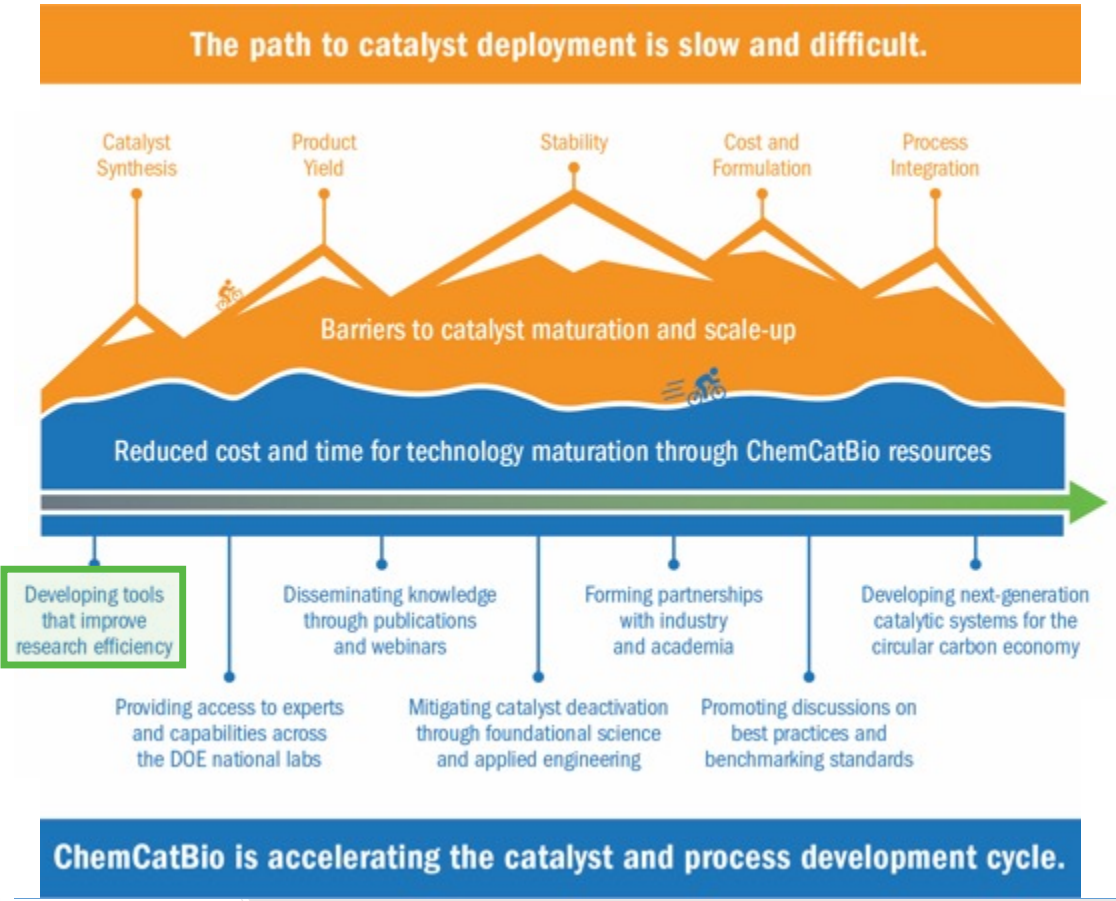
ChemCatBio
Chemical Catalysis for Bioenergy

DOE BETO 2023 Project Peer Review
ChemCatBio Data Hub
2.6.2.500

April 6, 2023
Catalytic Upgrading Session
Frederick Baddour
NREL

Project Overview – Accelerating Catalyst Discovery



eliable, directly comparable datasets
) are difficult to find

The Consequence: Data used in catalyst discovery is **often collected/computed from scratch**

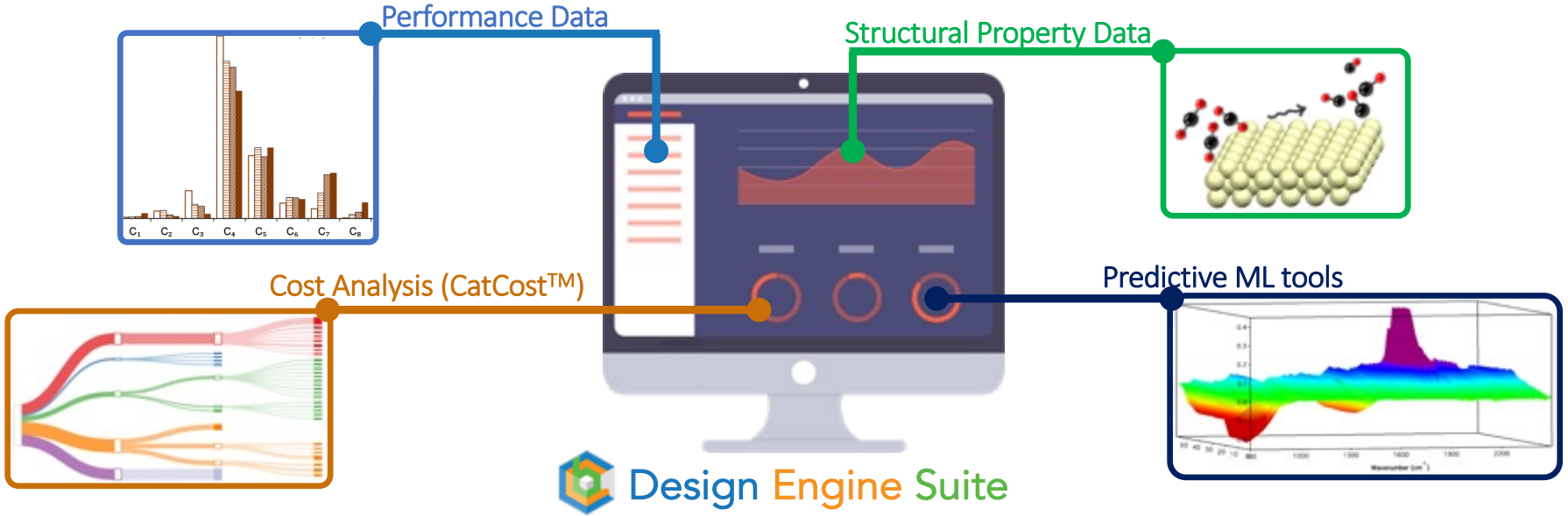
The Result: **Redundant calculations and**
s are performed repeatedly, with
l subset of published data entering
public domain

- ***Wasted time***
- ***Wasted resources***

The Goal: **To harness and curate** data to accelerate catalyst discovery

Project Overview: From Data Hub to a Catalyst Design Engine

To support and accelerate catalysis RD&D by addressing barriers with a suite of predictive analytical tools



*Integrating database technology from **Data Hub**, cost estimation from **CatCost** at the **frontier of machine learning** to transform catalyst design and deployment*

