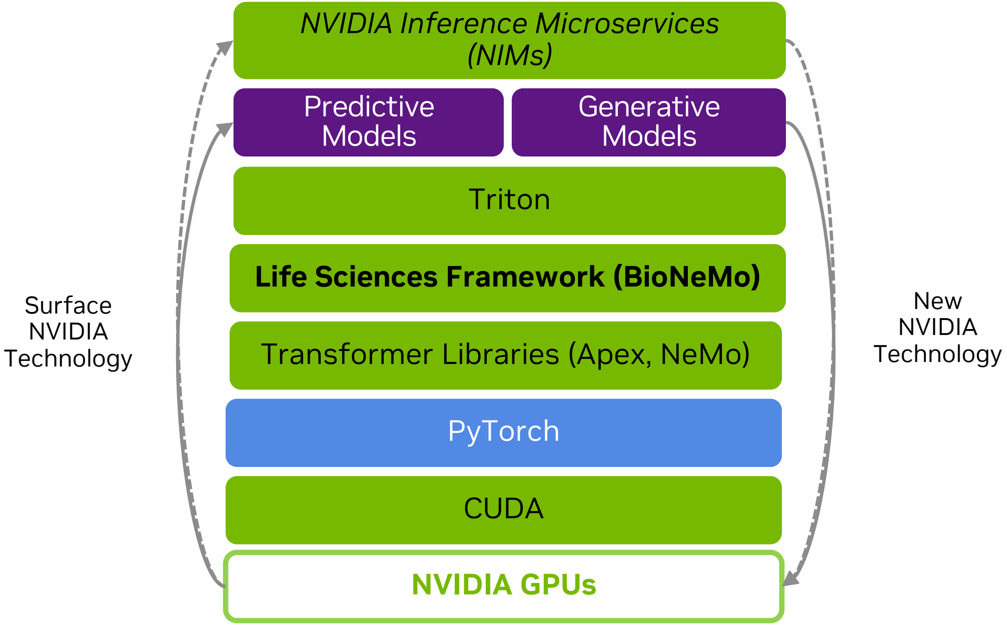
Jin, W., Wohlwend, J., Barzilay, R. & Jaakkola, T. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. *arXiv [q-bio.BM]* (2021)

Ferruz, N. *et al.* 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

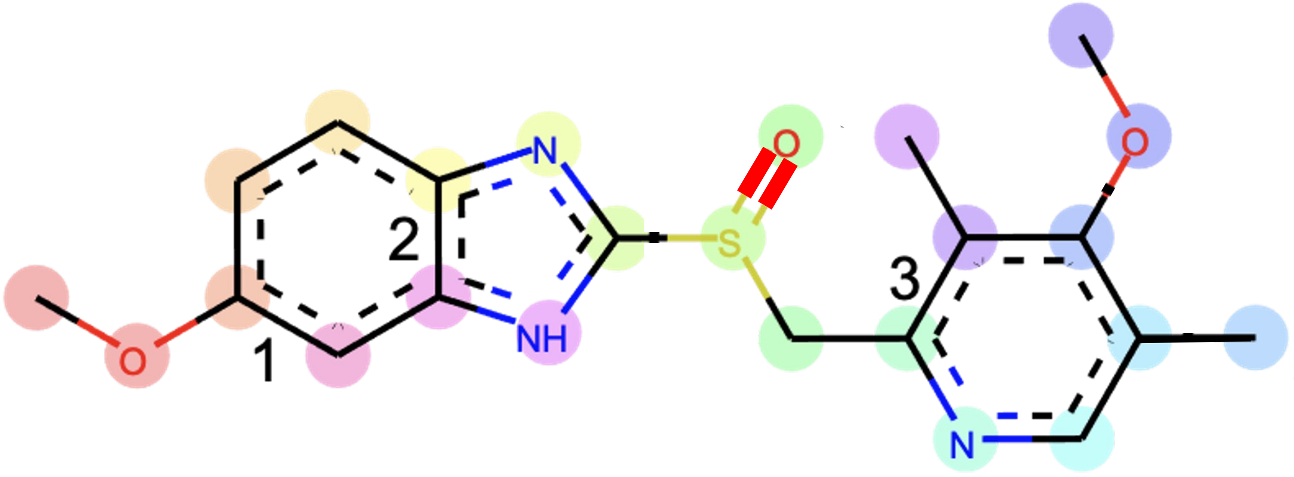
NVIDIA Generative AI Life Sciences Software Stack



NVIDIA Inference Microservices (NIMs)
Predictive Models
Generative Models
Triton
Life Sciences Framework (BioNeMo)
Transformer Libraries (Apex, NeMo)
PyTorch
CUDA
NVIDIA GPUs
Surface NVIDIA Technology
New NVIDIA Technology

