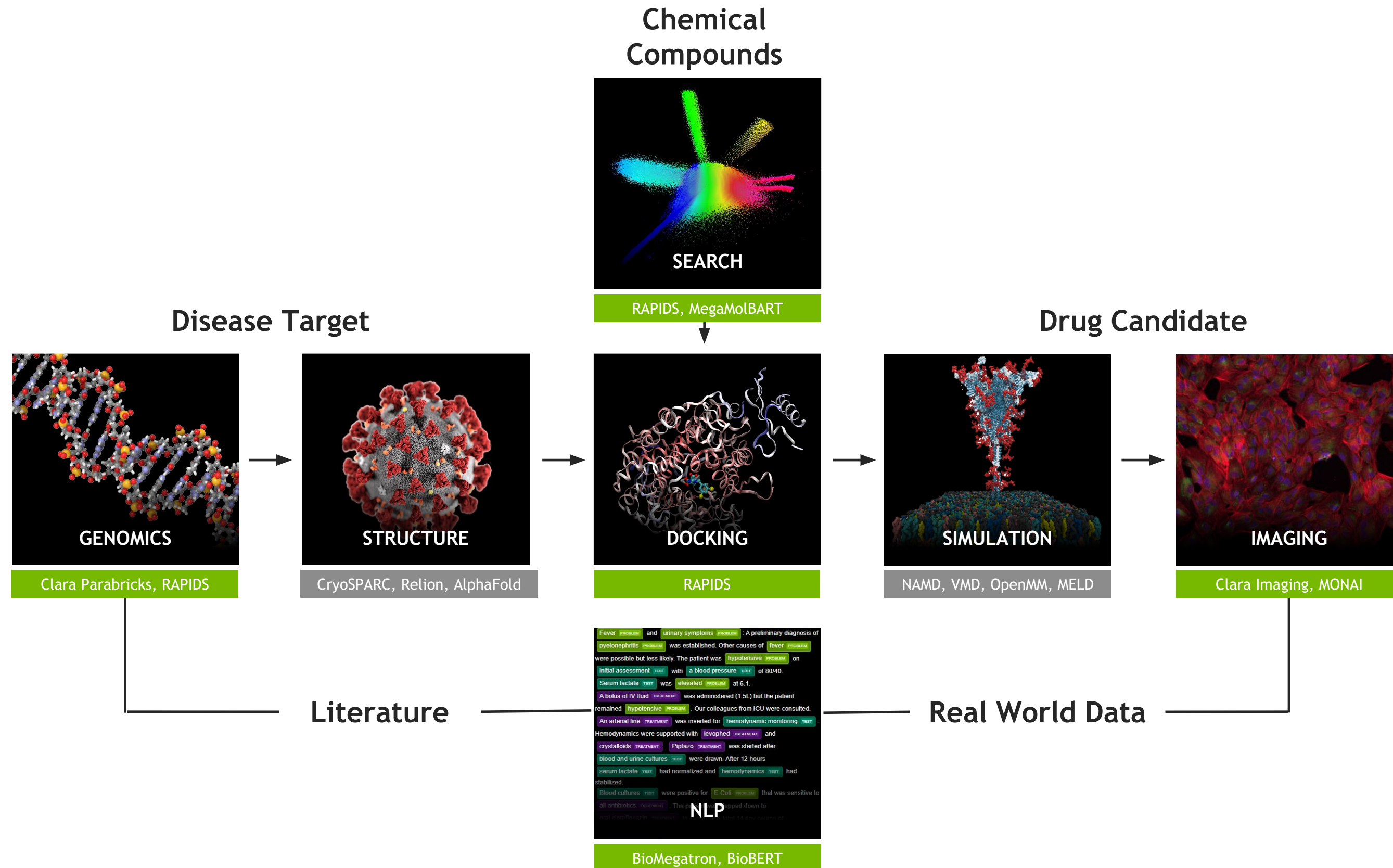


Exploring Molecular Space and Accelerating Drug Discovery with Clara Discovery and MegaMolBART

