ORCHESTRATING MULTI-TASK MATERIAL DESIGN CAMPAIGNS WITH ARTIFICIAL INTELLIGENCE

LOGAN WARD
Asst. Computational Scientist
Data Science and Learning Division
Argonne National Laboratory

8 March 2022
TMS 2022 – Anaheim, CA

AI HAS BEEN PART OF SOCIETY FOR DECADES

A great way for using human effort more effectively

■ Fire victim showered with gifts — Page 3A

■ Dance clubs have classes for all ages — Page 3B

THE New VOLUSIAN

© News-Journal Corporation

Incorporating The Sun News/The Enterprise

Volume 1, No. 55 Serving the Communities of West Volusia County Sunday, September 13, 1992

The mail must get through

By DINAH PULVER

LAKE MARY — When a letter leaves a home or business it begins a whirlwind journey through a high-tech maze of machinery designed to make sure it arrives at its destination in a timely manner.

Yesterday's historic small town post office cancellations have given way to fast-paced technological breakthroughs.

When a mail carrier picks up the mail, it is delivered to large laundry-type baskets at the post office, which in turn are driven in afternoon runs to the Lake Mary Mail Processing Center.

The center handles mail from all post offices in the 327 zip code area. Around 4:30 or 5 p.m., the mail starts arriving at the center and the cavernous facility becomes a vortex of bustling activity.

First, mail is placed on a culling belt that sorts packages from letters. Then it is sent onto an Advanced Facer Canceler System, which turns all the mail to face the same direction.

The facer separates the mail into three groups: script mail, machine printed mail, and bar-coded mail. The system cancels an average of 34,000 pieces of mail per hour and requires one operator.

Machine printed mail goes directly to an Optical Character Reader, while hand-addressed mail is sent to a letter-sorting machine. Bar-coded mail is sent to a bar-code sorter.

The multiline optical character reader "reads" an address block, sprays a matching bar-code on and sorts the letter. It finds the bar-code by consulting its address directory and spraying on the proper "zip plus four" bar-code. The reader averages 36,000 pieces per hour, or 12 per second and requires two operators.

A bar-code reader reads and sorts mail with bar-codes and sorts an average of 35,000 pieces per hour and requires two operators. The bar-code sorter can sort the mail right to its city of destination.

If the mail has no bar-code and cannot be read by the optical character reader, it could be channeled through the remote bar-coding system, where an image is taken of a letter in Lake Mary, and the picture is sent electronically to a remote location somewhere in the United States.

At the remote location, a person keys the information into a computer, which in turn matches that information with its data base and sends up a matching zip-plus-four bar-code.

When the bar-code reader receives the information it matches it with the piece of mail, sprays on a bar-code and sends the piece of mail back through the system.

Everything that cannot be processed by automated equipment must be done by a letter sorting machine or by hand.

On a multiposition letter sorting machine, mail pieces are passed in front of an operator who reads the information and keys in a portion of the address. Mail is processed at the same rate as the with the optical character reader but 18

SEE MAIL/ 5A

Automation speeds mail delivery to customers

DELTONA — In 1970, an act of Congress transformed the Post Office Department into the United States Postal Service to make the operation more business-like and halted tax subsidies.

of his postmaster, Wayne Dunn.

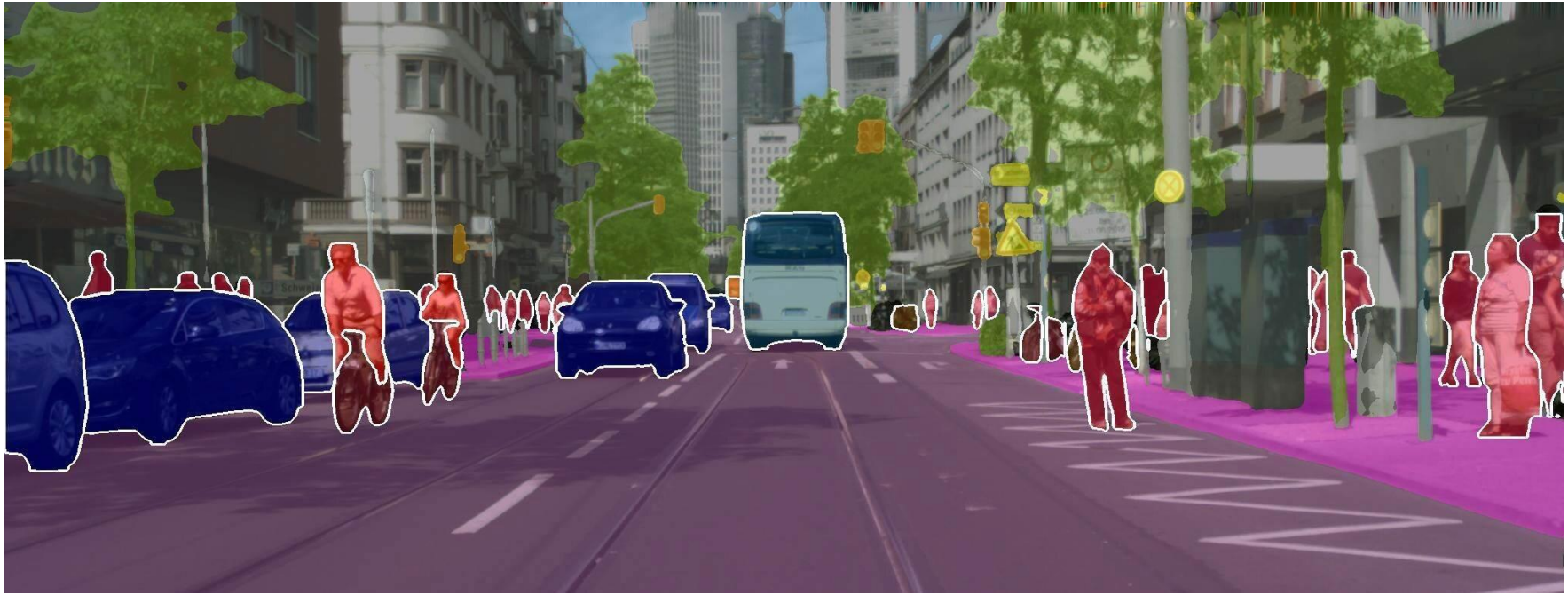
"This is a fast-paced, changing postal service we have now," Masiarczyk said. "A few years ago, most of this was unheard of, but

to a bar-code sorter.

The multiline optical character reader "reads" an address block, sprays a matching bar-code on and sorts the letter. It finds the bar-code by consulting its address directory and spraying on the proper "zip plus four" bar-code. The

… AND THE TOOLS ARE GETTING VERY GOOD

AI in more places than you might expect

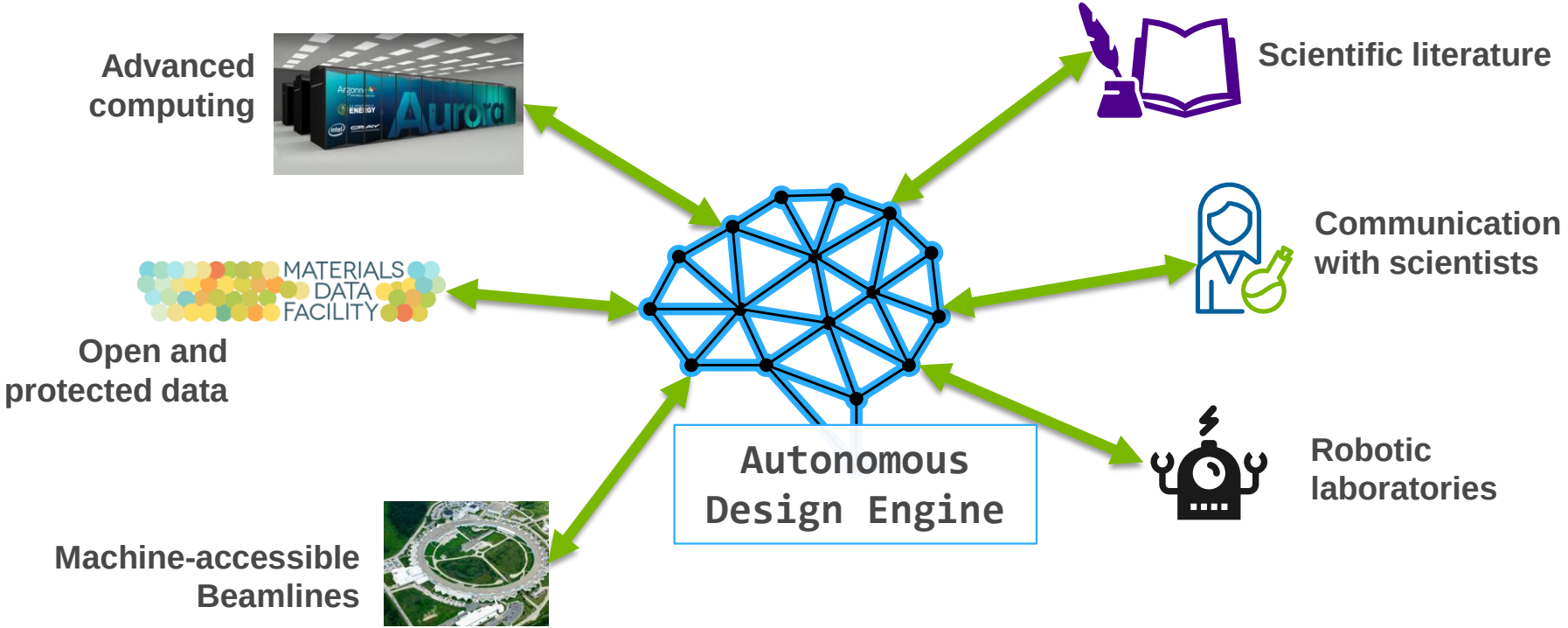


Perennial Question: What “human” tasks can we automate in science?