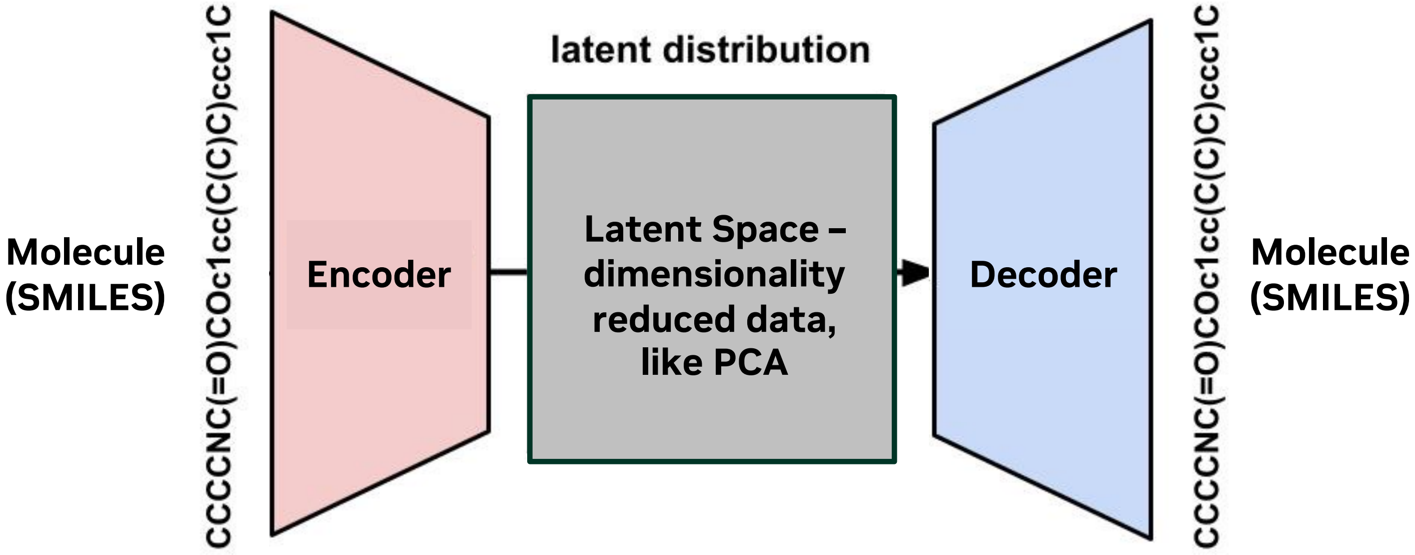
- Surface new technology from NVIDIA hardware and software; and feedback domain specific advancements to improve them
- GPU-accelerated life sciences frameworks, e.g. BioNeMo, depend on CUDA and accelerated deep learning libraries
- NVIDIA deployment libraries and (soon) microservices bring accelerated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

SMILES: a Natural Language Representation of Small Molecules



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

Anatomy of an Auto Encoder Model



Molecule (SMILES)

CCCCNC(=O)COc1cc(C(C)C)ccc1C

Encoder

latent distribution

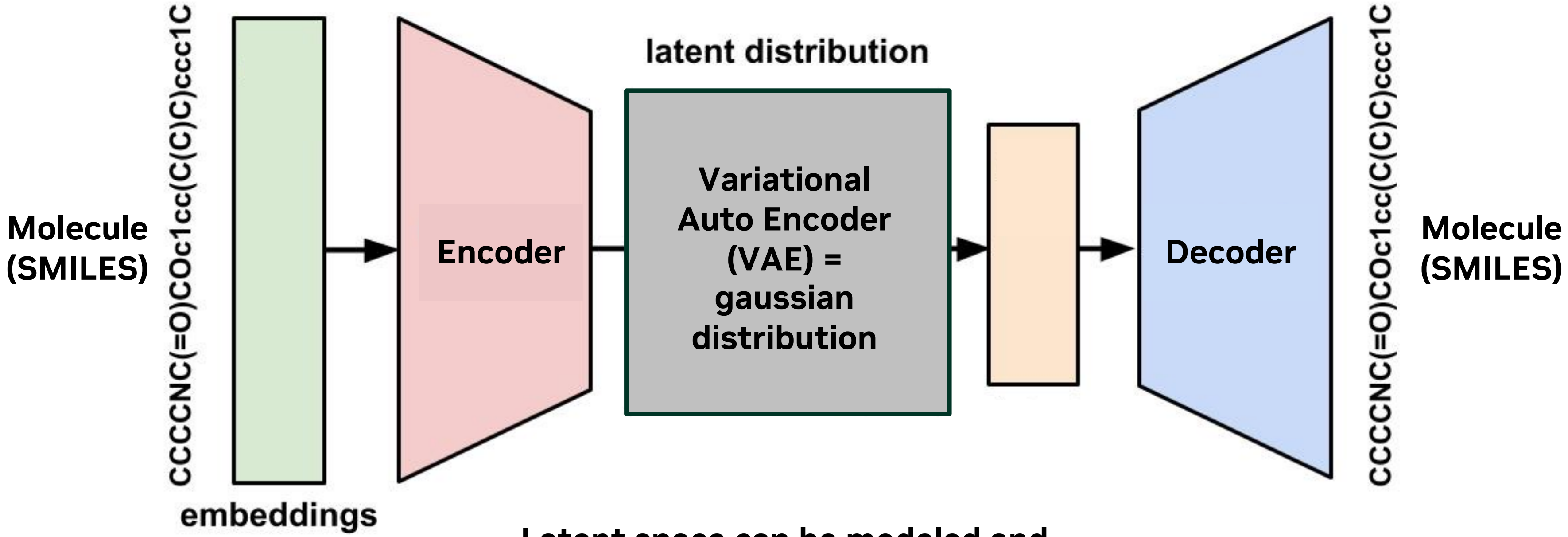
Latent Space – dimensionality reduced data, like PCA

Decoder

CCCCNC(=O)COc1cc(C(C)C)ccc1C

Molecule (SMILES)