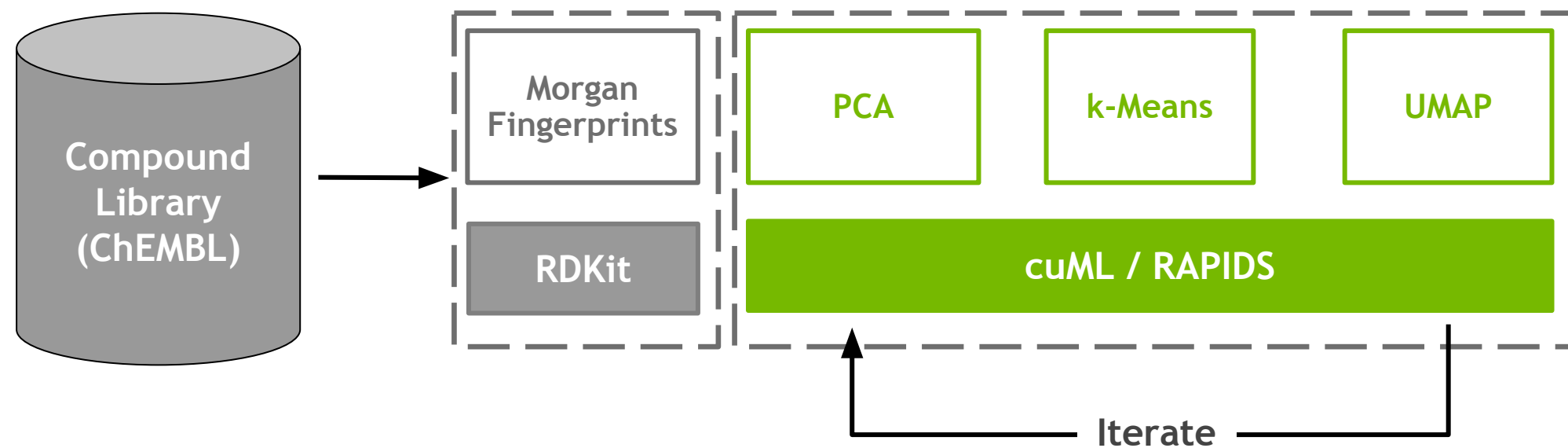
Michelle L. Gill, Ph.D.
4th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry Symposium
September 27, 2021

