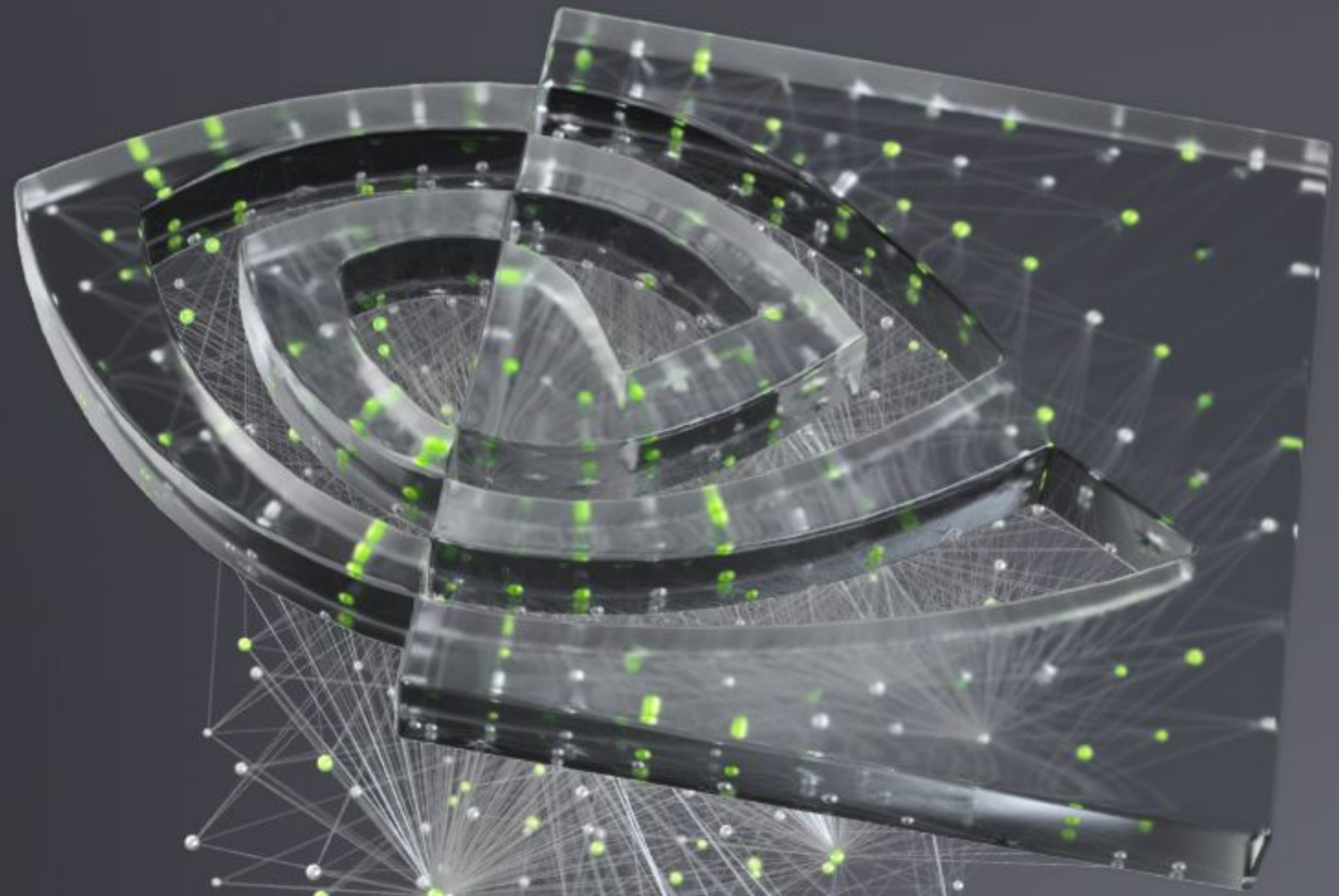


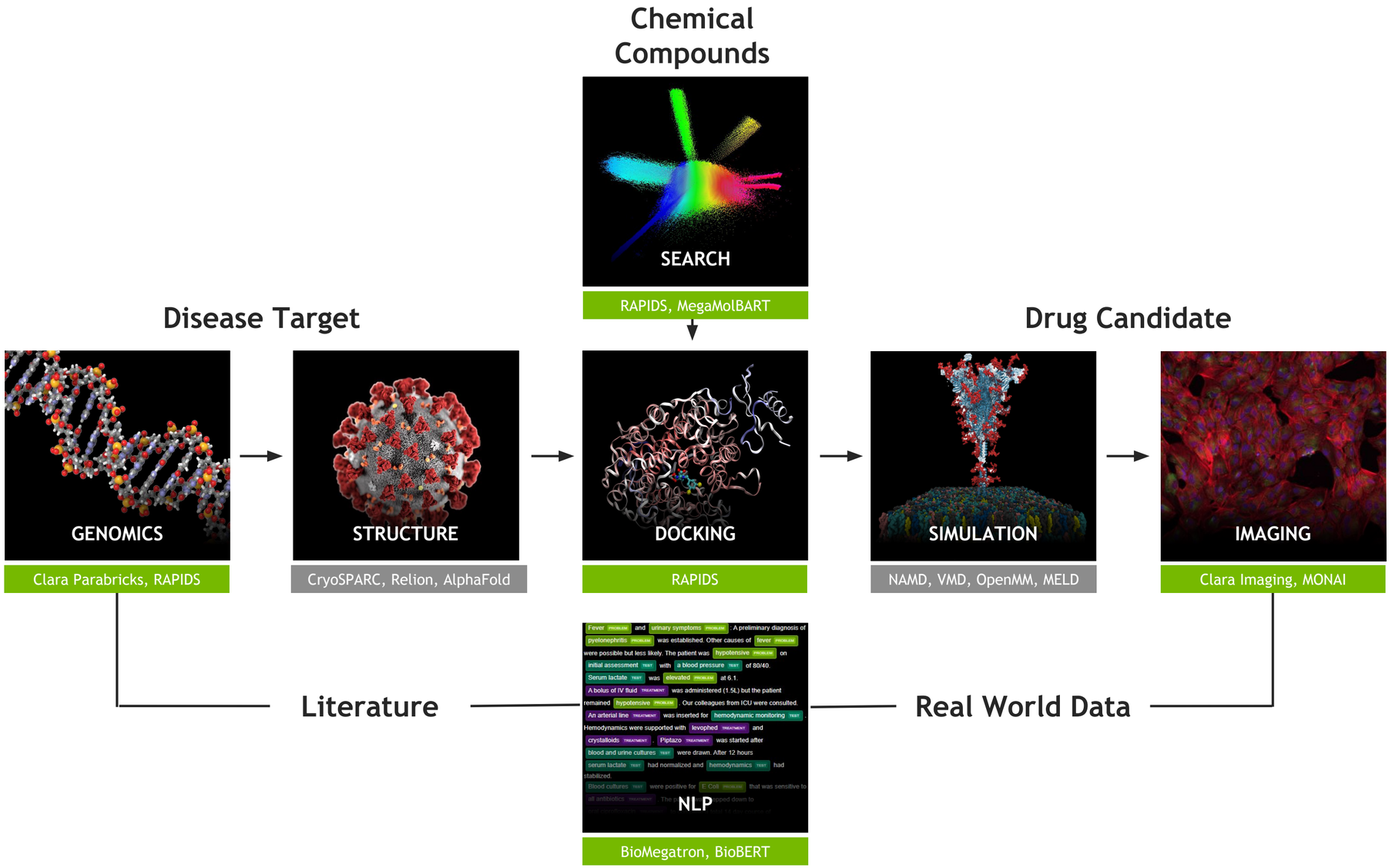
Exploring Molecular Space and Accelerating Drug Discovery with Clara Discovery and MegaMolBART

Michelle L. Gill, Ph.D.

