

# Autonomous Discovery of Separation Science

Ping Yang

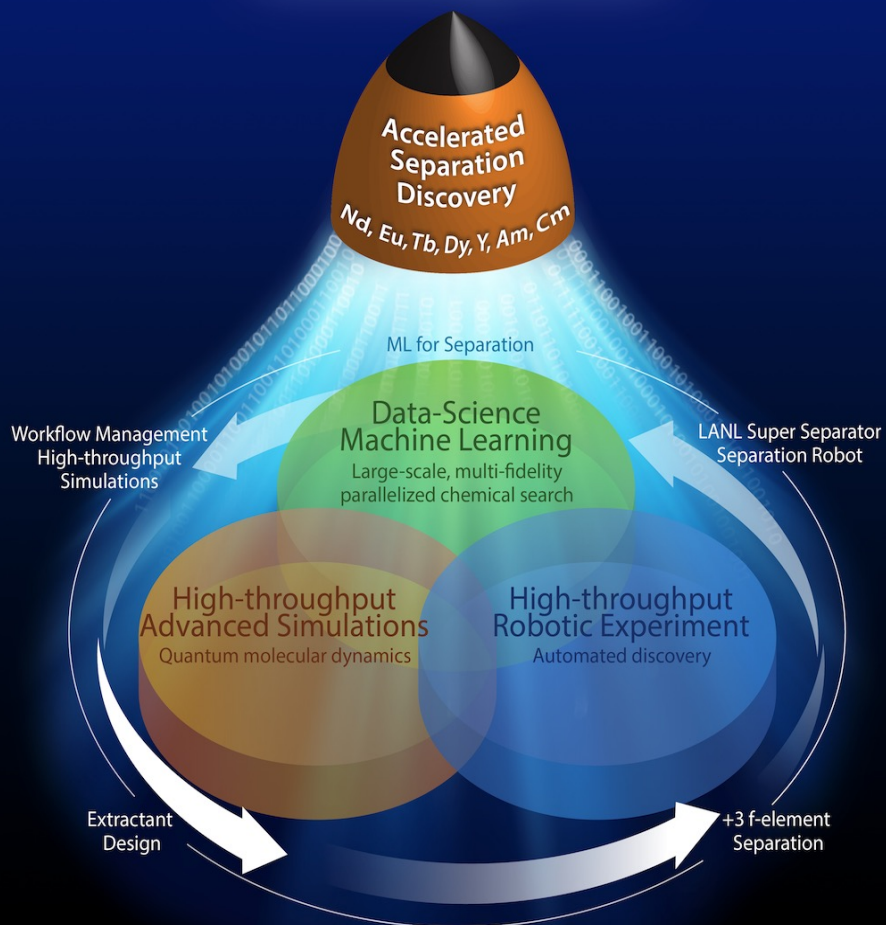
Los Alamos National Laboratory

Workshop III: Complex Scientific Workflows at Extreme Computational Scales

Part of the Long Program  
New Mathematics for Exascale

IPAM, UCLA  
May 1-5, 2023

Clean Energy Future



# Chemical Separation & Clean Energy

## Chemical Separations:

- Critical to almost every aspect of our daily lives from energy, the medications, to clean water
- Costs ~10 – 15% of total energy used in US

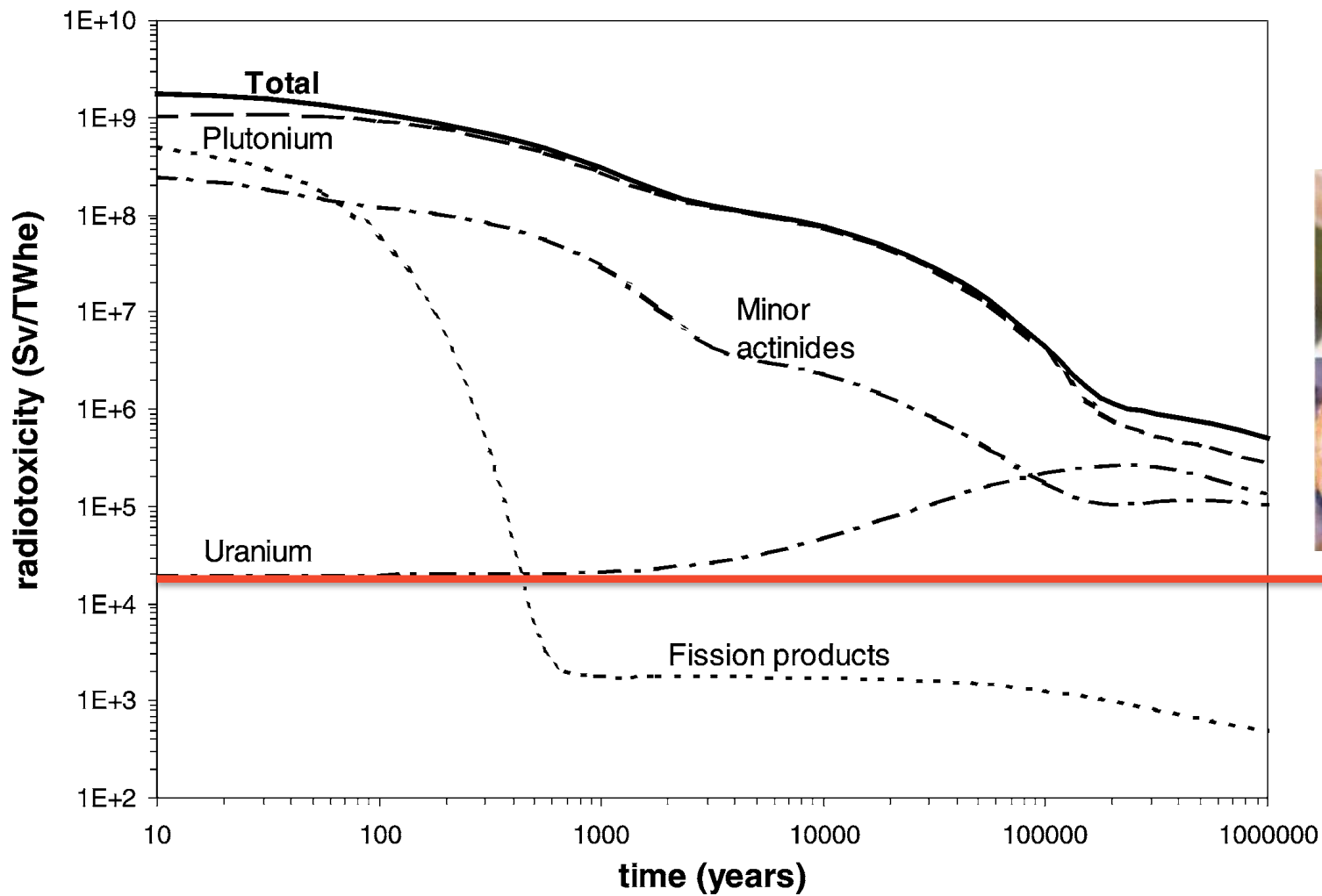


## Clean Energy and Nuclear Energy:

- Rare-Earth Elements (Nd, Ho, Dy, Eu, etc) needed for clean energy technology, such as wind turbines, EV motors, etc.
- CO<sub>2</sub> level → 420ppm, NE is a low-carbon footprint solution
- 20% of nation's electricity (55% of clean energy in US)
- 80,000+ metric tons of used nuclear fuel in US
- 2,000 metric tons increase each year
- ~\$\$B annually to deal with the waste
- More nuclear reactors under construction worldwide



# Radiotoxicity

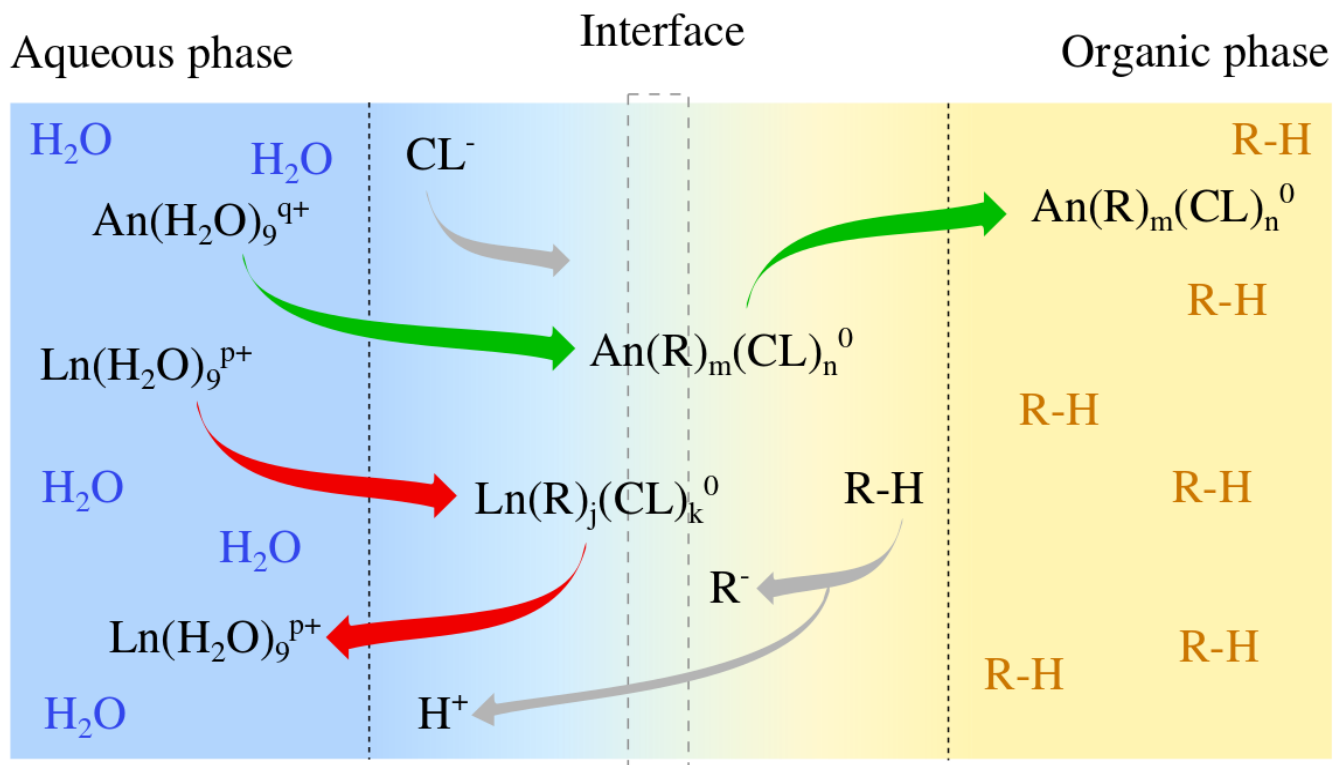


$An^{3+}$

$Ln^{3+}$

Madic, C. *et al*, C.R. Physique, 3, 797-811 (2002)

# An/Ln and Ln/Ln Solvent Extraction Process



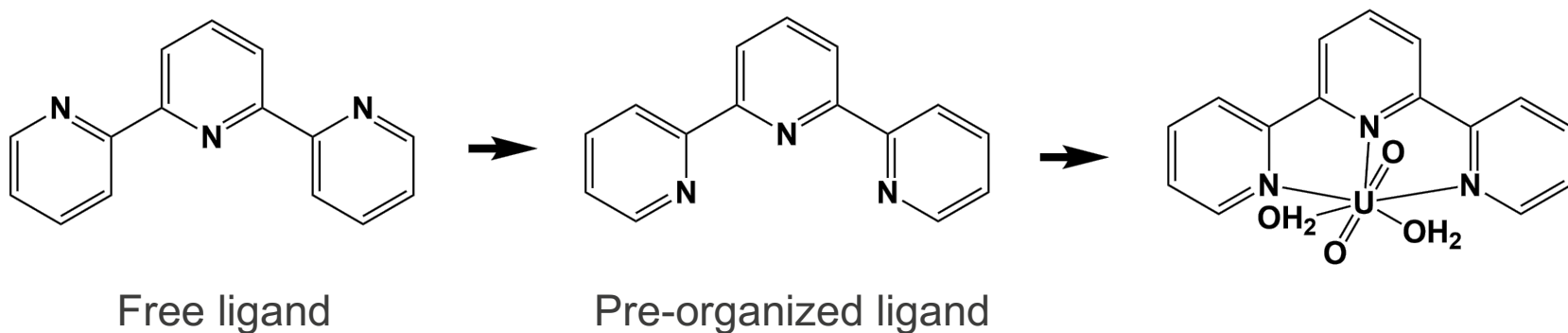
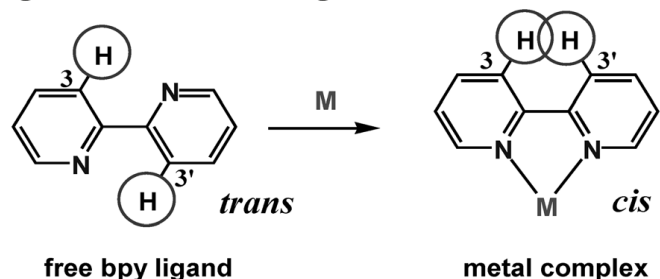
Challenges: chemical space is too vast to be manually explored by experiment or theory.

- Aqueous matrix
- Extractant solubility
- Binding kinetics
- Complexation stability
- Organic matrix
- Phase transfer catalysts
- Matrix effect (cooperative vs. blocked binding)
- Holdback agent
- Binding capacity
- Valence adjustment agent and rate
- Resin identity
- Bead size
- Solid-state support properties

# The Conventional Way of Discovering a New Extractant

Starting with the literature, observations, chemical intuition

Steric effects of nitrogen donor ligands

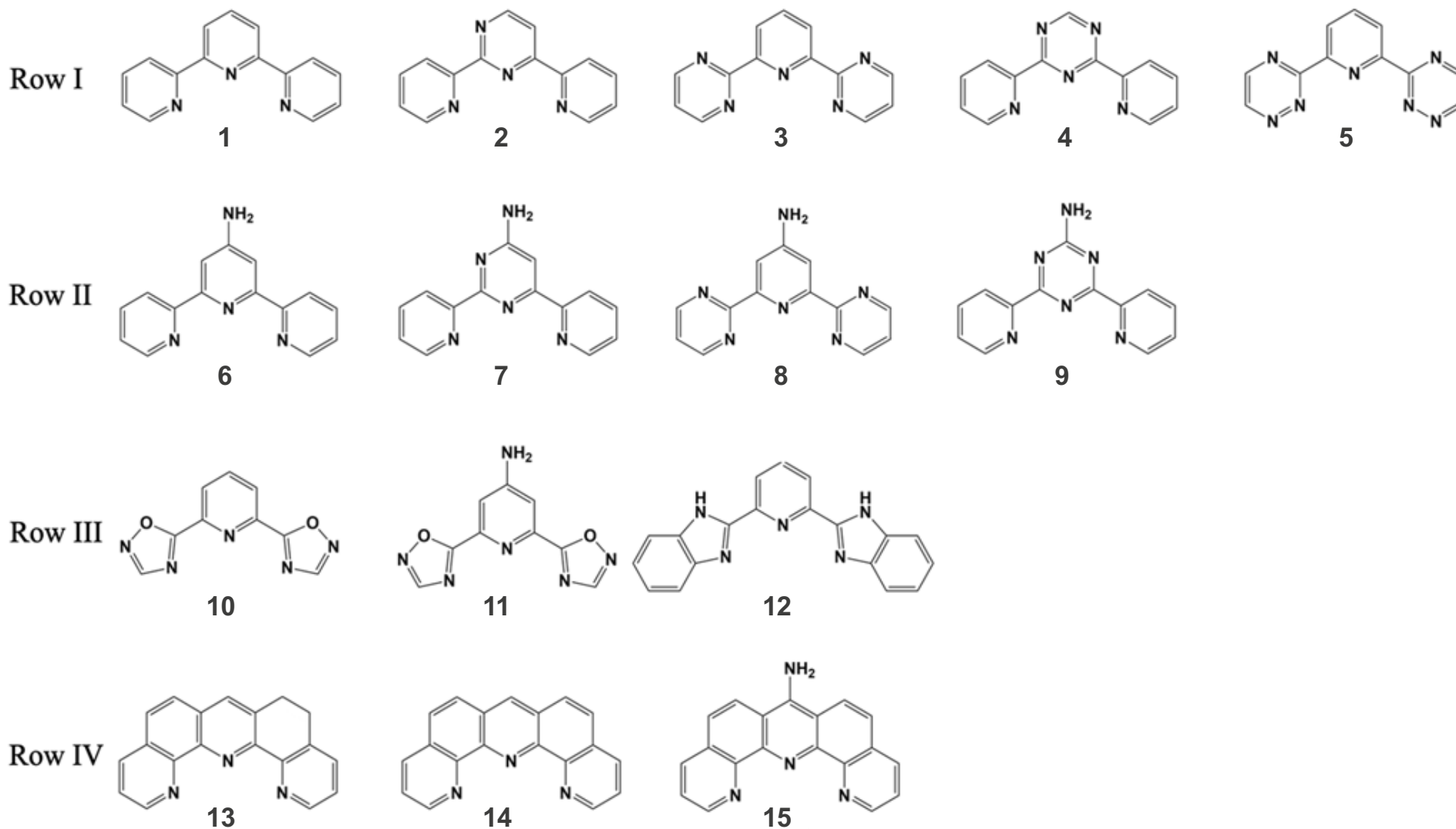


**Our hypotheses:**

- Preorganization  $\rightarrow$  reduce kinetic binding barrier  $\rightarrow$  improve efficiency
- Planarity  $\rightarrow$  reduce steric repulsion between H atoms  $\rightarrow$  improve efficiency
- Substitution Effects  $\rightarrow$  increase electron donating group  $\rightarrow$  improve efficiency

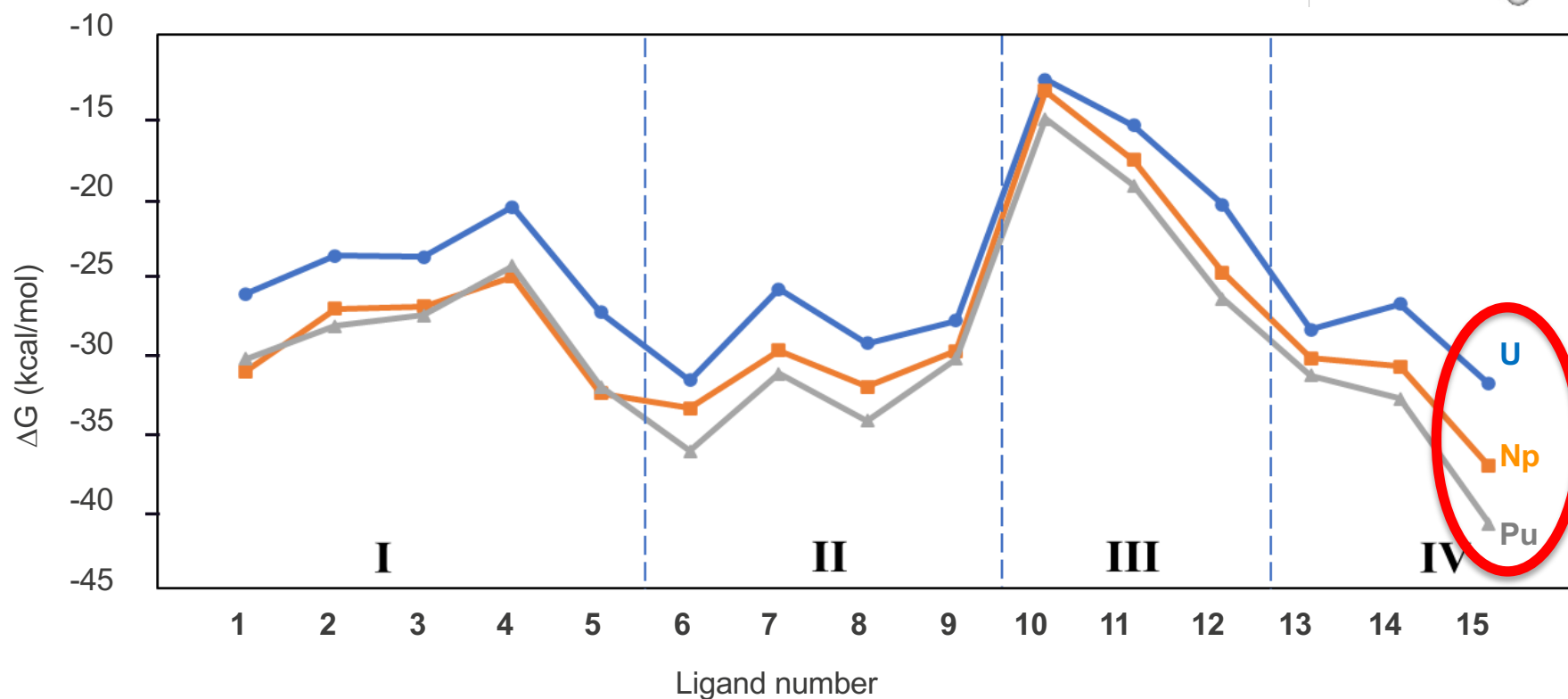
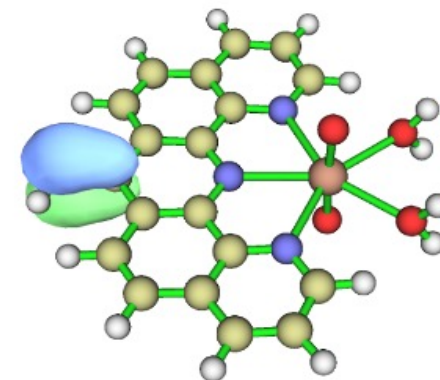
*R Hancock, Chem. Soc. Rev., 2013, 42, 1500-1524*

# Proposed Extractants Based on the Hypothesis



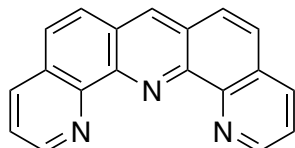
# Construct Investigation: Binding Energies of Extractants

$AnO_2^{2+}$ , An = U, Np, Pu  
as model systems for the screening process.



