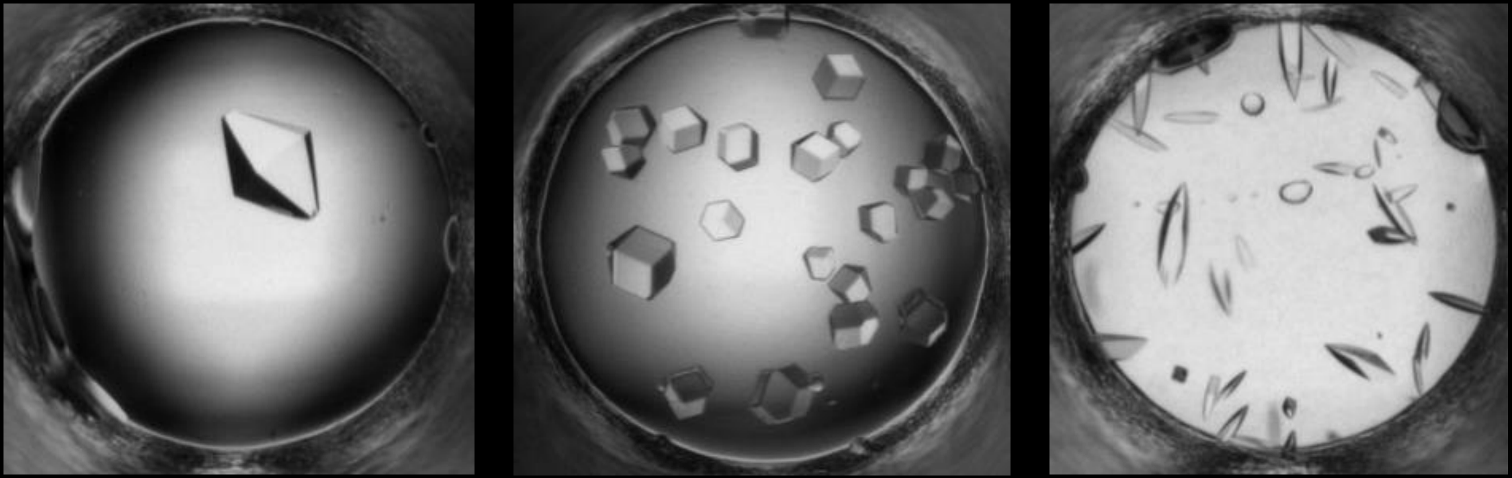
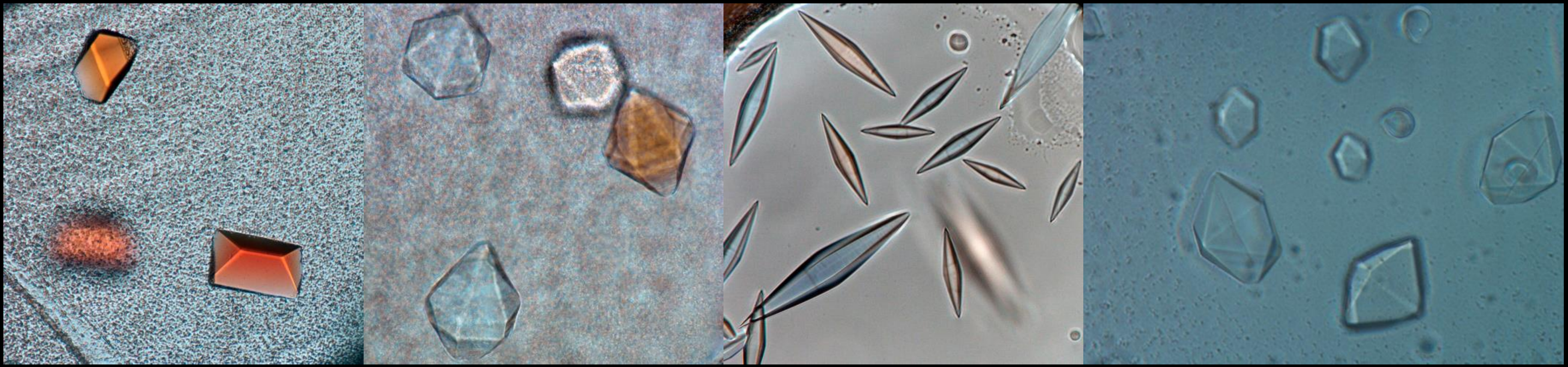
Efficient High-Throughput Crystallization

Edward H. Snell and Joseph R. Luft

Hauptman-Woodward Medical Research Institute



Crystallography Requires Crystals

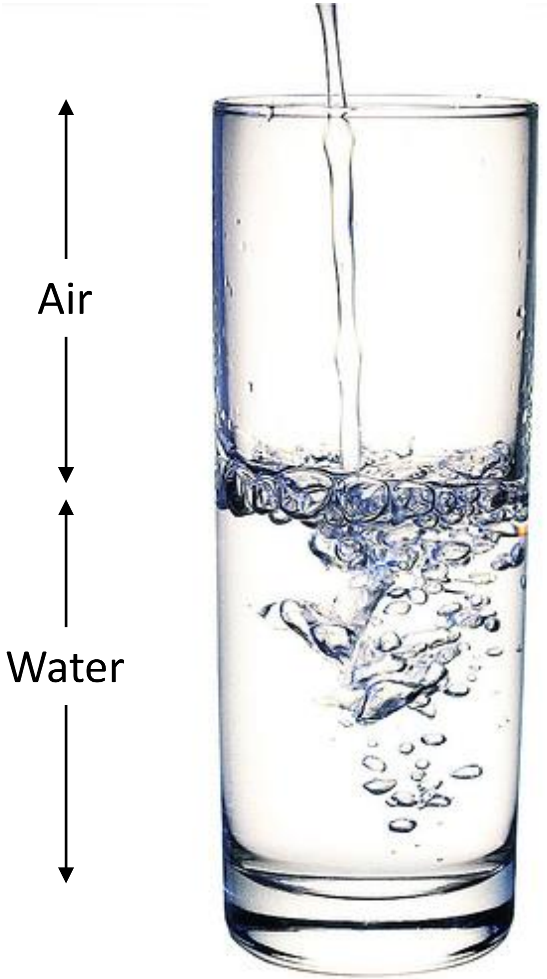


No crystal …

No crystallography ….

No crystallographer ….

Pessimists, Optimists, and Crystallographers



Air

Water

Consider a glass of water

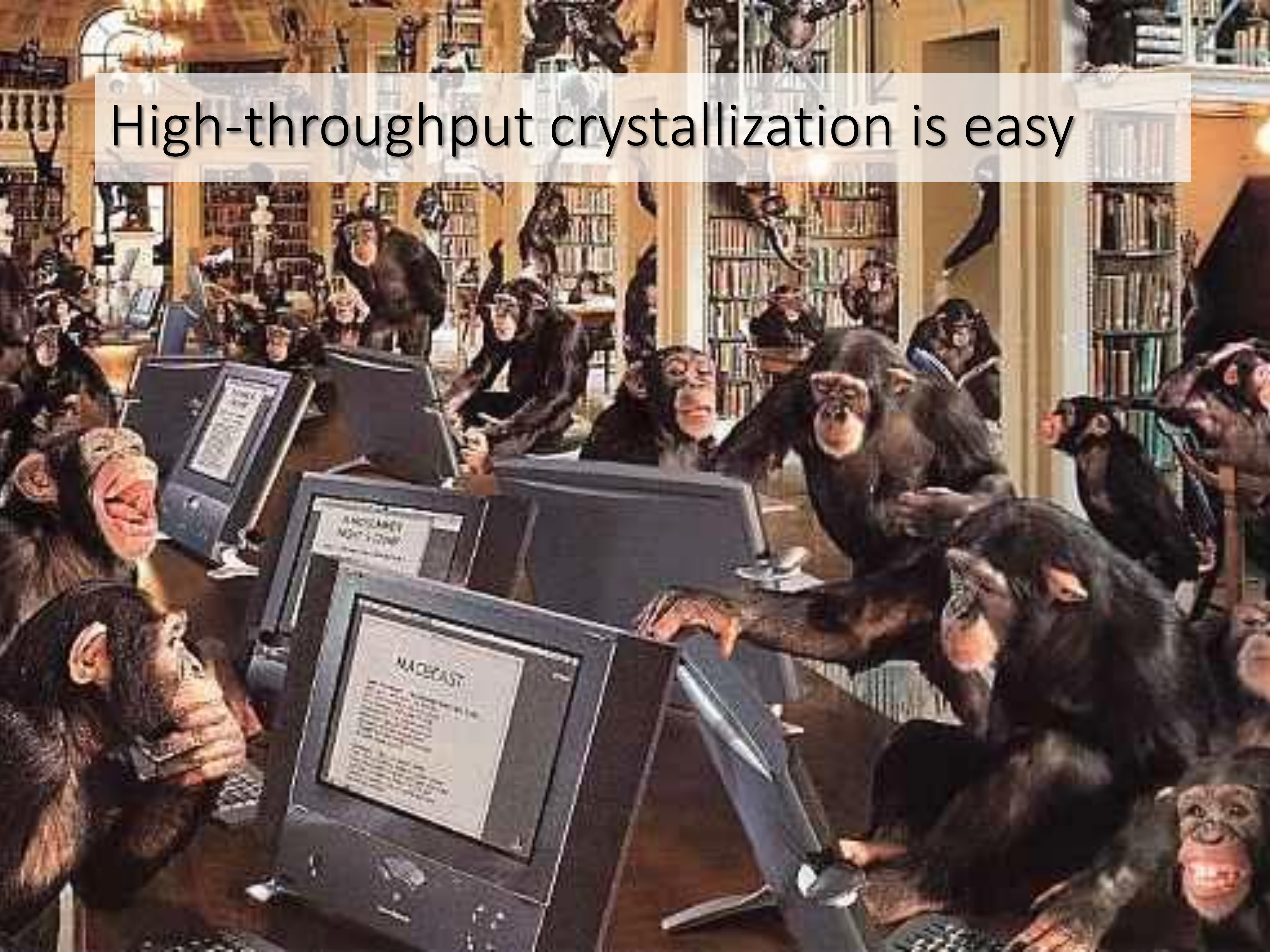
Pessimist
(the glass is half empty)

Crystallographer
(the glass is completely full)

Optimist
(the glass is half full)

Fantasy

High-throughput crystallization is easy



Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized

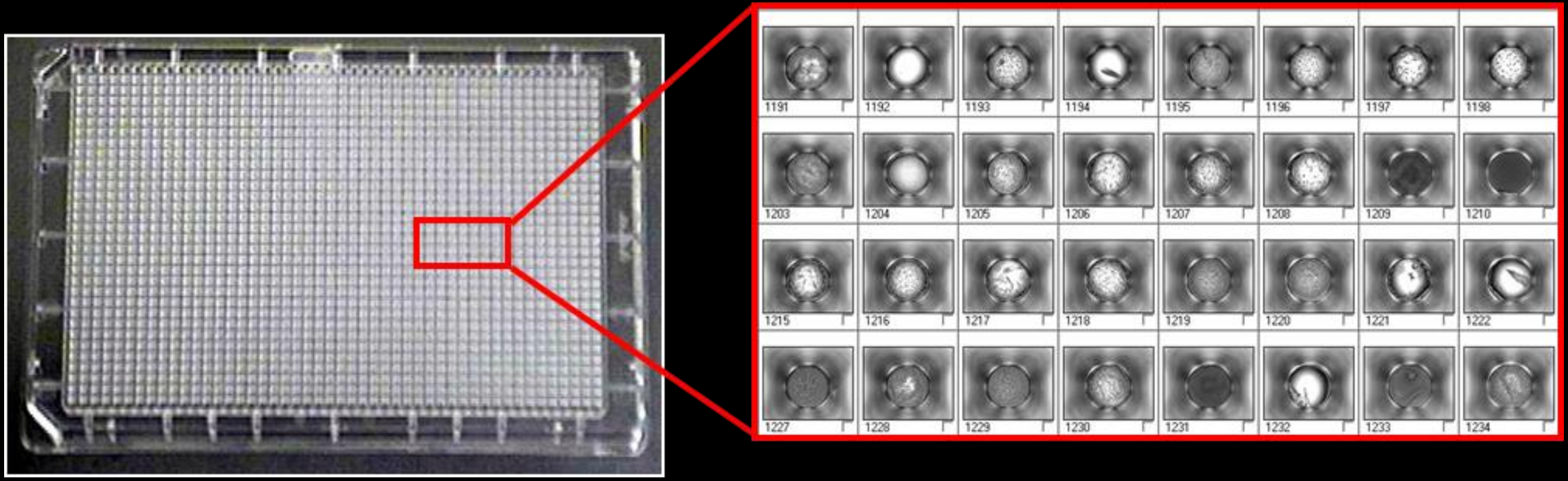
The Crystallization Screening laboratory at the Hauptman-Woodward Medical Research Institute

Since February of 2000 the High Throughput Search (HTS) laboratory has been screening potential crystallization conditions as a high-throughput service

The HTS lab screens samples against three types of cocktails:

1. Buffered salt solutions varying pH, anion and cation and salt concentrations
2. Buffered PEG and salt, varying pH, PEG molecular weight and concentration and anion and cation type
3. Almost the entire Hampton Research Screening catalog.

The HTSlab has investigated the crystallization properties of over 15,000 individual proteins archiving approximately 140 million images of crystallization experiments.

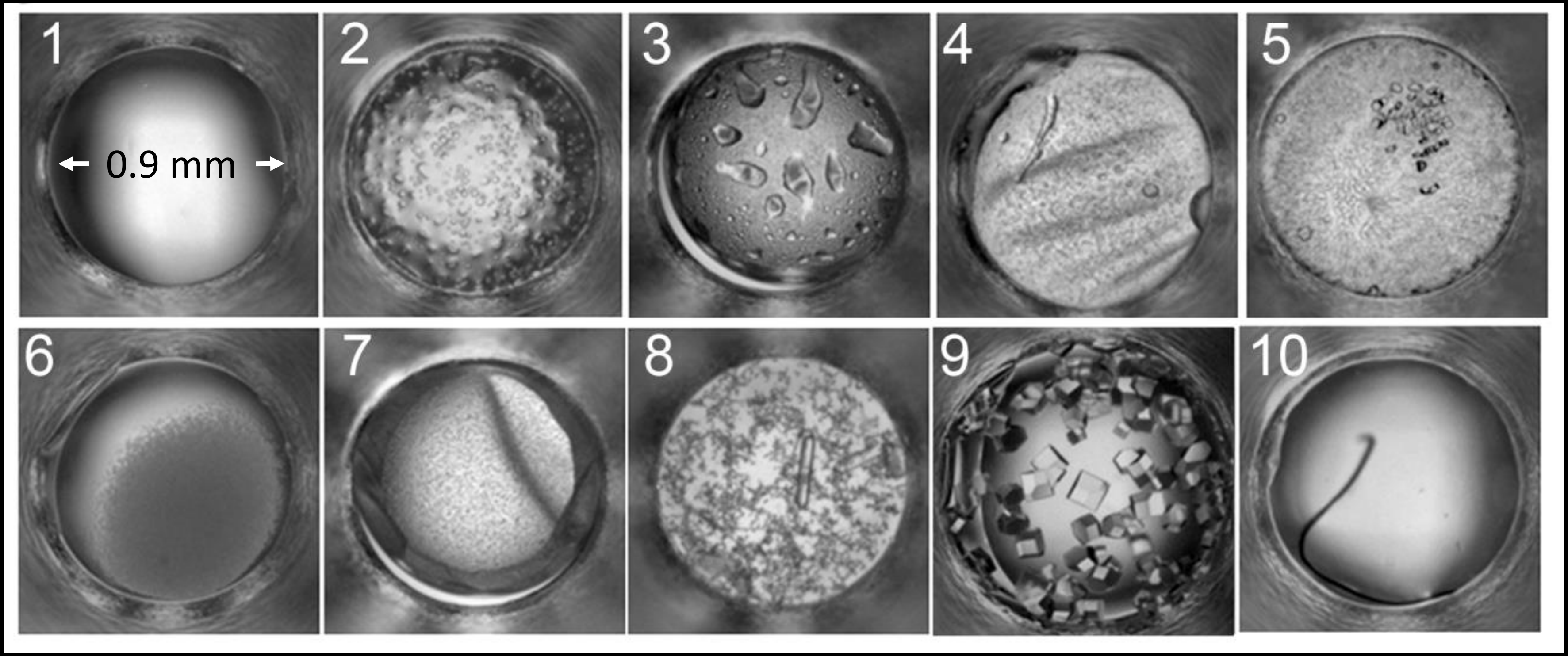


The crystallization method used is micro-batch under oil with 200 nl of protein solution being added to 200 nl of precipitant cocktail in each well of a 1536 well plate.

