PURDUE UNIVERSITY | The Data Mine

Merck: Natural Language Processing

The Data Mine Corporate Partners Symposium 2023

MERCK

Introduction

- Bioanalytical "BP" documents submitted by scientists & uploaded to app
- Can be both read and written

Result: Helps forecast future drug development procedures

- Terms extracted and stored into database
- Searches yield BP documents with attribute

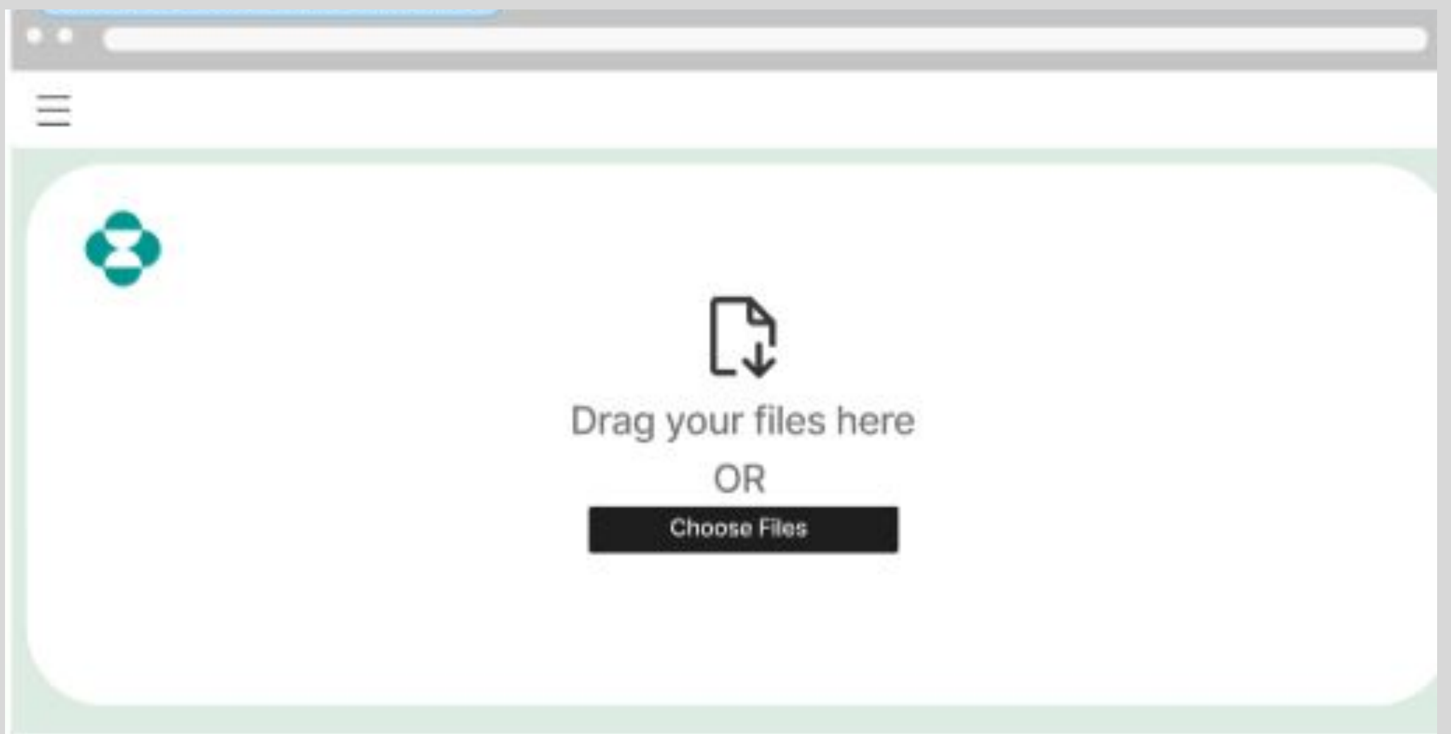
Methods

Researcher provides BP document with chemical information

BP-XXXX (Draft)

This analytical method is based on an automated 96-well format Extraction Method of drug from species matrix. MK-XXXX and stable isotope labeled internal standard (XXX) are chromatographed using chromatography and detected with tandem mass spectrometric detection employing a turbo ionspray (TIS) interface in the polarity ion mode.

