Function annotated from sequence alignment (FAFSA)

ChemE 545/546: UW DIRECT Courses for Data Scientists

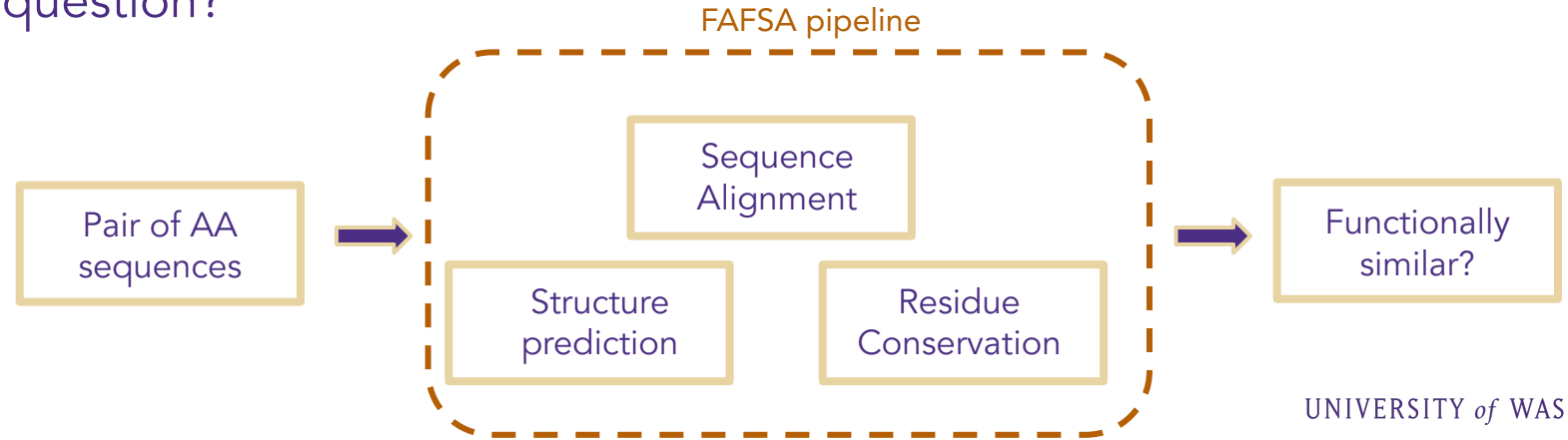
Group members: Humood, Logan, Ryan, Amin, Chau
15/03/23

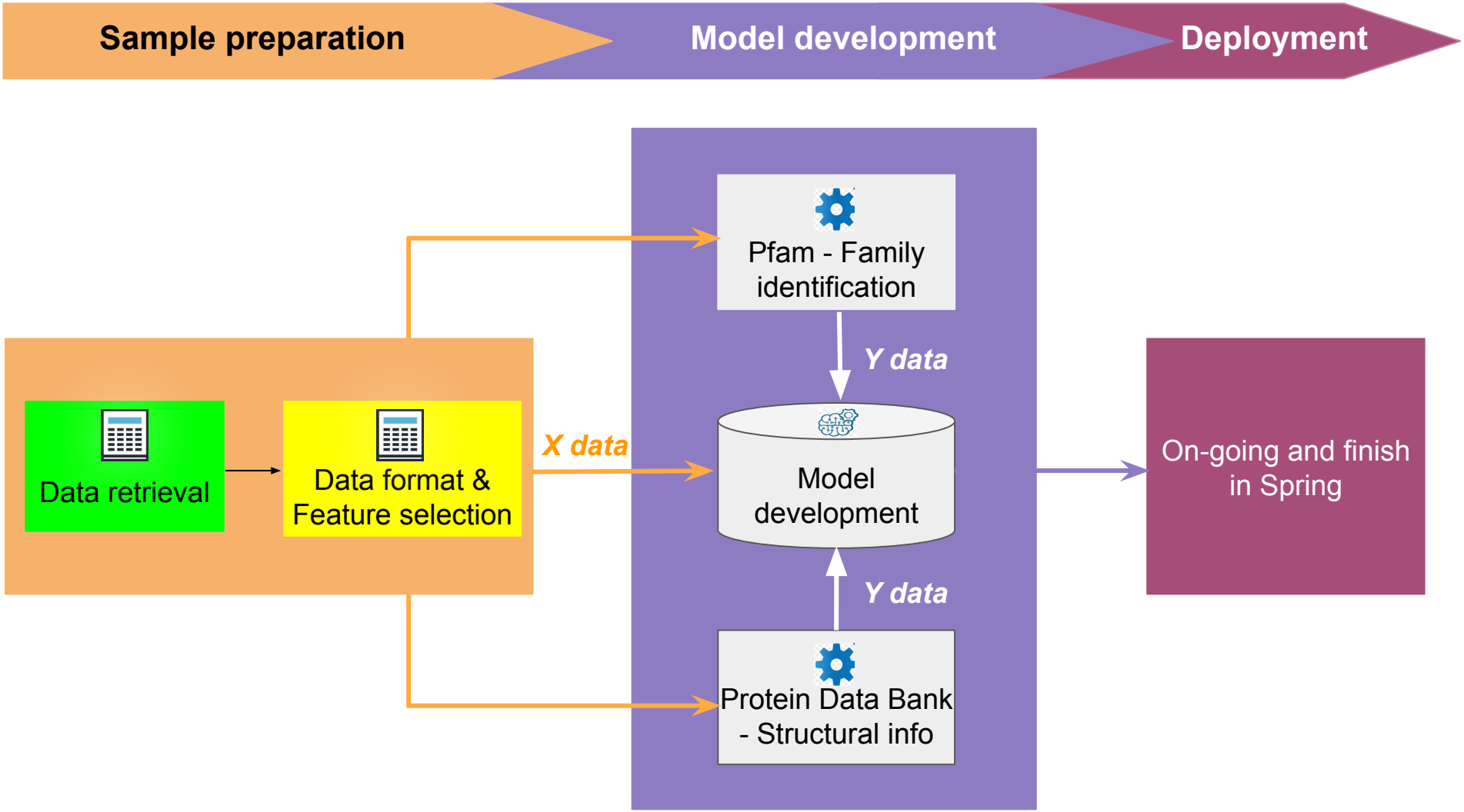
Why is this useful?

