# Real Time, GPU-Accelerated Analysis and Visualization in the Life Sciences

Michelle Gill, PhD and Avantika Lal, PhD
October 27, 2020

# Outline

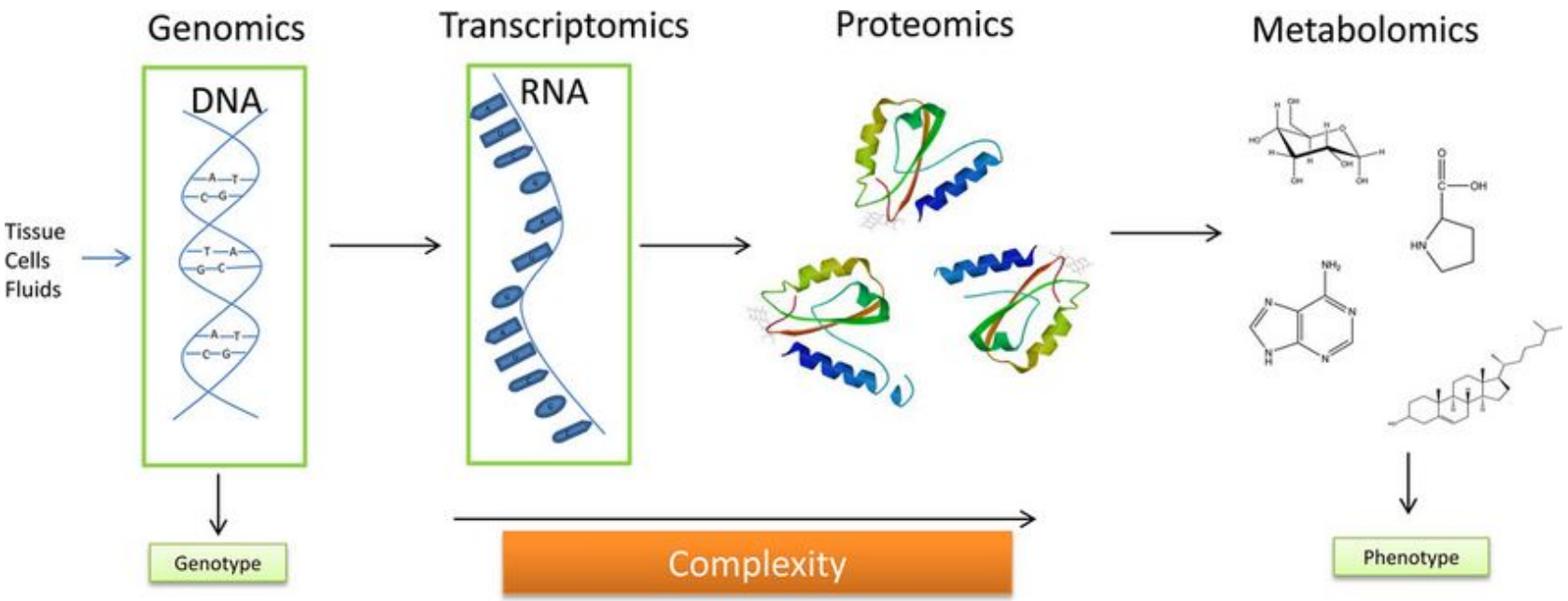
Overview of bioinformatics and the drug discovery process

GPU accelerated data science

Two examples of real-time, interactive clustering:

- Identifying SARS-CoV-2 susceptible cells in the human lung
- Virtual screen used to select potential SARS-CoV-2 inhibitors

# Layers of Bioinformatics Data

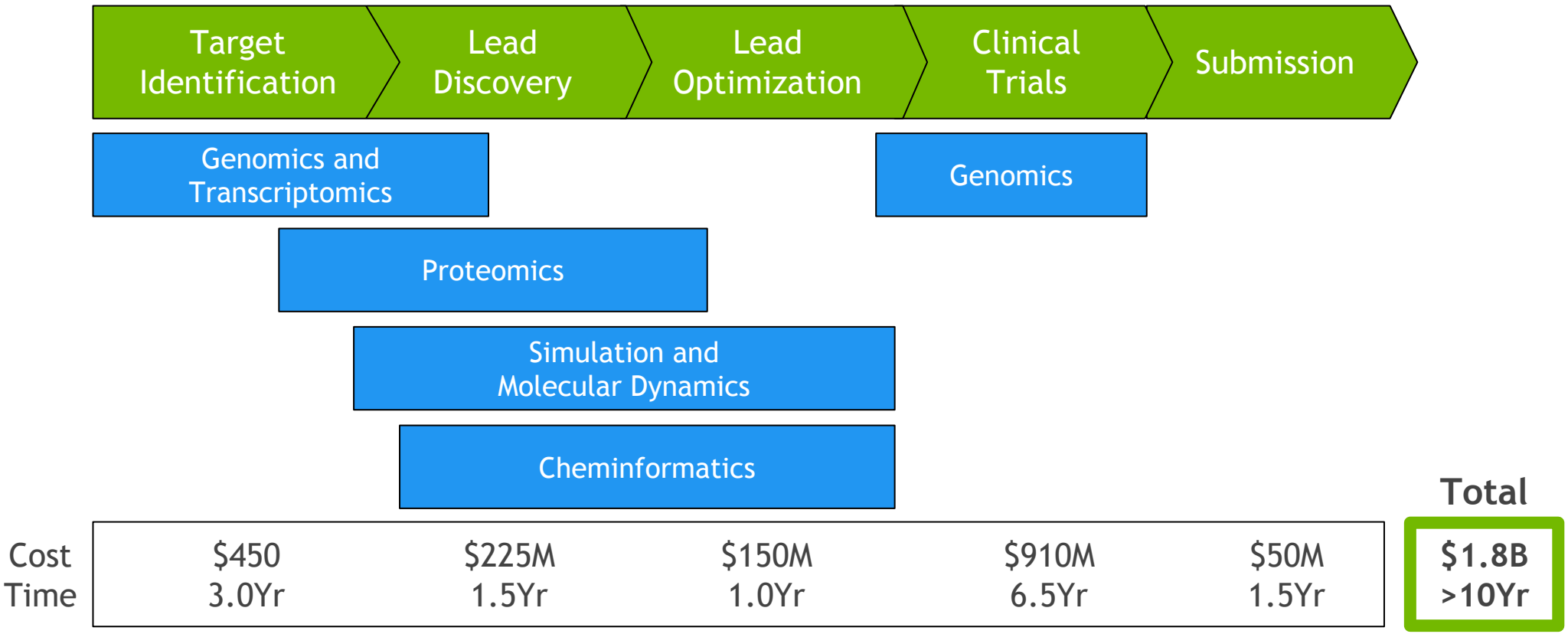



**Transcriptomics** measures the activity of genes

**Cheminformatics** is a family of techniques associated with the search and retrieval of chemical compounds

*Source: Minai-Tehrani, et al. Andrologia 48.6 (2016).*

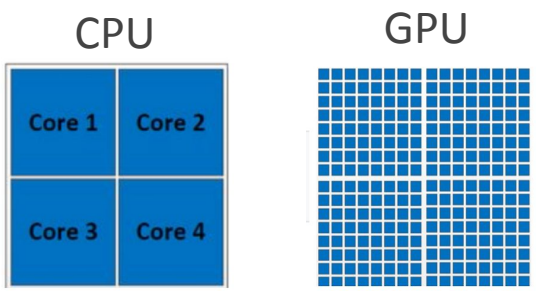
# Drug Discovery: an Informatics Perspective