Project Overview – The Data Hub

The Data Hub

- A requirement as an EMN consortium
- Envisioned as a data sharing tool and collaboration framework
- Began as a repository for scientific data generated in the ChemCatBio consortium
- Designed with advanced tools and visualization capabilities
- Enables storage of public and private datasets – curated by ChemCatBio

Data Hub – A framework for transformational tools enabling catalyst R&D

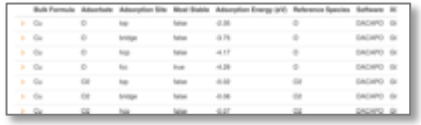
2018 Launch

CatCost

A free catalyst cost estimator

2018 Launch

The Catalyst Property Database (CPD)



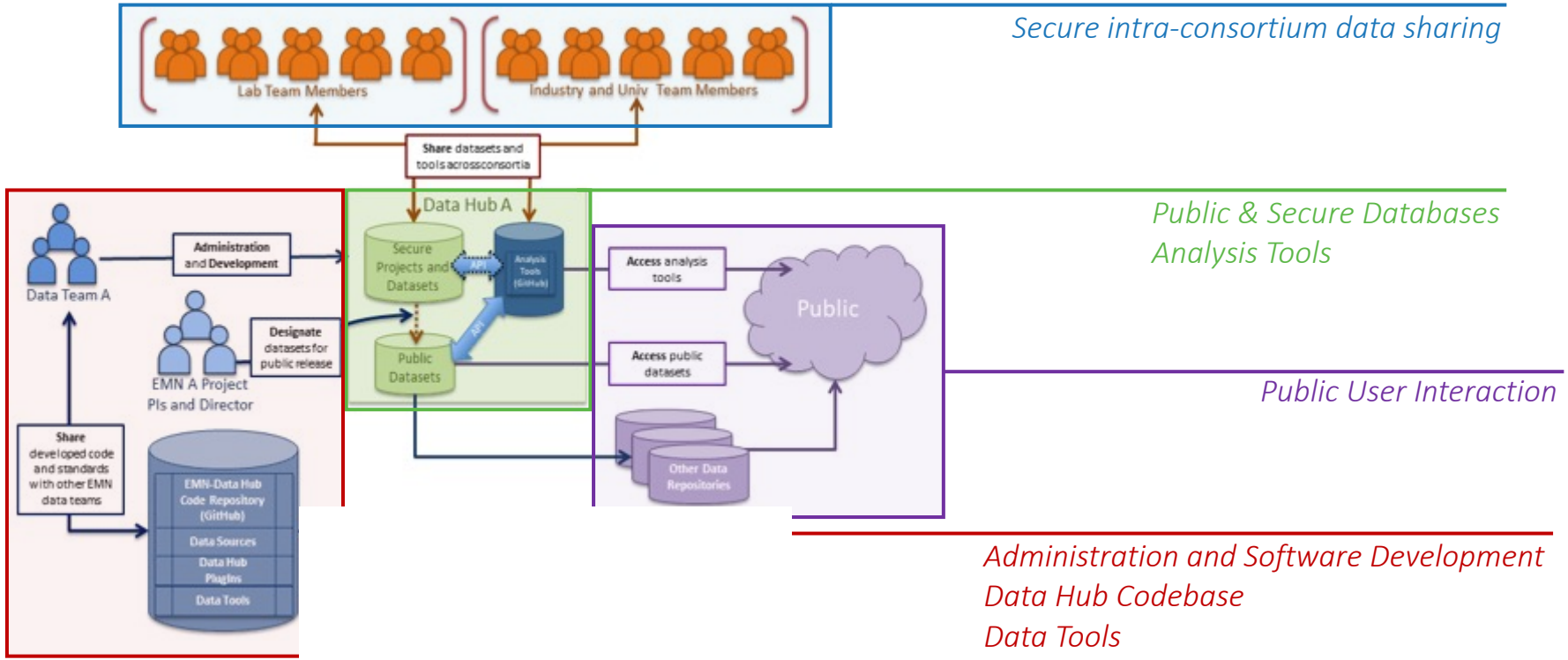
2020 Launch

datahub.chemcatbio.org

Public Release in 2018

The Data Hub is harnessing data to accelerate catalyst discovery

Project Overview – The Data Hub Framework



R. R. White, K. Munch, N. Wunder, N. Guba, C. Sivaraman, K. M. Van Allsburg, H. Dinh, C. Pailing, *Int. J. Adv. Comp. Sci. & App.* **2021**, *12*, 6, 657–667.

1 – Approach: Team and Collaborations

Management Plan: A project team with ***diverse, targeted expertise***

Fred Baddour, Ph.D.

PI

Experimentalist

CatCost R&D Lead

Carrie Farberow, Ph.D.

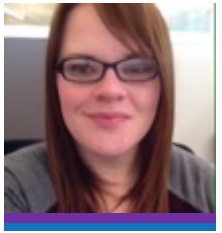
Technical expert

Computational researcher

Cody Wrasman, Ph.D.

Technical expert

Catalyst and process development



Rachel Hurst

Technical project lead, system architect

Nalinrat Guba, Ph.D.

Lead Developer

Software engineer (previously at Oracle)

Kurt Van Allsburg, Ph.D.

Former PI

Experimentalist

Experienced developer of R&D tools such as CatCost

Sean Tacey, Ph.D.

Technical expert

Computational researcher

Alicia Key

Software engineer with computational chemistry experience

Advisors: Josh Schaidle and Dan Ruddy (CCB Directors), Tom King (NREL UI designer), Nick Wunder (NREL web dev expert)

NREL Communications: Kathy Cisar, Erik Ringle

Project Tasks:

- *Task 1* – CPD Development: Computational Chemistry
- *Task 2* – CPD Development: Software & AI/ML Development
- *Task 3* – Data Hub Maintenance, Security, and Oversight

Project Collaborations:

- *CCPC – Atomistic Modeling Task* (PI: Carrie Farberow)
- *CatCost* (PI: Fred Baddour)
- *Catalyst Deactivation & Mitigation (*PI: Huamin Wang)

CCPC Consortium for Computational Physics and Chemistry U.S. DEPARTMENT OF ENERGY BIOENERGY TECHNOLOGIES OFFICE

CatCost

1 – Approach: Management Plan

Task 1 – Catalyst Property Database Development: Computational Chemistry ($175,000)

- Conceptualization of CPD features and applications, establish acceptance criteria, provide data test sets
- Identify new datasets for addition to CPD, develop data schema, python scripting for automating workflows
- External user interviews, user training & documentation, webinars/screencasts, quality control
- Define how AI/ML approaches should be implemented with Task 2

Task 2 – Catalyst Property Database Development: Software & AI/ML Development ($200,000)

- Implement new milestone-related features in CPD
- Bug fixes, usability improvements, UI enhancements in CPD and Data Hub
- Develop APIs and backend database improvements
- AI/ML method integration with database