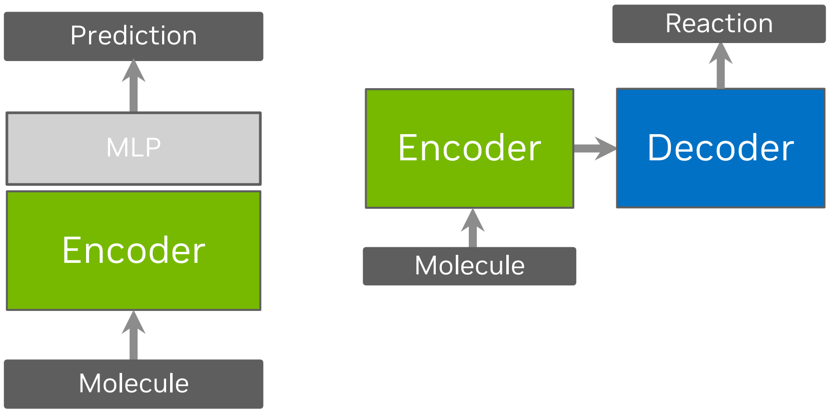
Deep Learning Models as Lego Blocks


Molecule (SMILES)
CCCCNC(=O)COc1cc(C(C)C)ccc1C
embeddings
Encoder
latent distribution
Variational Auto Encoder (VAE) = gaussian distribution
Decoder
CCCCNC(=O)COc1cc(C(C)C)ccc1C
Molecule (SMILES)

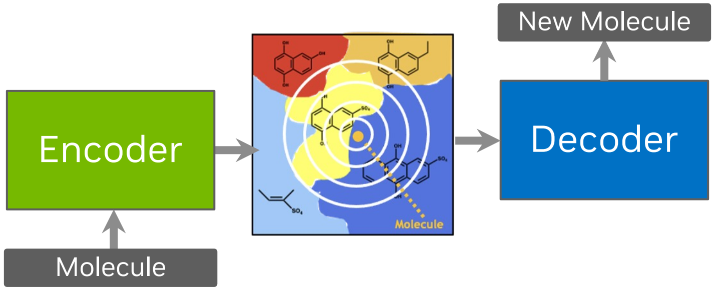
Latent space can be modeled and manipulated in many ways

Objectives of a Cheminformatics Foundation Model

Representation and Translation



Generation

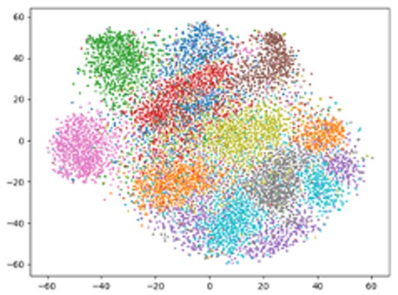


Cheminformatics foundation models can be applied to a wide range of predictive tasks (physical chemical properties, retrosynthesis) and the generation of novel molecules

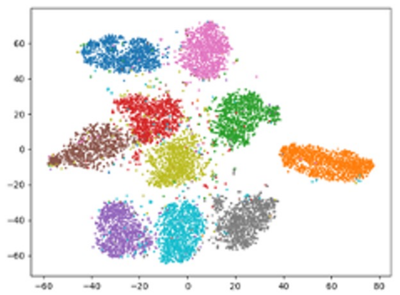
A Clustered Latent Space with Mutual Information Machine (MIM)

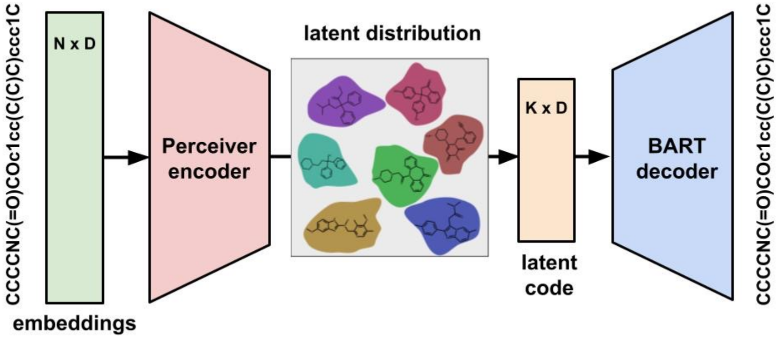
- A variational autoencoder (VAE) loss smooths the latent space resulting in blurring
- MIM loss results in a clustered space

VAE



MIM

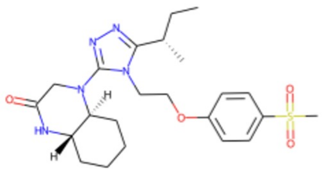




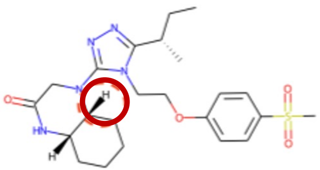
Danny Reidenbach, Micha Livne, Rajesh Illango