Wells are imaged before filling, immediately after filling then weekly for six weeks duration with images available immediately on a secure ftp server.

Several software utilities for viewing and analyzing data are available.

Outcomes



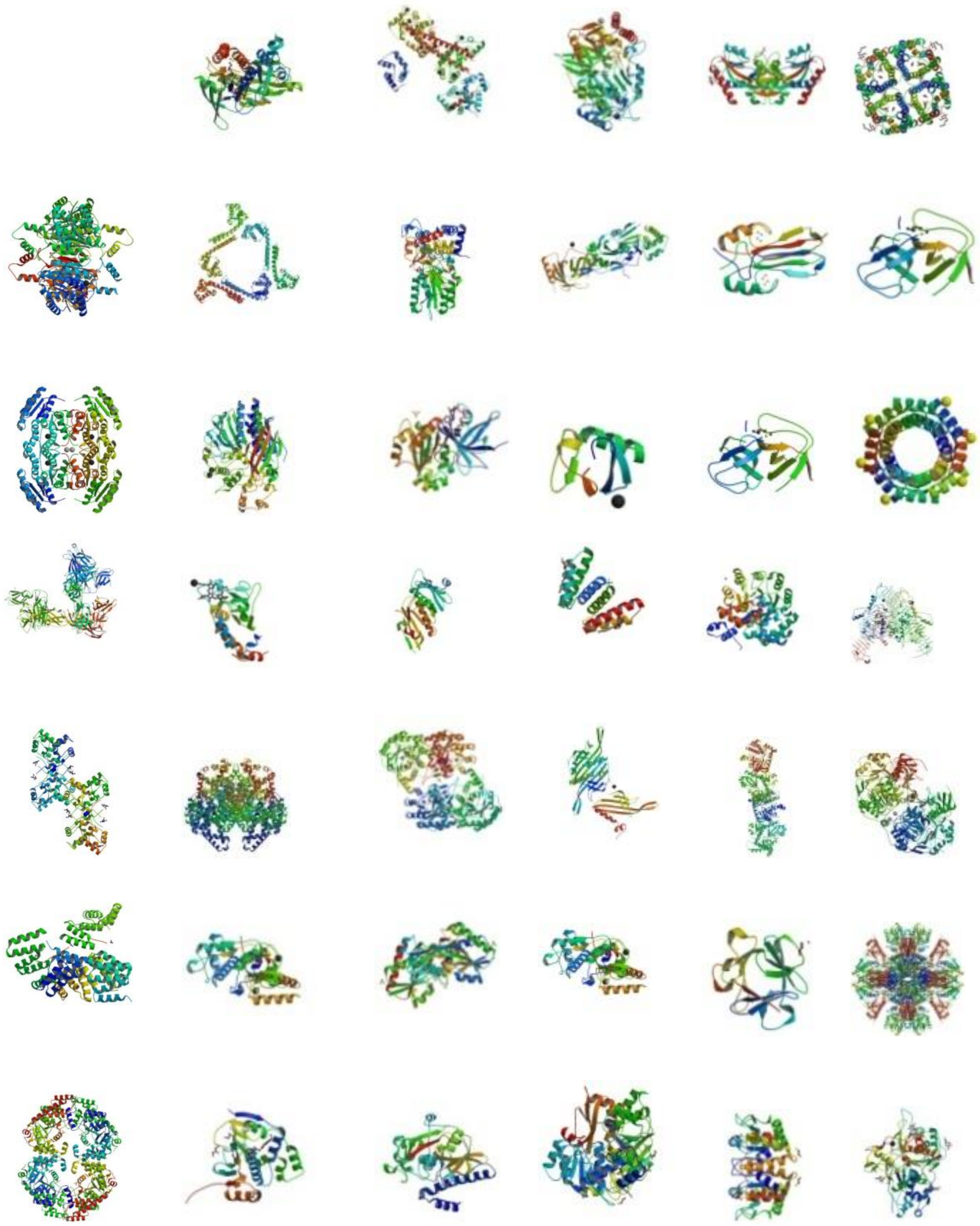
Born in Buffalo

Over 1,000 general biomedical laboratories world wide use the crystallization screening service with approximately 2,000 unique investigators.

Investigators are sent photographs of the results, analyze these images and perform their own optimization of any hits observed.

No information is released on targets. Progress is tracked by acknowledgements and citation searches. Currently no other metrics are used to measure success rates for the general biomedical community.

These images represent examples of structures from initial hits in the HTS laboratory.



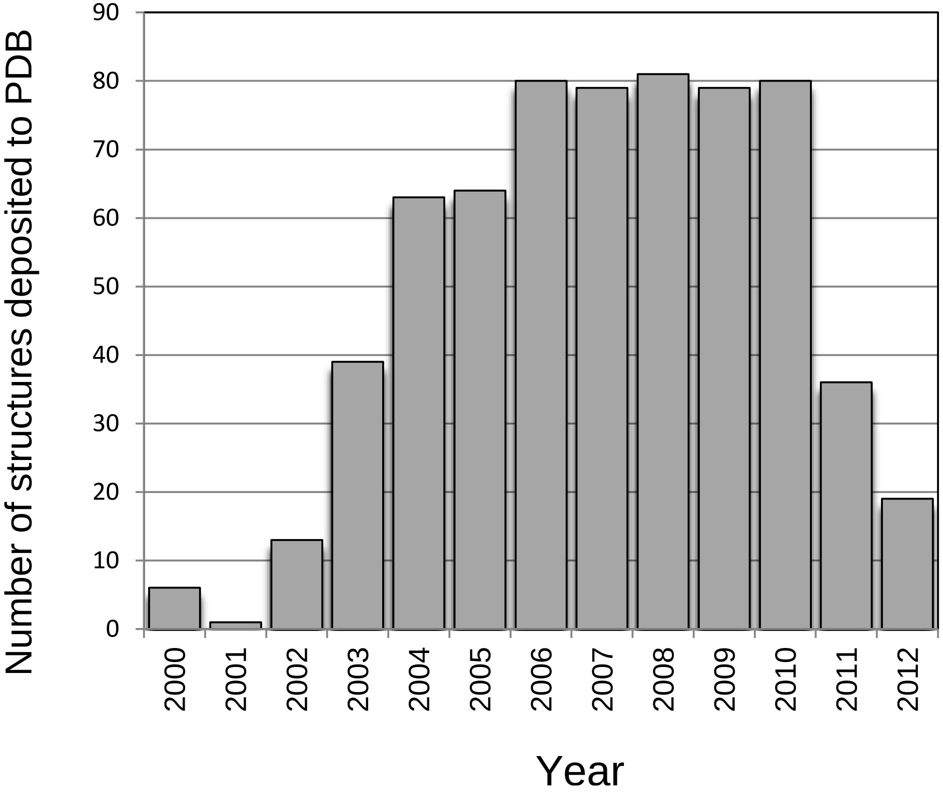
Where success is tracked.

For our Protein Structure Initiative partners both success and failure is tracked. In the case of NESG our initial screening hits enable on average 80 structures per year to be deposited to the PDB.

The graph demonstrates the ramp up of operations with maximum success reached from 2006 onward.

Our success rate from protein in the door to a crystallization hit leading to a PDB deposition is **22%**.

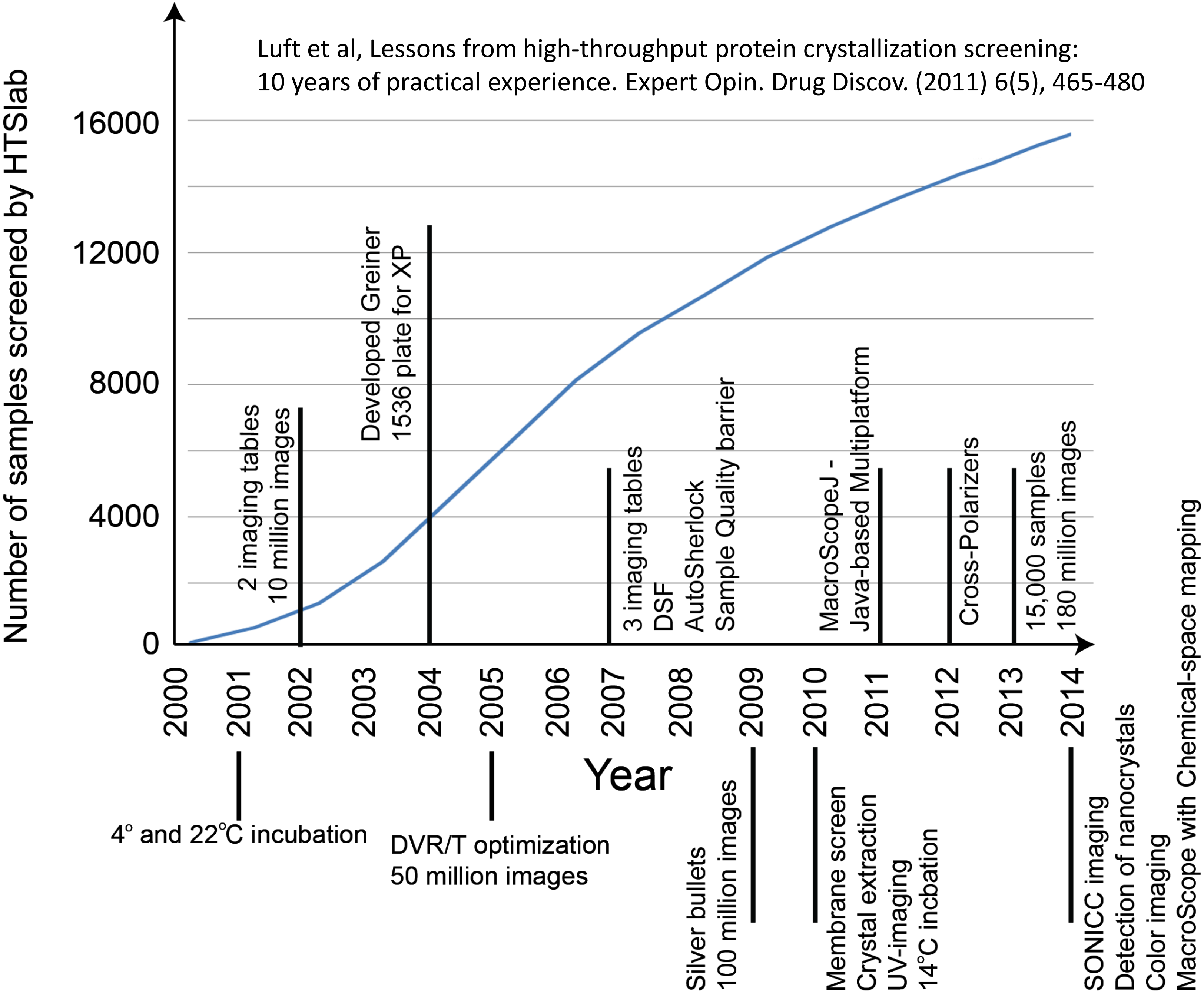
The NESG samples represent a special case in that they are well characterized beforehand – size exclusion chromatography, mass spec analysis and dynamic light scattering studies.



In 2011 we switched to PSI Biology – More difficult targets

Old data

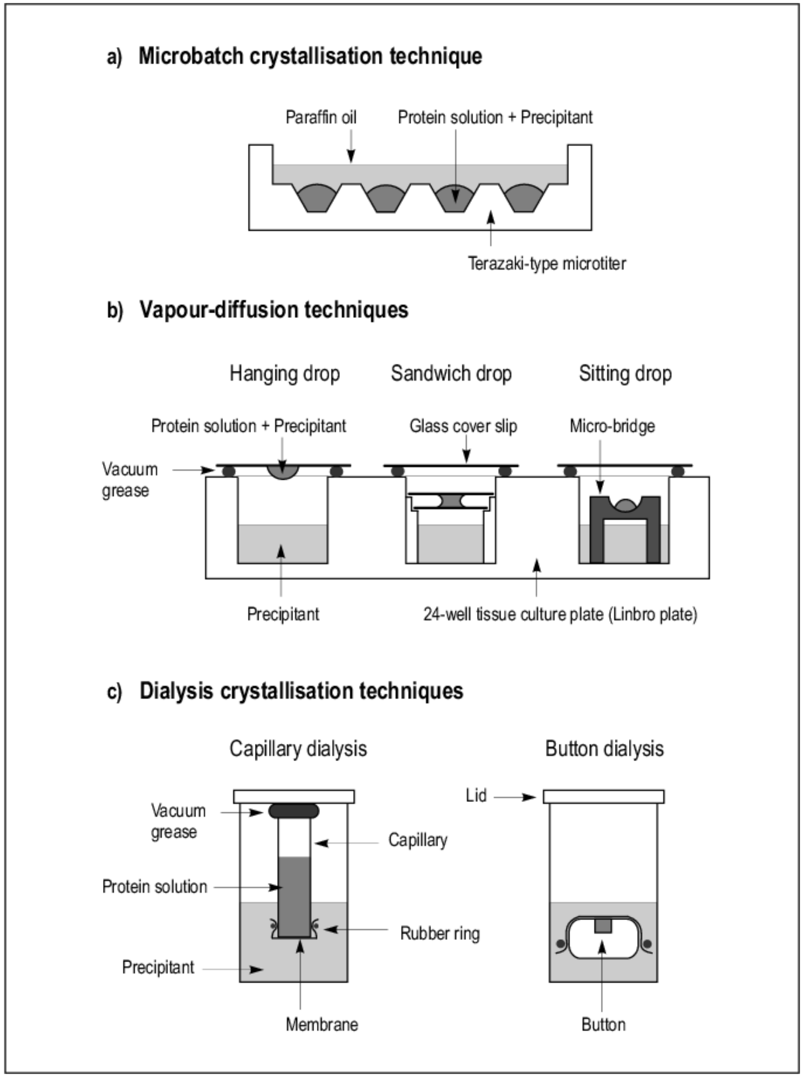
Luft et al, Lessons from high-throughput protein crystallization screening: 10 years of practical experience. Expert Opin. Drug Discov. (2011) 6(5), 465-480



Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized

Crystallizing Macromolecules



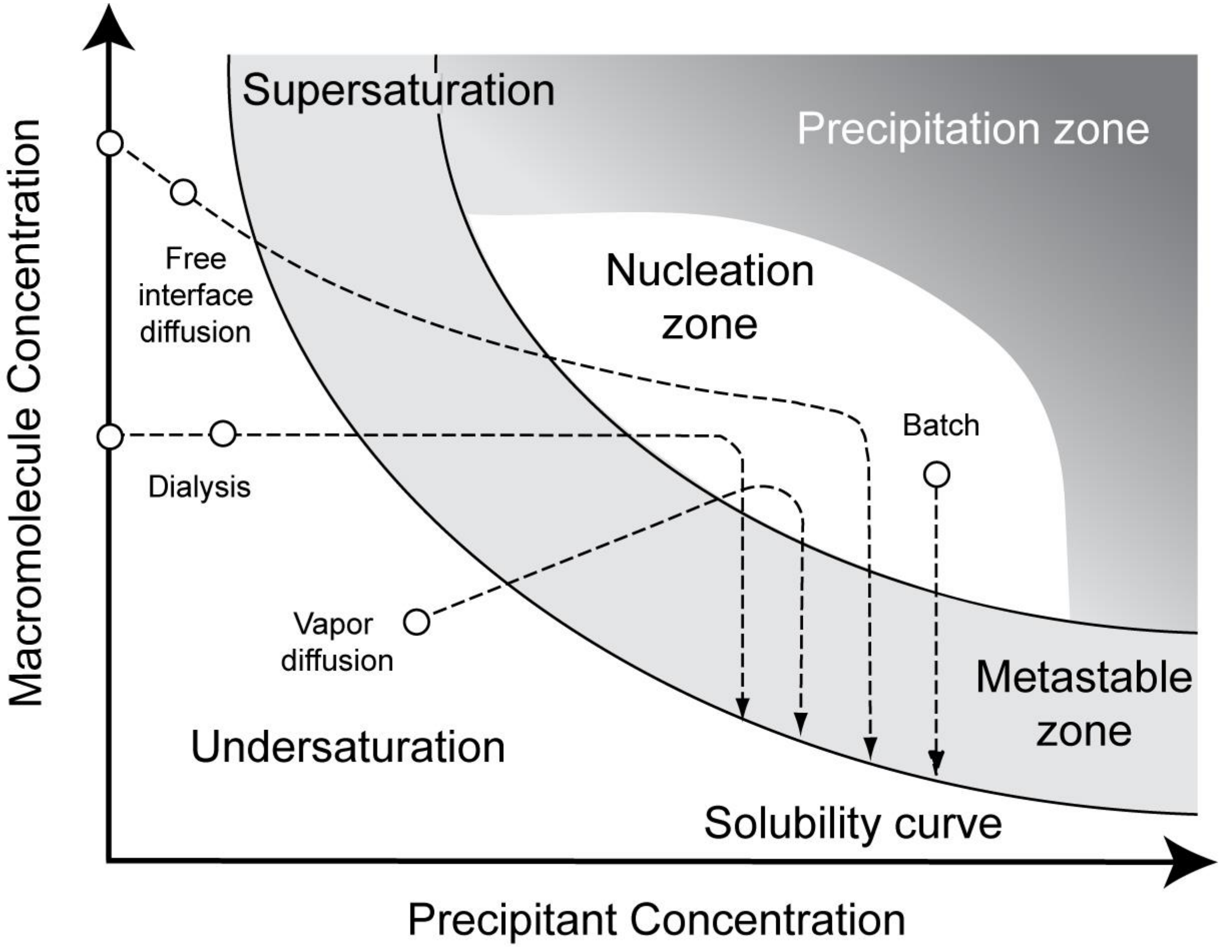
Many different methods but they all have things in common:

- They are designed to traverse the crystallization phase diagram.
- They use many different kinds of solutions to sample crystallization space at many points.

Which method?

- Vapor diffusion (most common)
  - Dynamic – samples wide physical chemical space
  - Can use small volumes
  - Reproducible
  - Multiple experiments in one drop
- Microbatch under oil (used by our laboratory)
  - Static – initial conditions highly defined
  - Sealed in one setup
  - Transportable
- Dialysis (less common)
  - Larger volumes
  - Difficult automated setup

Simplified phase diagram for crystallization



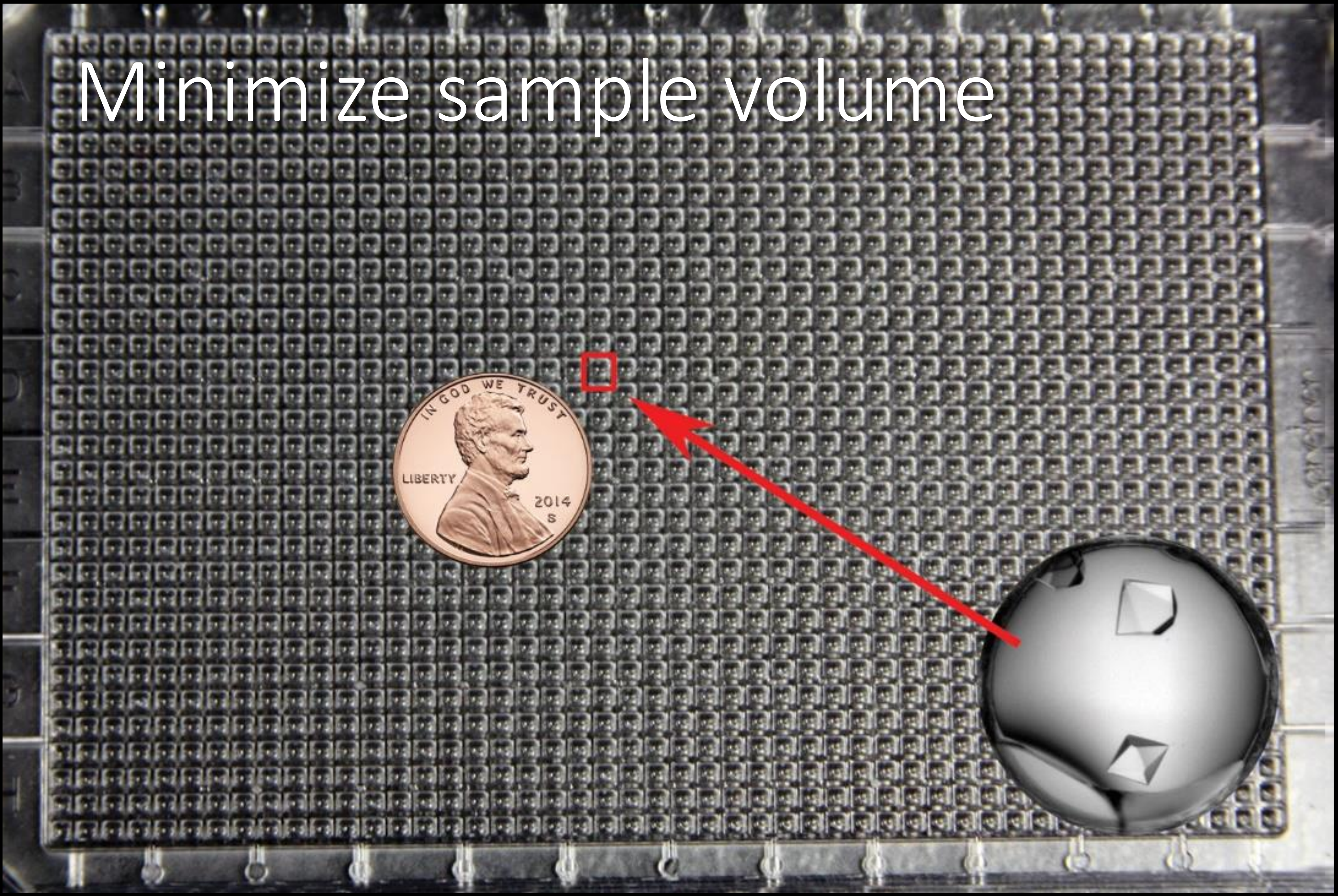
Soluble or membrane?

- There are different approaches to each type.
- At the Hauptman-Woodward High-throughput Screening Laboratory the same automated methodology is used for each but different sets of screening chemistries.
- Soluble proteins use a set of commercial and in-house designed screens.
- Membrane proteins prove the region around the critical micelle concentration (Koszleak-Rosenblum et al., Protein Science 18, 1828-1839, 2003).
- This talk just describes the soluble protein case

Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized

Minimize sample volume



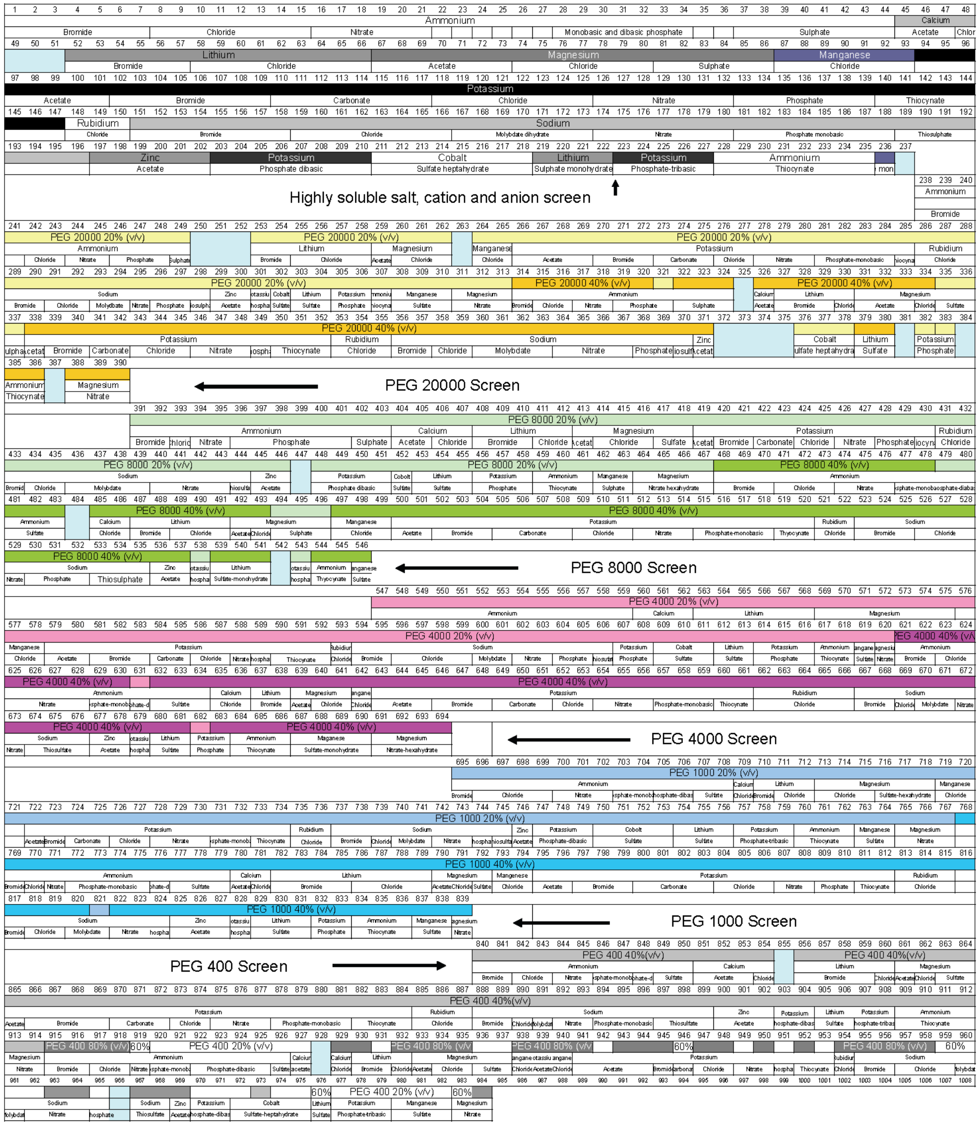
Minimize sample volume

- Each experiment uses 200 nl of protein.
- The concentration is typically a few mg/ml depending on solubility.
- Each experiment uses 200 nl of cocktail.
- 1,536 different conditions are set up.
- Total volume needed is ~400 µl
- The volume needed is larger than other methods due to the large number of screens used but the information content is high.

Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized

The HWI crystallization cocktail screen.



Highly soluble salt, cation and anion screen

PEG 20000 Screen

PEG 8000 Screen

PEG 4000 Screen

PEG 1000 Screen

PEG 400 Screen

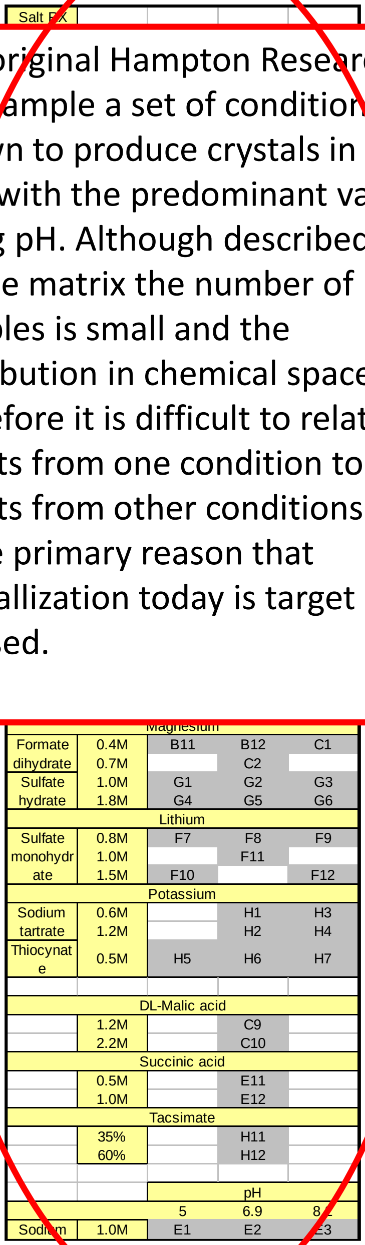
The 1536 diverse chemical cocktails (Luft et al., 2003). The 984 in-house conditions comprise a incomplete factorial sampling of 36 salts, eight buffers, and 5 different PEGs.

The remainder of 1536 cocktails are comprised of commercial screens available from Hampton Research. Specifically, in order of use; the Natrix Screen, Quick Screen, Nucleic Acid Screen, Sodium Malonate Grid, PEG/Ion, PEG 6000 Grid, Ammonium Sulfate Grid, Sodium Chloride Grid, HT Screen, Index and the SaltRx screen.

The Commercial Screens in the HWI crystallization cocktails

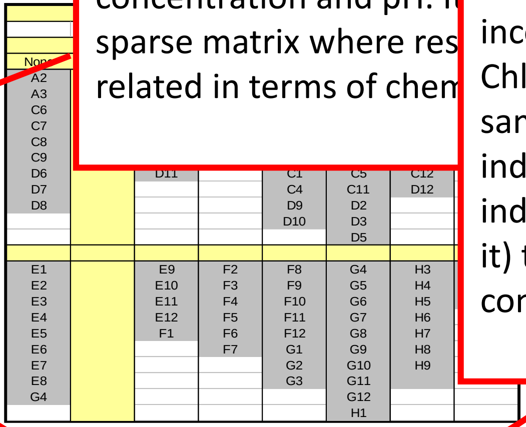
The commercial screens incorporate several distinct mechanisms of sampling the crystallization space. Examples are shown here.

