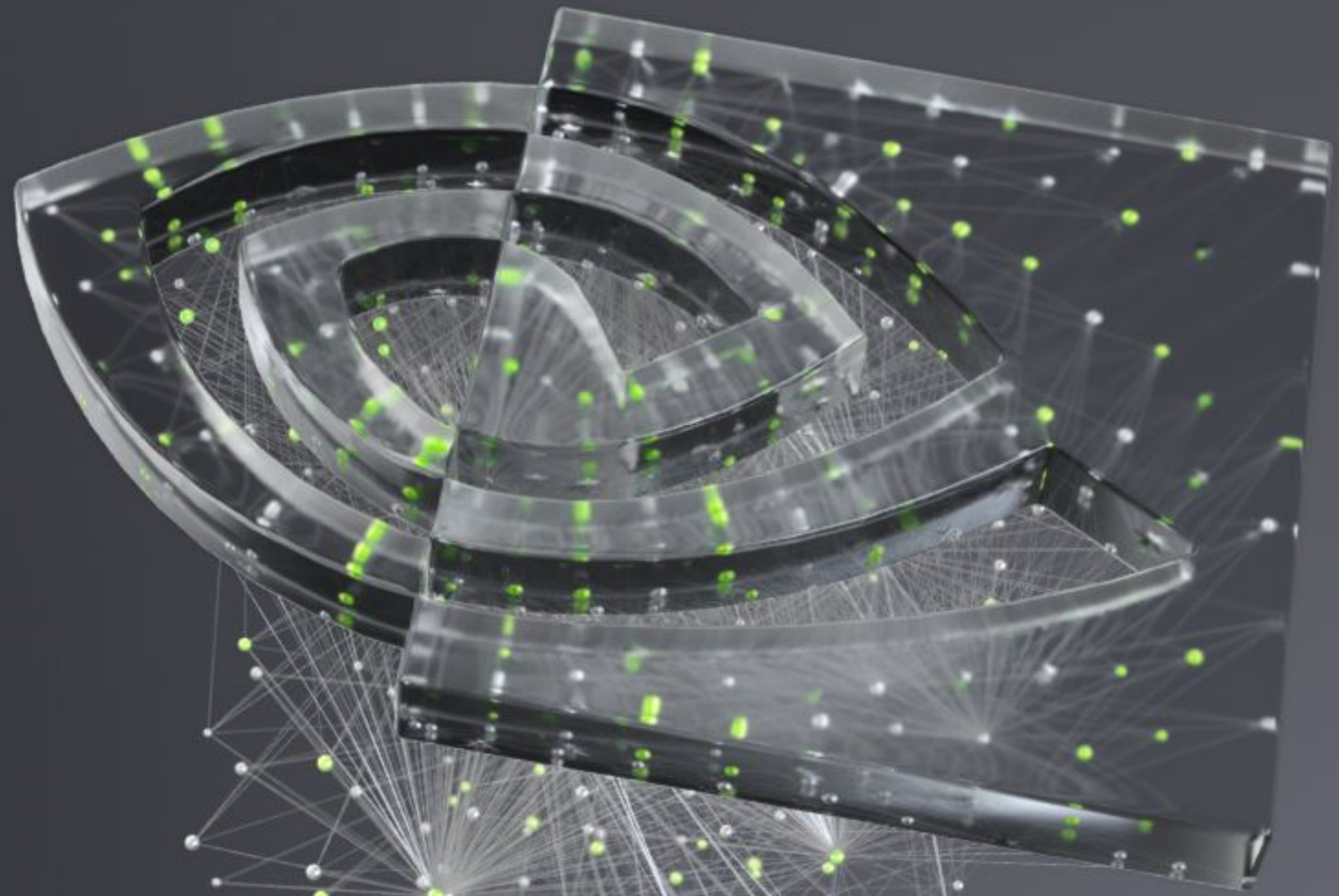


Exploring Molecular Space and Accelerating Drug Discovery on the GPU with Clara Discovery

Michelle L. Gill, Ph.D.
Senior AI Scientist and Tech Lead, Clara Discovery
Gates Foundation Grand Challenges: *Applications of Artificial Intelligence in Machine Learning for Drug Discovery*
November 10, 2021