MolMIM – Sampling Distance Can Be Tuned for Similarity

Small Perturbations

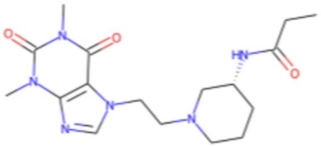


Seed Molecule

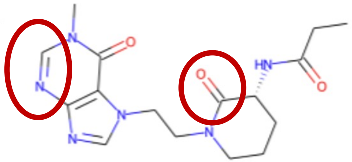


Sampled Molecule

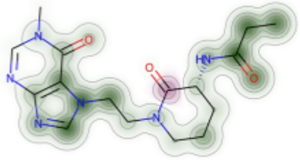
Larger Perturbations



Seed Molecule

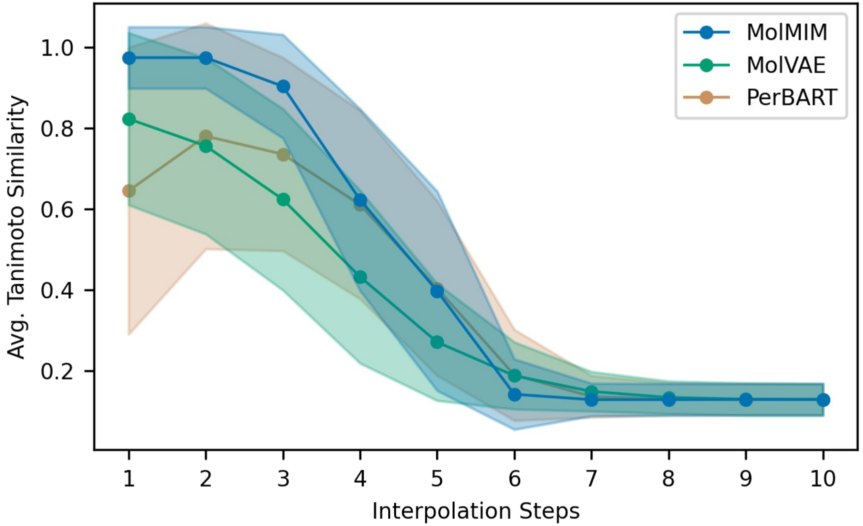


Sampled Molecule



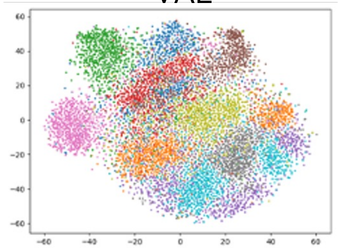
Similarity Map

Probing Latent Structure by Molecule Interpolation

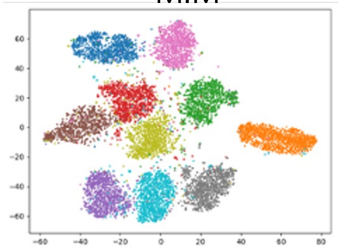


- Pairwise interpolations between 1,000 molecules performed at ten evenly spaced steps
- Similarity between starting molecule and each interpolated molecule calculated
- Molecules sampled from baseline models (PerBART, MolVAE) have reduced similarity at start and high variance at early interpolation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

VAE



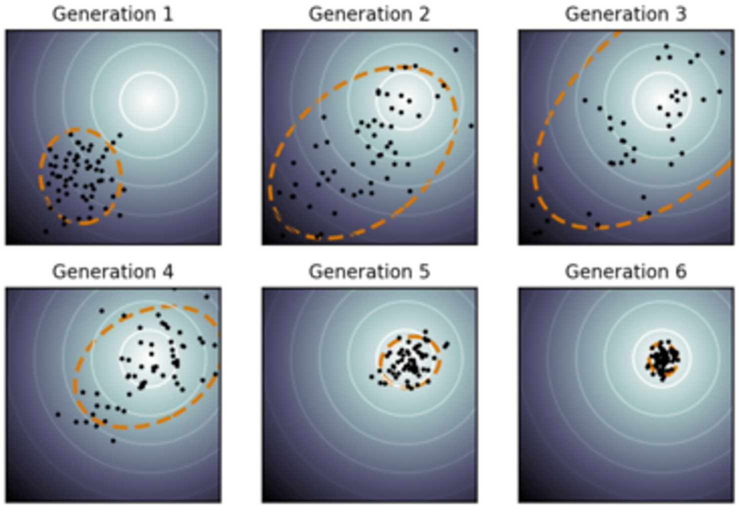
MIM



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

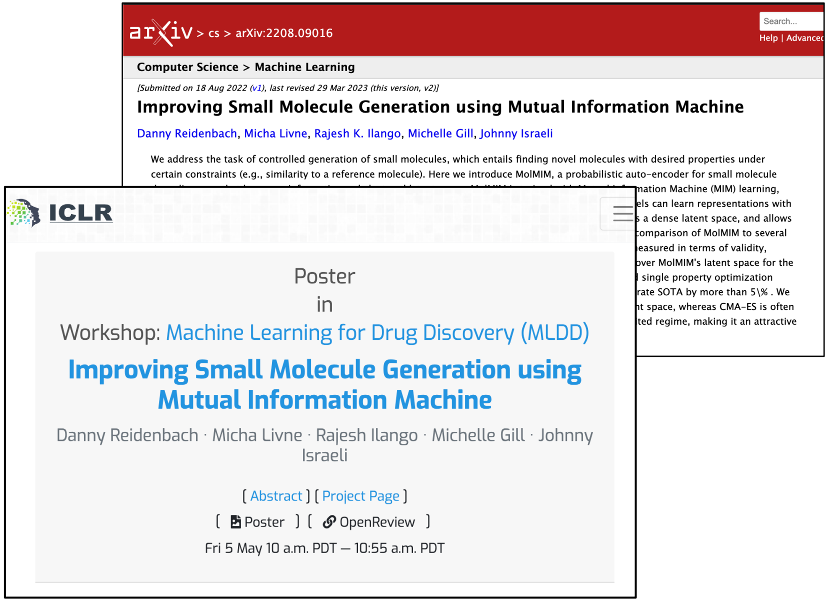
- Performed multi-objective molecule optimization to jointly optimize two molecular properties (QED and SA), and binding to two targets (JNK3 and GSK4β).
- Objective was to maximize success, novelty, and diversity metrics.
- Optimization methods:
  - *Random*: subset of randomly selected molecules
  - *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



arXiv > cs > arXiv:2208.09016

Search...