NVIDIA Clara Discovery

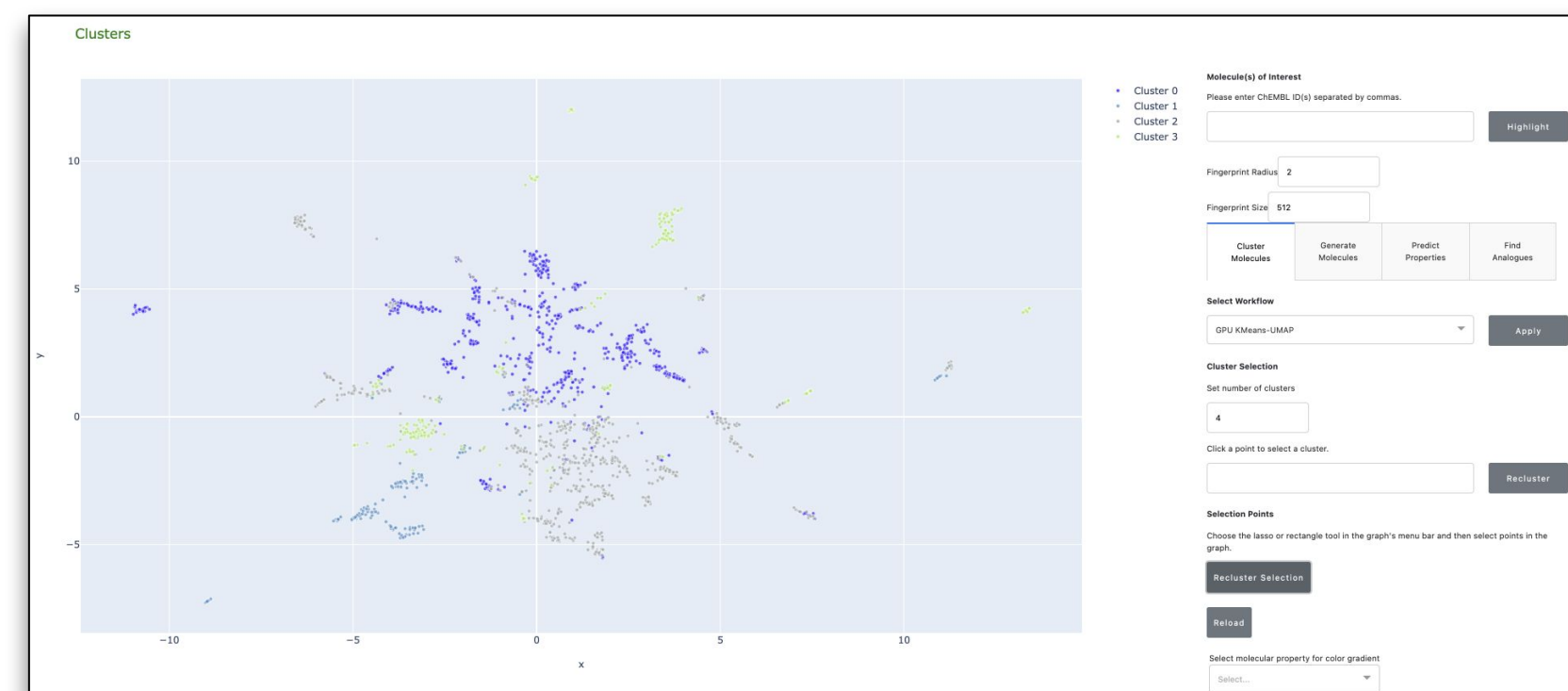


Interactive Clustering and Visualization

Workflow



Interface



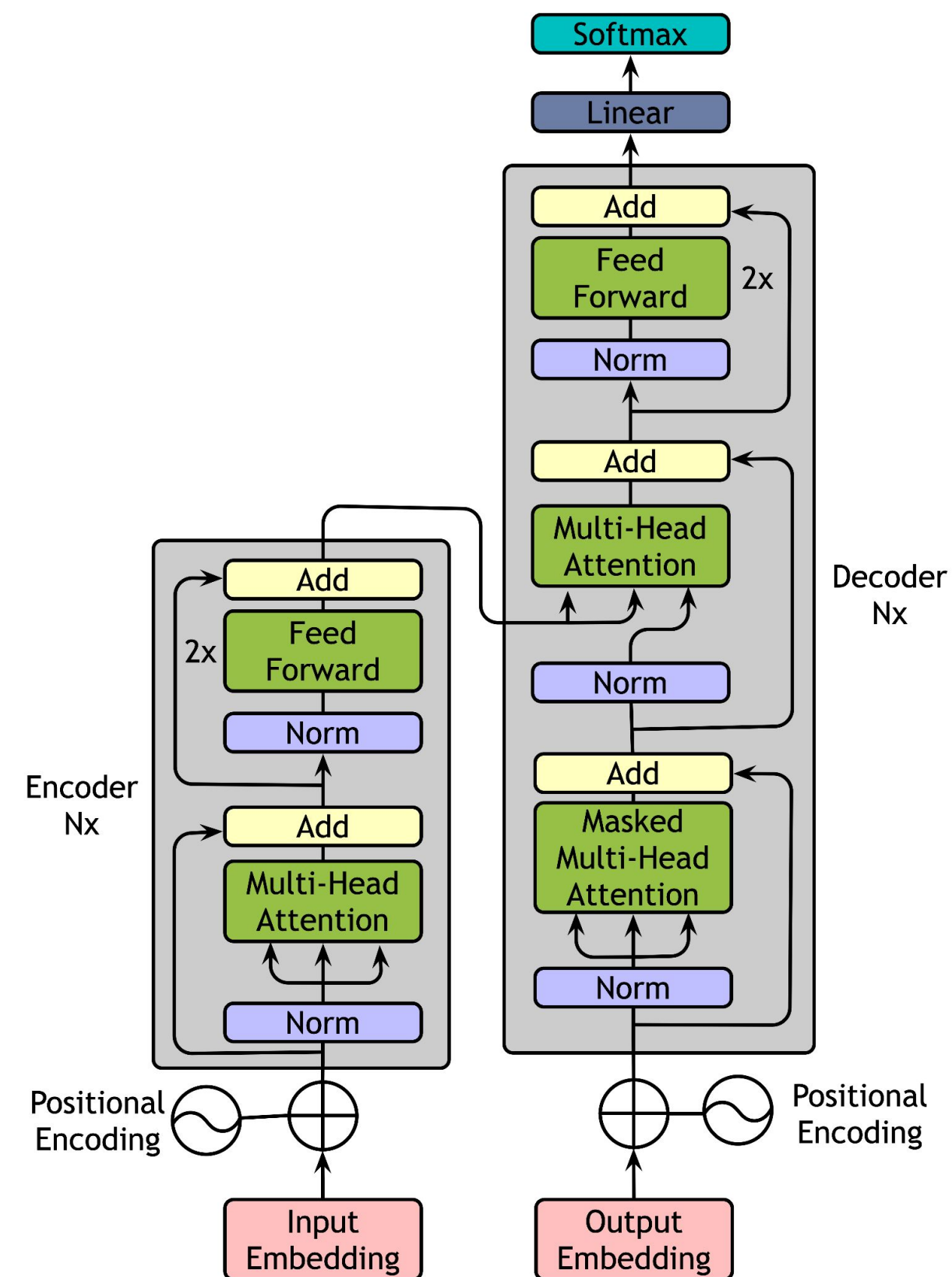
MegaMolBART Architecture

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

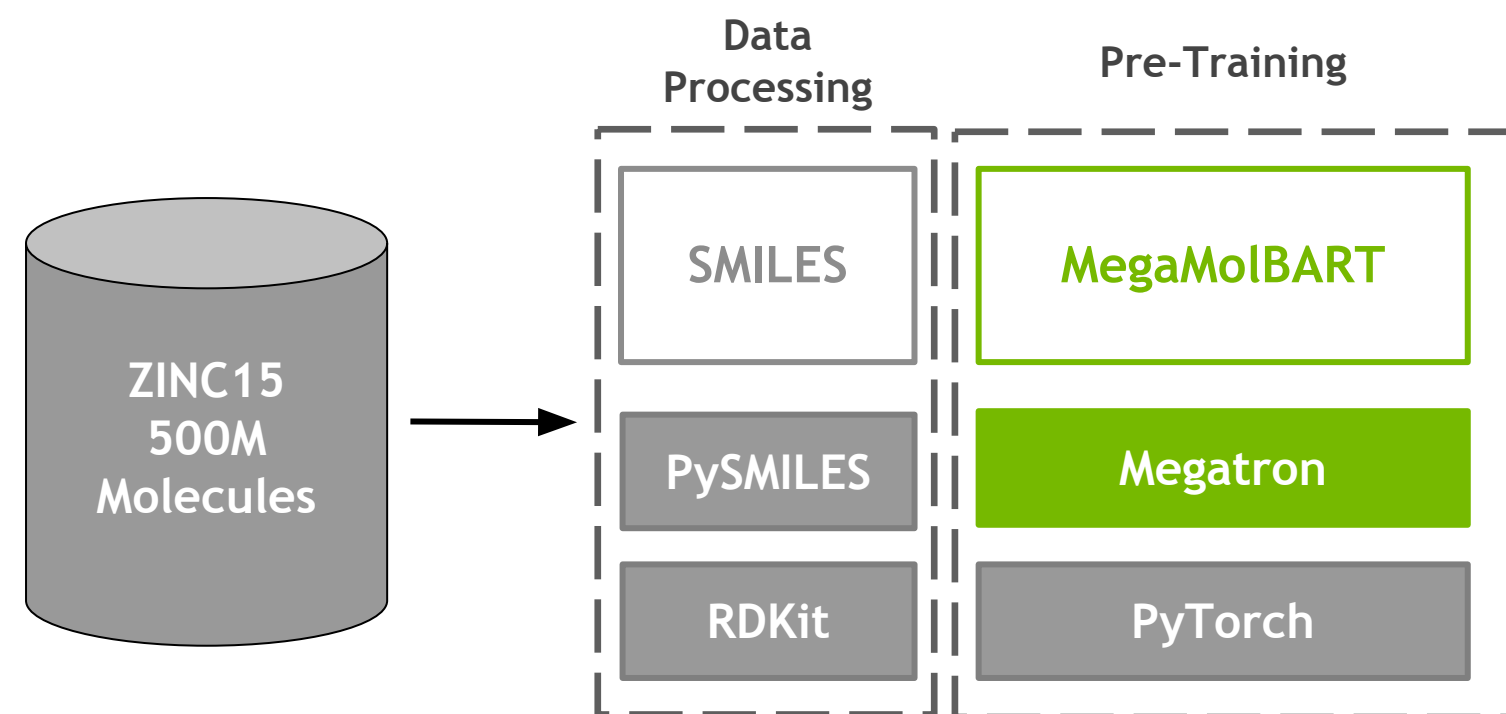
MegaMolBART is based on a BART (seq2seq) transformer -- bidirectional encoder and autoregressive decoder