Overview

Overview of Clara Discovery

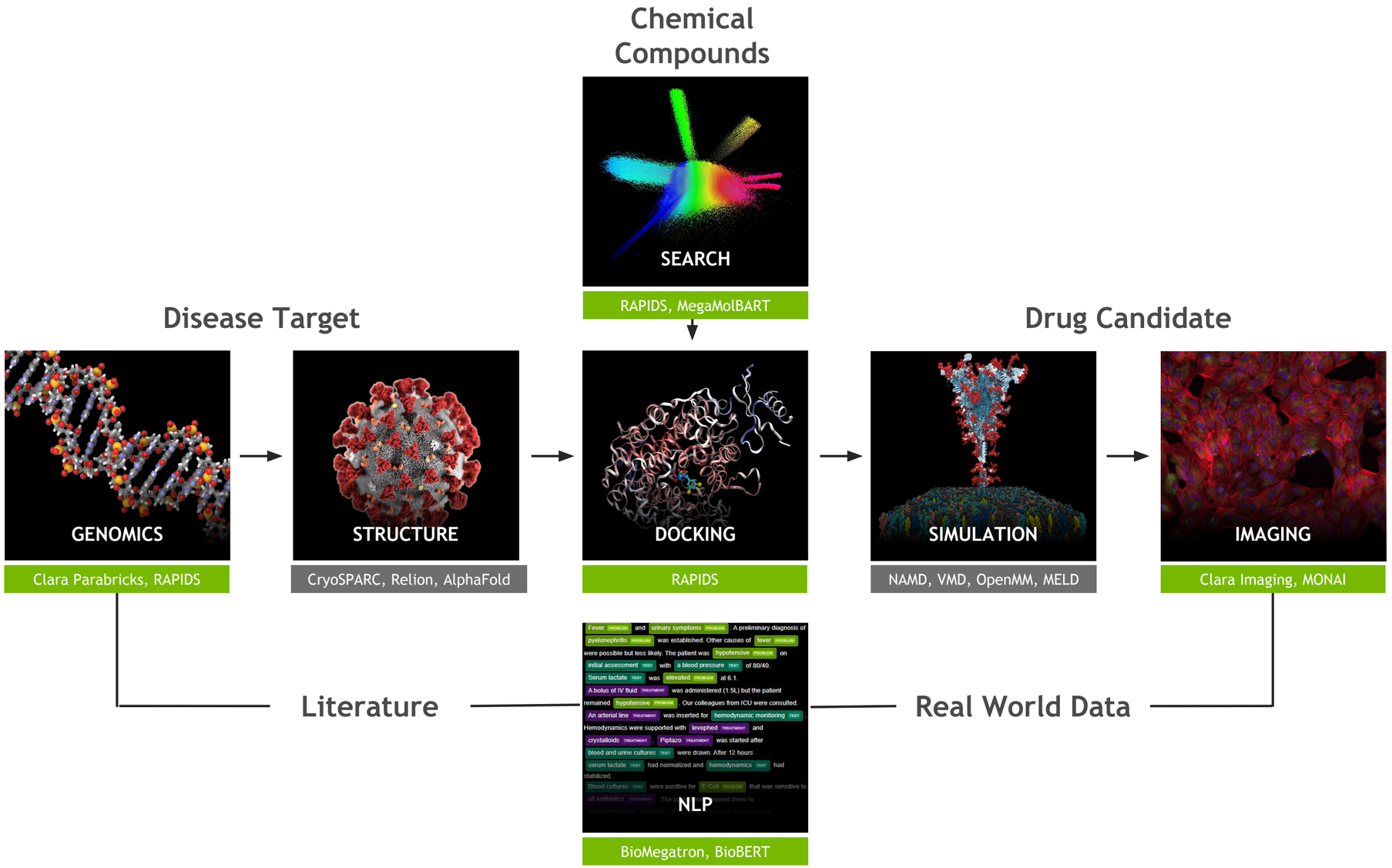
Interactive, accelerated machine learning and visualization

MegaMolBART architecture

Data augmentation and model pretraining

Impact of pretraining on downstream tasks

NVIDIA Clara Discovery



Motivation: Three Years for Design-Make-Test-Analyze Cycle

Hit Compound

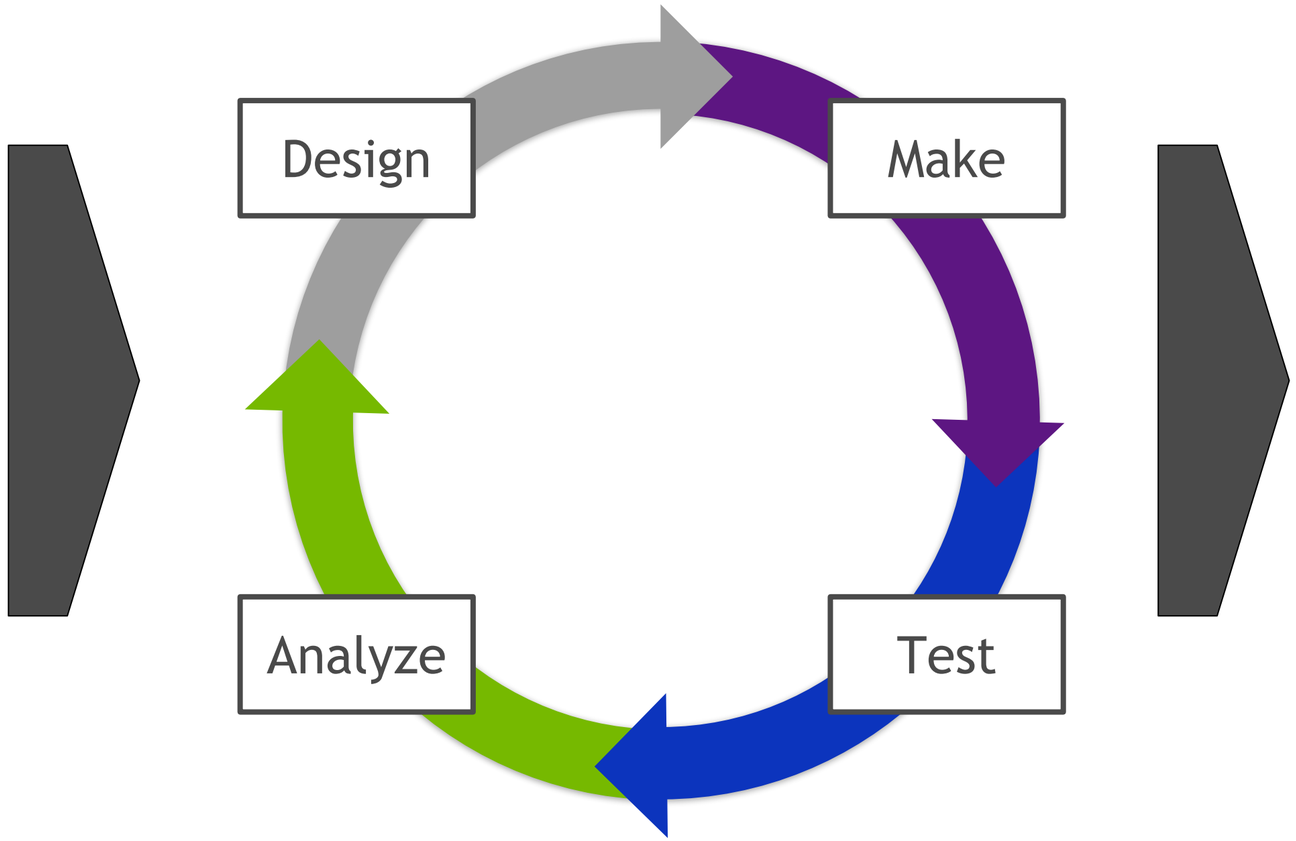
From HTS, DEL, fragment screen, or knowledge