Help | Advanced

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
mation Machine (MIM) learning,
els can learn representations with
s a dense latent space, and allows
comparison of MolMIM to several
easured in terms of validity,
over MolMIM's latent space for the
single property optimization
rate SOTA by more than 5\% . We
nt space, whereas CMA-ES is often
ted regime, making it an attractive

ICLR

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

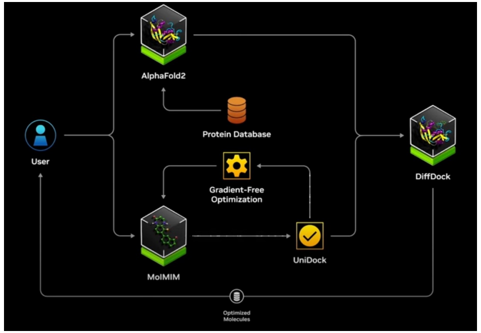
[Abstract] [Project Page]

[Poster] [OpenReview]

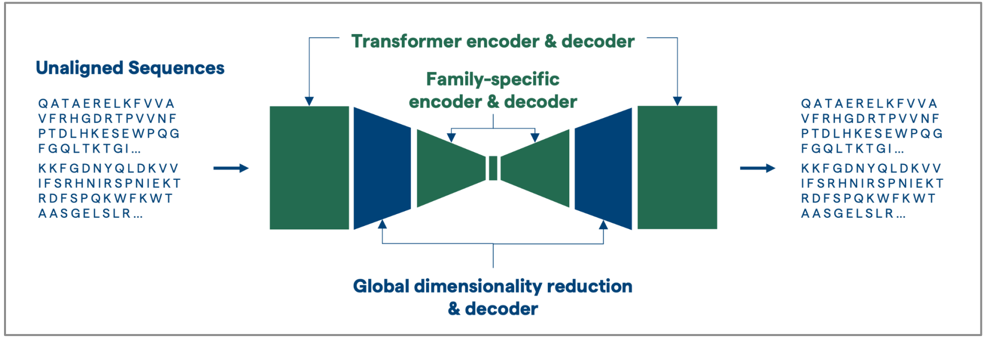
Fri 5 May 10 a.m. PDT — 10:55 a.m. PDT

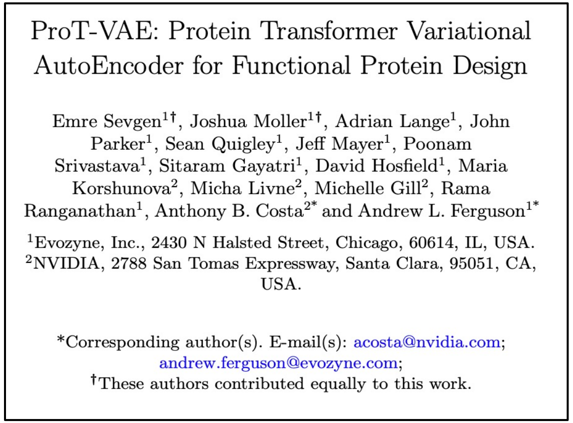
- MolMIM and controlled generation is hallmark feature of BioNeMo NIMs
- Model released on BioNeMo framework and accelerated inference workflows for controlled generation available soon on NIMs
- *On-going work:*
  - Improving encoder representations to make MolMIM well-rounded foundation model
  - Development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

MolMIM Featured in Jensen's 2024 GTC Keynote:



Improving Enzyme Function with Protein Language Models





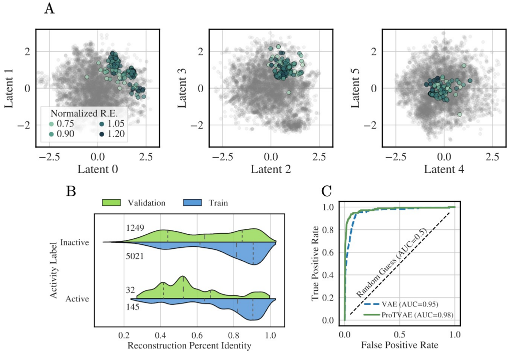
ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design

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*]

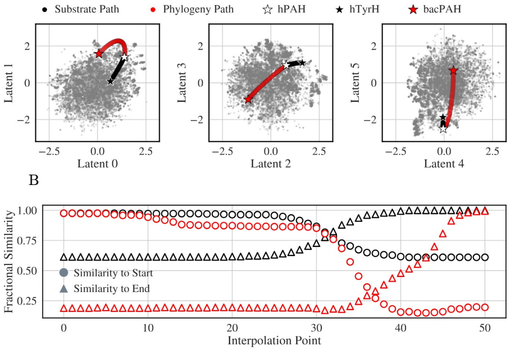
[1]Evozyne, Inc., 2430 N Halsted Street, Chicago, 60614, IL, USA.
[2]NVIDIA, 2788 San Tomas Expressway, Santa Clara, 95051, CA, USA.

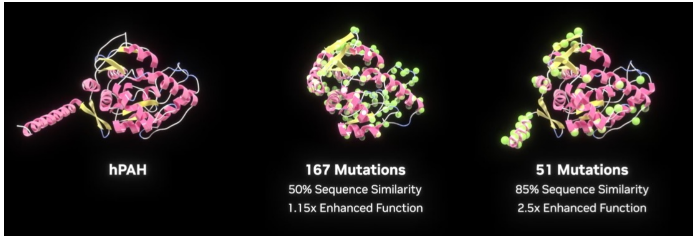
*Corresponding author(s). E-mail(s): acosta@nvidia.com; andrew.ferguson@evozyne.com;
†These authors contributed equally to this work.