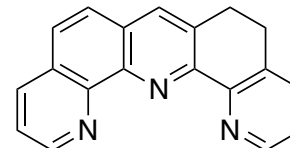
# Synthesis and Separation Experimental Validation

14

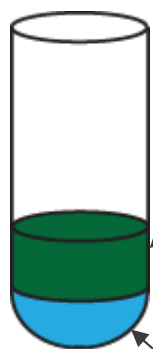


**Not soluble** in non-volatile organic solvents  
→ not great for liquid/liquid extraction

13, dpap

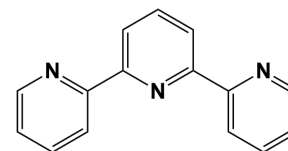


Precursor **is soluble** in *n*-octanol, still  
has the pre-organized N donors

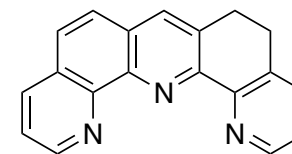


*n*-octanol + 20 mM "L" +  
1 M 2-bromohexanoic acid

0.01 M HNO<sub>3</sub> +  
<sup>241</sup>Am and <sup>155</sup>Eu tracers



"L" = terpy



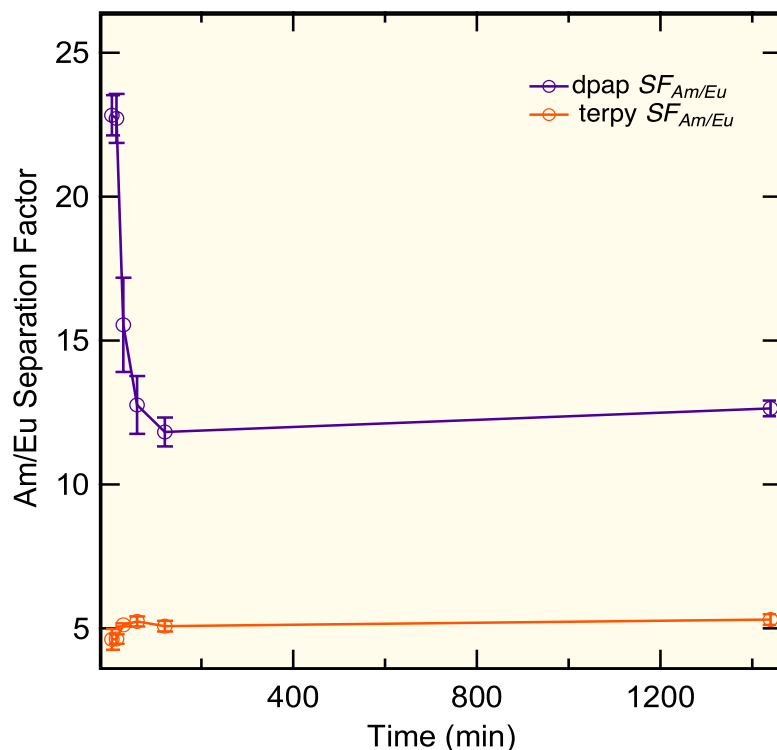
"L" = dpap

Close integration between theory and experiment  
is critical as cross-validation is highly needed!

X Zhang, SL Adelman, SA Kozimor, BW Stein, ER Batista, P Yang, et al **2022**, 61, 11556-11570



# Distribution Coefficients and Separation Factors

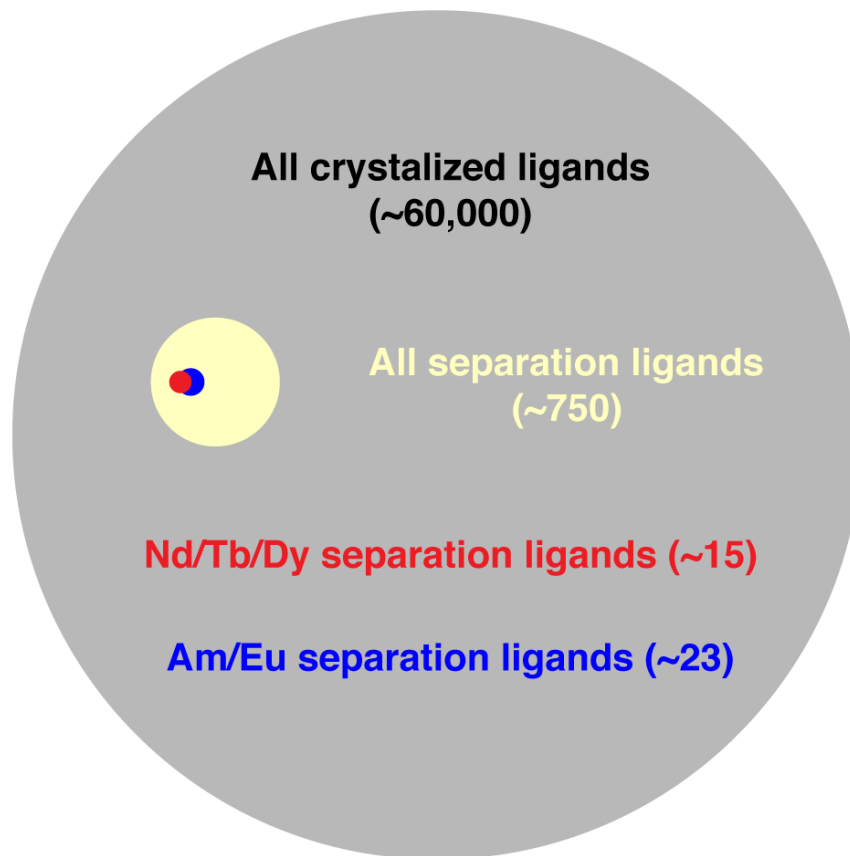
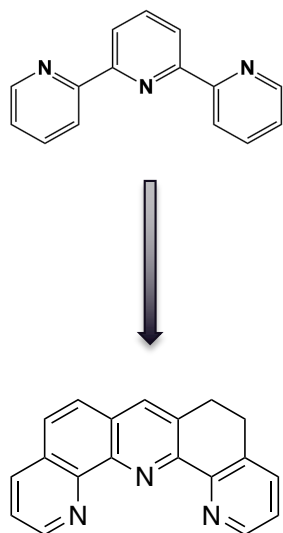


- At early times (<15 min), dpap is **~5x** better than terpy at separating Am from Eu
- Once equilibration has been reached, dpap is only **~2x** better than terpy.
- Larger separation factors with dpap are due to rapid and substantial transport of Am into organic phase
- This kind of fundamental knowledge will be used to design ligand features for the ML model

Time (min)	terpy			dpap			dpap/terpy		
	Am D	Eu D	Am/Eu SF	Am D	Eu D	Am/Eu SF	Am D	Eu D	<b>SF</b>
-	Am D	Eu D	Am/Eu SF	Am D	Eu D	Am/Eu SF	Am D	Eu D	<b>SF</b>
5	0.20	0.04	4.62	12.77	0.56	23	64	13	<b>5</b>
15	0.25	0.05	4.63	25.59	1.13	23	101	21	<b>5</b>
30	0.37	0.07	5.12	16.90	1.09	16	45	15	<b>3</b>
60	0.39	0.07	5.24	31.73	2.49	13	82	34	<b>2</b>
120	0.38	0.07	5.08	30.47	3.01	10	81	40	<b>2</b>
1440	0.33	0.06	5.31	39.55	3.13	13	122	51	<b>2</b>

# Local Exploration vs. the Vast Chemical Space

Full separation ligand design space  
( $\gg 10^{10}$ )

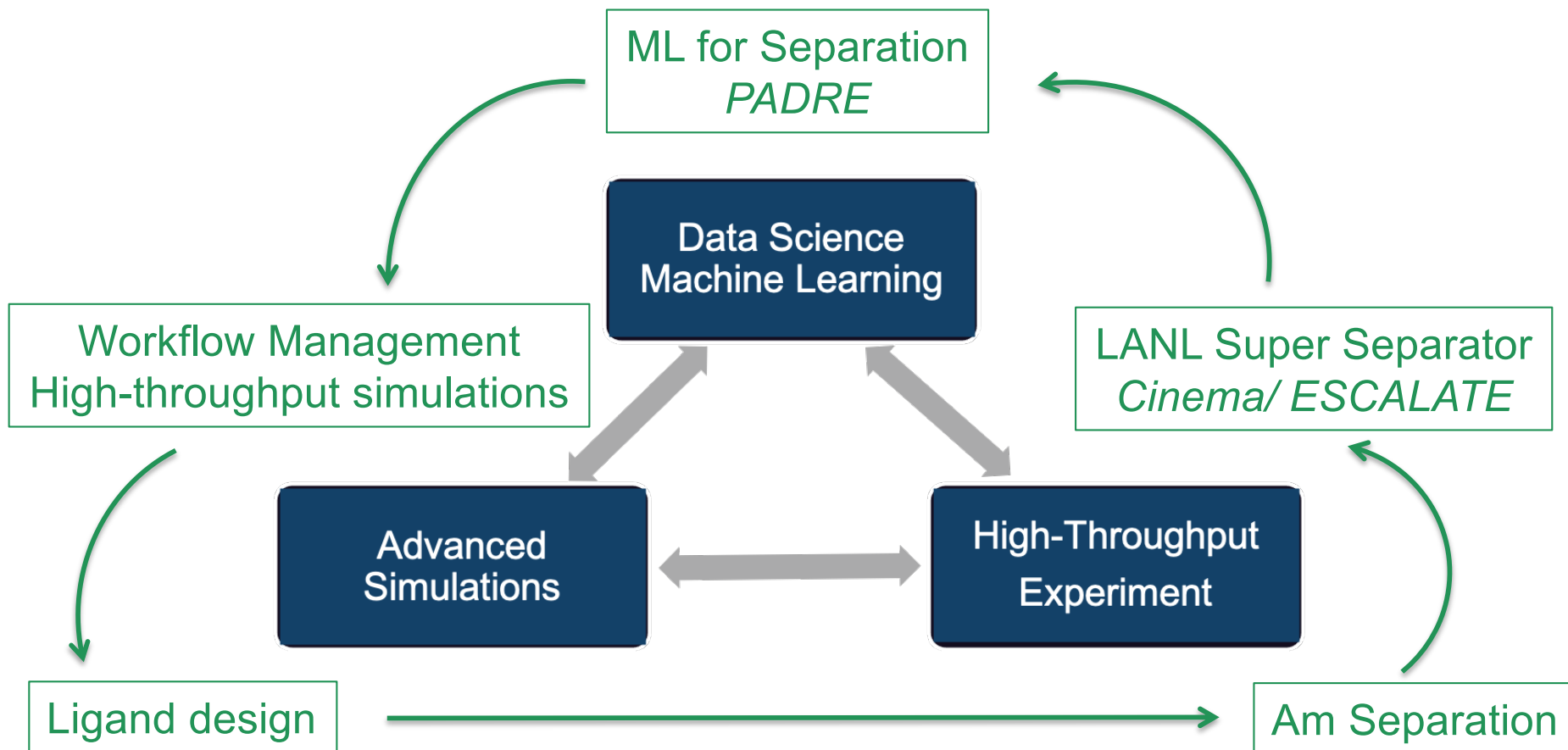


Missed chemistry

Nd/Tb/Dy 98% !

Am/Eu 97%!

# Approach: SeparationML



*Follow the data!*

**Complex workflow management is a necessity.**

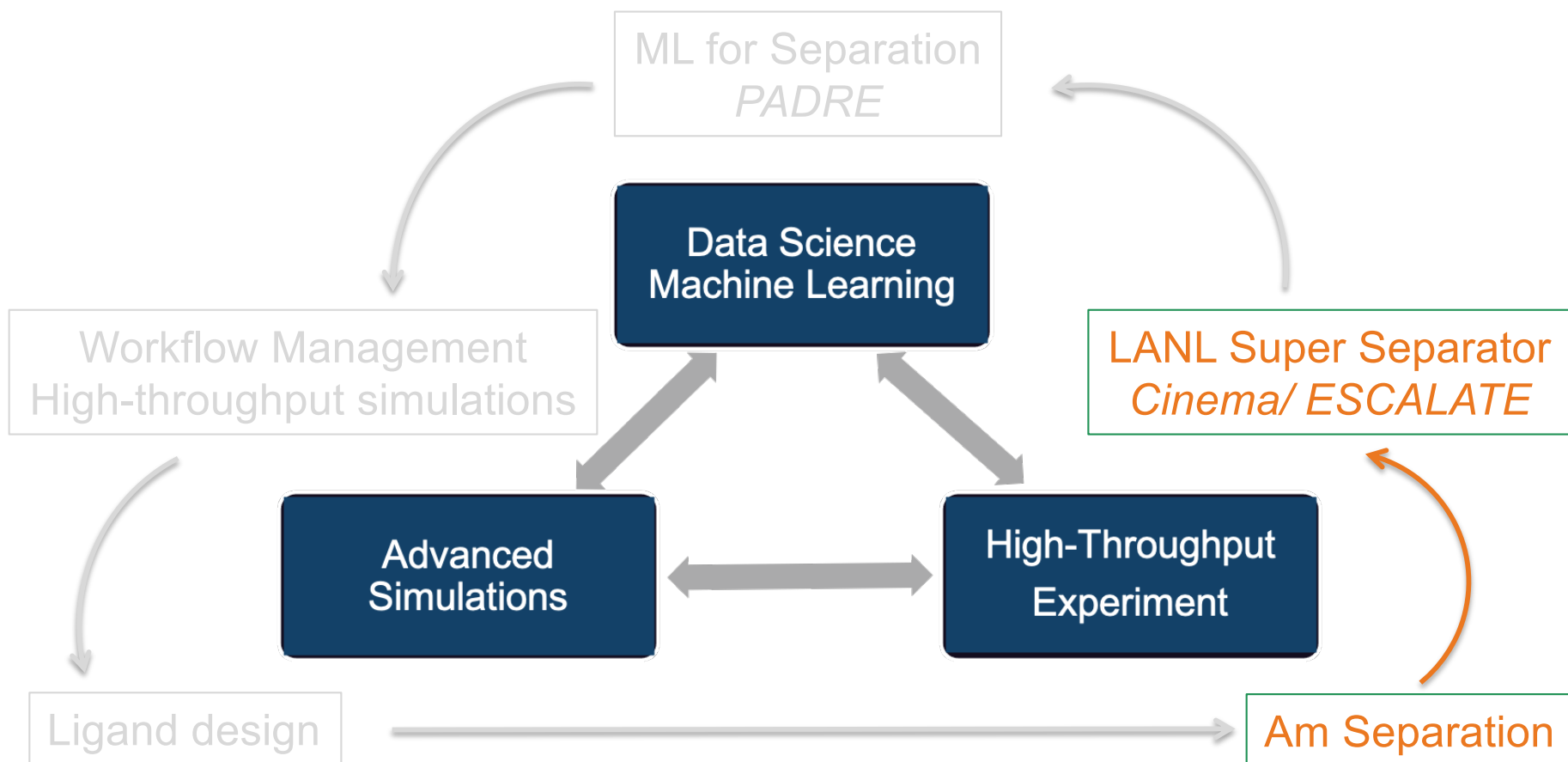
# Challenges of Applying Data-Science for An/Ln Separation

## What is Needed: Data!!!



- **High-throughput randomized uniform experimental data covering the vast chemical space.**
- **High-throughput theoretical data through advanced simulations**
  - Screen millions of ligands/extractants
  - Form chemically sensible An/Ln complexes
  - Manage thousands of calculations simultaneously
  - Quantum-based simulations for bond forming & breaking
  - Long-timescale molecular dynamics simulations across interfaces

