SYMPL: Sugar Yield Machine Predictive Learning from Biomass Feedstocks

Scott McCloskey[1,2,4], Dupeng Liu[2,4], Brian Taylor[3,5], Yinglei Han[3,5]

[1]Saddleback College, [2]Advanced Biofuels and Bioproducts Processing Development Unit, [3]Joint BioEnergy Institute, [4]Lawrence Berkeley National Laboratory, [5]Sandia National Laboratories

BERKELEY LAB Workforce Development & Education

U.S. DEPARTMENT OF ENERGY Office of Science

ABSTRACT

This project uses machine learning techniques to predict sugar yields from various biomass feedstocks, aiming to optimize renewable energy production. Our approach utilizes machine learning on a combination of data collected from the ABPDU laboratory and the data extracted from scientific literature to accelerate the development of efficient biomass pretreatment methods, contributing to sustainable biofuel production and the achievement of UN Sustainable Development Goal 7.

We first utilize large language models for data extraction from scientific literature to supplement data collected in our laboratory. Machine learning algorithms are used to perform predictions on the data. Additionally a fine-tuned chatbot was created that allows researchers to ask questions about the data, the data source, and make sugar yield predictions with a high accuracy of 0.94 R^2.

BACKGROUND INFO

The UN Sustainable Development Goal 7 aims to ensure access to affordable, reliable, sustainable and modern energy for all. Sustainable biofuel production technologies contribute to this goal, and accurate prediction of sugar yields from biomass feedstocks is crucial for advancing these technologies.

GOAL OF PROJECT

PREDICTING SUGAR YIELDS (GLUCOSE AND XYLOSE) FROM BIOMASS FEEDSTOCKS USING MACHINE LEARNING

INPUT & TARGET FEATURES

- Biomass Feedstock
- Temperature
- Pretreatment Method
- Reaction Time
- Glucose Yield
- Xylose Yield
- Lignin
- Glucan/Cellulose
- Xylan/Hemicellulose
- Moisture/Water
- Ash

KEY:
Feedstock
Pretreatment
Yield (Target)
Characterization

FLOWCHART

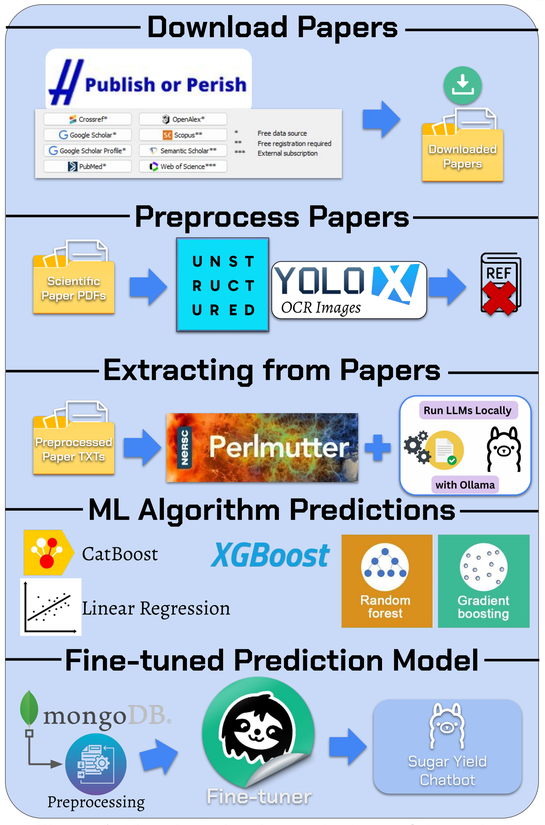

Figure 1: *Flowchart showing each phase of the project*

PROMPT ENGINEERING

Prompt Engineering Strategies Applied:

- Reduce Hallucinations by saying "N/A"
- Three-shot examples for format clarification
- Indications for document context in the prompt
- Prompt error correction if output format is invalid
- Adjusted Parameters: Temperature, Top_p, Top_k

Sugar Yield Prompt Example

```
Template = """Your goal is to extract data from the document context in the
format of the JSON schema for the biomass feedstock: {{feedstock}}. I'm
using a program to read the JSON output you provide. Return only JSON.
Do not return a list. Do not say any additional text.

Format your response with the following JSON schema pattern. Each value
in the schema should be filled out with the extracted information from the
document context. If the document does not contain the required
information for one of the values in the JSON, put 'N/A' as the value.
JSON Schema: {{schema}}

=== START OF DOCUMENT CONTEXT ===
{{document}}
=== END OF DOCUMENT CONTEXT ===

{% if invalid_replies and error_message %}
You already created the following output in a previous attempt:
{{invalid_replies}}
However, this doesn't comply with the format requirements from above and
triggered this error: {{error_message}}
Correct the output and try again. Just return the corrected output in JSON
without any extra explanations.
{% endif %}"""
```

Data Extraction Prompt Workflow:

- What feedstocks are in this document?
  - Is this a real biomass feedstock?
- Sugar yields and features for this feedstock?