SH3



hPAH





Maria Korshunova, Micha Livne

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

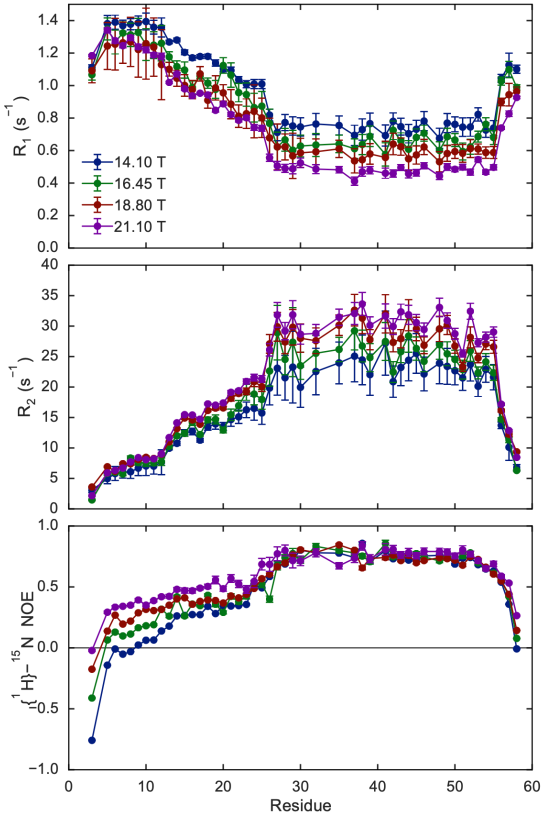
Outline

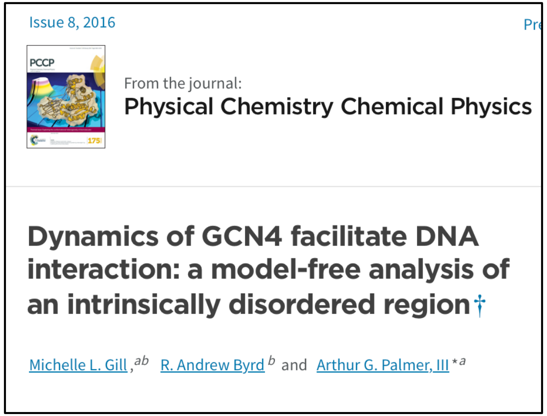
- Foundation model development for science -- small molecules, proteins, and genomics

- What I learned in Andy’s group; and advice for NMR spectroscopists and scientists in the age of AI

BYRD
FEST

There’s No Such Thing as Too Many Fields





Issue 8, 2016

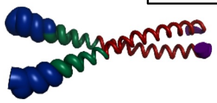
Pre

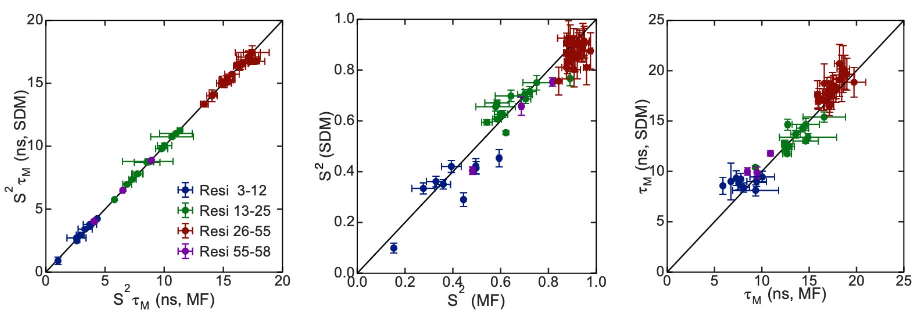
From the journal:

Physical Chemistry Chemical Physics

Dynamics of GCN4 facilitate DNA interaction: a model-free analysis of an intrinsically disordered region†

Michelle L. Gill,[ab] R. Andrew Byrd[b] and Arthur G. Palmer, III*[a]