Weakly active

Target unselective

Toxicity risk

Low metabolic stability



Candidate Drug

Highly potent

Effective for *in vivo* models

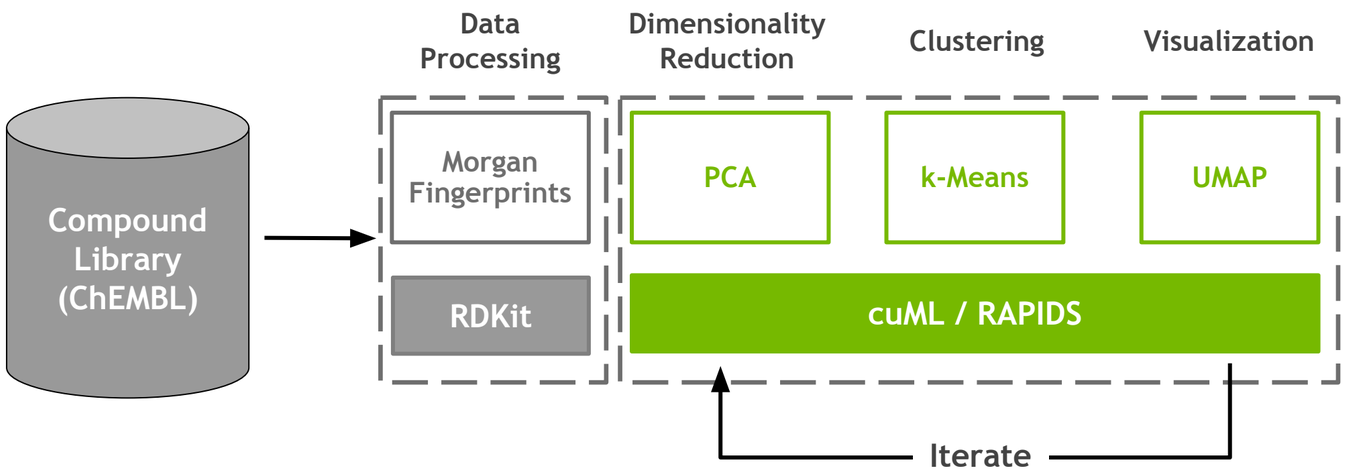
Metabolically stable

No toxicity issues

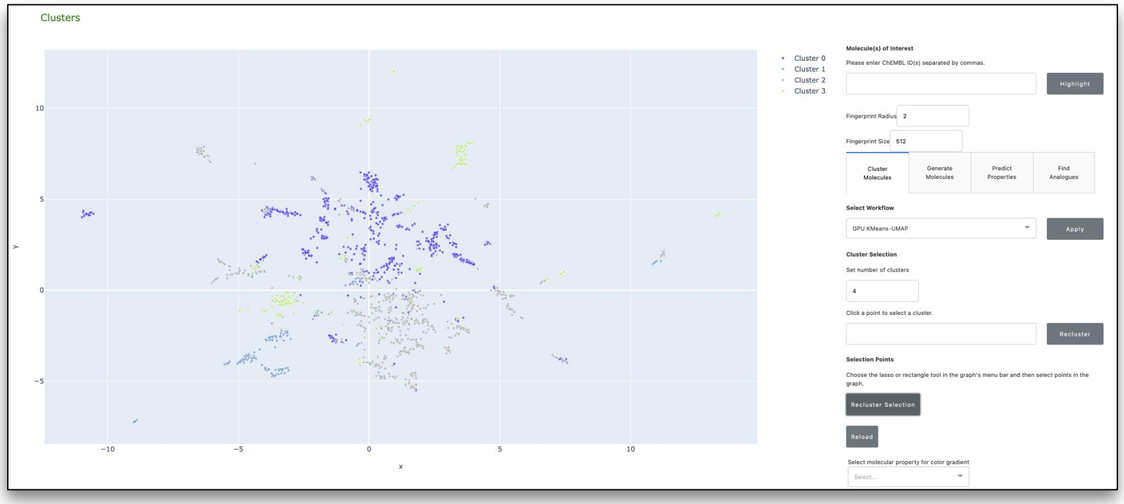
Multiple of DMTA cycles at 4-6 weeks/cycle
Transition between multiple labs

Interactive Clustering and Visualization

Workflow



Interface



plotly | Dash

Real-time, GPU-enabled clustering and visualization of clustering workflows

NVIDIA.