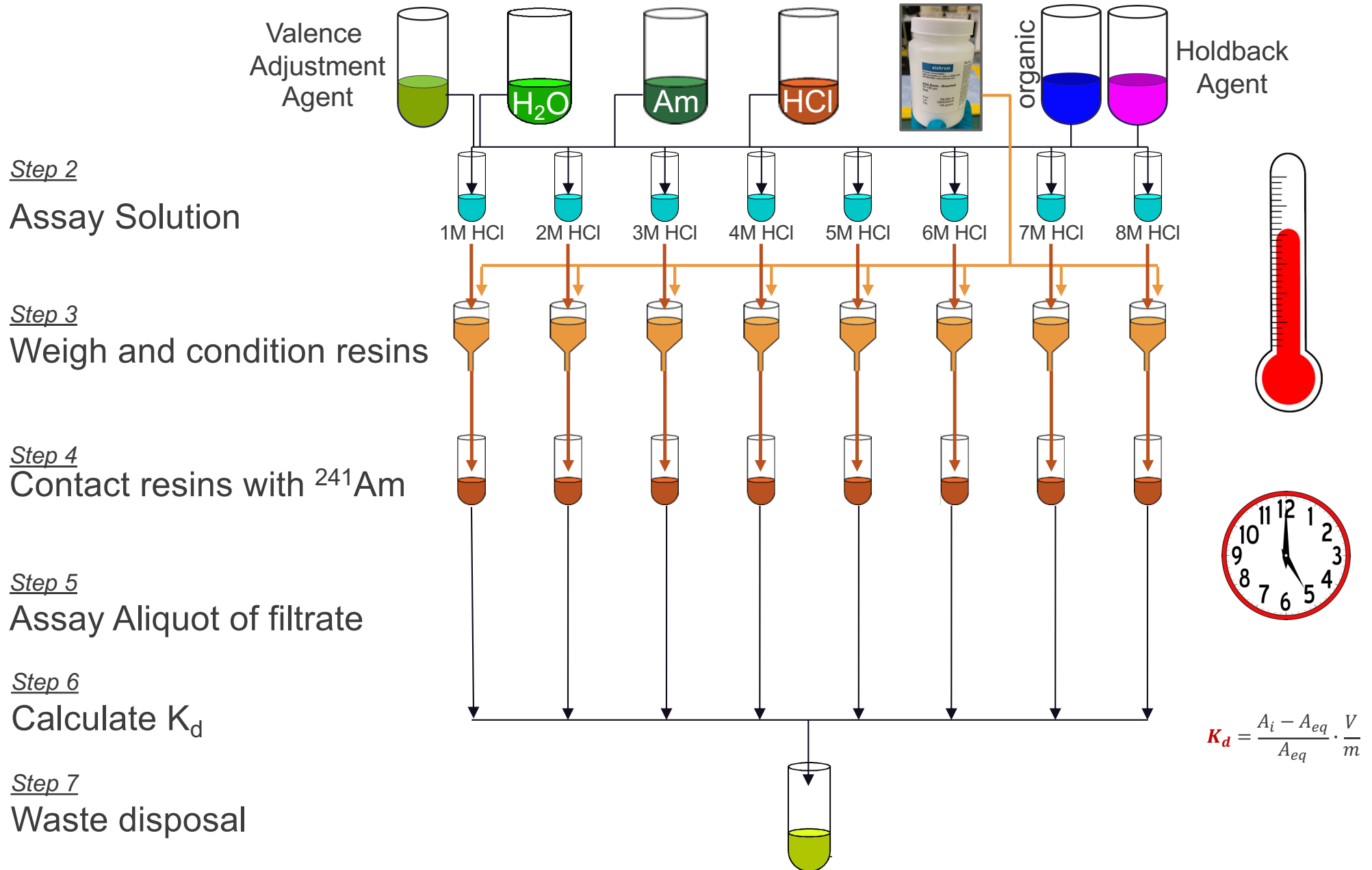
# Approach: SeparationML



**Follow the data!**

**Complex workflow management is a necessity.**

# Workflow of a $K_d$ Measurement Using Extraction Resins (in a regular lab)

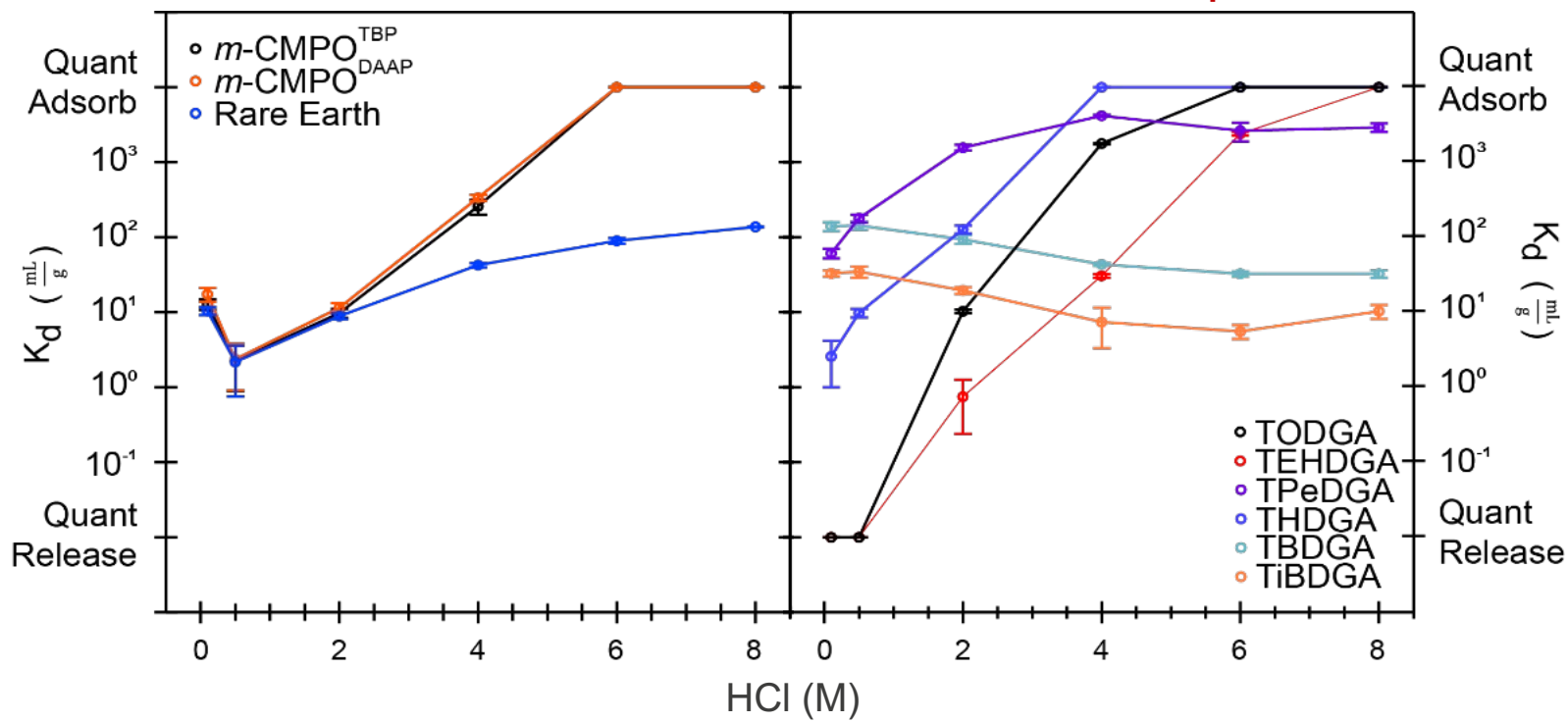


# Summary: Matrix Effects on Am Processing

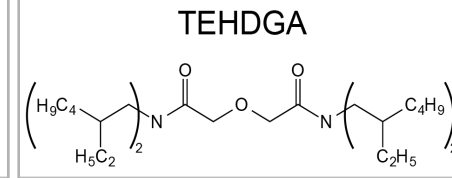
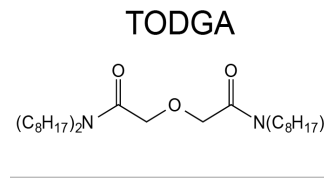
342 experiments later

## Seven Tested Properties

<b>Extractant identity</b>	Binding capacity
Phase transfer catalysts	Bead size
<b>Matrix effect</b>	Binding kinetics
<b>Aqueous matrix</b>	



- TEHDGA and TODGA  $\gg$   $m\text{-CMPO}^{\text{TBP}}$  and  $m\text{-CMPO}^{\text{DAAP}}$   $\gg$  Rare Earth (RE) resin.
- TEHDGA and TODGA also provide a release mechanism via controlling acid concentration



# Automating Separations: LANL Super Separator



Installed in February 2021



*Increases  
throughput*



*Human error*



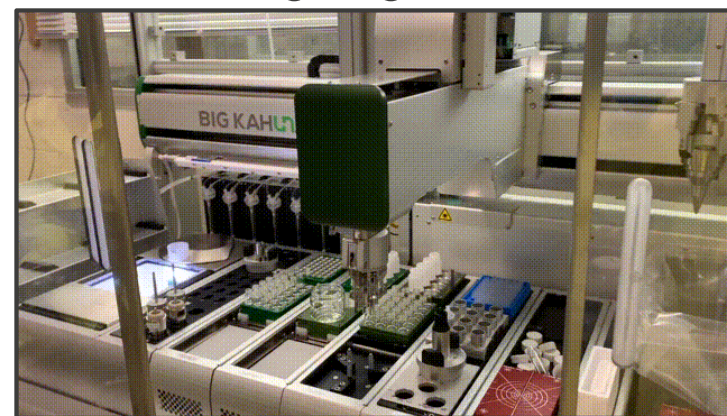
*Efficient*



*Cost Saving*

- Makes it easier to develop a new process.
- Makes it easier to optimize an operational process with the confines of an existing safety envelope.

## Weighing Solids



2.0 mg of resin  $\pm$  0.2 mg

## Dispensing Solutions

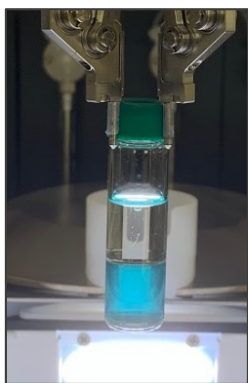
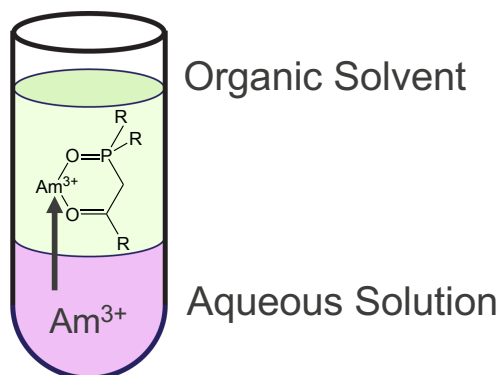


10 to 10000 mL of solution  
with 2% to 10% error

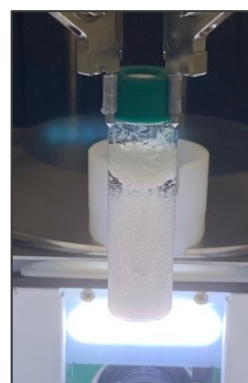
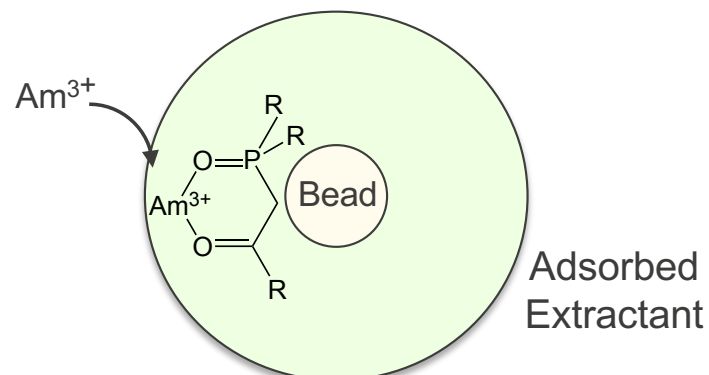


# Automating Separations: LANL Super Separator

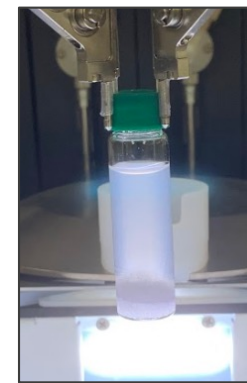
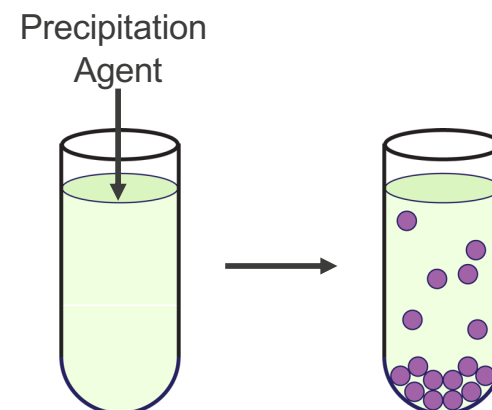
## Solvent Extraction



## Extraction Chromatography



## Selective Precipitation

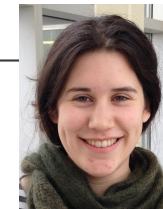


- Automation increases throughput and minimizes human error.
- Commissioned U and Th; transuranics soon

Stosh Kozimor

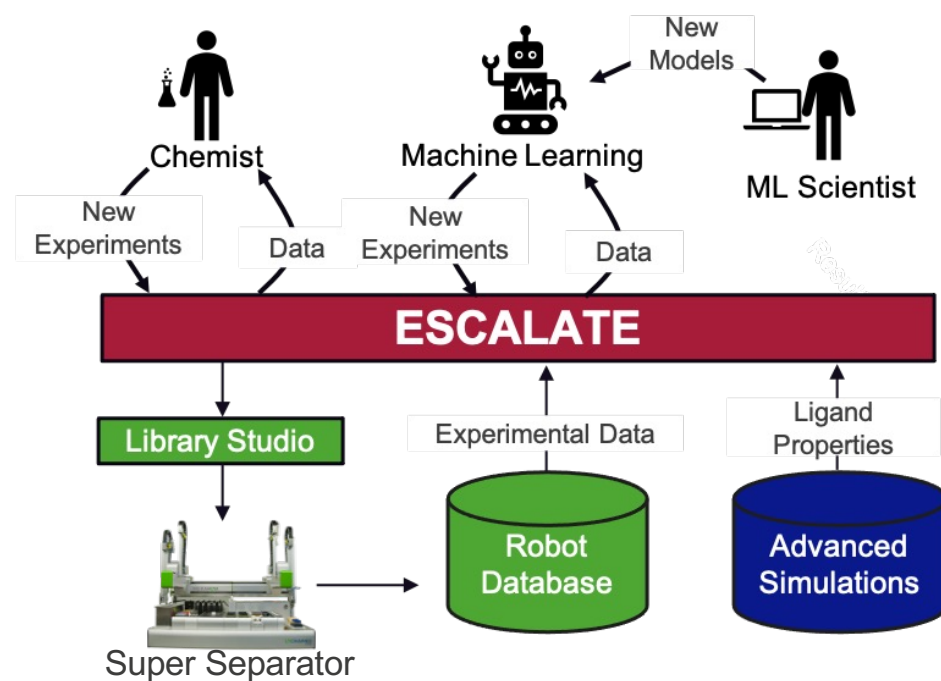
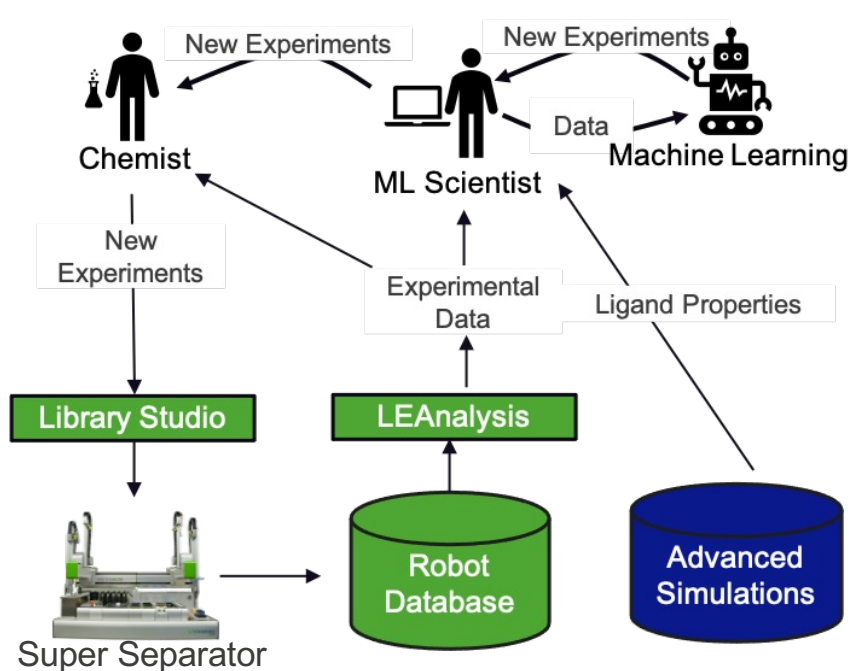


Sara Adelman



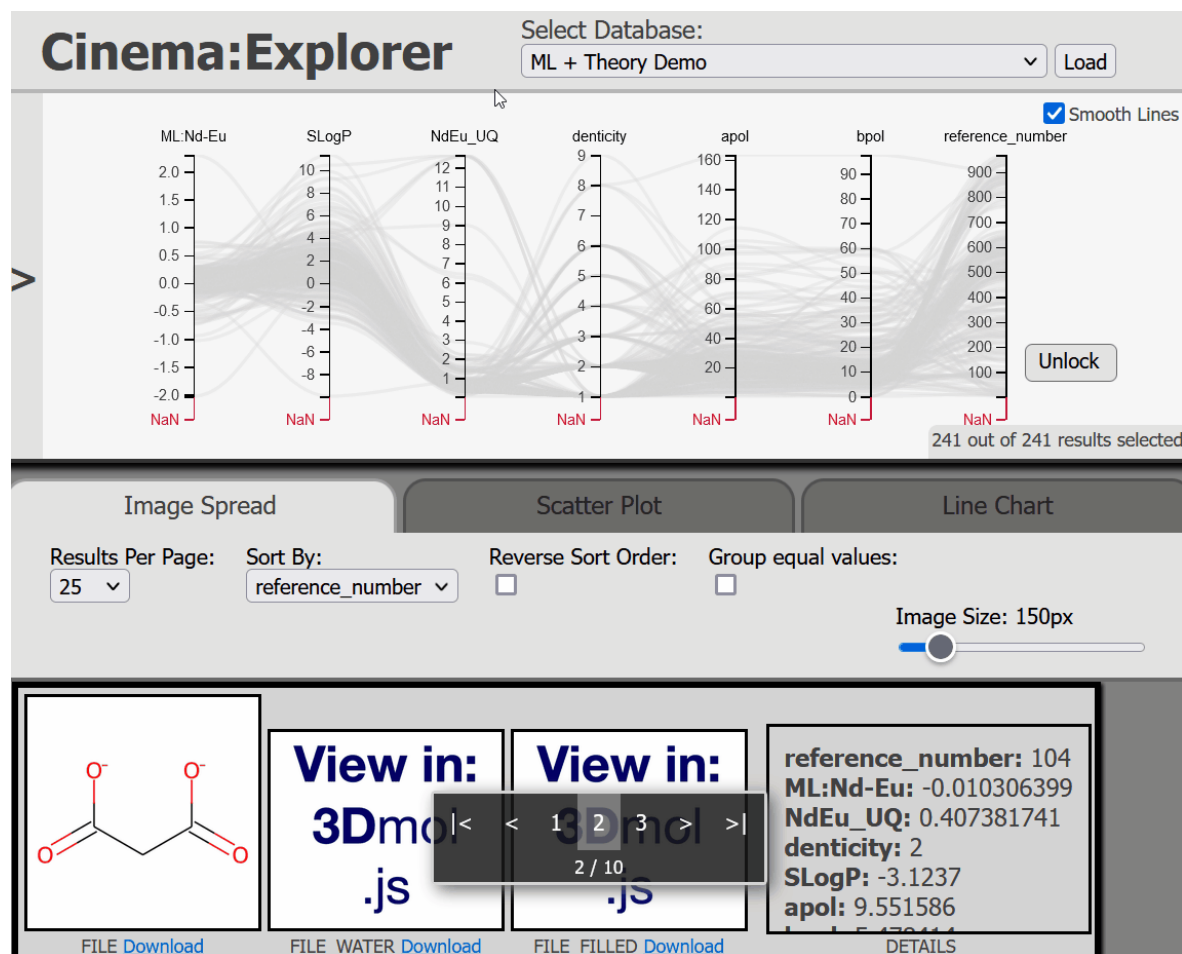
# Interfacing the robot with ML: ESCALATE

- Provides an abstraction layer between chemists + ML model with the Robot
  - Reduce complicated interaction between many pieces of software and databases
- A major software engineering task
  - Enables closed-loop interaction between ML and the Robot
    - Data from robot curated to ML model
    - API for humans or algorithms to specify new experiments
  - Enables easy visualization of large dataset through single web application



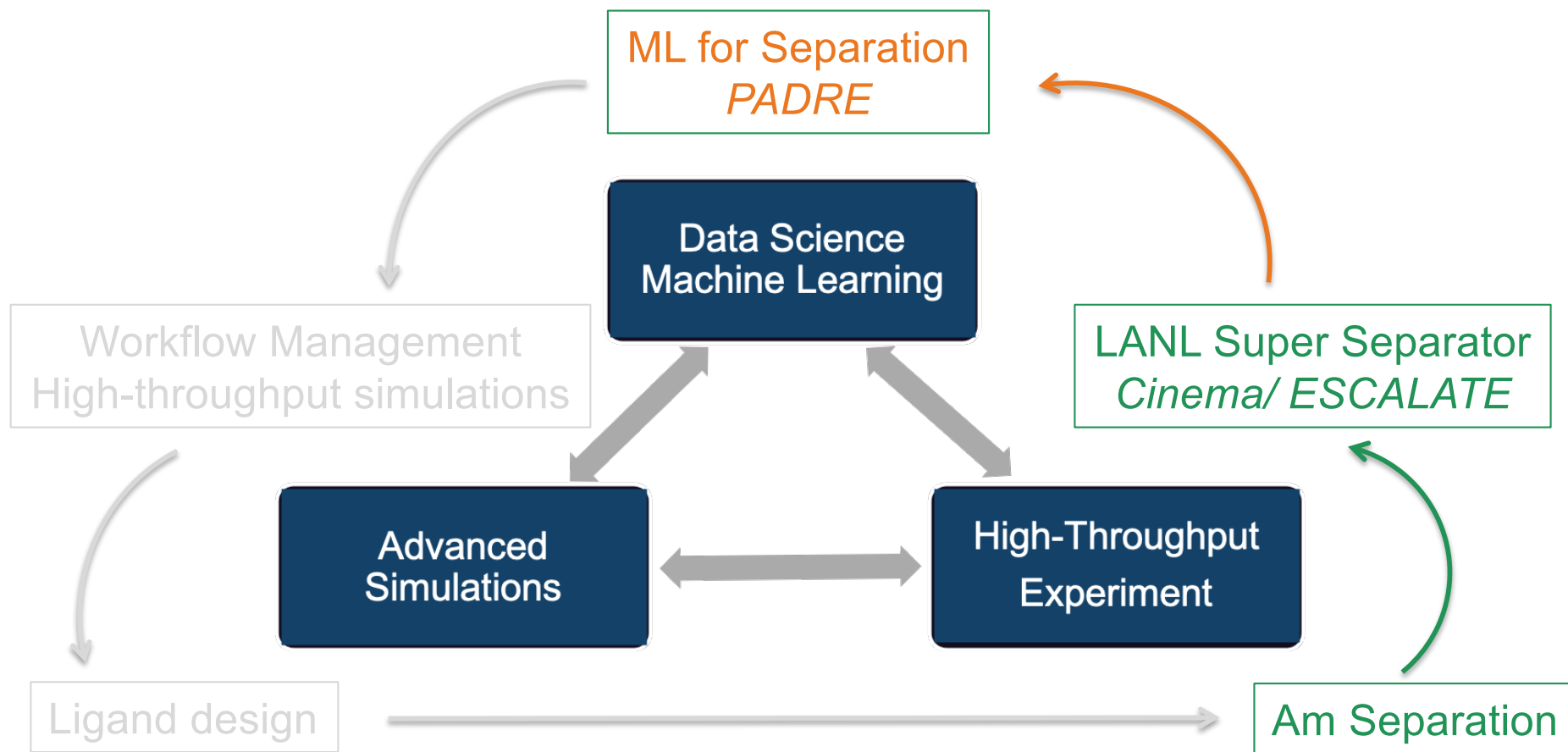
# Interactive Virtualization of High-dimensional Data