Researcher uploads document into the app

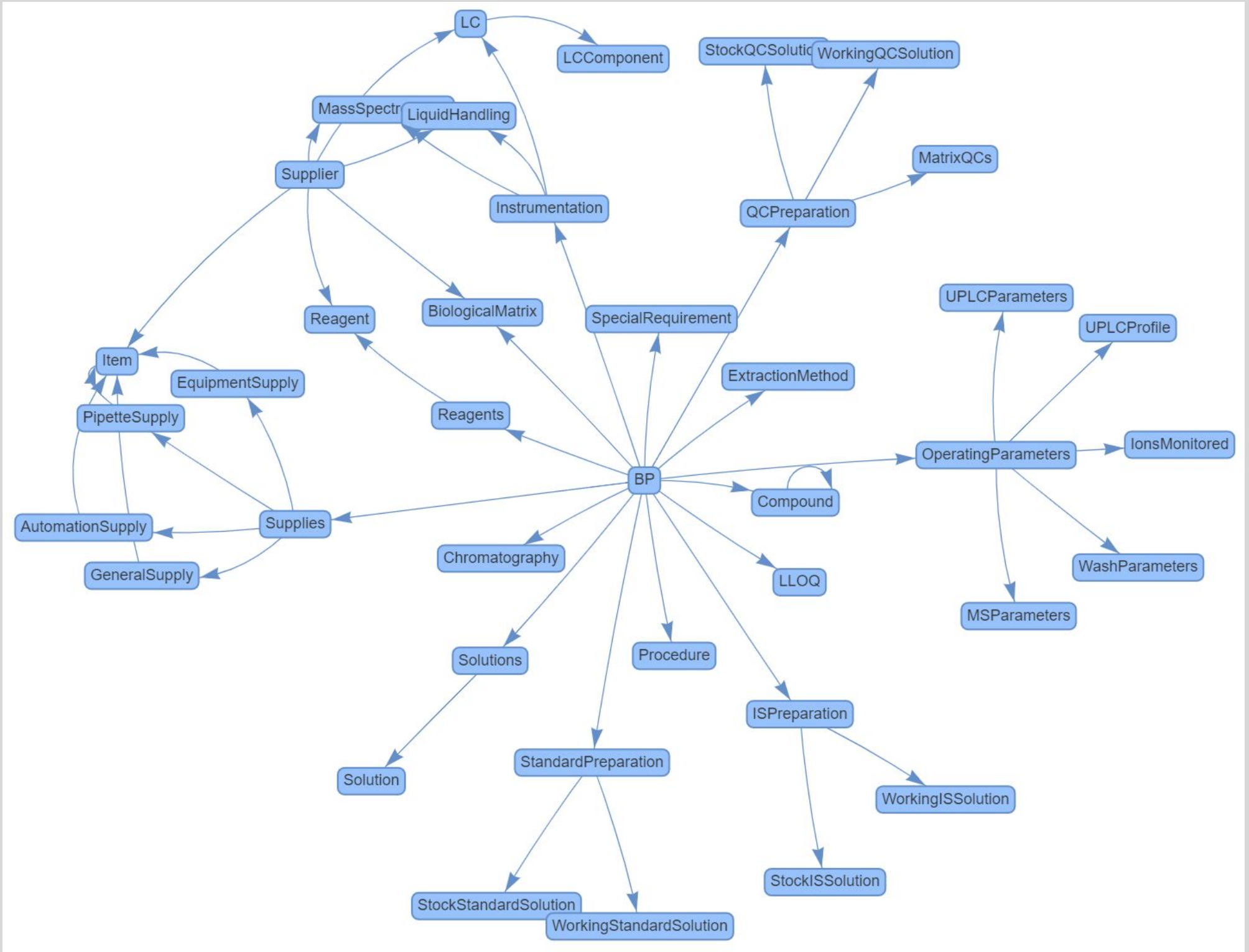


The app extracts the key terms identified in 'blue' using RegEx

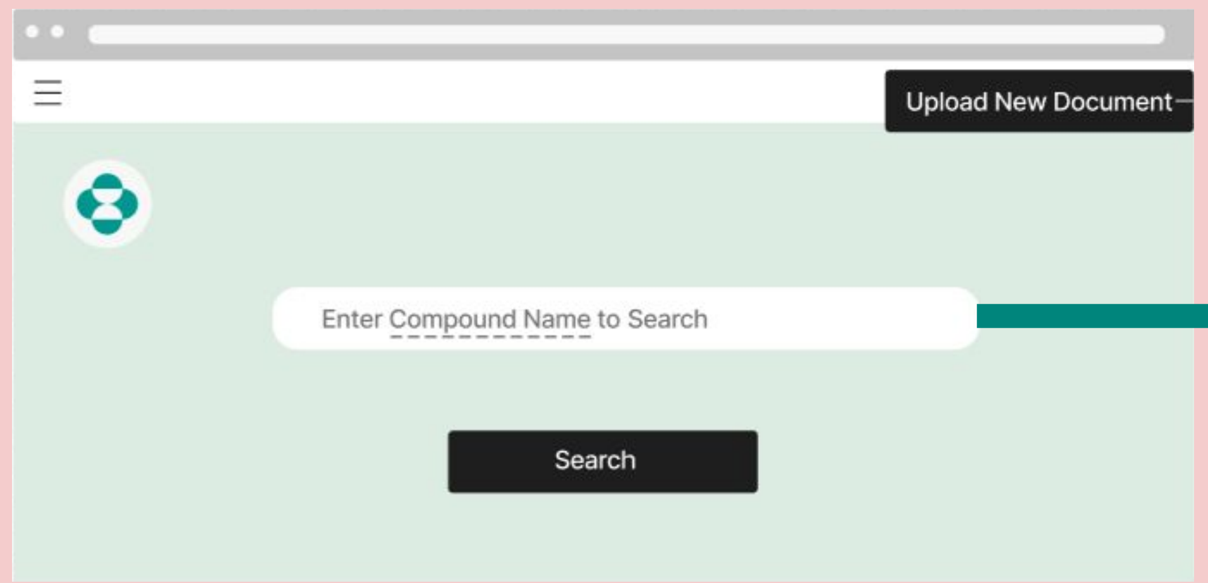
BP-XXXX (Draft)

This analytical method is based on an automated 96-well format **Extraction Method** of drug from **species matrix**. MK-**XXXX** and stable isotope labeled internal standard (**XXX**) are chromatographed using **chromatography** and detected with tandem mass spectrometric detection employing a **turbo ionspray (TIS)** interface in the **polarity** ion mode.

Extracted terms are stored in the Graph Database

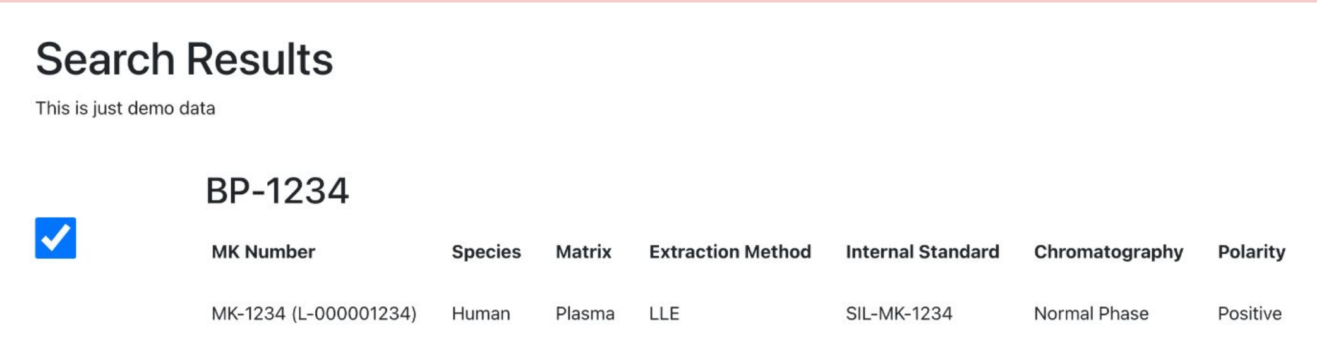


Frontend



Fuzzy search allows for accurately returned results regardless of variations.

Original	Variations	
Human	hummann	**Misspelling**
	hUmAN	**Casing**
	humans	**Other**



Search Results

This is just demo data

☑ BP-1234

MK Number	Species	Matrix	Extraction Method	Internal Standard	Chromatography	Polarity
MK-1234 (L-000001234)	Human	Plasma	LLE	SIL-MK-1234	Normal Phase	Positive

- BP documents ranked with relevancy
- Checklist-style downloading

Modeling

Where RegEx Works

BP-XXXX (Draft)

Using a 20 **mL** plasma sample from **humans**. The samples are stored at 70**°C**. The interface is in the **positive** ion mode.

Where RegEx Fails:

BP-XXXX (Draft)

Using a 20 **µL** plasma sample from **Homosapiens**. The samples are __ stored at 70**°c**. The interface is in the **POSITIVE** ion mode.