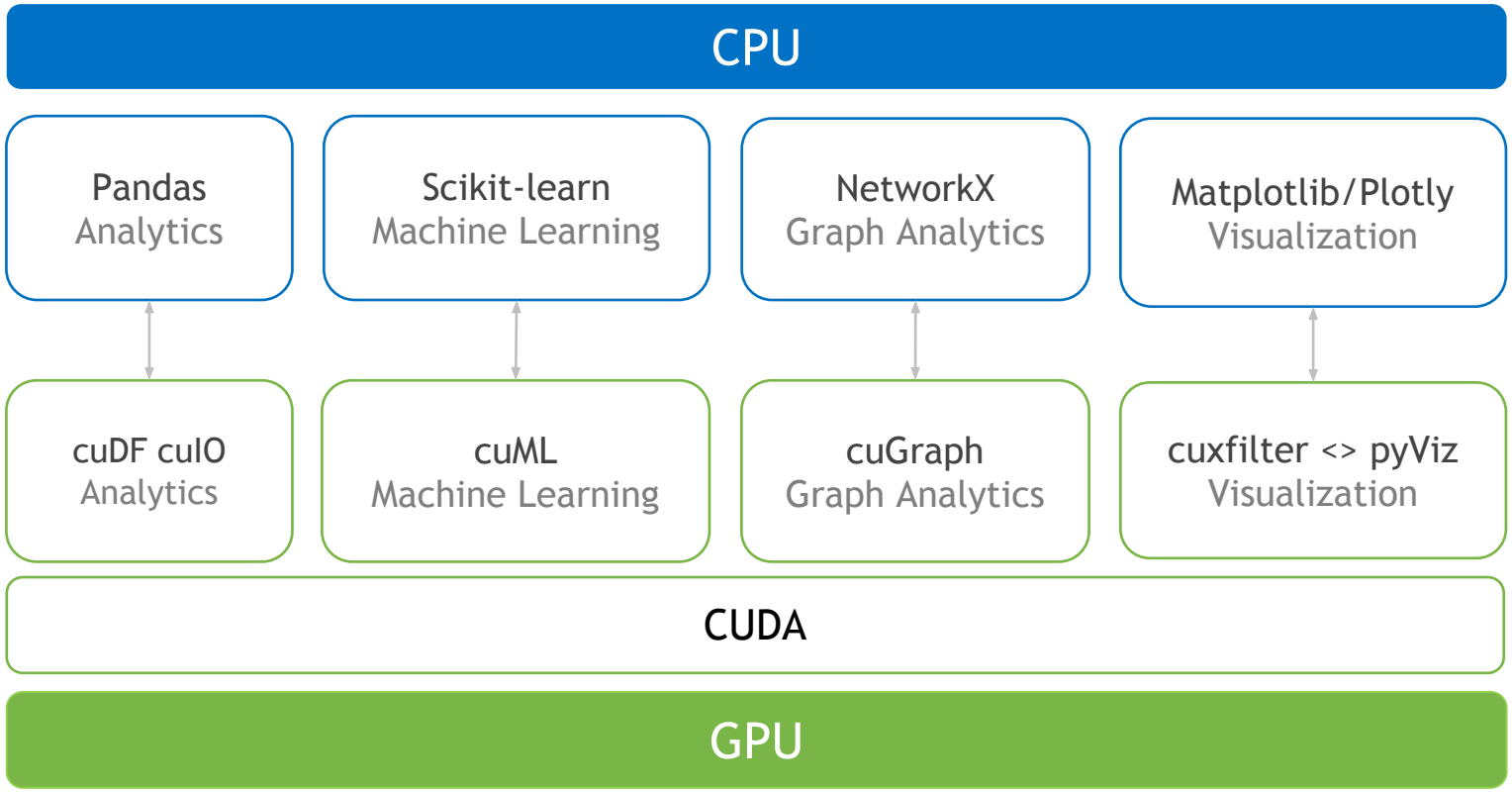

## $1.8B and >10 Years to Bring a Drug to Market

*Source: Zurdo. Pharmaceutical BioProcessing. 1.1 (2013).*

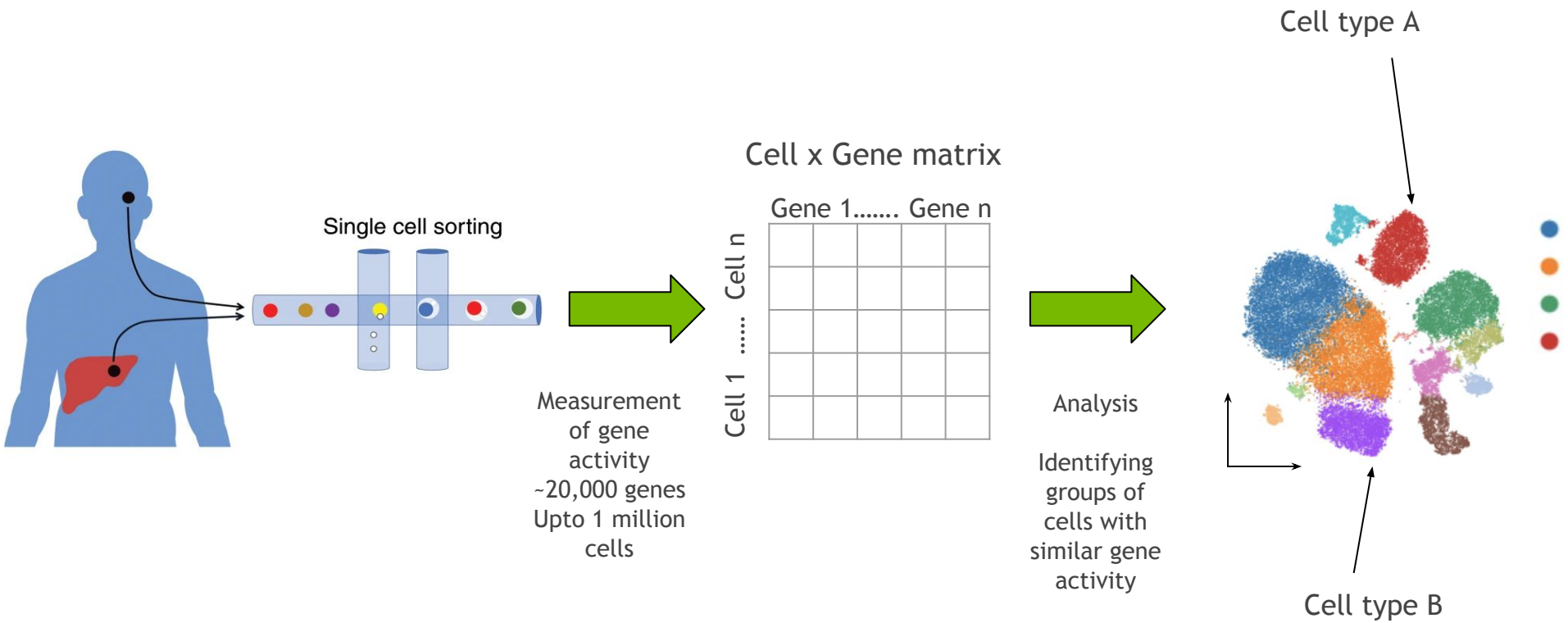
# RAPIDS Accelerates Scientific Discovery on the GPU