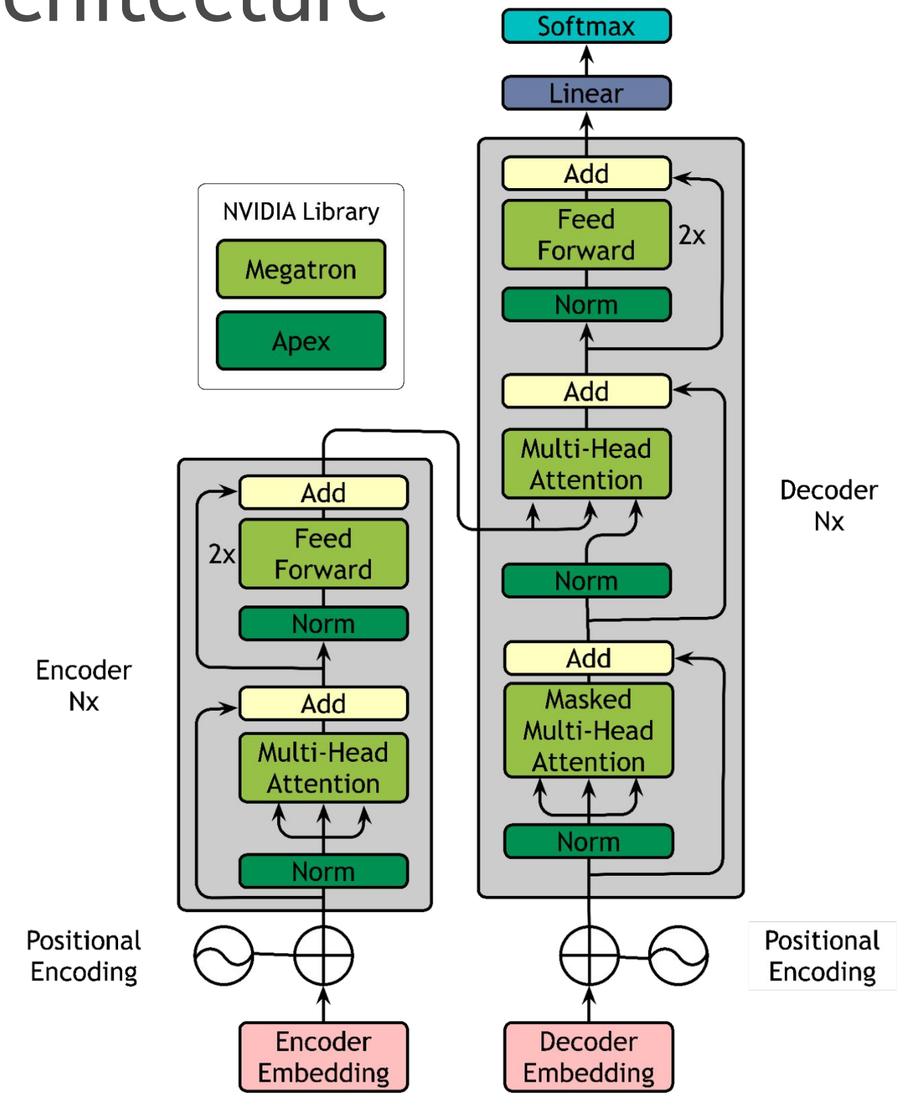
MegaMolBART Architecture

MegaMolBART is a transformer-based model for small molecule drug discovery

MegaMolBART is based on a BART (seq2seq) transformer -- bidirectional encoder and autoregressive decoder