The original Hampton Research 1+2 sample a set of conditions known to produce crystals in the past with the predominant variable being pH. Although described as a sparse matrix the number of samples is small and the distribution in chemical space wide therefore it is difficult to relate results from one condition to results from other conditions. This is the primary reason that crystallization today is target focused.



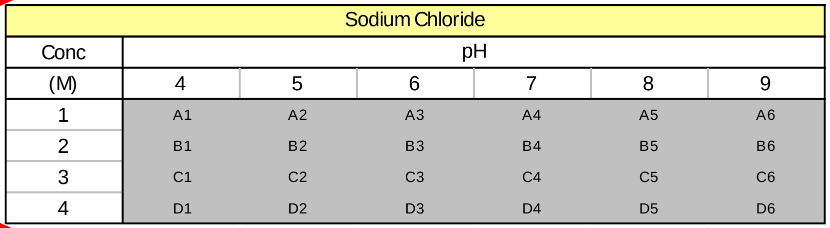
Salt Rx				
Magnesium				
Formate dihydrate	0.4M	B11	B12	C1
	0.7M		C2	
Sulfate hydrate	1.0M	G1	G2	G3
	1.8M	G4	G5	G6
Lithium				
Sulfate monohydrate	0.8M	F7	F8	F9
	1.0M		F11	
	1.5M	F10		F12
Potassium				
Sodium tartrate	0.6M		H1	H3
	1.2M		H2	H4
Thiocynate	0.5M	H5	H6	H7
DL-Malic acid				
	1.2M		C9	
	2.2M		C10	
Succinic acid				
	0.5M		E11	
	1.0M		E12	
Tacsimate				
	35%		H11	
	60%		H12	
		pH		
		5	6.9	8.2
Sodium	1.0M	E1	E2	E3

The SaltRx screen samples 22 crystallization salts with varying concentration and pH. It sparse matrix where res related in terms of chem



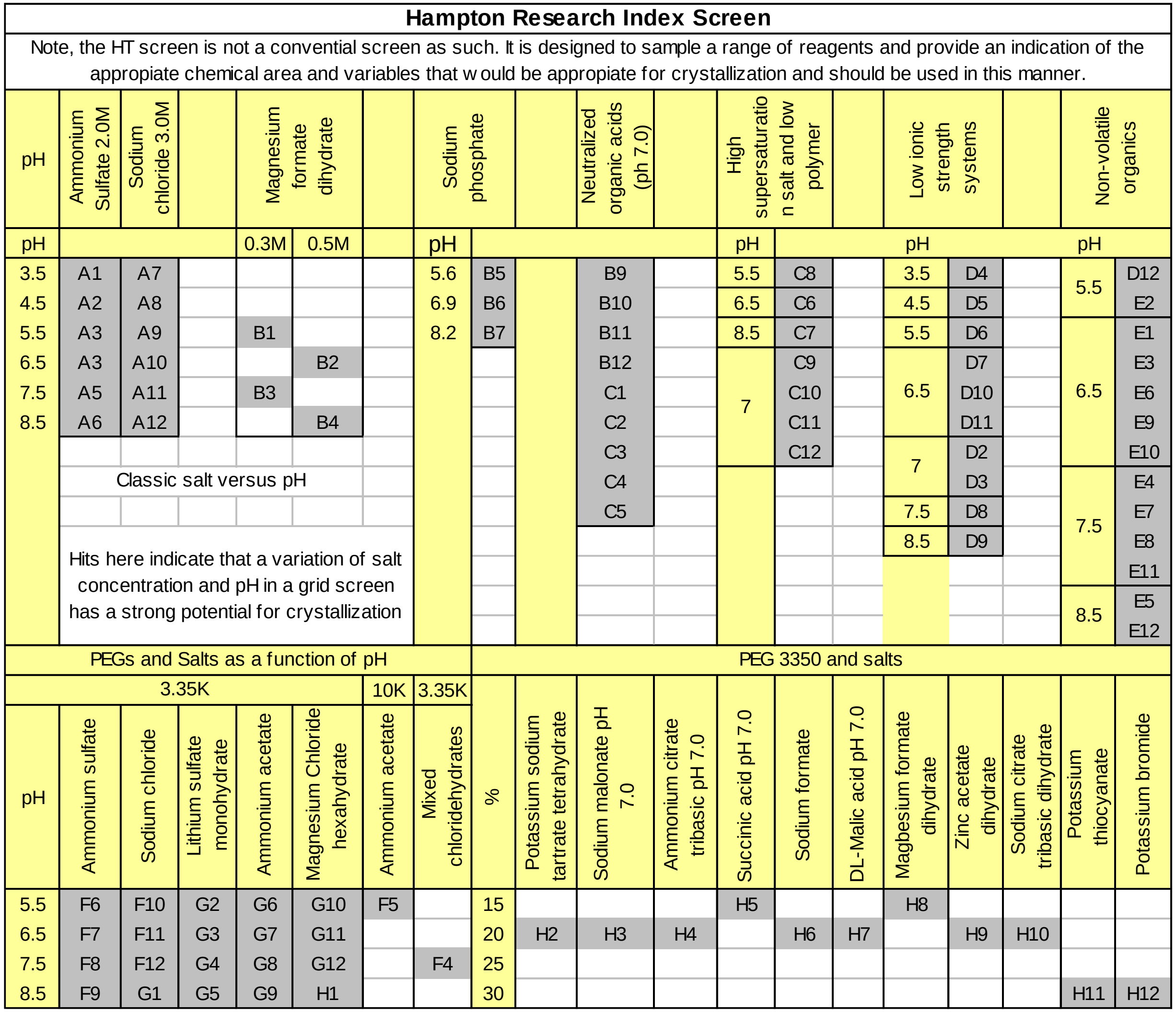
None						
A2						
A3						
C6						
C7						
C8						
C9						
D6		D11		C1	C5	C12
D7				C4	C11	D12
D8				D9	D2	
				D10	D3	
					D5	
E1		E9	F2	F8	G4	H3
E2		E10	F3	F9	G5	H4
E3		E11	F4	F10	G6	H5
E4		E12	F5	F11	G7	H6
E5		F1	F6	F12	G8	H7
E6			F7	G1	G9	H8
E7				G2	G10	H9
E8				G3	G11	
G4					G12	
					H1	

A number of Grid screens are incorporated, in this case Sodium Chloride. These provide a fine sampling of a small subset of individual conditions and serve to indicate the sensitivity (or lack of it) to small changes in precipitant conditions.



Sodium Chloride						
Conc	pH					
(M)	4	5	6	7	8	9
1	A1	A2	A3	A4	A5	A6
2	B1	B2	B3	B4	B5	B6
3	C1	C2	C3	C4	C5	C6
4	D1	D2	D3	D4	D5	D6

A special case – The Hampton Research Index Screen



Hampton Research Index Screen

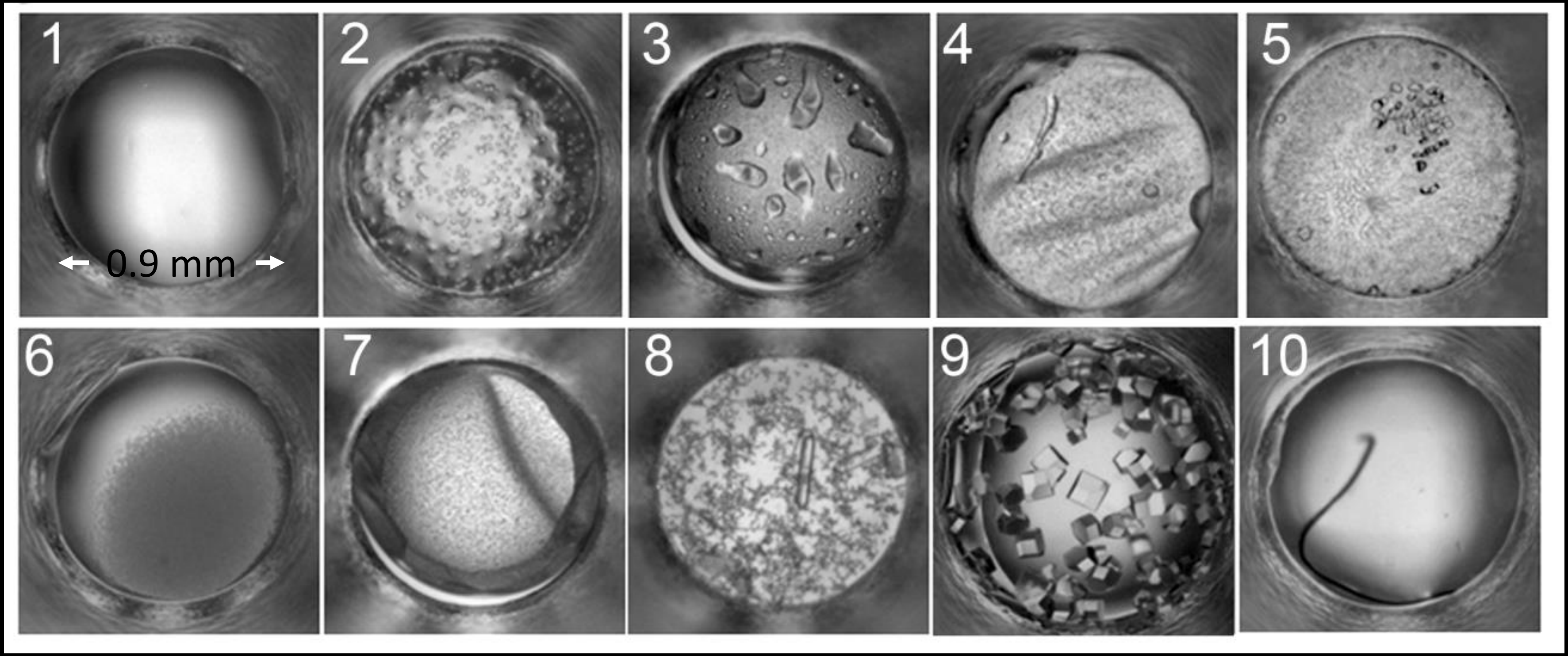
Note, the HT screen is not a convential screen as such. It is designed to sample a range of reagents and provide an indication of the appropiate chemical area and variables that w ould be appropiate for crystallization and should be used in this manner.

pH	Ammonium Sulfate 2.0M	Sodium chloride 3.0M		Magnesium formate dihydrate			Sodium phosphate			Neutralized organic acids (ph 7.0)		High supersaturation salt and low polymer			Low ionic strength systems			Non-volatile organics	
pH				0.3M	0.5M		pH					pH			pH			pH	
3.5	A1	A7					5.6	B5		B9		5.5	C8		3.5	D4		5.5	D12
4.5	A2	A8					6.9	B6		B10		6.5	C6		4.5	D5			E2
5.5	A3	A9		B1			8.2	B7		B11		8.5	C7		5.5	D6		6.5	E1
6.5	A3	A10			B2					B12		7	C9		6.5	D7			E3
7.5	A5	A11		B3						C1			C10			D10			E6
8.5	A6	A12			B4					C2			C11			D11			E9
										C3			C12		7	D2			E10
	Classic salt versus pH									C4						D3		7.5	E4
										C5					7.5	D8			E7
	Hits here indicate that a variation of salt concentration and pH in a grid screen has a strong potential for crystallization														8.5	D9			E8
																			E11
																		8.5	E5
																			E12

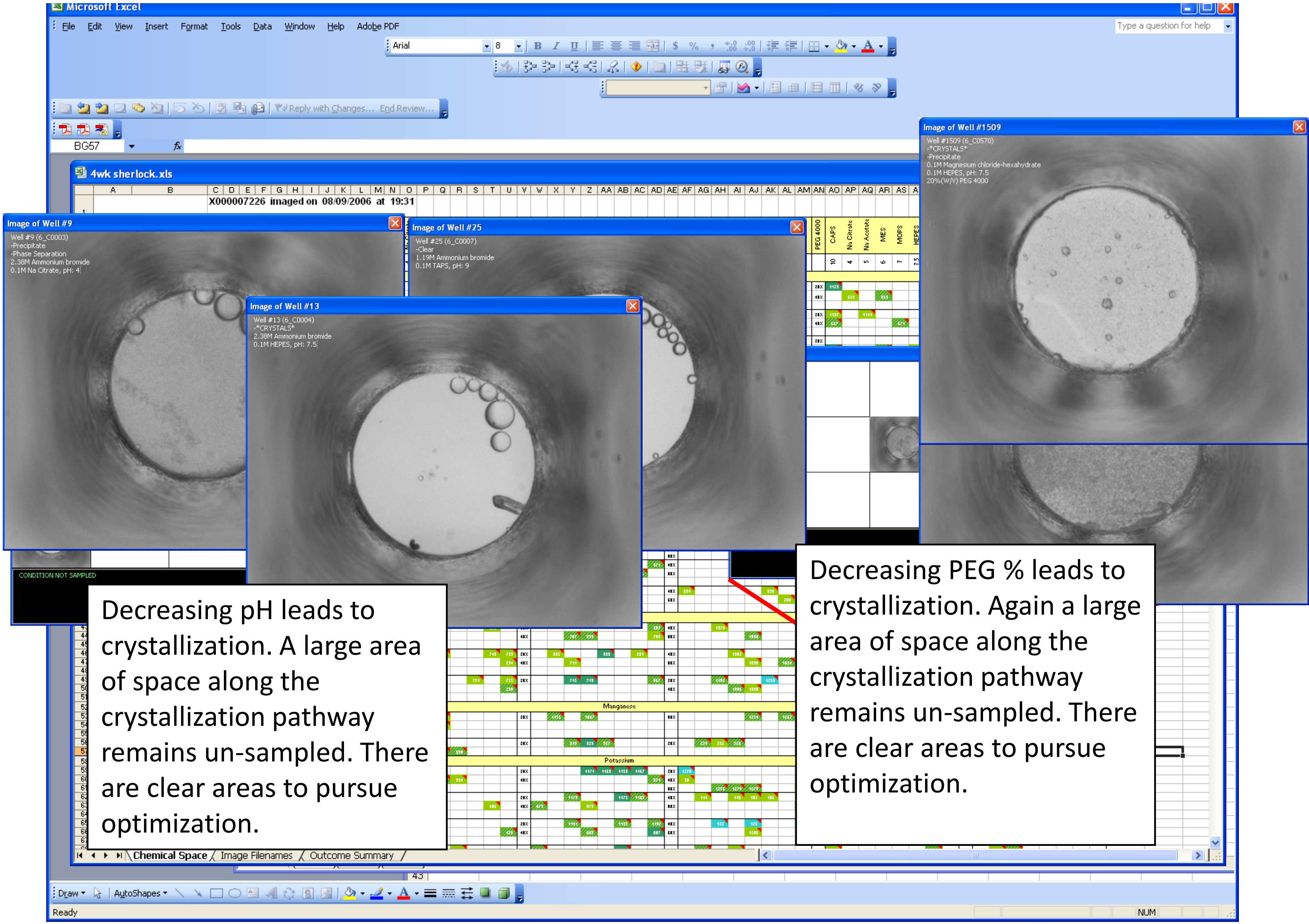
PEGs and Salts as a function of pH								PEG 3350 and salts											
3.35K						10K	3.35K												
pH	Ammonium sulfate	Sodium chloride	Lithium sulfate monohydrate	Ammonium acetate	Magnesium Chloride hexahydrate	Ammonium acetate	Mixed chloridehydrates	%	Potassium sodium tartrate tetrahydrate	Sodium malonate pH 7.0	Ammonium citrate tribasic pH 7.0	Succinic acid pH 7.0	Sodium formate	DL-Malic acid pH 7.0	Magbesium formate dihydrate	Zinc acetate dihydrate	Sodium citrate tribasic dihydrate	Potassium thiocyanate	Potassium bromide
5.5	F6	F10	G2	G6	G10	F5		15				H5			H8				
6.5	F7	F11	G3	G7	G11			20	H2	H3	H4		H6	H7		H9	H10		
7.5	F8	F12	G4	G8	G12		F4	25											
8.5	F9	G1	G5	G9	H1			30										H11	H12

Coarse test for chemical conditions likely to produce crystallization

Imaging



The volume is designed such that the complete drop is within the depth of focus.
Imaging takes place before the protein is setup (a control), immediately after and then at one week intervals for 6 weeks.


