



3.1.1.009 Developing Hydrotreating Models using Machine Learning

March 23, 2021
System Development and Integration
2021 BETO Peer Review

Dr. Mariefel V. Olarte
Chemical Engineer IV
Pacific Northwest National Laboratory

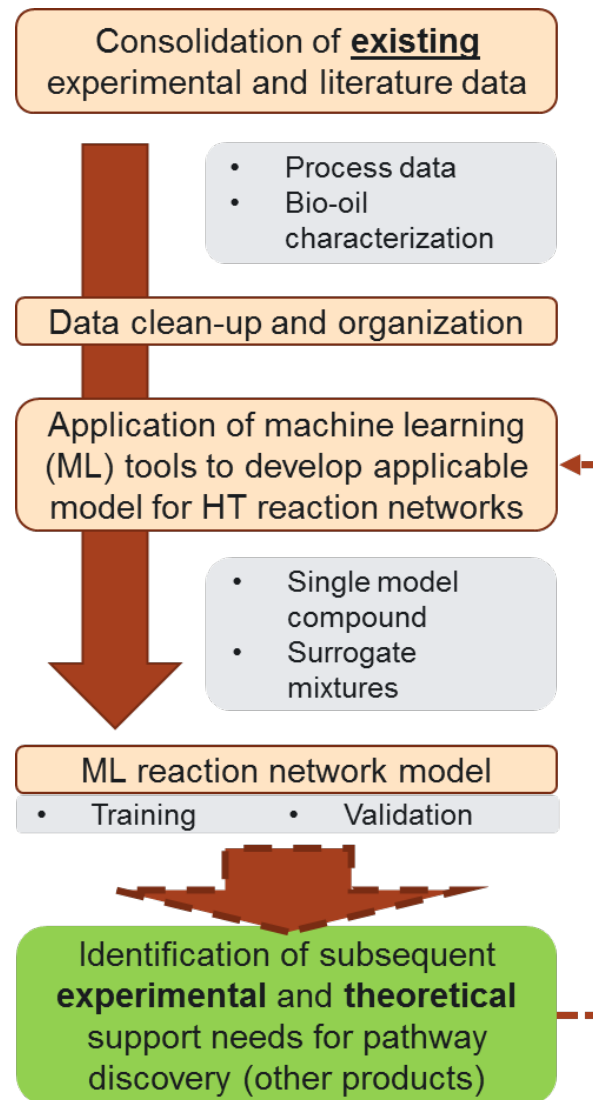


PNNL is operated by Battelle for the U.S. Department of Energy



Project Overview

BIG PICTURE: Create a framework to develop a hydrotreating (HT) reaction network using machine learning (ML) tools to model and predict expected HT conversions given specific bio-oil and biocrude inputs for use directly by researchers and new bio-oil upgrading companies, with potential in-house data-customization for more established companies



Hydrotreating

Target 3-year end goal: Develop **ML algorithms** that incorporate literature, computed and experimental data **that can predict upgraded oil fractions** based on feed chemical information and operating conditions such as temperature, pressure, and catalyst

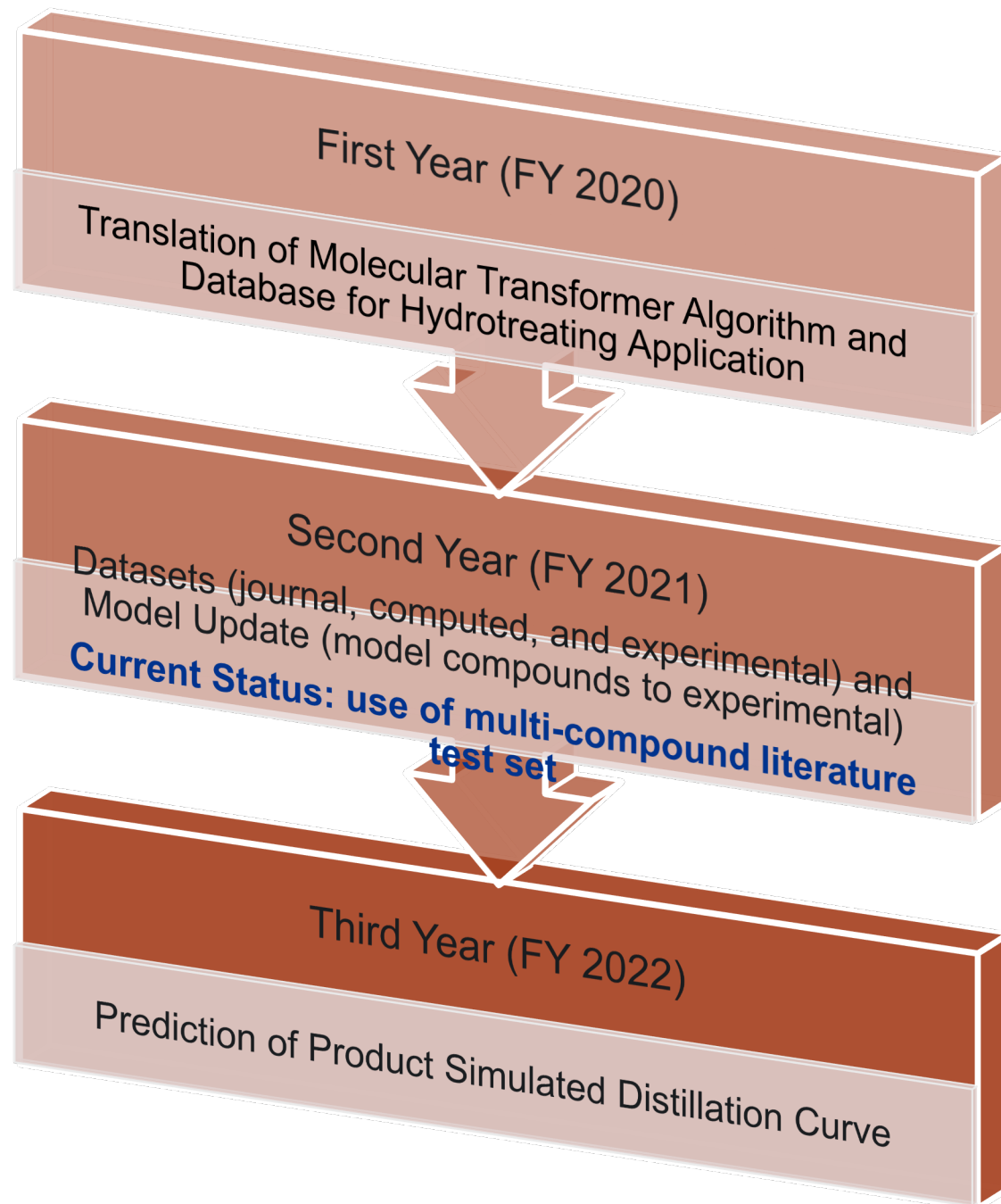
Why hydrotreating?

- Hydrotreating is **necessary** to convert thermochemically-produced biomass liquids into hydrocarbon fuels
- The underlying hydrotreating chemistries can potentially be **predicted based on chemical functional groups and reaction families**

Why machine learning ?

- Potential for exposing **non-intuitive trends and correlations** between existing data
- Ability to bring together information from disparate datasets

1 Management – Project and Task Structure



- **Project Start:** FY 2020
- **FY22 Initial End Project Goal:** Develop a model that will have a predictive accuracy of at least 70% for a key product attribute, such as simulated distillation curve

INPUT

Literature review and research, compilation (publicly available, computed, and experimental data), data clean-up and ingestion, and database development

Ensuring syntax and algorithm compatibility

Code development, **training**, **validation**, and **testing**

OUTPUT

Predict product quality based on experimental data, and provide insight for process operation

1 Management – Key Personnel and Roles

- **Key personnel: Diverse Team**

- **Ms. Sudha Eswaran (Computer Scientist)**

- ✓ Translation of original Molecular Transformer* and accompanying US Patent Office datasets into MongoDB**
- ✓ Development of algorithm: Code writing, training, validating, and testing

- **Dr. Robert Rallo (Chemist, Data Scientist)**

- ✓ Co-PI, Data Science expert

- **Dr. Mariefel V. Olarte (Chemical Engineer, Experimentalist)**

- ✓ PM and Co-PI, Domain (hydrotreating) expert
- ✓ US Department of Energy Science Undergraduate Laboratory Internship (SULI) mentor

- **Ms. Alexzabria Starks (Chemist, Intern)**

- ✓ SULI Intern (10 weeks), Building dataset of reaction SMILES*** from the literature and boiling point ranges from a gas chromatography-mass spectrometry

* Molecular Transformer – baseline algorithm; University of Cambridge, IBM; <https://github.com/pschwillr/MolecularTransformer>

