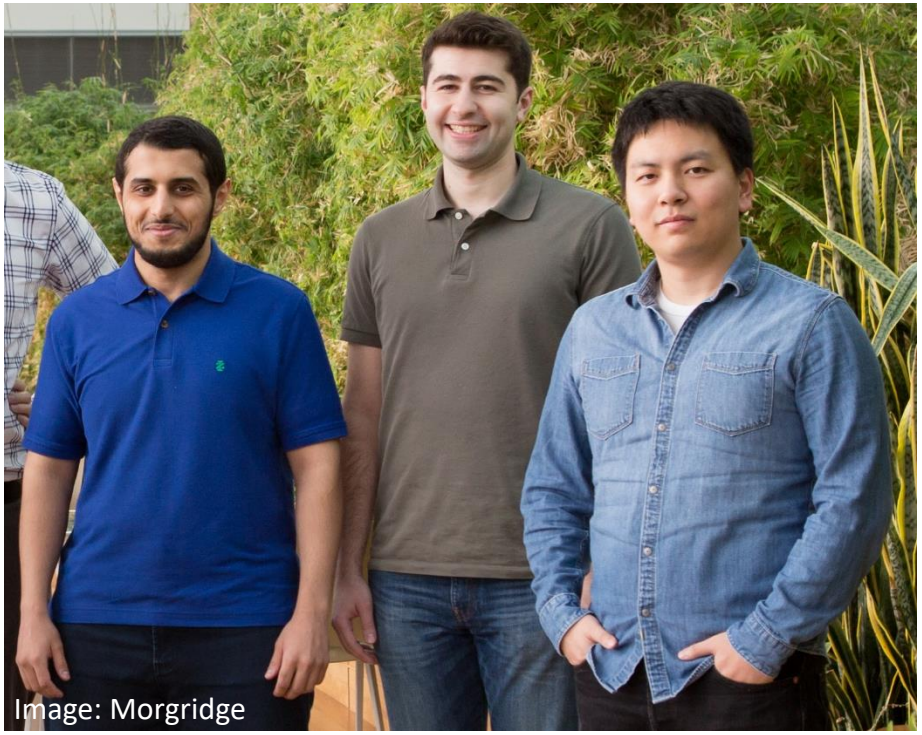
HTCondor for machine learning in biology

Anthony Gitter
University of Wisconsin-Madison
Morgridge Institute for Research
May 22, 2018

gitter@biostat.wisc.edu www.biostat.wisc.edu/~gitter @anthonygitter

Drug discovery and GPU computing



Moayad Alnammi Shengchao Liu

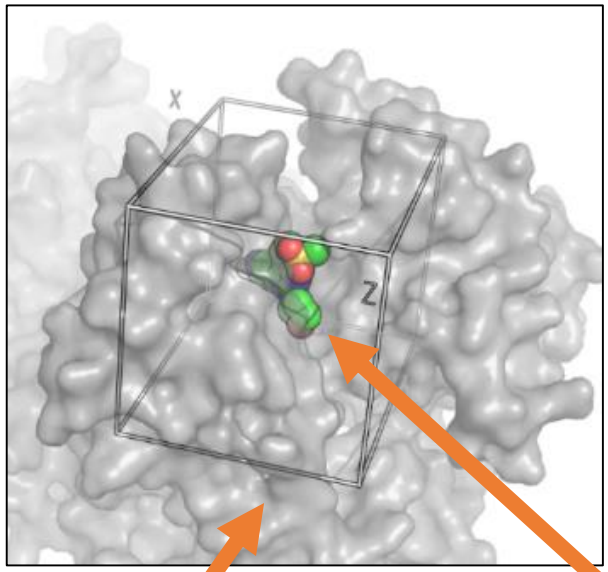
Sam Gelman

High-throughput chemical screening

Given a protein of interest, identify chemicals that may have the ability to control the protein

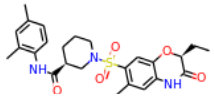
Computational chemical prioritization

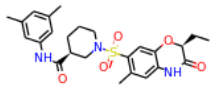
New protein target

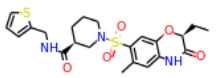


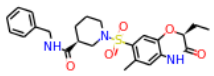
> 1 million compounds

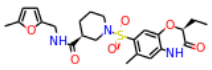
Suggested test order

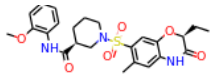

5*


30390


201


980000


2*

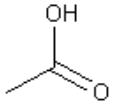
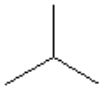
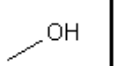
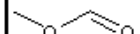
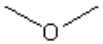
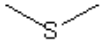

