Eric Pitman Summer Workshop in Computational Science

Intro to R

Project Introduction

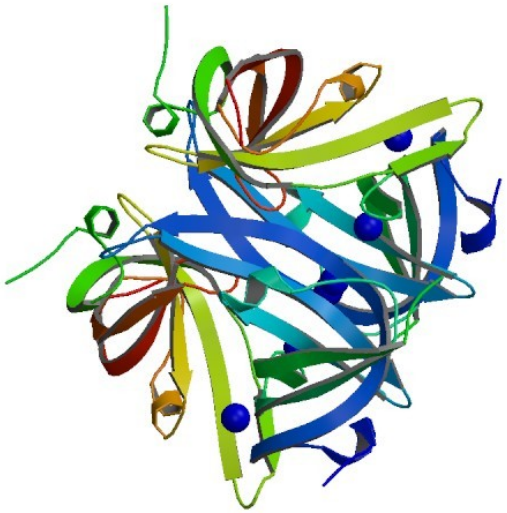
Jeanette Sperhac & Amanda Ruby

Center for Computational Research
University at Buffalo The State University of New York

Introducing the Workshop Project

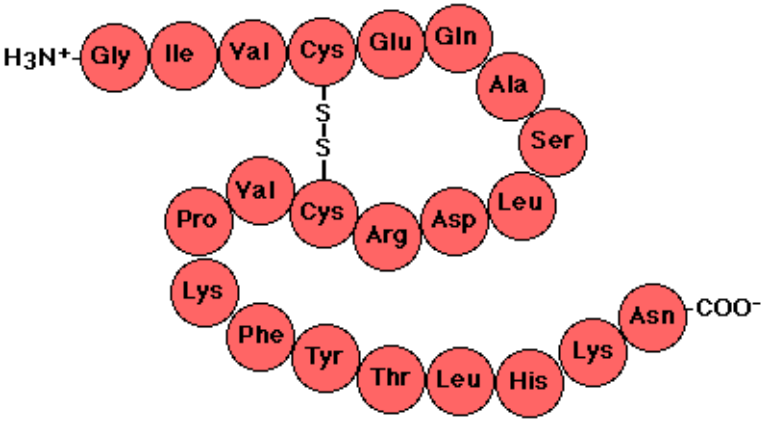
Here's what we'll cover:

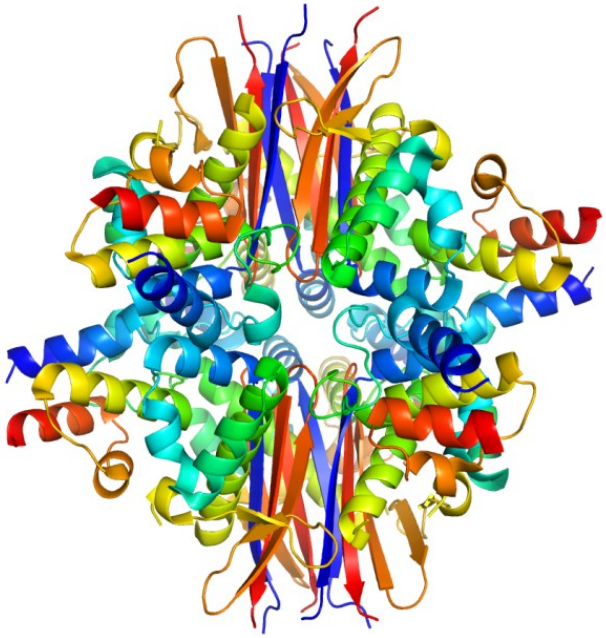
- The story of the HWI protein crystallization data
- The Questions
- So what's a classifier?
- Inside the dataset
- The Project in RStudio
- Exploring the Proteins
- What you'll need



Proteins

Proteins are large biological molecules, composed of long chains of amino acids



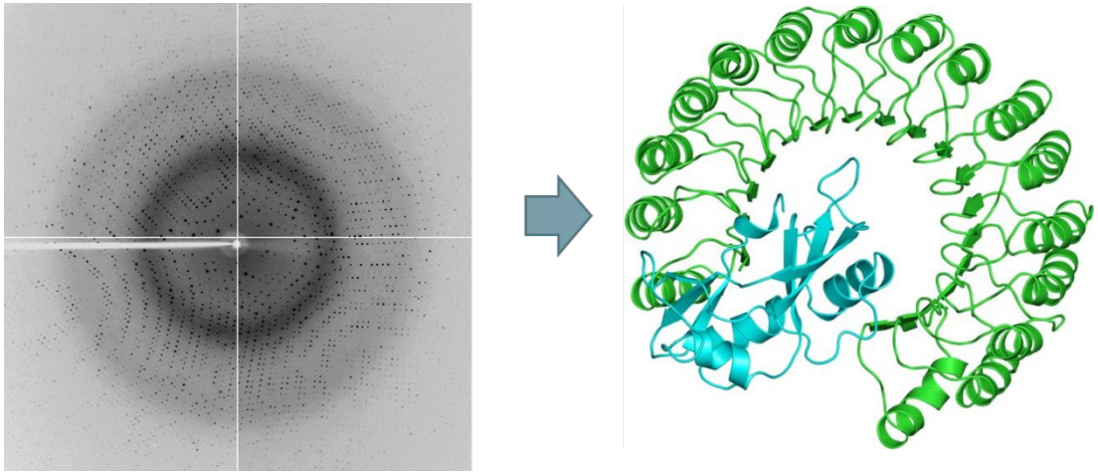


These chains fold into complex structures, allowing the protein to perform biological tasks

STRUCTURE AND FUNCTION ARE VERY CLOSELY RELATED

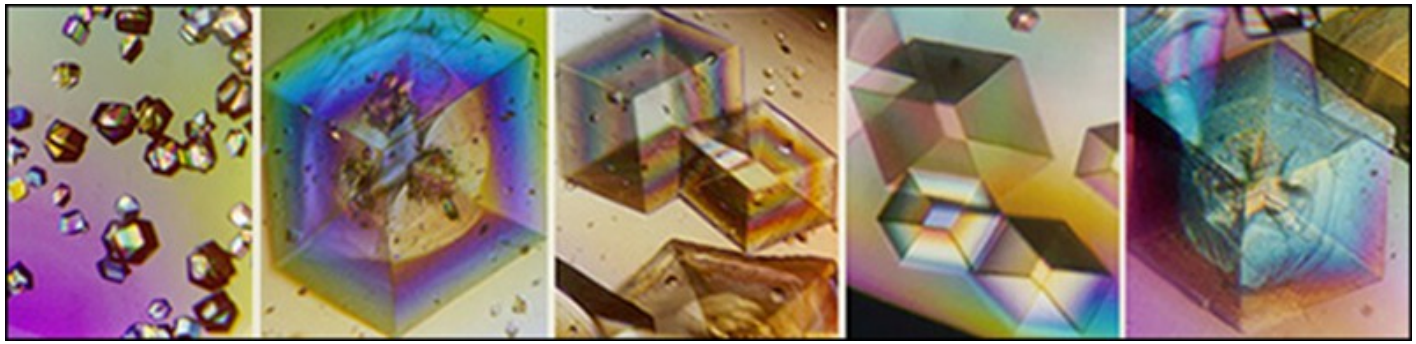
X-Ray Crystallography

- Protein crystals are bombarded with x-rays. The x-rays are diffracted, giving structural biologists a way to determine structure
- Herbert A. Hauptman developed the mathematical model to convert reflections to molecular structures
- The process requires *protein crystals*