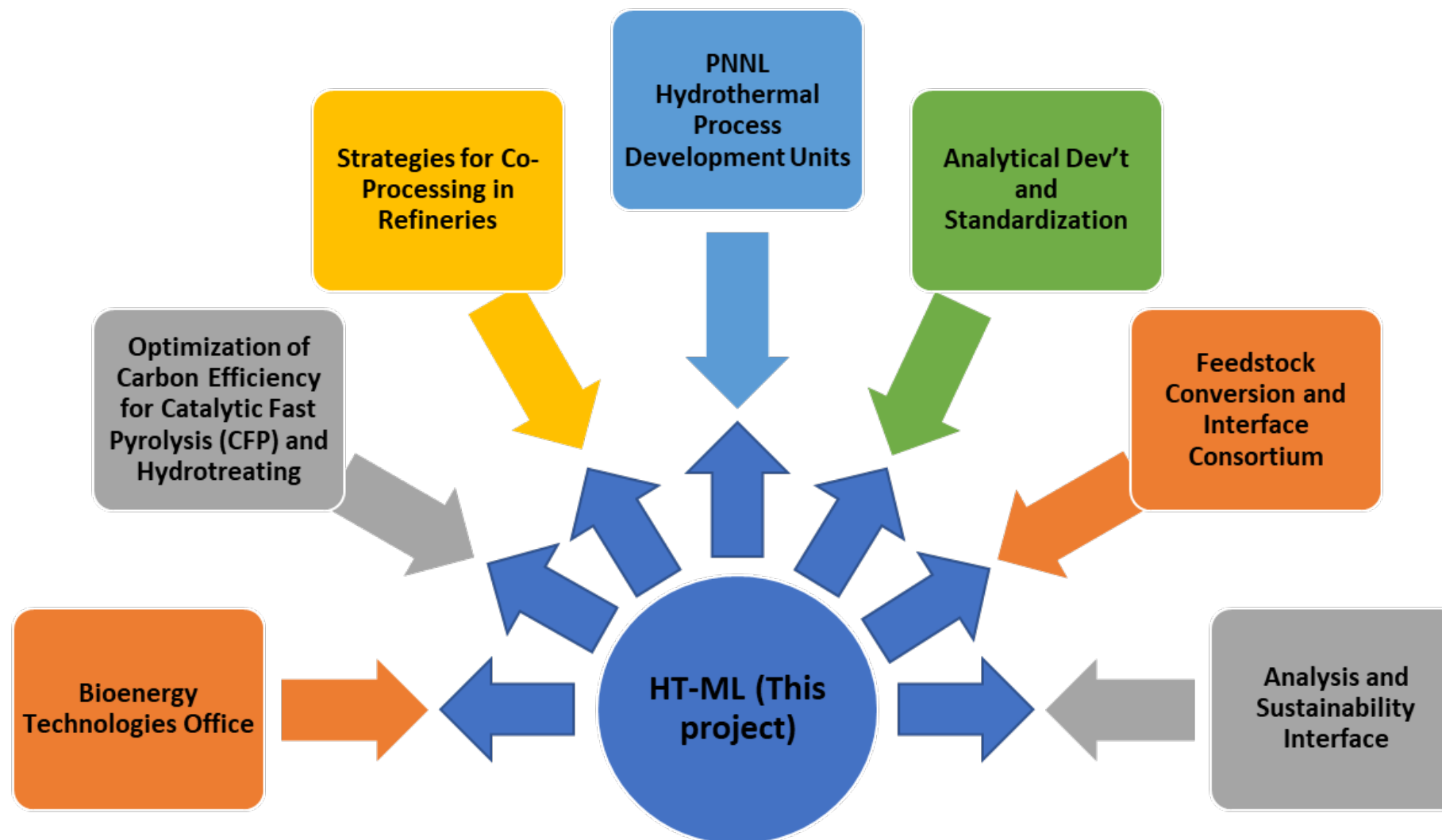
** MongoDB – database

*** SMILES – Simplified Molecular Input Line Entry System

1 Management – Communication of data and knowledge

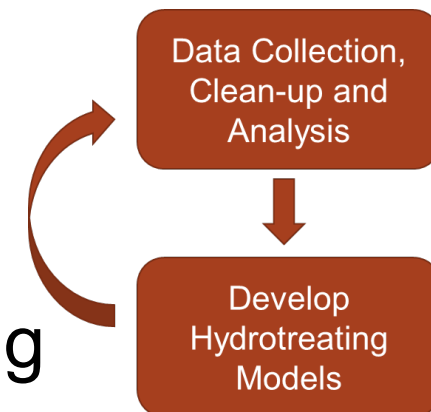
- **Interaction with other projects:**

- This project **relies** on experimental data from other projects
- In the 2nd year, rescope focused on gathering more data



- Leverage data from other projects to generate conversion data to improve ML algorithm
- Opportunity to **tease out non-intuitive trends/correlations** that can inform data-donor projects

1 Management – Project Risk Mitigation

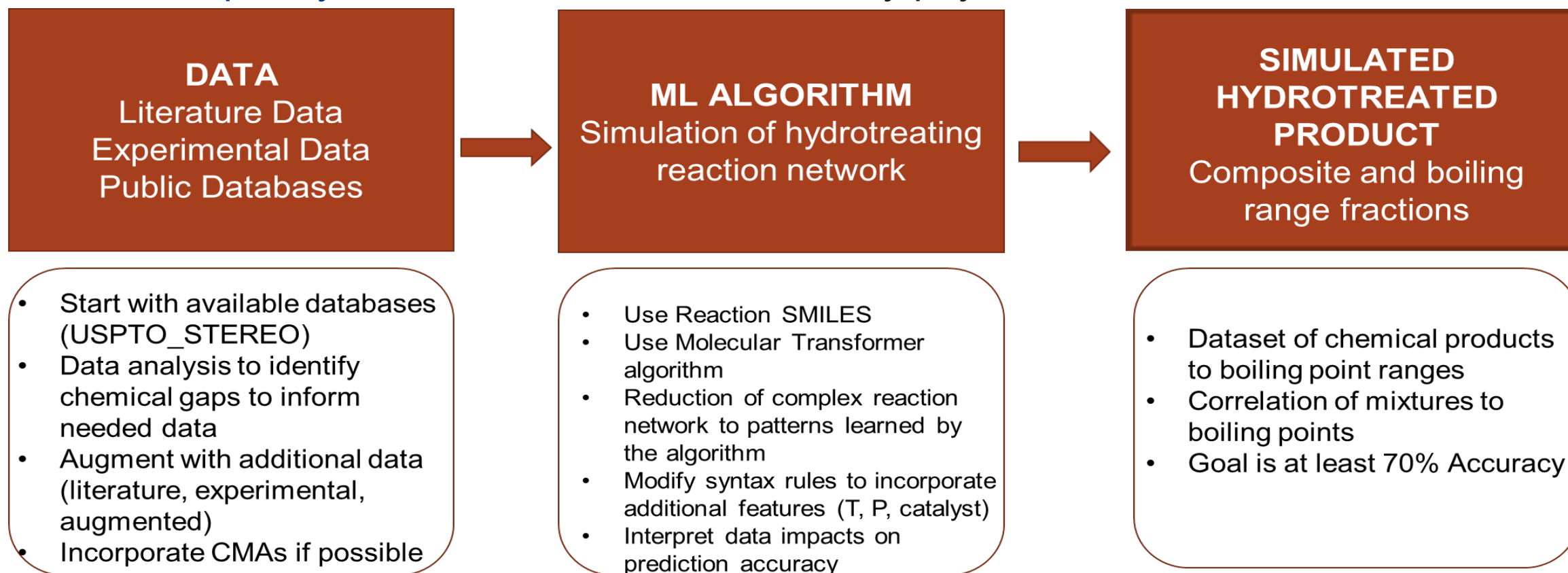


- **Focus:** Develop a machine learning (ML) tool to capture underlying chemistries that are represented in hydrotreating reactions
- **Ancillary:** Fit physico-chemical understanding of trends observed in reactions across literature and in experiments
- **Risks and Mitigations:**

Risk	Mitigation
Data quality and/or volume is not sufficient	<ul style="list-style-type: none"> • We are looking at various sources of data, including publicly available datasets, literature, and website (e.g., PNNL Environmental Molecular Sciences Laboratory (EMSL) Arrows) curation and rely on other projects for their experimental and calculated data results. • We are actively interpreting the impact of the volume and type of data in our model's predictive capability.
Low model performance	<ul style="list-style-type: none"> • Weekly project meetings aside from one-on-one discussions. • Conduct mini-algorithm experiments to gain better insight and improve ML interpretability.