458

…

Chemical compound ≈ key

Protein ≈ lock

Chemical representations for machine learning

1	1	0	1	0	0	0	0
							

2-D searching tutorial

Chemception

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Wikipedia SMILES

PubChem

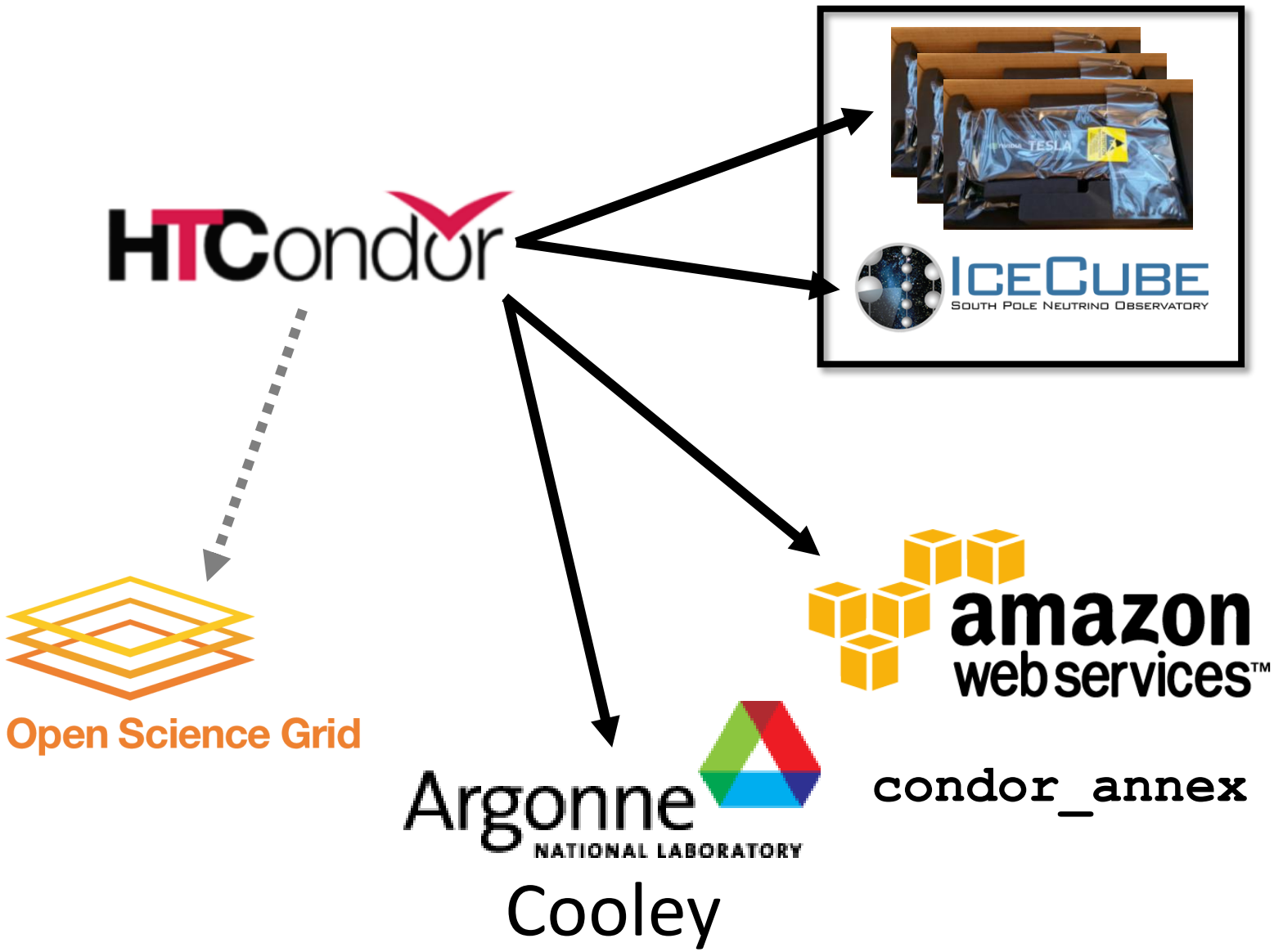
See Ching *et al*. 2018

Training neural networks

We Us

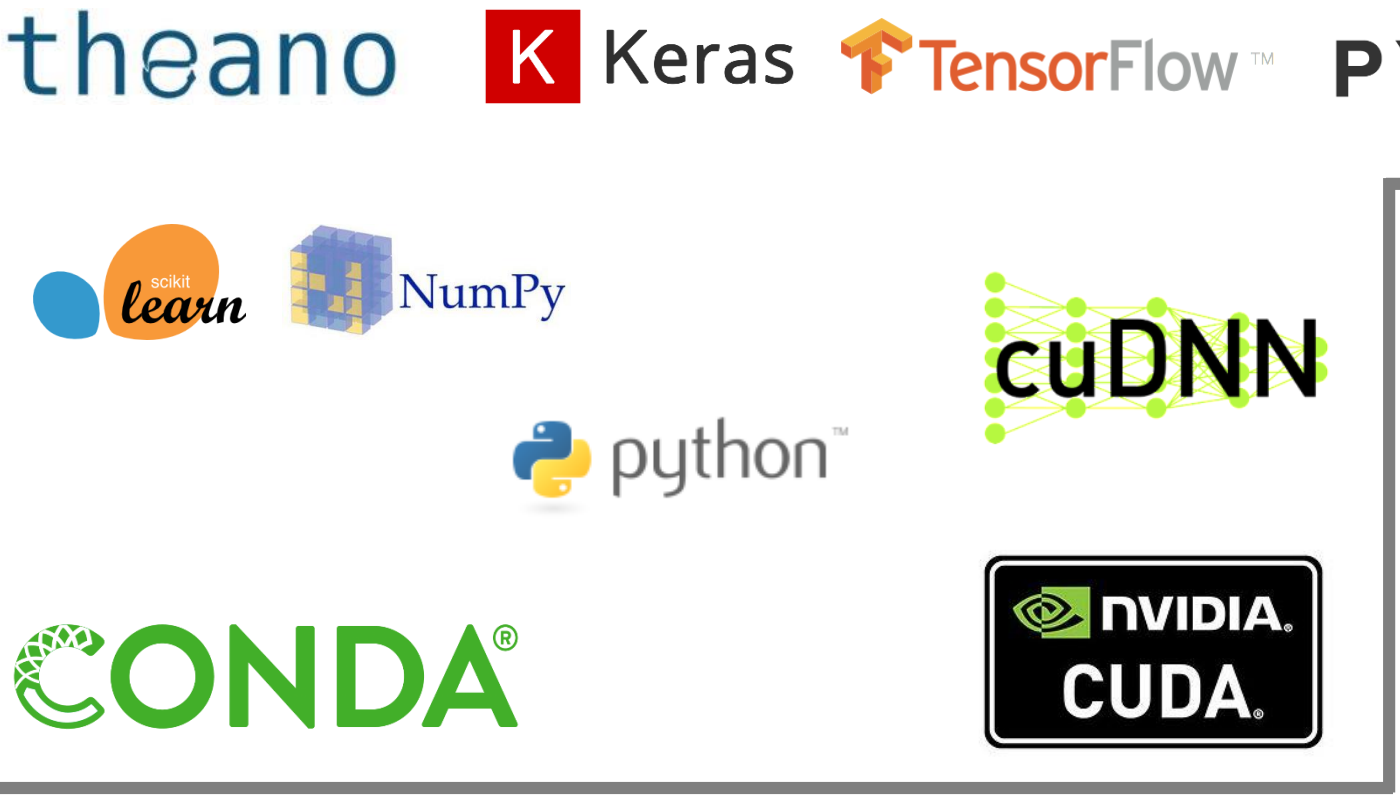
So ls

Expanding our GPU capacity



Software dependencies

Chemical screening software



NVIDIA driver XXX.YY

Tradeoffs of conda

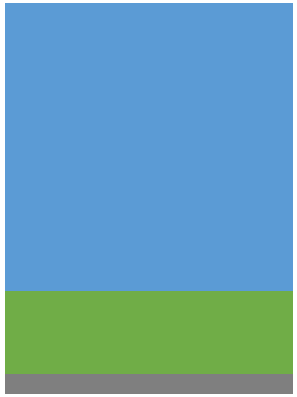
Pros:

- Easy to install Anaconda
- Environments are shareable
- Easy to update packages in environment

Cons:

- Still depend on system libraries
- **`conda_submit gpu-job.sub`**
- **`condor install numpy`**