4th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry Symposium

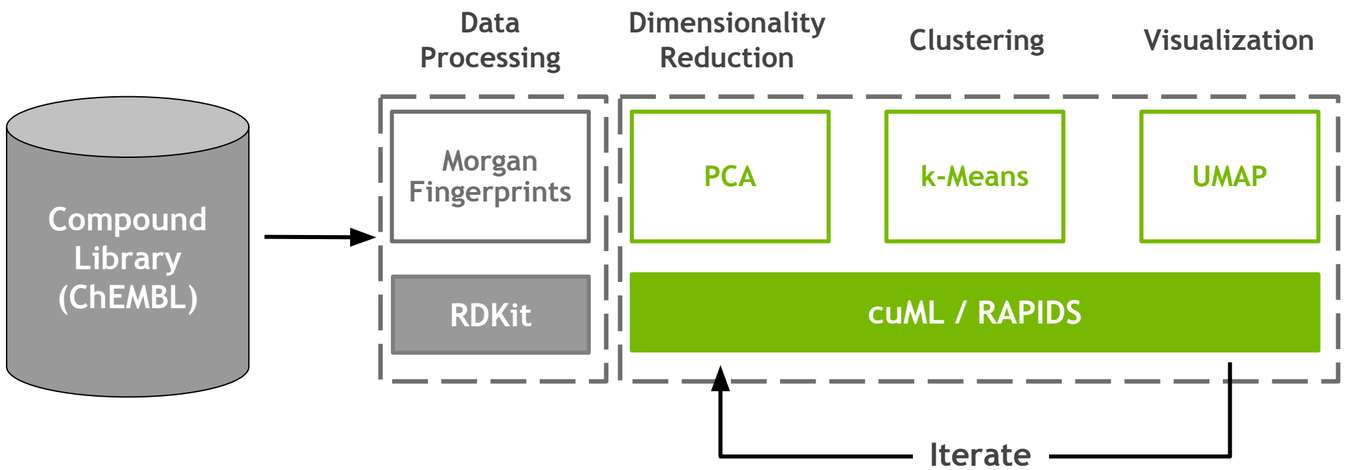
September 27, 2021

NVIDIA Clara Discovery

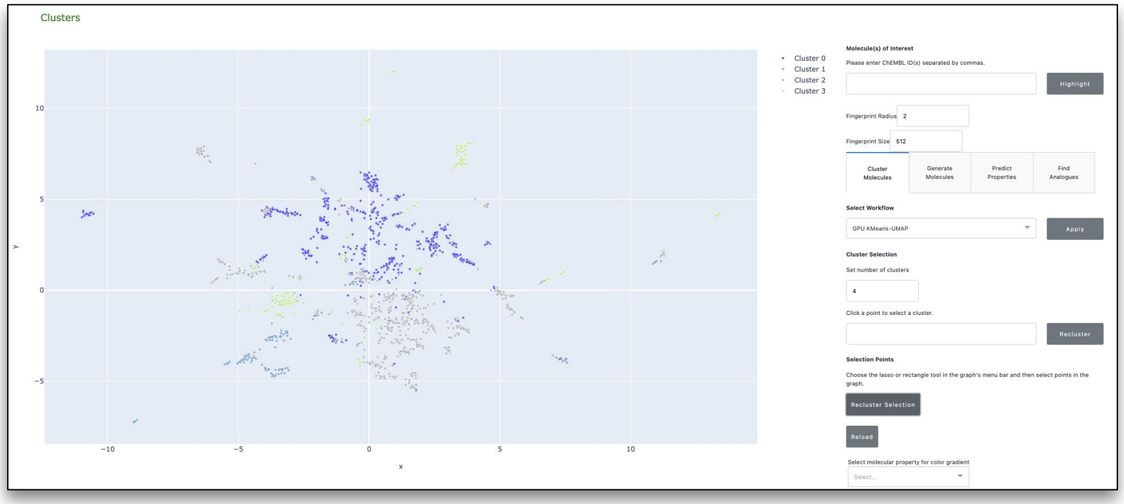


Interactive Clustering and Visualization

Workflow



Interface



plotly | Dash

NVIDIA