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for help
Arial 8
Reply with Changes... End Review...
BG57
4wk sherlock.xls
Chemical Space / Image Filenames / Outcome Summary
Draw AutoShapes
Ready NUM
PEG 4000 CAPS Na Citrate Na Acetate MES MOPS HEPES
Manganese
Potassium
CONDITION NOT SAMPLED

X000007226 imaged on 08/09/2006 at 19:31

Image of Well #9

Well #9 (6_C0003)
-Precipitate
-Phase Separation
2.38M Ammonium bromide
0.1M Na Citrate, pH: 4

Image of Well #25

Well #25 (6_C0007)
-Clear
1.19M Ammonium bromide
0.1M TAPS, pH: 9

Image of Well #13

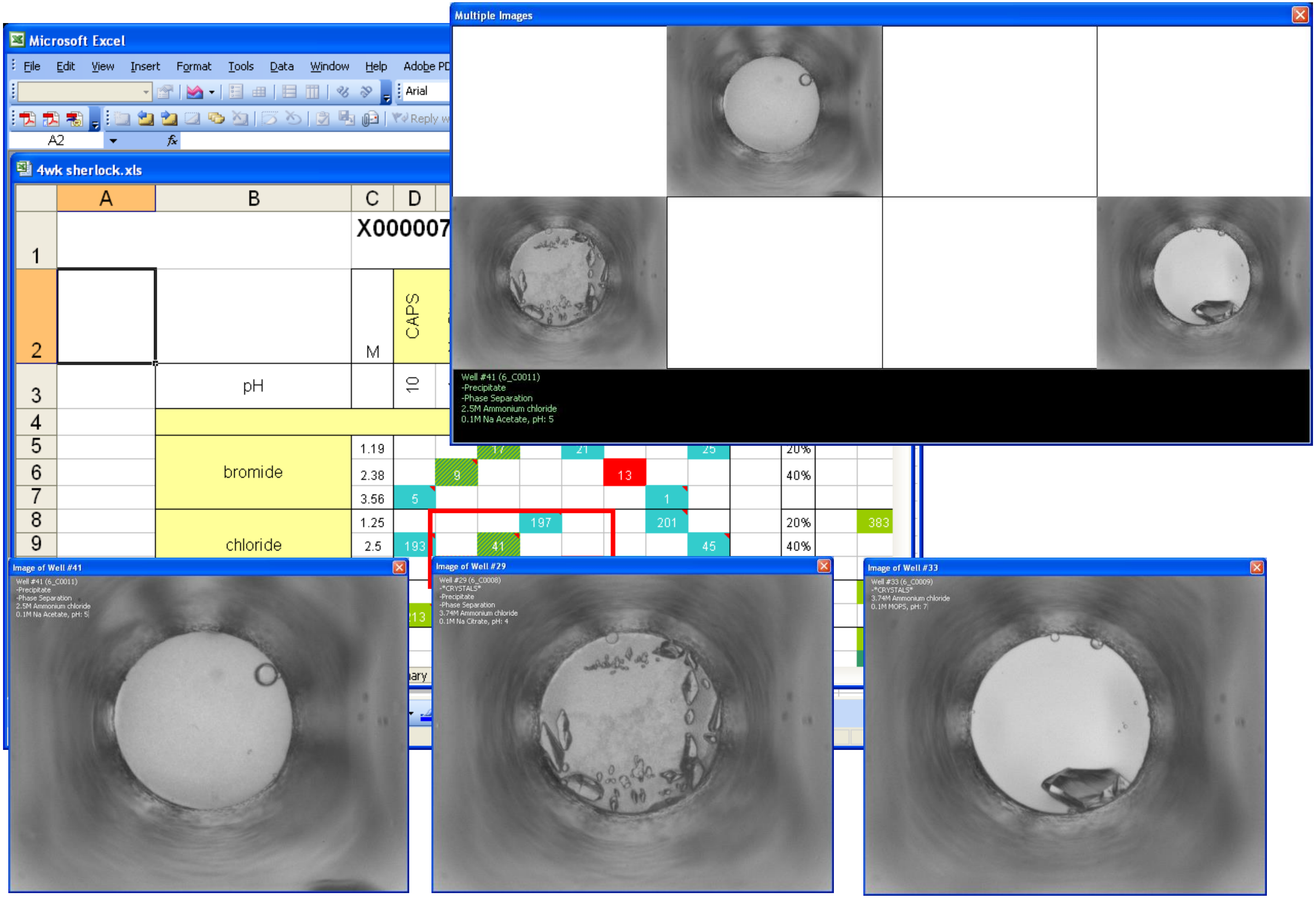
Well #13 (6_C0004)
-*CRYSTALS*
2.38M Ammonium bromide
0.1M HEPES, pH: 7.5

Image of Well #1509

Well #1509 (6_C0570)
-*CRYSTALS*
-Precipitate
0.1M Magnesium chloride-hexahydrate
0.1M HEPES, pH: 7.5
20%(W/V) PEG 4000

Decreasing pH leads to crystallization. A large area of space along the crystallization pathway remains un-sampled. There are clear areas to pursue optimization.

Decreasing PEG % leads to crystallization. Again a large area of space along the crystallization pathway remains un-sampled. There are clear areas to pursue optimization.



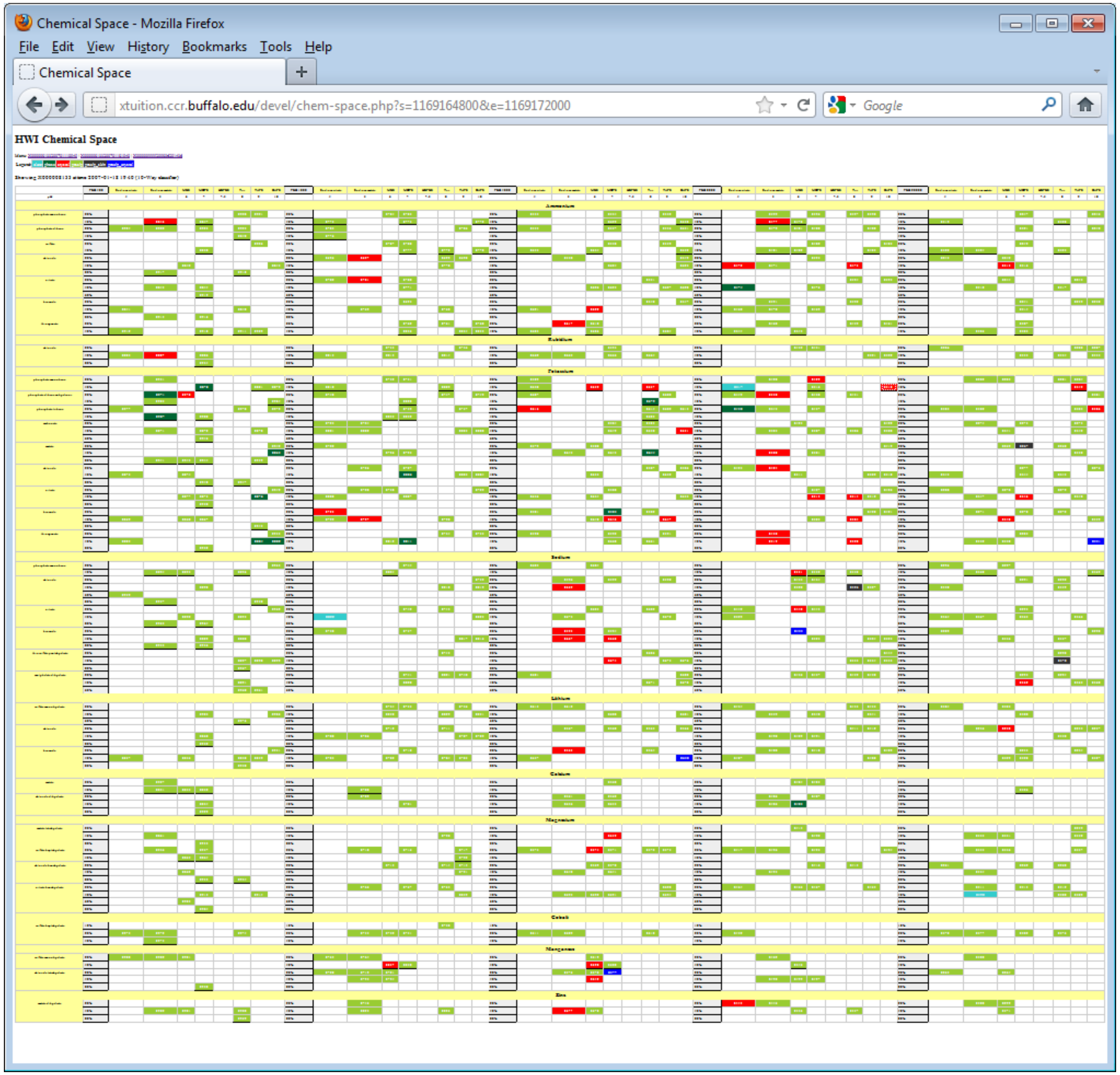
	A	B	C	D											
1			X000007												
2			M	CAPS											
3		pH		10											
4															
5		bromide	1.19			17		21			25		20%		
6			2.38		9				13				40%		
7			3.56	5						1					
8		chloride	1.25				197			201			20%		383
9			2.5	193		41					45		40%		

Well #41 (6_C0011)
-Precipitate
-Phase Separation
2.5M Ammonium chloride
0.1M Na Acetate, pH: 5

Well #41 (6_C0011)
-Precipitate
-Phase Separation
2.5M Ammonium chloride
0.1M Na Acetate, pH: 5

Well #29 (6_C0008)
-*CRYSTALS*
-Precipitate
-Phase Separation
3.74M Ammonium chloride
0.1M Na Citrate, pH: 4

Well #33 (6_C0009)
-*CRYSTALS*
3.74M Ammonium chloride
0.1M MOPS, pH: 7



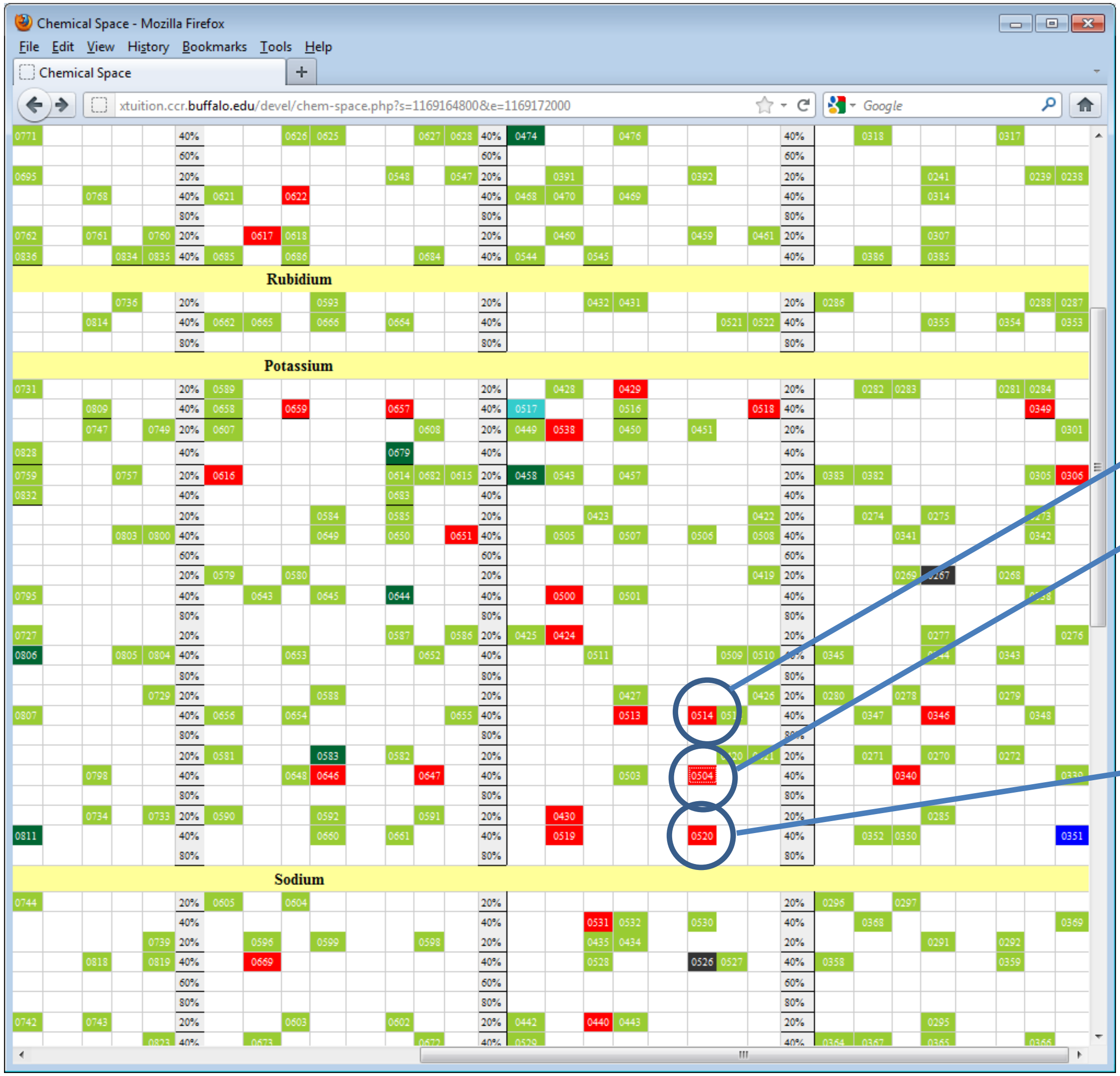
Chemical Space - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
Chemical Space
+
xtuition.ccr.buffalo.edu/devel/chem-space.php?s=1169164800&e=1169172000
Google

HWI Chemical Space

Showing [illegible] (10-Way classifier)