Predictive models perform well in experimental tests



Train on 75k chemicals, PriA-SSB inhibition
Choose among many possible models

Models select 250 of 25k new chemicals

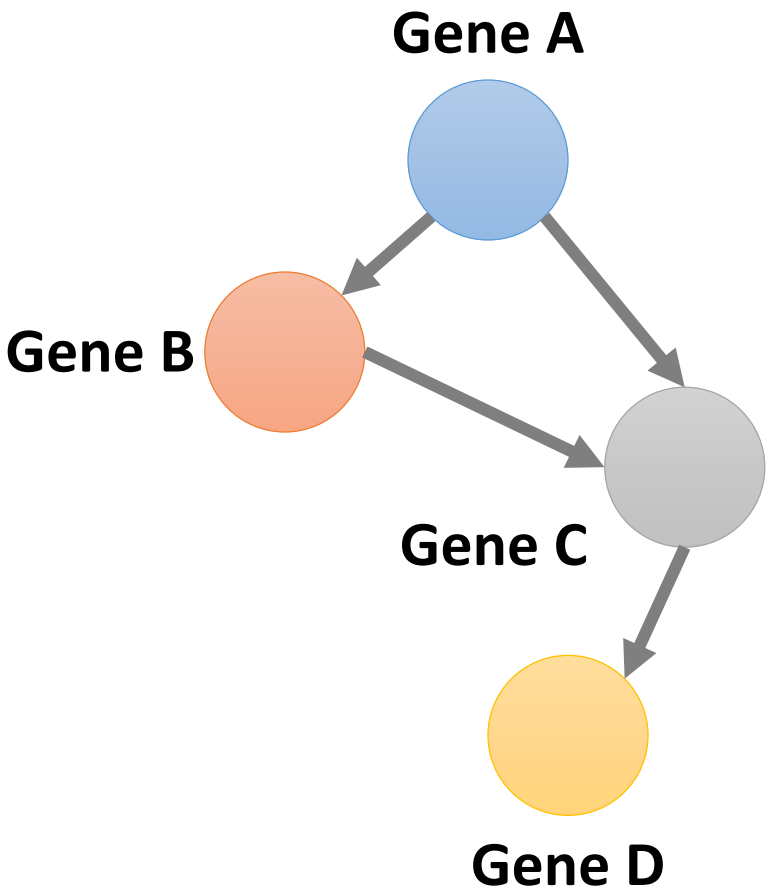
- 64 active chemicals in the 25k
- Model we selected is the best, finds 40 of the 62
- Random forest outperforms the neural networks

- Now testing on much larger chemical libraries

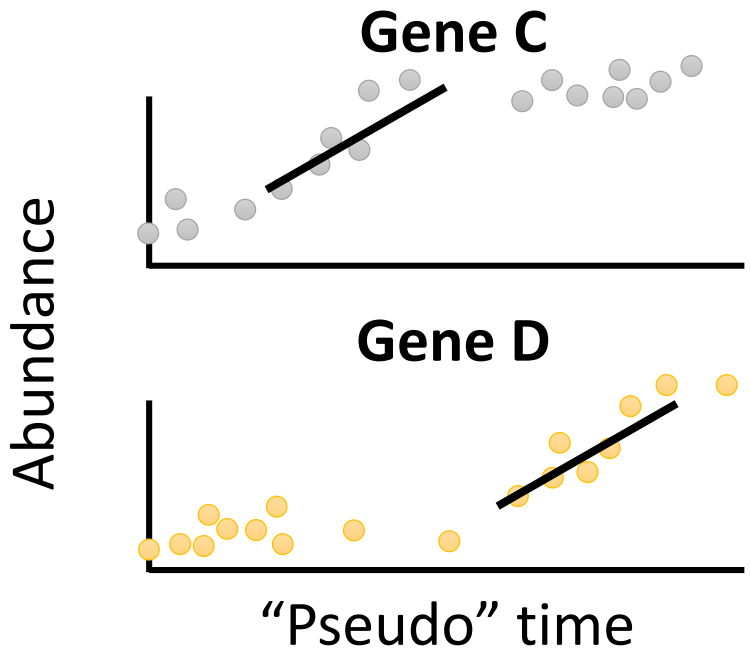
Gene networks and single cell expression