When RegEx fails, use…

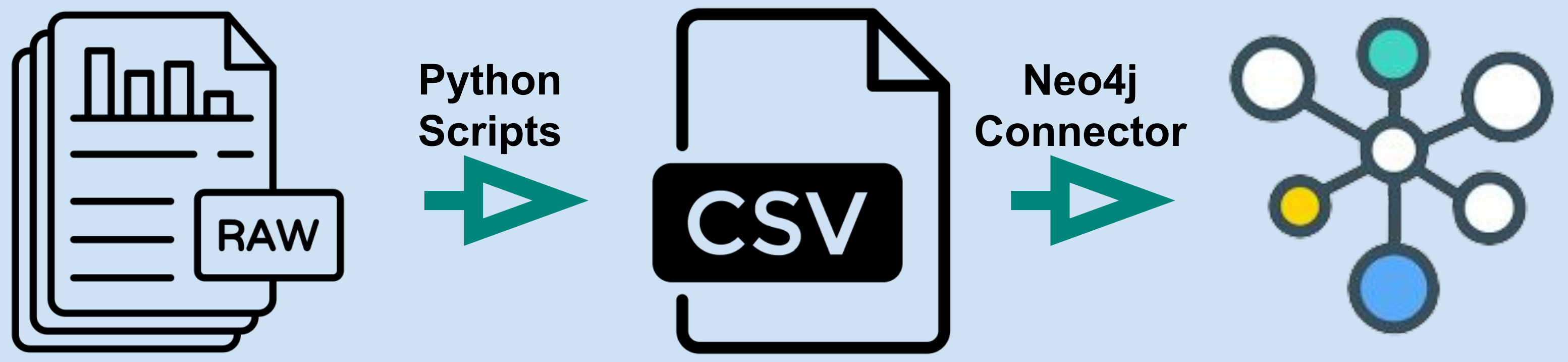
1) BERT: A general model meant to allow computers to better understand ambiguous language

2) ChatGPT: Artificial intelligence chatbot capable of retrieving the targeted words when prompted

BP	Matrix	Extraction Method	Chromatography	Ionization Method	Polarity	Regression Model
BP-0001	Plasma	protein precipitation	reversed phase	turbo ionspray	positive	linear

RegEx, BERT, and ChatGPT outputs compared and stored into CSV for best results

Database



Node Class:

- Formats attributes as nodes
- Follows graph database relationship structure

GraphPopulator Class:

- Uses nodes generated from node class to populate database
- Omits "repeated" nodes, creating relationships between BPs

As the graph database grows, more relationships are established between existing chemicals, which will facilitate future drug development.

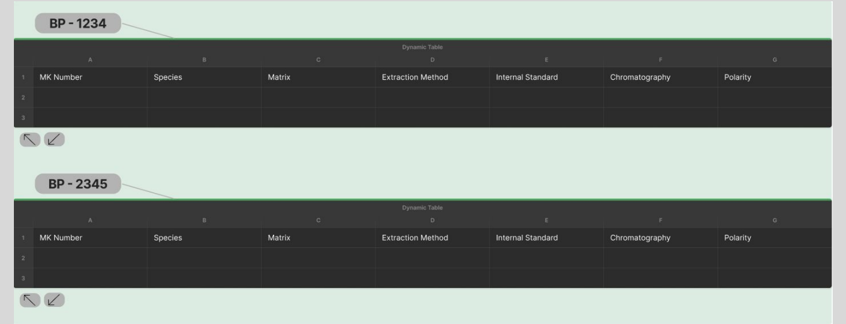
Conclusion

We created an app that…

- Extracts data of interest from provided BP documents using RegEx, Bert, and ChatGPT
- Maps extracted data to Neo4j-hosted graph database
- Provides a front-end application through React that has upload, search, and download functionalities

Next Steps

- Incorporate dynamic tables into search results page (arrangeable and filterable)



BP - 1234

Dynamic Table

	A	B	C	D	E	F	G
1	MK Number	Species	Matrix	Extraction Method	Internal Standard	Chromatography	Polarity
2							
3							

BP - 2345

Dynamic Table

	A	B	C	D	E	F	G
1	MK Number	Species	Matrix	Extraction Method	Internal Standard	Chromatography	Polarity
2							
3							

- Modeling improvements (bettering parsing scripts & BERT model)

Hugging Face

Acknowledgements

Merck: Xiang Yu, Terri Bui, Kenneth Wilson, Rajesh Desai

Purdue Data Mine: Dr. Mark Ward, Maggie Betz, Nicholas Rosenorn, Kali Lacy, Mihir Ahlawat

Fall Semester Members: Dante Moon, Yi Zhang, Shah Mahir

Sarah Firestone, Brian Tam, Adhvitha Sivaganesh, Arsh Batth, Jagannath Prabhakaran, Cheyenne Ward, Pradyun Kamaraju, Lingyu Li, Varun Manish Karlekar, Jayden Cheung