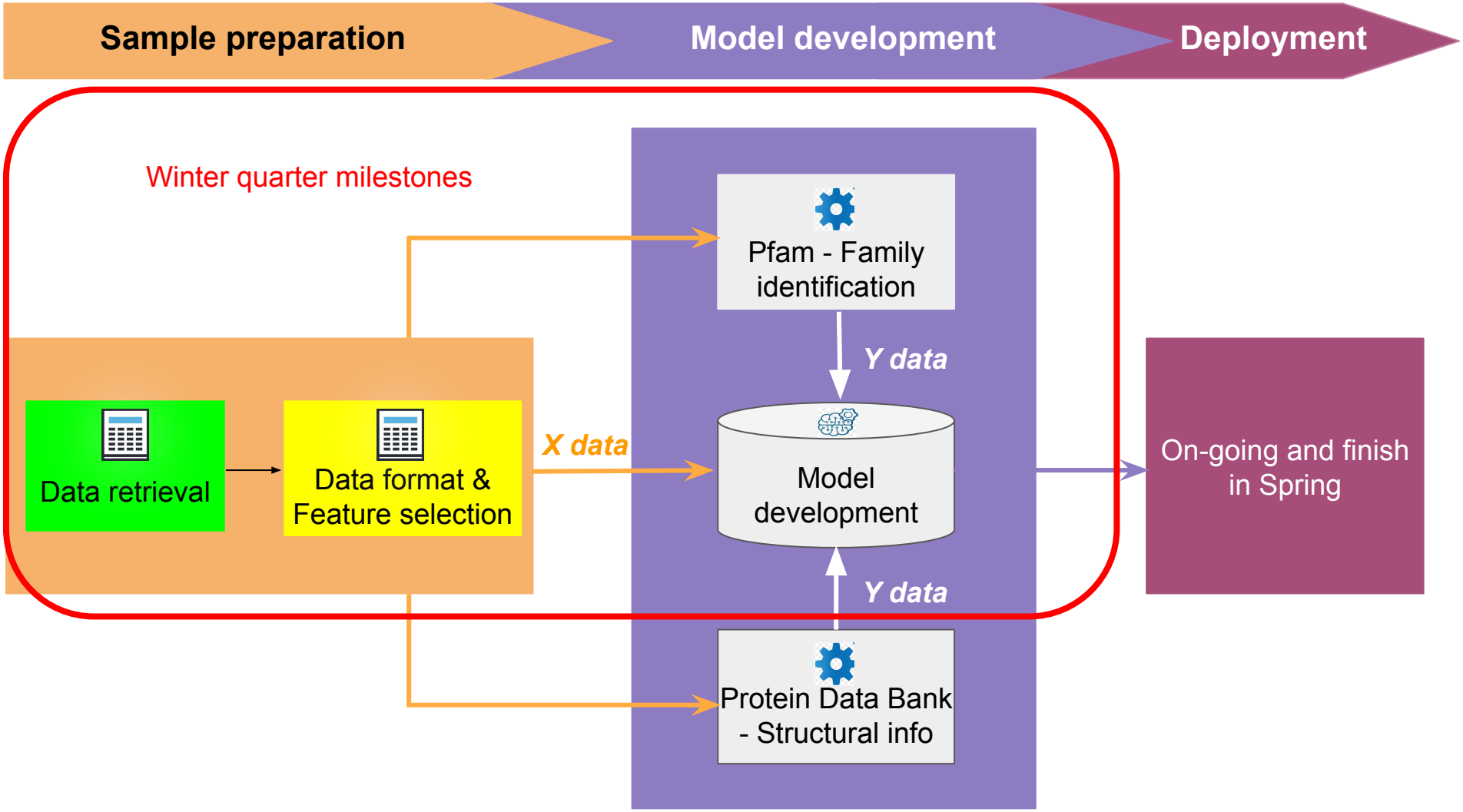
Protein pair validation is problematic in the current state:

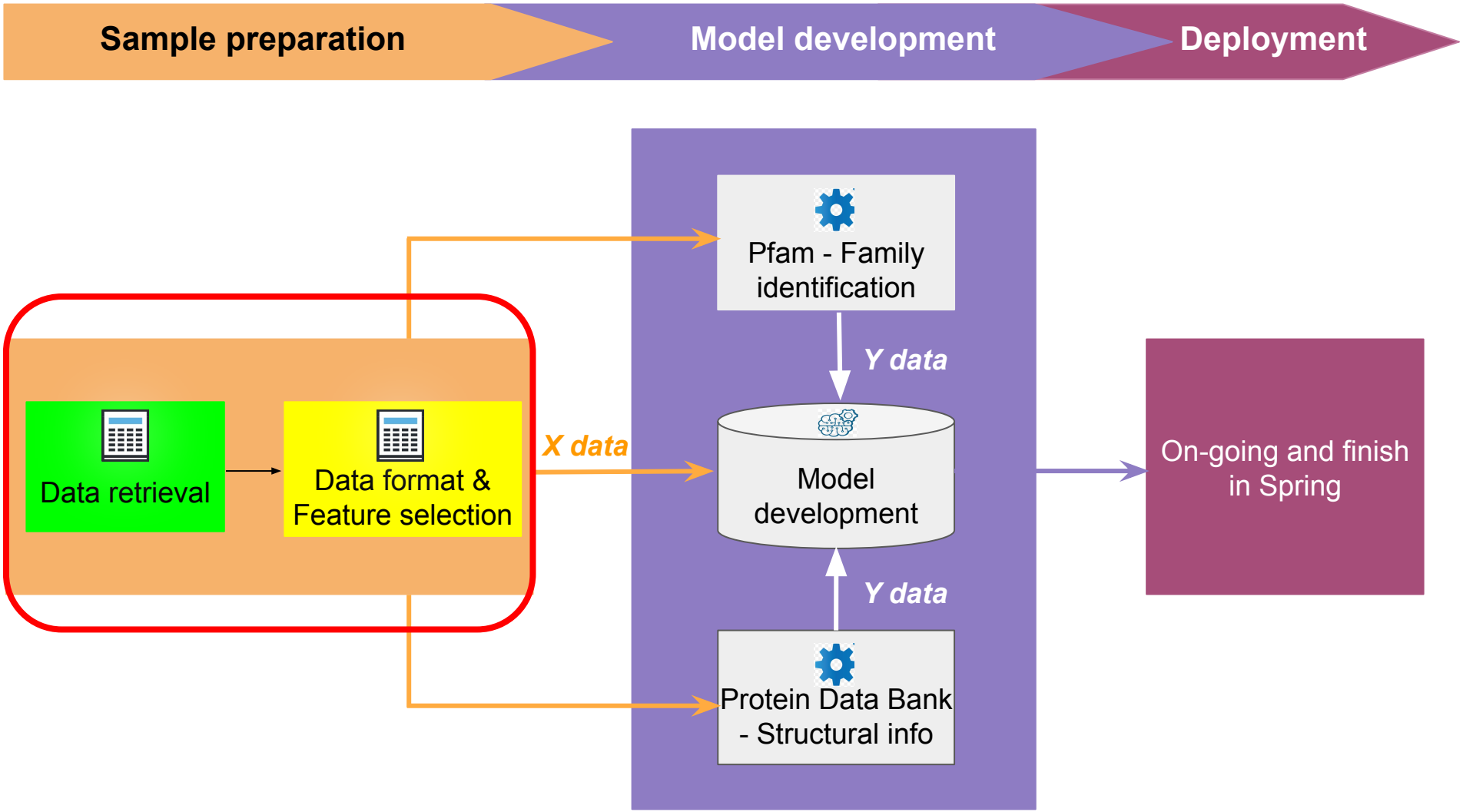
-It's time consuming and resource intensive
-Proteins are related through many non-trivial parameters
-Many software applications relate proteins through some of these parameters