- Cinema Science tool allows interactive data analysis and visualization of theory, ML, and experimental data, especially large volume and high-dimension
- 3D molecular viewer, regular expression queries, other advanced features added by SeparationML team
- Collaboration with DOE-ASCR, ECP, SciDAC programs:  
*Data Science at Scale.*



<https://cinemascience.github.io/>

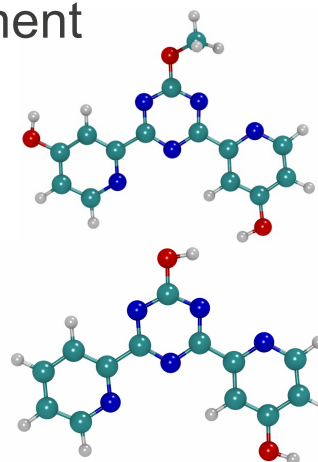
# Approach: SeparationML



# The ML Model: Motivating Observation

- Errors in different models often cancel: calculated energy differences are often much more precise than absolute values.
- “Chemistry informed ML”: Can this be applied to the ML models as well?
- We have developed a “pairwise” ML method that addresses key challenges of:
  - uncertainty quantification
  - limited size datasets

Experiment

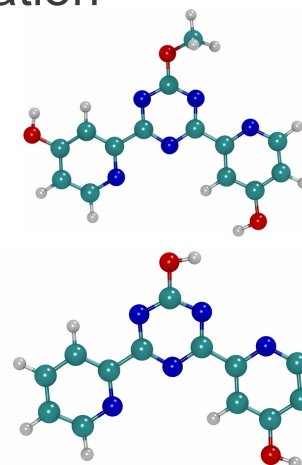
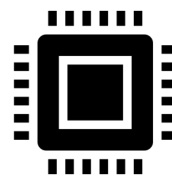


110 kcal/mol

100 kcal/mol

10 kcal/mol

Computation



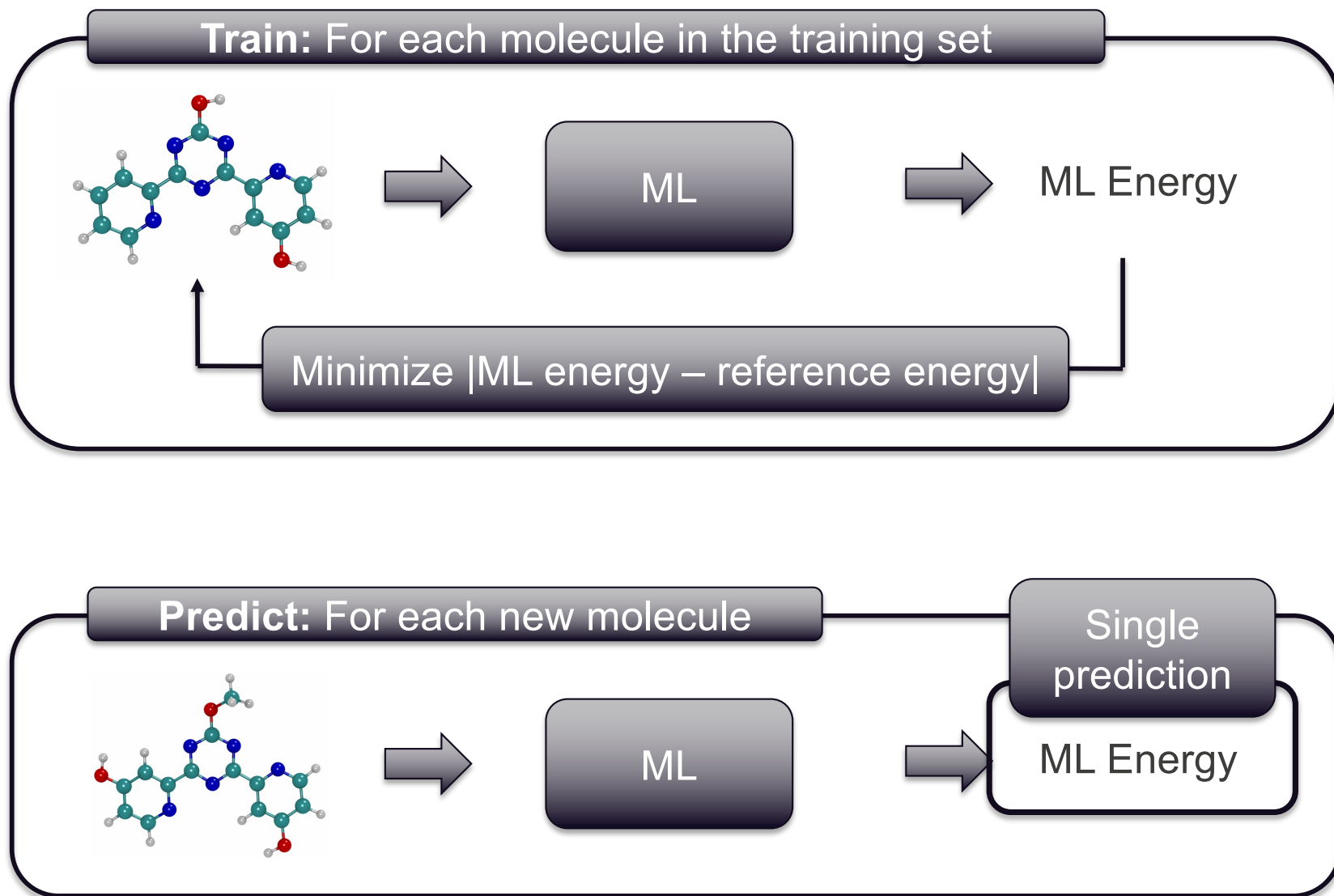
31 kcal/mol

20 kcal/mol

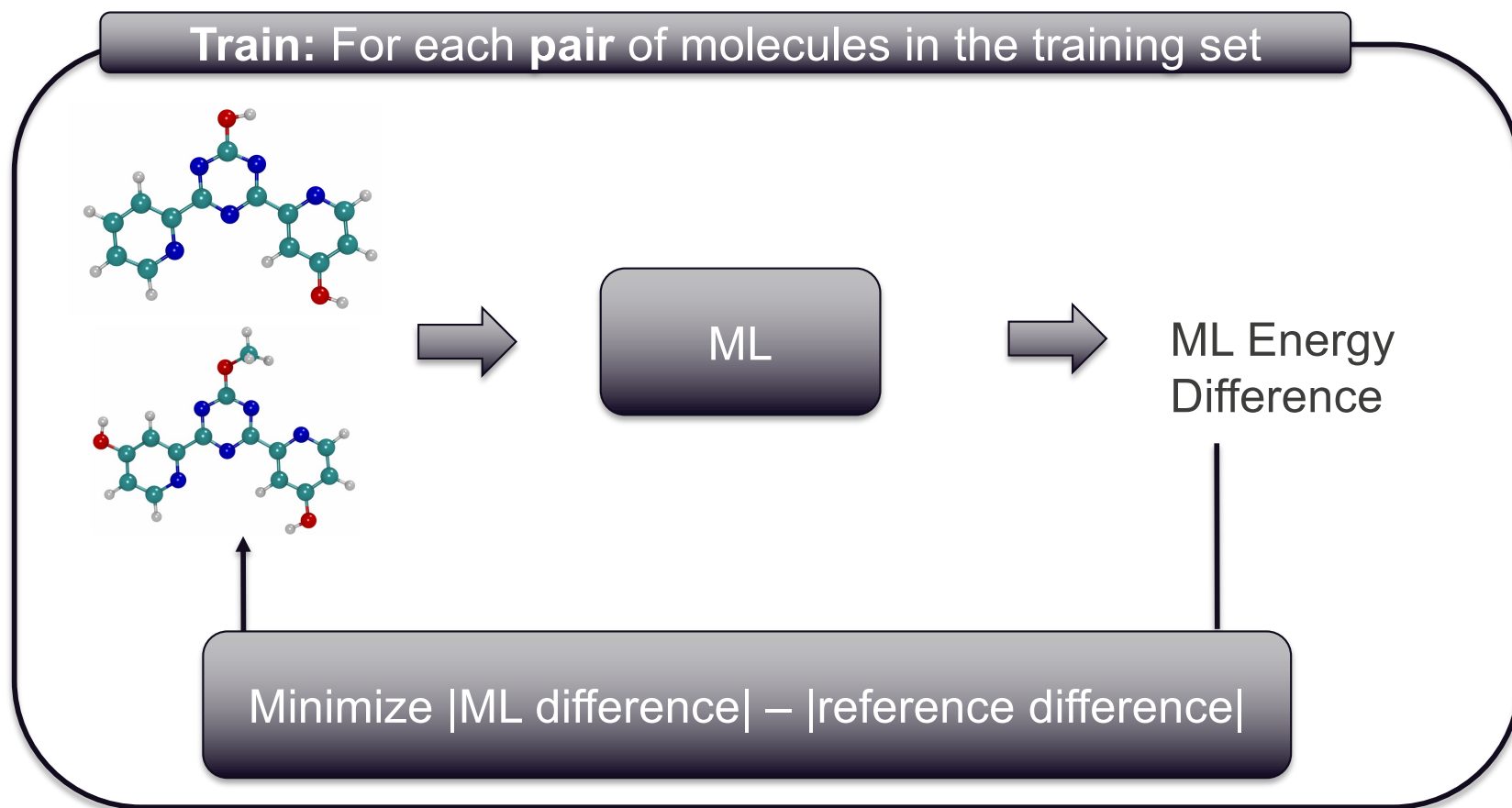
11 kcal/mol

*M Tynes, W Gao, DJ Burrill, ER Batista, D Perez, P Yang, N Lubbers, J. Chem. Inf. Model, 2021, 61, 3846-3857*

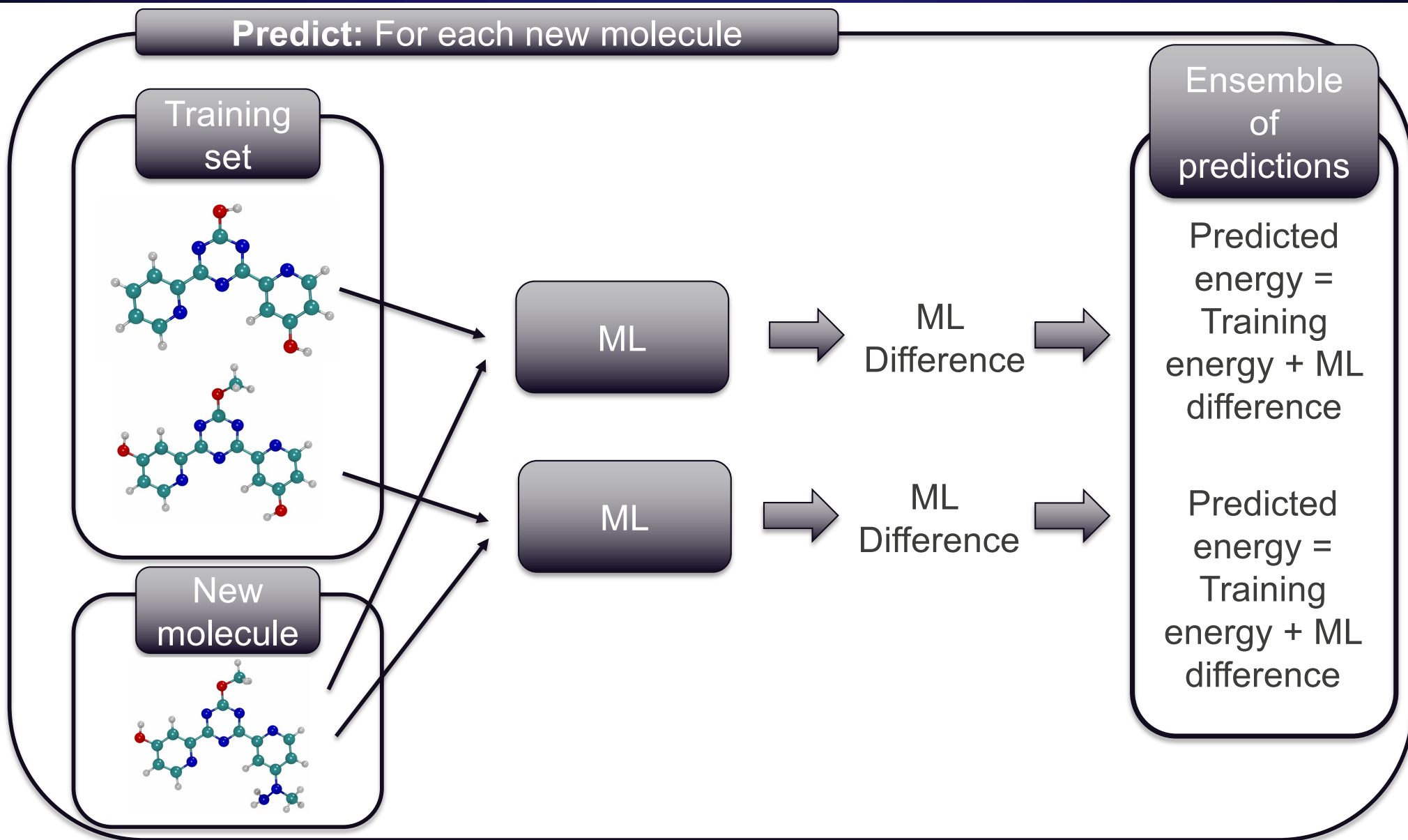
# Conventional ML



# ML Model: Pairwise Difference Regression (PADRE)

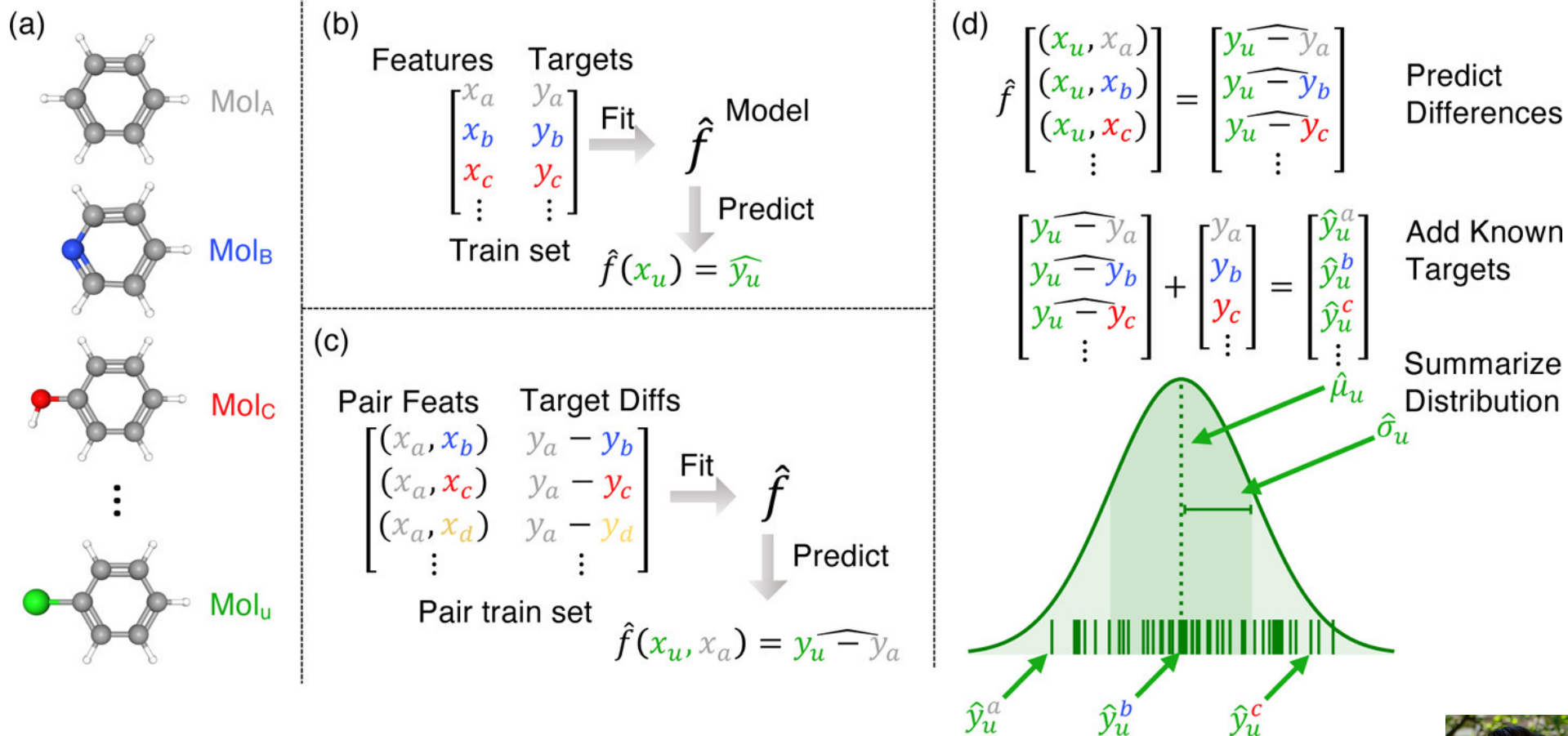


# ML Model: Pairwise Difference Regression (PADRE)





# PADRE Algorithm Overview



M Tynes, W Gao, DJ Burrill, ER Batista, D Perez, P Yang, N Lubbers,  
*J. Chem. Inf. Model*, **2021**, 61, 8, 3846–3857

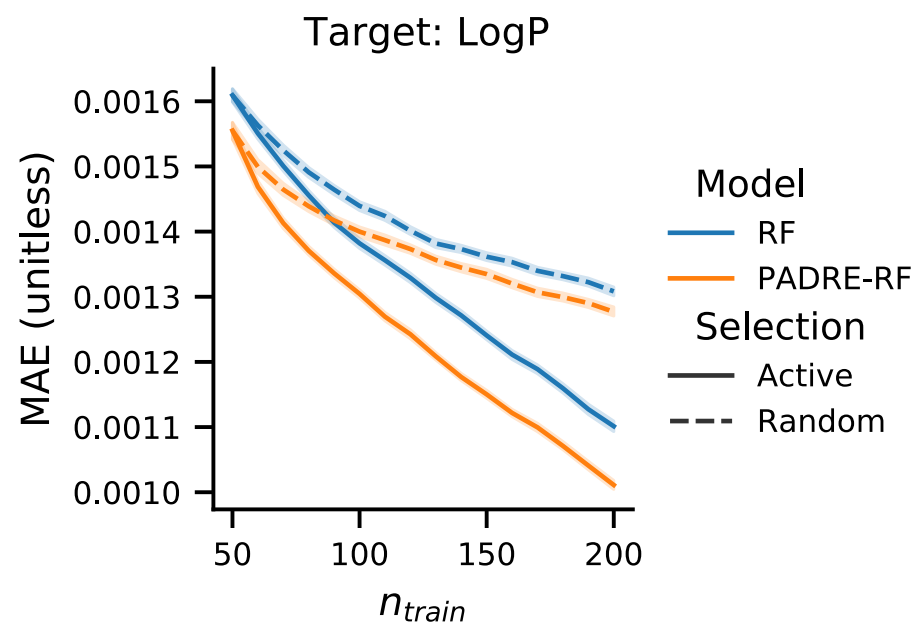
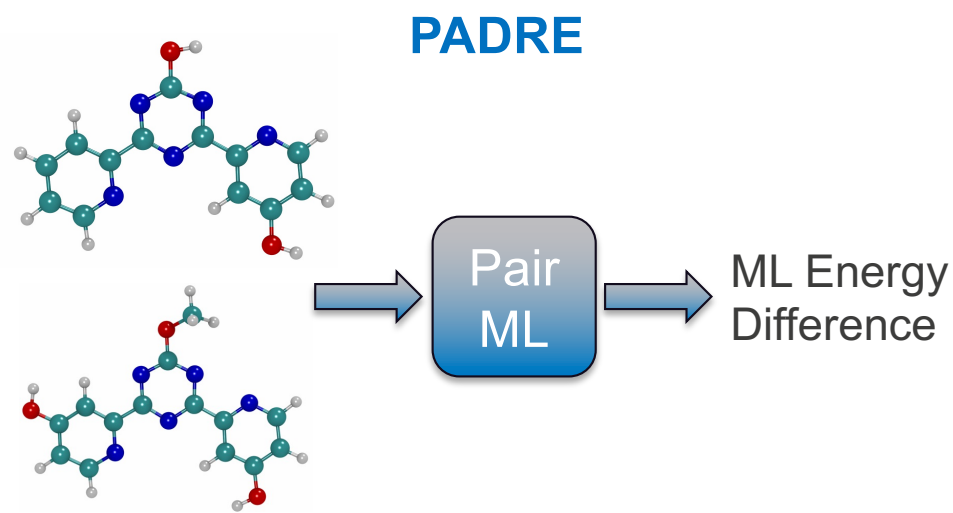
Mike  
 Tynes