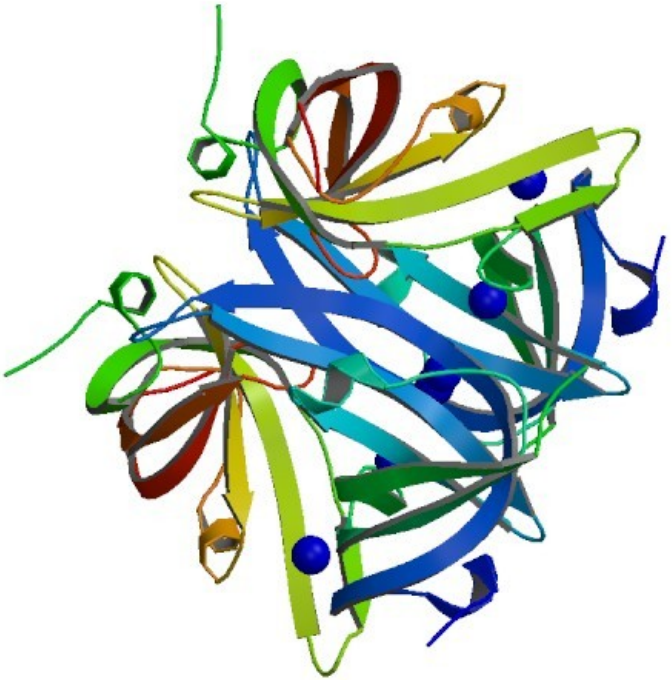
Crystallization

- Trying to get a protein to “crash out” of solution, in an orderly fashion
- Requires a precise set of conditions, which vary from protein to protein
- Included in this precise set of conditions: a chemical cocktail
- Cocktails are combined into screens (==generations)



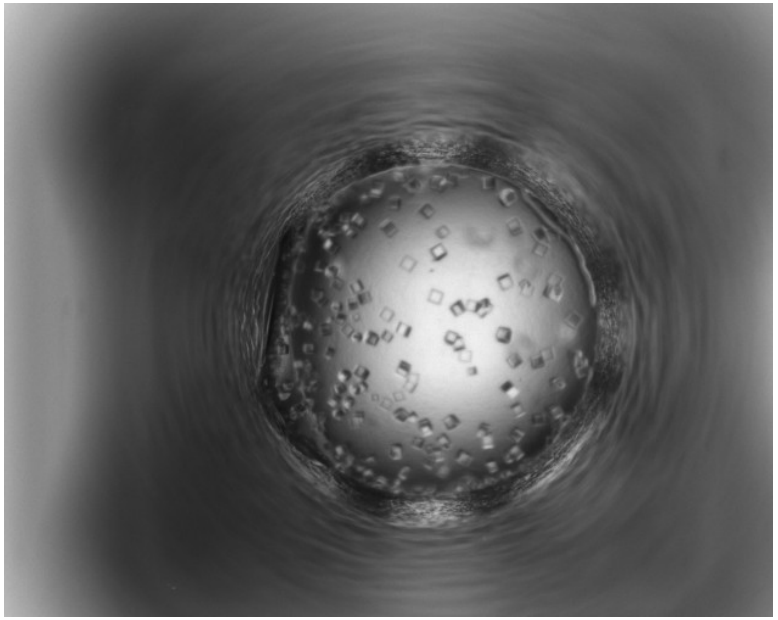
Introducing the Workshop Project

To achieve this (protein structure):

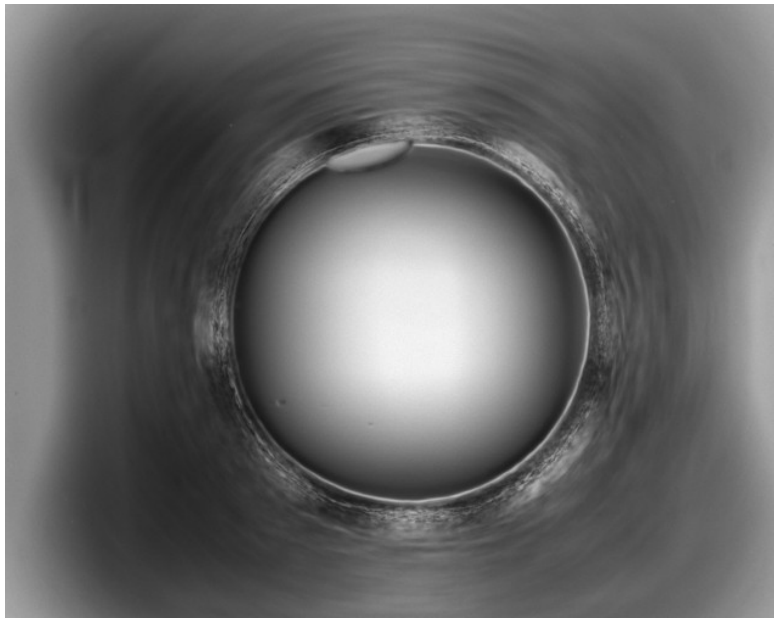


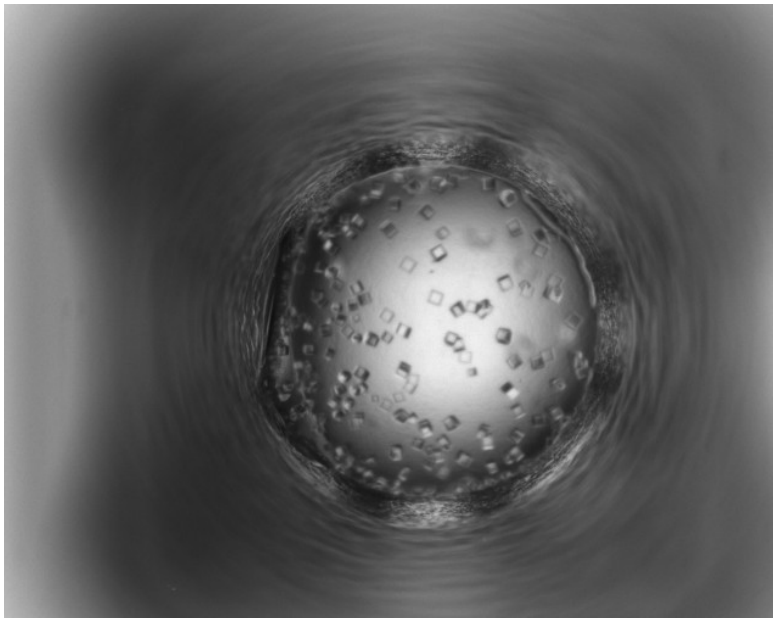
Protein Data Bank ID: 3dm3

We must get this (pure, crystallized protein):



HWI Protein Crystallization Data





Which sample contains a crystallized protein?