What if one tool could combine these resources to answer a simple question?




Sample preparation
Model development
Deployment
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
Y data
Protein Data Bank - Structural info
On-going and finish in Spring

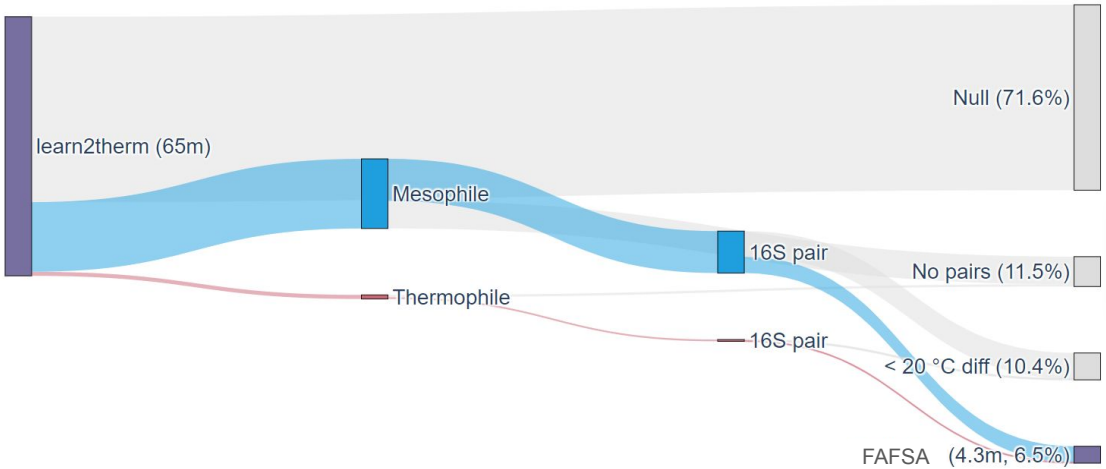

Sample preparation
Model development
Deployment
Winter quarter milestones
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
Y data
On-going and finish in Spring
Protein Data Bank - Structural info


Sample preparation
Model development
Deployment
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
Y data
On-going and finish in Spring
Protein Data Bank - Structural info

Data retrieval from DuckDB

-Protein pairs derived from meso/thermophilic organisms with high 16S alignment.