2 Approach – Project Overview

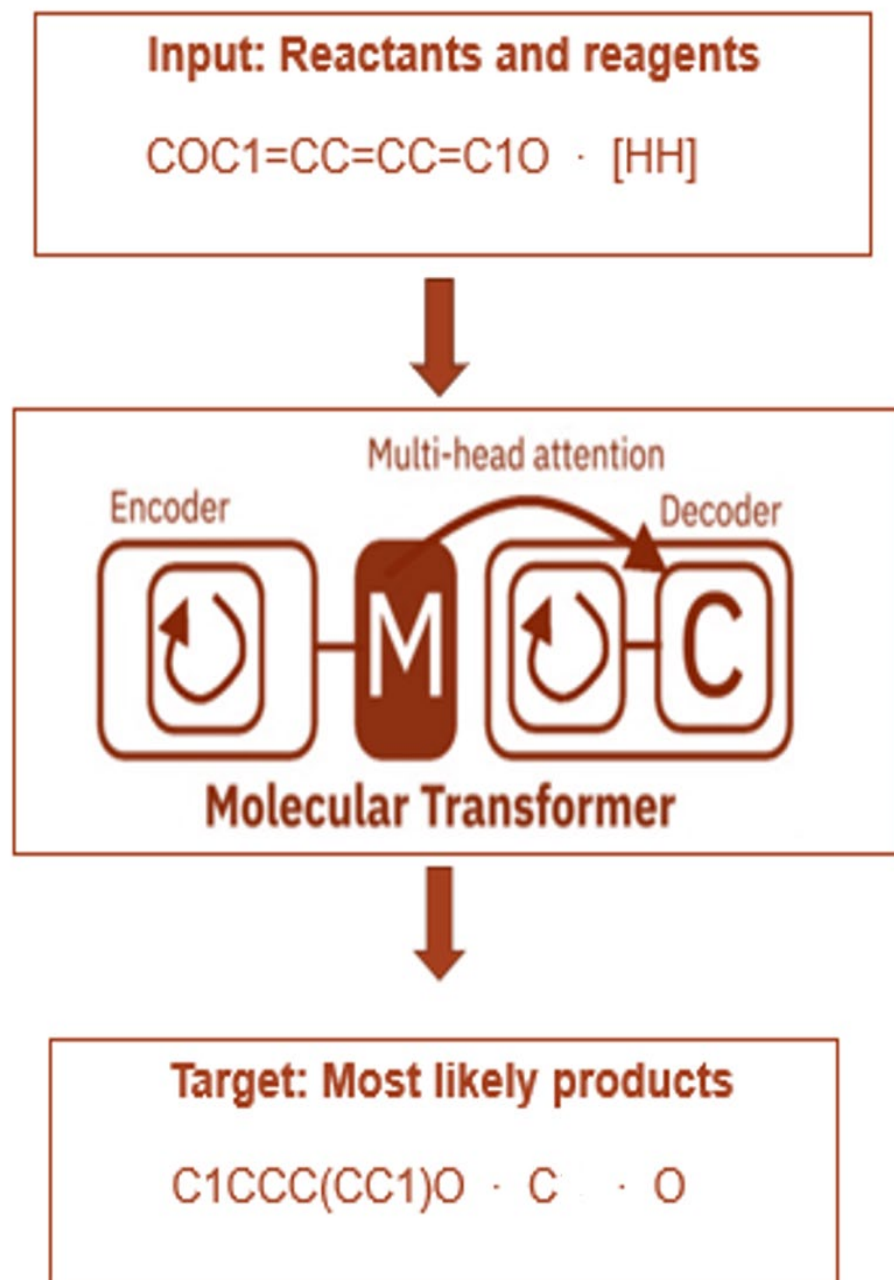
- **Goal:** Develop a machine learning tool to capture underlying chemistries that are represented in hydrotreating reactions
 - **Hypothesis:** Chemical transformations are expected to be finite but are largely affected by feed inputs
 - Build a framework on predicting single model compound hydrotreating reactions and then increase complexity as needed to be able to identify physico-chemical correlations



* CMA – critical material attributes

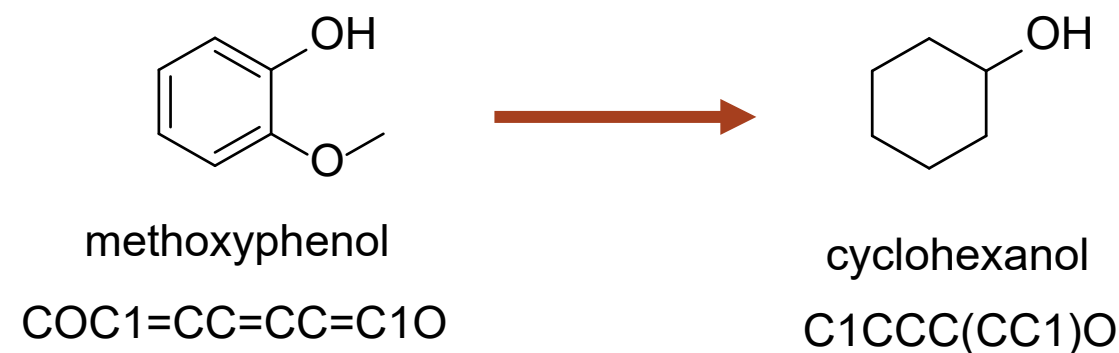
* SMILES - simplified molecular-input line-entry system

2 Approach – Molecular Transformer Overview



• Molecular Transformer (MT)

- an ML model inspired by **language translation**, accurately **predicts** the outcomes of organic reactions and **estimates** the confidence of its own predictions
- Convert chemical structures as “**words**” called Simplified Molecular Input Line Entry System (**SMILES**) string to form “**sentences**” called **Reaction SMILES**
- **Novelty:** 1st time applied to model hydrotreating



Reaction SMILES: COC1=CC=CC=C1O · [HH] > C1CCC(CC1)O · C · O

2 Approach – Data and Algorithm Metrics

US patents into a searchable database (USPTO Stereo)

United States Patent [19] [11] 3,931,181
Kompis et al. [45] Jan. 6, 1976

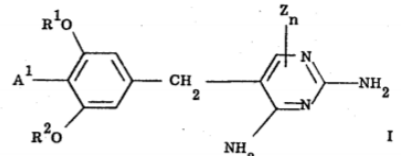
[54] 2,4-DIAMINO-5-BENZYLPIRIMIDINES [57] ABSTRACT
[75] Inventors: Ivan Kompis, Oberwil; Gerald Rey-Bellet, Basel; Guido Zanetti, Fullinsdorf, all of Switzerland
[73] Assignee: Hoffmann-La Roche Inc., Nutley, N.J.
[22] Filed: July 16, 1974
[21] Appl. No.: 489,050

[30] Foreign Application Priority Data
July 27, 1973 Switzerland..... 10995/73

[52] U.S. Cl. 260/256.4 N; 260/256.4 C; 260/256.5 R; 424/251

[51] Int. Cl.² C07D 239/00

[58] Field of Search..... 260/256.4 C, 256.4 N, 256.5 R



ADD DATA VIEW

```

_id: ObjectId("5e5844b1df6804f66916b264")
source: Object
  documentId: "US03931181"
  paragraphText: "A mixture of 12 g. of 4-(chloroformyl)-2,6-diethoxy-benzoic ethyl este..."
  reactionSmiles: "[CH2:1]([O:3][C:4](=[O:20])[C:5]1[C:10]([O:11][CH2:12][CH3:13])=[CH:9]..."
productList: Object
  product: Object
    role: "product"
    molecule: Object
      identifier: Arr
        0: Object
          dictRef: '
          value: "C(
        1: Object
          dictRef: '
          value: "Ir
          entityType: "ex
          appearance: "co
          state: "oil"
      reactantList: Object
        reactant: Array
          0: Object
            role: "react
            count: "1"
            molecule: Ob
            identifier:
            entityType:
          1: Object
          2: Object
      spectatorList: Object
      reactionActionList: Object
  
```

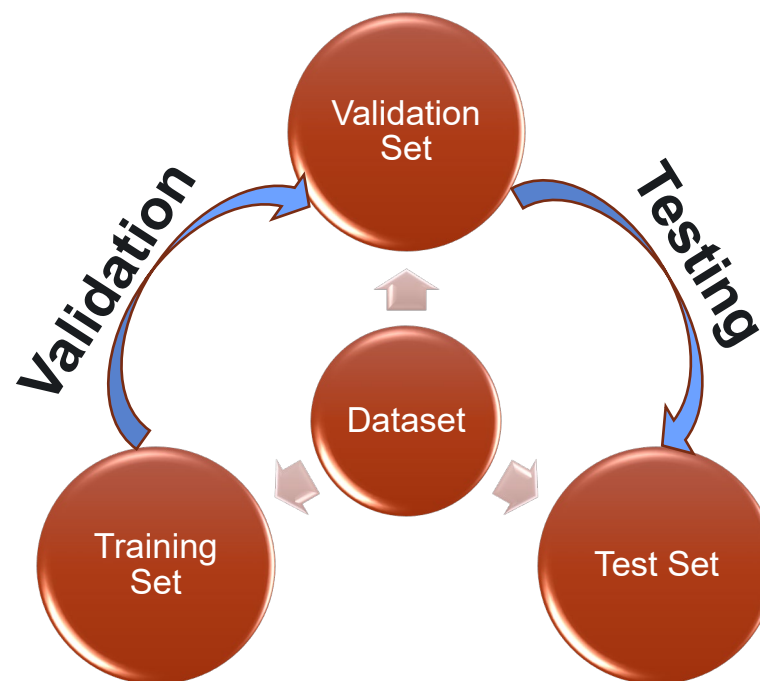
Dataset metrics:

- Manually check for errors
- Was the correct reaction translated from the patent?
- Was the reaction correctly translated into SMILES*?