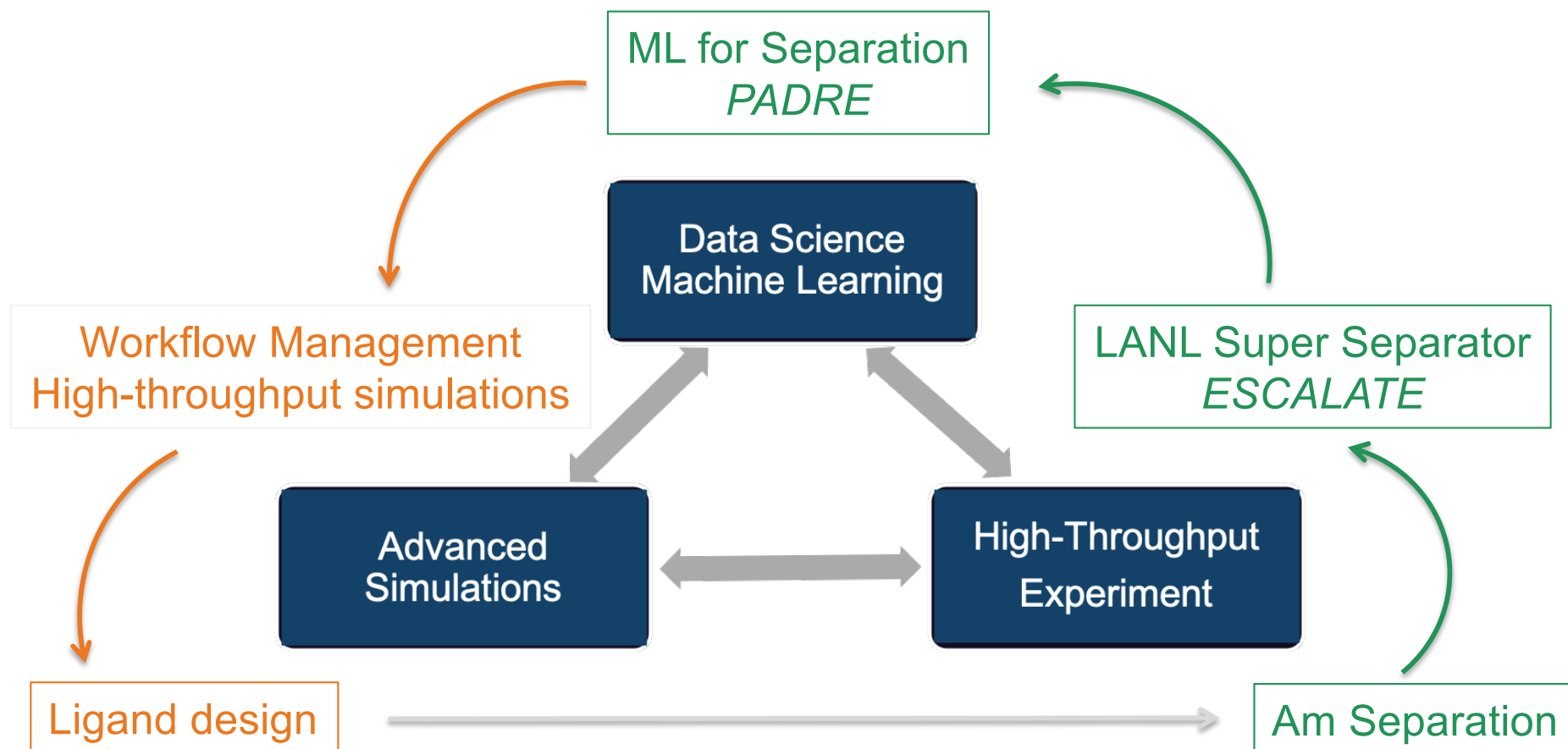
# PADRE: Features & Performance

A chemistry-informed ML model  
Pairwise difference regression:

- *Predict relative quantities, e.g. Separation Factors*
- $n^2$  data augmentation
  - Train on all pairs
  - Improve prediction by 10%
- ML meta-algorithms for arbitrary base regressors
  - Simple RF+PADRE is competitive with more complex NN
  - Redox, solubility, and BE
- Provides a useful UQ metric
  - Competitive with state-of-the-art chemical candidate selection

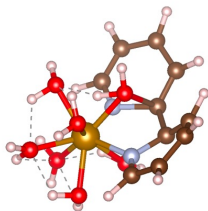


# Approach: SeparationML

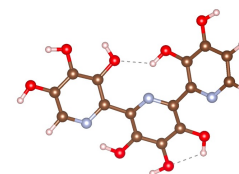


# High-throughput Quantum Simulations

**Architector**, Generate chemically-sensible extractant-actinide complexes



**Vulcan**, Suggest new extractants factoring in synthesizability, pH, solubility ...



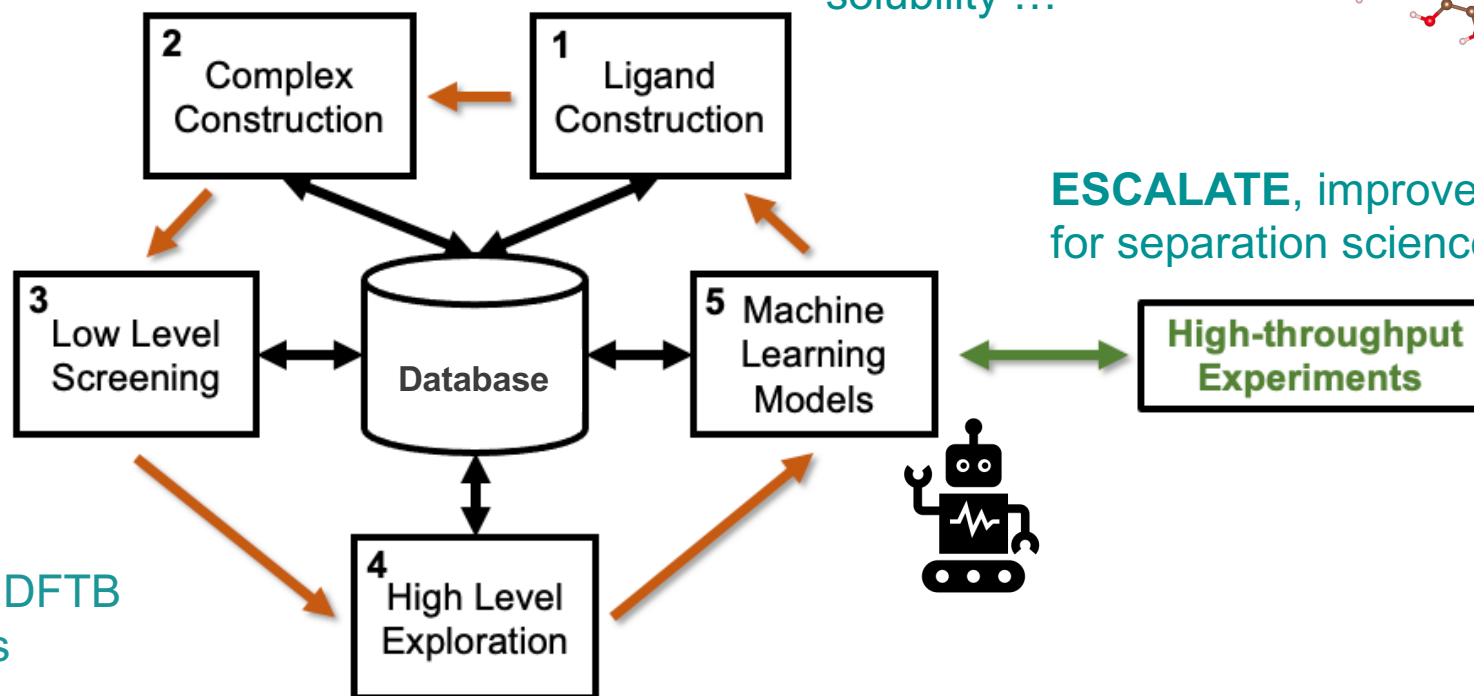
**Mineva**, calculate topological chemical descriptors

**EPO**, train DFTB parameters

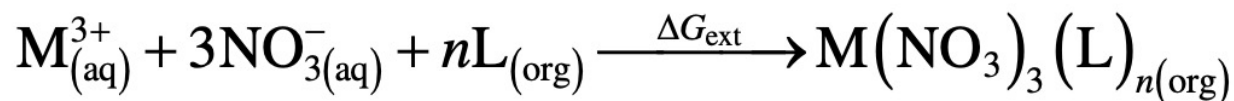
**ADF/Gaussian**, DFT calculations

**pyiron**, complex workflow management package for all-level calculations simultaneously using HPC to enable automatus discovery

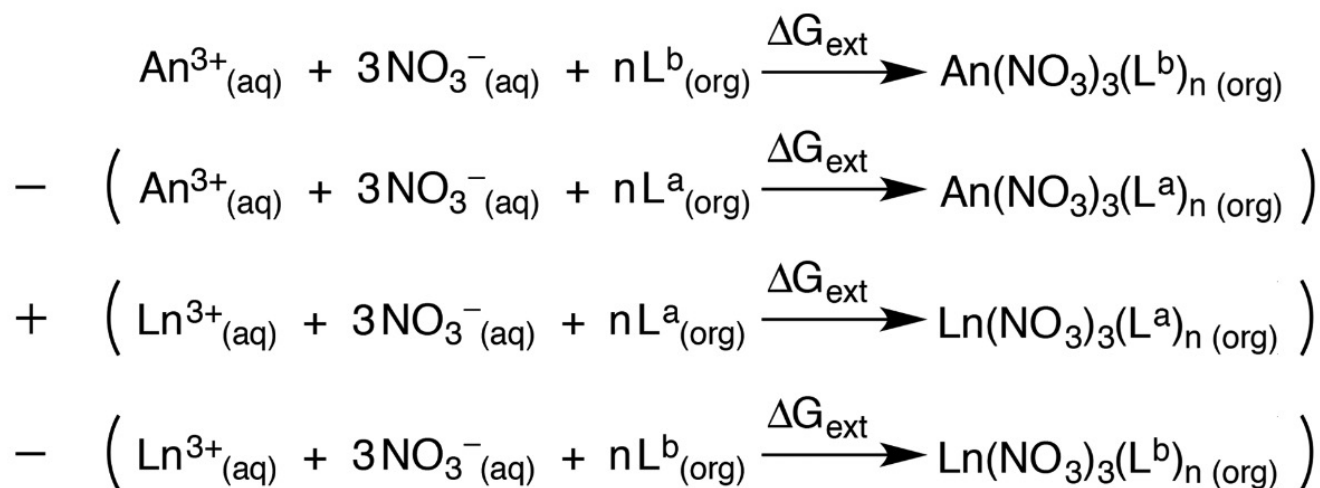
**ESCALATE**, improved for separation science



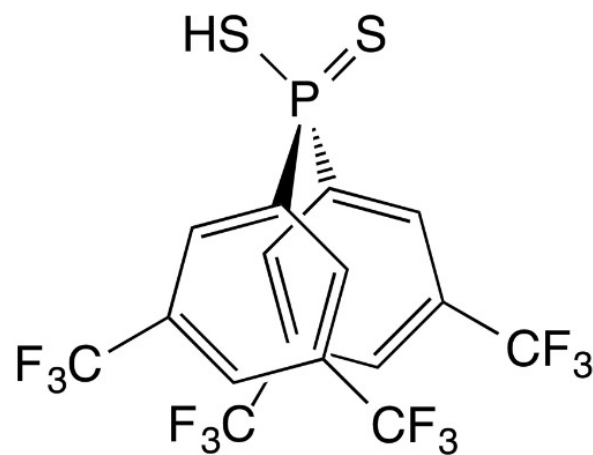
# Calculating Separation Factors



$$SF_{An/Ln} = D_{An}/D_{Ln} = \exp(-\Delta\Delta G_{ext}/RT)$$



# An Example of Calculating Separation Factors



L1

$SF_{Am/Eu}$

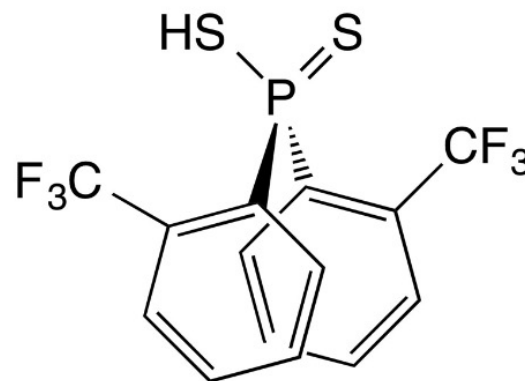
20

$\Delta\Delta G_{ext}$

-1.8

$\Delta\Delta\Delta G_{ext}$

-5.0



L2

100,000

-6.8

kcal/mol

# Workflow for Calculating Separation Factors

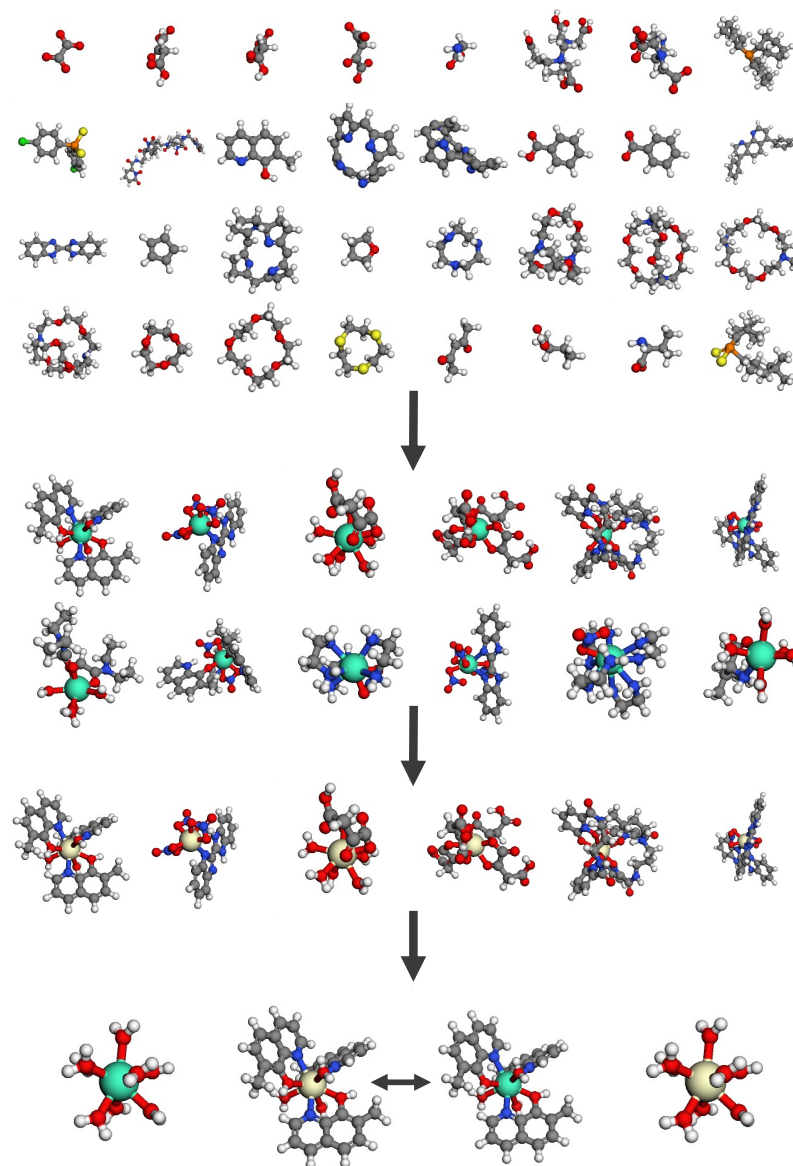
New extractants

Architector build complexes

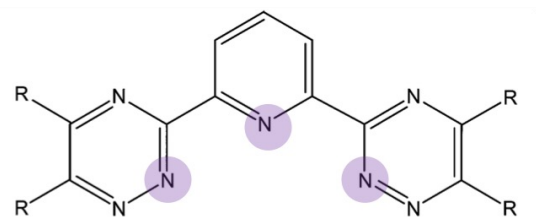
pyiron/ADF to DFT-relax 5-lowest GFN2-xTB conformers

Swap Metal for other Metals + Re-optimize

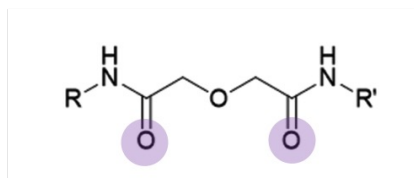
Calculate  $\Delta\Delta G_{\text{rxn}}$  Swapping Metals/Ligands



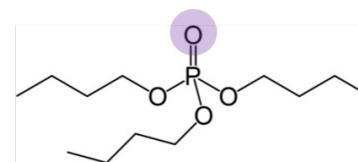
# A Must: Building Chemically Sensible 3D Structures of f-Element Complexes



Bis-triazinyl-pyridine ligands (BTP)

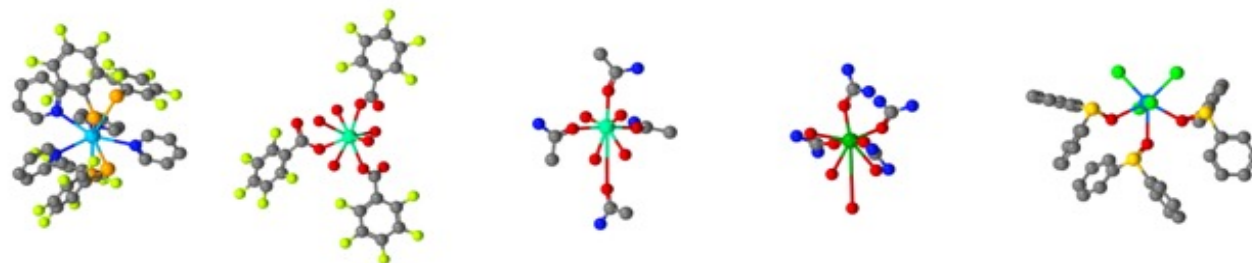


Diglycolamide (DGA)



Tributyl-phosphate (TBP)

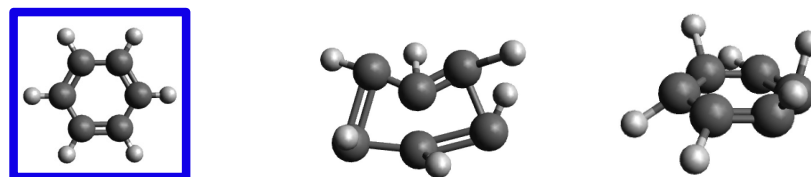
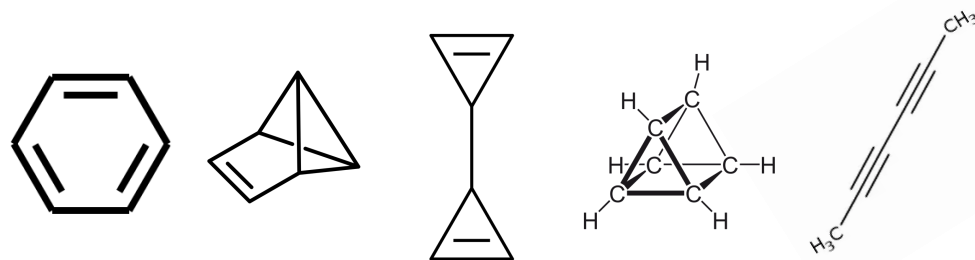
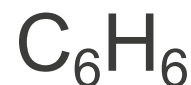
How to form 3D structures as a function of ligands, metal, coordination number, counter ions?