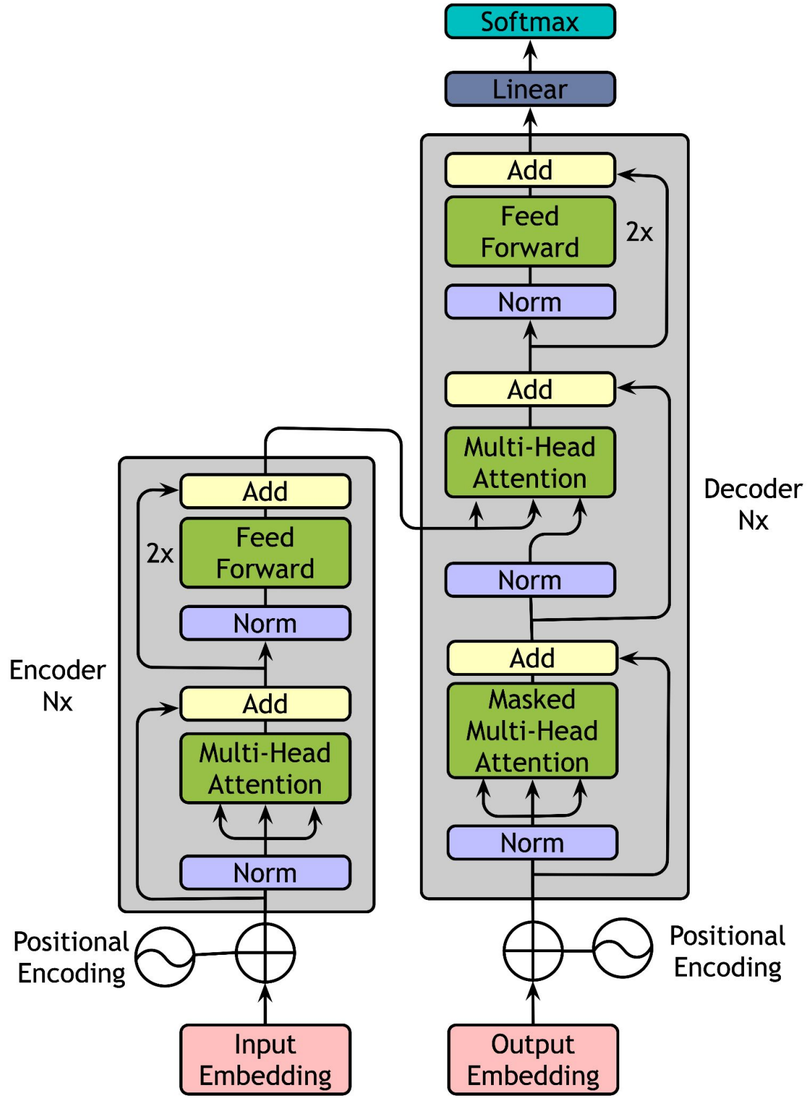
MegaMolBART Architecture

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

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

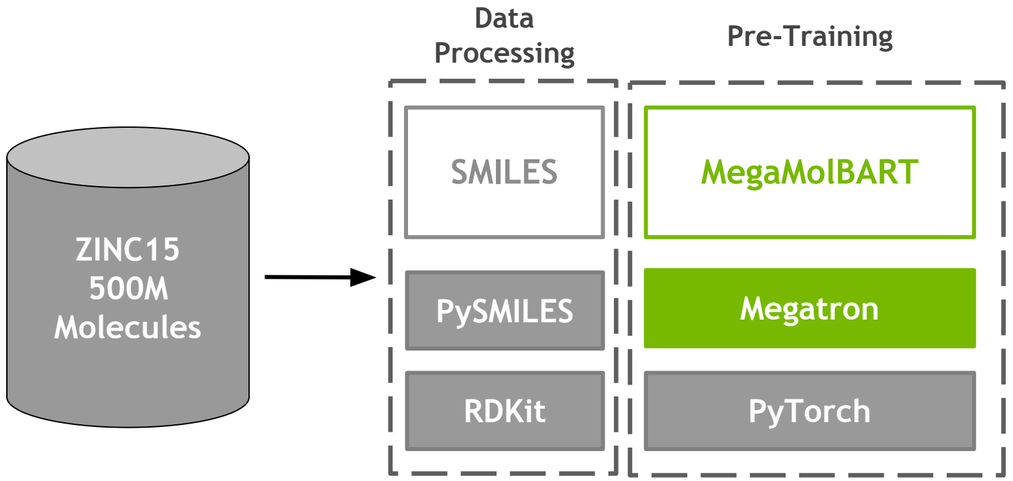
Developed in collaboration with AstraZeneca

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.