Developed in collaboration with AstraZeneca

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



Pre-Training of MegaMolBART

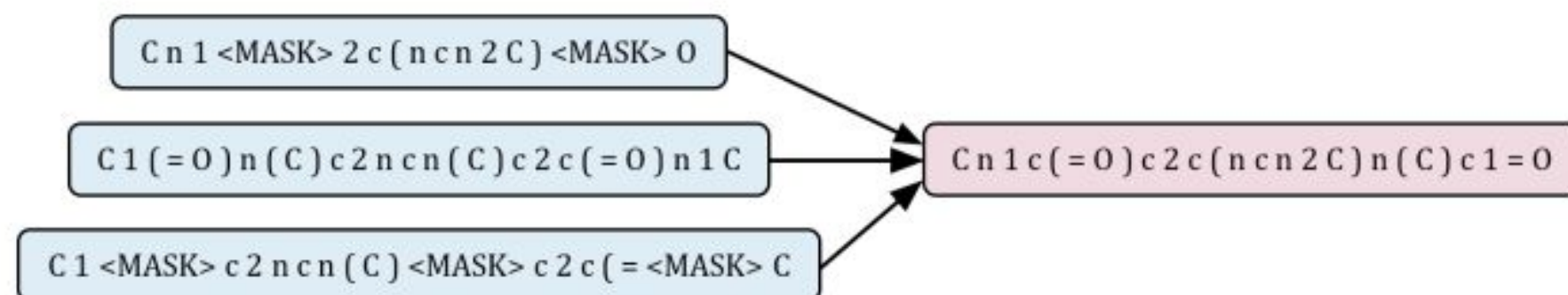


Pre-training performed on ZINC15 -- tranche from reactive, annotated molecules with molecular weight $\leq 500\text{Da}$, and $\text{LogP} \leq 5$