# Generating 3D Structure In Chemistry

- Degrees of abstraction in chemistry:
  - 1D = Chemical Formula
  - 2D = Molecular Graph
  - 3D = Atomic Positions  
( Electronic Structure!)
- Moving from 3D->1D = relatively easy
- Moving from 1D->3D = much harder!
- 2D -> 3D complexity alone is NP-hard
- d-block: molsimplify, DENOPTIM, Molassembler
- f-block: no tools available



molsimplify



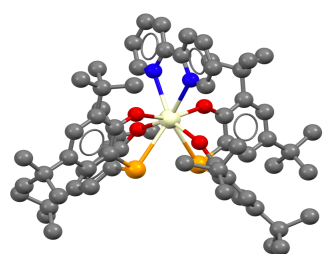
Molassembler

# Architector Design Overview

## Inputs

Metal, Coordination Number, Ligands, Coordinating atoms (CAs)

**Example:** Cambridge Structural Database Refcode = CONPEC



Metal: **Ce**

Coordination Number: 8

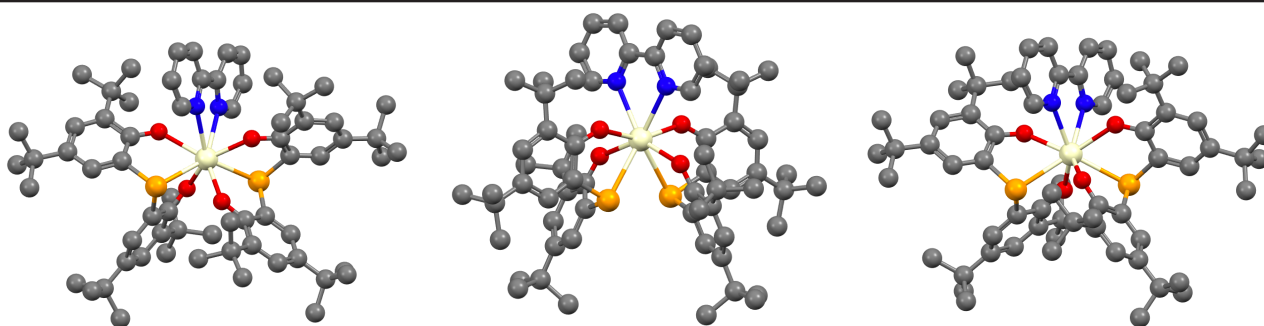
Ligands (SMILES, CAs):

1. c1ccc(nc1)c1cccn1, (4, 11)
2. CC(C)(C)c1cc([Se]c2cc(cc(c2[O-])C(C)(C)C)C(C)(C)C)c([O-])c(c1)C(C)(C)C, (7,14,24) (x2)

**User interactions**

## Outputs

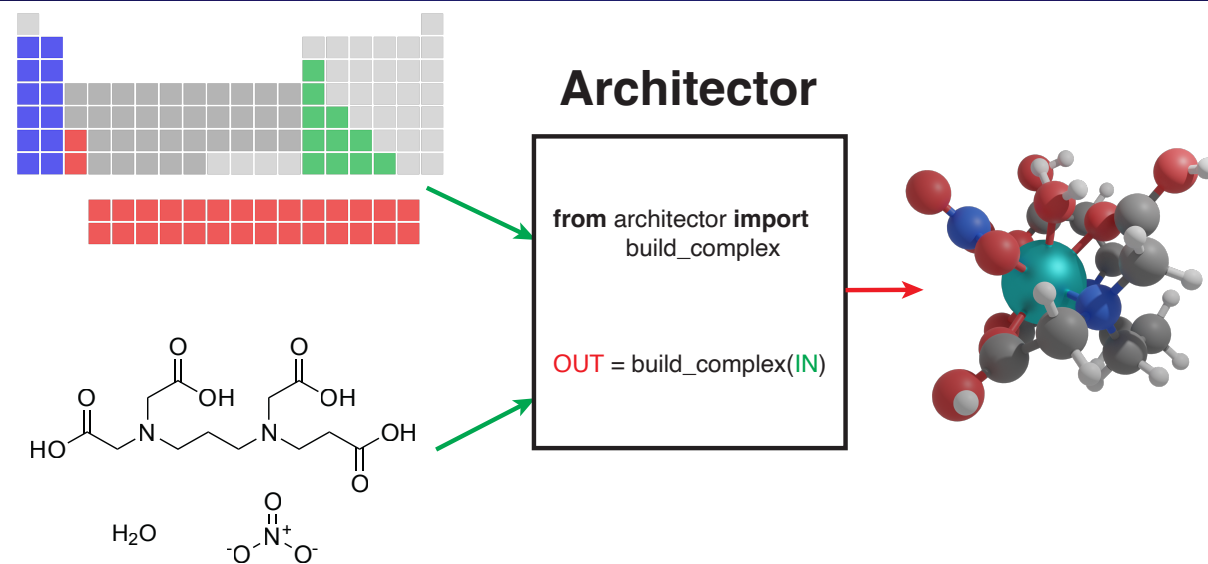
Ranked-order (by GFN2-xTB energy) complexes



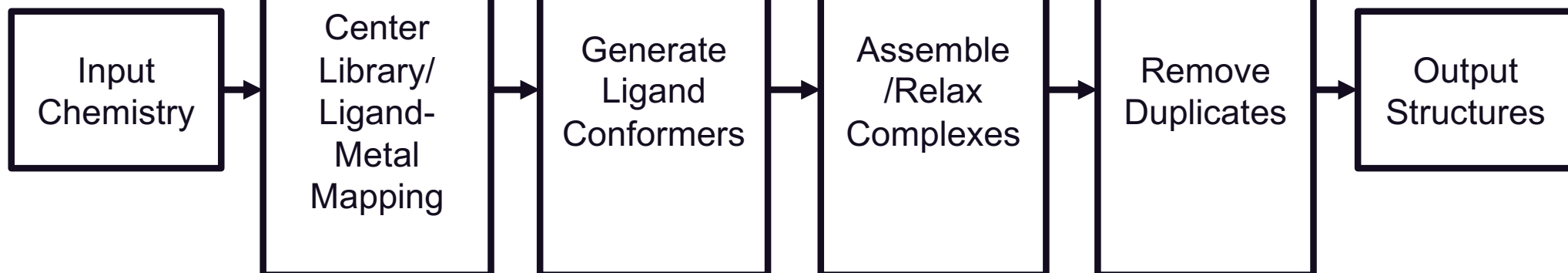
Michael Taylor

<https://github.com/lanl/Architector>

# Architector – 3D Molecule Generation Workflow



Iterate over Metal Center Symmetries



# Computational Tools Needed for the Architector Workflow



**Python:**  
Input/Output  
Dictionary / YAML  
Handles Connections  
Between Software



## Architector

```
from architector import  
build_complex  
  
OUT = build_complex(IN)
```



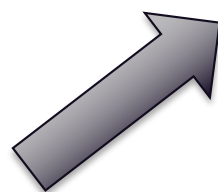
**Atomic Simulation Environment:**

1. Non-forcefield Calculations
2. Structural Data



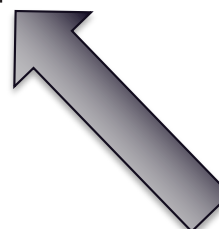
**OpenBabel:**

1. SMILES Conversion including charge handling
2. Ligand initial conformer generation
3. Forcefield Relaxations



**Extended Tight Binding:**

1. Tight Binding Calculations
2. GFN- Forcefield Calculations

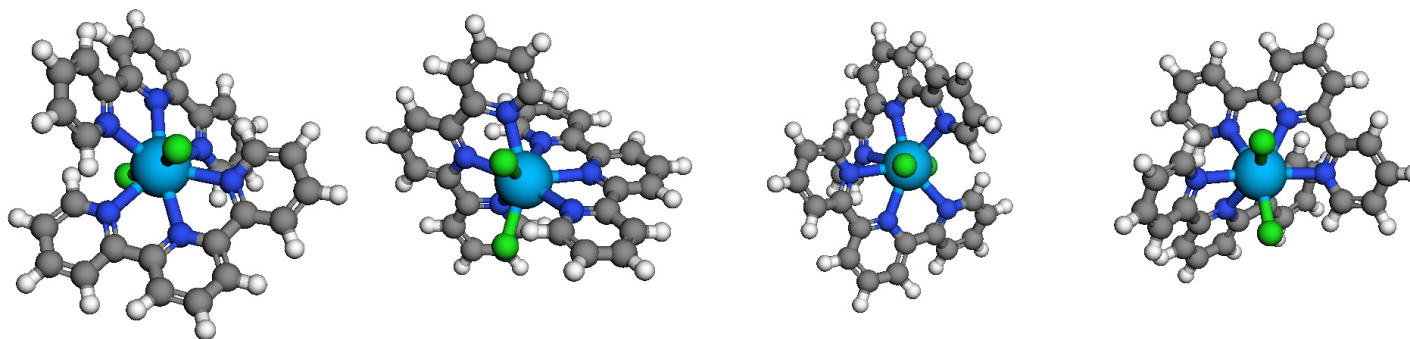


# Architector Visualization

```
In [1]: from architector import build_complex, view_structures
```

```
In [3]: out = build_complex({'core':{'metal':'Th', 'coreCN':8},  
                             'ligands':['terpy']*2+['chloride']*2,  
                             'parameters':{'metal_ox':4}})
```

```
In [4]: view_structures(out)
```

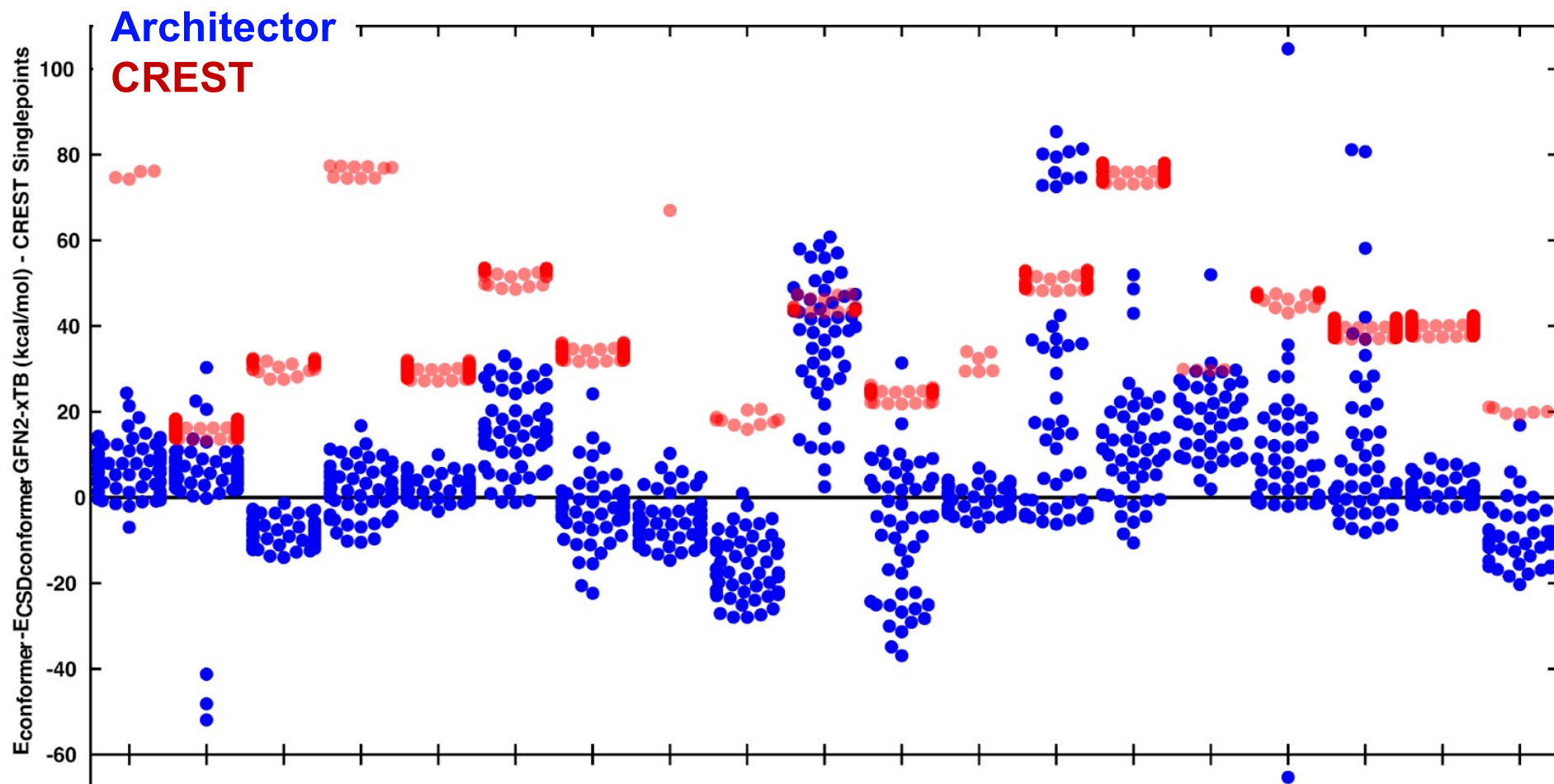


- Visualization routine integrated with jupyter notebooks – broad use-cases for visualization (dynamic grid visualization)

**3Dmol.js**

<https://github.com/lanl/Architector>

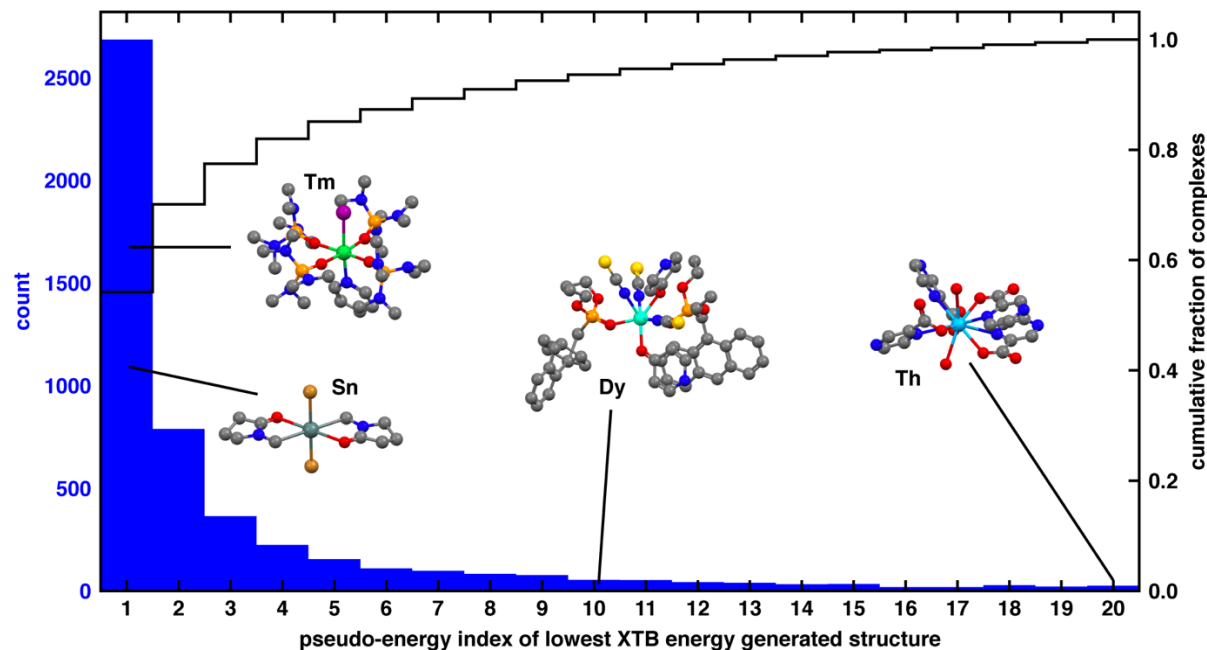
# Finding New Conformers Beyond CSD Database



- Gas phase or liquid phase conformers can dramatically differ from crystalized structures. Conformer sampling can be vital.
- CREST<sup>1</sup> needs 3D structure as input, vs. Architector only needs 2D input

# Architector Performance

- Benchmarked on over 6,500 experimental structures.
- “Embarrassingly parallel” 99% finished in under 12 hours on 500 cores.
- Vast majority (95%) produced at least one conformer within 10 kcal/mol or lower than experimental structures.
- Diverse conformer generation for higher-energy symmetries.



<https://github.com/lanl/Architector>

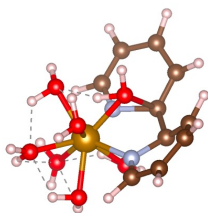
**conda install -c conda-forge architector**

Browser-based webserver for the community is coming. –Based on NERSC SPIN

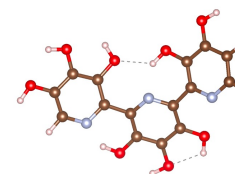


# High-throughput Quantum Simulations

**Architector**, Generate chemically-sensible extractant-actinide complexes



**Vulcan**, Suggest new extractants factoring in synthesizability, pH, solubility ...



**Mineva**, calculate topological chemical descriptors

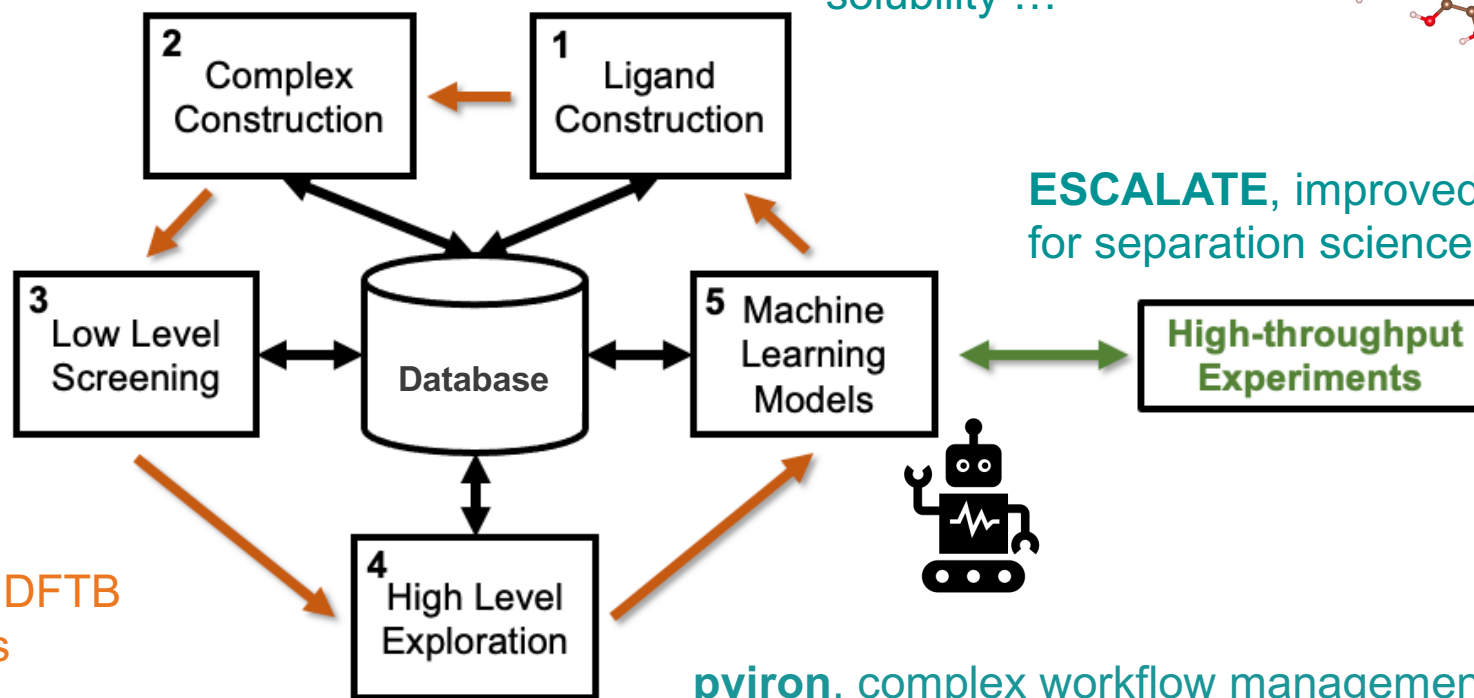
**EPO**, train DFTB parameters

**ADF/Gaussian**, DFT calculations

**pyiron**, complex workflow management package for all-level calculations simultaneously using HPC to enable automatus discovery