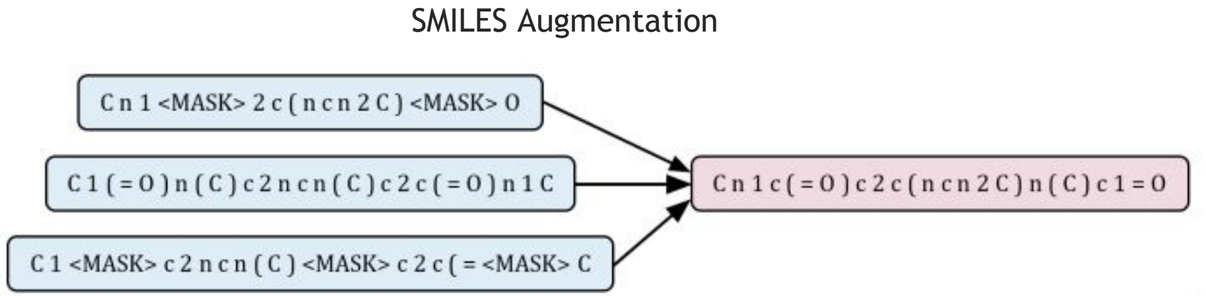
Pre-Training of MegaMolBART



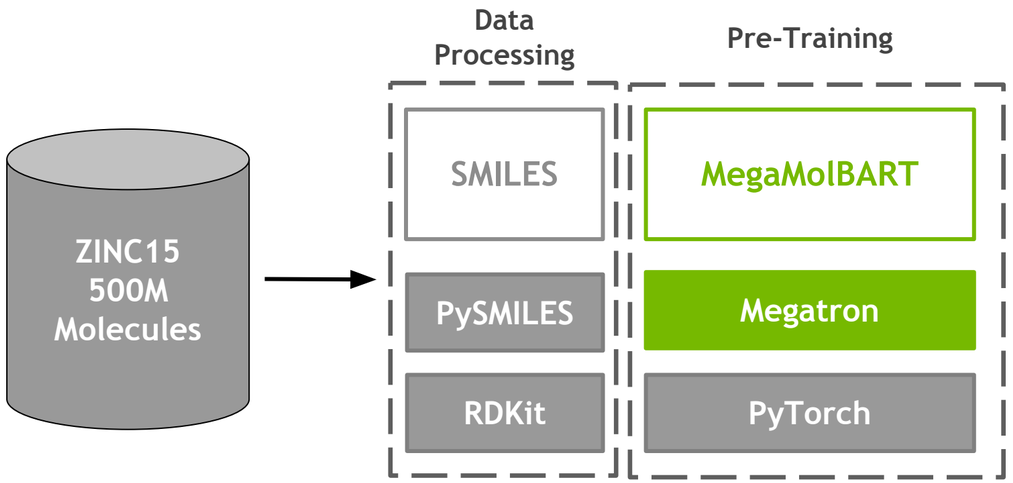
Pre-training performed on ZINC15 -- tranche from reactive, annotated molecules with molecular weight ≤ 500Da, and LogP ≤ 5

Train, validation, and test splits were 99% / 0.5% / 0.5%

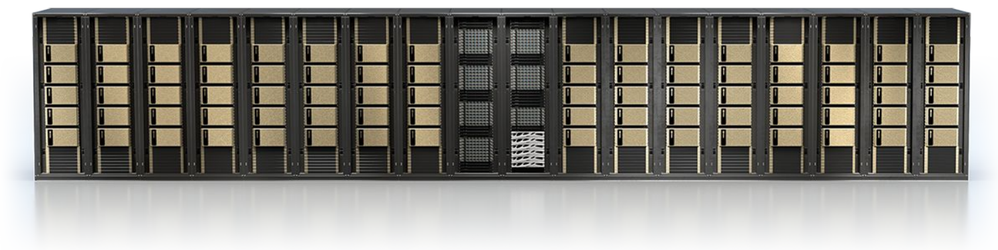
SMILES molecules were masked and enumerated (randomized) during training



Pre-Training of MegaMolBART



DGX SuperPOD / Cambridge-1



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

Train, validation, and test splits were 99% / 0.5% / 0.5%

SMILES molecules were masked and enumerated (randomized) during training

Trained on DGX SuperPOD -- upto 8 nodes x 8 A100 GPUs

AstraZeneca concurrently developing on Cambridge-1

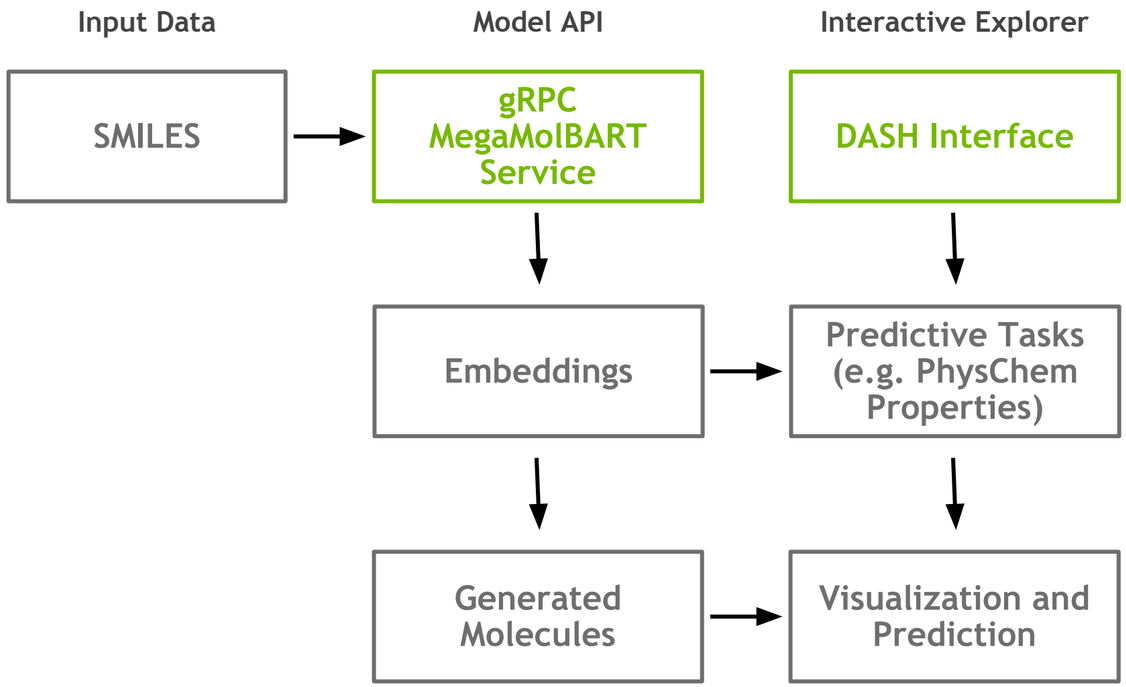
MegaMolBART Model Service

Model inference tasks available via gRPC API:

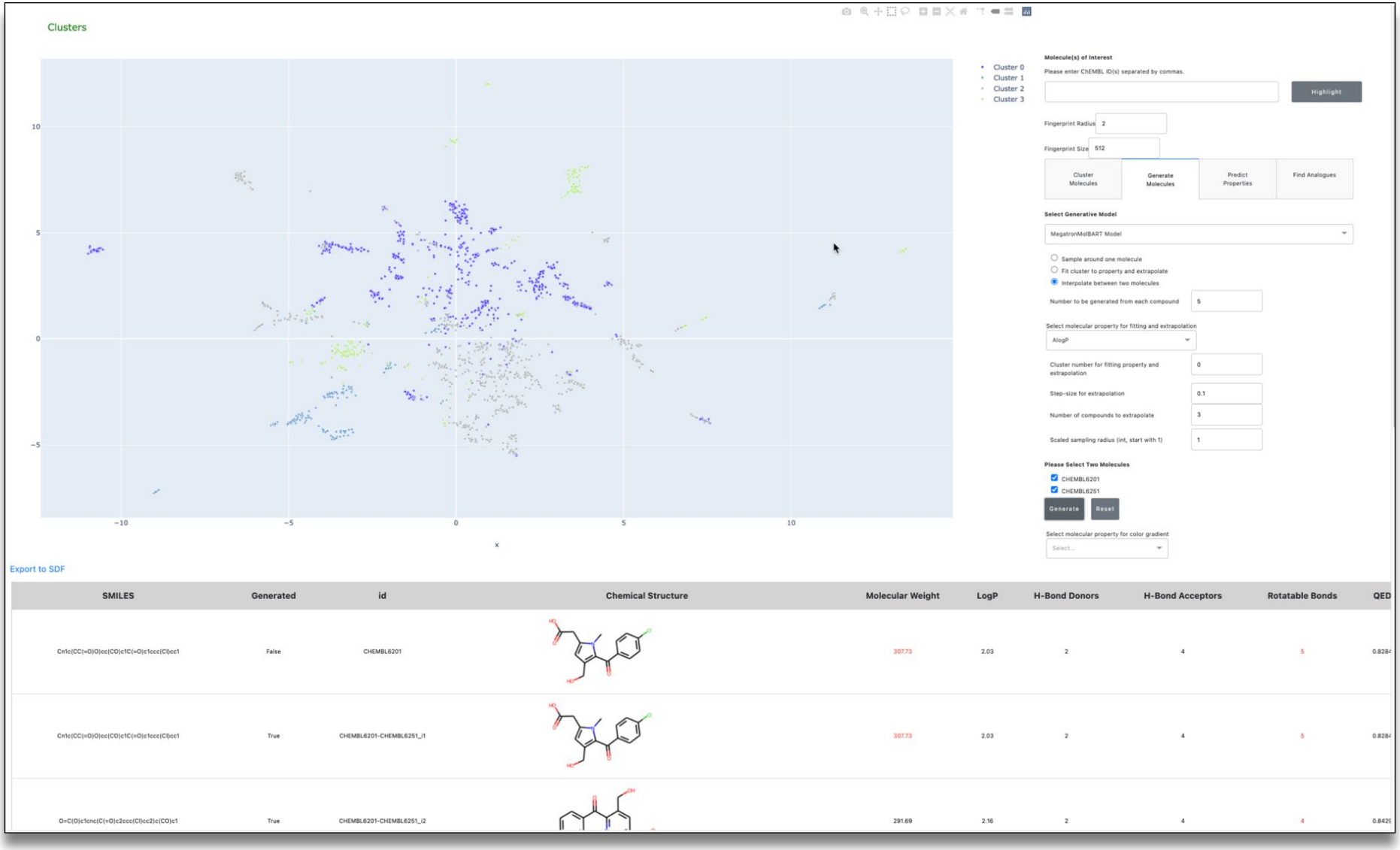
- Learned embeddings from SMILES molecule(s)
- Molecule generation from SMILES molecule(s)