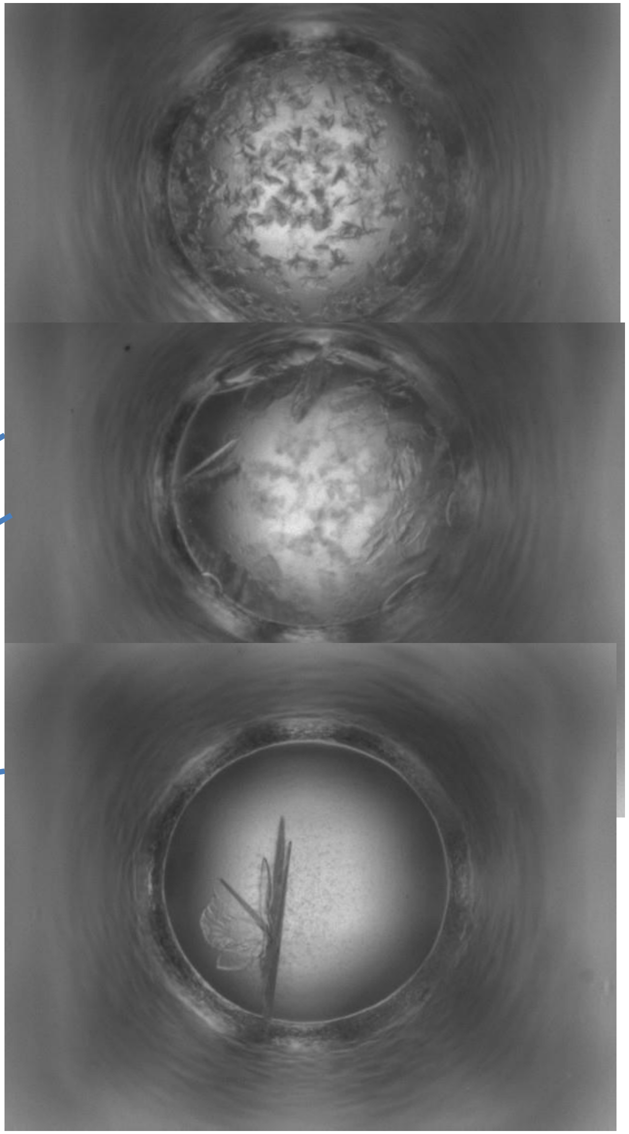
[illegible]

http://xtuition.ccr.buffalo.edu/devel/chem-space.php


Chemical Space - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Chemical Space
+
Google

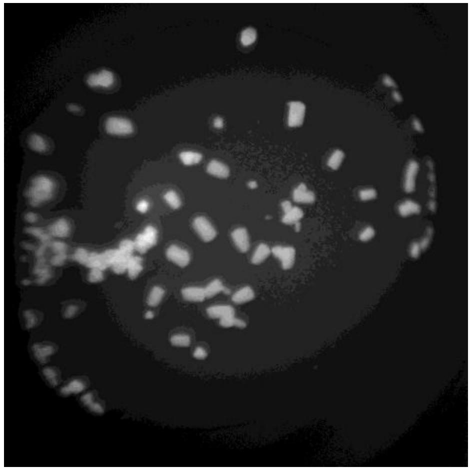
xtuition.ccr.buffalo.edu/devel/chem-space.php?s=1169164800&e=1169172000

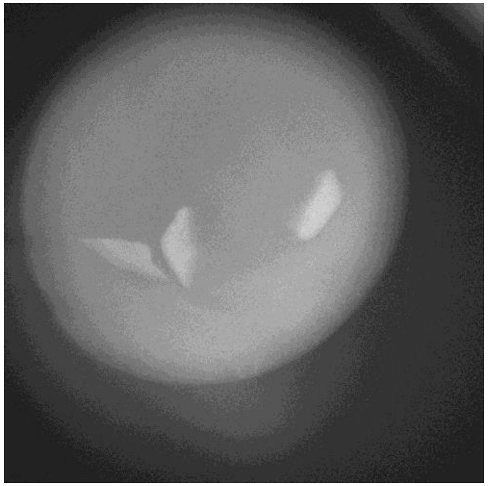
0771					40%			0626	0625			0627	0628	40%	0474			0476					40%		0318				0317		
					60%									60%									60%								
0695					20%						0548		0547	20%		0391				0392			20%				0241			0239	0238
		0768			40%	0621		0622						40%	0468	0470		0469					40%				0314				
					80%									80%									80%								
0762		0761		0760	20%		0617	0618						20%		0460				0459		0461	20%				0307				
0836			0834	0835	40%	0685		0686				0684		40%	0544		0545						40%		0386		0385				
Rubidium																															
			0736		20%				0593					20%			0432	0431					20%	0286						0288	0287
		0814			40%	0662	0665		0666		0664			40%							0521	0522	40%				0355		0354		0353
					80%									80%									80%								
Potassium																															
0731					20%	0589								20%		0428		0429					20%		0282	0283			0281	0284	
		0809			40%	0658		0659			0657			40%	0517			0516				0518	40%							0349	
		0747		0749	20%	0607						0608		20%	0449	0538		0450		0451			20%								0301
0828					40%						0679			40%									40%								
0759			0757		20%	0616					0614	0682	0615	20%	0458	0543		0457					20%	0383	0382					0305	0306
0832					40%						0683			40%									40%								
					20%				0584		0585			20%			0423					0422	20%		0274		0275			0273	
			0803	0800	40%				0649		0650		0651	40%		0505		0507		0506		0508	40%			0341				0342	
					60%									60%									60%								
					20%	0579		0580						20%								0419	20%			0269	0267		0268		
0795					40%		0643		0645		0644			40%		0500		0501					40%							0338	
					80%									80%									80%								
0727					20%						0587		0586	20%	0425	0424							20%				0277				0276
0806			0805	0804	40%			0653				0652		40%			0511				0509	0510	40%	0345			0344		0343		
					80%									80%									80%								
				0729	20%				0588					20%				0427				0426	20%	0280		0278			0279		
0807					40%	0656		0654					0655	40%				0513		0514	0512		40%		0347		0346			0348	
					80%									80%									80%								
					20%	0581			0583		0582			20%							0420	0421	20%		0271		0270		0272		
		0798			40%			0648	0646			0647		40%				0503		0504			40%			0340					0339
					80%									80%									80%								
		0734		0733	20%	0590			0592			0591		20%		0430							20%				0285				
0811					40%				0660		0661			40%		0519				0520			40%		0352	0350					0351
					80%									80%									80%								
Sodium																															
0744					20%	0605		0604						20%									20%	0296		0297					
					40%									40%			0531	0532		0530			40%		0368						0369
				0739	20%		0596		0599			0598		20%			0435	0434					20%				0291		0292		
		0818		0819	40%		0669							40%			0528			0526	0527		40%	0358					0359		
					60%									60%									60%								
					80%									80%									80%								
0742		0743			20%			0603			0602			20%	0442		0440	0443					20%				0295				
				0823	40%		0673					0672		40%	0529								40%	0364	0367		0365			0366	

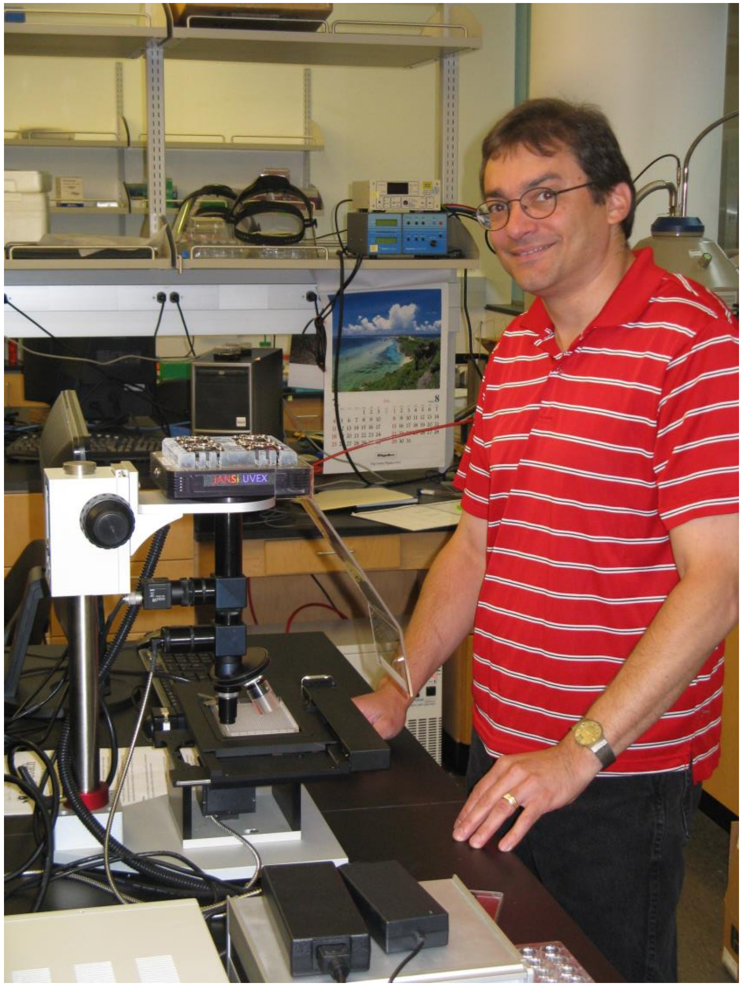


http://xtuition.ccr.buffalo.edu/devel/chem-space.php

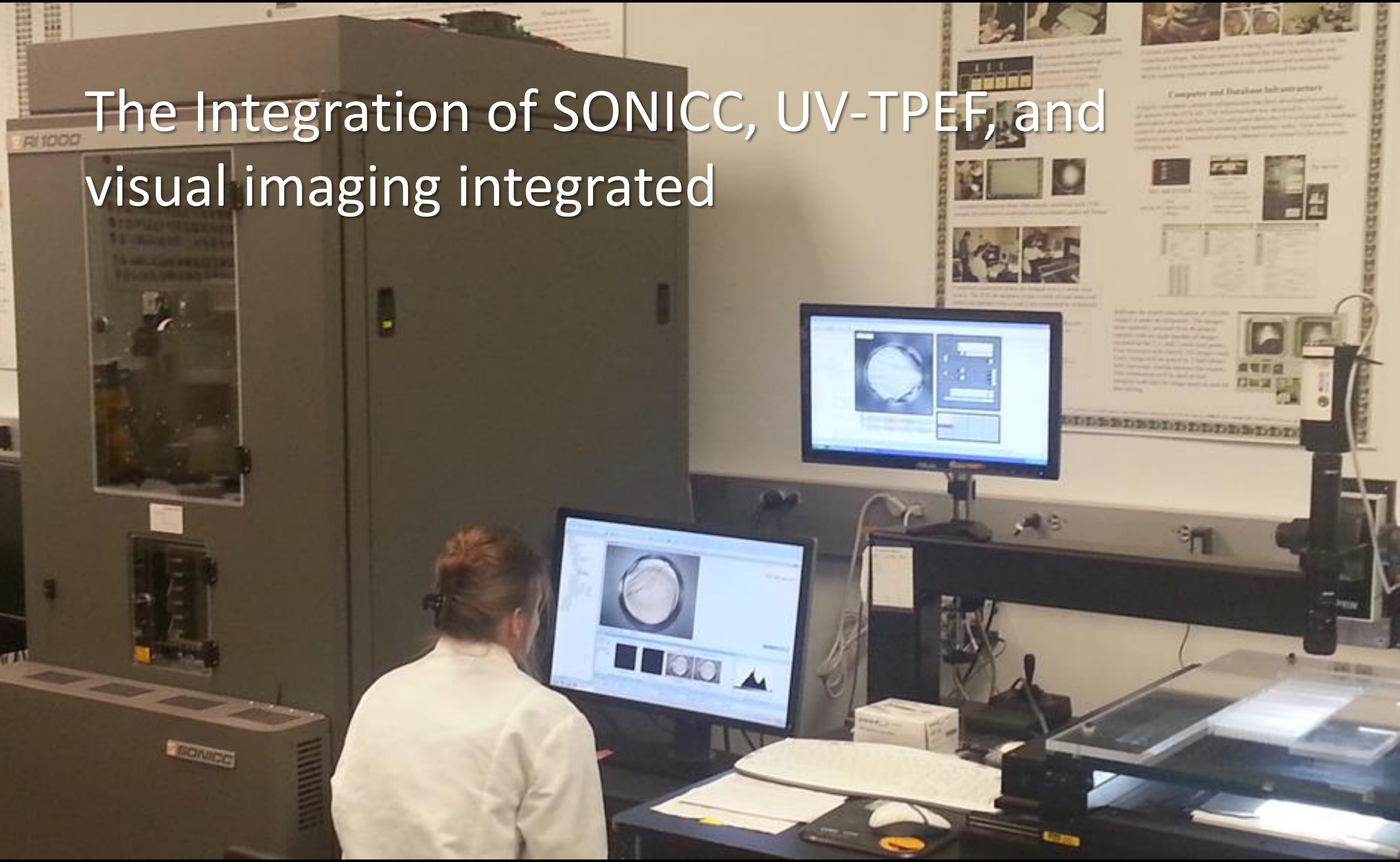
UV imaging – is it protein?







The Integration of SONICC, UV-TPEF, and visual imaging integrated



A major advance in imaging technology can identify submicron crystals

Using SONICC and UV-TPEF we can observe and verify protein crystals < 1 micron in size.

~80% of proteins in PDB low-symmetry generate SHG

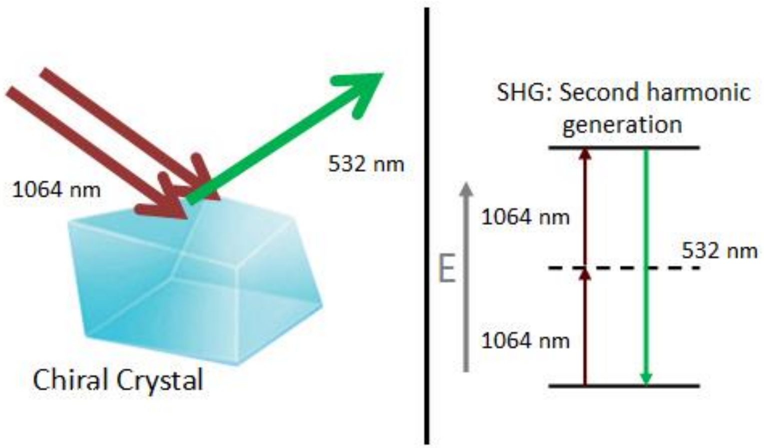


Figure 1. Two photons of IR (1064 nm) interact with a chiral crystal to generate SHG (532 nm).