**ESCALATE**, improved for separation science

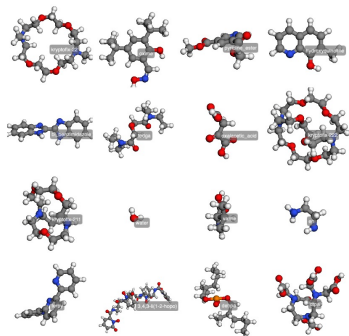
**High-throughput Experiments**





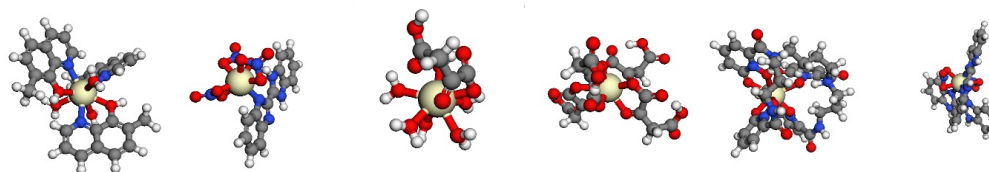
# Large-Scale Quantum-based Simulations are needed

## Screening of large number of structures

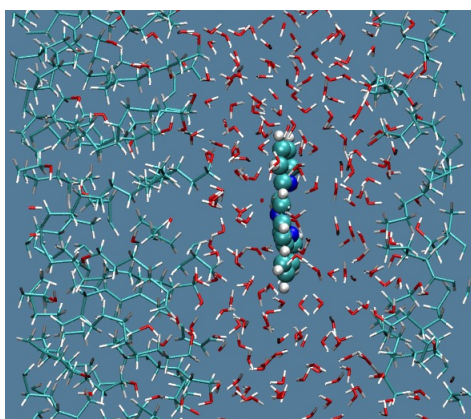


12 extractants, pH depend forms  
2 metals, 2 counter ions, 5 conformers

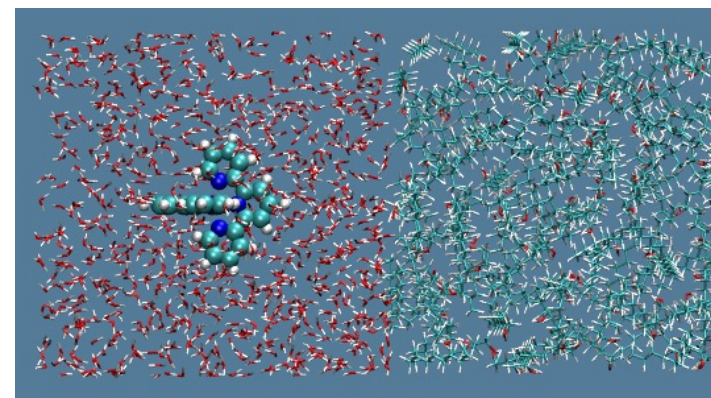
~2200 DFT jobs



Dynamic behavior needs long time scale simulations to cross the interface.



1907 atom  
terpy + H<sub>2</sub>O + octanol



6150 atom ML<sub>2</sub> + H<sub>2</sub>O + octanol  
Largest SCC-DFTB calculation!

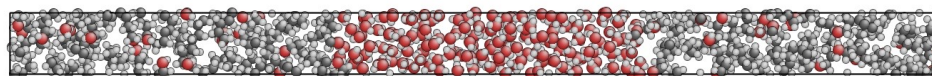
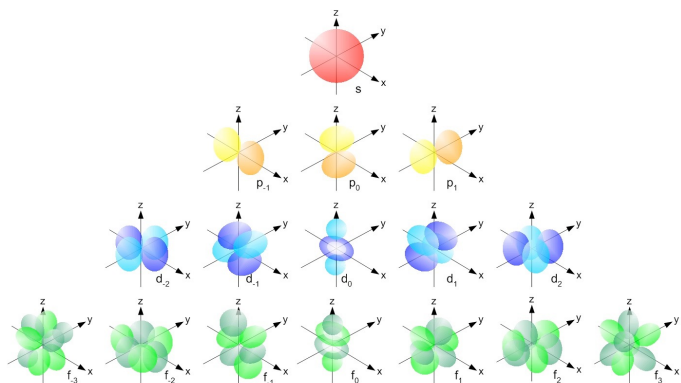
# Density Functional Tight Binding (DFTB)

- DFTB is a semi-empirical method derived from DFT

$$E_{DFTB} = \underbrace{\sum_i^{occ} \langle \Psi_i | \hat{H}_0 | \Psi_i \rangle}_{\text{Band structure energy}} + \underbrace{\frac{1}{2} \sum_{\alpha, \beta}^N \gamma_{\alpha\beta} \Delta q_{\alpha} \Delta q_{\beta}}_{\text{Charge transfer energy}} + \underbrace{\frac{1}{2} \sum_{\alpha}^N \sum_l W_l m_l^2}_{\text{spin energy}} + \underbrace{\frac{1}{2} \sum_{\alpha, \beta}^N V_{rep, \alpha, \beta}}_{\text{empirical repulsive pair potential}}$$

- Why DFTB (LATTE)?

- Fast:** Roughly ~100X faster than DFT
  - Linear-scaling algorithm available for large systems
  - XLBOMD formalism removes the expensive SCF iterations at each time step
- Accurate**
  - DFT level accuracy for forces and energies with good parameterization
  - Self-consistent-charge ensures describing the quantum nature of chemical bonds



1560 atom H<sub>2</sub>O + octanol

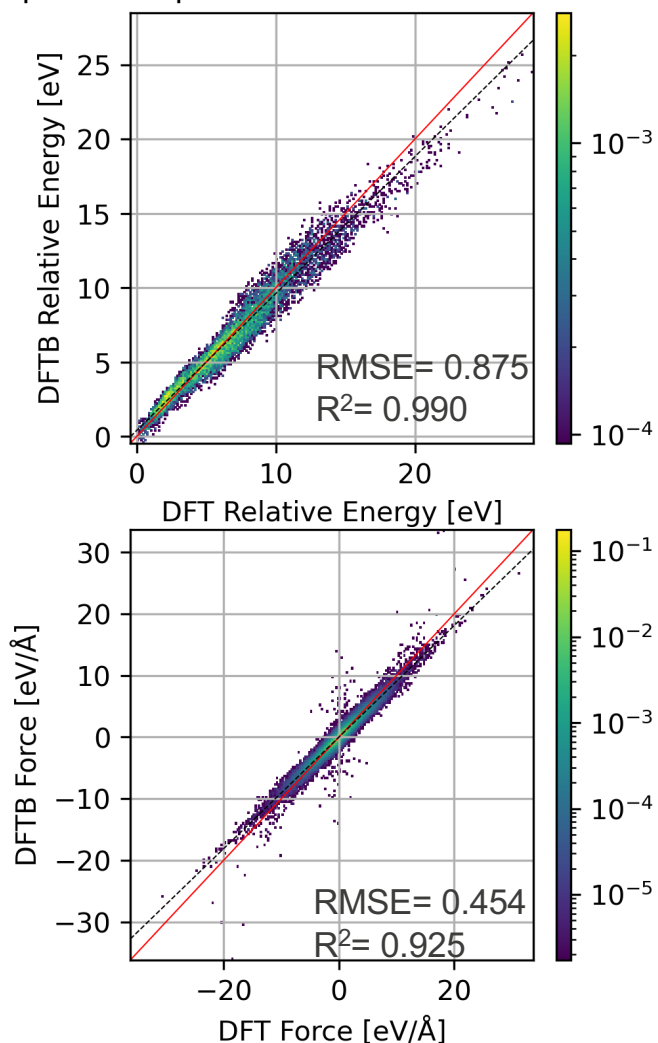
Method	Time/time step (s)
Full convergence + diag	1.100
Zero SCF + diag	0.590
Zero SCF + sparse SP2	0.427
Zero SCF + sparse SP2 (GPU)	0.552



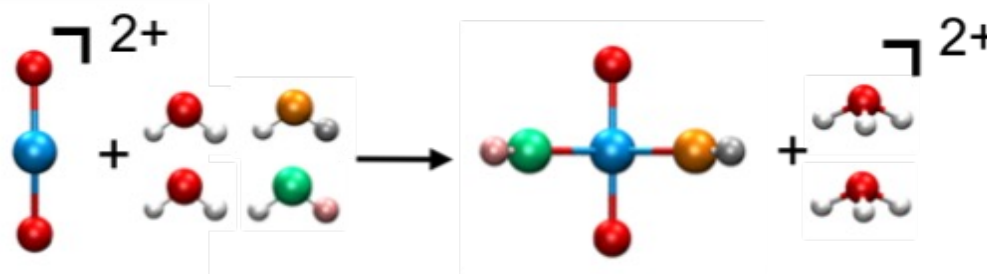
Marc Cawkwell  
LANL

# DFTB Energies, Forces, and Reactivity

Optimized parameters on n=1-8 clusters



## DFTB for Uranium Reactivity



	KS-DFT	DFTB
$\text{UO}_2^{2+} + 2\text{H}_2\text{O} \rightarrow [\text{UO}_2(\text{OH})]^+ + \text{H}_3\text{O}^+$	-7.69	-8.64
$\text{UO}_2^{2+} + 4\text{H}_2\text{O} \rightarrow [\text{UO}_2(\text{OH})_2] + 2\text{H}_3\text{O}^+$	-8.02	-9.44
$[\text{UO}_2(\text{OH})]^+ + 2\text{H}_2\text{O} \rightarrow [\text{UO}_2(\text{OH})_2] + \text{H}_3\text{O}^+$	-0.33	-0.80

Challenges of DFTB parameterization:

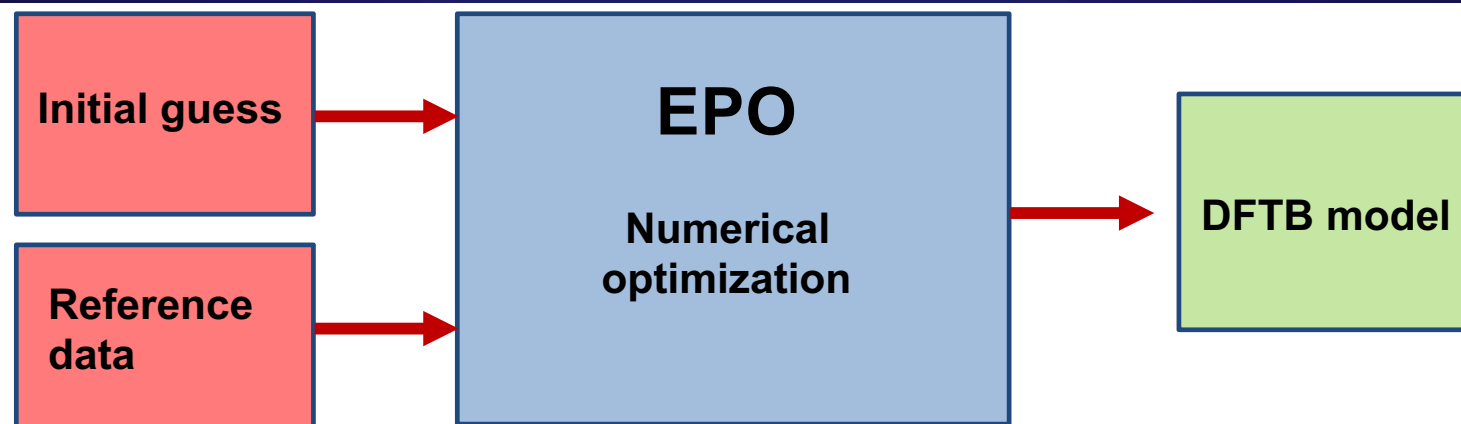
- Large number of parameters
- Limited availability of training data for large chemical space

Th-O: Liu, Aguirre, Cawkwell, Batista, Yang, *to be submitted*.

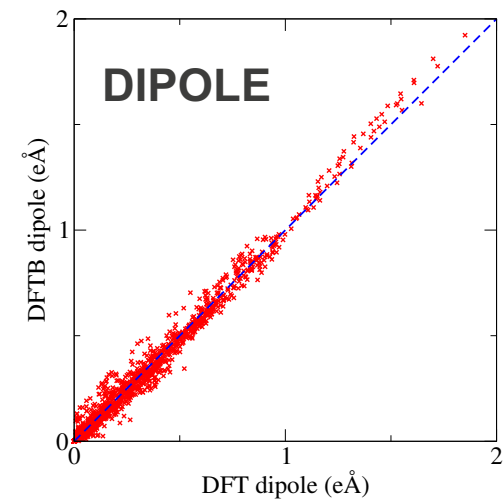
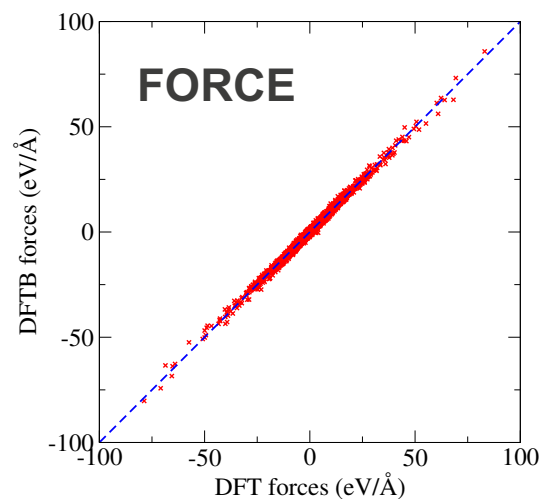
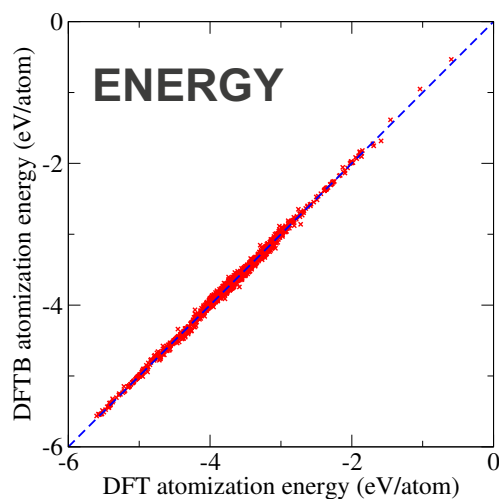
U-O-H: Carlson, Cawkwell, Batista, Yang, JCTC. 2020, **16**, 3073

Am-O-H: Taylor, Burrill, Cawkwell, Lubbers, Batista, Yang, *to be submitted*

# EPO: Parameterization of DFTB for f-Elements

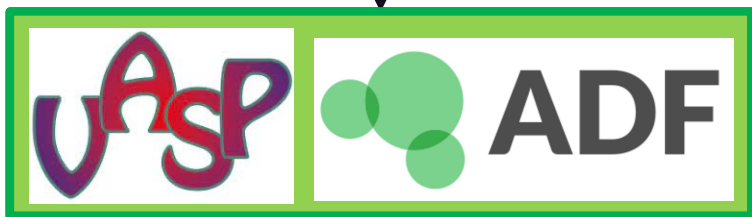


- The ‘EPO’ code iteratively improves a DFTB parameterization to minimize an objective function that measures the errors in the binding energy and forces vs. DFT.
- Terms needing parameterization are radial dependences of bond integrals and pair potentials, Hubbard U, and on-site energies.

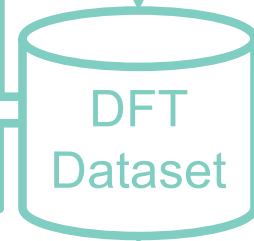
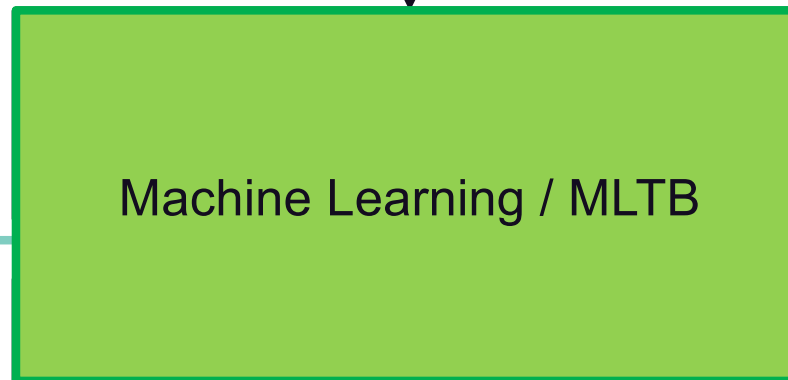
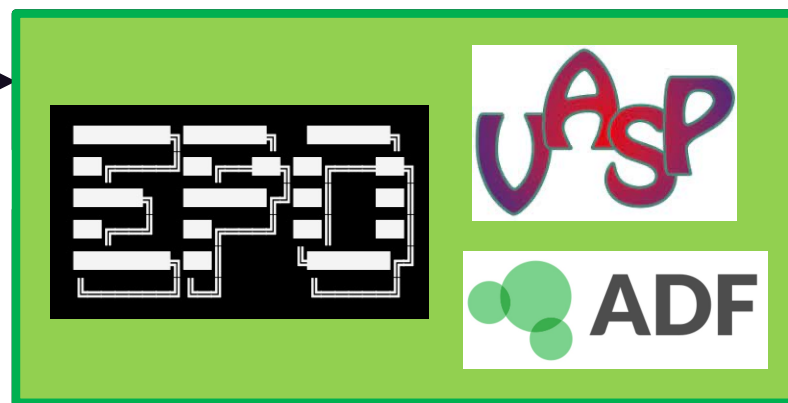


# DFTB Training Workflow

Initial Steps:



Optimization Loop :

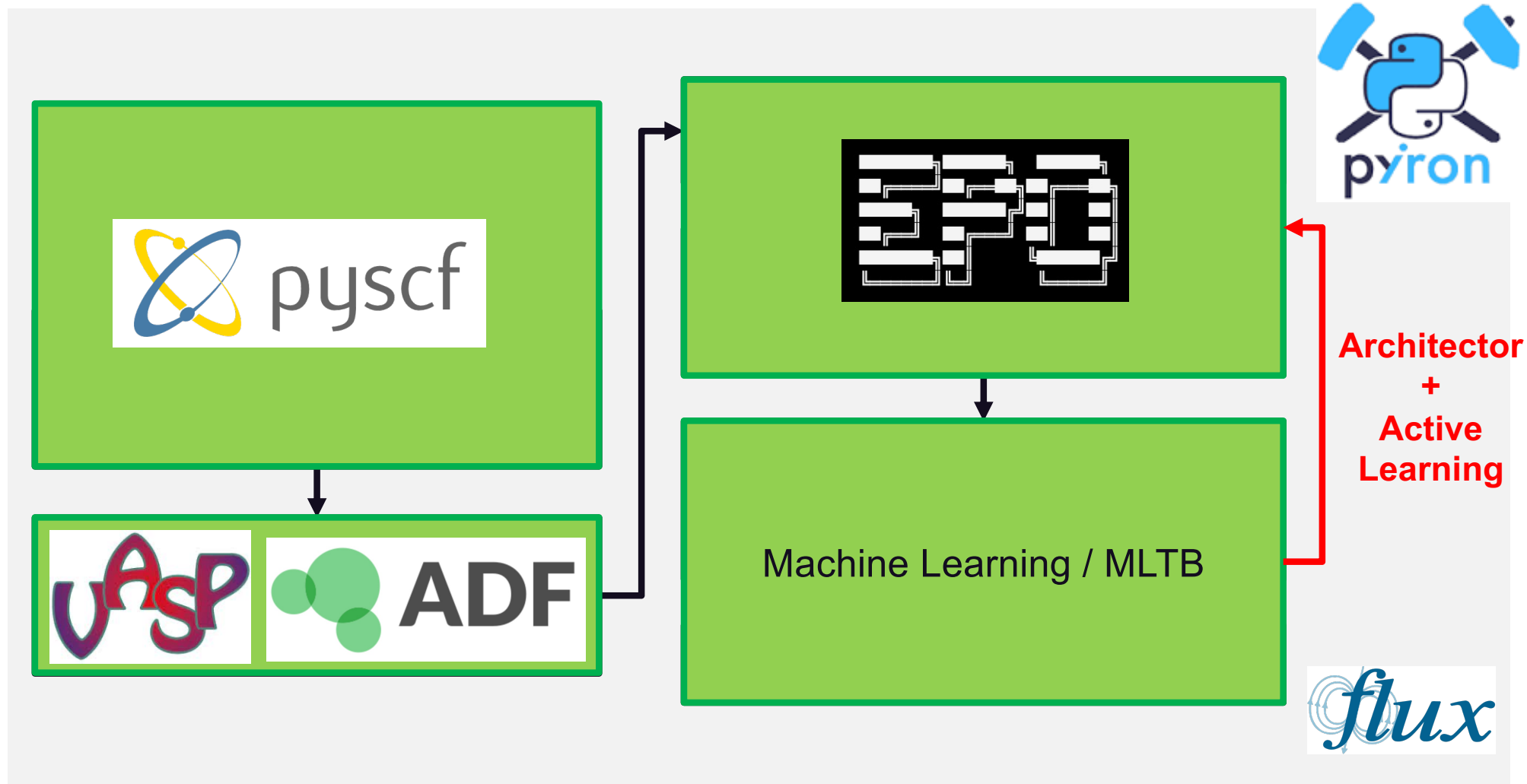


no  
MLTB  
converged?  
yes

- Traditional method requires extensive user knowledge and careful training
- Manual effort required for each black/teal arrow



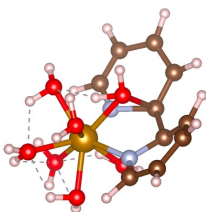
# DFTB Training Workflow – Automation Improvements!