Workflows and outputs are integrated into the interactive explorer

Designed for user customization -- predictive tasks incorporated into workflow



Demo of Interactive Explorer and Molecule Generation

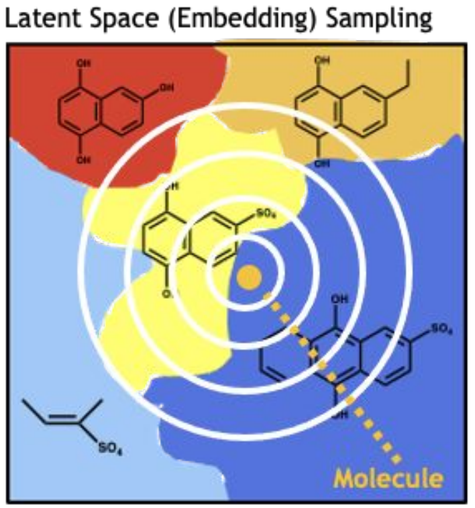


What's Next?

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

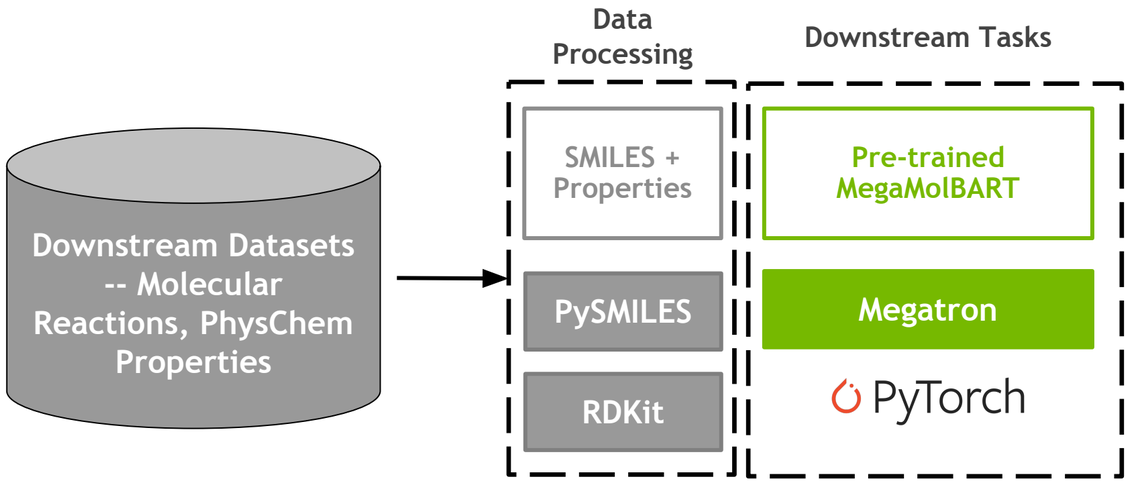
Attention Heads	Layers	Hidden Size	Feed Forward	Parameters
8	4	256	1024	10M
8	6	512	2048	45M
16	8	1024	4096	230M
...	...	...	...	...

Improved molecule generation -- development of novel model architectures

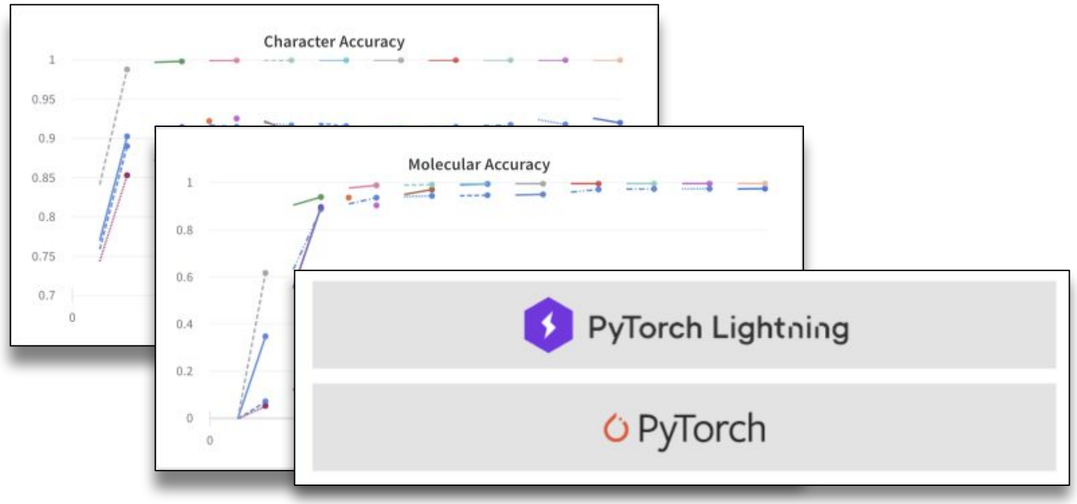


What's Next?