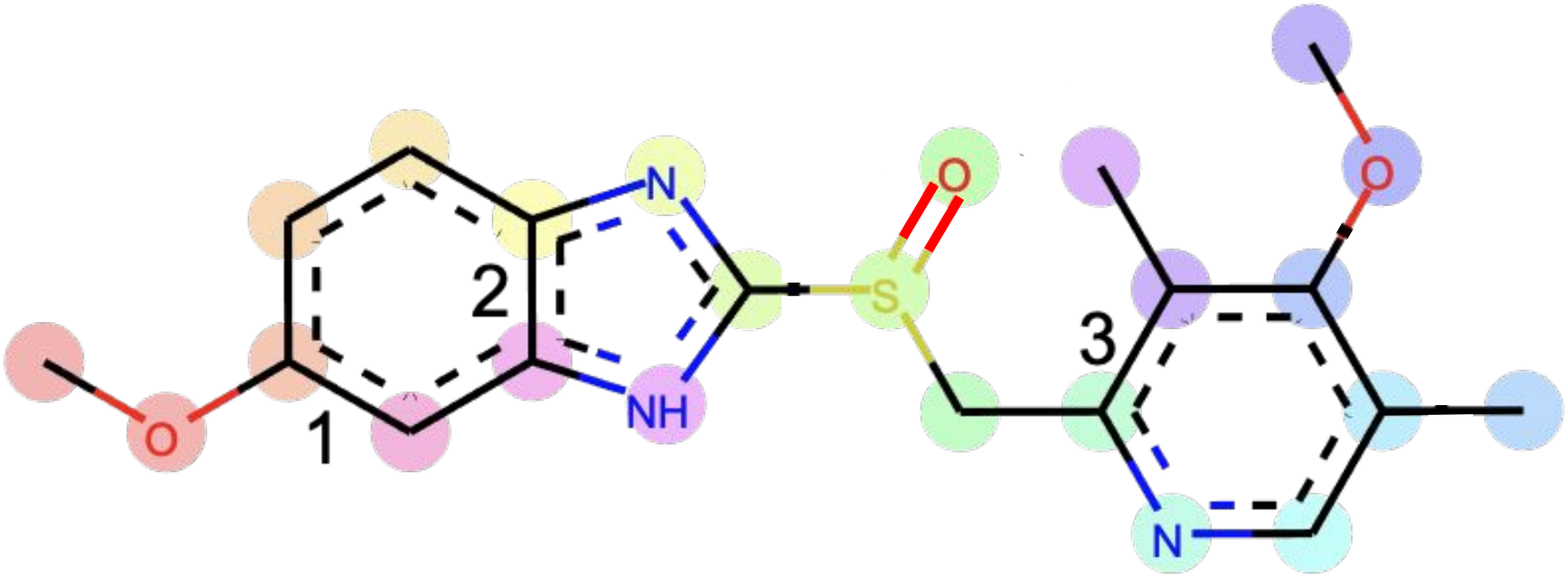
Developed in collaboration with AstraZeneca

Built on NVIDIA's Megatron framework to enable training and inference at scale



Ross, I., Spyridon, D., Jiazhen, H. & Esben, B. Chemformer: A Pre-Trained Transformer for Computational Chemistry. ChemRxiv (2021) doi:10.33774/chemrxiv-2021-v2pnn.

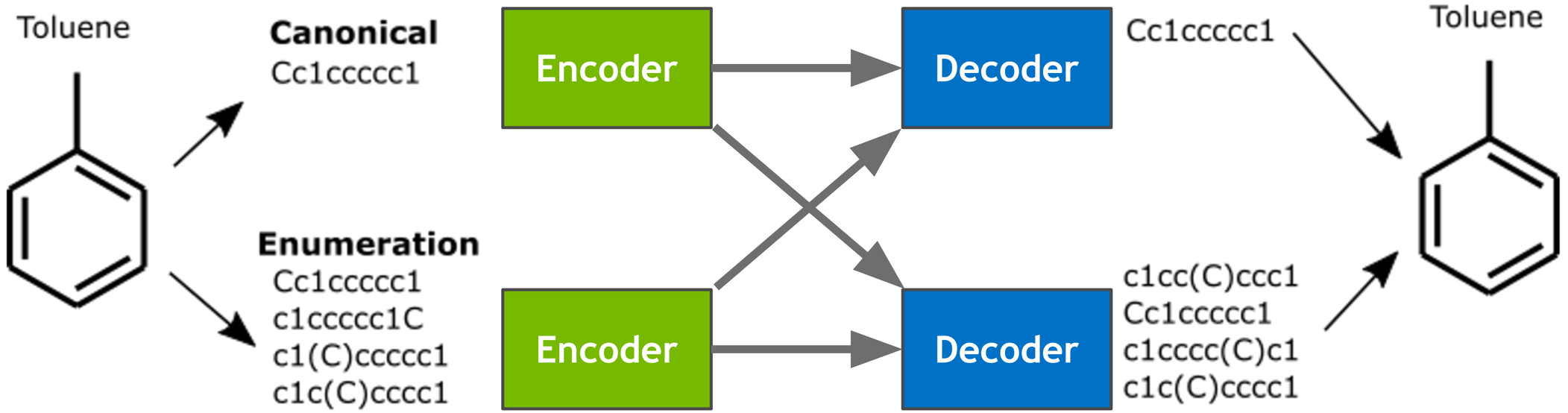
Natural Language Processing for Deep Learning in Cheminformatics



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

SMILES enable small molecules to be represented as text
Leverage deep learning advancements in natural language processing