*Simplified Molecular Input Line Entry System

USPTO Dataset

- USPTO_Stereo – US patents from 1976 – Sept 2016; subset of the original published database by Lowe; Contains 1.0 M reactions
- Dataset divided into three: Training Set, Validation Set, and Test Set

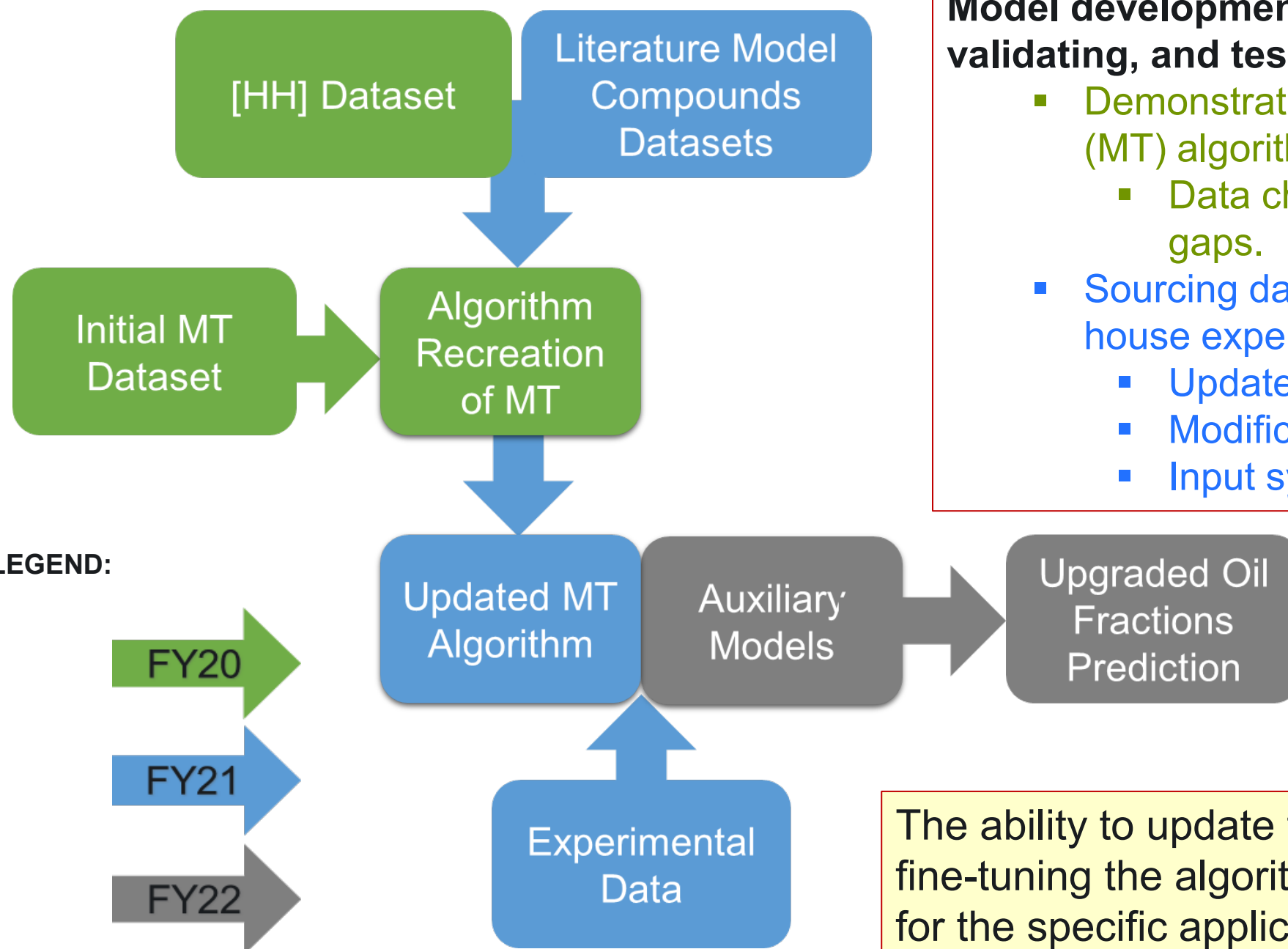


Algorithm metrics:

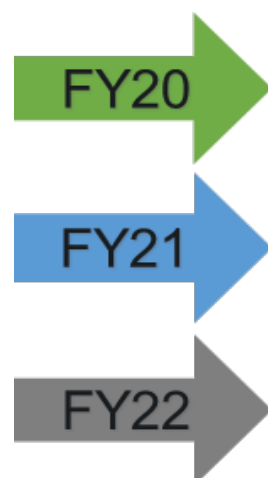
- Accuracy of predicted compounds, > 70%
- Prediction confidence threshold, > 0.5

Lowe DM. Extraction of Chemical Structures and Reactions from the Literature. PhD Thesis, University of Cambridge, 2012

2 Approach – Algorithm Development



LEGEND:



Model development involves several stages of training, validating, and testing the algorithm.

- Demonstrating recreation of Molecular Transformer (MT) algorithm.
 - Data characterization and analysis to identify gaps.
- Sourcing data from literature, computing, and in-house experimental data
 - Update MT algorithm
 - Modification/creation of helper scripts
 - Input syntax adjustment, as needed

Future work needed to reach end-of-project goal:

- Development of auxiliary models

The ability to update the models with new data will lead to fine-tuning the algorithms for higher predictive capability for the specific application.

2 Approach – Potential Challenges, Solutions, and Go/No-Go

Potential Challenges	Mitigations
Molecular Transformer (MT) dataset insufficient for hydrotreating (HT) application	Sourcing of additional quality data from literature and actual experiments
Quality data is not available	Additional data is being sourced from multiple projects, including in-house experimental data
Data availability not in the form and syntax required by the algorithm (original MT syntax has SMILES* string only)	Data pre-processing and syntax development to include other parameters such as temperature, pressure, and catalyst information
MT cannot capture complex reaction network	Design a segmented algorithm: Predict all possible correct HT product structures with MT and then constrain with another algorithm using operational data such as T, P, and catalyst information

Go/No-Go decision point

Name	Date
Model update attains target accuracy of +20% over preliminary model after incorporation of additional HT-specific data	3/30/2021

*Simplified Molecular Input Line Entry System

3 Impact – Big Picture and Short Term

- **Big Picture:**

- To provide an adopter the ability to **consider alternatives** to a disrupted supply chain or when feedstock diversification is needed to optimize costs while **ensuring that target quality products can still be met by their existing infrastructure**
- Having an accurate predictive model that can leverage both literature and available experimental data that will **reduce the need for costly experiments to test each possible alternative feedstocks**

- **Short-term (3-Years):**

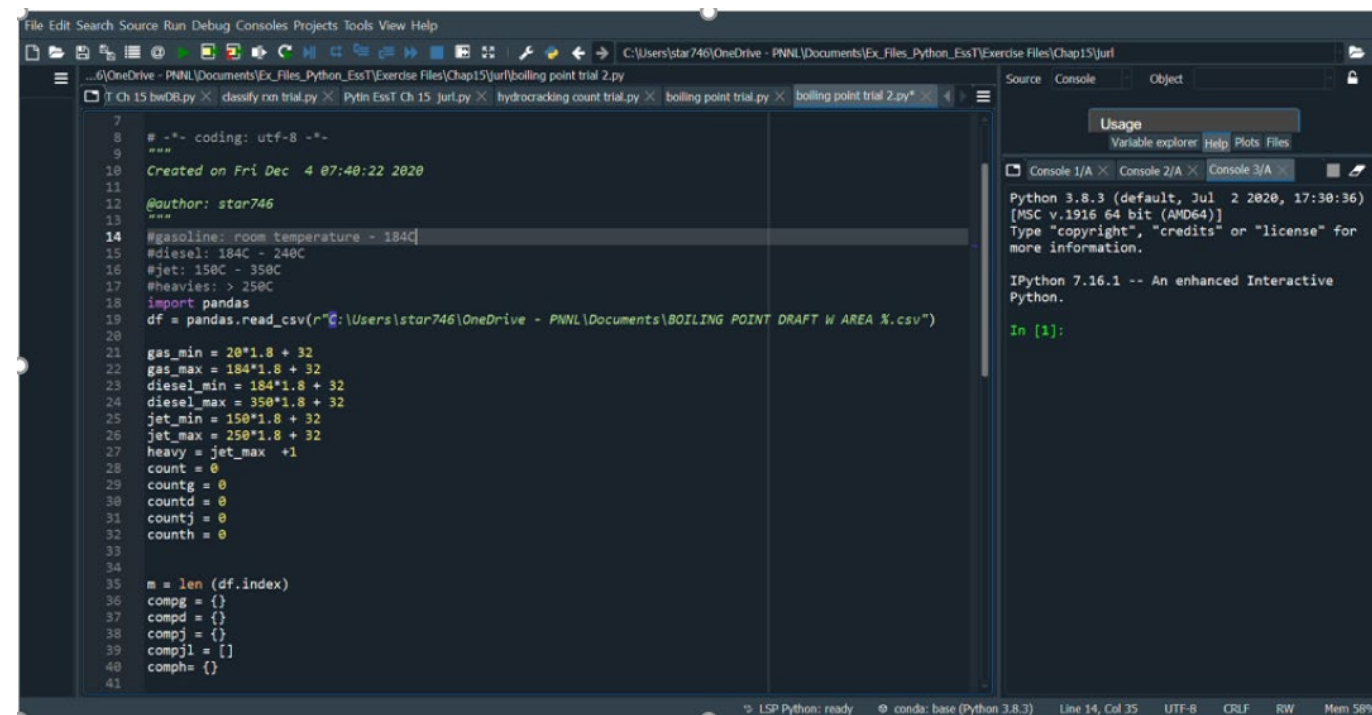
- Potential for **less computationally expensive model** compared to purely theoretical calculations
- Leverage extensive amount of **data already generated in other projects** to gain new conversion insight by (1) identifying **chemical gaps** to target data generation, (2) **streamlining** expensive experiments, and (3) providing **first pass prediction** of impact by input/feed change

