HAUPTMAN-WOODWARD
MEDICAL RESEARCH INSTITUTE

Building a Database of Crystallization Conditions

Max Dudek, Sarah EJ Bowman

HWI, University at Buffalo Department of Biochemistry

University at Buffalo

ABSTRACT

Crystallization is a critical step in the determination of biomolecular structures. To crystallize, a protein is placed in a particular set of chemical compounds (a crystallization cocktail) which may differ significantly between structures. At the High-Throughput Crystallization Screening Center at HWI, a 1536-cocktail microbatch-under-oil screen is used to determine conditions for protein crystallization. These initial crystal hits can be optimized for better crystals. Though automatic scoring systems are being worked on, currently the 1536 wells must be checked manually for crystals, making the screening process time consuming. If a method could be developed to narrow down the set of initial screening cocktails, the crystallization process would take much less time.

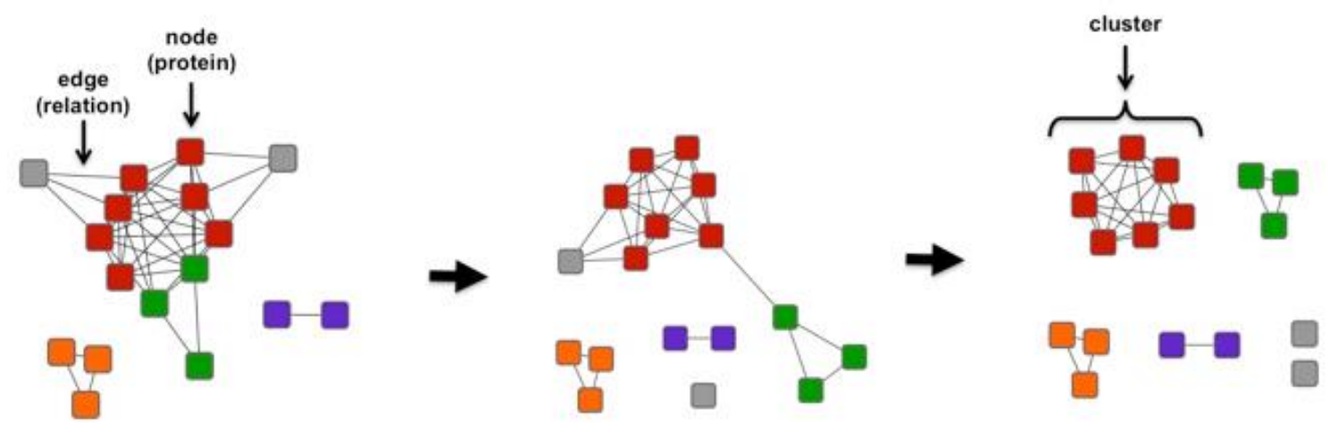
To find correlations between protein sequence and crystallization conditions, a dataset needs to be created, which can be computationally searched through to find patterns in the data. No such database currently exists in a consistent format with standardized chemical names. The "crystallization details" field in the Protein Data Bank (PDB) consists of a sentence describing the conditions, and is not easily searchable. Using the API of the PDB, we have created a Python-based database of crystallization conditions and protein sequences. The database is able to parse the crystallization details and extract the individual compounds and concentrations in a consistent format. It also is able to standardize the names of the chemical compounds using a compound dictionary.

OBJECTIVES

In this project, we created an easily searchable database which contained information on crystallization conditions in a consistent format, and sequences for statistical analysis. We also created a compound dictionary to standardize the names of chemical compounds. Information was downloaded from the Protein Data Bank (PDB) and the xTuition database from HWI.