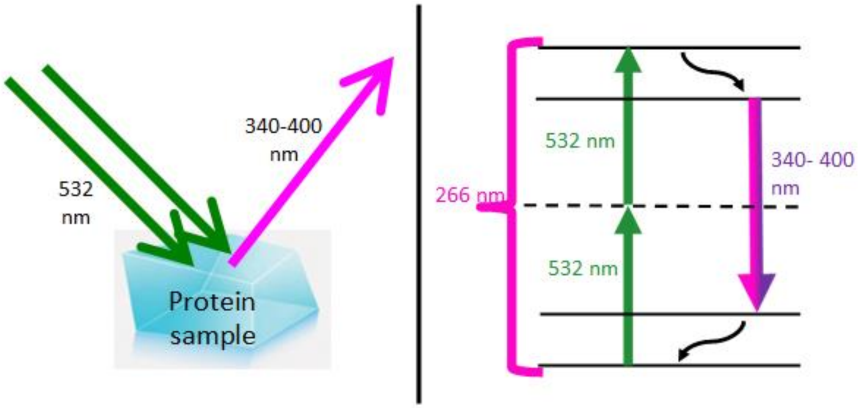


Figure 2. Depiction of UV-TPEF where two photons of green interact with a protein sample to generate UV excited fluorescence

http://www.formulatrix.com/products/protein-crystallography-tools/sonicc/how.html

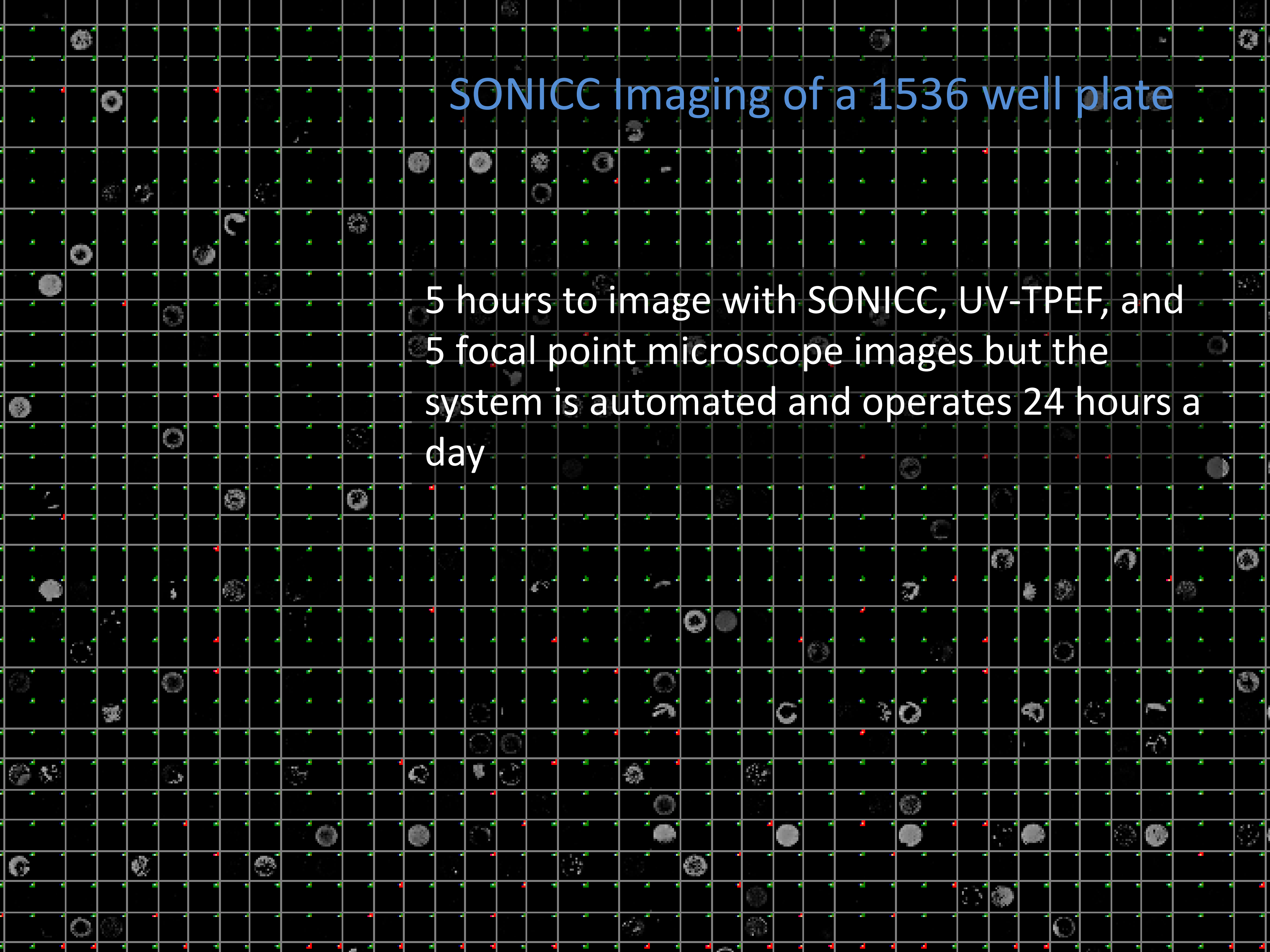
SONICC and UV-TPEF are well described elsewhere

Second-Order Nonlinear Optical Imaging of Chiral Crystals. David J. Kissick, Debbie Wanapun, and Garth J. Simpson. *Annu Rev Anal Chem*. 2011 ; 4: 419–437.

Two-photon fluorescence imaging of impurity distributions in protein crystals. Caylor, C. L., Dobrianov, I., Kimmer, C., Thorne, R. E., Zipfel, W. & Webb, W. W. (1999). Phys. Rev. E, 59, R3831–R3834

We’ll talk about their application

SONICC Imaging of a 1536 well plate

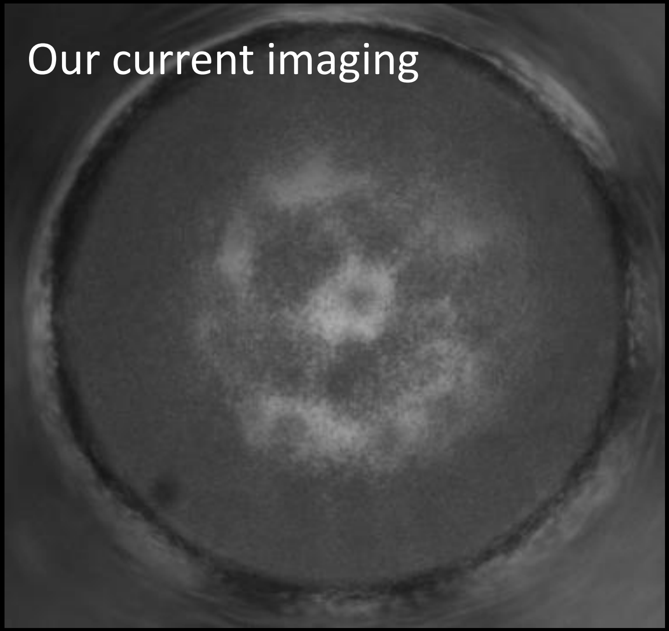


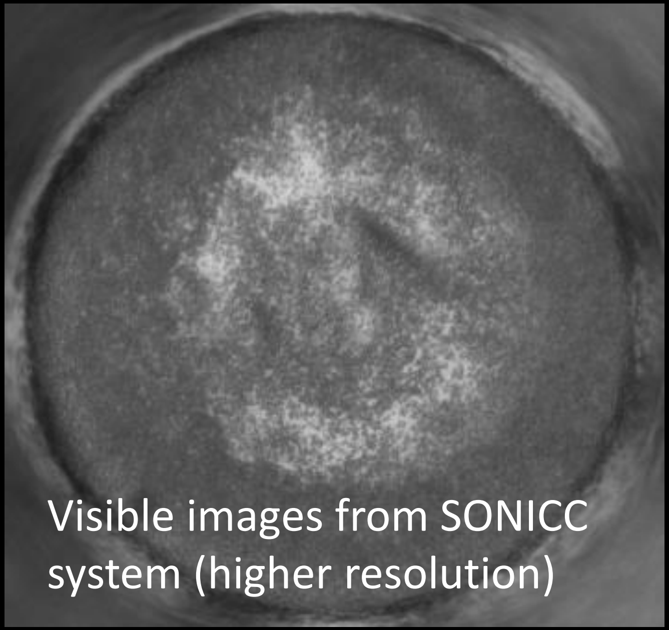
5 hours to image with SONICC, UV-TPEF, and 5 focal point microscope images but the system is automated and operates 24 hours a day

One protein in detail to lay out the experiment

Protein 1, part of the pyruvate dehydrogenase protein complex

Hampton Research PEGRx HT-F4, 4% (v/v) 2-Methyl-2,4 pentanediol, 0.1 M Citric Acid pH=3.5 20% (w/v) PEG 1500 produced the following:


Our current imaging

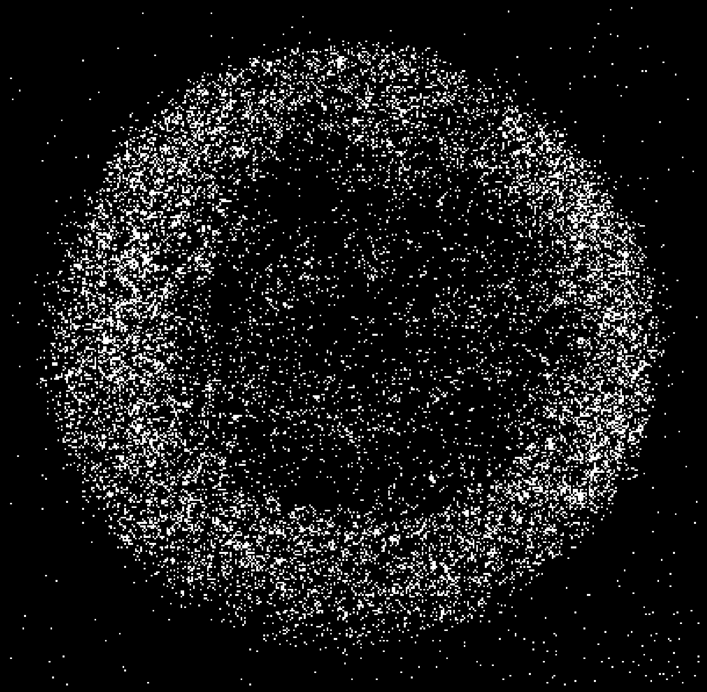

Visible images from SONICC system (higher resolution)

Immediately after the protein is added to the cocktail

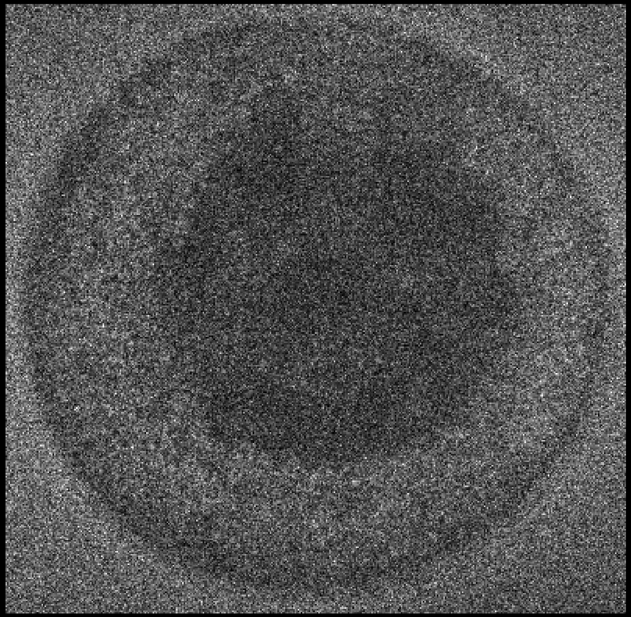
Initial use of SONICC and UV imaging

Protein 1, part of the pyruvate dehydrogenase protein complex

Hampton Research PEGRx HT-F4, 4% (v/v) 2-Methyl-2,4 pentanediol, 0.1 M Citric Acid pH=3.5 20% (w/v) PEG 1500 produced the following:



SONICC SHG image



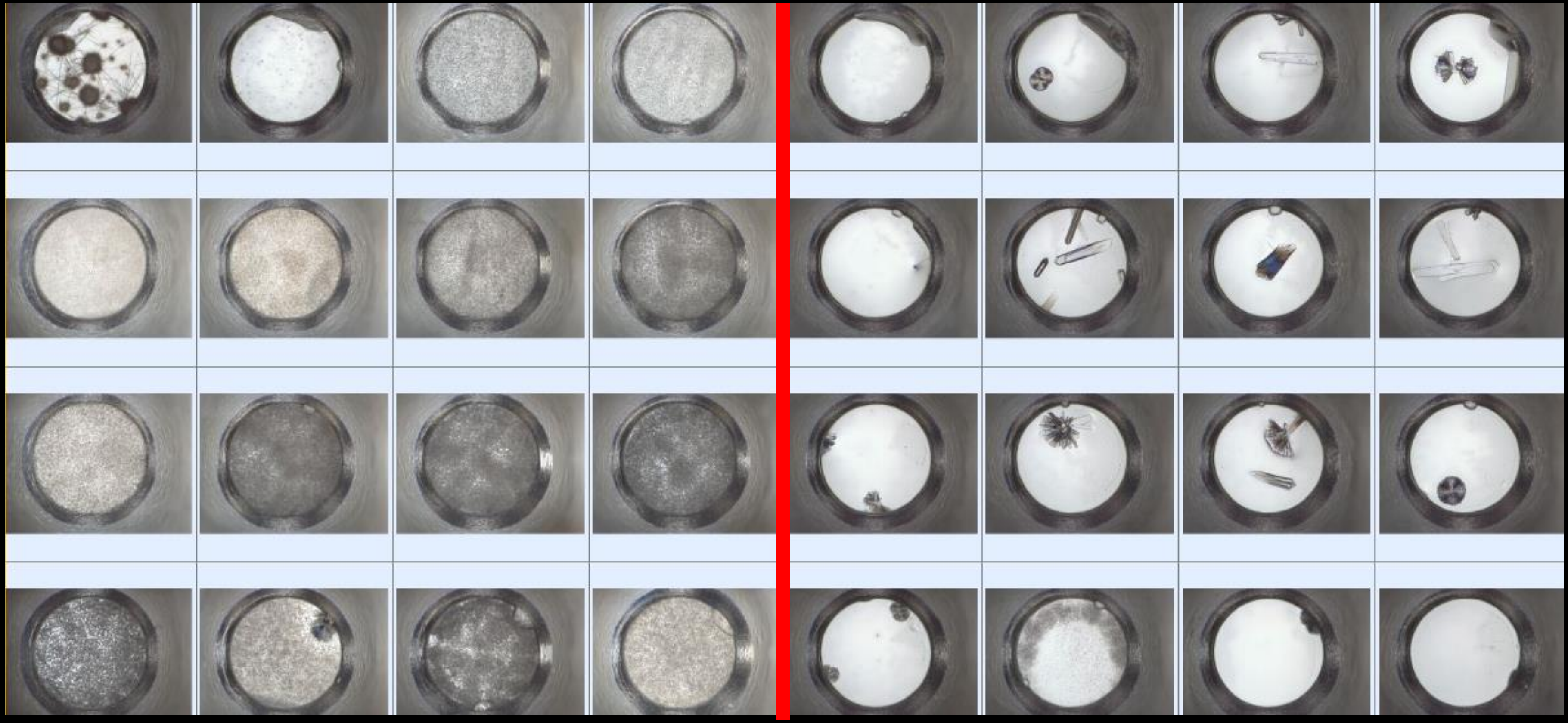
UV-TPEF image

Protein 1, part of the pyruvate dehydrogenase protein complex

Hampton Research PEGRx HT-F4, 4% (v/v) 2-Methyl-2,4 pentanediol, 0.1 M Citric Acid pH=3.5 20% (w/v) PEG 1500