Protein Representation

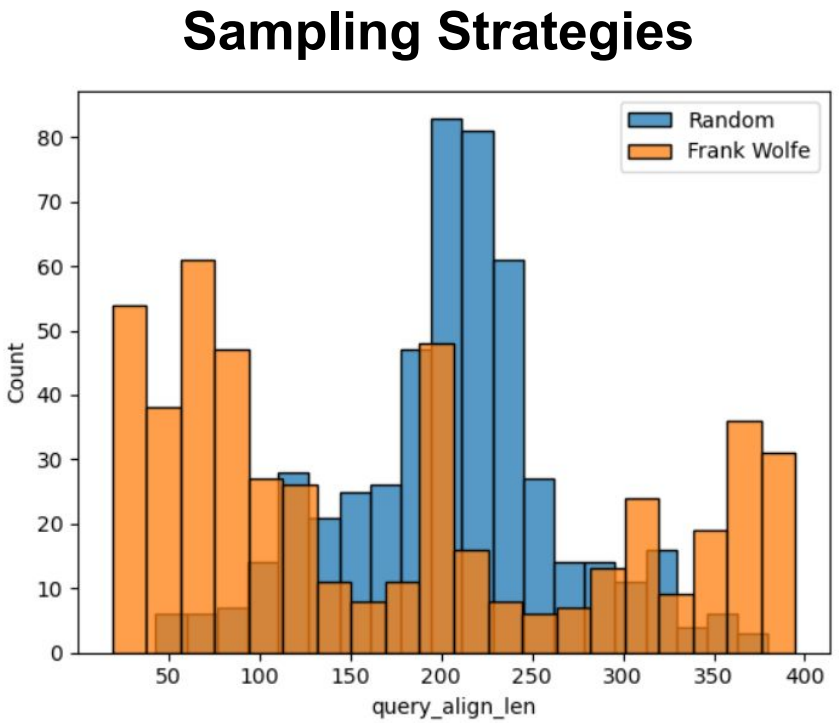


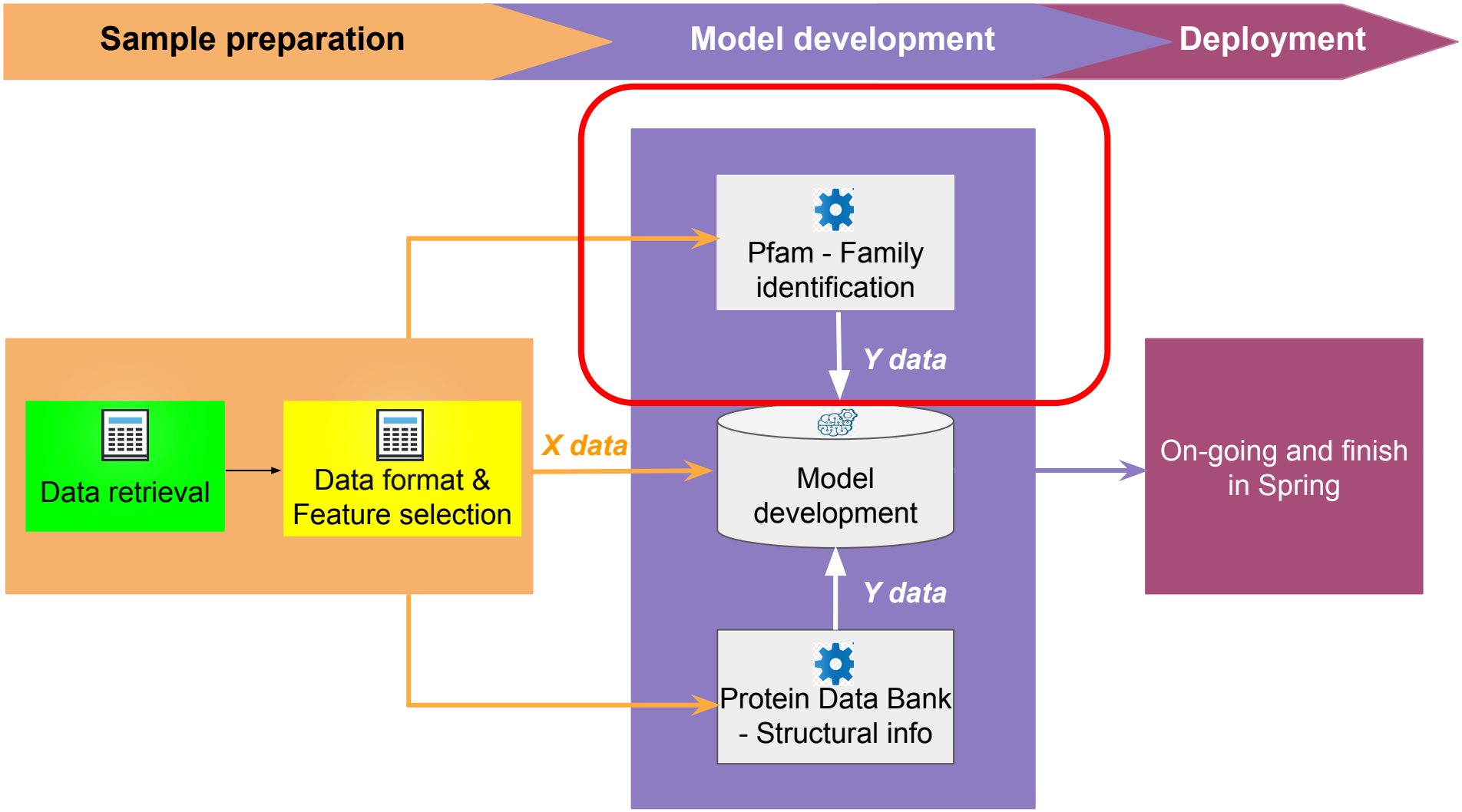
FAFSA

- **> 20°C growth temp difference**
- **4.3 million proteins**
- **54 million pairs**
- **1.8k species**

Maximizing sample information

- Many features have large tailed distributions.
- D-optimal design (Frank-Wolfe) creates a training set with more information about fringe data points.
- This is computationally expensive because there are multiple features.



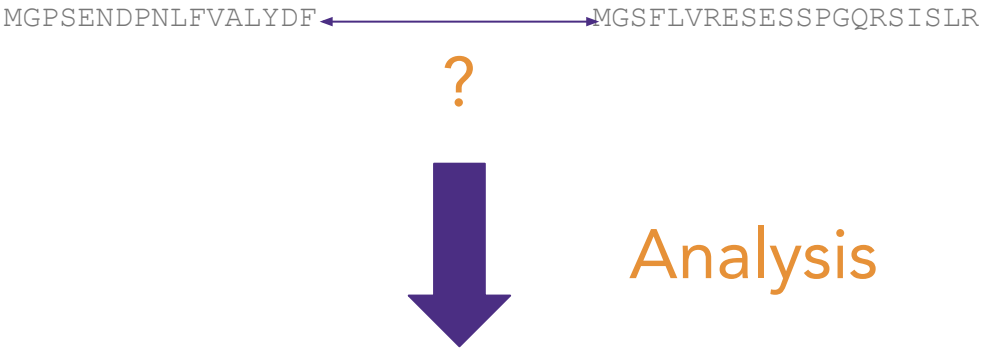

Sample preparation
Model development
Deployment
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
Y data
Protein Data Bank - Structural info
On-going and finish in Spring

Finding functional information

Goal:

1) Use HMMER and pfam to compute family information for protein pairs
2) Parse and filter the output generated to make dataframe if a pair is functional or not

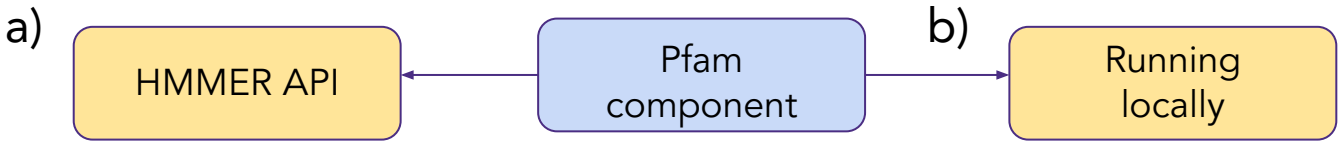
Upstream: Sampling component
Downstream: ML function prediction



1) Yes, they're functional pairs. Appends domain HMMER information to dataframe
2) No, they're not functional pairs. Appends a false boolean to functional pair

Computing functional information from protein pairs

Given two sequences, are we able to see if they're functional or not?

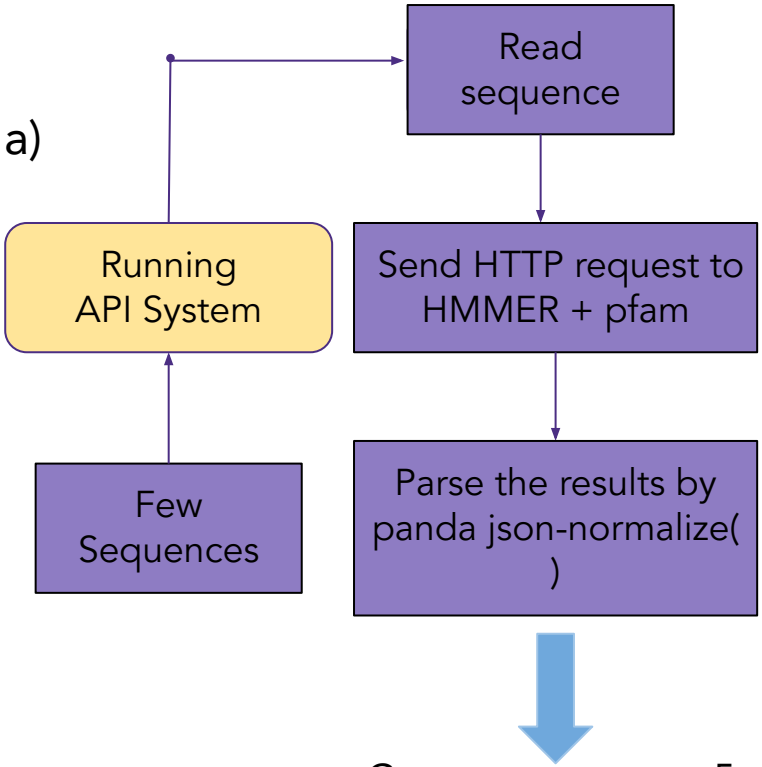


		Run with API	Run locally
1	User has a few sequences	✓	✗
2	Users want few dependencies	✓	✗
3	User doesn't want server variability	✗	✓

This choice is dependent on the user!