Source: Techxplore (2020)

PROGRAMMING AN "AI STAFF SCIENTIST"

What do I need to automate to perform materials design?



Achieving this vision will require innovation in computing fabric and AI technologies

WE DON’T HAVE AN “ARTIFICIAL SCIENTIST” YET

So what am I going to talk about then?

Two objectives for my talk today:

1. Deep-dive on a specific project where we use AI in materials design
   Go through the nuances of what such a project looks like

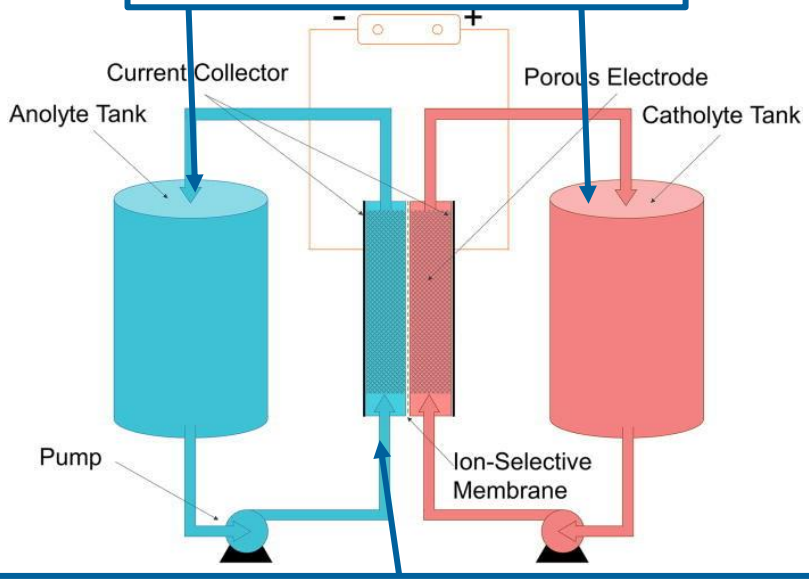
2. Discuss emerging technologies for “AI for Science” at Argonne
   We’re working hard to make AI in science a reality for all

FIRST: WHAT HAVE WE DONE?

OUR MAIN FOCUS: REDOX FLOW BATTERIES

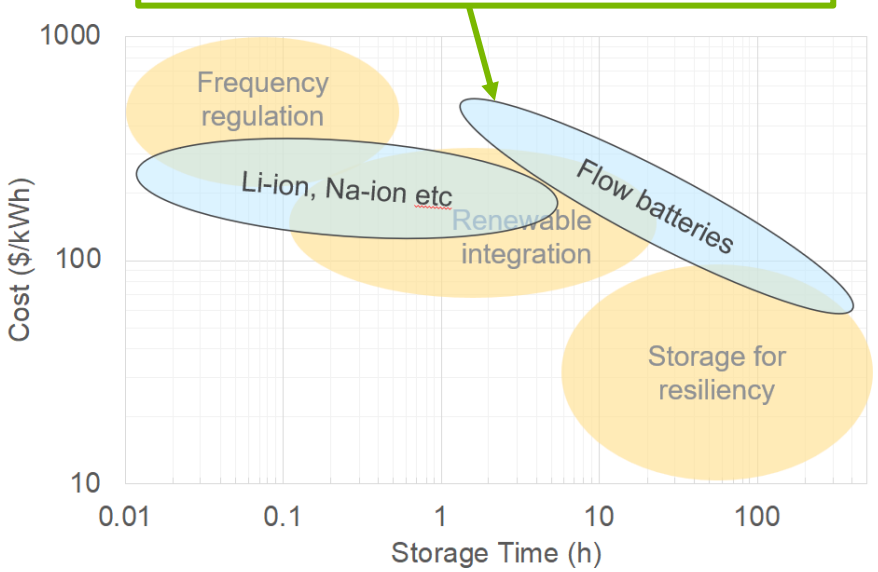
Part of a sustainable energy future

Energy stored in molecules that hold electric charge: “gas tank”



Store/release energy at the current collector: “engine”

Cheaply upgrade storage capacity – Just buy a bigger gas tank and more gas



Key problem: What molecules do I use to hold electric charge? (“fuel”)

Figures: (left) Wikipedia, (right) V. Srinivasan (Argonne)

CAN WE SOLVE IT WITH A BIG COMPUTER?

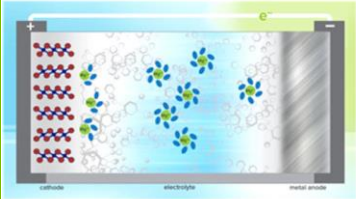
This probably sounds familiar to you, if you attend TMS

Numerous, constantly-changing battery chemistry

Insanely-intractable search space

Multivalent Intercalation

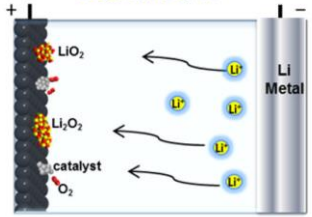
Replace monovalent Li+ with di- or tri-valent ions: Mg^{2+}, Al^{3+}, . . .
Double or triple capacity stored and released



- Electro- and chemical stability
- Dissolves salt
- Desolvates cation readily

Chemical Transformation

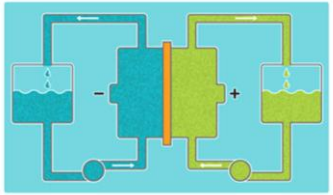
Replace intercalation with high energy chemical reaction:
Li-S, Li-O, Na-S, . . .



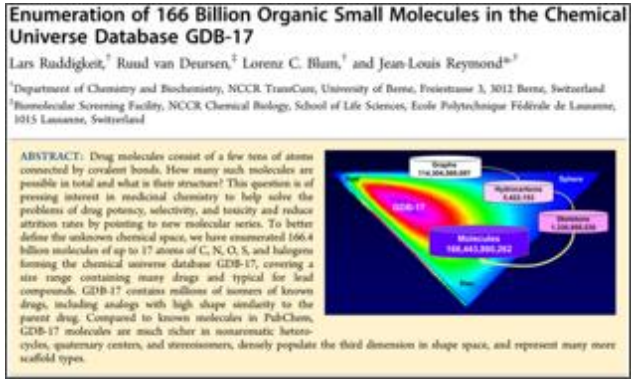
- Electro- and chemical stability
- Dissolves salt

Redox Flow

Replace solid electrodes with liquid solutions or suspensions:
lower cost, higher capacity, greater flexibility



- Electro-stability
- Dissolves salt and redox molecules
- Redox species with wide redox windows



Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17

Lars Ruddigkeit,[†] Ruud van Deursen,[‡] Lorenz C. Blum,[†] and Jean-Louis Reymond*[,†]

[†]Department of Chemistry and Biochemistry, NCCR TransCure, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland
[‡]Biomolecular Screening Facility, NCCR Chemical Biology, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