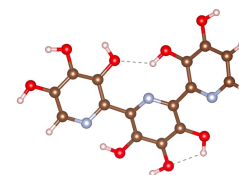
- Extensive Workflow Management Development is underway.

# High-throughput Quantum Simulations

**Architector**, Generate chemically-sensible extractant-actinide complexes



**Vulcan**, Suggest new extractants factoring in synthesizability, pH, solubility ...



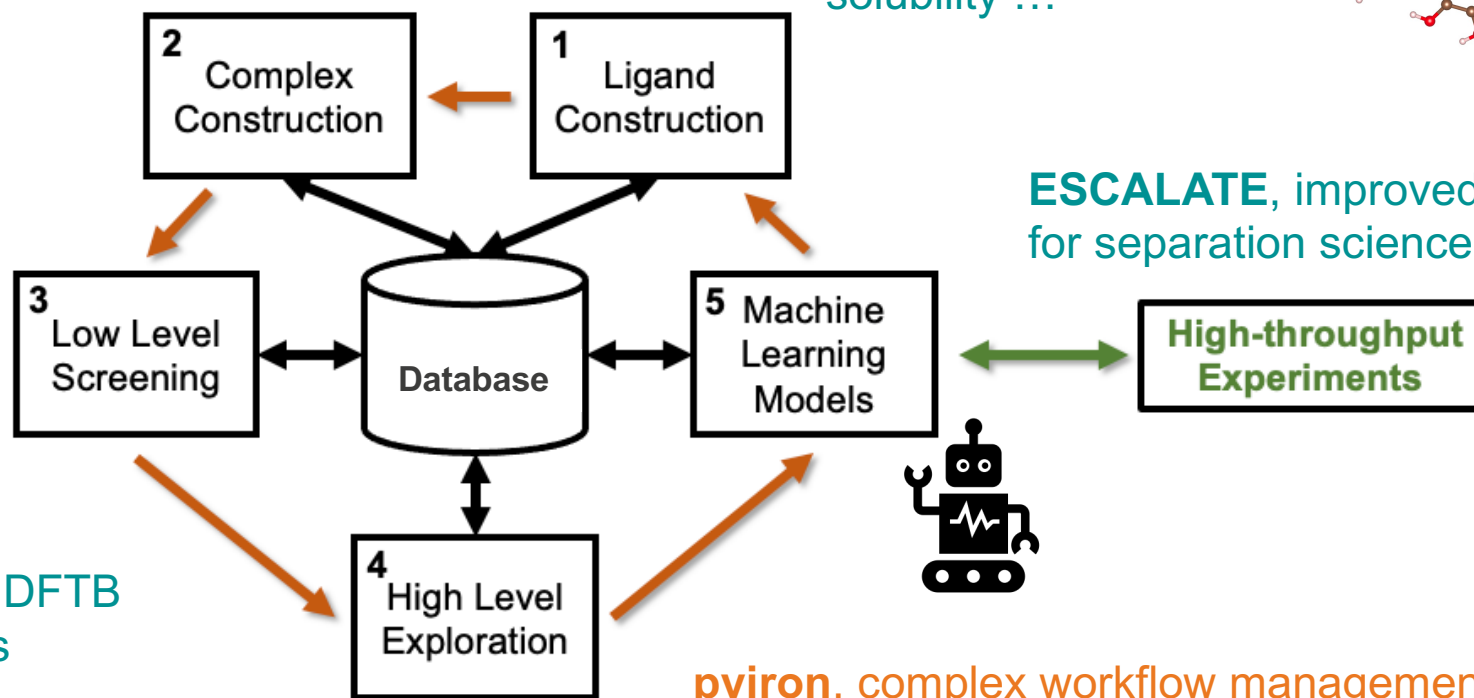
**Mineva**, calculate topological chemical descriptors

**EPO**, train DFTB parameters

**ADF/Gaussian**, DFT calculations

**pyiron**, complex workflow management package for all-level calculations simultaneously using HPC to enable automatous discovery

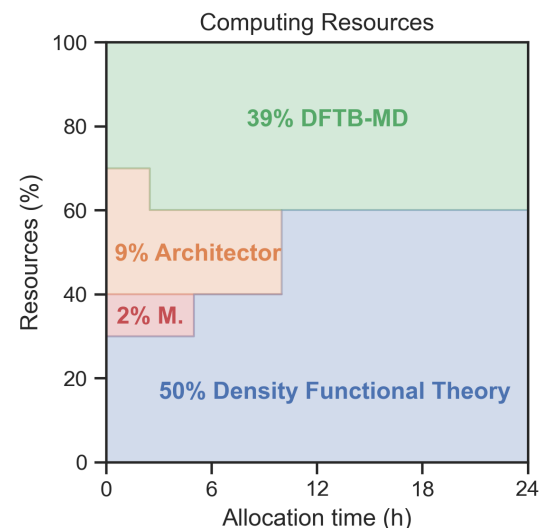
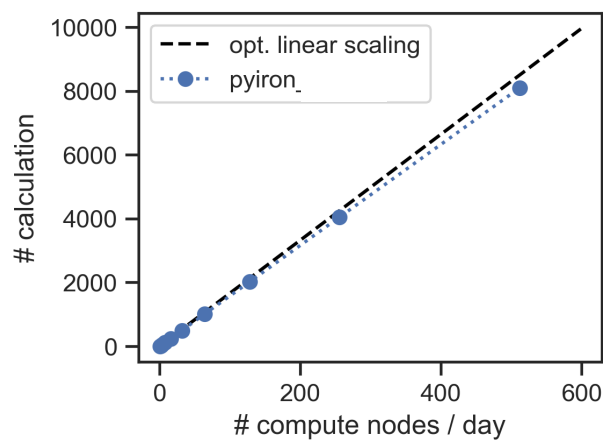
**ESCALATE**, improved for separation science



# Efficient Workflow Management for High-throughput Chemistry Simulations

## Enable up-scaling of high-throughput simulations for chemistry

- Developed dynamically-scaling workflow
- Provides simulation protocol provenance
- New interfaces to chemistry codes: ADF, xTB, Gaussian, etc.

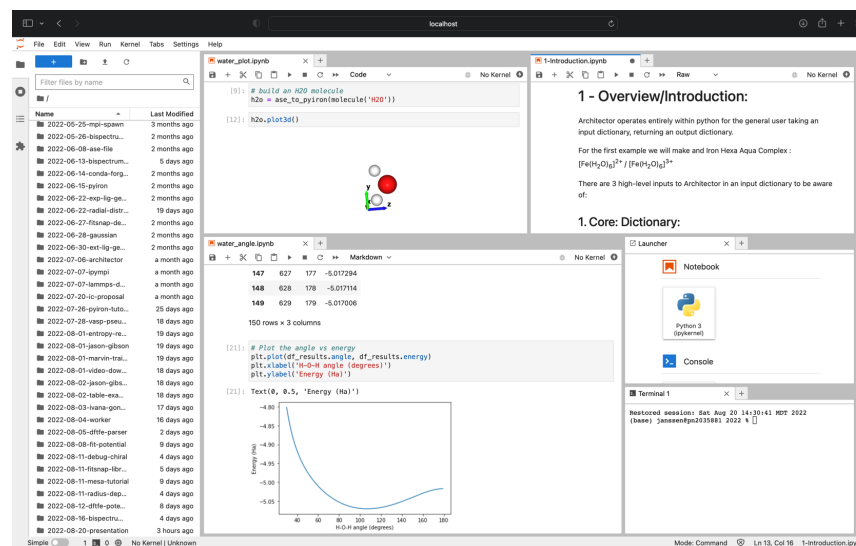


## Features:

- Highly scalable
- User-friendly interface
- Independent of HPC infrastructure



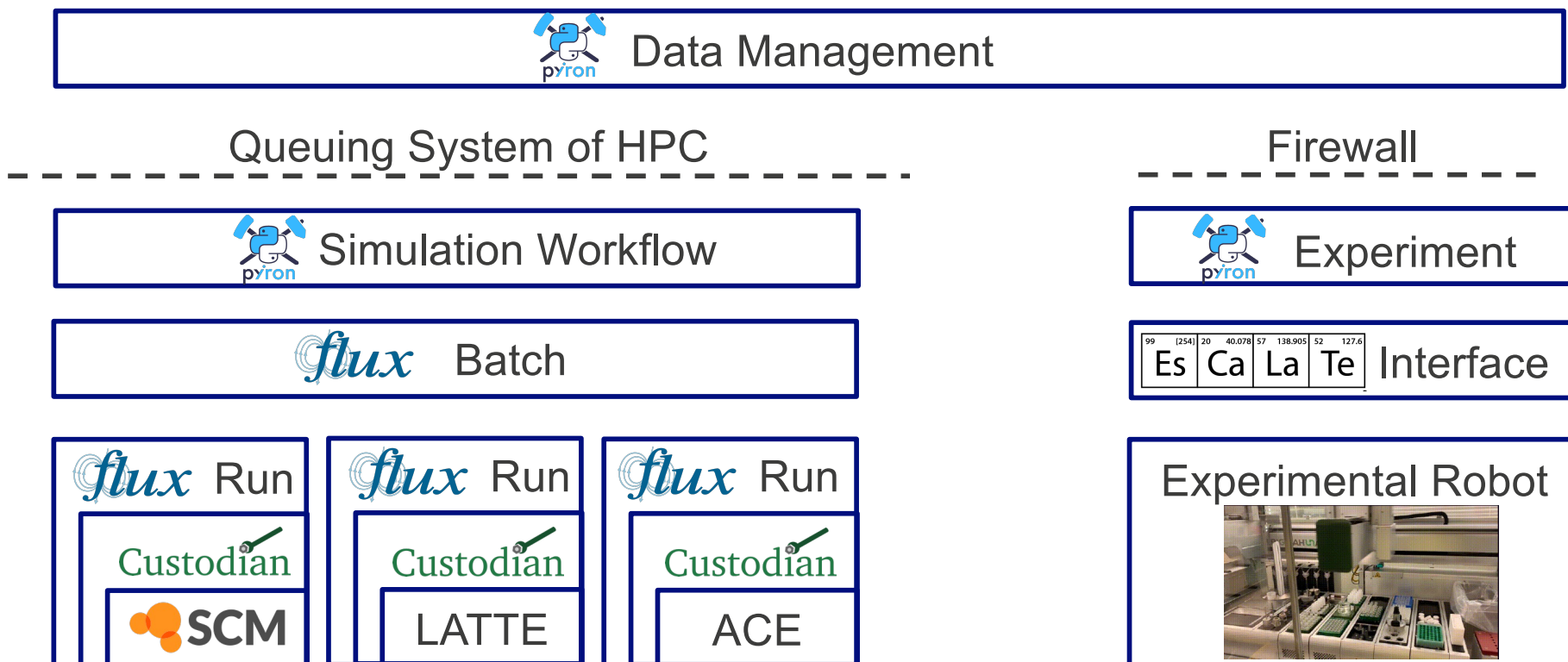
J. Janssen, et al., Comp. Mat. Sci.  
161 (2019) <http://pyiron.org>





# High-throughput Chemistry Simulations at Scale

Construct a complex *hierarchical workflows* with the building blocks you've heard about in this workshop:



Promising solution coming from this NME IPAM long program!

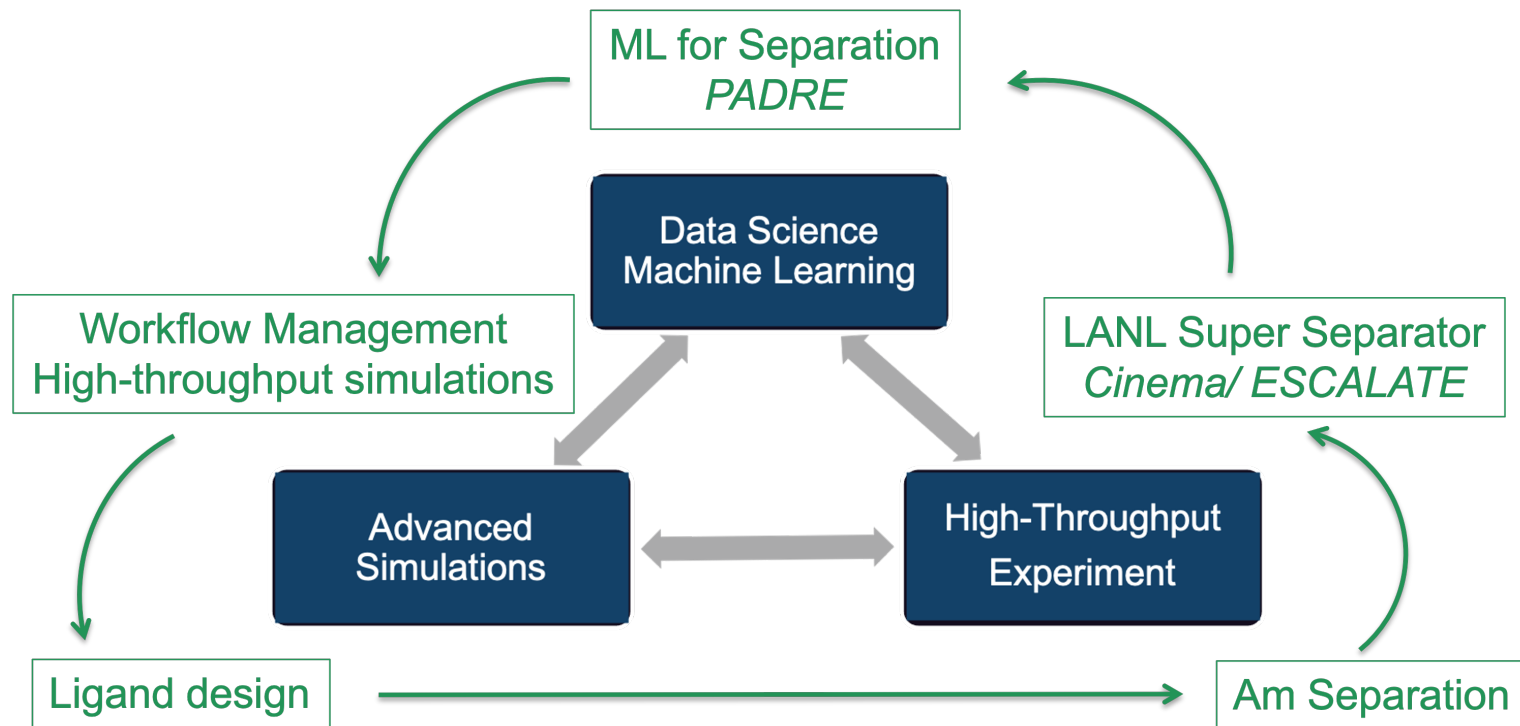
Michael Taylor  
Jan Janssen  
Danny Perez

# Exascale Future: What is needed?

- Exascale needs:
  - Integration/Centralization of databases across multiple compute resources
  - Strategies for handling calculation/experimental failures.
  - Prioritization strategies for “promising” chemistries (e.g. identify and prioritize chemistries on the fly)
  - Metrics for quality of computational predictions vs. experimental validations.
  - **Hierarchical workflows** need to be interoperable, adaptive, robust such as pyiron+flux



# Summary



**Robust and interoperative hierarchical workflow management is the key to accelerate discovery of separation science.**

# Acknowledgements

## Team :

**Sara Adelman**, Brian Arko, Enrique Batista, Nathan Bessen, Daniel Burrill, **Marc Cawkwell**, Wenhao Gao, Zach Jones, **Stosh Kozimor**, **Jan Janssen**, Jiyoung Lee, Chang Liu, Nick Lubbers, **Danny Perez**, **Josh Schrier**, Benjamin Stein, **Michael G Taylor**, **Michael Tynes**, Yufei Wang, Xiaobin Zhang

## Collaborators:

Prof. Jen Shafer, Colorado School of Mines  
Prof. Graeme Henkelman, University of Texas Austin

## Funding:

DOE BES Separation Program  
DOE Nuclear Energy Program  
LANL Seaborg GRA and Postdoc Fellowships  
LANL CNLS GRA and Postdoc Fellowships  
LANL ISTI GRA and AML Summer School

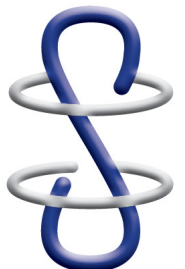


U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science

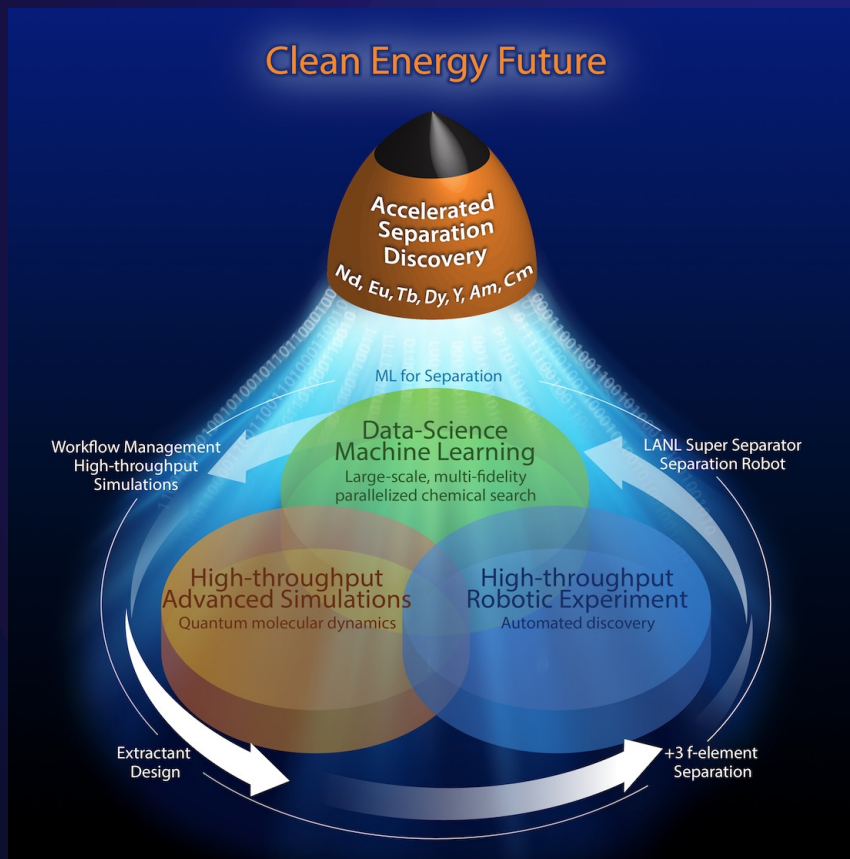
Office of  
**NUCLEAR ENERGY**

## Facilities:



# Thank you!

[pyang@lanl.gov](mailto:pyang@lanl.gov)



Delivering science and technology  
to protect our nation  
and promote world stability

LA-UR# 22-21082