3 Impact – Educational Outreach



Hydrotreating

- Students mentored in this project:
 - get inspired to explore the [convergence](#) of computer science, chemistry, and chemical engineering
 - work with a graduate student and an undergraduate Science Undergraduate Laboratory Internship (SULI) intern
 - [cross-fertilize](#) with established Lab personnel and promote a continual learning environment.



```
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\star746\OneDrive - PNNL\Documents\Ex_Files_Python_EssT\Exercise Files\Chap15\Jurl
...6\OneDrive - PNNL\Documents\Ex_Files_Python_EssT\Exercise Files\Chap15\Jurl\boiling point trial 2.py
T Ch 15 bwDB.py x classify rxn trial.py x Pytn EssT Ch 15 Jurl.py x hydrocracking count trial.py x boiling point trial.py x boiling point trial 2.py* x
7
8 # -*- coding: utf-8 -*-
9
10 Created on Fri Dec 4 07:48:22 2020
11
12 @author: star746
13
14 #gasoline: room temperature - 184C
15 #diesel: 184C - 240C
16 #jet: 150C - 350C
17 #heavies: > 250C
18 import pandas
19 df = pandas.read_csv(r"C:\Users\star746\OneDrive - PNNL\Documents\BOILING POINT DRAFT W AREA %.csv")
20
21 gas_min = 20*1.8 + 32
22 gas_max = 184*1.8 + 32
23 diesel_min = 184*1.8 + 32
24 diesel_max = 350*1.8 + 32
25 jet_min = 150*1.8 + 32
26 jet_max = 250*1.8 + 32
27 heavy = jet_max + 1
28 count = 0
29 countg = 0
30 countd = 0
31 countj = 0
32 counth = 0
33
34
35 m = len(df.index)
36 compg = {}
37 compd = {}
38 compj = {}
39 compj1 = []
40 comph = {}
41
```

Usage
Variable explorer Help Plots Files
Console 1/A x Console 2/A x Console 3/A x
Python 3.8.3 (default, Jul 2 2020, 17:30:36)
[MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 7.16.1 -- An enhanced Interactive Python.
In [1]:

Student learned python programming.

4 Progress and Outcomes - Milestones

- **FY 2020 Milestones:**

- ✓ 12/31/19 – Outline (achieved)
- ✓ 03/31/20 – Recreation of Molecular Transformer (MT) implementation (achieved)
- ✓ 06/30/20 – Analysis of the MT dataset to identify chemical gaps (achieved)
- ✓ 09/30/20 – Implement hydrotreating-related reaction as test set showing 50% accuracy (partially achieved)

- **FY 2021 Milestones:**

- **Inclusion of additional data into existing database**

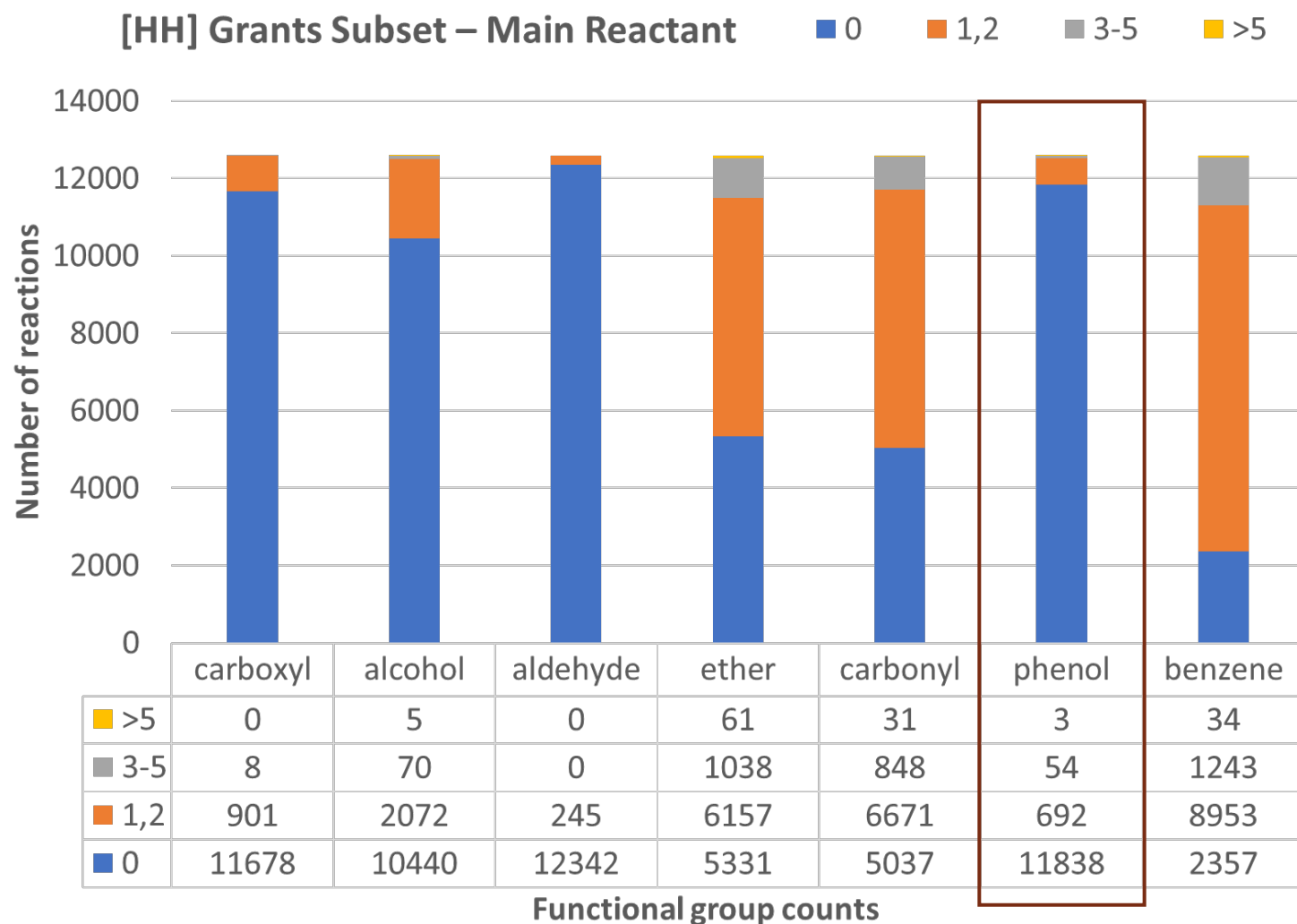
- ✓ Starting database – USPTO database (achieved)
- ✓ 12/31/20 – Additional literature data (achieved)
- ✓ 03/31/21 – Demonstrate at least 50% accuracy for test compounds. (achieved) Simulated kinetic and thermodynamic data (e.g., PNNL EMSL Arrows)
- ✓ 06/30/21 – In-house experimental data from other projects.

- **Re-training, validating, and testing of machine learning algorithm**

- ✓ 09/30/21 – Demonstrate at least 70% accuracy for test compounds
- ✓ Understand the impact of additional data
- ✓ Are we improving the accuracy of the model? Why?

Satisfied project milestones. Go/No-Go milestone achieved.