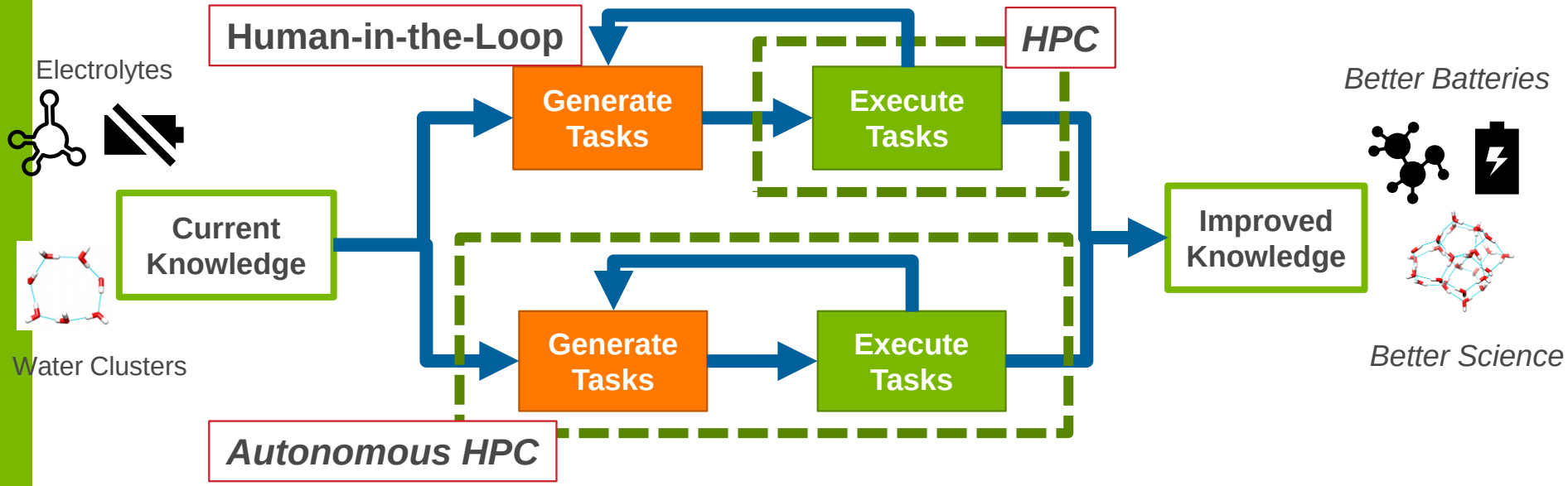
ABSTRACT: Drug molecules consist of a few tens of atoms connected by covalent bonds. How many such molecules are possible in total and what is their structure? This question is of pressing interest in medicinal chemistry to help solve the problems of drug potency, selectivity, and toxicity and reduce attrition rates by pointing to new molecular series. To better define the unknown chemical space, we have enumerated 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens forming the chemical universe database GDB-17, covering a size range containing many drugs and typical for lead compounds. GDB-17 contains millions of isomers of known drugs, including analogs with high shape similarity to the parent drug. Compared to known molecules in PubChem, GDB-17 molecules are much richer in nonaromatic heterocycles, quaternary centers, and stereoisomers, densely populate the third dimension in shape space, and represent many more scaffold types.

Many design requirements....

How can we use a "worlds-first" Exascale computer to design redox molecules *quickly*?

BIGGER COMPUTERS GIVE US *DIFFERENT* PROBLEMS

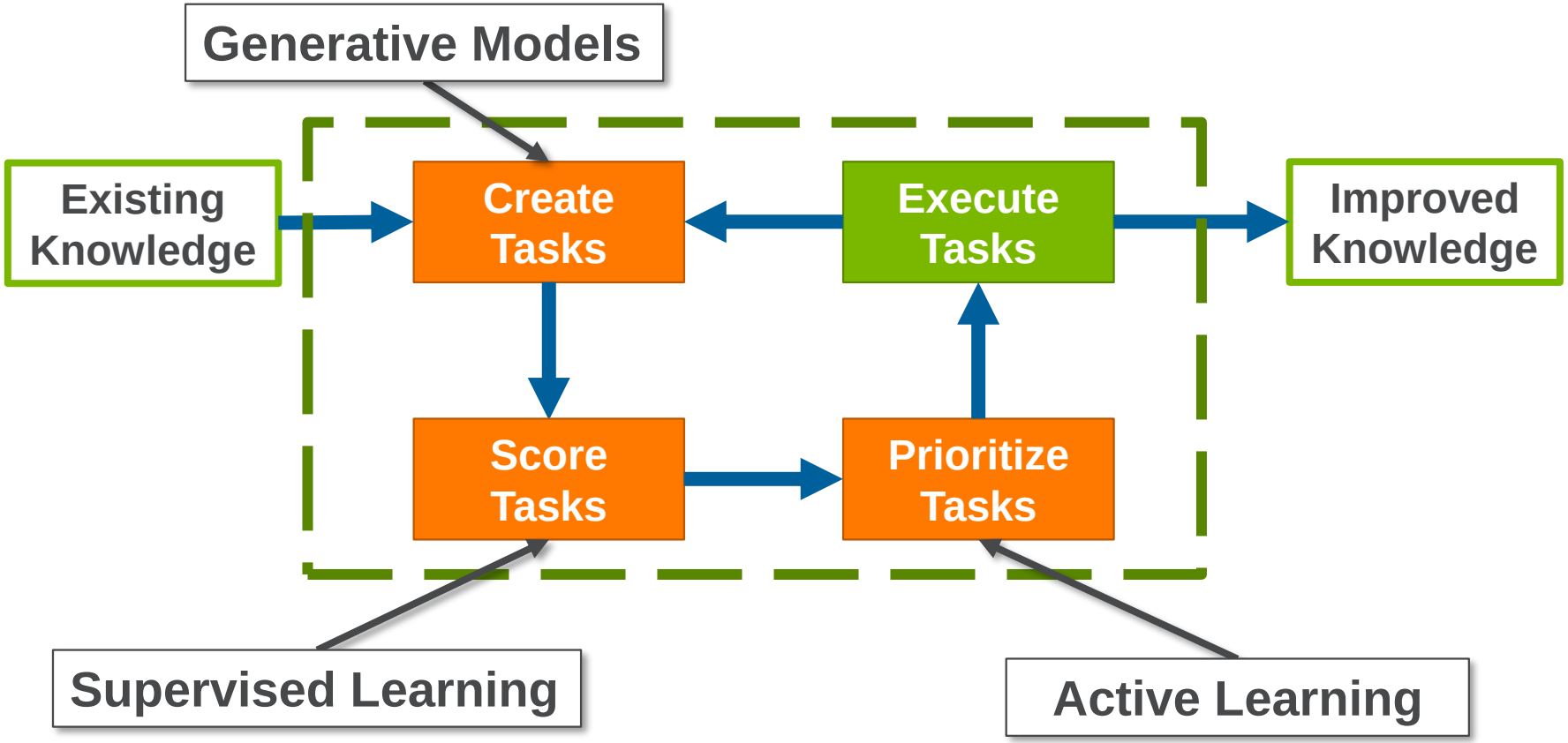
The New Problem: Humans are **slow and not getting any faster**



Our goal: **Replace computational scientist with algorithms**

WHAT DO WE NEED TO MAKE THIS HAPPEN?

A recipe for me to not have to go to work



PUT IN SPECIFIC TERMS, WE DEVELOPED…

1. Database of electrolyte properties

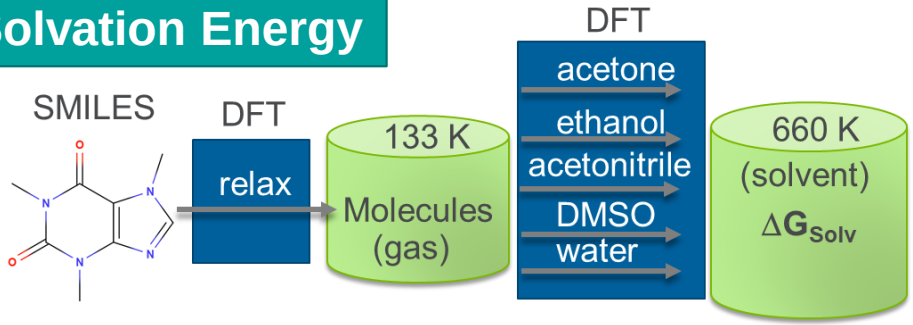
2. Suite of predictive models for electrolyte properties

3. **Adaptive High-Throughput Screening Application** ← The newest part!

STEP 1: GATHER (AND PUBLISH) DATA

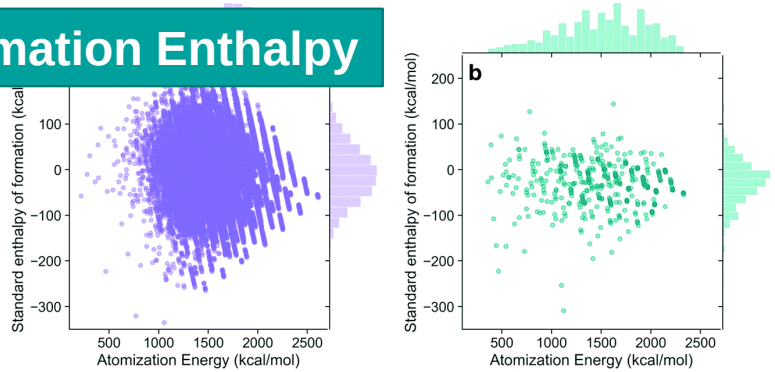
Key properties: Solvation, ionization potential, thermochemistry