Graphics Processing Units (GPUs) perform thousands of operations in parallel using CUDA

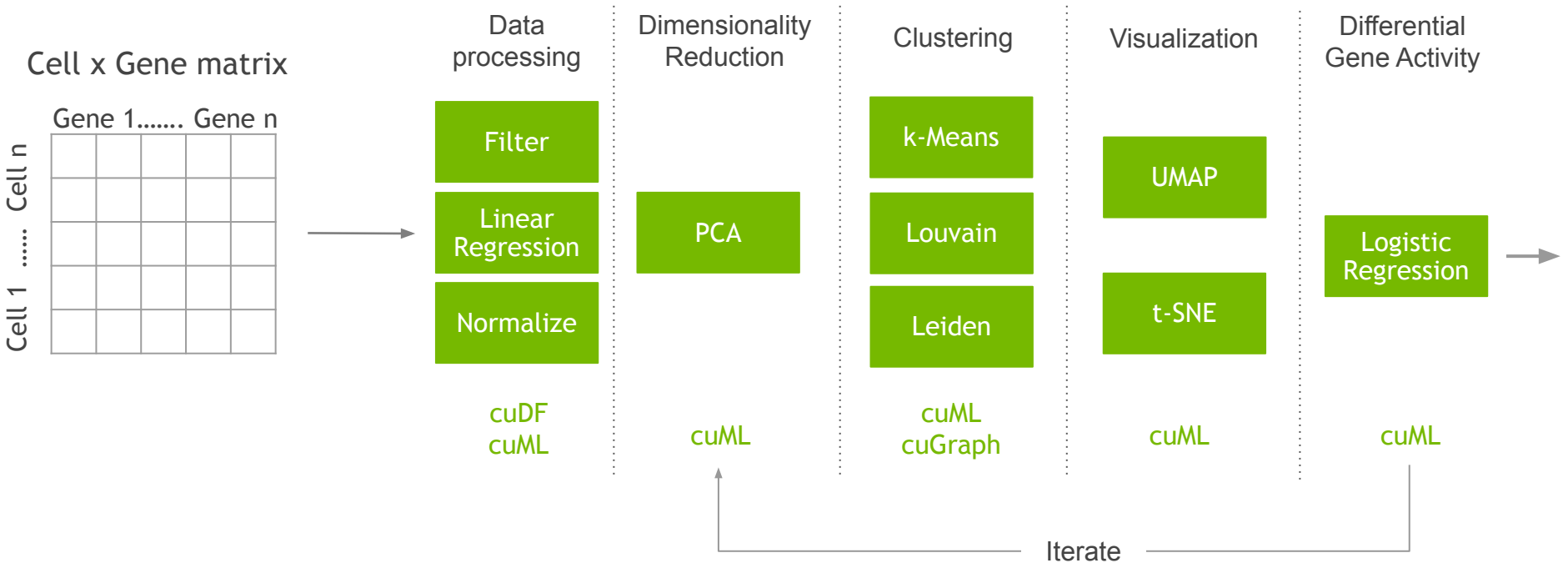


# Single-Cell Transcriptomics

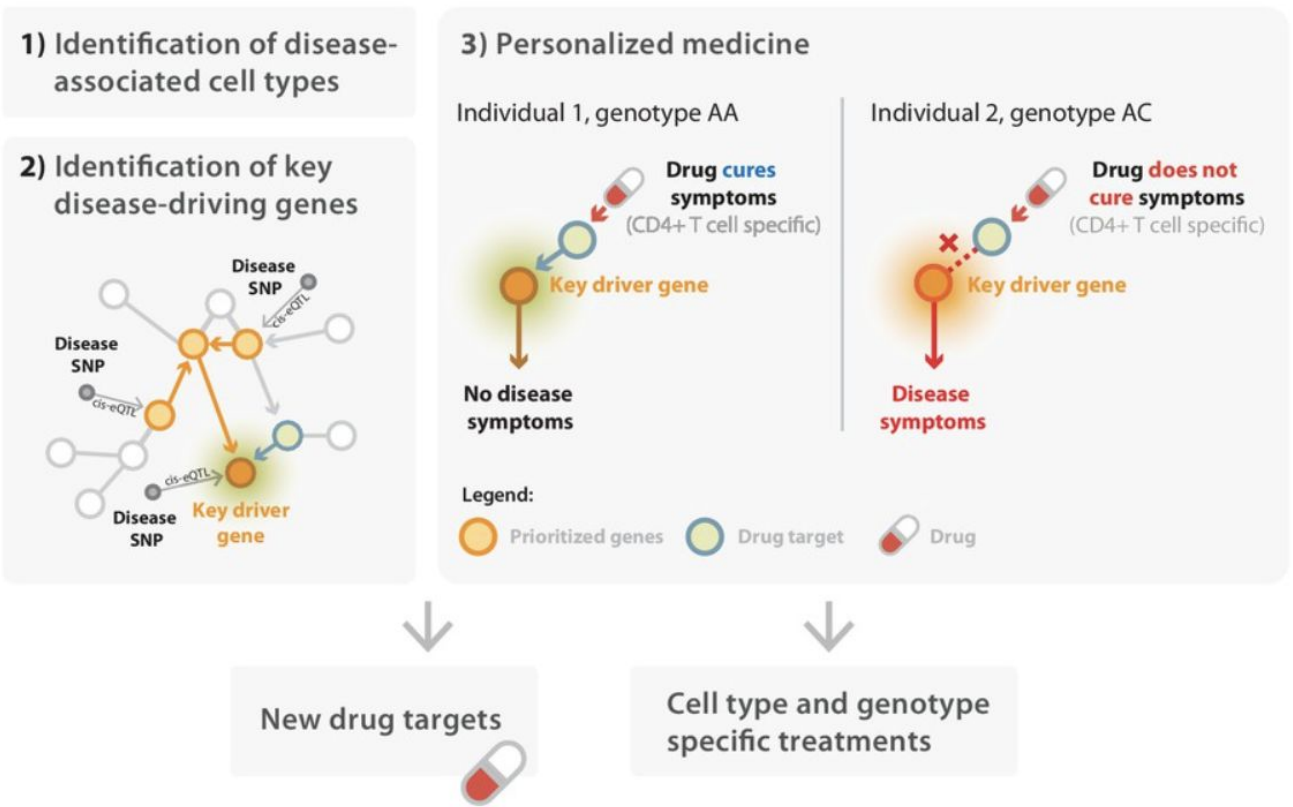


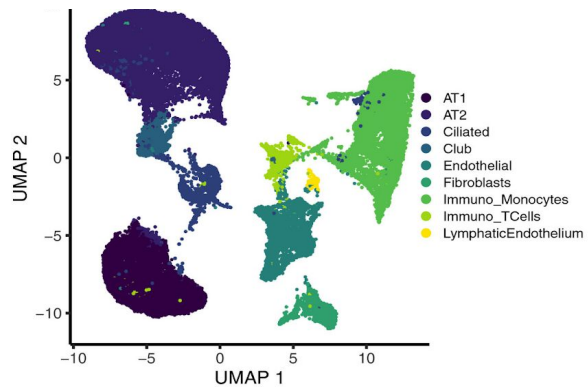

*Source: Durante et al., 2020*