4 Progress and Outcomes – Identification of Chemical Spaces in the Existing Dataset

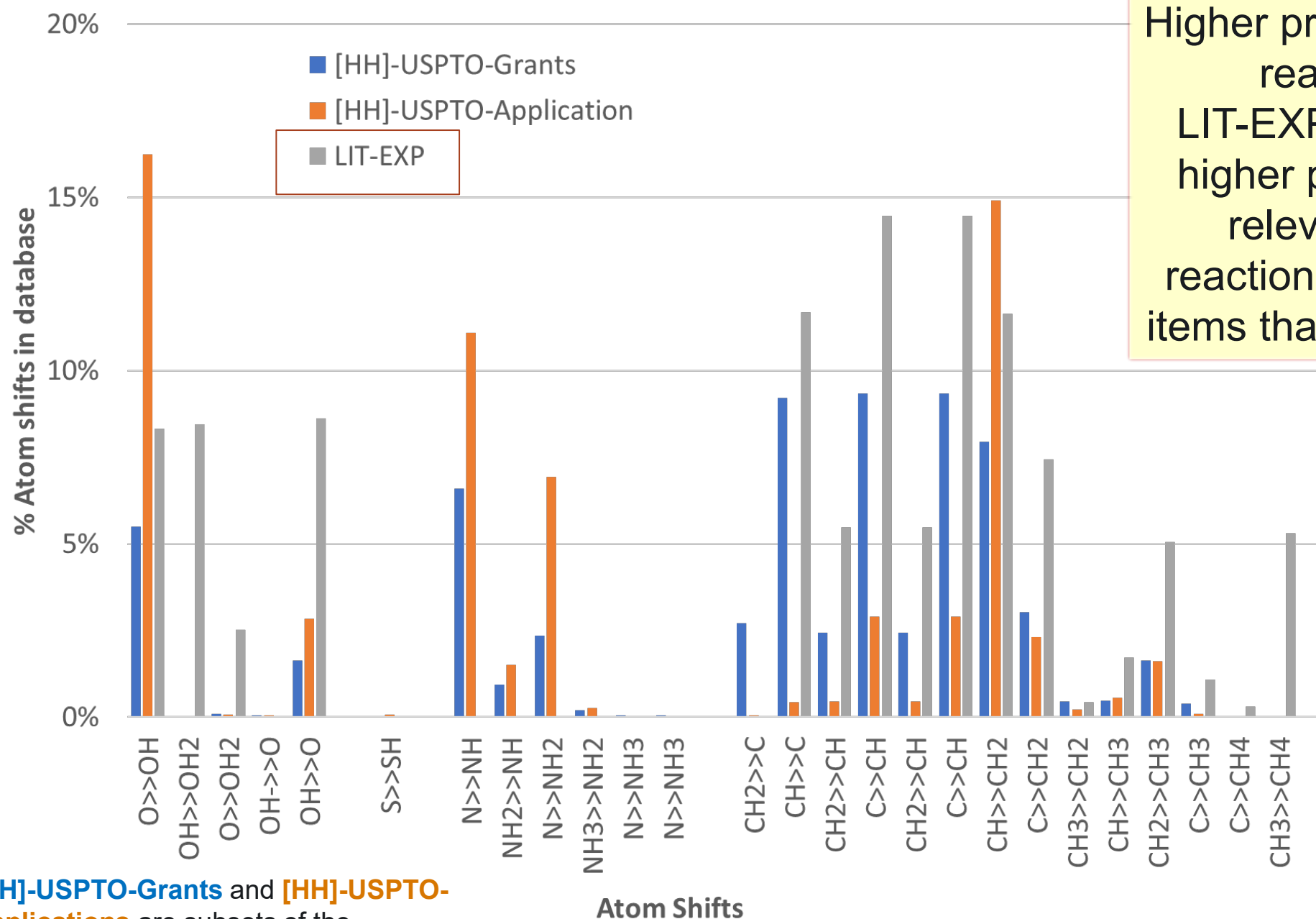


	% O (by ³¹ P NMR)*
Aliphatic Alcohol	8
Phenolic	4
Carboxyl	2

- **Reasons for querying the chemical space of the existing dataset:**
 - Identify missing data needed to augment the existing datasets
 - Guide and focus the subsequent data collection
 - Inform which future experiments are needed to collect additional data and improve prediction

• **Example:** In lignocellulosic bio-oils, we expect to see more phenolics. There is a potential scarcity in this data region.

4 Progress and Outcomes – Comparison of Functional Group Reactions in New Dataset



Higher prediction accuracy of the LIT-TEST reactions due to the addition of LIT-EXP dataset is likely because of the higher percentage of hydrotreating (HT) relevant reactions and similar type reactions in LIT-EXP, despite much lower items than Molecular Transformer dataset.

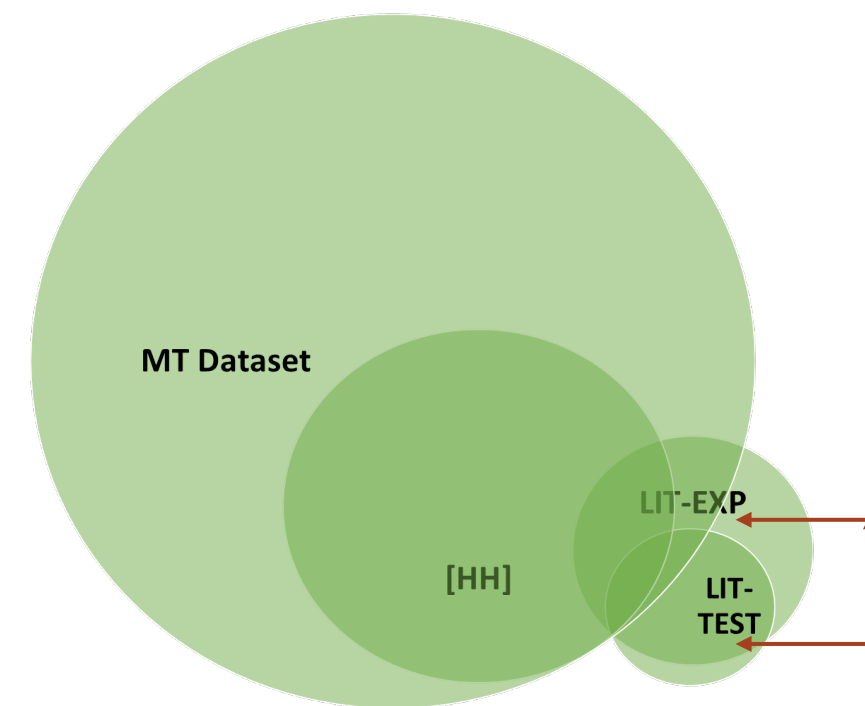
Observations:

- LIT-EXP tend to have **higher proportion of reducing atom shifts:**
 - $O \rightarrow H_2O$ and $O \rightarrow H_2O$
 - Reduction of C species
- Upon inspection, $OH^- \rightarrow O$ and $OH \rightarrow O$ are atom shifts involved in **multiple reactions in one molecule**, e.g., the aromatic ring where the OH is attached was reduced to an alkyl ring.

[HH]-USPTO-Grants and [HH]-USPTO-Applications are subsets of the USPTO_STEREO database.

4 Progress and Outcomes – List of Datasets

Initial Molecular Transformer (MT) Dataset	Number of Items
<ul style="list-style-type: none"> USPTO_MIT (MT Dataset) 	888 k reactions
Hydrotreating/Hydrogenation-Related Dataset, (HH)	
<ul style="list-style-type: none"> USPTO dataset filtered for reactions with H₂ as one of the reactants 	21 k reactions
Literature Single Compounds Dataset	
<ul style="list-style-type: none"> Augmented data of 5 distinct biomass-derived single model compound reactions from 1 journal article, LIT-TEST Single compound reaction manually extracted from 10 journal articles, LIT-EXP 	113 reactions; 5 distinct reactions
	395 reactions; 85 distinct reactions
Manually Added Test Reactions	
<ul style="list-style-type: none"> 2 additional test reactions not in the original LIT-TEST 	2 distinct reactions



The information from the data not found in the original MT dataset improved prediction accuracy for HT-related (LIT) reactions. **Fine tuning on specific data.**

4 Progress and Outcomes – Preliminary Machine Learning Model Performance

- Recreation of the Molecular Transformer (MT) implementation

Dataset	Train	Validation	Test	Accuracy
MT dataset (Pre-trained Model)	818 k	30 k	40 k	90.4%

Use of hydrotreating-related reactions as Test Set

- MT training data applied to LIT-TEST

Dataset	Train	Validation	Test	Accuracy
MT dataset (Pre-trained Model)	818 k	30 k	113	17.7%

LIT-TEST reactions

- Anisole (methoxybenzene) + H₂ -> Phenol + Methane
 - COc1ccccc1.[H][H]>>c1ccc(cc1)O.C
- 2-methoxyphenol + H₂ -> 1,2-dihydroxybenzene + Methane
 - COc1ccccc1O.[H][H]>>c1ccc(c(c1)O)O.C
- 1,2-dihydroxybenzene + H₂ -> Phenol + H₂O
 - c1ccc(c(c1)O)O.[H][H]>>c1ccc(cc1)O.O
- Phenol + H₂ -> Benzene + H₂O
 - c1ccc(cc1)O.[H][H]>>c1ccccc1.O
- Phenol + H₂ -> Cyclohexane + H₂O
 - c1ccc(cc1)O.[H][H]>>C1CCCCC1.O

Decrease in accuracy suggests that LIT-TEST specific reaction centers are not represented in the MT dataset.

4 Progress and Outcomes – Preliminary Machine Learning Model Performance

Use of hydrotreating-related reactions (LIT-TEST) as Test Set

- Molecular Transformer (MT) training data applied to LIT-TEST

Dataset	Train	Validation	Test	Accuracy
MT dataset (Pre-trained Model)	818 k	30 k	113 (augmented)	17.7%

- MT + LIT-EXP as training data applied to LIT-TEST

Dataset	Train	Validation	Test	Accuracy
MT dataset + LIT-EXP	818 k + 394	30 k	113 (augmented)	34.5%

- MT + enhanced LIT-EXP as training data applied to enhanced distinct LIT-TEST

Dataset	Train	Validation	Test	Accuracy
MT dataset + LIT-EXP + 2	818 k + 394 + 2	30 k	5 (distinct) + 2	57%

- enhanced LIT-EXP – addition of two reactions from LIT-TEST not found in LIT-EXP
- enhanced LIT-TEST – addition of two reactions (different from above) not found in LIT-TEST

- Go/No-Go milestone (additional 20% accuracy) achieved.
- **Insight:** The type of additional training data is important.

Future work: Identify a metric that measures quality of additional data.

INCREASING ACCURACY

Quad Chart Overview

Timeline

- Project Start: October 1, 2019
- Project End: September 30, 2022

	FY20	FY21	Active Project
DOE Funding	\$ 150,000	\$ 85,000	\$ 235,000

Project Partners

- Collaboration with projects 3.4.3.304, 2.4.2.305, 2.2.2.301, 2.5.2.302, 1.2.2.807, 2.1.0.301

Barriers addressed

ADO-A: Process Integration

ADO-G: Co-Processing with Petroleum Refineries

Project Goal

Develop a machine learning tool that can model and predict expected hydrotreating (HT) conversions given specific bio-oil and biocrude inputs.