APPLICATIONS

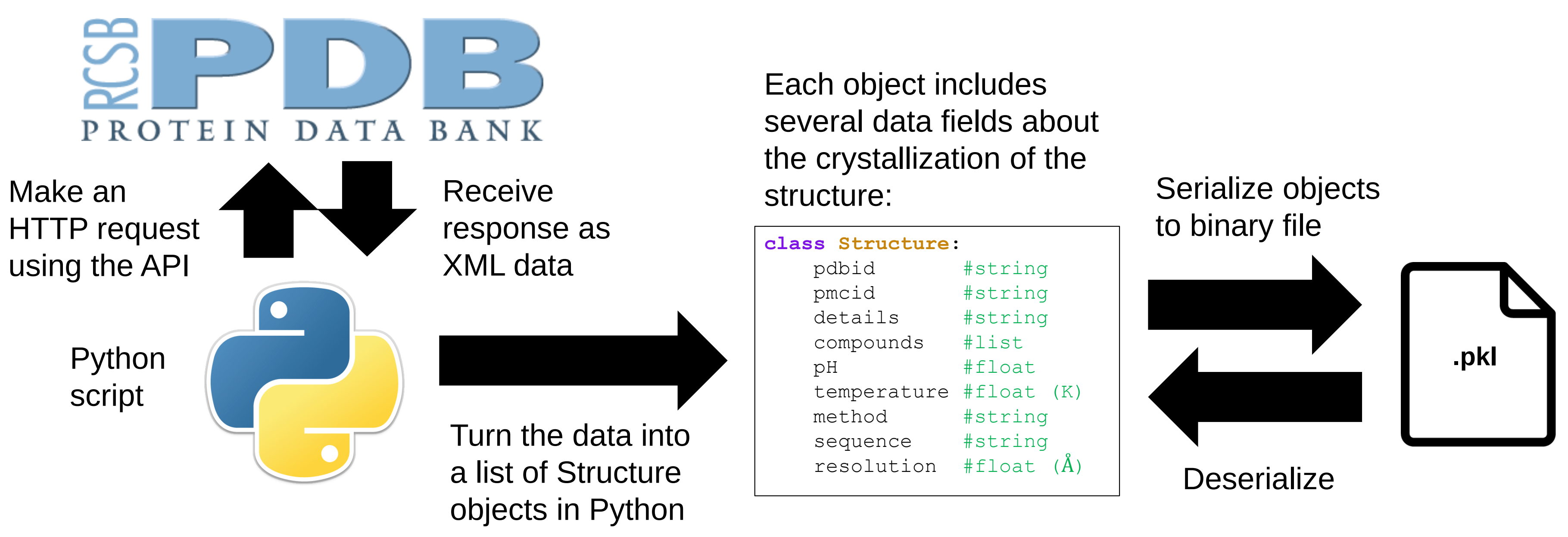
This database can be used to create Sequence Similarity Networks (SSN), which cluster structures together based on their sequence:



Clusters can also be created for crystallization conditions, and this can allow patterns to be discovered between sequence and conditions, which would allow for predictions of conditions based on sequence. Furthermore, the database can be coupled with HWI's autoscoring system, for the subset of structures that were initially crystallized at HWI.

METHODS – BUILDING THE DATABASE

Step 1: Download structure data from the PDB and create Python objects


PROTEIN DATA BANK
Make an HTTP request using the API
Receive response as XML data
Python script
Turn the data into a list of Structure objects in Python
Each object includes several data fields about the crystallization of the structure:
Serialize objects to binary file
.pkl
Deserialize

```
class Structure:
    pdbid          #string
    pmcid          #string
    details        #string
    compounds      #list
    pH             #float
    temperature    #float (K)
    method         #string
    sequence       #string
    resolution     #float (Å)
```

Step 2: Parse crystallization details into consistent list of compounds and concentrations

Below is an example of the information in the "crystallization details" field of a structure in the PDB:

```
15-18% (w/v) PEG 3350, 25mM MgCl2, 100mM NH4Cl, 5mM DTT and 0.1M MES, pH
6.5, VAPOR DIFFUSION, HANGING DROP, temperature 277K
```

A detail parsing function extracts compounds and concentrations from ~113k structures in the PDB which include details about chemical compounds, and creates a list of every compound followed by its concentration

```
['PEG 3350', '16.5% w/v', mgcl2', '25',
'nh4cl', '100', 'DTT', '5', mes', '100.0']
```