Tracked & prioritized using Agile methods

Task 3 –Data Hub Maintenance, Security, and Oversight ($25,000)

- Security upgrades to Data Hub and Catalyst Property Database
- Managing site hosting with Amazon Web Services

1 – Approach: Risk Identification and Mitigation

Risk 1: Database Does Not Match User Needs (irrelevant / wrong features)

Mitigation: Go/No-Go milestone (FY21Q2) focuses on seeking expert / potential user feedback on development direction & pitfalls

Risk 2: Database Is Too Difficult To Use or Does Not Justify Required Effort

Mitigation: FY21 Go/No-Go sought feedback from software devs & UI experts, and ChemCatBio stakeholders. Drafted extensive documentation

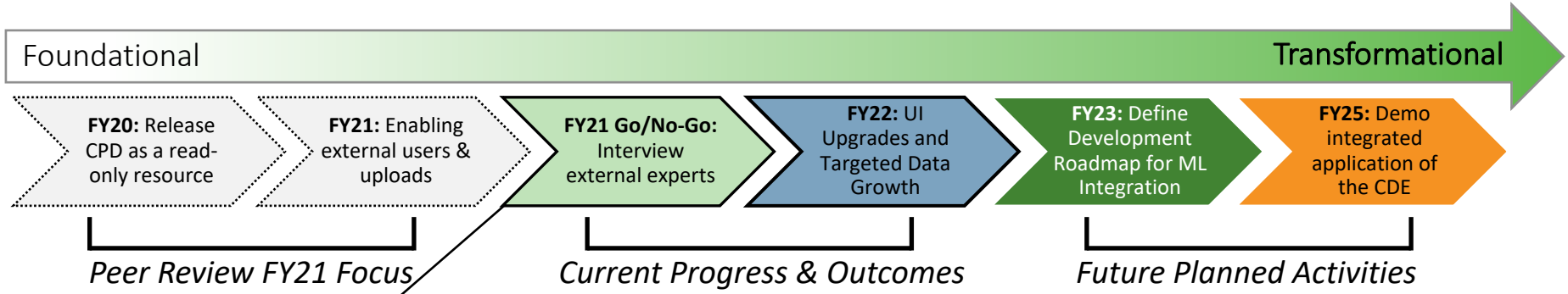
Risk 3: Data Quality/Quantity Issues

Mitigation: Established curation & training (FY21Q3). Leveraged partnerships with CCPC, CCB, external interviewees to gain buy-in & users

Risk 4: Data gaps in Catalyst Property Database

Mitigation: Go/No-Go FY24; Collab with CCPC modeling projects to identify/fill datagaps; leverage high-throughput computation & AI/ML approaches; utilize literature experimental data; collaborate with core technology projects

1 – Approach: Development Plan



FY21 Go/No-Go:

Interviewed 10+ experts to evaluate CPD development aligns with the needs of potential users and adjust as necessary.

Focus areas for interviews, to support CPD innovation:

- Preferred features
- Experience with competing solutions
- Strengths-Weaknesses-Opportunities-Threats
- Preferences for UI, scripting, etc.

Critical Milestone Efforts

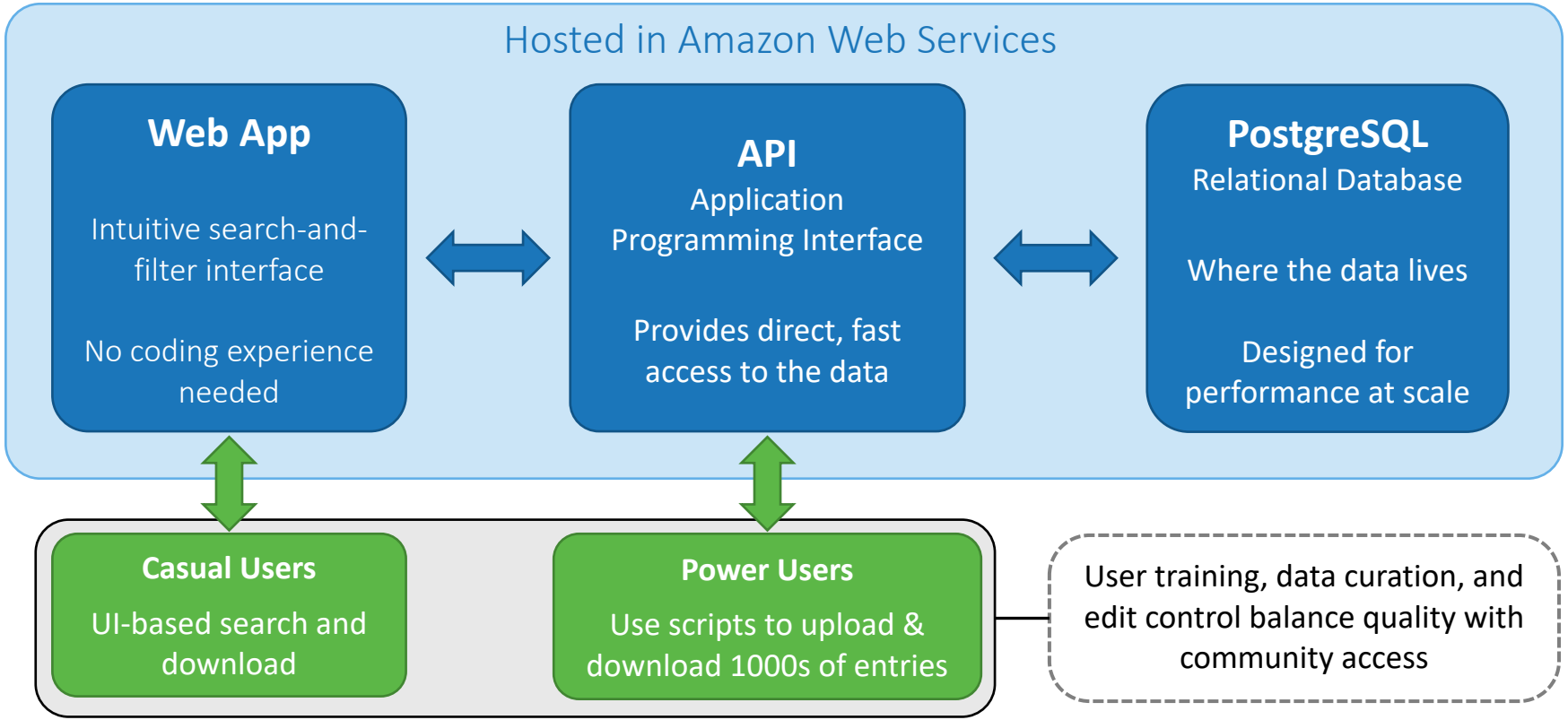
CPD Data Growth and Usability Enhancements

- FY21Q3/4: Documentation and batch upload capability
- FY22Q2: CPD UI upgrades address Go/No-go feedback
- FY22Q4: Catalyst deactivation mitigation resource
- FY23Q3: Release reference species interconversion feature