Solvation Energy



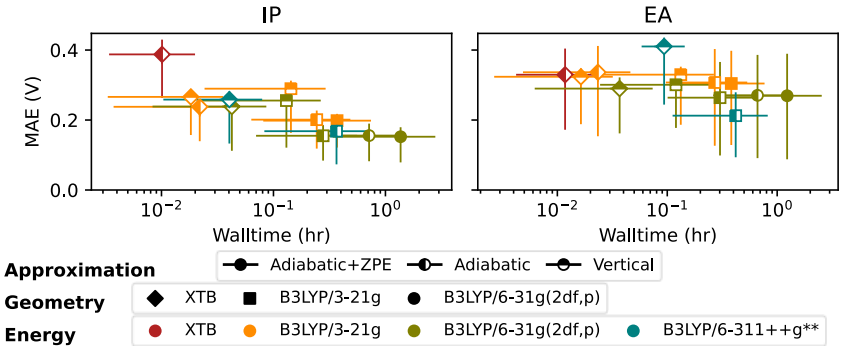
Ref: Ward et al. JCP:A (2021)

Formation Enthalpy



Ref: Narayanan et al. Chem Sci (2019)

Redox Potentials

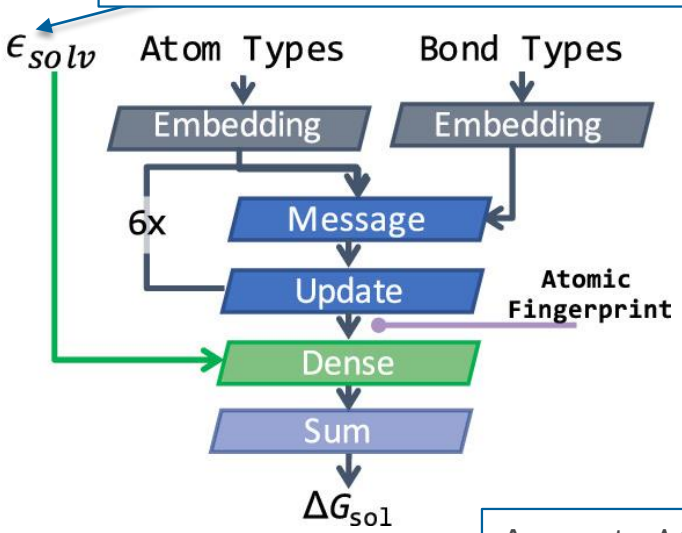


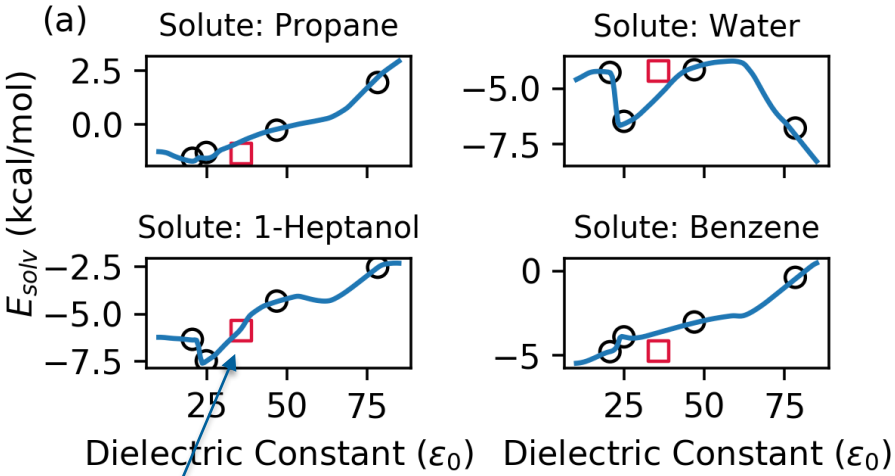
Ref: *in preparation*

STEP 2: CREATE SUITE OF ML MODELS

Be able to predict properties with minimal effort (~1 ms)

Use properties of solute *and solvent*



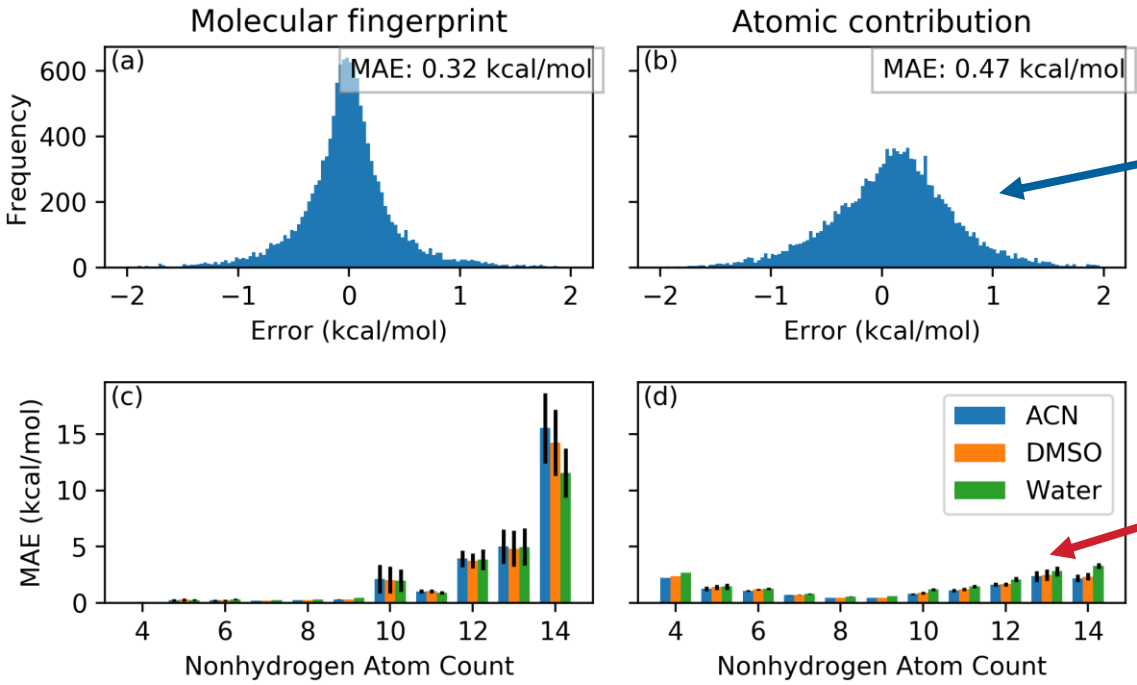


Accurate ΔG_{solv} in other solvents

Ref: Ward et al., JPCA (2021)

VALIDATE AS IF YOU MEAN TO USE IT

We want to use this model on large molecules



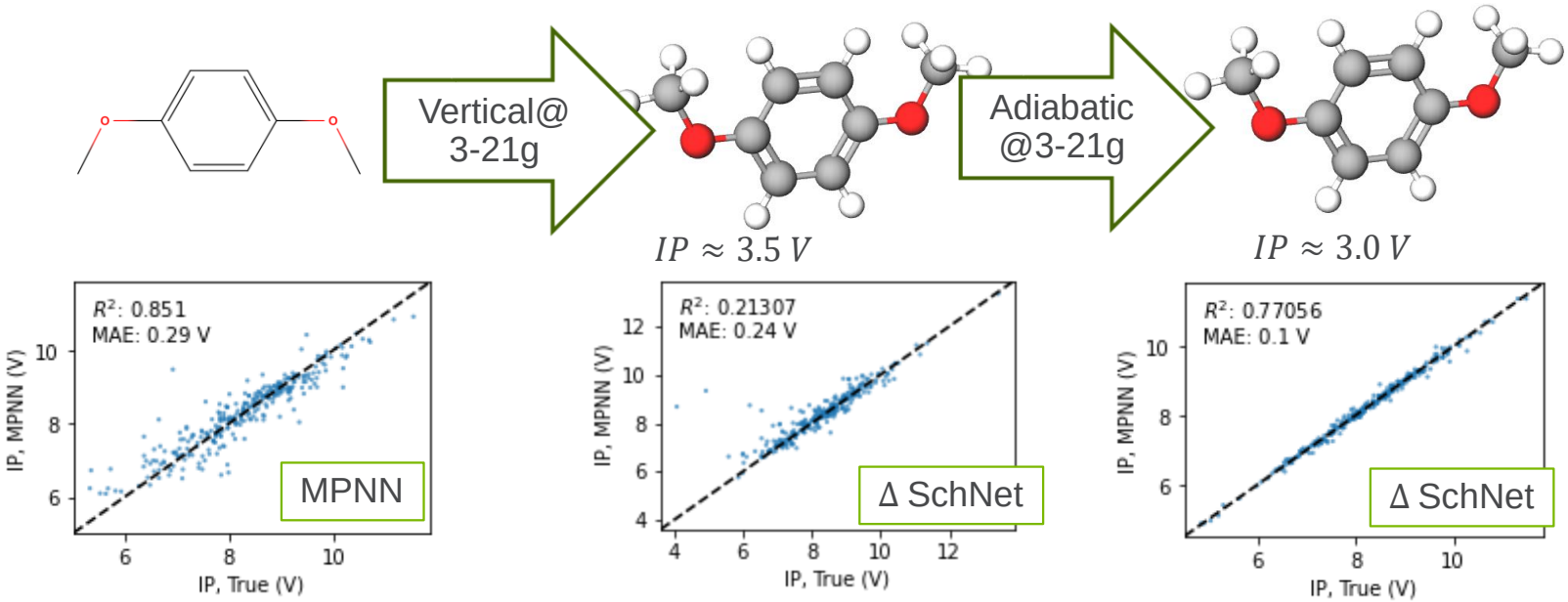
Trade accuracy on original dataset set

for accuracy on search space

Ref: Ward et al., JPCA (2021)