Computing functional information from protein pairs

a)




```python
def hmmerscanner(df: pd.DataFrame, k: int):

    if k > 300:
        print("Use local function for the number of sequences more than 300.")
        return pd.DataFrame()

    # Create an empty DataFrame to store the results.
    results_df = pd.DataFrame()

    # Loop through the sequences to check them.
    for i in range(k):
        # This is for meso protein sequences; we can change that in the future according to our request.
        sequence = df['m_protein_seq'][i]

        # Send an HTTP request to the HMMER API to get information for the current sequence.
        url = 'https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan'
        headers = {'Content-Type': 'application/x-www-form-urlencoded',
                   'Accept': 'application/json'}
        data = {'hmmdb': 'pfam', 'seq': f'>seq\n{sequence}'}
        data = urllib.parse.urlencode(data).encode('ascii')
        response = requests.post(url, headers=headers,
                                 data=data, allow_redirects=False)
        redirect_url = response.headers.get('Location')

        if redirect_url is None:
            # If the server doesn't work, show this error.
            print("Error: No redirect URL found in response.")
        elif redirect_url == 'late':
            # Raises an exception if the status is pending for too long.
            response.raise_for_status()
            time.sleep(180)
            raise IOError("Error notice after 3 minutes.")
        else:
            response2 = requests.get(redirect_url, headers=headers)

            # Put the results in the empty DataFrame.
            results = response2.json()
            hits = results['results']['hits']
            dfff = pd.json_normalize(
                hits, 'domains', ['acc', 'name', 'score', 'evalue', 'pvalue', 'desc'])
            dfff.insert(0, 'sequence', sequence)
            dfff = dfff.set_index('sequence')
            results_df = pd.concat([results_df, dfff])
            if redirect_url == 'late':
                # Raises an exception if the status is pending for too long.
                response2.raise_for_status()
                time.sleep(180)
                raise IOError("Error notice after 3 minutes.")

    return results_df
```

Results from API System

sequence	alisqacc	aliIdCount	alirfline	is_included	alihmmname	bitscore	display	ievalue	alisqto	aliSim	...	pvalue
NFRLWPTS		35		1	DCD	28.693352	0.0	6.0e-07	111	0.660550	...	-36.480112
NFRLWPTS		23		0	DCD	15.748980	NaN	0.0051	175	0.672131	...	-36.480112
NFRLWPTS		30		1	dUTPase	40.322865	1.0	0.0	178	0.663636	...	-32.277306
FMARREGR		41		1	AAA	51.984688	1.0	1.0e-13	175	0.690476	...	-40.228785
FMARREGR		41		1	AAA_5	43.017456	0.0	0.0	171	0.669118	...	-34.13238

823 rows × 221 columns

Results from API System

```
# Show the first row and the information of the columns
pd.set_option('display.max_rows', None) # Show all details of the first row
print(Result_test.iloc[0])
```

```
alisqacc
aliIdCount                                                                27
alirfline
is_included                                                               1
alihmmname                                                       Sigma70_r2
bitscore                                                         67.940552
display                                                                1.0
ievalue                                                                0.0
alisqto                                                                113
aliSim                                                           0.884058
jali                                                                   113
bias                                                                  1.66
ienv                                                                    45
cevalue                                                                0.0
significant                                                            1.0
alimline                  l++ + p+v++l rr+l+++a ae+++Qe+++ +wr+++rfdp++g...
alihmmfrom                                                              1
clan                                                             CL0123
aliL                                                                  206
alihindex                                                           16681
is_reported                                                             1
alintseq
jenv                                                                  114
alimmline
alihmmacc                                                      PF04542.17
oasc                                                                0.96
aliaseq                   LYDLLAPRVYGLIRRVLRDPALAEEVTQEVLVEVWRRAARFDPAQG...
alihmmto                                                               70
aliId                                                            0.391304
alippline                 688899****************************************...
alimodel                  lverylplvkrlarrllgsgadaeDlvQegflrlwraverfdperg...
aliM                                                                   71
iali                                                                   45
alicsline                 HHHHHHHHHHHHHHHHCTCHHHHHHHHHHHHHHHHHHHHGCCTTTC...
aliSimCount                                                            61
alihmmdesc                                                     Sigma-70 region 2
alisqdesc
```

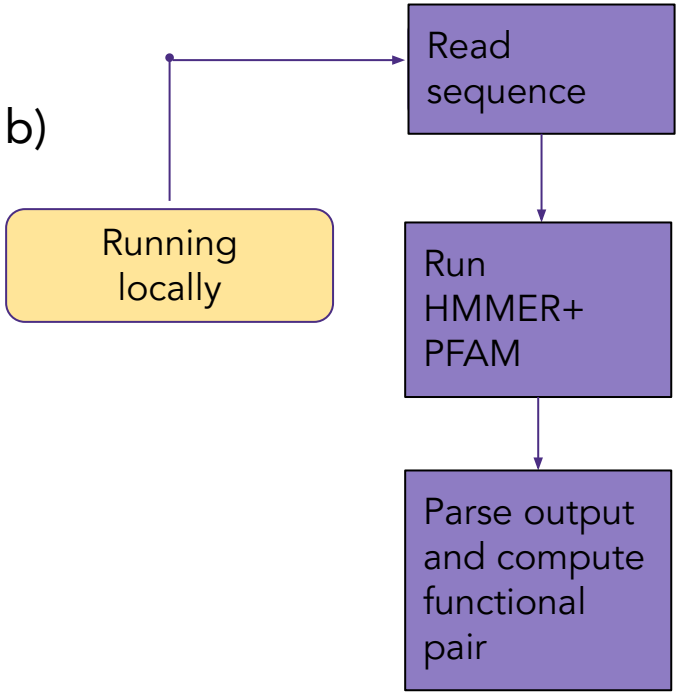
```
outcompeted                                                             0.0
alisqname                                                              >seq
alisqfrom                                                               45
uniq                                                                   1.0
aliN                                                                    70
acc                                                            PF04542.17
name                                                             Sigma70_r2
score                                                                68.9
evalue                                                                0.0
pvalue                                                           -52.678535
desc                                                           Sigma-70 region 2
```