End of Project Milestone

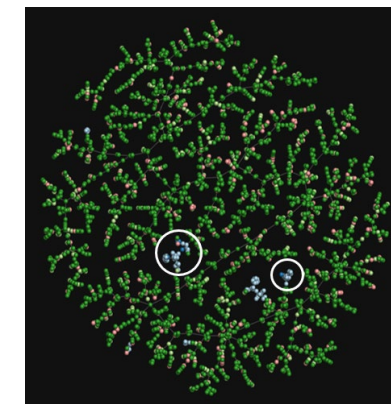
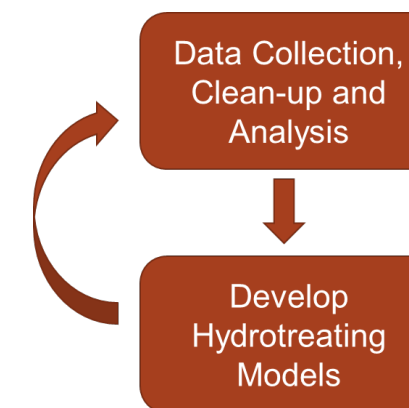
By 2022, we will develop a machine learning (ML) model that describes hydrotreating of HTL biocrude through a reaction network framework, with at least a predictive accuracy of 70% for a key product attribute, such as simulated distillation curve.

Funding Mechanism

Lab Call 2019

Acknowledgement

- Dr. Robert Rallo
 - Ms. Sudha Eswaran
 - Ms. Alexzabria Starks
 - Mr. Alan Cooper
 - Dr. Asanga Padmaperuma
 - Ms. Corinne Drennan
-
- The US DOE Bioenergy Technologies Office (TM: Ms. Liz Moore) for funding our efforts.



- **Overview: GOAL:** Create a framework to develop hydrotreating (HT) reaction network using machine learning (ML) tools to model and predict expected HT conversions given bio-oil and biocrude inputs without experimentation.
- **Management:** New project. Assembled a diverse team in this multi-disciplinary project.
- **Approach:** Assemble data from various sources. Leverage data from other projects/efforts. Understand the impact of type of available data on model prediction accuracy and eventually, correlate operational data with product quality.
- **Impact:** Initial effort to apply natural language processing (NLP)-based ML application to HT reaction networks. Potential for less expensive computational requirement. Inform experimental work and identify chemical data gaps. Educational outreach.
- **Progress and Outcomes:** Developed new datasets. Improved accuracy from 17.7% to 57%.

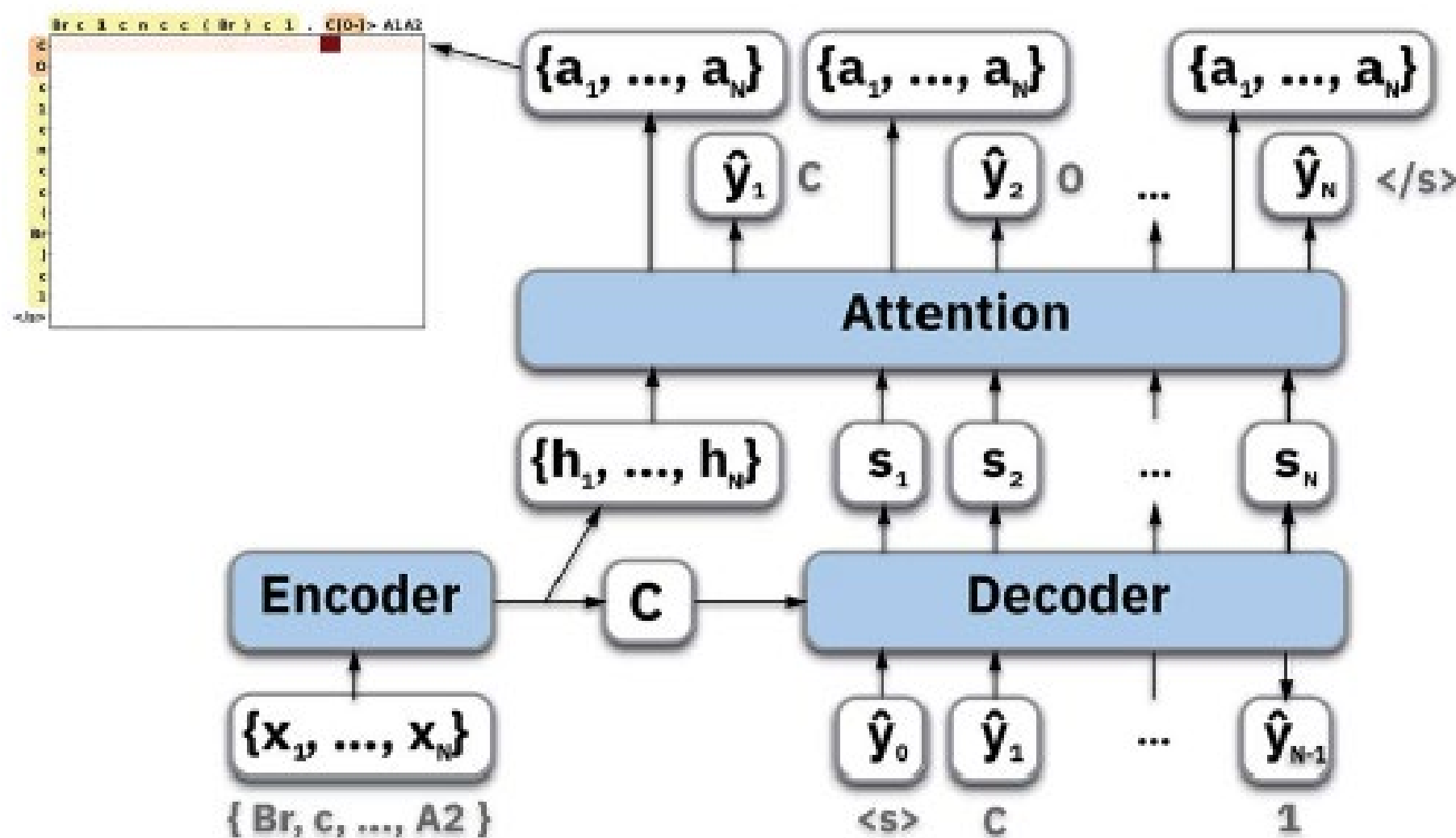


Pacific Northwest
NATIONAL LABORATORY

Thank you



2 Approach – Molecular Transformer Algorithm



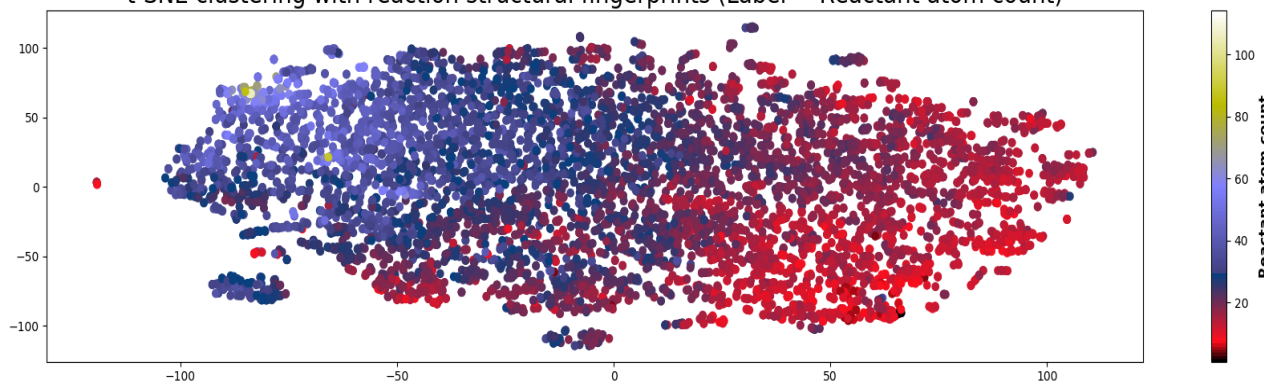
Attention-based seq2seq model

- Bi-directional long short-term (LSTM) encoder was used
- Use of attention allowed for complex long-range dependencies between multiple tokens