TOGETHER: MULTI-FIDELITY MODELING

Our message-passing networks are one piece of the puzzle

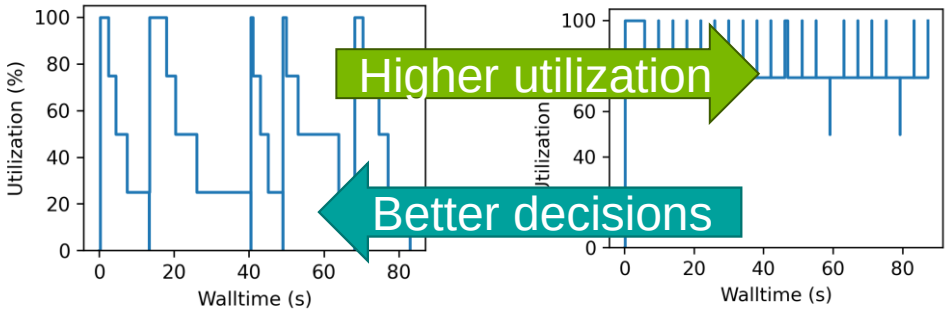


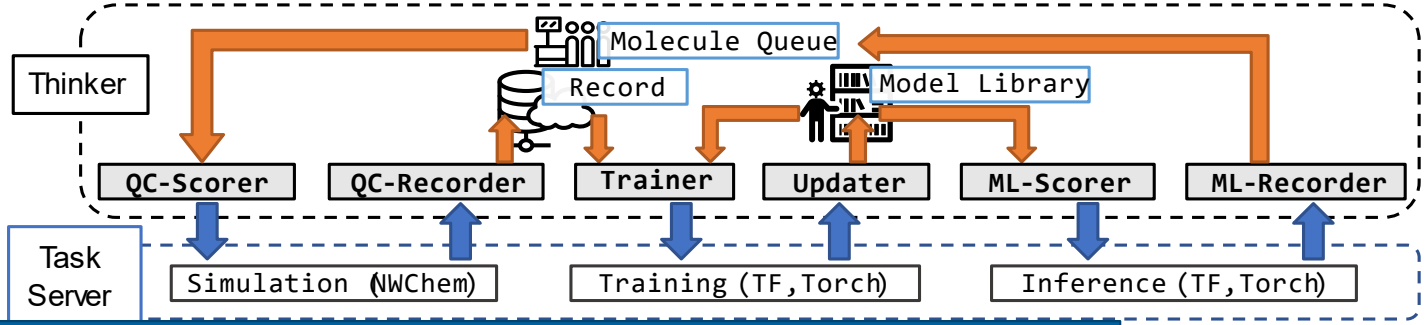
Gradually better estimates of IP as we do more computations

INGREDIENT 3: MIX SIMULATION AND AI

Combine until humans precipitate out of solution

Key issue: Autonomous experiment requires intelligent policies





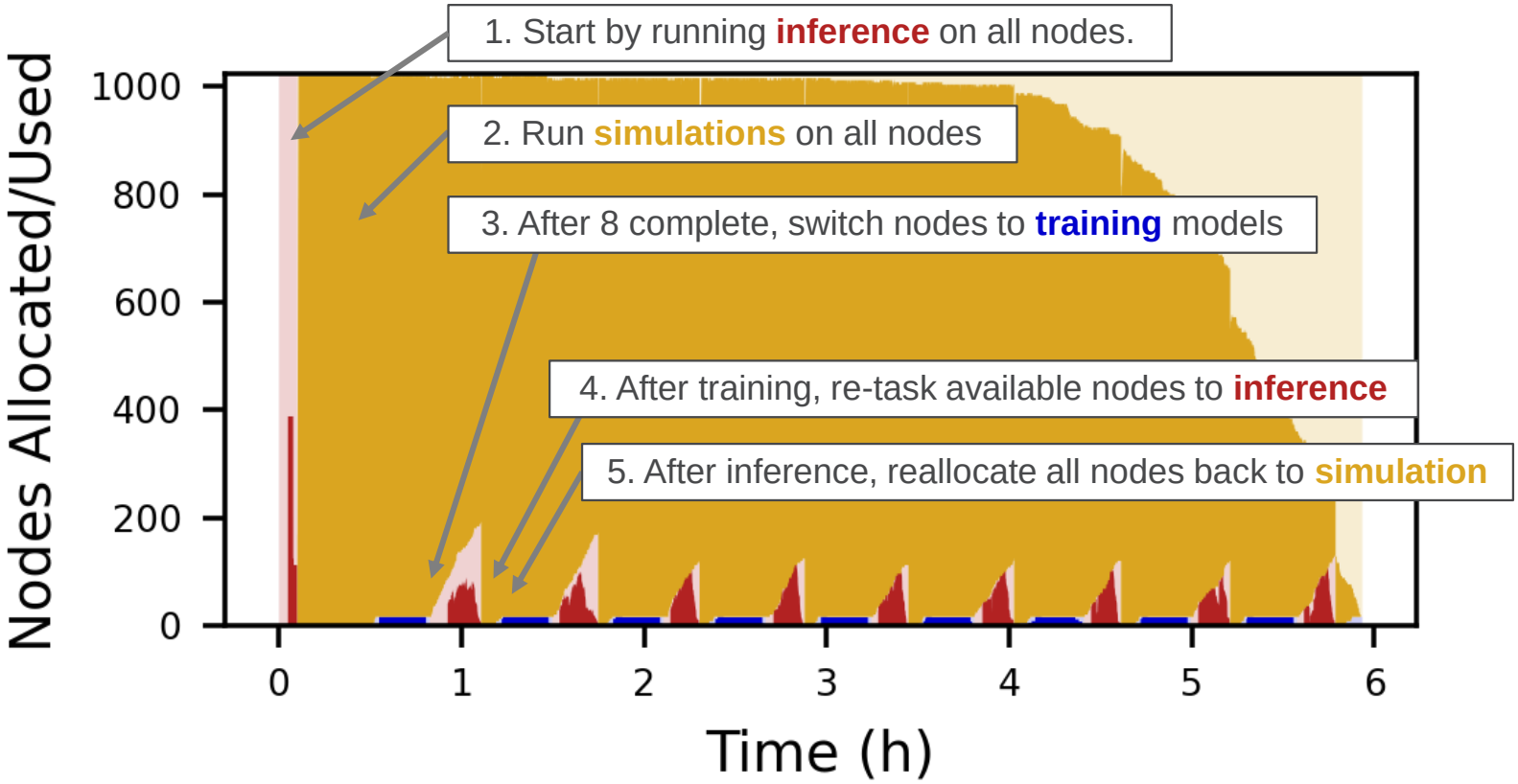
Our solution: Encode scientific process as intelligent "agents," simple tasks

Ref: Ward et al. ML4HPC @ SC21 (2021)

I gave a whole talk on the software, Colmena, at SC21 (see YouTube)