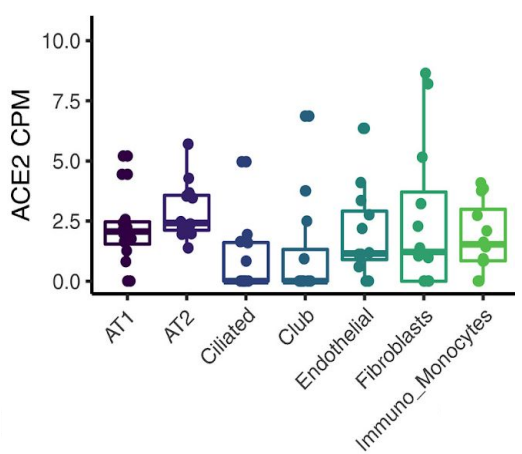
# Single-cell Transcriptomics Workflow



# Single-Cell Transcriptomics in Medicine and Drug Discovery

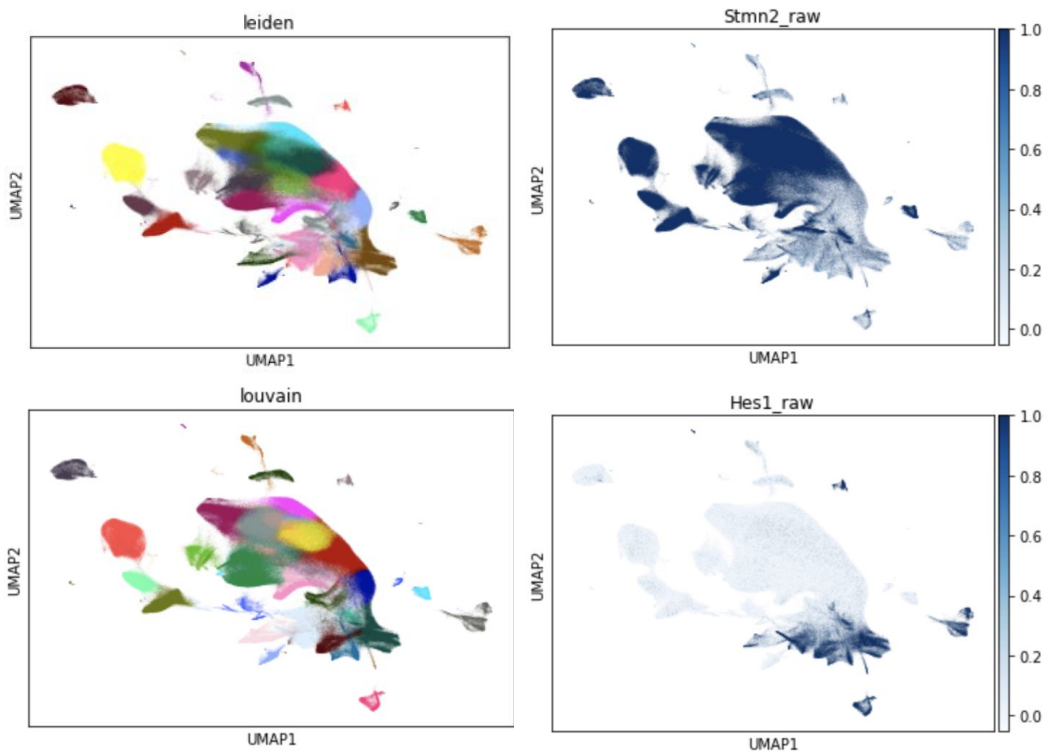



*Sources: van der Wijst, et al. "Science Forum: The single-cell eQTLGen consortium." eLife 9 (2020); Lukassen et al. (2020).*