Predictive tasks based on model embeddings -- physchem properties, reaction prediction, retrosynthetic synthesis



Improved user experience -- automation of data processing, pre-training and downstream tasks



Where to Get It:
Clara Discovery Release V0.1.3

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

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 model developed in collaboration with AstraZeneca and trained at scale

All tools are open source and freely available

Acknowledgements

Abe Stern, PhD

Rajesh Ilango

Venkatesh Mysore, PhD

Johnny Israeli, PhD

Rob Brisk, MRCP, PhD

Hassan Sirelkhatim

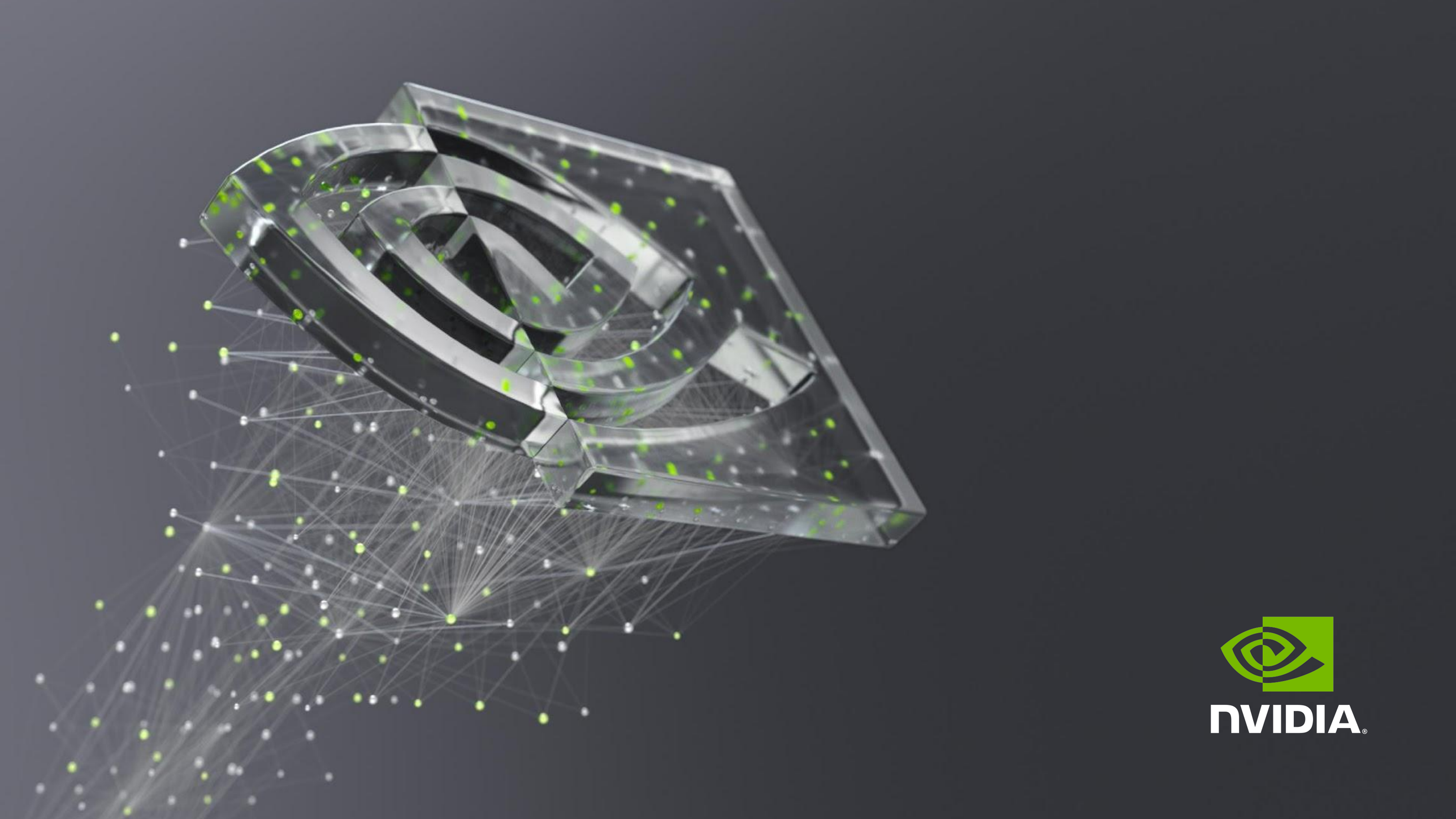
Myriem Demouth

Esben Jannik Bjerrum, PhD

Ross Irwin

Jiazhen He, PhD

Ola Engkvist, PhD


NVIDIA