Implementation of Data Science for Predictive Capabilities

- FY23Q4: Define scope for AI/ML integrations
- FY24/25: Deliver and demonstrate an AI/ML model capability within the CPD

1 – Approach: Database Architecture



*Implemented modern design to **maintain a performant database as it scales***

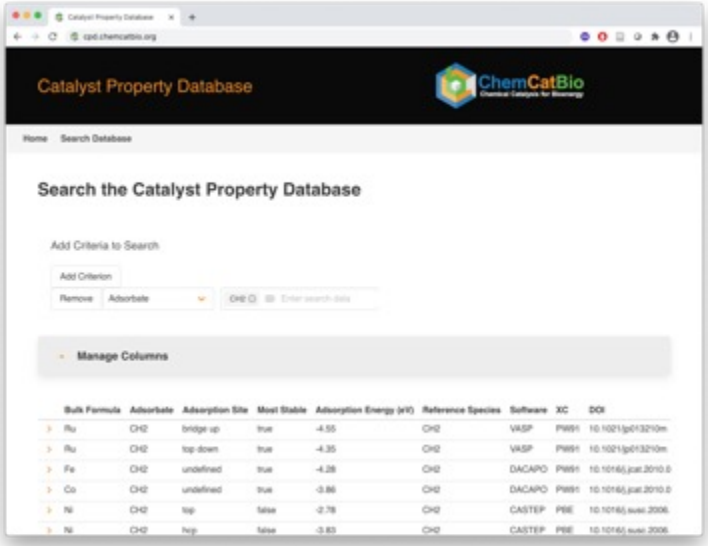
1 – Approach: The Catalyst Property Database

Since FY21, the Data Hub project is focused on developing the **Catalyst Property Database** as a ***collaboration and discovery tool***

The Catalyst Property Database (CPD)

- A centralized, searchable repository of catalyst properties
- Publicly accessible to view and upload

  Uploads subject to quality control
- Initial release: DFT-computed, published adsorption energies for intermediates on catalyst surfaces



cpd.chemcatbio.org

Free and public R&D resource

DFT = density functional theory

2 – Progress: Gather External User Feedback

Interviewed 10 experts spanning computational and experimental catalysis R&D for feedback on:

1. Utility of databases like CPD
2. Usability and gaps in existing data resources
3. Strengths-weaknesses-opportunities analysis of CPD
4. Preferred modes of interacting with a database

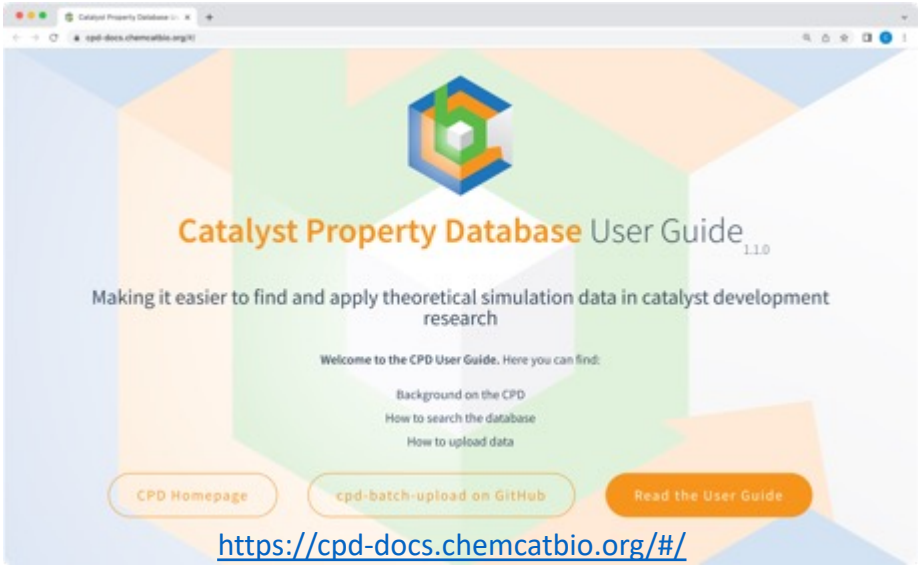
There's a lot of reinventing the wheel…Often, I'll find 3 groups have studied the same thing and published 3 different answers. Having a database for validation and benchmarking would be valuable.

I have previously scraped data from databases such as the NIST webbook, and having an API is really helpful.

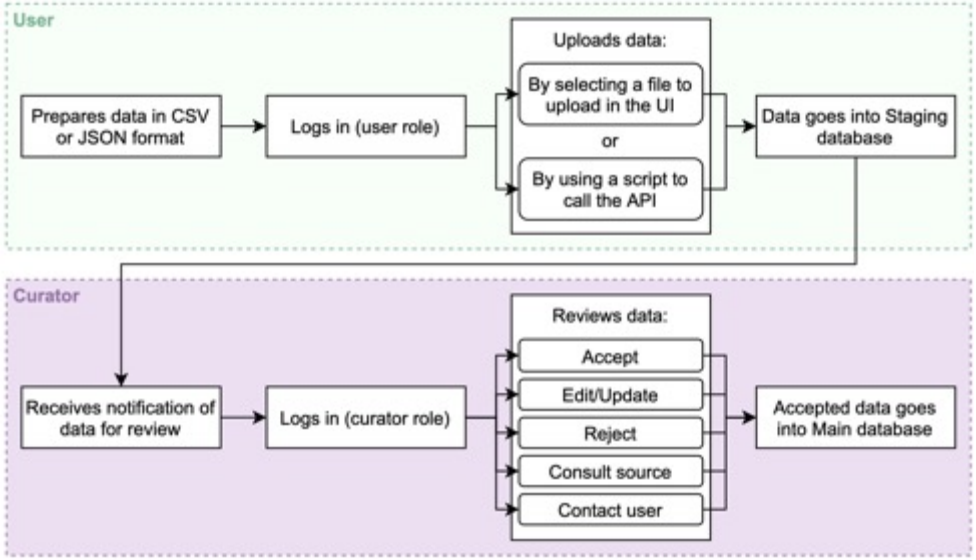
Key Takeaways and Lessons Learned:

- **Strengths Include:** Breadth of both data and sources; use of high-quality, published (i.e., vetted) data; accessibility to non-experts including experimentalists
- **Suggested Improvements:** UI improvements to make features more intuitive; CSV data export; documentation; batch data upload
- **Challenges:** Data growth; maintaining data quality
- General Enthusiasm regarding utility of Reference Species Interconversion