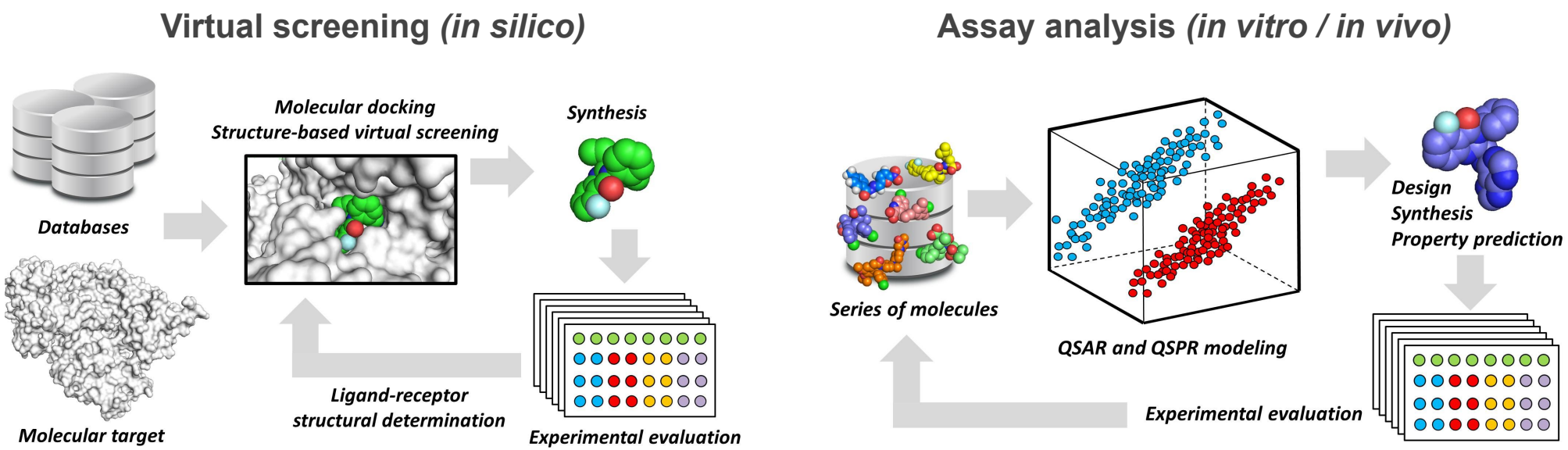
# Five Hours -> Twelve Minutes with GPUs




| Step | CPU runtime m5a.12xlarge Intel Xeon Platinum 8000, 48 vCPUs | GPU runtime g4dn.12xlarge T4 16 GB GPU (Acceleration) | GPU runtime p3.8xlarge Tesla V100 16 GB GPU (Acceleration) |
|---|---|---|---|
| Preprocessing | 4337 | 344 (13x) | 336 (13x) |
| PCA | 29 | 28 (1.04x) | 23 (1.3x) |
| t-SNE | 5833 | 134 (44x) | 38 (154x) |
| k-means (single iteration) | 113 | 13.2 (8.6x) | 2.4 (47x) |
| KNN | 670 | 106 (6.3x) | 55.1 (12x) |
| UMAP | 1405 | 87 (16x) | 19.2 (73x) |
| Louvain clustering | 573 | 5.2 (110x) | 2.8 (205x) |
| Leiden clustering | 6414 | 3.7 (1733x) | 1.8 (3563x) |
| Re-analysis of subgroup | 249 | 10.9 (23x) | 8.9 (28x) |
| End-to-end notebook run (steps above + data load and additional processing) | 19908 | 912 | 702 |
| Price ($/hr) | 2.064 | 3.912 | 12.24 |
| Total cost ($) | 11.414 | 0.991 | 2.388 |

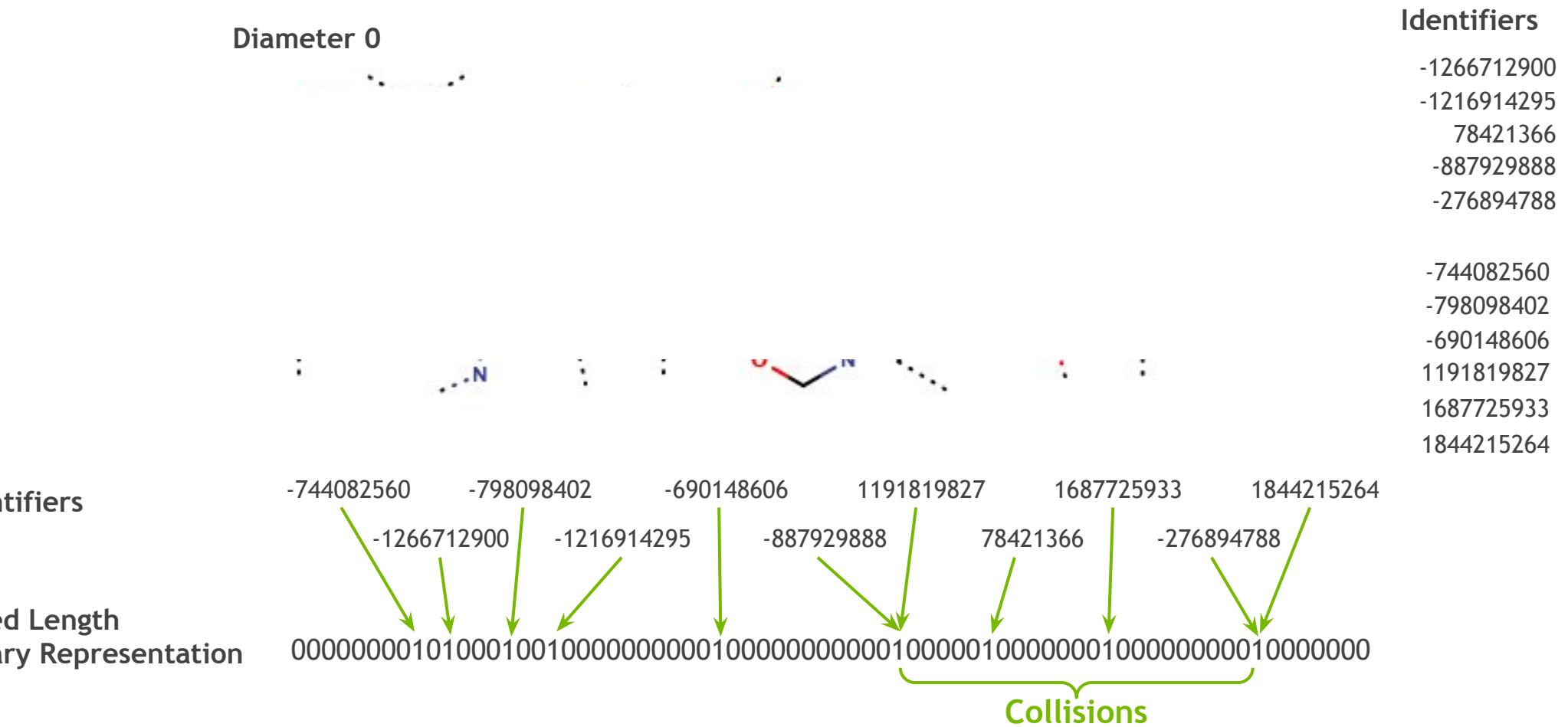
# Real-Time, Interactive Browsing of Human Lung Cells