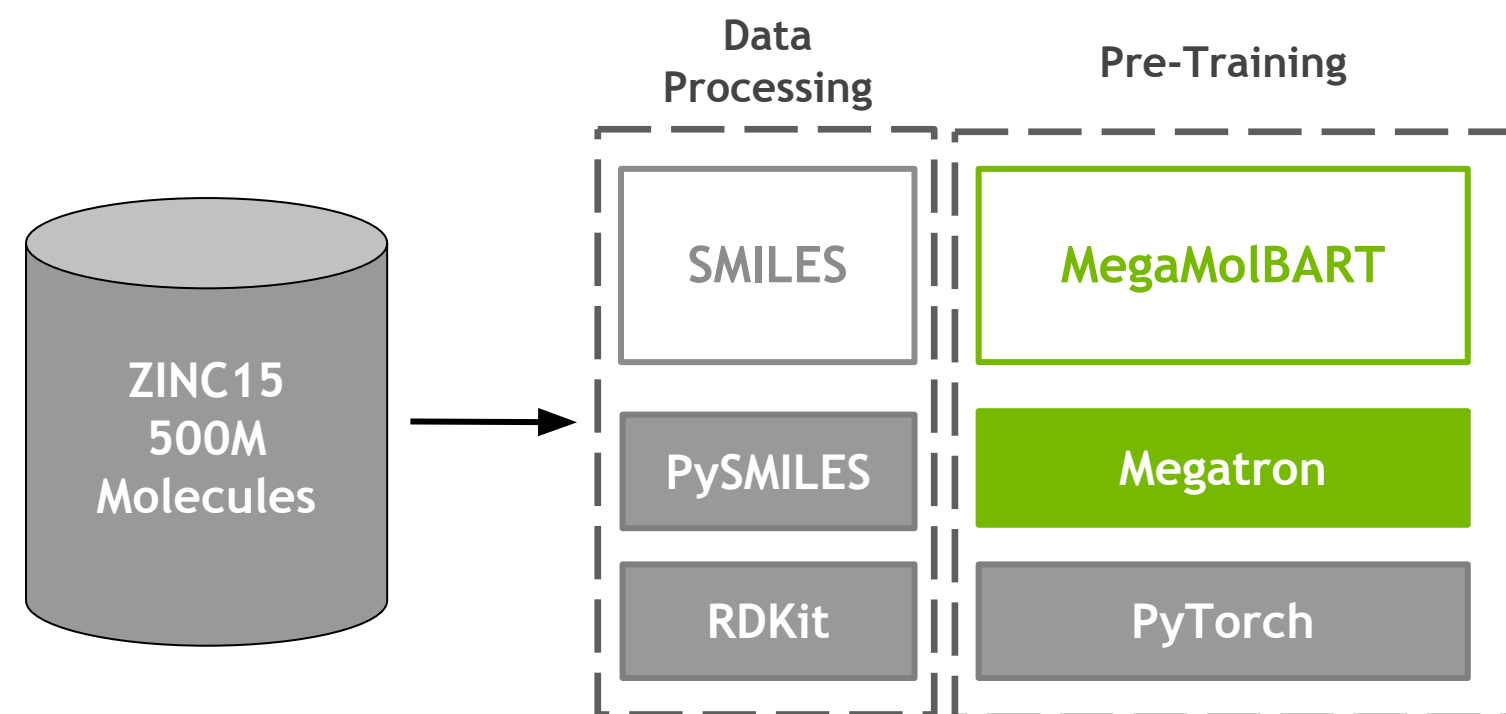
Train, validation, and test splits were 99% / 0.5% / 0.5%

SMILES molecules were masked and enumerated (randomized) during training

SMILES Augmentation



Pre-Training of MegaMolBART



DGX SuperPOD / Cambridge-1



Pre-training performed on ZINC15 -- tranche from reactive, annotated molecules with molecular weight $\leq 500\text{Da}$, and $\text{LogP} \leq 5$