2 – Progress: Documentation, Training, and Data Curation



https://cpd-docs.chemcatbio.org/#/



Created public, wiki-style documentation website

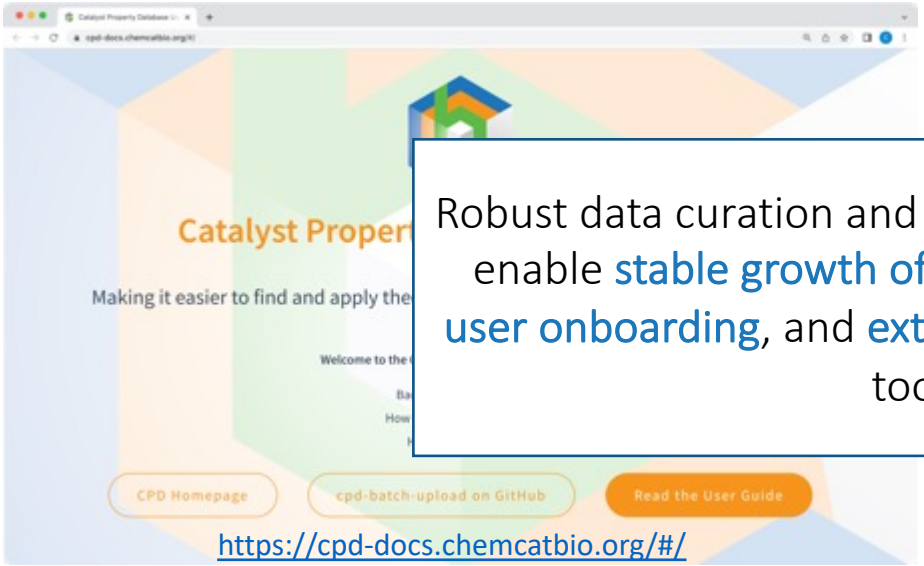
- Detailed guide for using and searching the CPD
- Easily expanded and updated

Public webinar produced to engage community
www.chemcatbio.org

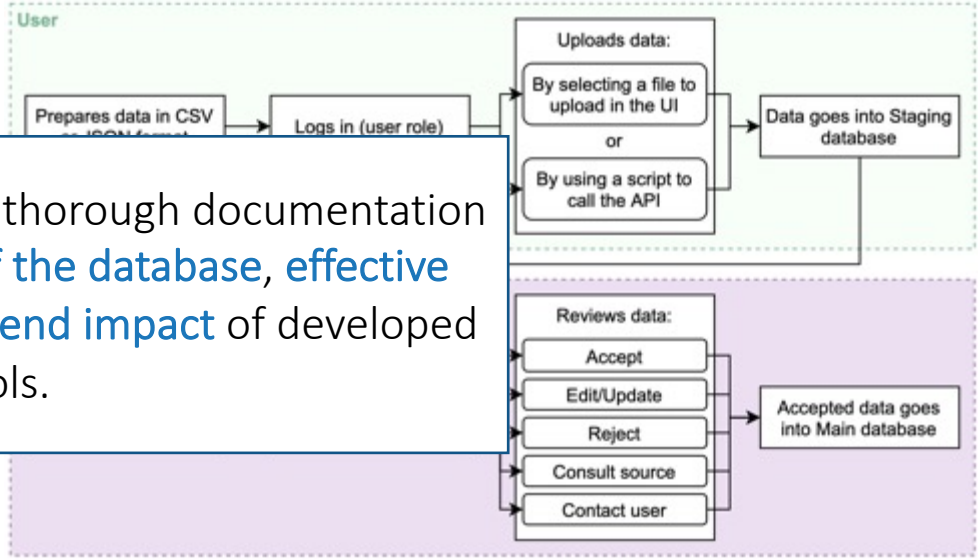
Developed data curation plan

- Defines steps and roles for user and curator
- Allows comparison with existing database data
- Possible future work: development of curation tools to accelerate process and identify data redundancies

2 – Progress: Documentation, Training, and Data Curation



https://cpd-docs.chemcatbio.org/#/



Robust data curation and thorough documentation enable stable growth of the database, effective user onboarding, and extend impact of developed tools.

Created public, wiki-style documentation website

- Detailed guide for using and searching the CPD
- Easily expanded and updated

Public webinar produced to engage community
www.chemcatbio.org

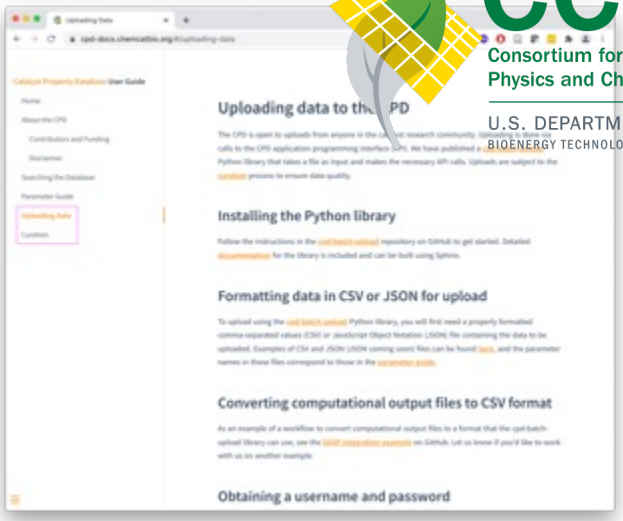
Developed data curation plan

- Defines steps and roles for user and curator
- Allows comparison with existing database data
- Possible future work: development of curation tools to accelerate process and identify data redundancies

2 – Progress: Established Batch Upload Workflow

- Released complete workflow, ‘cpd-batch-upload’ Python library for uploading datasets to the CPD: https://github.com/NREL/cpd-batch-upload
- In collaboration with the CCPC, released a computational chemistry software integration example
  - Directly converts data in output files for the Vienna Ab initio Simulation Package (VASP) to CPD-compatible CSV file that seamlessly integrates with the cpd-batch-upload Python library
  - Upload > 100 new adsorption energy entries in < 1 h
  - Possible future work: expand to include other software

- Added user authentication to the API upload feature to map new data additions to users
- Corresponding documentation added to the CPD User Guide
- Google analytics enabled for tracking external user site visits



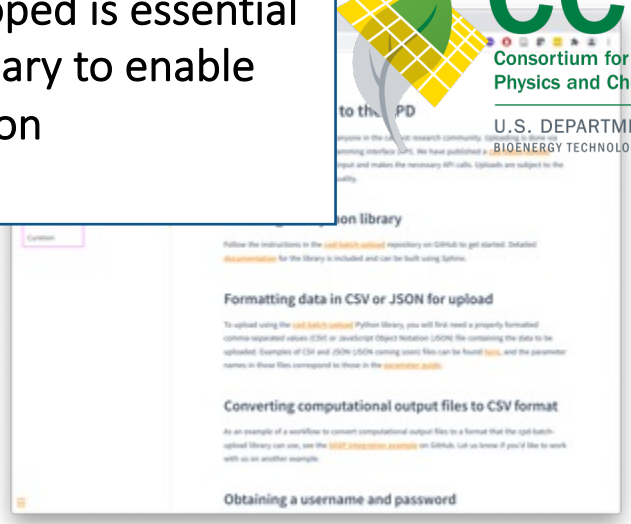
CCPC
Consortium for Computational Physics and Chemistry
U.S. DEPARTMENT OF ENERGY
BIOENERGY TECHNOLOGIES OFFICE