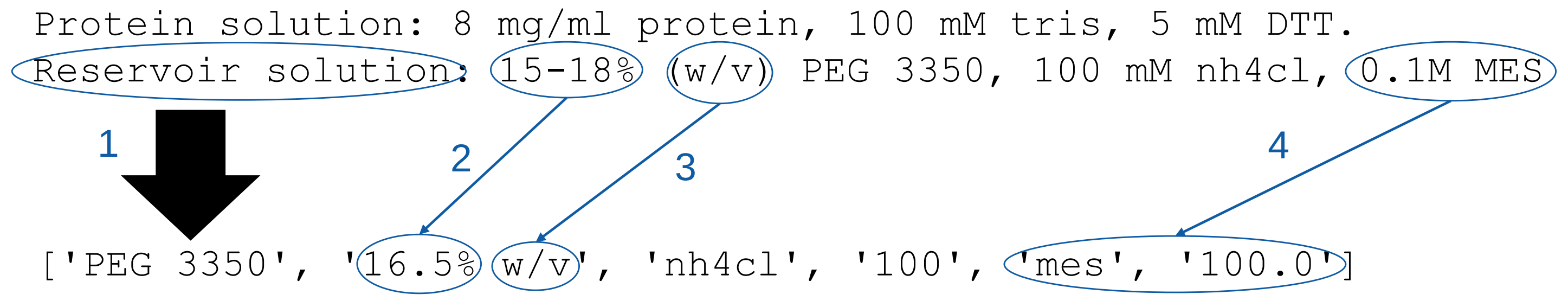
A compound dictionary standardizes the names of compounds.
~88k structures have all of their compounds recognized by the dictionary

```
['PEG 3350', '16.5% w/v', 'magnesium chloride', '25',
'ammonium chloride', '100', 'DTT', '5', 'MES', '100.0']
```

Features of the detail parsing function:

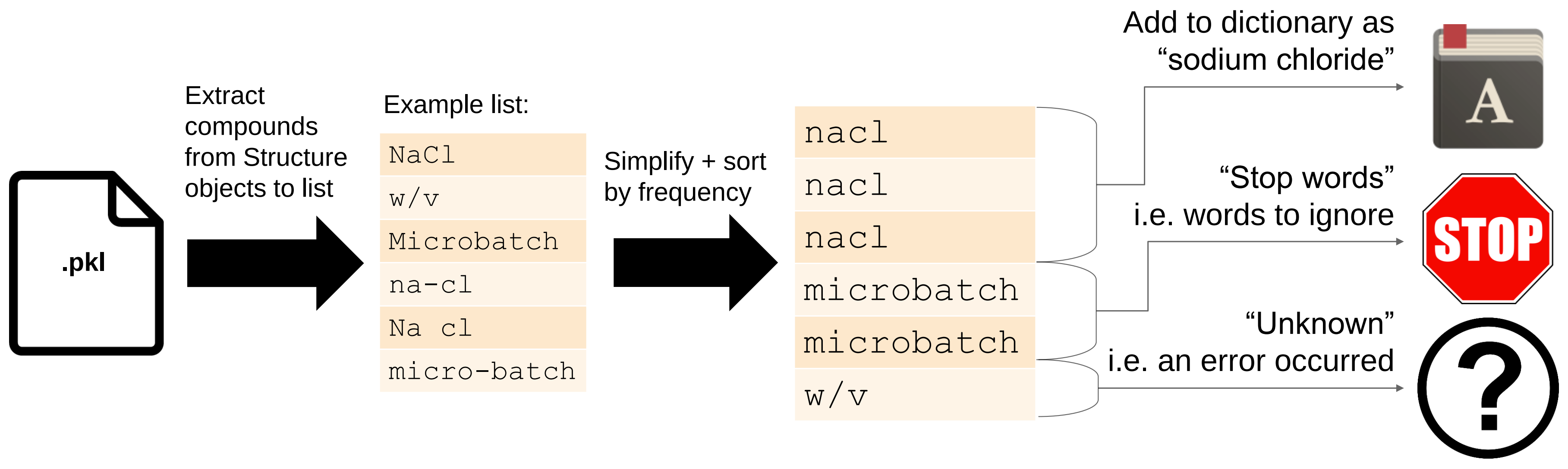
1. Isolate reservoir solution from protein solution/cryoprotectant/soaking
2. Average ranges of concentrations
3. Associate percent (%) concentrations with (w/v) or (v/v) specifications, if available
4. Convert molar (M) and micromolar (μM) concentrations to mM
5. Standardize names of compounds with chemical dictionary

Example details:



```
Protein solution: 8 mg/ml protein, 100 mM tris, 5 mM DTT.
Reservoir solution: 15-18% (w/v) PEG 3350, 100 mM nh4cl, 0.1M MES
['PEG 3350', '16.5% w/v', 'nh4cl', '100', 'mes', '100.0']
```