(Can we make this decision for 3 million samples?)

Real data, local data

HWI lab work

CCR processing

1. Laboratory work: HWI

- A protein is placed in 1536 wells on an experimental plate.
- A different chemical cocktail is added to each well.
- Photos of each well are taken at different timepoints.

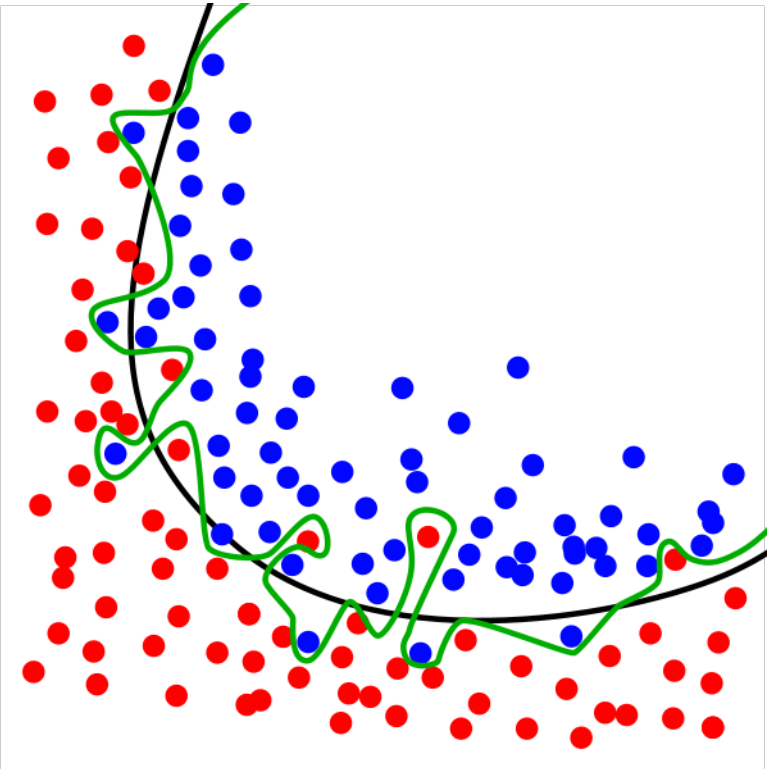
Which proteins crystallized in which cocktails, at which timepoints?

2. Automatic Classifier: CCR

An automatic classifier helps identify crystallized samples:

- Compute measures that describe the photo of each sample.
- Classifier assigns probability that each sample contains crystallized protein.

3. But...



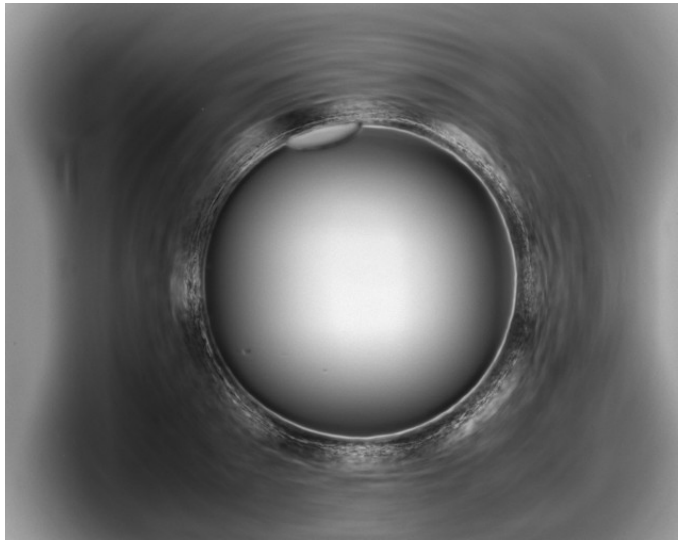
The classifier is only ~70% accurate, so:

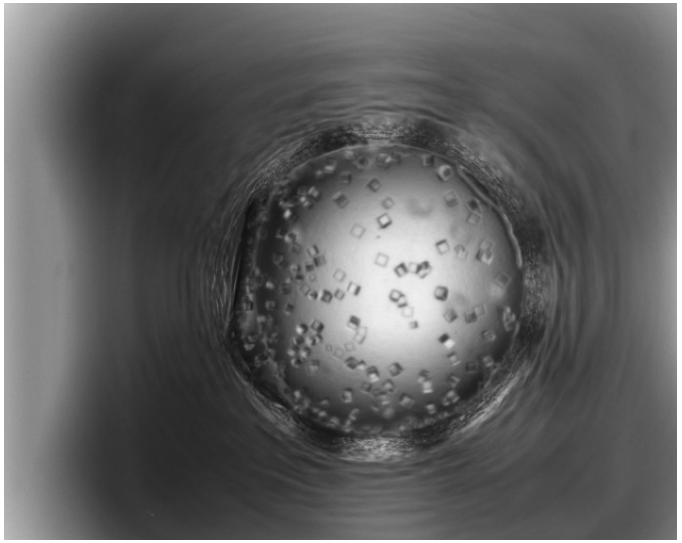
Human expert classifies each sample. This is the final word.

39936 experimental records

We have a photograph of each experiment:

- 13 proteins
- In 1536 chemical cocktails
- At 2 different time points

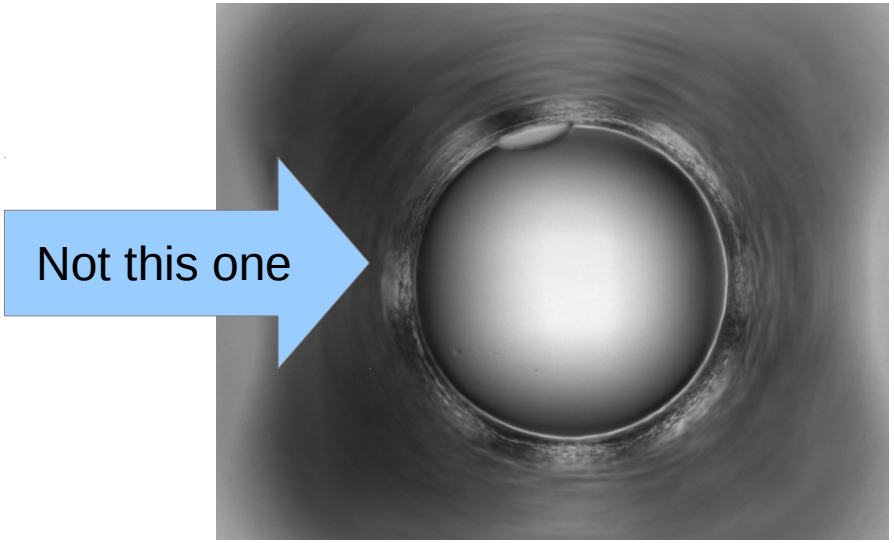


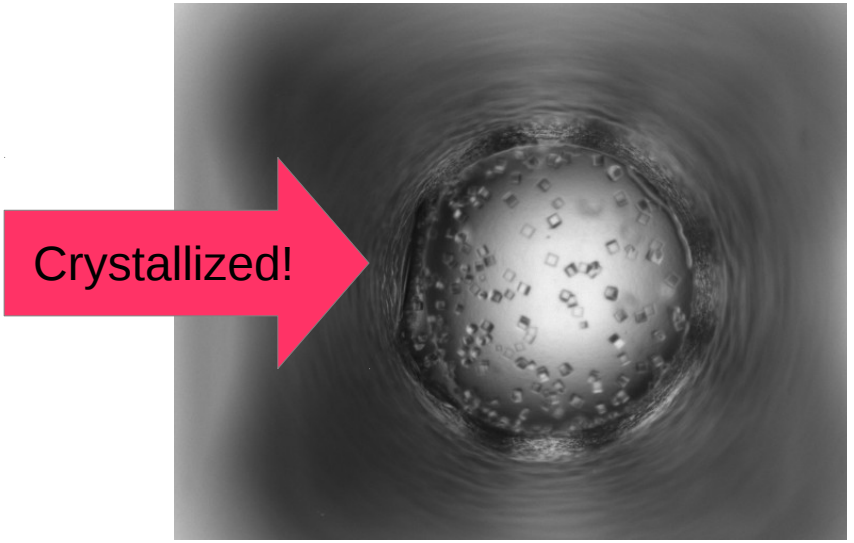


39936 experimental records

We have information about each one:

- Classifier score (value between 0 and 1)
- Human assignment (*truth*): crystal/non-crystal
- Some other stuff...

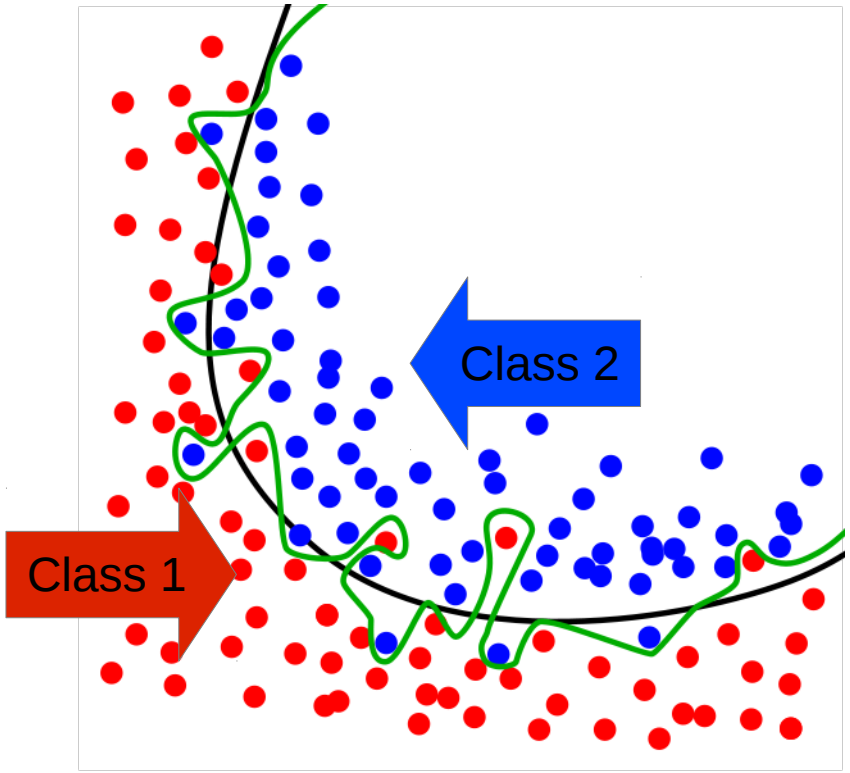




Automated Classifiers

This classifier has assigned data points into two classes, 1 and 2.

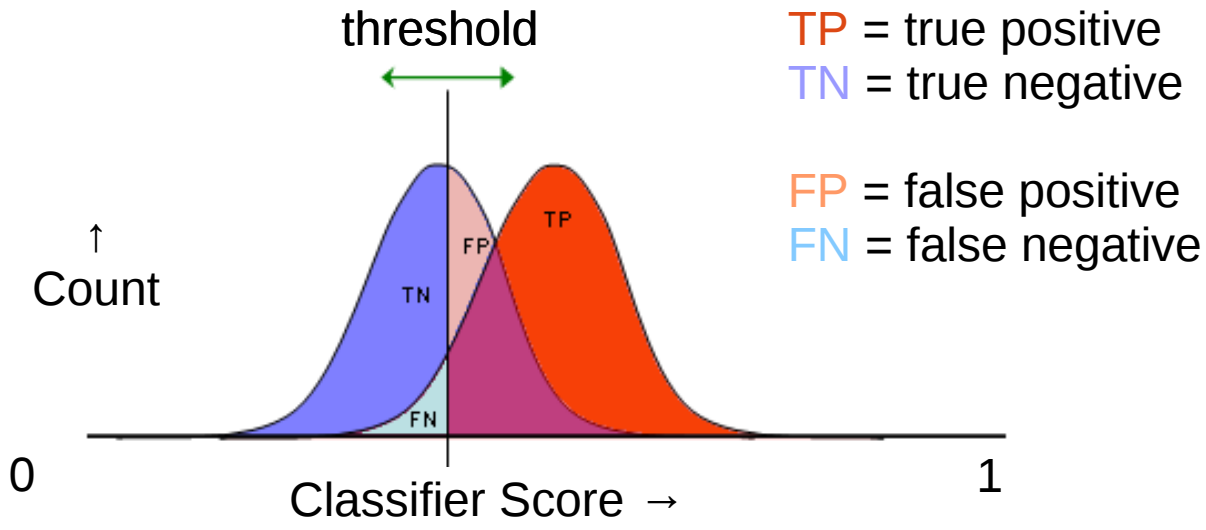
Automated Classifiers



How does this work?