2 – Progress: Established Batch Upload Workflow

- Released complete workflow, 'cpd-batch-upload' Python library for uploading datasets to the CPD: https://github.com/NREL/cpd-batch-upload
- In collaboration with the CCPC, released a computational chemistry software integration example
  - Directly converts data in output files for the Vienna Ab initio Simulation Package (VASP) to CPD-compatible CSV file that seamlessly integrates with the cpd-batch-upload Python library
  - Upload > 100
  - Possible futu
- Added user au
  feature to ma
- Corresponding documentation added to the CPD User Guide
- Google analytics enabled for tracking external user site visits

Bulk data upload workflow developed is essential to ingest data quantities necessary to enable AI/ML implementation

CCPC
Consortium for Computational Physics and Chemistry
U.S. DEPARTMENT OF ENERGY
BIOENERGY TECHNOLOGIES OFFICE



2 – Progress: Data Growth and UI Updates

Piloted use of the open-source natural language processing (NLP) tool, ASReview (https://asreview.nl/) to prioritize journal articles for data mining

- Training set: 100 articles [79 relevant, 21 irrelevant]
- Test set: 994 articles identified via keyword search
- Results: ASR accurately predicted quality of articles for data-mining reducing time in required to manually identify the most relevant articles

FY20 – 22: >300% increase in quantity of data

Created catalyst deactivation mitigation resource dataset in collaboration with the CCPC

- Generated 1000+ data points, via high-throughput calculations using NREL's HPC resource, describing catalyst binding of poisons common to biomass/waste conversion process

User Interface Updates

- Visual updates to modernize interface
- Filters, pre-populated with editable default values
- Downloadable results in JSON format
- Improved search/filter options

ASReview Results

Top 15

Rank	#Datapoints
1	160
2	36
3	48
4	92
5	8
6	43
7	20
8	8
9	12
10	8
11	0
12	0
13	100
14	12
15	29
False positives: 2/15	

Bottom 15

Rank	#Datapoints
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
False negatives: 0/15	

Data Quantity

	Median # Datapoints	Average # Datapoints
CPD Existing Dataset	27	38
ASReview Top 15	29	44

2 – Progress: Data Growth and UI Updates

Piloted use of the open-source natural language processing (NLP) tool, ASReview (https://asreview.nl/) to prioritize journal articles for data mining — + Database Scale

- Training set: 100 articles [79 relevant, 21 irrelevant]
- Test set: 994 articles identified via keyword search
- Results: ASR accurately predicted quality of articles for data-mining reducing time in required to manually identify the most relevant articles

FY20 – 22: >300% increase in quantity of data

Created catalyst deactivation mitigation resource dataset in collaboration with the CCPC — + Database Scope

- Generated 1000+ data points, via high-throughput calculations using NREL's HPC resource, describing catalyst binding of poisons common to biomass/waste conversion process

User Interface Updates — + Database Utility

- Visual updates to modernize interface
- Filters, pre-populated with editable default values
- Downloadable results in JSON format
- Improved search/filter options

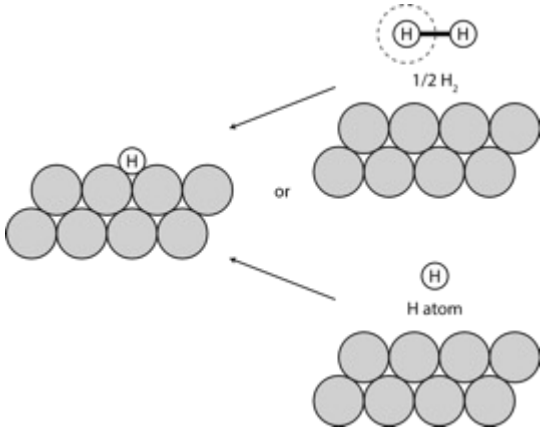
Database expansion efforts identified as essential to (1) link with ***experimental results*** and (2) ***leverage AI/ML tools***

2 – Progress: Reference Species Interconversion

Challenge for utilization of CPD data for benchmarking and big data applications: Computed adsorption energies, the critical DFT output included in the Catalyst Property Database, may be reported with different reference species

Adsorption energy (E_i^*, in eV; 1 eV = 96.5 kJ/mol) for atomic H on a Pt(111) surface calculated with different gas-phase references.

Gas-phase reference	E_i^* (eV)	|Difference| (eV)
H	–2.80	2.26
½H2	–0.54	

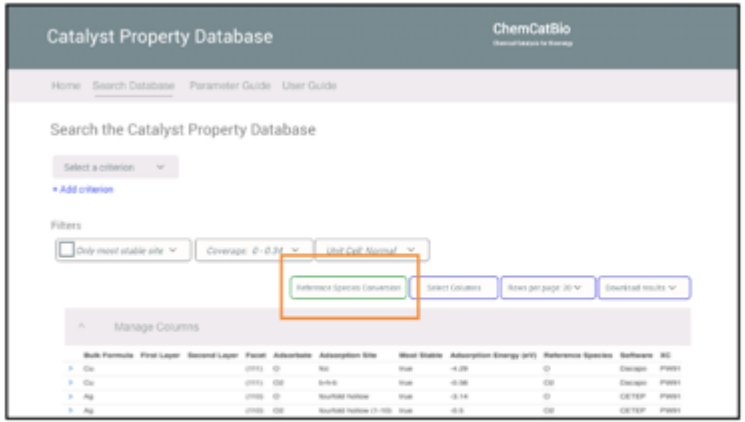


CPD Solution: Create a Reference Species Interconversion feature to enable interconversion between compatible reference species sets.

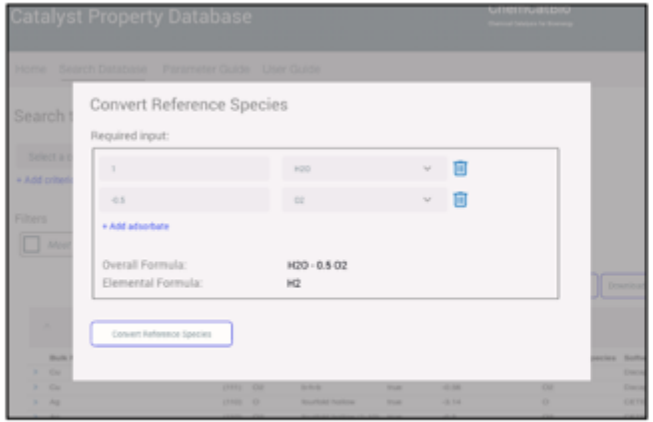
This is a key differentiator not found in any public database or resource.