# Cheminformatics in Drug Discovery



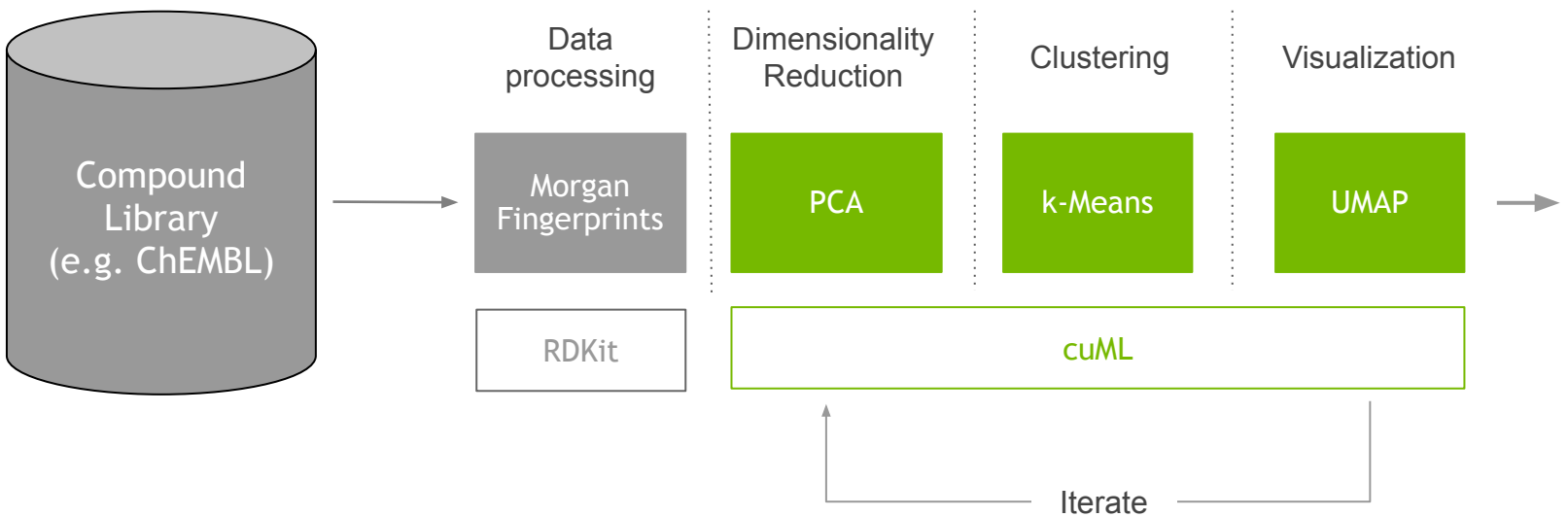

Source: Ferreira and Andricopulo. Front. Pharmacol. (2018) 9.1278.

# Featurization of Chemicals with Morgan Fingerprints




*Source: Rogers and Hahn. J. Chem. Inf. Model. (2010) 50.5.*
*Image: ChemAxon ECFP documentation*

# Cheminformatics Workflow



# Virtual Screen of COVID-19 Drug Candidates

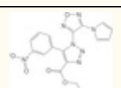
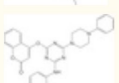
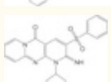
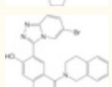
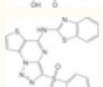
FEBS openbio

FEBS PRESS science publishing by scientists

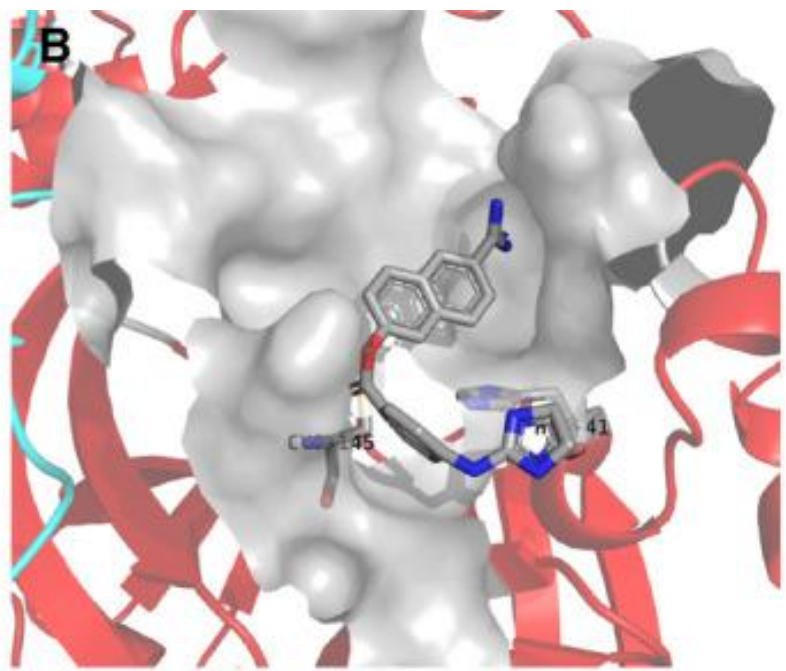
**Potential anti-SARS-CoV-2 drug candidates identified through virtual screening of the ChEMBL database for compounds that target the main coronavirus protease**

Motonori Tsuji

Institute of Molecular Function, Misato-shi, Saitama, Japan

| CHEMBL ID | Structure | Target |
|---|---|---|
| CHEMBL1559003 |  | Survival motor neuron protein |
| CHEMBL2237553 |  | *Aspergillus niger* |
| CHEMBL1511674 |  | Histone-lysine *N*-methyltransferase MLL |
| CHEMBL3260476 |  | Heat shock protein HSP 90-alpha |
| CHEMBL1170272 |  | Serotonin 6(5-HT6) receptor |

3-Chymotrypsin-Like Protease from SARS-CoV-2 (PDB ID 6Y2G) with Sepimostat (CHEMBL114586)