- Classifier is run multiple times on each data point
- The results are converted to a probability that a crystal was seen

Classifier Performance



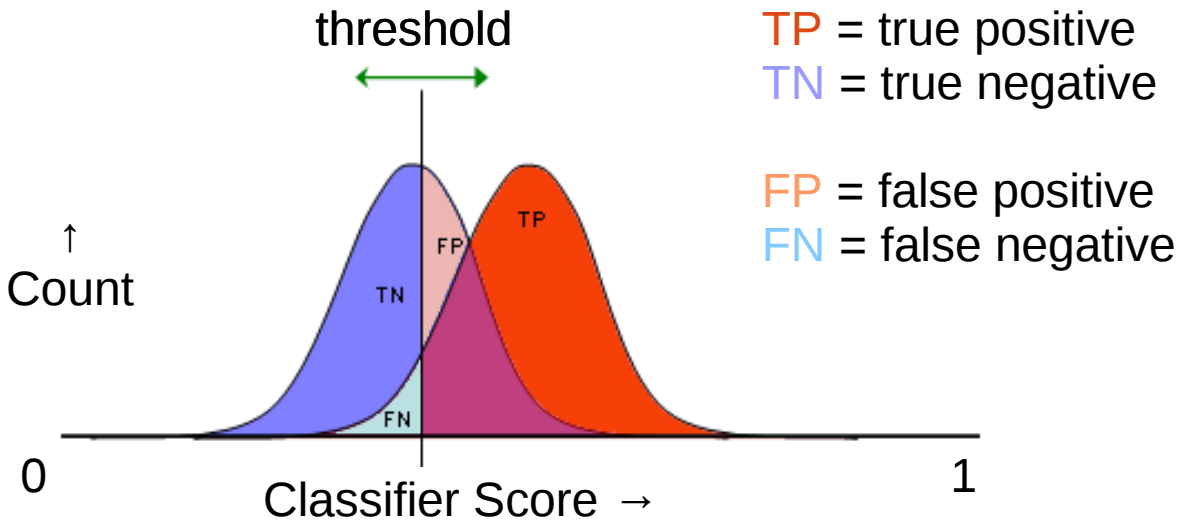
Our classifier assigns scores that lie between 0 and 1.

We can overplot two kernel density plots:

- One of human-classified crystal examples
- One of human-classified non-crystal examples

...both with classifier score on the x axis

Classifier Performance

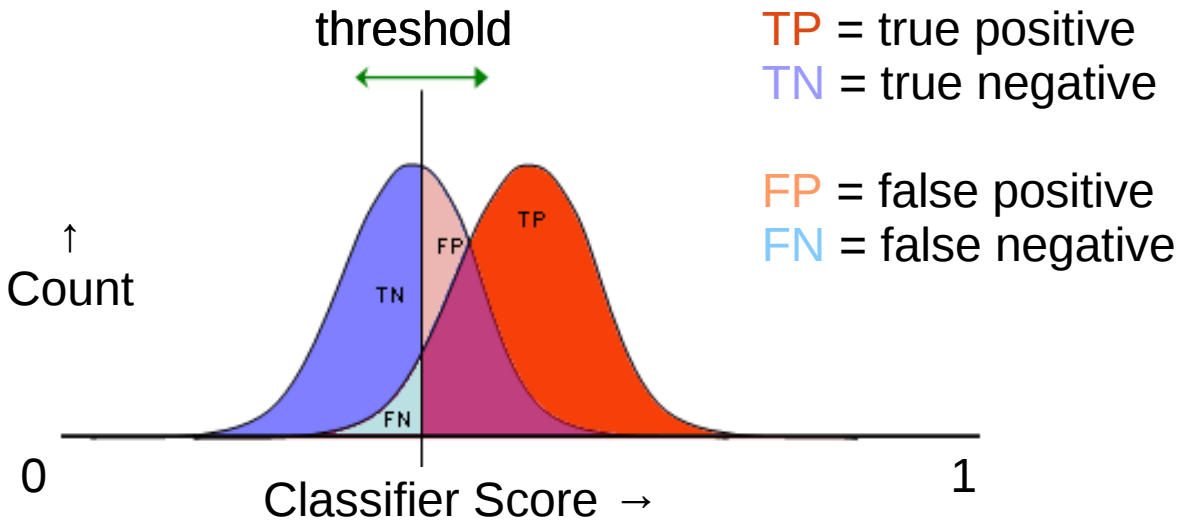


Pick a threshold (cut point) for classifier scores:

- above it, consider all examples to be positive (crystal)
- below it, consider all examples to be negative (not crystal).

No matter where we cut, we get some classifications wrong!

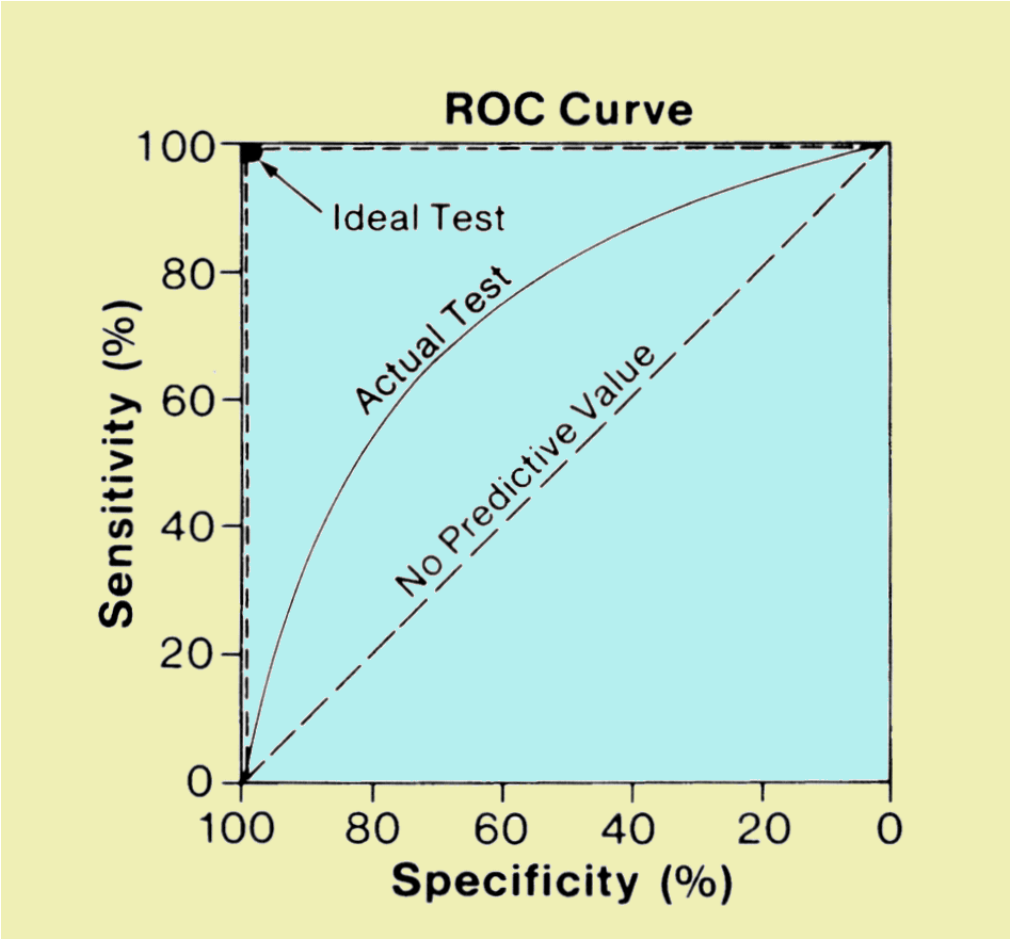
Classifier Performance



After we pick a threshold, we can see four types of outcomes:

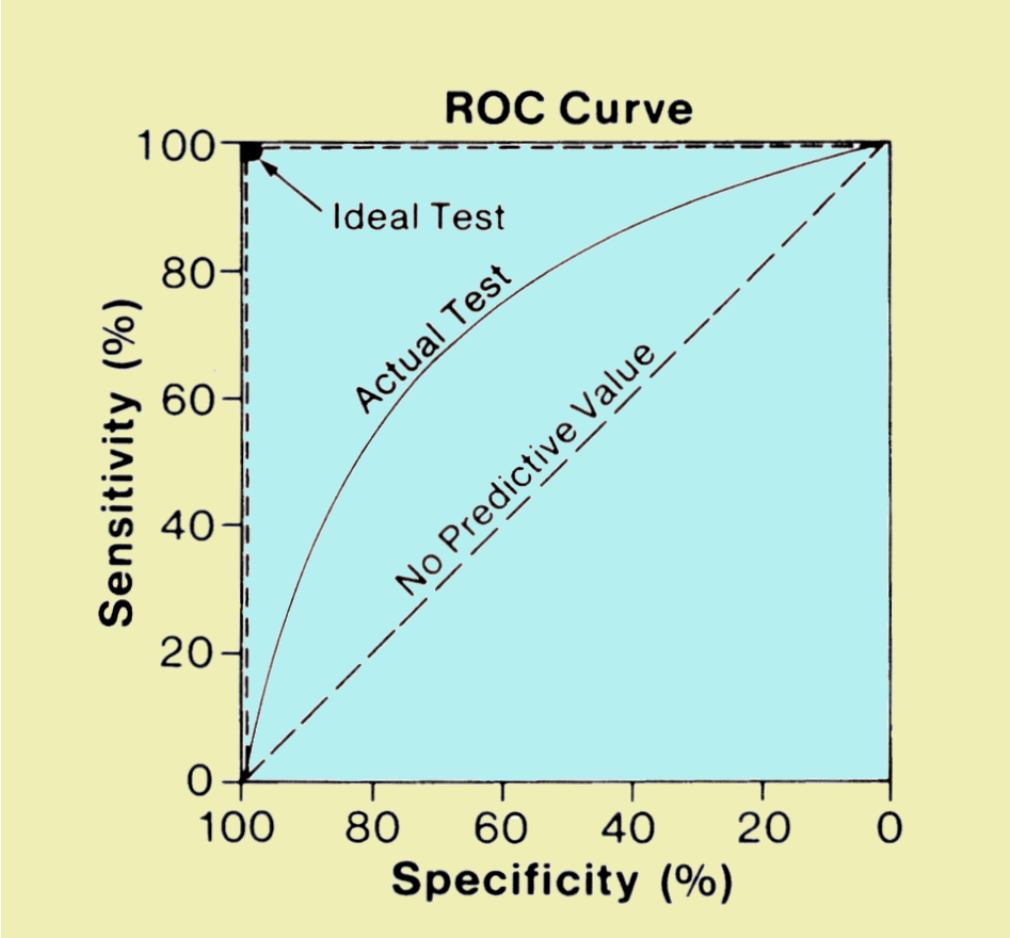
- True and False Positive (crystal)
- True and False Negative (not crystal).

ROC plots and classifiers



- Evaluate the performance of a binary classifier
- Vary the threshold or cut point, vary the sensitivity/specificity ratio

Sensitivity vs. Specificity



- Specificity (true negative rate): the ability to classify non-crystals as non-crystals.
- Sensitivity (true positive rate): the ability to classify crystals as crystals

Meet your HWI dataset

The data frames:

- experiment: 39936 rows
- sample: 13 proteins
- drop: 3930 rows (1536 chemical cocktails)
- expUrl: 39936 rows (one per experiment)

$$13 \times 1536 \times 2 = 39936$$

proteins cocktails timepoints experiments

Experiment data: 39936 rows

experiment data frame describes the crystallization experiments:

- read_no: 39936 values, identifies experiment
- sample_id, plate_no: 13 values, identifies protein
- week_no: 2 timepoints
- cocktail_no: 1536 unique values
- human_crystal: crystal or not?
- class3_crystal: probability of crystal

$$\underset{\text{proteins}}{13} \times \underset{\text{cocktails}}{1536} \times \underset{\text{timepoints}}{2} = \underset{\text{experiments}}{39936}$$

Sample data: 13 different proteins

sample data frame describes the proteins:

- Identify each protein:
  - sample_id, p_number, targetdb_status, name
- experimental_molecular_weight
- concentration of the protein
- seq: sequence of amino acids in protein
- seq_len: number of amino acids
- targetdb_ref: PDB (Protein Data Bank, rscb.org)
- targetdb_status: unique protein code on PDB

Drop data: 3930 rows

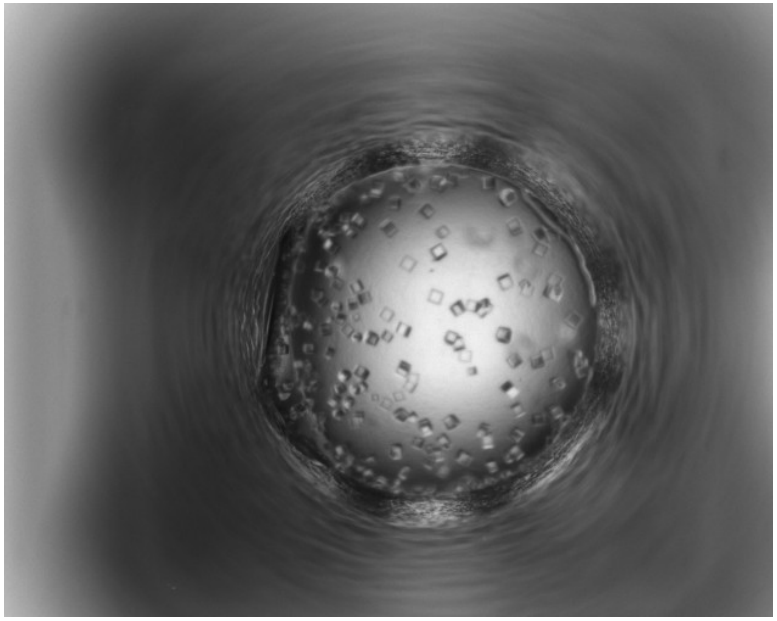
drop data frame describes the components of the chemical cocktails:

- Identify each chemical cocktail (1536 in total):
  - cocktail_no + solution_component_no
- concentration of the solution component
- name of the solution component
- ph: pH of the solution component

expUrl data: 39936 rows

expUrl data frame contents:

- read_no: identifies experiment
- image_url: URL of the experiment image



RStudio: Project Data

HWI data are text files with csv format.