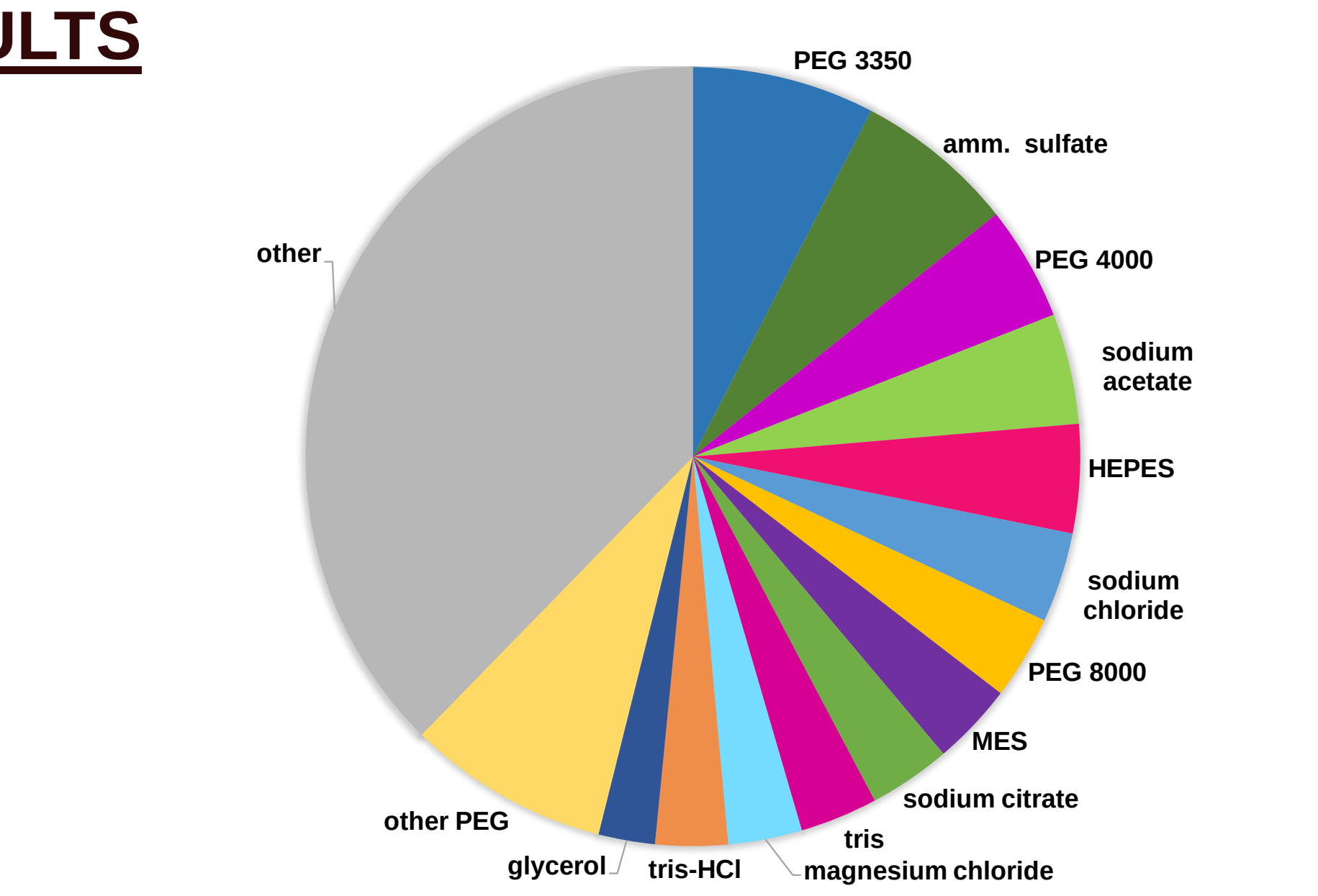
Step 3: Build a dictionary to standardize the names of chemical compounds


.pkl
Extract compounds from Structure objects to list
Example list:
NaCl
w/v
Microbatch
na-cl
Na cl
micro-batch
Simplify + sort by frequency
nacl
nacl
nacl
microbatch
microbatch
w/v
Add to dictionary as "sodium chloride"
A
"Stop words" i.e. words to ignore
STOP
"Unknown" i.e. an error occurred
?