*Source: Tsuji. FEBS OpenBio 10.6 (2020).*

# Real-Time, Interactive Browsing of COVID-19 Drug Candidates

# Acknowledgements

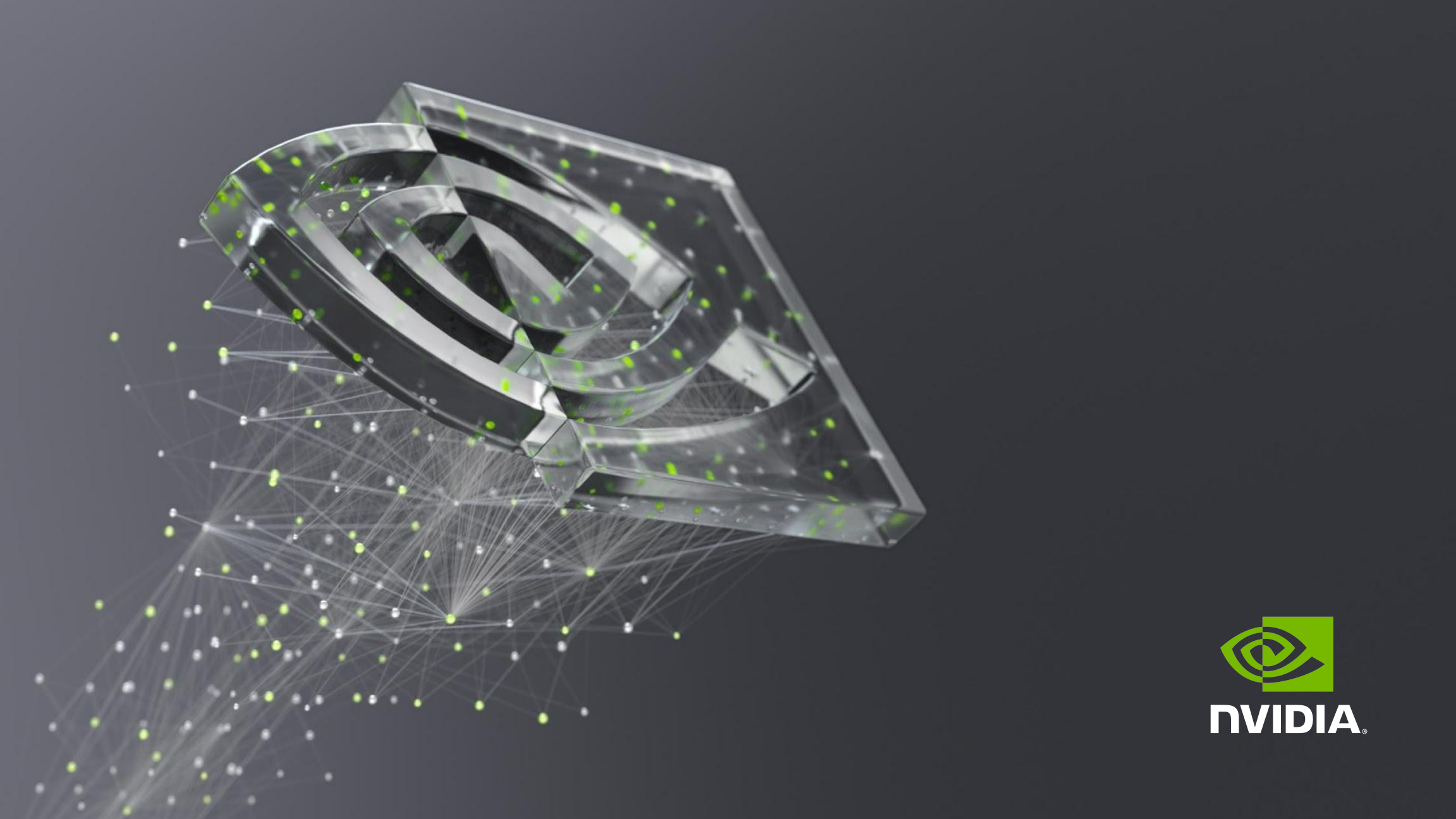
Pat Walters, RELAY Therapeutics
Abe Stern, NVIDIA
Rajesh Ilango, NVIDIA
Corey Nolet, NVIDIA
Taurean Dyer, NVIDIA
John Zedlewski, NVIDIA
Johnny Israeli, NVIDIA

## GitHub Repos

clara-parabricks/rapids-single-cell-examples
NVIDIA/cheminformatics

## Twitter

Michelle: @modernscientist
Avantika: @lal_avantika


NVIDIA

# Who We Are

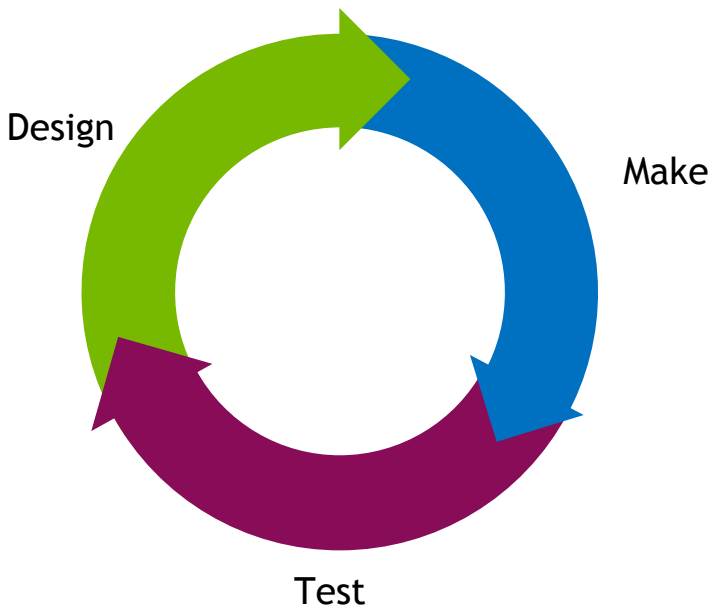
Michelle Gill, PhD
Senior Scientist - Deep Learning,
Proteomics, and Cheminformatics

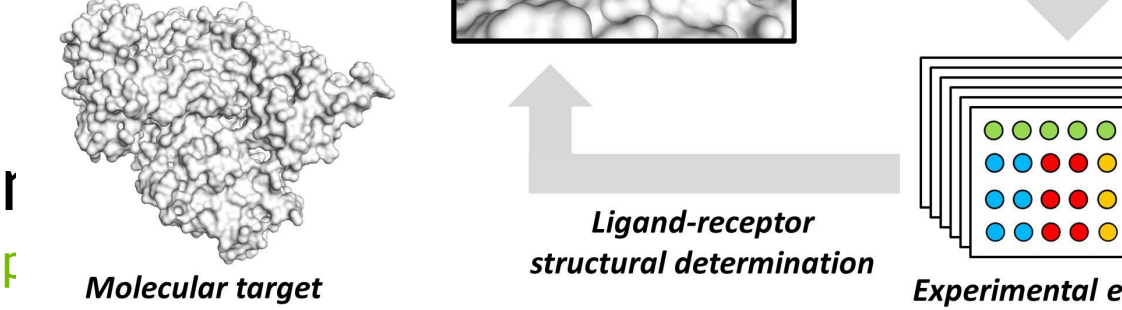
Avantika Lal, PhD
Senior Scientist - Deep Learning,
Genomics