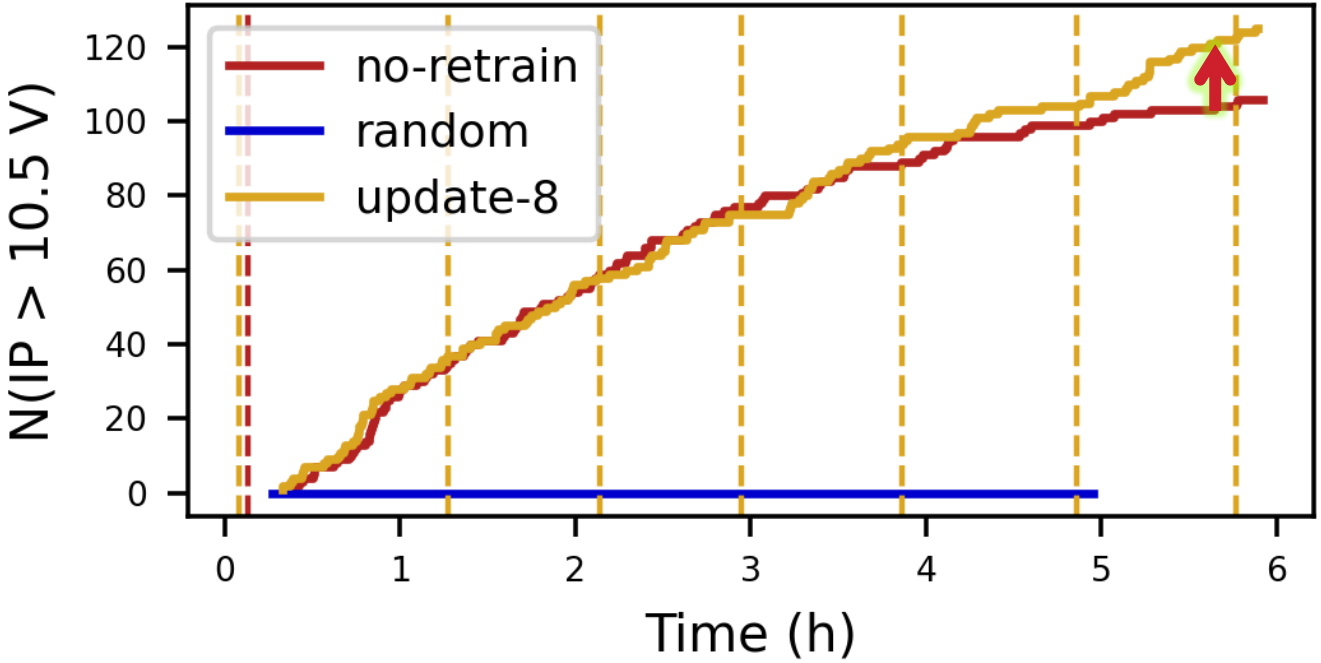
FOR GOOD MEASURE: RUN AT SCALE

1024 nodes ~ 1 MW electricity



INTELLIGENT POLICIES YIELD BETTER SCIENCE

without any human intervention



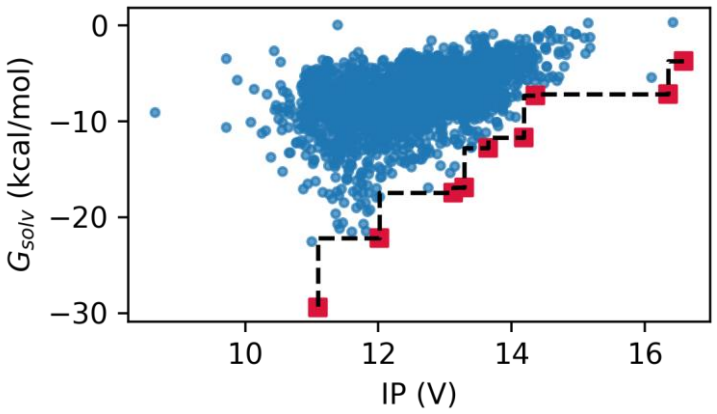
Found 10% more high-performing molecules with same allocation size

OUR NEXT STEPS: MULTI-OBJECTIVE, MULTI-FIDELITY

Gradually encode more knowledge

RUNNING MULTIPLE TASKS…

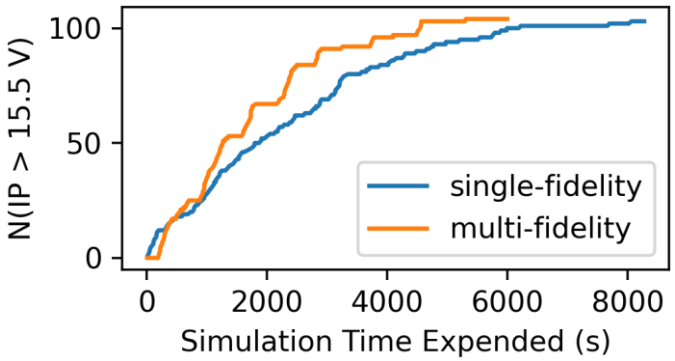
Very necessary for materials



Great reading: Agarwal et al. Chem Mat. (2021)

AT DIFFERENT LEVELS OF ACCURACY

Another tool for acceleration



Great reading: Woo et al. ArXiv: 2019.11683

TAKE-HOME POINTS

What is it I was hoping to show off?

Target problem: Redox flow battery design

How did we approach it? Autonomous HPC

1. Generated datasets for each problem
2. Created a suite of models to replace human inference
3. Encoded experimental planning/actions into autonomous agents

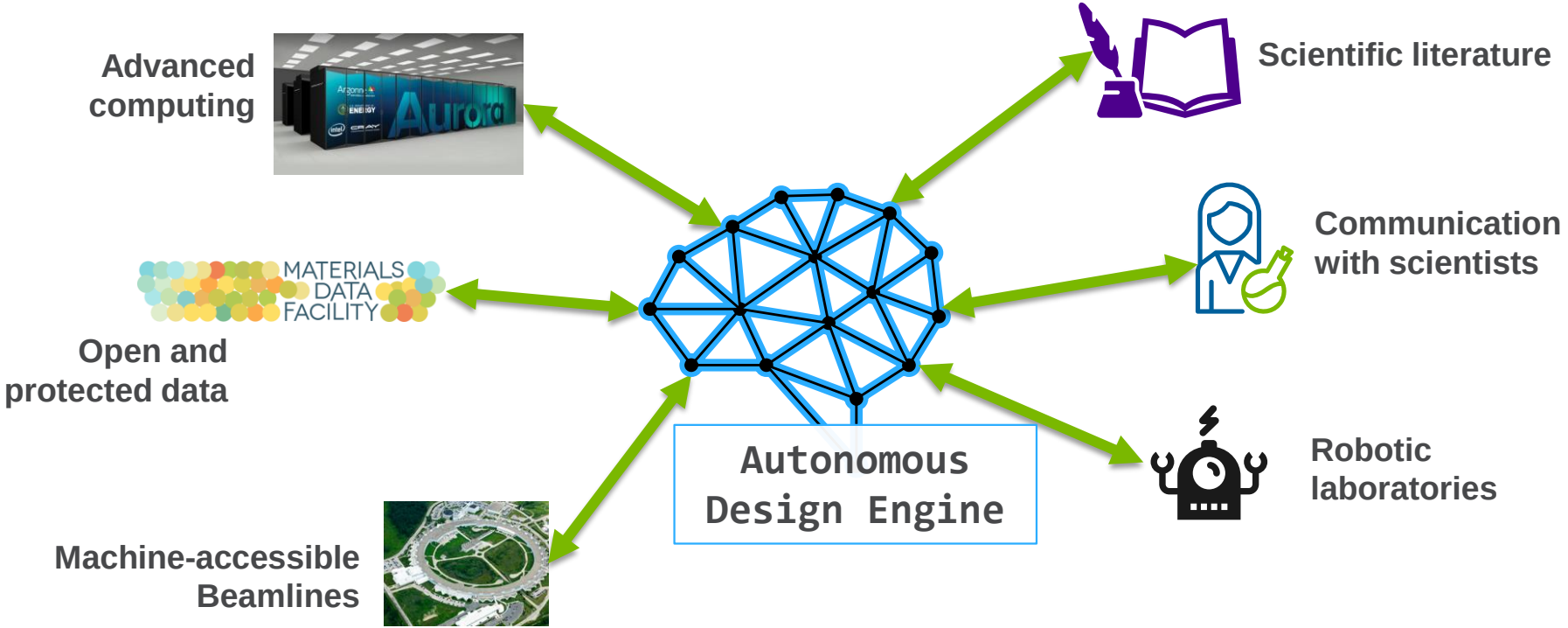
What did we learn?

- Validation on search space **is critical**
- Can achieve ~10% boost to performance with on-line learning

SECOND: WHERE ARE WE GOING?

PROGRAMMING AN "AI STAFF SCIENTIST"

Linking Argonne's resources with computing and AI



What is going on in these areas!?