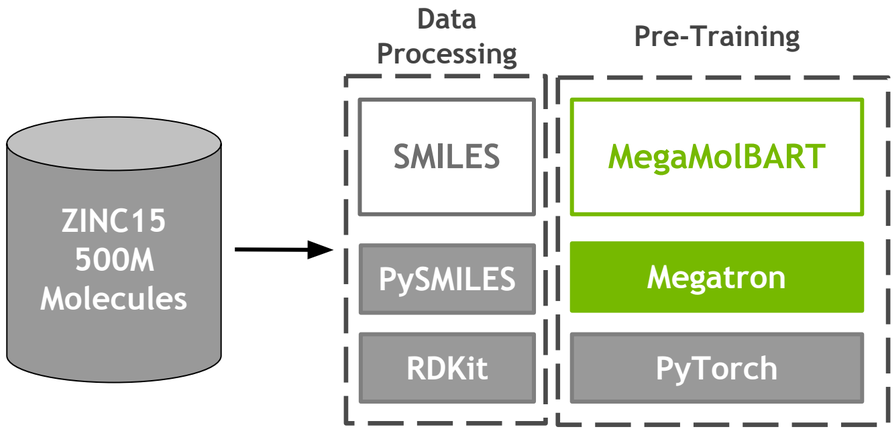
Data Augmentation for SMILES Based Deep Learning



SMILES enumeration and masking improve molecular embeddings

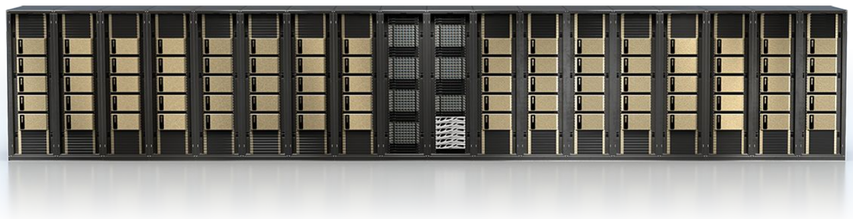
Bjerrum, E. J. & Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. Biomol 8, 131 (2018).

Pre-Training of MegaMolBART at Scale



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

DGX SuperPOD / Cambridge-1

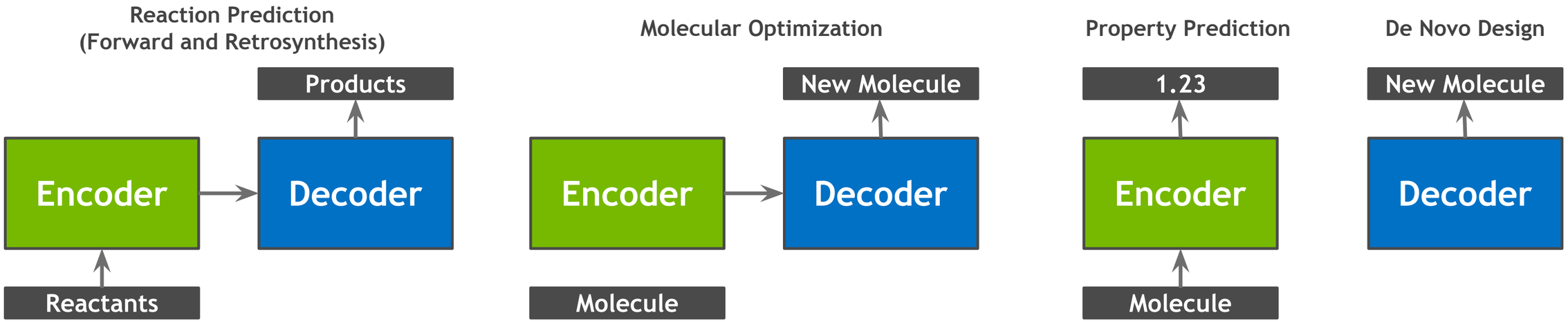


Pre-training performed on ZINC15 -- tranche from reactive, annotated molecules with molecular weight ≤ 500Da, and LogP ≤ 5