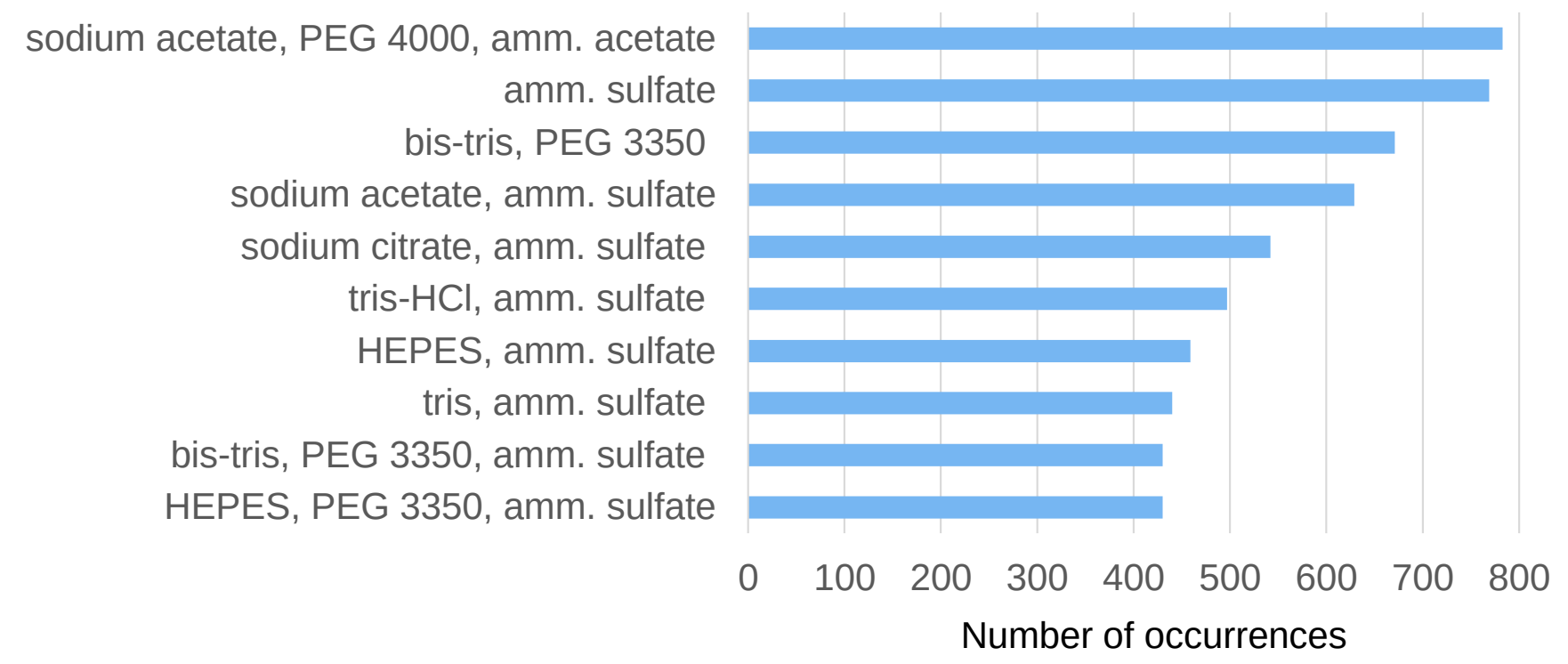
RESULTS

The Database

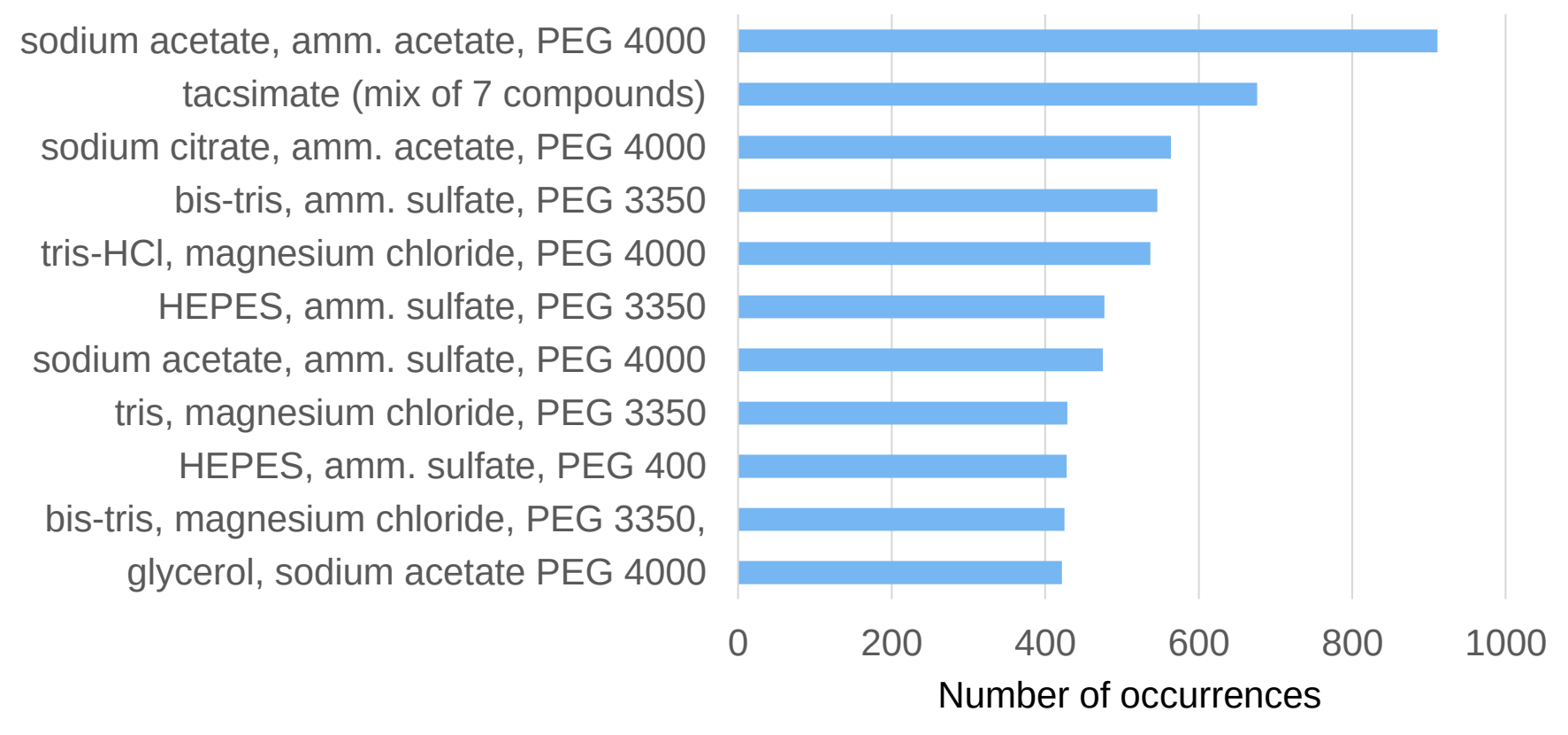
- There are ~142k structures total in the PDB
- ~113k structures include relevant chemical details
- ~88k (77%) are "sensible"
  - "Sensible" means that all compounds extracted from the detail parser are recognized by the dictionary – indicating that there were no errors
- In the compound dictionary:
  - All compounds appearing > 30 times are recognized
  - ~680 entries, or various names for compounds
  - ~350 unique compounds
  - 323k (88%) out of the 365k "compounds" extracted from the 113k structures are recognized



13 most common crystallization compounds (out of 350) in PDB
Out of all compounds extracted from the 88k "sensible" structures



Most common complete sets of compounds for crystallization in PDB
Out of the 88k "sensible" structures