Simple access to data / models

FOUNDRY
DATA, MODELS, SCIENCE

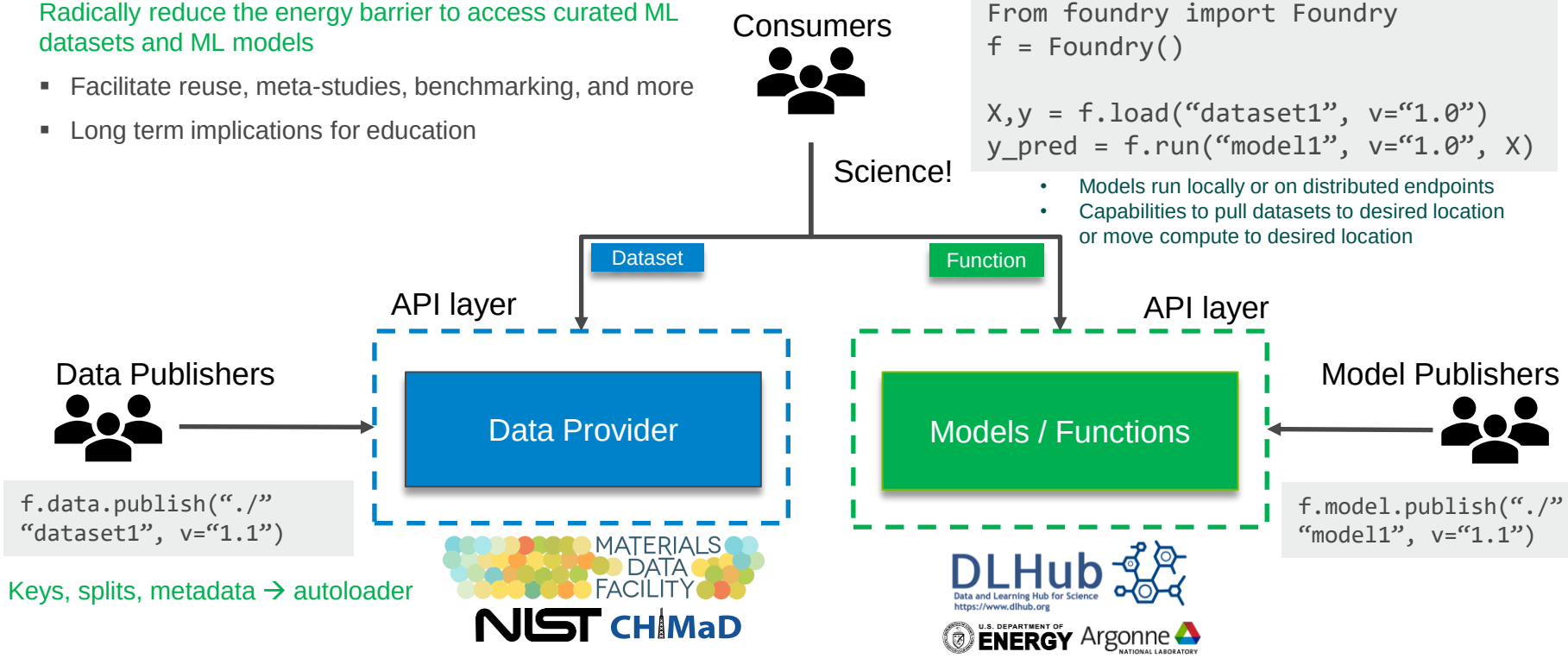
Radically reduce the energy barrier to access curated ML datasets and ML models

- Facilitate reuse, meta-studies, benchmarking, and more
- Long term implications for education

```
From foundry import Foundry
f = Foundry()

X,y = f.load("dataset1", v="1.0")
y_pred = f.run("model1", v="1.0", X)
```

- Models run locally or on distributed endpoints
- Capabilities to pull datasets to desired location or move compute to desired location



(Dane Morgan, Paul Voyles, Michael Ferris, Aristana Scourtas, KJ Schmidt, Marcus Schwarting, Ben Blaiszik)

Molecular design datasets are available

a Load Dataset

```
from Foundry import Foundry

f = Foundry()
f.load("10.18126/jos5-wj65", globus=globus)
```

b Understand Dataset Contents

G4MP2 Estimates of Solvation Energy in Multiple Solvents

• g4mp2_energy	target	Ha	G4MP2 Internal energy at 298.15K
• g4mp2_enthalpy	target	Ha	G4MP2 Enthalpy at 298.15K
• g4mp2_free	target	Ha	G4MP2 Free eergy at 0K
• g4mp2_atom	target	Ha	G4MP2 atomization energy at 0K
• sol_acetone	target	kcal/mol	Solvation energy, acetone
• sol_acn	target	kcal/mol	Solvation energy, acetonitrile
• sol_dmso	target	kcal/mol	Solvation energy, dimethyl sulfoxide
• sol_ethanol	target	kcal/mol	Solvation energy, ethanol
• sol_water	target	kcal/mol	Solvation energy, water

c Use Data

```
res = f.load_data()
X,y = res['train']
```

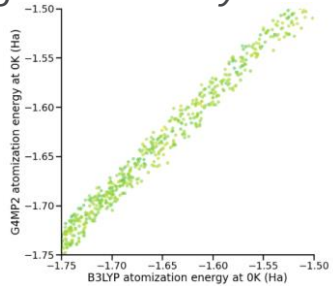
X.head()

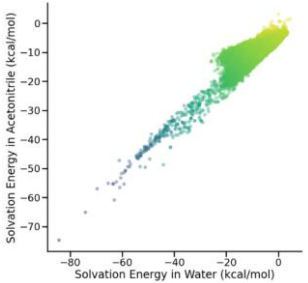
	smiles_0	smiles_1	inchi_0	inchi_1	xyz	atomic_charges	A	B
0	C	C	InChi=1S/CH4/h1H4	InChi=1S/CH4/h1H4/n	5\nH4 C\n-0.012698 1.085804 0.008001\nH 0.0...	[-0.535689, 0.133920999999999998, 0.13392199999...	157.71180	157.709970
1	N	N	InChi=1S/H3N/h1H3	InChi=1S/H3N/h1H3/n	4\nH3 N\n-0.040426 1.024108 0.062564\nH 0.0...	[-0.707143, 0.235719999999999998, 0.2357199999...	293.60975	293.541110
2	O	O	InChi=1S/H2O/h1H2	InChi=1S/H2O/h1H2/n	3\nH2 O\n-0.034360 0.977540 0.007602\nH 0.0...	[-0.589706, 0.294853, 0.294853]	799.58812	437.903860
3	C#C	C#C	InChi=1S/C2H2/c1-	InChi=1S/C2H2/c1-	4\nH2 C2\n-0.599539 0.000000 0.000000	[-0.207018999999999998, -0.207018999999999998, 0...	0.00000	35.610036
4	[H]C([H])	[H]C([H])	InChi=1S/C2H4/c1-	InChi=1S/C2H4/c1-	6\nH4 C2\n-0.000000 0.000000 0.000000\nH 0.0...	[-0.049656, -0.188473, 0.238127999999999998]	0.00000	44.593883

y.head()

	g4mp2_hf298	g4mp2_0k	g4mp2_energy	g4mp2_enthalpy	g4mp2_free	g4mp2_atom	sol_acetone	sol_acn	sol_dmso	sol_
0	-17.642516	-40.427662	-40.424791	-40.423846	-40.447329	-0.625083	0.3624	0.4569	1.2154	
1	-10.280320	-56.478971	-56.476107	-56.475163	-56.498045	-0.439864	-3.0186	-3.0297	-2.5934	
2	-57.552864	-76.355852	-76.353017	-76.352073	-76.374154	-0.349181	-4.2803	-4.2132	-4.1604	
3	54.235405	-77.212309	-77.209392	-77.208448	-77.231319	-0.619715	-1.9940	-2.1527	-0.8402	
4	30.659525	-93.312546	-93.310021	-93.309077	-93.331907	-0.483424	-3.7116	-3.8149	-3.0400	

Database to allow training of predictive ML models to identify promising Li-ion battery electrolyte materials





Ward, Logan; Dandu, Naveen; Blaiszik, Ben; Narayanan, Badri; Assary, Rajeev S.; Redfern, Paul C.; Foster, Ian; Curtiss, Larry A.

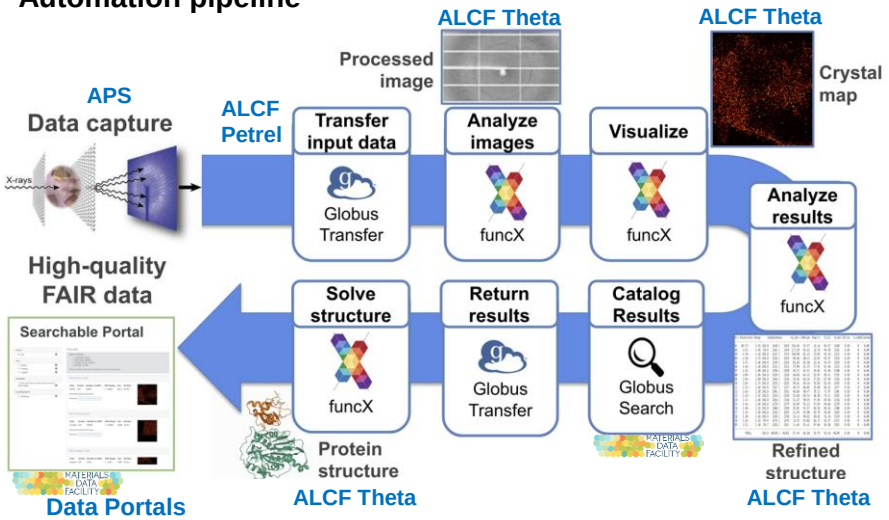
U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

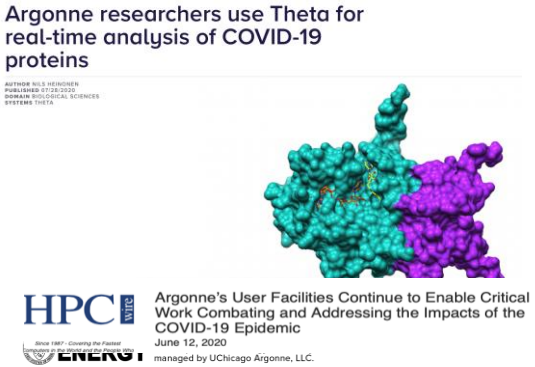
Solving Protein Structures 10-100x Faster

by linking together APS and ALCF with intelligent data services

- **Developed new automation pipeline** to collect data, analyze and visualize the data, solve protein structure and load results into a searchable portal for realtime feedback
- **Achieved over 10-100x speed up** in time to solution of protein structures at APS beamline
- **Leveraged unique DOE facilities** at Advanced Photon Source (SBC Sector 19) and ALCF (Theta/ ThetaGPU, Petrel, and Data Portals)

Automation pipeline





Deposited first results in open repositories

"These data services have taken the time to solve a structure from weeks to days and now to hours"

Darren Sherrell, SBC beamline scientist APS Sector 19

(Chard, Vescovi, Foster, Blaiszik, Sherrell, Joachimiak, et al.)