2 – Progress: Reference Species Interconversion

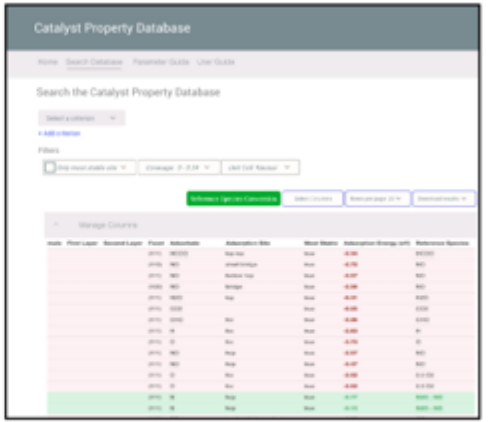
Wireframes developed depict how the RSI feature will be implemented



RSI Button



RSI Modal Pop-up



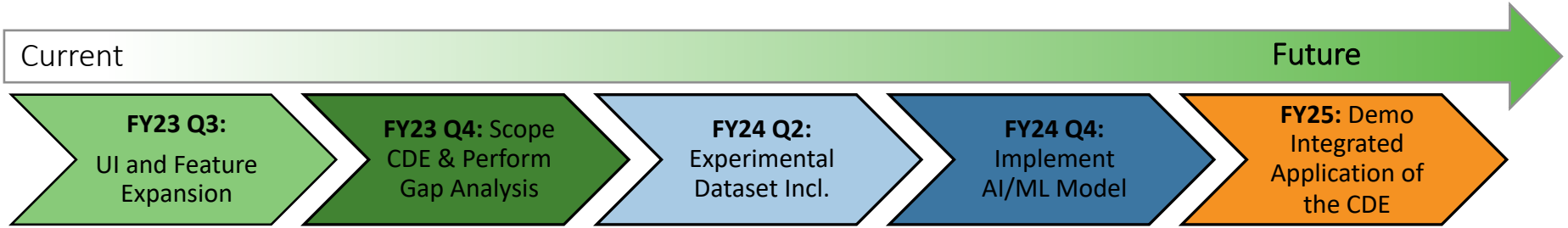
Results Table

Step 1: User filters data set for reference species interconversion and clicks ‘Reference Species Interconversion’ button

Step 2: Modal pop-up allows user to input desired reference species and provides relevant formula

Step 3: Results table displays converted adsorption energy values in green and values that could not be converted in red

2 – Progress: Future Feature Development and Demonstration



Development Roadmap & Applied Demonstration of Integrated Catalyst Design Engine:

FY23Q3: User Interface Improvements and Functionality Expansion

FY23Q4: Define Scope and Conduct Feature Gap Analysis for Catalyst Design Engine Vision

FY24Q2: Implement Capability to Include Experimental Catalytic Data in CPD

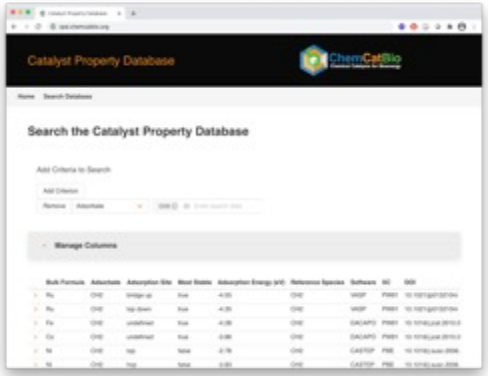
FY24Q4: Implement and Validate AI/ML Model Utilizing the CPD

FY25Q4 (*End of Project*): Demonstrate the Integrated Application of the Catalyst Design Engine to a Technology Challenge within ChemCatBio

3 – Impact Diverse R&D Applications of the CPD

The Catalyst Property Database has diverse R&D applications that are expanding as the database grows:

cpd.chemcatbio.org



10–1,000 entries

Computational Researcher
Data Validation/
Benchmarking

1,000–100,000 entries

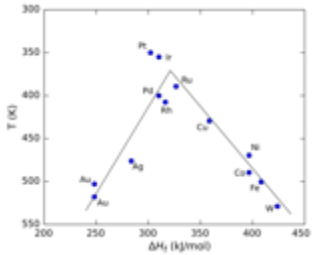
High-Throughput Calculations

1,000–100,000 entries

CCPC
Consortium for Computational Physics and Chemistry
U.S. DEPARTMENT OF ENERGY
BIOENERGY TECHNOLOGIES OFFICE

1M+ entries

Catalysis ML



Reactivity Descriptor Discovery

3 – Impact: Faster & Cheaper Catalyst Discovery

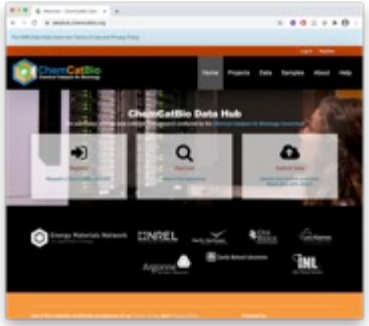
Every year, **more experimental and computational catalyst data is generated**, but the methods and **tools to locate, organize, and apply this data have not kept up**.

The CPD is advancing the state of the art for application of computational data:

- Stop spending time and money re-creating data that already existed
- Enable new approaches to catalyst discovery that require large datasets

By harnessing the power of data, the Data Hub and the Catalyst Property Database are accelerating catalyst discovery

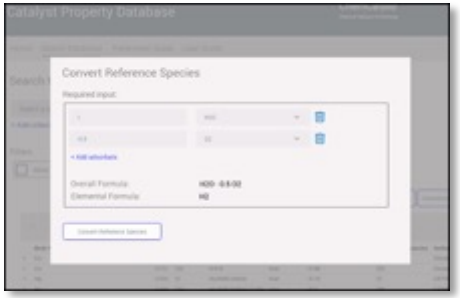
Two Publicly Available Tools Released to Accelerate Catalyst R&D



datahub.chemcatbio.org

cpd.chemcatbio.org

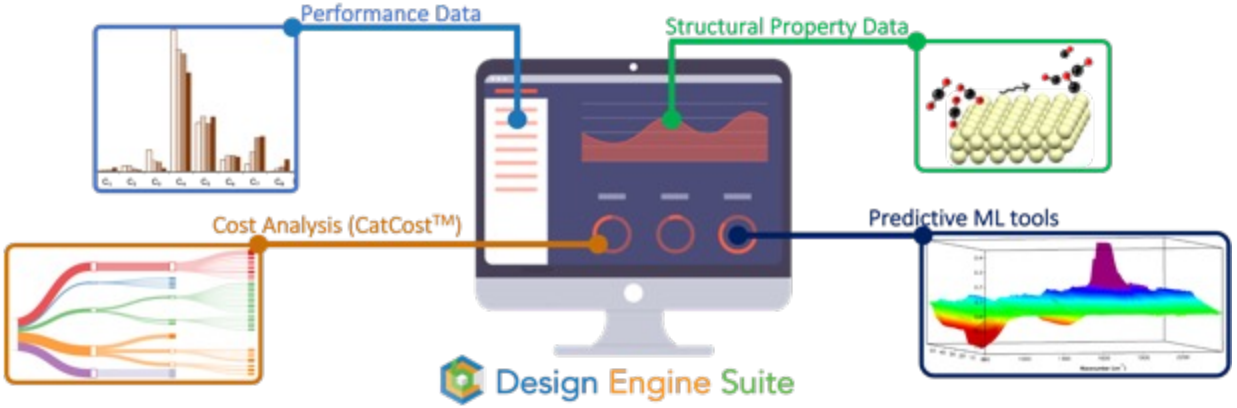
UI Improvements, data import scripts, RSI tool, and thorough documentation development



enhance usability, user onboarding, and impact

Summary

Vision: Integrating database technology from **Data Hub**, cost estimation from **CatCost** at the **frontier of machine learning** to transform catalyst design and deployment