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

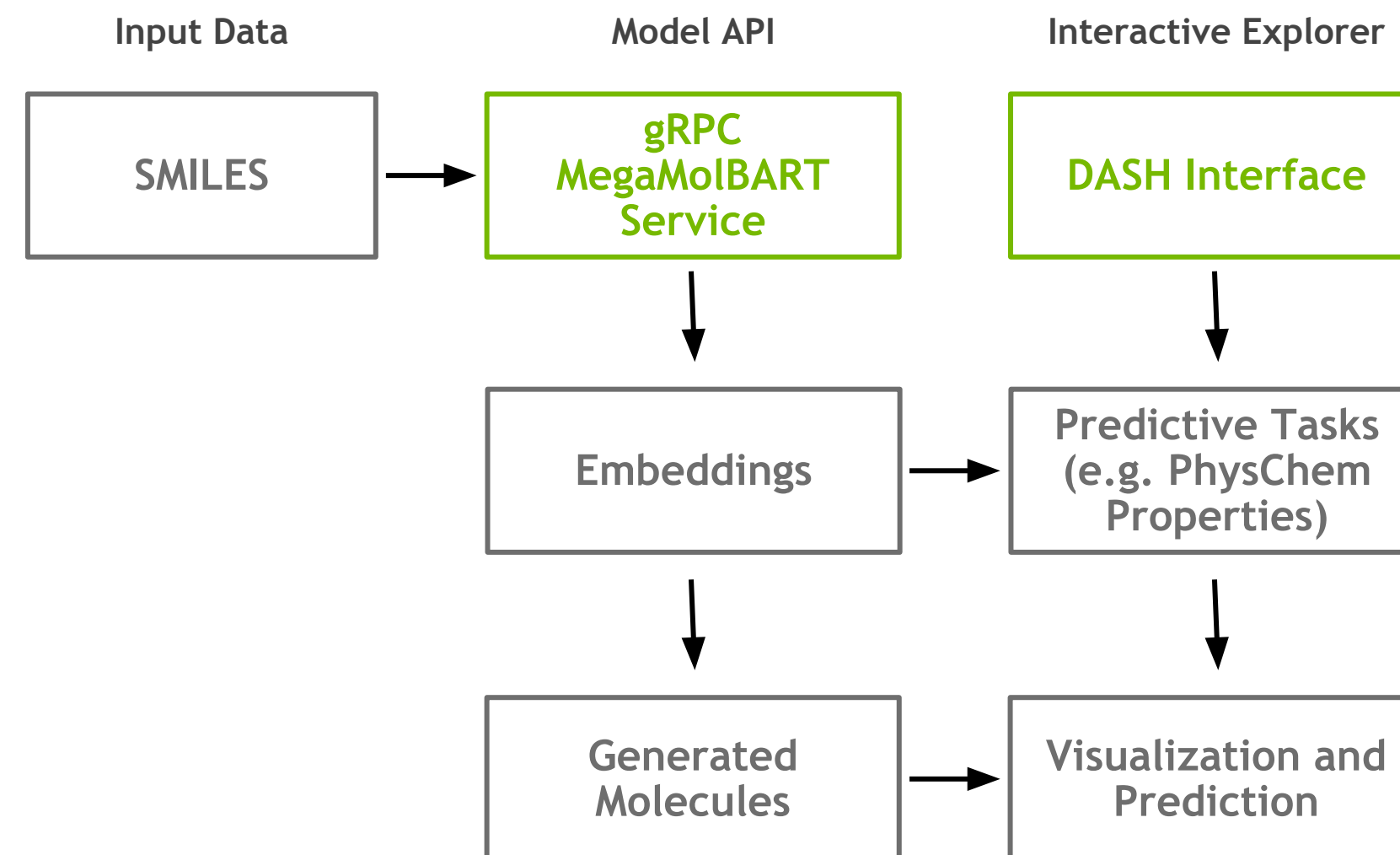
Model inference tasks available via gRPC API:

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

Clusters

Molecule(s) of Interest

Please enter ChEMBL ID(s) separated by commas.

Fingerprint Radius:

Fingerprint Size:

Cluster Molecules | **Generate Molecules** | Predict Properties | Find Analogues

Select Generative Model

MegatronMolBART Model

Sample around one molecule
 Fit cluster to property and extrapolate
 Interpolate between two molecules

Number to be generated from each compound:

Select molecular property for fitting and extrapolation

AlogP

Cluster number for fitting property and extrapolation:

Step-size for extrapolation:

Number of compounds to extrapolate:

Scaled sampling radius (int, start with 1):

Please Select Two Molecules

CHEMBL6201

CHEMBL6251

Select molecular property for color gradient

Select...