- Formatted to load straight into an R data frame
- Columns are labeled by name

csv == comma separated values

Rstudio and csv: Two ways to load data

1. RStudio Workspace:

- Select Import Dataset: From Text File
- Select a .csv file to Open
- Use Heading=Yes

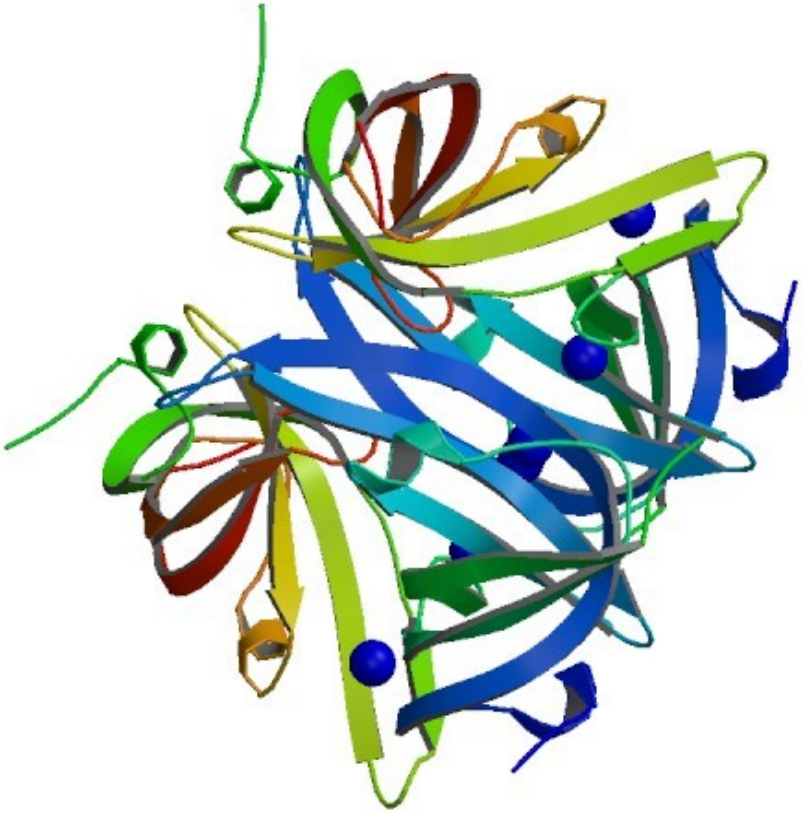
2. Or, from the command line:

- Set the Working Directory
- Load command:

```
> drop=read.csv("drop.csv")
```

Exploring Proteins

Search rscb.org



Protein Data Bank ID: 3dm3

Exploring Proteins: rscb.org

RCSB Protein Data Bank - RCSB PDB - 3DM3 Structure Summary - Mozilla Firefox

www.rcsb.org/pdb/explore/explore.do?structureId=3dm3

RCSB PDB PROTEIN DATA BANK → PDB-101

A MEMBER OF THE PDB | EMDataBank
An Information Portal to Biological Macromolecular Structures
As of Tuesday Jun 18, 2013 at 5 PM PDT there are 91550 Structures | PDB Statistics

Search | Advanced | Browse

Everything | Author | Macromolecule | Sequence | Ligand

e.g., PDB ID, molecule name, author

Search History , Previous Results

PDB-101 Hide
- Structural View of Biology
- Understanding PDB Data
- Molecule of the Month
- Educational Resources
- Author Profiles

MyPDB Hide
- Login to your Account
- Register a New Account
- MyPDB Help Page

Home Hide
- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Deposition Hide
- All Deposit Services
- Electron Microscopy
- X-ray | NMR
- Validation Server

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

Crystal structure of a domain of a Replication factor A protein, from Methanocaldococcus jannaschii. NorthEast Structural Genomics target MjR118E

DOI:10.2210/pdb3dm3/pdb

3DM3

PSI Protein Structure Initiative

Display Files
Download Files
Share this Page

Primary Citation

Crystal structure of a domain of a Replication factor A protein, from Methanocaldococcus jannaschii. NorthEast Structural Genomics target MjR118E

Seetharaman, J., Su, M., Maglaqui, M., Janjua, H., Ciccosanti, C., Xiao, R., Nair, R., Everett, J.K., Acton, T.B., Rost, B., Montelione, G.T., Tong, L., Hunt, J.F.

Journal: To be Published

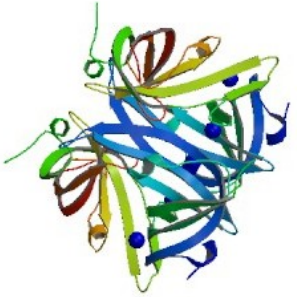
PubMed ID is not available

Molecular Description Hide

Classification: Replication
Structure Weight: 35296.17

Molecule: Replication factor A
Polymer: 1 **Type:** protein **Length:** 105
Chains: A, B, C
Fragment: residues 170-274