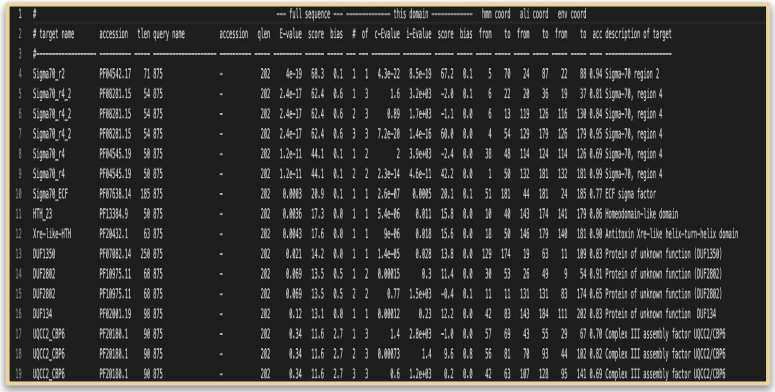
```
Name: MAESGTSRRADHLVPVPGPDAEPPAVADELLRAVGRGDEQAFGRLYDLLAPRVYGLIRRVLRDPALAEEVTQEVLVEVWRRAARFDPAQGSANAWVFTIAHRRAVDRVRAE
QKAADRTVRAGAAALDSPYDSVADEVSGRLERRQVRHCLDALTGLQREVVTLAYYQGHSYPQVAELLKTPLGTVKTRMRDGLIRLRDCLGVEATA, dtype: object
```

Computing functional information from protein pairs

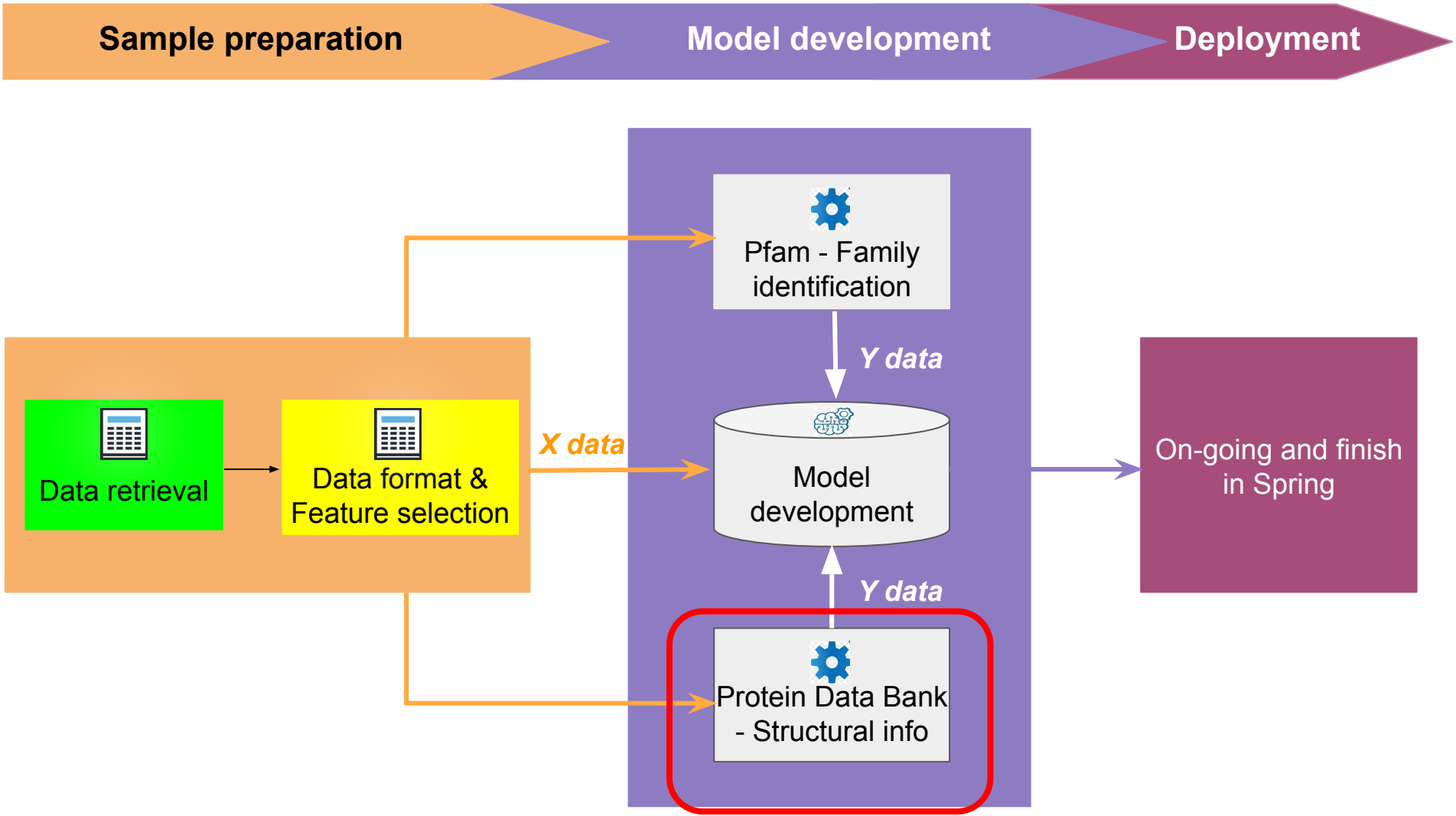
Given two sequences, are we able to see if they're functional or not?

b)