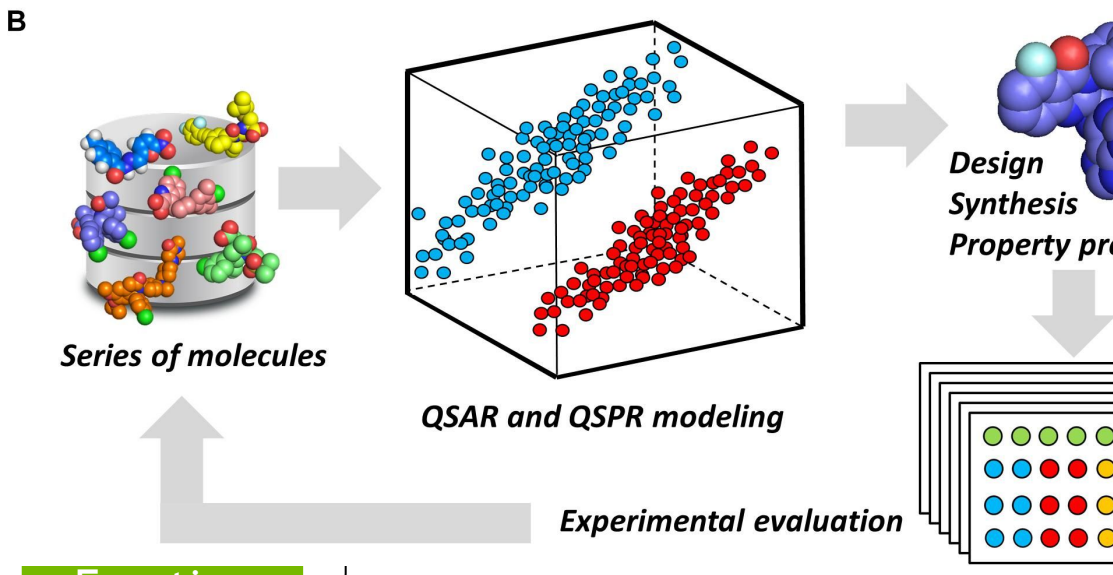
# Real-Time, Interactive Browsing of COVID-19 Drug Candidates

# Cheminformatics ir

Subtitle Op



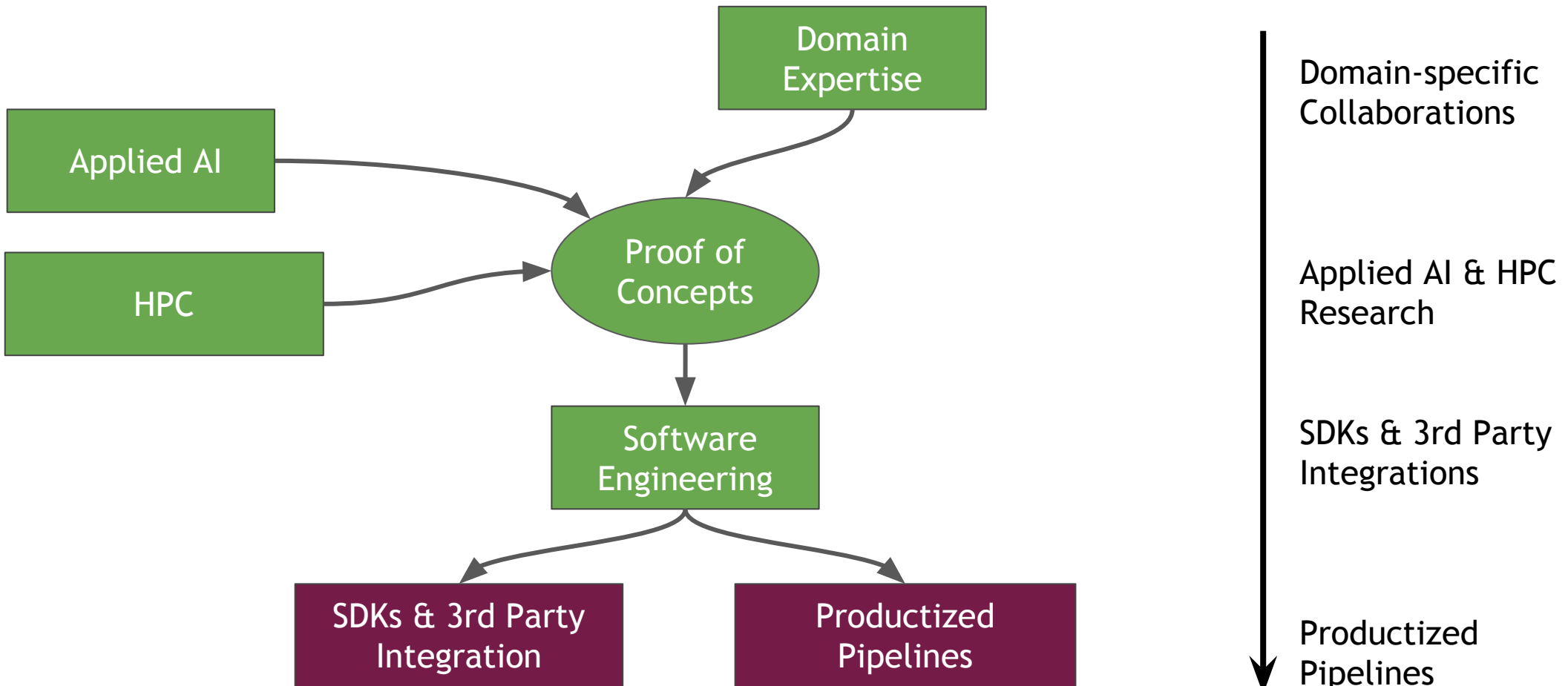

Function

Safety

ADMET

QSAR

# How We Work


Domain Expertise
Applied AI
HPC
Proof of Concepts
Software Engineering
SDKs & 3rd Party Integration
Productized Pipelines
Domain-specific Collaborations
Applied AI & HPC Research
SDKs & 3rd Party Integrations
Productized Pipelines

# CONTENT SLIDE

## Subtitle Optional

Body/bullet text with no bullet icon

Use 14 pt Trebuchet font

No sub-bullets

No more than five bullets; one idea per bullet

Example of highlighted text

# SAME AS PREVIOUS SLIDE

## But Includes Logo and Page Number in Lower Right Corner

# BULLET ICON

- Body/bullet text WITH a bullet icon option

# NVIDIA CONFIDENTIAL

## Use This for Slides That Include “Confidential” Information/Data

# TITLE ONLY SLIDE


SECTION DIVIDER OR
TRANSITION SLIDE