Approach: Focused development on the CPD as a ***collaboration and discovery tool***, while building out database growth, curation, and validation capabilities

Outcomes: Performed extensive user feedback campaign and integrated into a development roadmap for FY23–FY24. Expanded database and functionality, ***focused on key differentiators***

Impact: ***Accelerating the Catalyst R&D cycle*** through targeted data curation and layering analysis tools onto the state-of-the-art Data Hub Database

Quad Chart Overview

Timeline

- ***Project Start: 10/1/2022***
- ***Project End: 09/30/2025***

	FY22 Costed	Total Award FY23-FY25
DOE Funding	*(10/01/2021 – 9/30/2022)* *$372,932*	*$400k/y Budget Authority* *$1.2 M*
Project Cost Share	N/A	N/A

TRL at Project Start: 2
TRL at Project End: 4

Project Goal

Enable ChemCatBio and the bioenergy industry to accelerate the catalyst and process development cycle through development of publicly available advanced analytics tools. Develop the CPD as a significant, respected resource accelerating catalysis R&D. Demonstrate the Catalyst Design Engine vision of predictive catalyst design using theoretical and experimental data as well as cost information.

End of Project Milestone

Demonstrate an integrated application of the Catalyst design Engine. Us AI/ML, including at least two distinct data sets (e.g., theoretical adsorption energies, experimental conversion/selectivity) within the Catalyst Property Database, to address a current technology challenge in biomass/waste conversion process development (e.g., mitigate deactivation, reduce process cost, reduce process severity, process intensification). Collaborate with at least one ChemCatBio project to obtain experimental data to validate the model predictions and summarize lessons learned to inform future applications of the CDE.

Funding Mechanism

BETO FY23 National Laboratory Call - Subtopic 2c

Project Partners

- **N/A**

This research was supported by the DOE Bioenergy Technology Office under Contract no. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory

This work was performed in collaboration with the Chemical Catalysis for Bioenergy Consortium (ChemCatBio, CCB), a member of the Energy Materials Network (EMN)

U.S. DEPARTMENT OF ENERGY | Office of ENERGY EFFICIENCY & RENEWABLE ENERGY
BIOENERGY TECHNOLOGIES OFFICE

Data Hub Team

Carrie Farberow
Nalinrat Guba
Sean Tacey
Matt Jankousky
Alicia Key
Rachel Hurst
Tom King
Qiyuan Wu
Tuong Bui
Cody Wrasman
Nick Wunder
Kris Munch
Courtney Pailing

Advisors

Kathy Cisar
Erik Ringle
Josh Schaidle
Dan Ruddy

BETO

Trevor Smith
Sonia Hammache
Jesse Glover
Andrea Bailey
Nicole Fitzgerald

ChemCatBio
Chemical Catalysis for Bioenergy
Energy Materials Network
U.S. Department of Energy

Data Hub: Informed by 2021 Peer Review

"Collecting data has been one of the most important tasks for developing artificial intelligence technologies for catalyst development. "

- Growing the data in CPD while maintaining data quality to effectively train ML models is a critical element of ongoing efforts within this project

"It could be beneficial if… this project uses natural language processing tools to automate or semi-automate the extraction of information from the literature

- The use of the open-source natural language processing (NLP) tool, ASReview, was piloted to prioritize journal articles for data mining

"This is important work and should include an education/training campaign in conjunction to realize the full impact"

- This is aligned well with fy22 efforts that included the development of documentation, training, a data curation strategy and public outreach via webinar

FY21Q2 Go/No-Go

Outcome: ”Go”

Confirm Computational Catalyst Property Database Development Direction Through External Outreach: Use feedback from external experts to confirm that Computational Catalyst Property Database development plans align with the needs of potential users and adjust if necessary.

Criteria – Interview 10+ experts and compile their feedback on:

1. Their preferred uses for databases like CCPD
2. Experience using and gaps left by competing solutions, such as Catalysis Hub
3. Strengths-weaknesses-opportunities-threats analysis of CCPD
4. Preferred modes of interacting with a tool like CCPD (e.g., coding languages, UI preferences)

High Level Take-Aways:

- **Strengths:** Breadth of both data and sources; use of high-quality, published (i.e., vetted) data; accessibility to non-experts including experimentalists
- **Suggested improvements:** UI improvements to make features more intuitive; CSV data export; documentation; batch data upload
- **Challenges:** Data growth; maintaining data quality
- *General enthusiasm regarding utility of Reference Species Interconversion*

Publications, Patents, Presentations, Awards, and Commercialization

Publications

- B. E. Petel, K. M. Van Allsburg, F. G. Baddour, “Cost-Responsive Optimization of Nickel Nanoparticle Synthesis” Advanced Sustainable Systems, **2023**, *accepted*.
- R. R. White, K. Munch, N. Wunder, N. Guba, C. Sivaraman, K. M. Van Allsburg, H. Dinh, C. Pailing, “Energy Material Network Data Hubs” *Int. J. Adv. Comp. Sci. & App.* **2021***, 12*, 657.
- S. A. Tacey, M. A. Arellano-Trevino, K. Van Allsburg, C. A. Farberow, “High-throughput screening of catalyst impurity adsorption for biomass upgrading applications” *Scientific Data* – (in prep.)

Software Record

- NREL SWR-21-47 “Catalyst Property Database” (Feb 4, 2021)

Presentations

- K. Van Allsburg, “Reduce, Reuse, Recycle: Data Benchmarking and Accessibility for Faster Research With the Catalyst Property Database” ChemCatBio Webinar (October 13, 2021)
- K. Van Allsburg, “Reduce, Reuse, Recycle: Data Benchmarking and Accessibility for Faster Research with the Catalyst Property Database” 27th North American Catalysis Society Meeting, New York, NY (May 2022)
- K. Van Allsburg, S. A. Tacey, and C. A. Farberow, “Accelerating Research with the Catalyst Property Database” Poster, 27th North American Catalysis Society Meeting, New York, NY (May 2022)