Goes to component 5

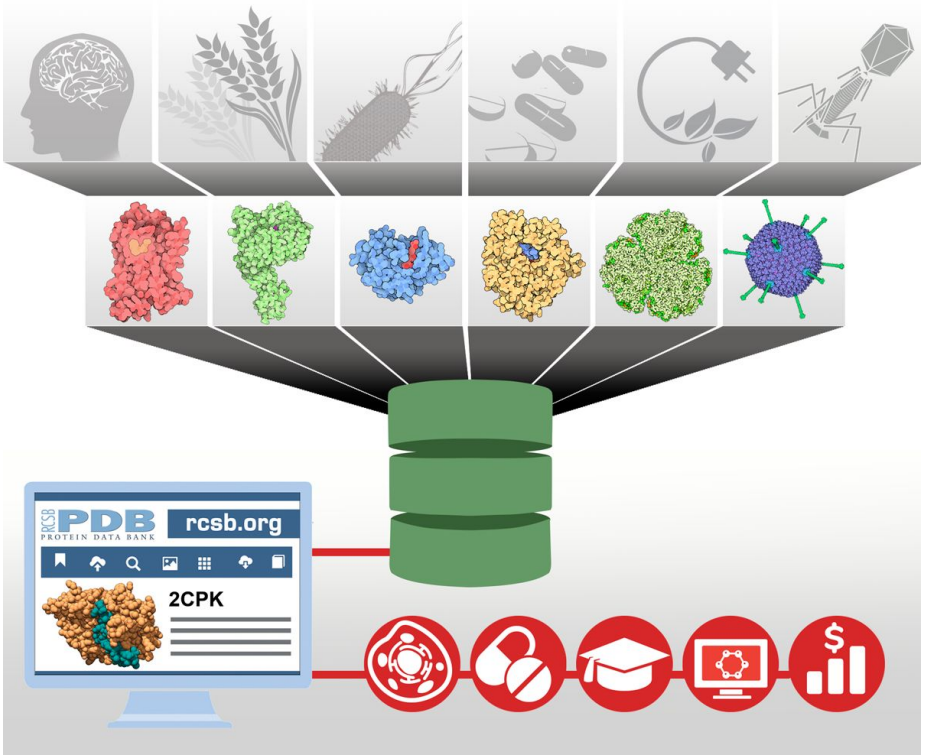

Sample preparation
Model development
Deployment
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
On-going and finish in Spring
Y data
Protein Data Bank - Structural info

Finding molecular structure information

Goal:

1) Query into Protein Data Bank (PDB)
2) Obtain lists of proteins with structural similarity

Upstream: Sampling component
Downstream: ML function prediction



Finding molecular structure information

Progress:

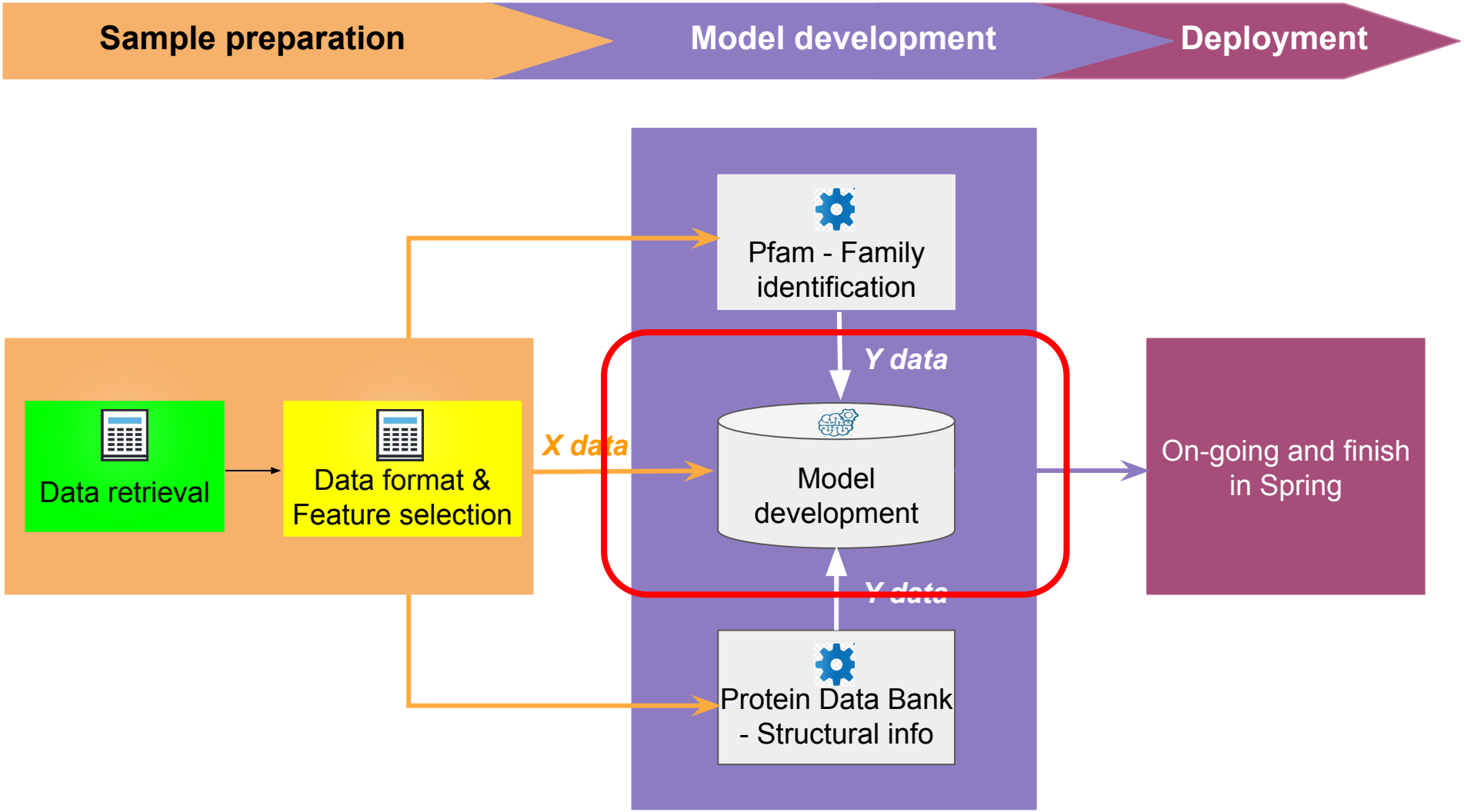
1) Successfully access PDB
2) Obtain lists of proteins with sequence similarity

Next:

1) Modify the search and return to obtain output based on structures.