[Export to SDF](#)

SMILES	Generated	id	Chemical Structure	Molecular Weight	LogP	H-Bond Donors	H-Bond Acceptors	Rotatable Bonds	QED
<chem>Cc1c(cc(=O)c(C)cc1)C(=O)c1ccc(Cl)cc1</chem>	False	CHEMBL6201		307.73	2.03	2	4	5	0.8284
<chem>Cc1c(cc(=O)c(C)cc1)C(=O)c1ccc(Cl)cc1</chem>	True	CHEMBL6201-CHEMBL6251_1		307.73	2.03	2	4	5	0.8284
<chem>O=C(O)c1cnc(C(=O)c2ccc(Cl)cc2)c(CO)c1</chem>	True	CHEMBL6201-CHEMBL6251_2		291.69	2.16	2	4	4	0.8425

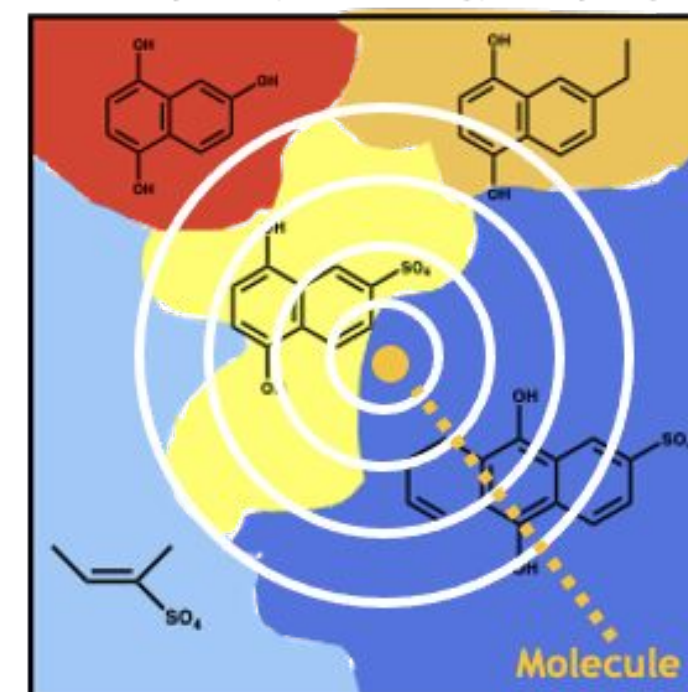
What's Next?

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
...

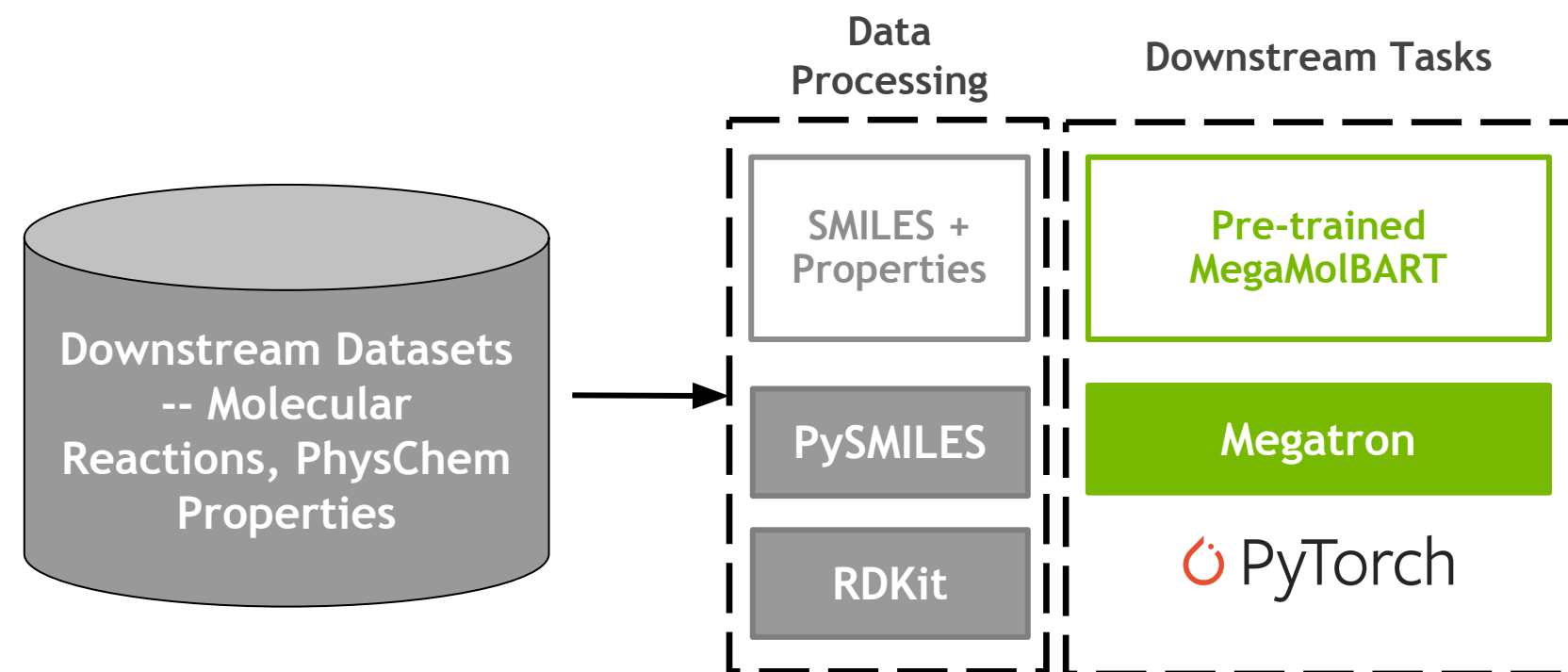
Improved molecule generation -- development of novel model architectures