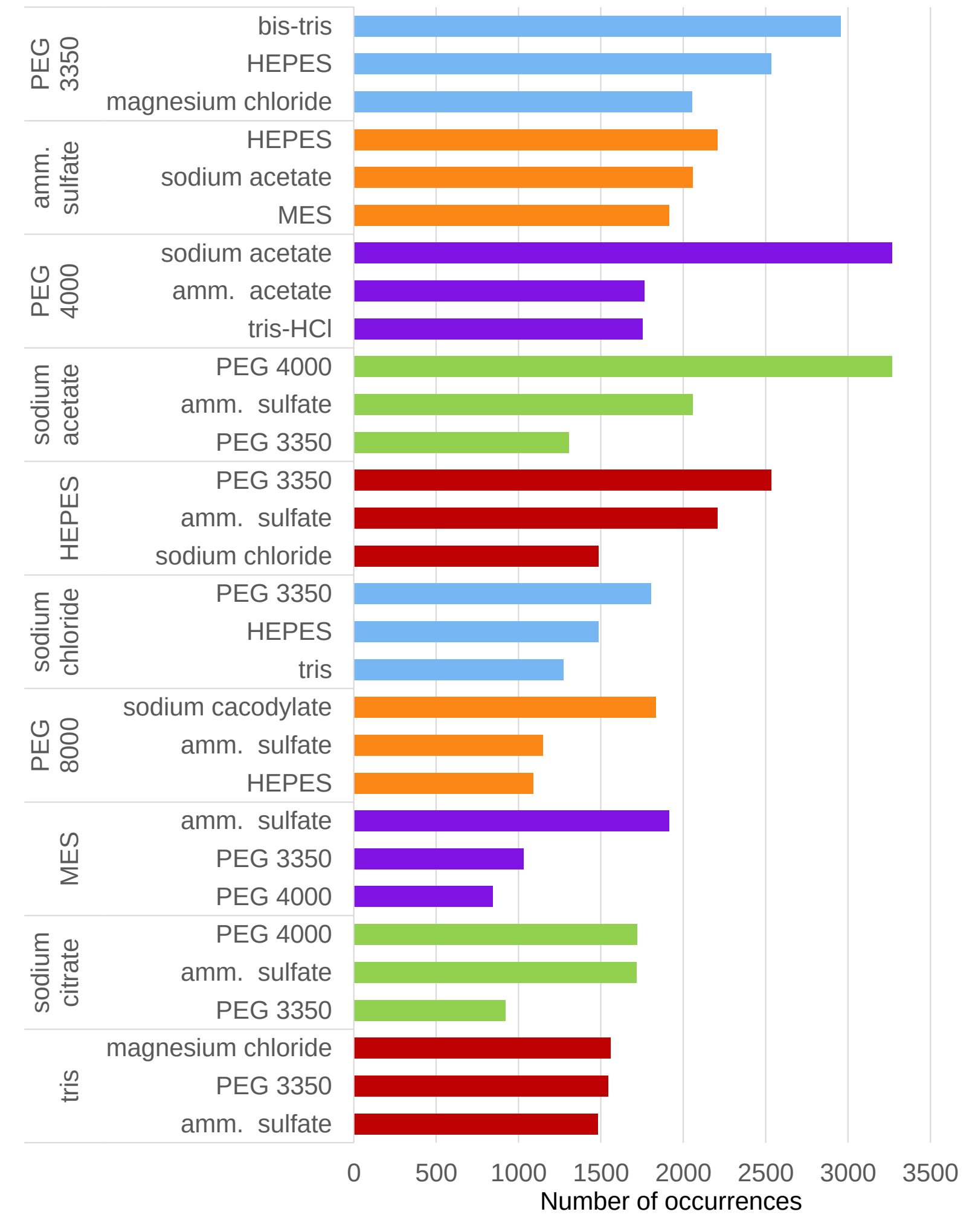
Most common 3-compound subsets within chemical sets
Out of the 88k "sensible" structures

Limitations

- Isolation of reservoir solution is unreliable
  - Compounds from protein solution may be included
  - Relevant compounds may be excluded
- Assumptions:
  - If only an anion is listed as a compound, the cation is assumed to be sodium
    - Ex. 0.1 M acetate = 100 mM sodium acetate
  - Assume the drop ratio is 1:1
  - Guesswork is required on compounds/mixtures with limited details
    - Ex. Assume "amino acids" = Morpheus amino acids
- Compound dictionary only recognizes 88% of extracted compounds
  - There are almost 200 unique unrecognized 'compounds' appearing 20-30 times in the PDB

Examples of Typos and Inconsistencies which the Detail Parser Can't Recognize

- Missing spaces/unexpected punctuation
  - `6% PEG 4000.30% glycerol` → `'PEG 400030', '% glycerol'`
  - `PEG 10 000` → `'PEG 10'`
- Not including enough information
  - `6% PEG, 5% glycerol` (No molecular weight specified)
- Unexpected concentration syntax
  - `PEG 4000 (1:3, W/V)` → `['PEG 4000', '1', 'w/v', None]`
- Mistakes and typos within the details
  - `PEG &k, 0.1X PEG MME 2000, pottasium chloride`



3 most common chemical 'partners' for the top 10 most common compounds
Out of the 88k "sensible" structures

GitHub Repository

A GitHub repository containing all of the code for this project, along with detailed documentation on how the database was created and how to use it, can be found at:

https://github.com/maxdudek/crystallizationDatabase

ACKNOWLEDGEMENTS

Berman, HM, et al. (2000). The Protein Data Bank, Nucleic Acids Research, 28 (1), 235.

Luft, JR, et al. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J. Struct. Biol. 142, 170.