Trained on DGX SuperPOD -- 4–8 nodes x 8 A100 GPUs

AstraZeneca concurrently developing on Cambridge-1

MegaMolBART Downstream Predictive Tasks

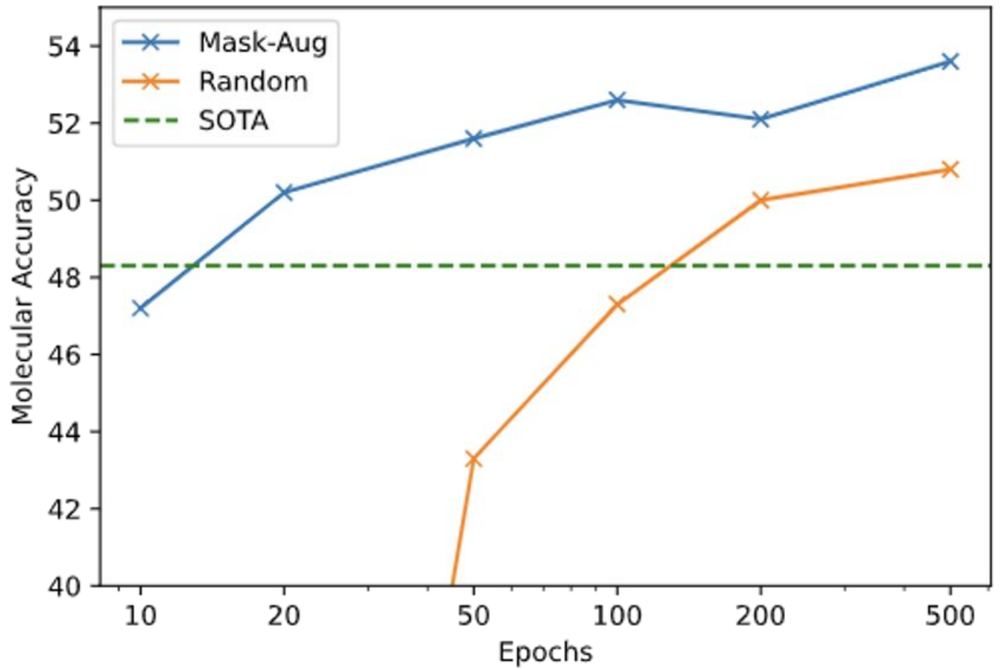


Reaction prediction and molecular optimization utilize encoder and decoder

Molecular property prediction from decoder embeddings

De novo molecular design based on decoder

Pre-Training Increases Performance for Retrosynthesis Prediction



Pre-Training	Forward Prediction (%)	Retrosynthesis (%)	Molecular Optimization (%)
-	91.1	50.8	69.5
Mask	91.2	52.1	**72.1**
Augmentation	91.1	51.8	71.2
Combined	**91.8**	**53.6**	69.7
SOTA	91.1*	48.3*	66.6%.‡

SOTA: Tetko, Igor V., et al. "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis." Nature communications 11.1 (2020): 1-11.
Random: Randomly initialized weights.

* Tetko, Igor V., et al. "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis." Nature communications 11.1 (2020): 1-11.
‡ He, J. et al. Molecular optimization by capturing chemist's intuition using deep neural networks. J Cheminformatics 13, 26 (2021).

Pre-trained model outperforms SOTA in fewer than 20 epochs (<30 min)

Based on top-1 molecular accuracy for retrosynthesis prediction

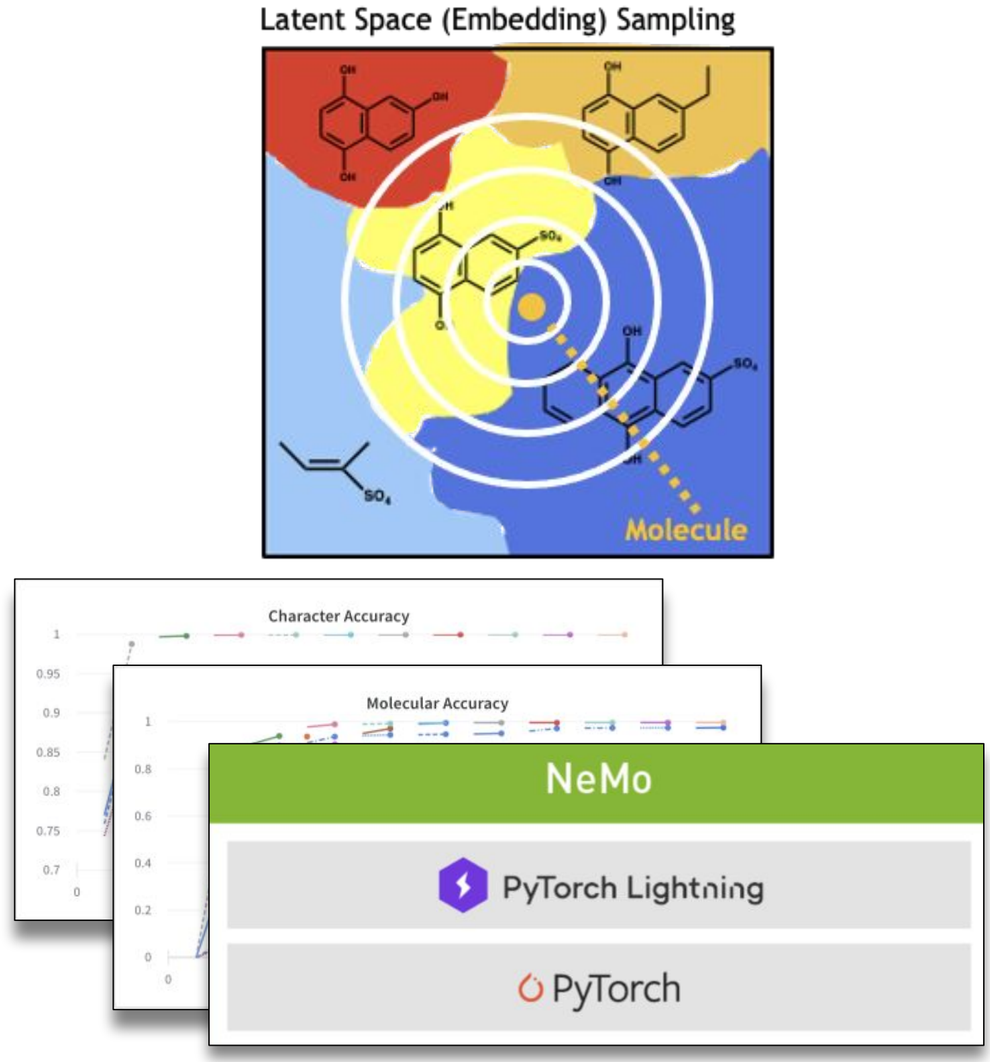
What's Next?

Scaling MegaMolBART -- what are the limits to larger models?