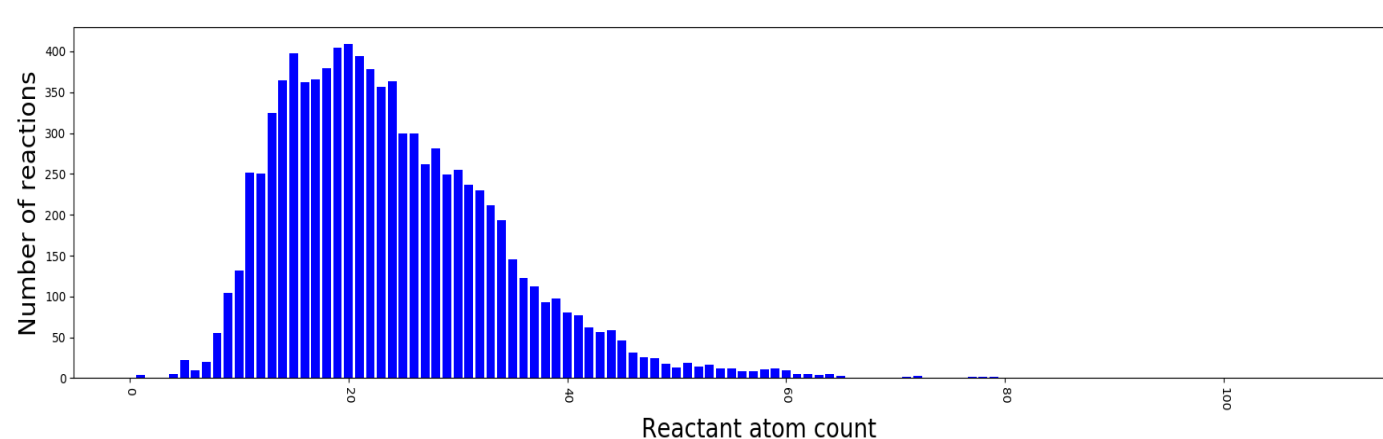
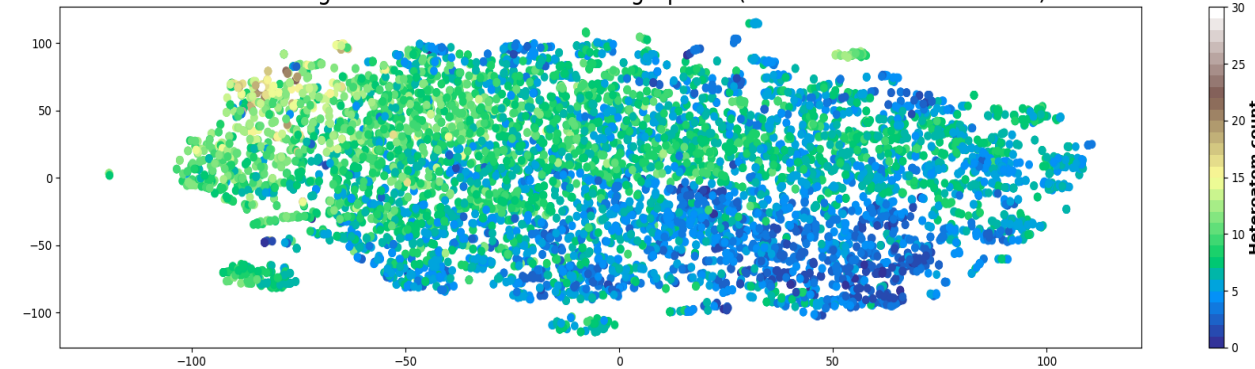
4 – Identification of Existing Chemical Space in the Molecular Transformer Datasets

- Reasons for querying the chemical space of the existing dataset:
 - Determine the baseline
 - Identify missing data needed to augment the existing dataset
 - Explain the impact of additional data

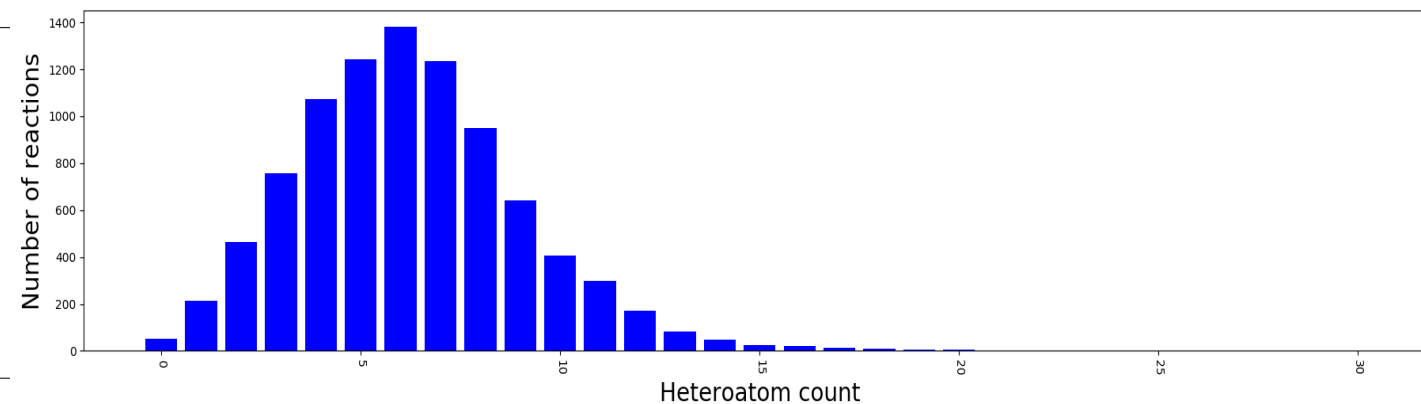
t-SNE clustering with reaction structural fingerprints (Label = Reactant atom count)



t-SNE clustering with reaction structural fingerprints (Label = Heteroatom count)

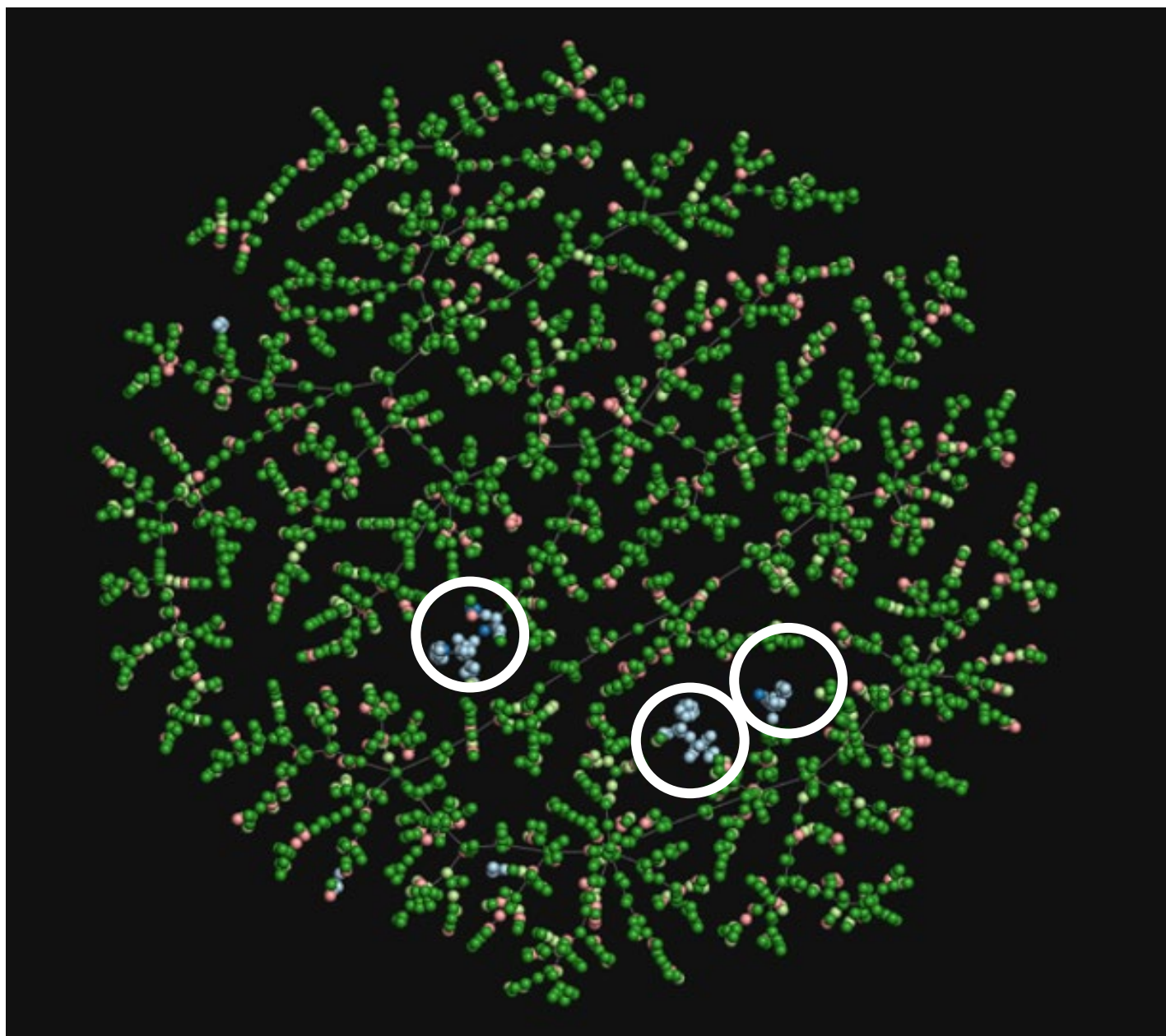


Reactant atom counts. 81% between 15 – 41 atoms.



Heteroatom counts. Median is 6.

4 Progress and Outcomes – Chemical Similarity in Training and Test Data Sets Likely Contribute to Improved Performance



[HH] dataset, a subset of Molecular Transformer dataset, is compared with LIT datasets.

Improved accuracy with the addition of **LIT-EXP** as training set (394/85) is likely due to its **similarity** (encircled regions) with the **LIT-TEST (113/5)**.

Small amount of relevant training data (394) can improve accuracy of model originally trained on large but disparate dataset (818k)

- LIT-EXP
- LIT-TEST
- [HH]-TEST
- [HH]-TRAIN
- [HH]-VAL

Future work: Addition of relevant **computed and in-house** derived experimental data to improve accuracy