If You Can’t Collect Enough Data, Simulate It

Minimize $\|f\|_{l1}$ **subject to** $Rx = b$

L1

$$\|f\|_{l1} = \sum |f_k|$$

$$|f_k| = \sqrt{f_{k,r}^2 + f_{k,i}^2}$$

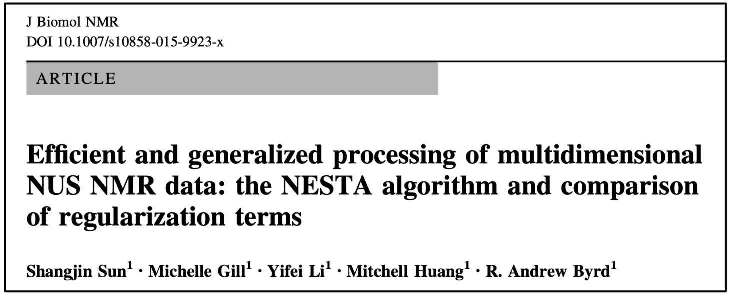
IRL1

$$\|f\|_{irl1} = \sum \omega_k |f_k|$$

$$\omega_k^{i+1} = 1/(|f_k|^i + \varepsilon)$$

Gaussian-SL0

$$\|f\|_{sl0} = \sum (1 - e^{-0.5|f_k|^2/\sigma^2})$$

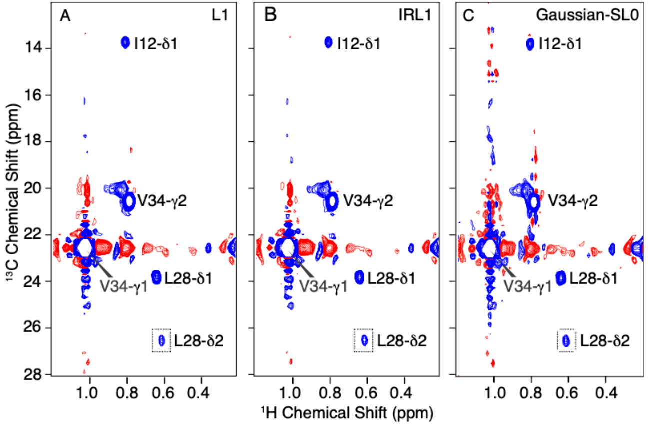


J Biomol NMR
DOI 10.1007/s10858-015-9923-x

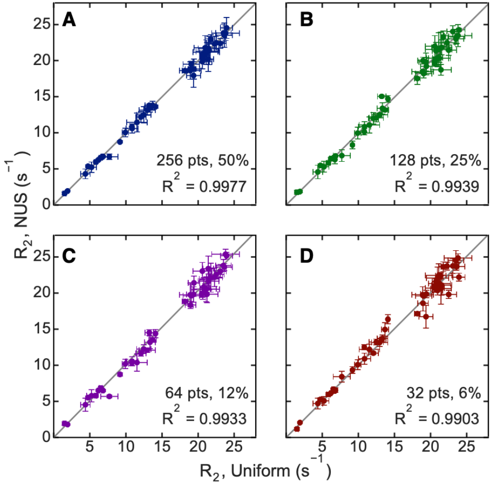
ARTICLE

Efficient and generalized processing of multidimensional NUS NMR data: the NESTA algorithm and comparison of regularization terms

Shangjin Sun[1] · Michelle Gill[1] · Yifei Li[1] · Mitchell Huang[1] · R. Andrew Byrd[1]



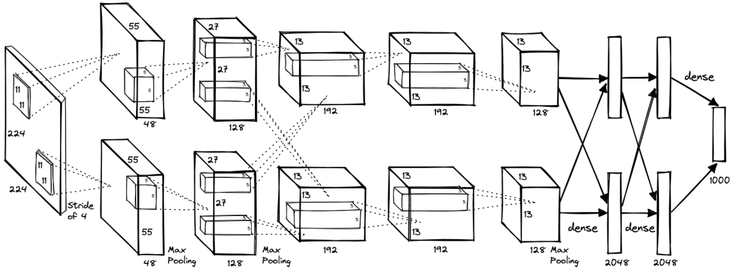
4D HMQC-NOESY-HMCQ of gp78 CUE Domain

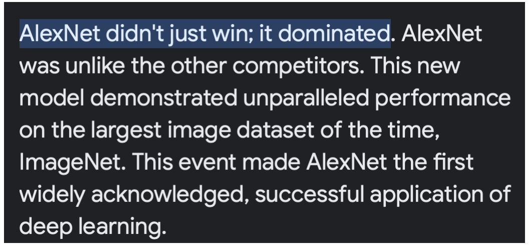


NUS in GCN4 R_2 Measurements
(unpublished)

Don’t Miss the Forest Through the (NMR) Peaks

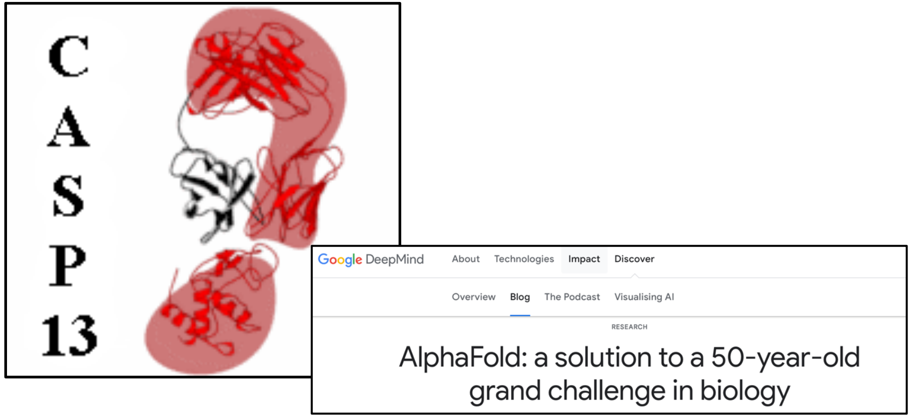
AlexNet Won ImageNet Challenge in 2012



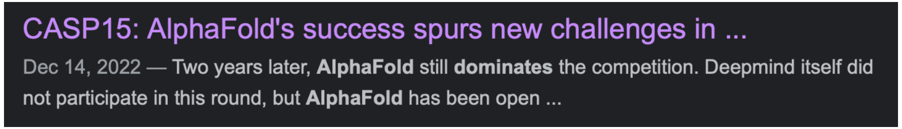


AlexNet didn't just win; it dominated. AlexNet was unlike the other competitors. This new model demonstrated unparalleled performance on the largest image dataset of the time, ImageNet. This event made AlexNet the first widely acknowledged, successful application of deep learning.