COMPARING LARGE LANGUAGE MODELS

LLM	Company	Context Size	VRAM
Mixtral:8x22b-q4	Mistral AI	32k (Sliding)	80 GB
qwen2:72b-instruct-q6_K	Alibaba Cloud	128k	64 GB
llama3-gradient:70b-q8	Gradient	up to 1M	75 GB
command-r-plus:104b-q8	Cohere	128k	110 GB

Figure 2: *Large Language Models (LLMs) compared and used for parsing scientific papers.*

These LLMs were chosen based on these requirements:

- Large enough context size for scientific papers
- High benchmark scores for long context
- Recently released
- VRAM <= 80GB

Models were also quantized to lower the amount of VRAM to fit on an individual 80GB A100 when running on the NERSC Perlmutter Supercomputer. For command-r-plus, the context size was lowered to 80GB of VRAM to run on a H100.

MACHINE LEARNING PREDICTIONS

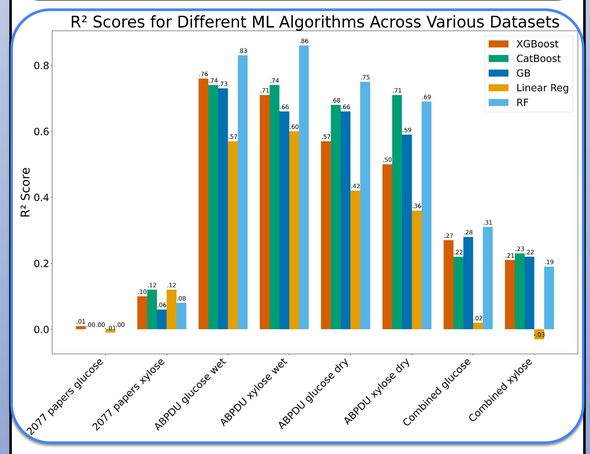

Figure 3: *R^2 (R-Squared) results of five different machine learning algorithms. Tested on the Scientific Papers (SP) dataset, the ABPDU datasets, and a combined dataset between the SPs and the ABPDU Dry Dataset. A higher R^2 score indicates a better prediction score.*

FEATURE ANALYSIS

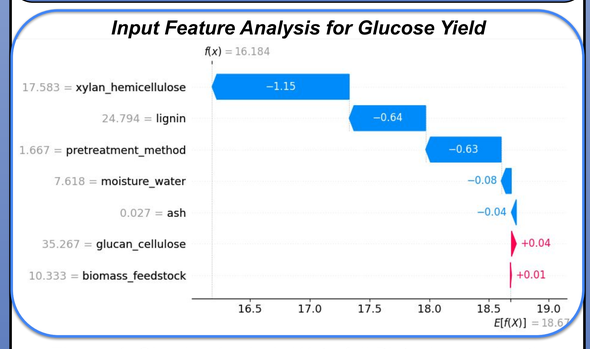

Figure 4: *Shapley Analysis Waterfall Plot, showcasing the importance/correlation of input features for glucose yields*

FINE-TUNED MODEL

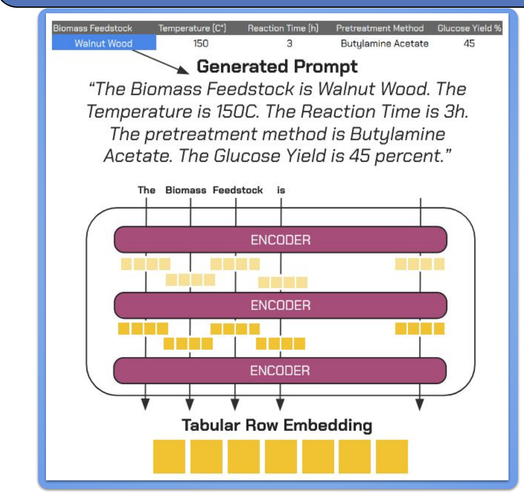

Figure 5: *The process of converting a tabular dataset into a paragraph structure for preprocessing a fine-tuned LLM dataset*

Evaluating the Fine-tuned Model

Finetuned Dataset	ABPDU Wet	ABPDU Dry	Scientific Papers
Glucose RMSE	3.44%	10.42%	36.94%
Glucose R^2	0.94 R^2	0.88 R^2	-0.67 R^2
Xylose RMSE	1.71%	14.43%	36.51%
Xylose R^2	0.95 R^2	0.72 R^2	-0.59 R^2
Evaluation Loss	0.142	0.195	0.551

Figure 6: *After fine-tuning Llama3 using Unsloth, the model was evaluated on a 20% split using human-like questions*

NEXT STEPS

- Predicting sugar yields on combinations of biomass feedstocks

ACKNOWLEDGEMENTS

Thank you to my mentors at the Lawrence Berkeley National Laboratory and the staff and other interns for supporting me throughout my internship.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internship (CCI) program.

STEM CORE · SADDLEBACK COLLEGE · ABPDU