MLLSDRDLVSEIKSGDLSLEPFEPALLQPS...

	identifier	score	sequence_identity	evalue	bitscore
0	2QXX_1	1.0	0.790	9.055000e-94	304
1	4A6A_1	0.992308	0.784	3.197000e-93	302
2	2QLP_1	0.842308	0.788	1.157000e-79	263
3	1XS1_1	0.365385	0.431	1.379000e-36	139
4	2V9X_1	0.361538	0.426	3.537000e-36	138
5	1XS4_1	0.357692	0.426	9.071000e-36	137
6	1XS6_1	0.357692	0.426	9.071000e-36	137
7	2J4Q_1	0.357692	0.426	9.071000e-36	137
8	2J4H_1	0.35	0.426	3.184000e-35	135
9	4XJC_1	0.2	0.338	6.627000e-22	96


Sample preparation
Model development
Deployment
Pfam - Family identification
Y data
Data retrieval
Data format & Feature selection
X data
Model development
On-going and finish in Spring
Y data
Protein Data Bank - Structural info

Training and validating a model to predict protein function

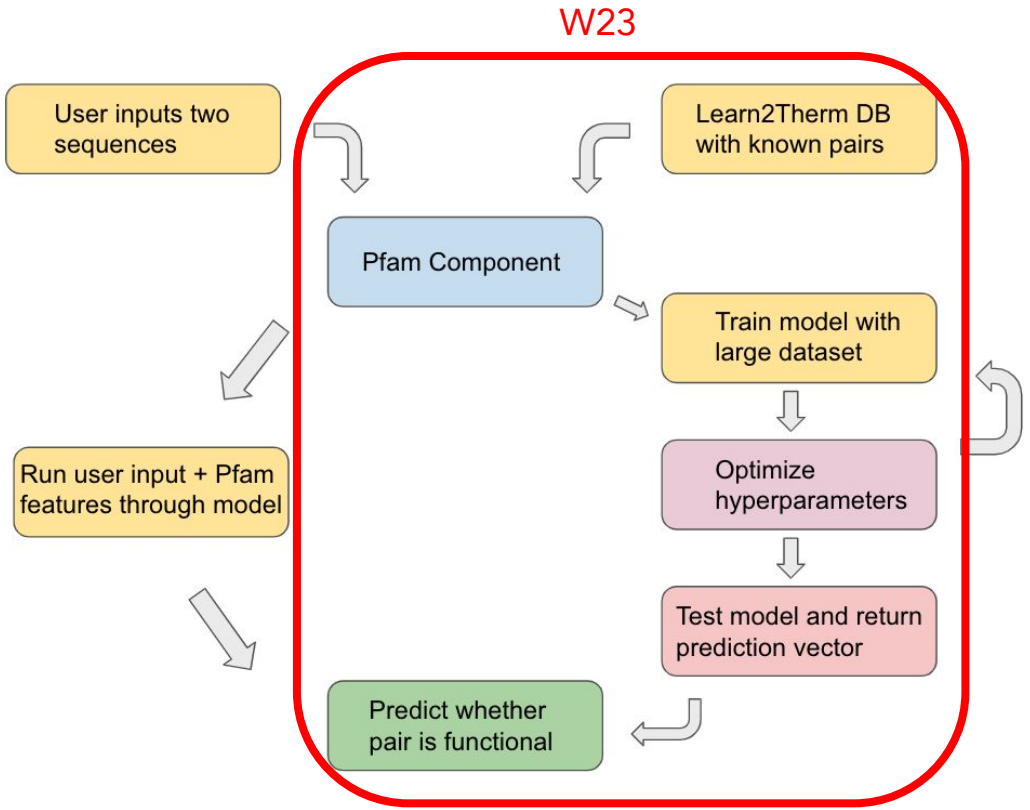
Goals:

1. Train a RandomForestClassifier on Learn2Therm DB in order to predict whether a protein pair is functional

2. Create robust pipeline that can handle expanded feature space

Upstream: Pfam + Data Retrieval components

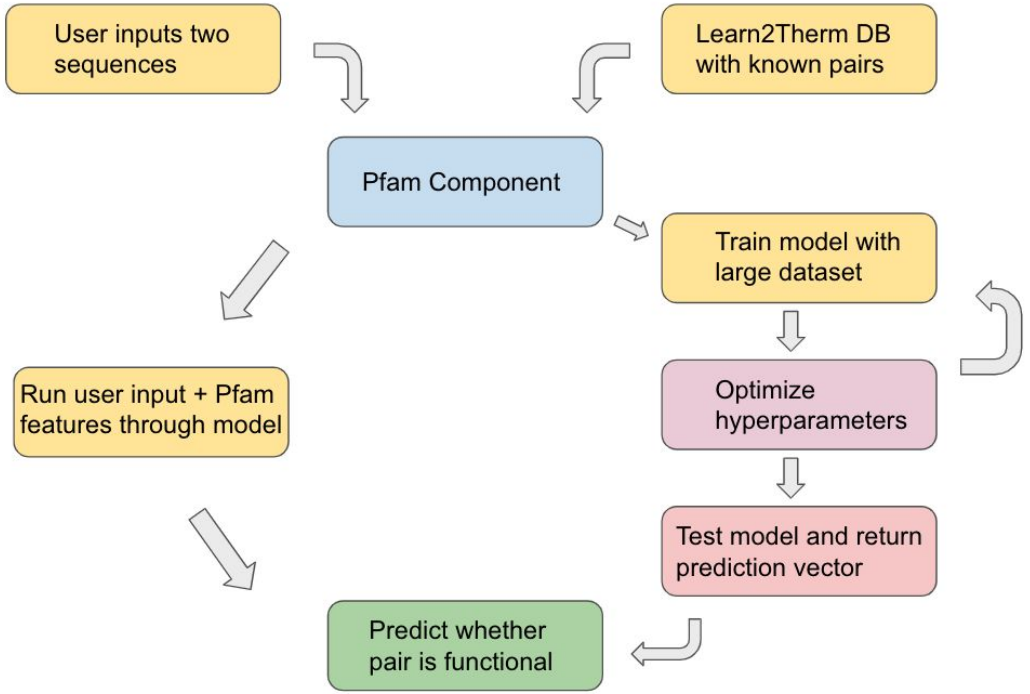
Downstream: Functionality analysis (future work)



Training and validating a model to predict protein function

Ongoing work:

1. Combine output of MSA and structural scraping to improve strength of model
2. Use model to predict functionality from a user input of two AA sequences
3. Develop metrics of functionality to be reported in addition to Boolean classification



Summary

- Obtained a training dataset from an existing SQL database.
- Ran the HMMER algorithm against Pfam database.
- Queried sequence and got output from PDB.
- Trained and validated a RandomForestClassifier to predict protein pair functionality.

Future work

- Make a technical validation pipeline that explores MSAs (evolutionary history) and structure (e.g. SWISS-model) purely from sequence pair inputs from users
    - Component 0-2: Run BLAST locally to obtain local alignment from users' protein pairs + Sample analytics
    - Component 3: More sophisticated parsing to get functional information for various pairs
    - Component 4: Integration with downstream and upstream components
    - Component 5: Increase feature space of pair functionality prediction model

Acknowledgement

Consultation and Code Review:

David Beck (PhD)
Nels Schimek
Orion Dollar

Development of Learn2Therm DB & EXTENSIVE project consultation:

Evan Komp

Questions?