AlphaFold Won CASP13 in 2018

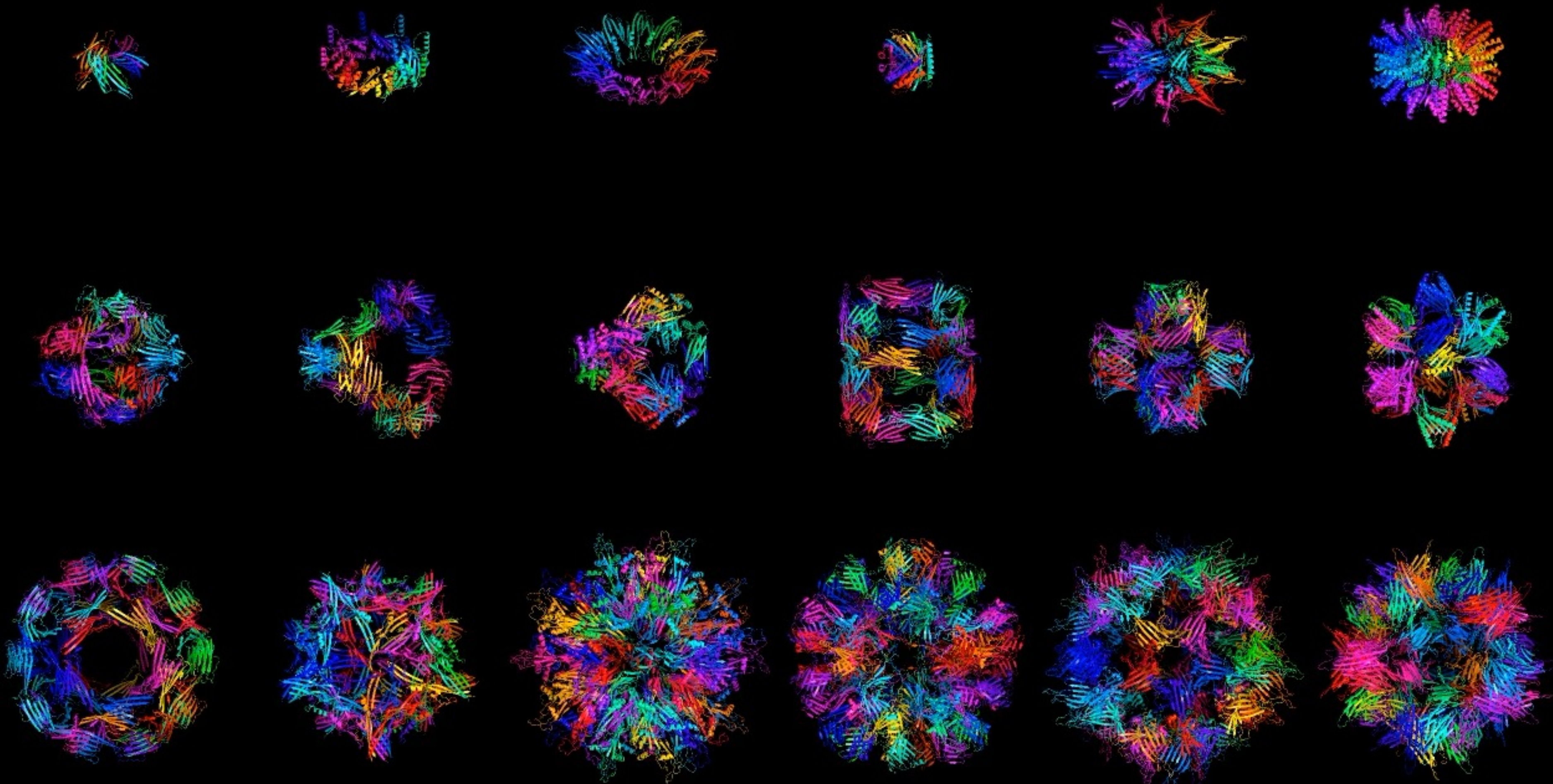


AlphaFold: a solution to a 50-year-old grand challenge in biology



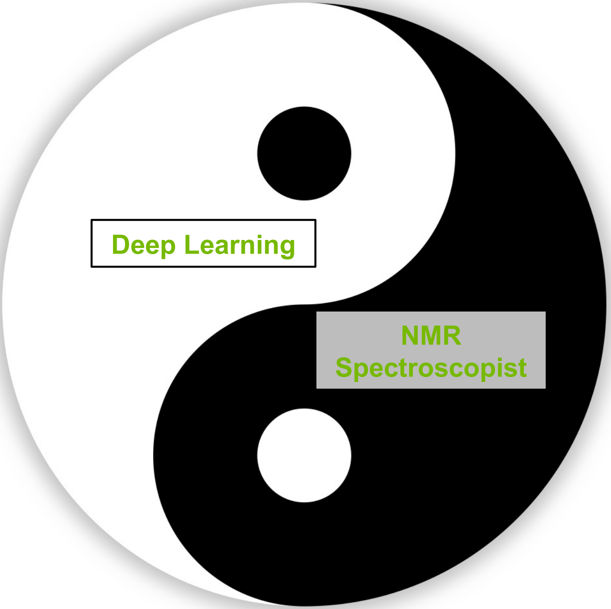
CASP15: AlphaFold's success spurs new challenges in ...

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



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

NMR and Deep Learning are Complementary



AlphaFold is an (Awesome) Tool, Not a Panacea: Open Challenges That NMR Can Address

- Incorporation of dynamics and intrinsically disordered regions in structure prediction
- Study of multimeric proteins with ligands and/or co-factors
- Achievement of structure resolution suitable for drug discovery
- Improved prediction of protein – protein interactions
- Influence of post-translational modifications on structure
- …

Thank You, Andy!

BYRD FEST