Latent Space (Embedding) Sampling

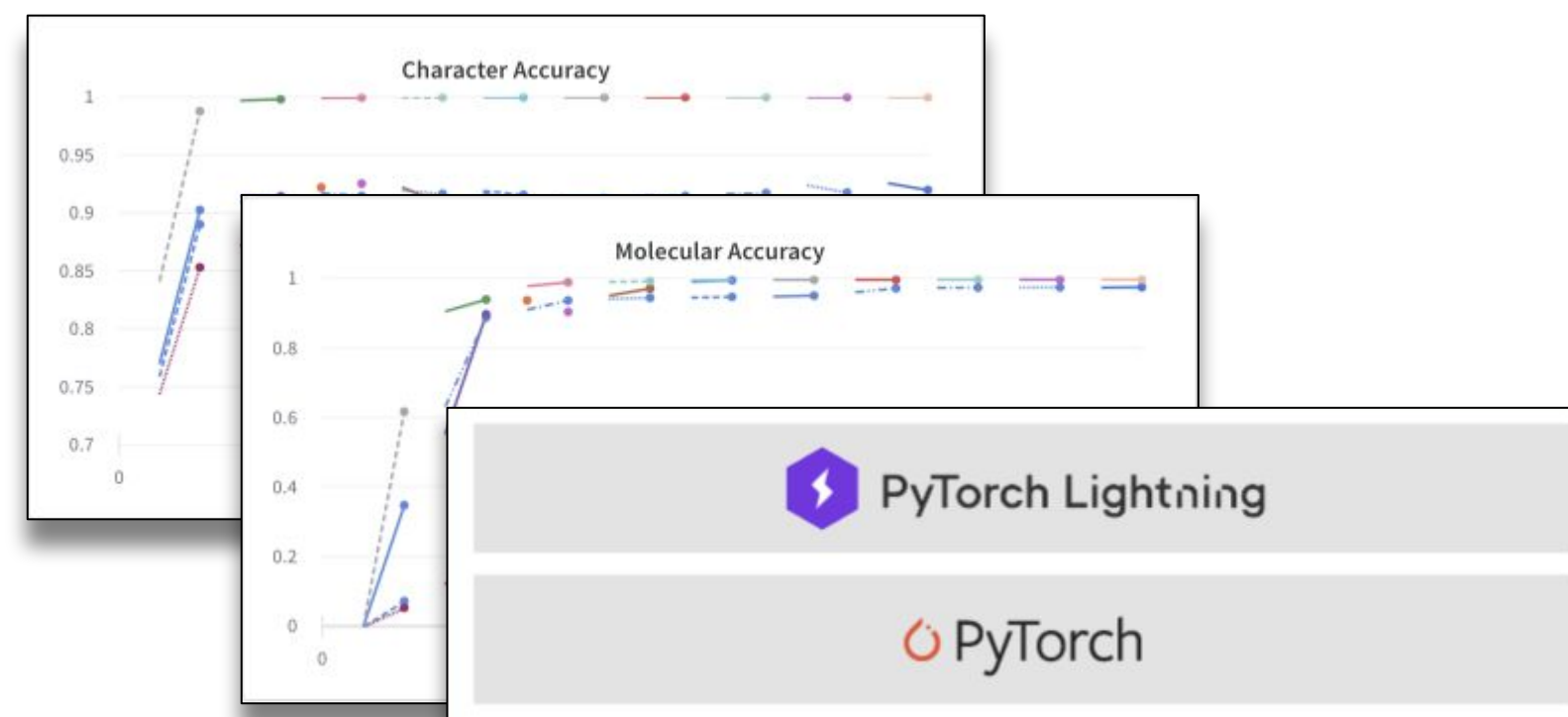


What's Next?

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

Featurizer Service

Interactive Explorer

Tutorials

<https://ngc.nvidia.com/catalog/resources/nvidia:clara:cheminformatics>

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

<https://ngc.nvidia.com/containers/nvidia:clara:megamolbart>

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