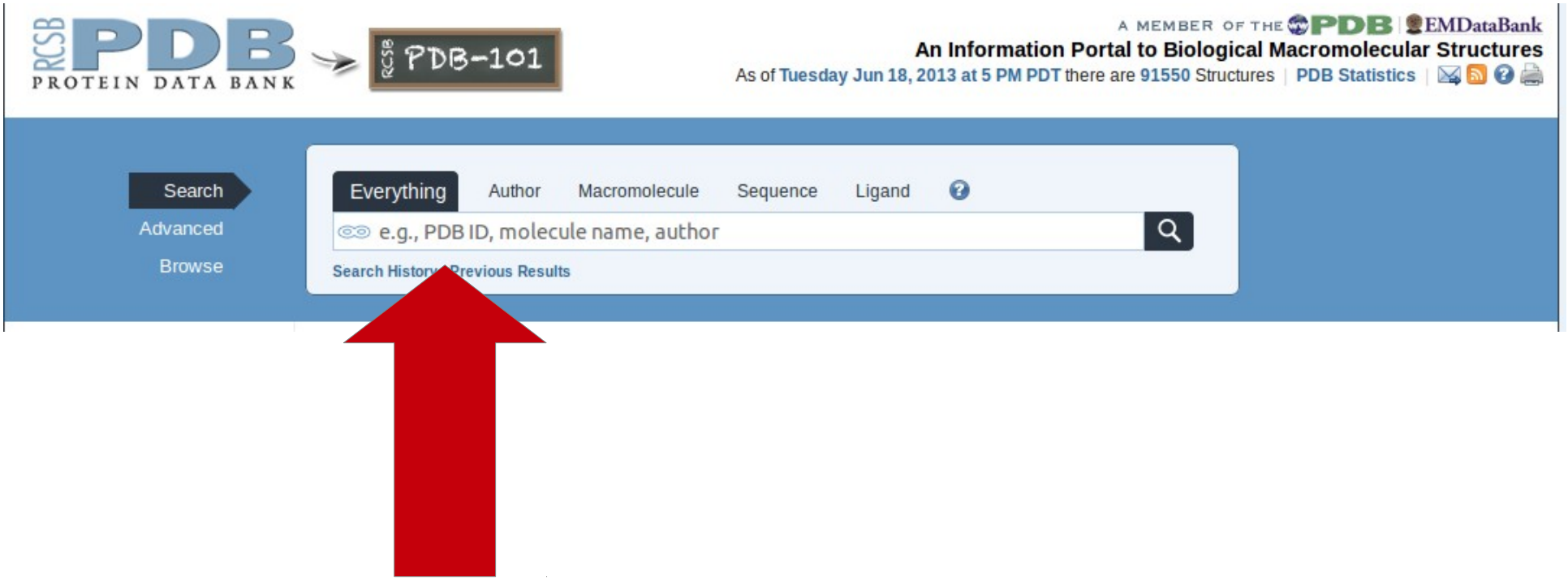
Biological Assembly 1



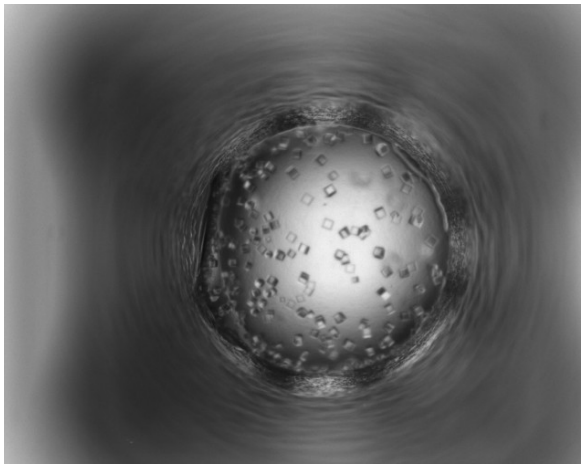
View in 3D | More Images...

Find: 66 | Previous | Next | Highlight all | Match case

Exploring Proteins: rscb.org



Exploring Proteins



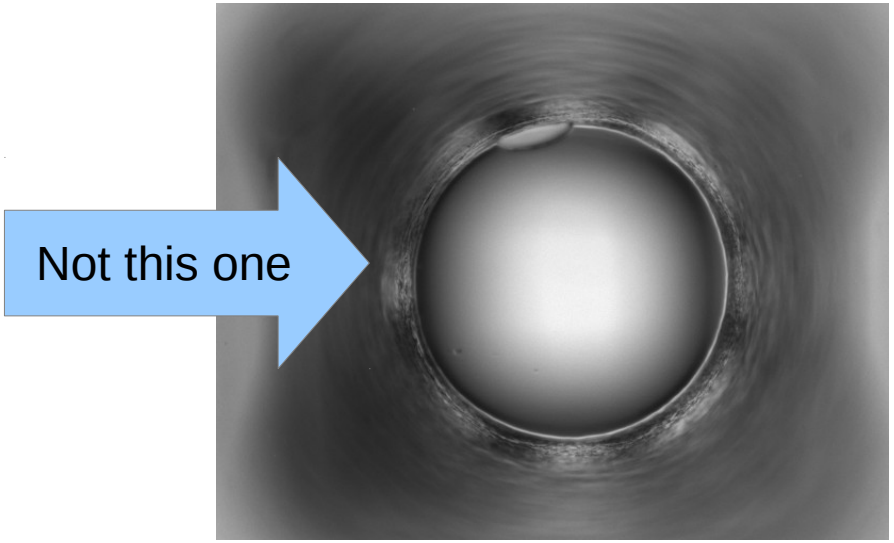
Check the data frame expUrl for links to the experimental images.

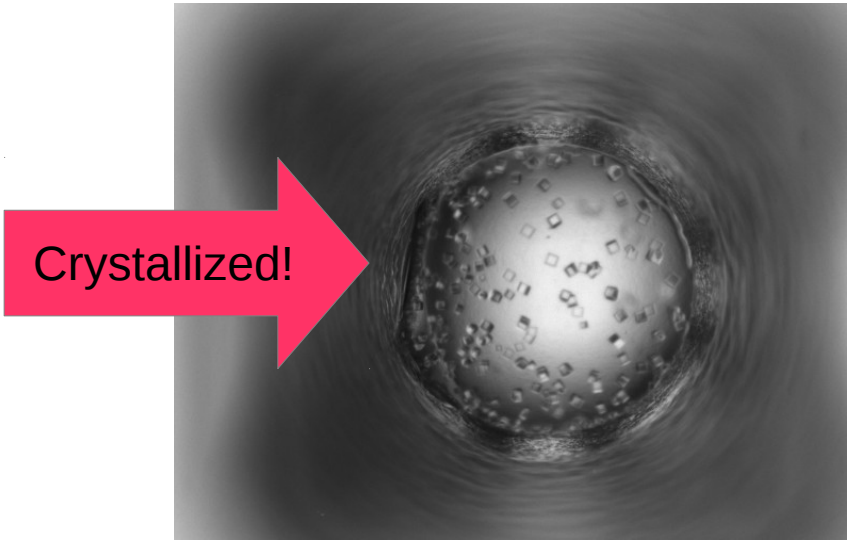
Can you identify the crystallization state?

	read_no	image_url
1	X000009589069620080118122 6	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589069620080118122 6.jpg
2	X000009589038420080118123 1	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589038420080118123 1.jpg
3	X000009589000120080118123 5	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589000120080118123 5.jpg
4	X000009589048820080118122 8	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589048820080118122 8.jpg
5	X000009589050420080118122 8	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589050420080118122 8.jpg
6	X000009589092620080118122 2	http://xtuition.ccr.buffalo.edu/image-data/X000009589/X000009589200801181215/X000009589092620080118122 2.jpg

The questions

1. How does the automated *classifier* perform?
2. What trends do we see in the *timeseries* data?
3. What trends do we see in crystallization of the proteins across the different *cocktails*?
4. What can we learn about the *proteins*?





Project: you'll need...

- Data load
  - read.csv() R function
- File transfer: WebDAV
  - Copy files to your workstation from hpc2.org
  - Mostly useful for creating your presentation slides
- factorify() function
  - Apply to a data frame after subsetting; updates factor levels
  - source generalFactorifyFunction.R

1. Classifier Performance Focus

- Data file: experiment.csv
- Epi R package

  library(“Epi”)
- R functions:
  - sapply()
  - Kernel density estimation: density()
  - data.frame()
- R graphics:
  - hist(), plot()
- Provided function compareDensityPlots()

2. Timeseries Crystallization Focus

- Data files:
  - experiment.csv
  - sample.csv
- R functions:
  - merge(), length(), table(), round()
- R graphics:
  - boxplot(), hist(), pie(), barplot(), legend()

3. Cocktail by Protein Focus

- Data files:
  - Experiment.csv
  - Sample.csv
  - Drop.csv
- R functions:
  - table(), unique(), sum(), length(), which(), dim(), merge()
- R graphics: boxplot(), pie(), barplot()

4. Protein Exploration Focus

- Data files:
  - expUrl.csv, experiment.csv, sample.csv
- R functions:
  - plot() – for scatterplots
  - lm() – linear model to fit data to a line
  - abline() – plot a linear model
  - cor() – examine correlations
- PDB (protein data bank): rscb.org
- R Package: “bio3d”:
  - library(“bio3d”)