Atul Deshpande

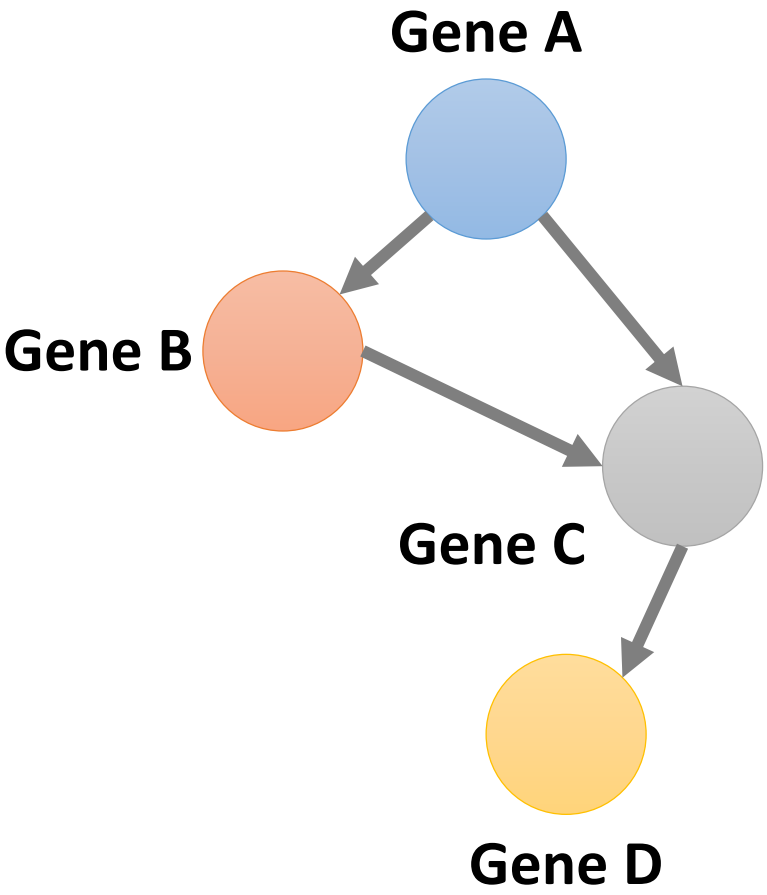
What are the relationships among genes inside a cell?



Measure snapshots of gene abundance



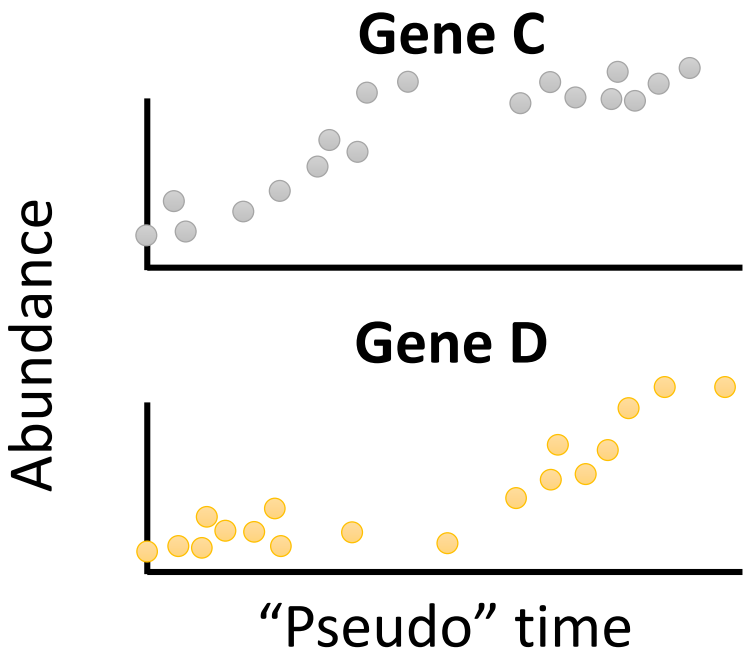
For each gene, learn which genes control it



	A	B	C	D
A		X	X	
B			X	
C				X
D				

Divide network inference into many small computational jobs

	D
A	?
B	?
C	?
D	?



Open Science Grid

100 batches of genes X
10 random bootstraps X
100 parameter combinations

Acknowledgments

UW-Madison, CHTC

- Lauren Michael
- Christina Koch
- Miron Livny
- Todd Miller
- Aaron Moate
- Tim Cartwright
- Jaime Frey
- Greg Thain
- Neil Van Lysel
- Dakota Chambers

Researchers, collaborators

- Moayad Alnammi
- Atul Deshpande
- Sam Gelman
- Shengchao Liu
- Other group members
- Spencer Ericksen
- Scott Wildman
- Michael Hoffmann
- Andrew Voter
- James Keck
- Ron Stewart

Funding and computing resources

- UW-Madison Center for High Throughput Computing
- NVIDIA
- NIH Commons Credits
- Center for Predictive Computational Phenotyping NIH U54 AI117924
- UW Carbone Cancer Center NIH P30 CA014520
- UW-Madison VCRGE with funding from WARF
- PhRMA Foundation
- Morgridge Institute for Research
- NSF CAREER 1553206