Argonne NATIONAL LABORATORY

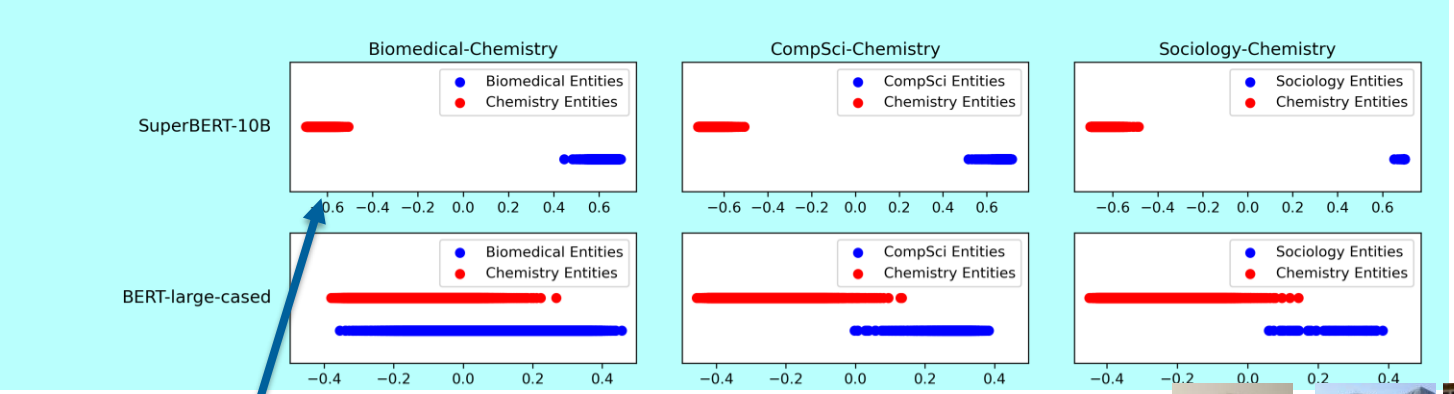
LARGE-SCALE LEARNING FROM LITERATURE

Trade computing time for human labeling effort, “ScholarBERT”

Key Concept: Semi-supervised/transfer learning

1. Train on easy-to-acquire data: “fill in the missing word”
2. Fine-tune the model on a task that requires effort: “identify polymers”

Simple approach: Train a huge model on many papers



Big benefit: Separate terms from different fields easily

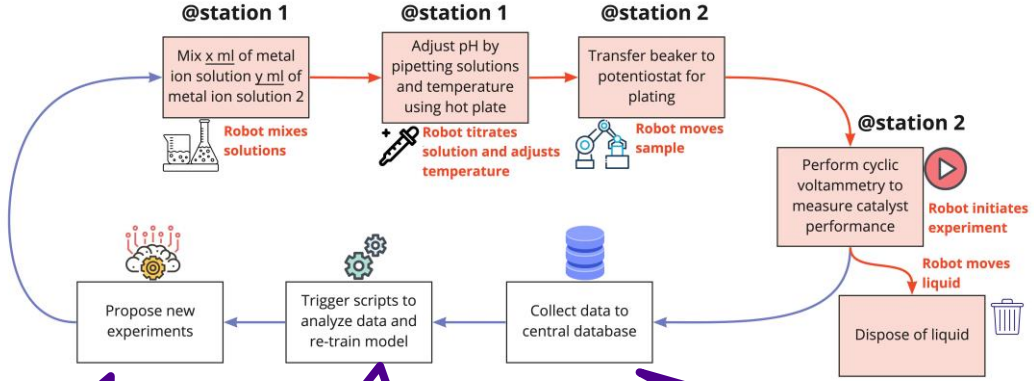


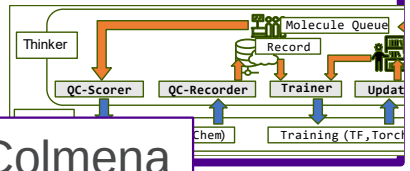
Ref: Hong et al. Fron. Mol. Bio. (2021); Hong et al. JOM (2021)

BUILDING AN “AUTONOMOUS USER FACILITY”

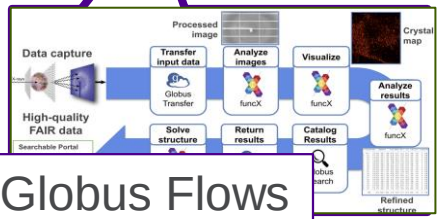
Our vision: Easily repurpose-able robots for any challenge

Breaking complex science into “stations” and “skills”





Colmena



Globus Flows

Argonne Discovery Cloud

That can be composed together with scientific computing and data infrastructure

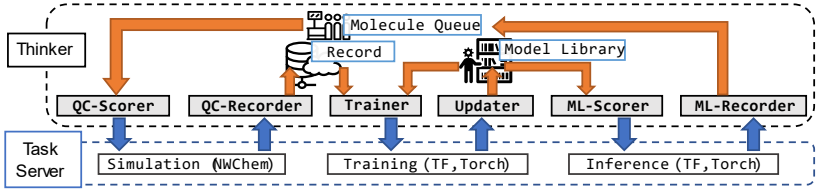
TAKE-HOME POINTS

TODAY'S SUCCESSES

Autonomous HPC for Material Design

Example Application: Flow Batteries

Approach: Encode "science" as series of simple actions



TOMORROW'S GOALS

Autonomous National Labs

What will we need:

1. Robust data infrastructure (MDF)
2. Analysis pipelines (Gladier)
3. Data from manuscripts (ScholarBERT)
4. Autonomous labs

EDUCATION IN “AI FOR MATERIALS”

Another big focus of mine

- “Applied AI for materials engineering” – Course at UChicago PME
  GitHub: https://github.com/wardlt/applied-ai-for-materials
  YouTube: Recordings of WI21 Lectures

- ALCF AI for Science tutorial series: alcf.anl.gov/alcf-ai-science-training-series
- MRS SP22 Tutorial on Battery Data Science
- AI Educators Slack: [sign up link] (thanks, Jason Hattrick-Simpers!)

THANK YOU TO TEAM!

More than I can fit on one slide

Argonne: ExaLearn – Using AI with HPC
Yadu Babuji, Ben Blaiszik, Ryan Chard, Kyle Chard, Ian Foster, Greg Pauloski, Ganesh Sivaraman, Rajeev Thakur

Argonne: JCESR – Molecular modeling for batteries
Rajeev Assary, Larry Curtiss, Naveen Dandu, Paul Redfern, Hieu A Doan

MolSSI – Workflows for quantum chemistry
Lori A. Burns, Daniel Smith, Matt Welborn, *many other open-source contributors*

PNNL: ExaLearn – Graph algorithms for learning
Sutanay Choudhury, Jenna Pope, Sotiris Xantheas

Argonne ALCF – AI, Data and Simulation on HPC
Murali Emani, Alvaro Vazquez-Mayagoitia, Venkat Vishnawath

BNL: ExaLearn – Optimal experimental design
Frank Alexander, Shantenu Jha, Kris Reyes, Li Tan, Byung-Jun Yoon, *and more*

Argonne SDL – Robots for materials and other science
Rajeev Assary, Anthony Averca, Ben Blaiszik, Tom Brettin, Ian Foster, Mark Hereld, Raf Vescovi, Jie Xu

UChicago/UIUC/UW-Madison – Data infrastructure and NLP
Aswathy Ajith, Ben Blaiszik, Kyle Chard, Ian Foster, Ben Galewsky, Zhi Hong, Ryan Jacobs, Dane Morgan, Greg Pauloski, KJ Schmidt, Marcus Schwarting, Aristana Scourtas

THANK YOU TO FUNDERS!

Highlight work from many projects

Molecular Design: JCESR and Exascale Computing Project

Data Infrastructure: NSF CSSI, NIST CHiMaD

Natural Language Processing: NIST CHiMaD

Autonomous Labs: Argonne LDRD