Implement and test downstream tasks in MegaMolBART

Improve *de novo* molecule generation -- development of novel model architectures

Automation of data processing, pre-training and downstream tasks



Conclusions

Clara Discovery is a collection of tools and frameworks that accelerate drug discovery

The interactive explorer provides a framework for visualizing and customizing workflows

MegaMolBART is a seq2seq transformer pre-trained at scale using augmented SMILES

Pre-training produces SOTA performance on downstream tasks

All tools are open source and freely available

Where to Get It:
Clara Discovery Release V0.1.4

Resource	
MegaMolBART Weights	https://ngc.nvidia.com/catalog/resources/nvidia:clara:cheminformatics
Featurizer Service	https://ngc.nvidia.com/models/nvidia:clara:megamolbart
Interactive Explorer	https://ngc.nvidia.com/containers/nvidia:clara:megamolbart
Tutorials	https://ngc.nvidia.com/containers/nvidia:clara:cheminformatics_demo
	https://github.com/NVIDIA/cheminformatics

Acknowledgements

Abe Stern, PhD

Rajesh Ilango

Venkatesh Mysore, PhD

Johnny Israeli, PhD

Rob Brisk, MRCP, PhD

Hassan Sirelkhatim

Myriem Demouth

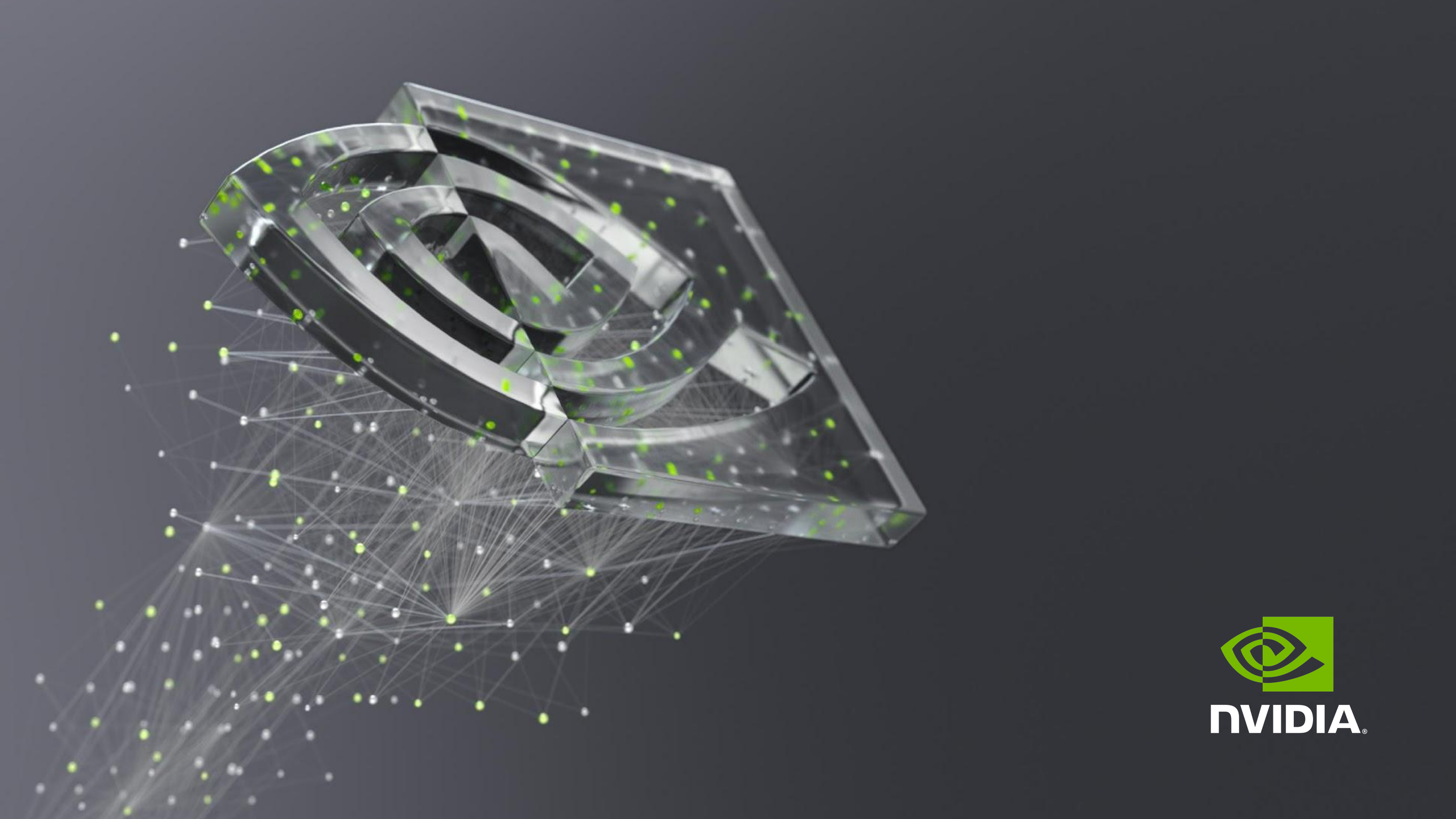
Esben Jannik Bjerrum, PhD

Ross Irwin

Jiazhen He, PhD

Spyridon Dimitriadis

Ola Engkvist, PhD


NVIDIA