X14163- Full[P]- 10mg/ml

X14163- [P]/2- 5mg/ml



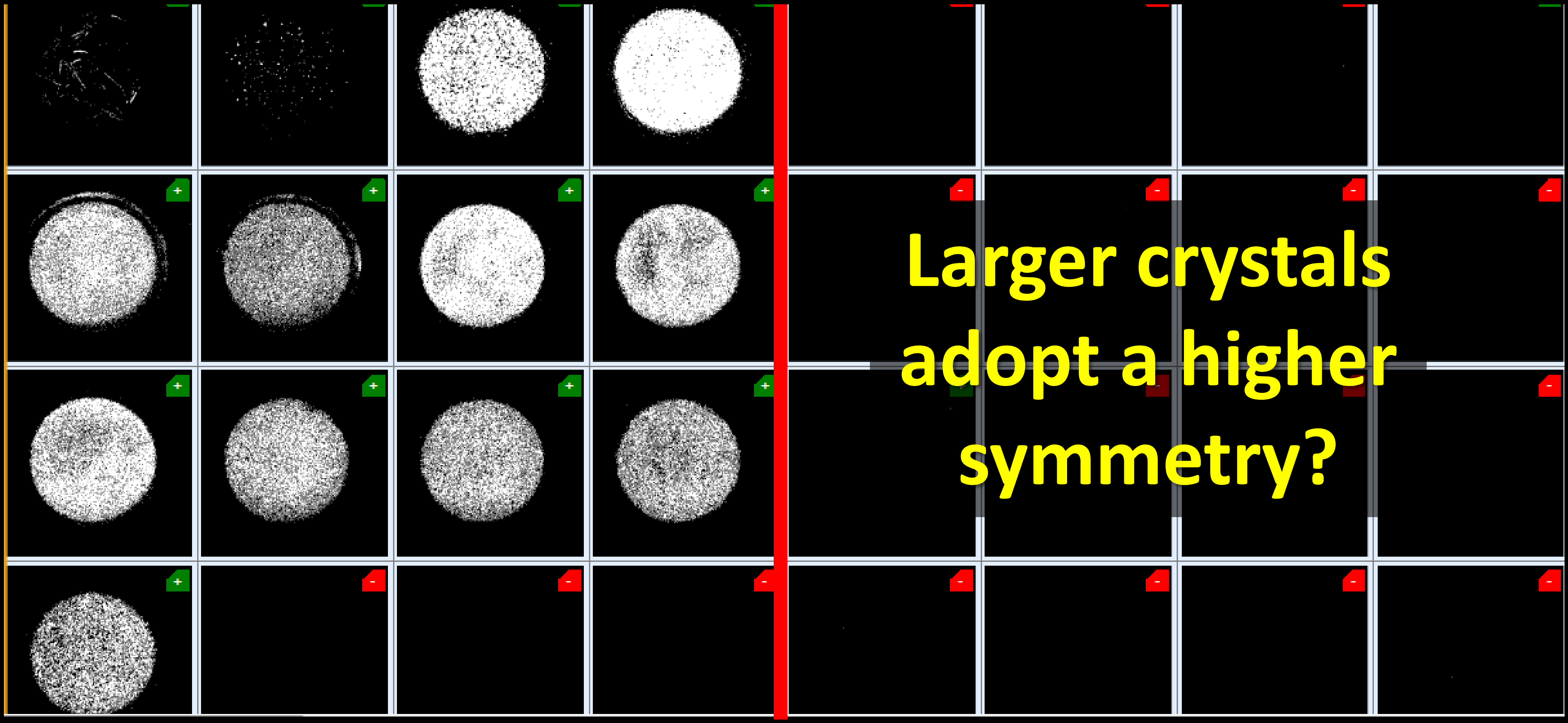
Visible at 4wk

Protein 1, part of the pyruvate dehydrogenase protein complex

Hampton Research PEGRx HT-F4, 4% (v/v) 2-Methyl-2,4 pentanediol, 0.1 M Citric Acid pH=3.5 20% (w/v) PEG 1500

X14163- Full[P]- 10mg/ml

X14163- [P]/2- 5mg/ml



Larger crystals adopt a higher symmetry?

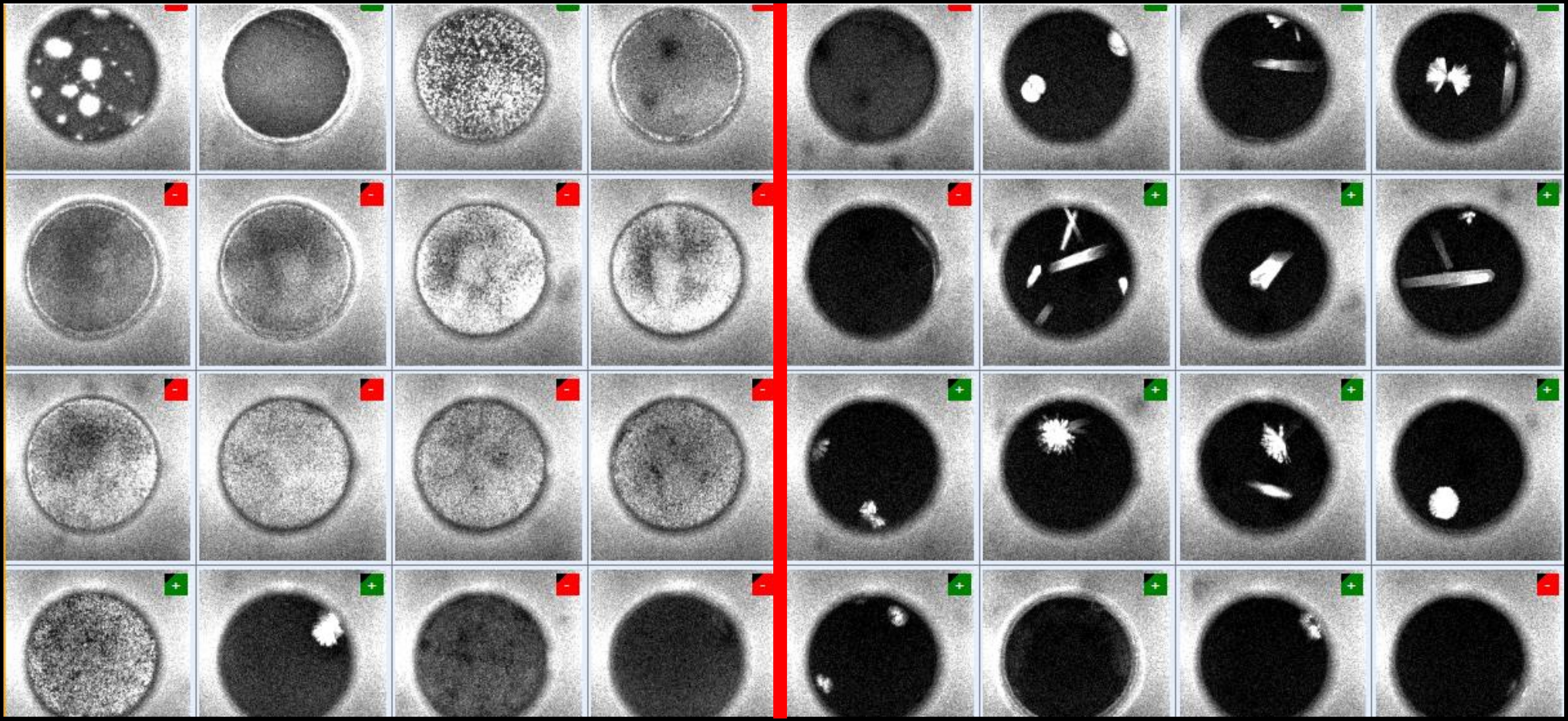
SHG at 4wk

Protein 1, part of the pyruvate dehydrogenase protein complex

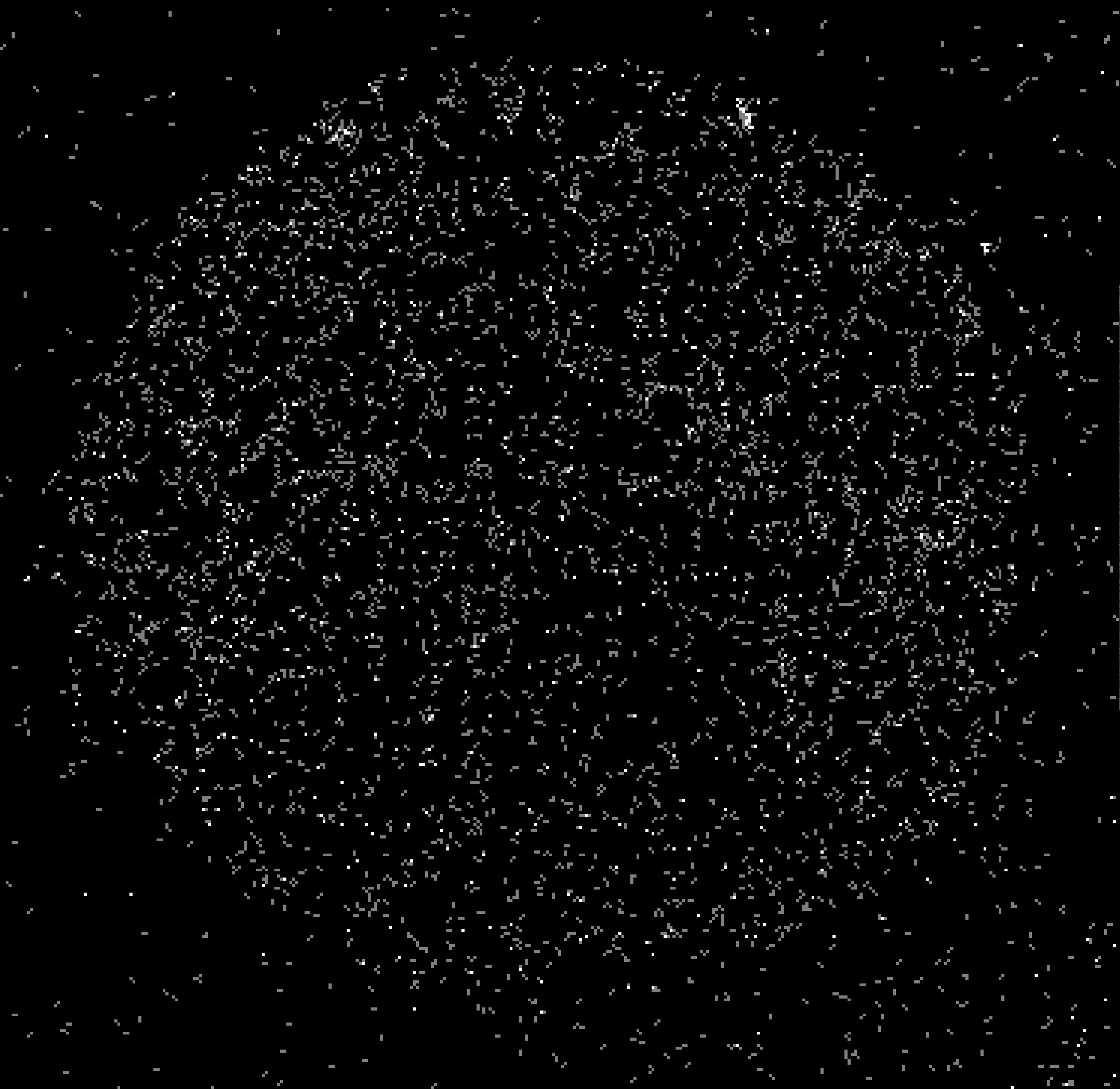
Hampton Research PEGRx HT-F4, 4% (v/v) 2-Methyl-2,4 pentanediol, 0.1 M Citric Acid pH=3.5 20% (w/v) PEG 1500

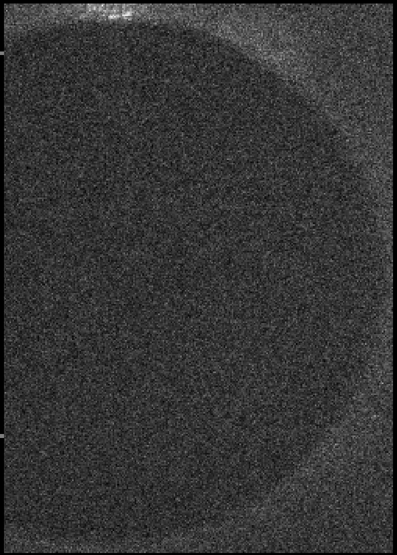
X14163- Full[P]- 10mg/ml

X14163- [P]/2- 5mg/ml



UV-TPEF at 4wk





UV-TPEF

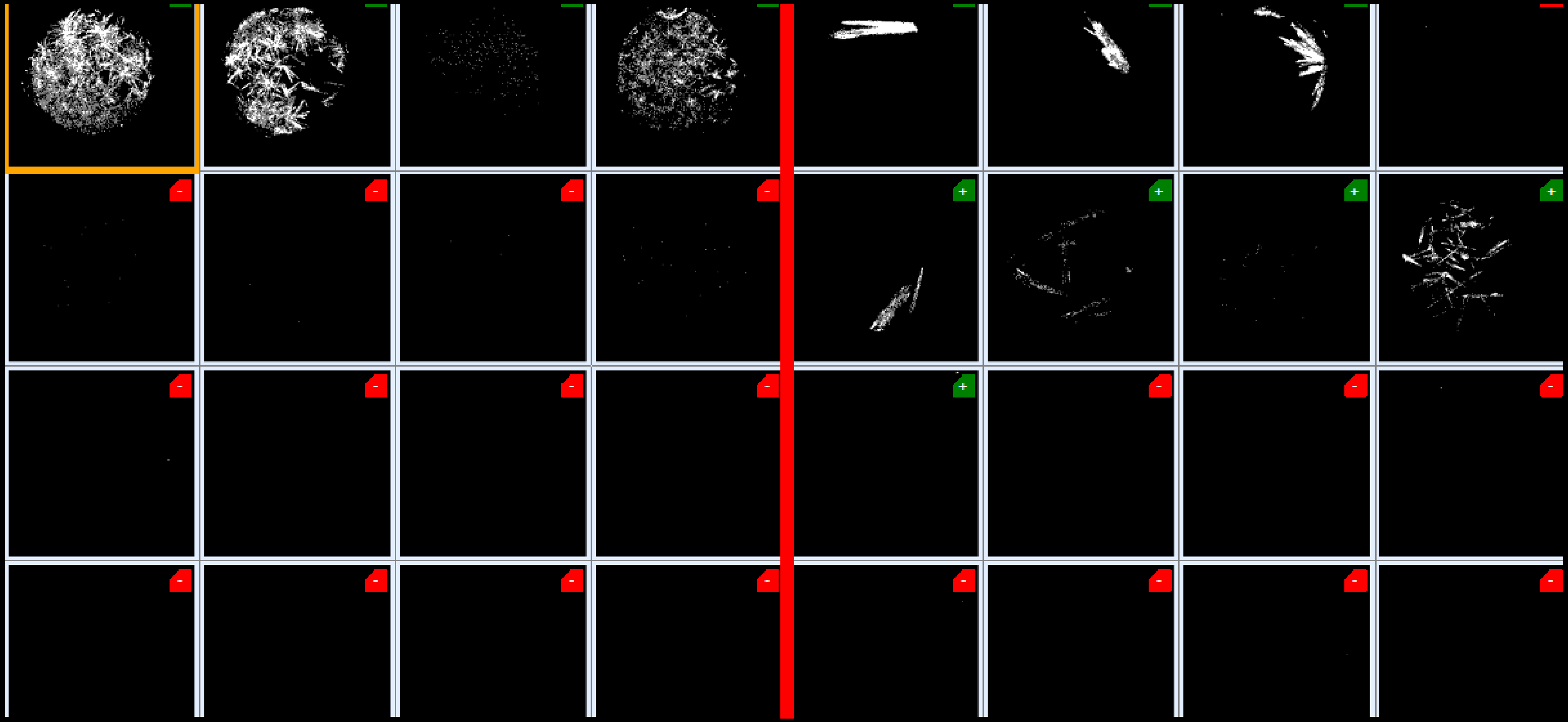
Hampton Research Ionic Liquids

5%(w/v) 1-Butyl-3-methyimidazolium dicyanamide

Protein 2

X14164- Full[P]-10 mg/ml

X14164- [P]/2- 5 mg/ml



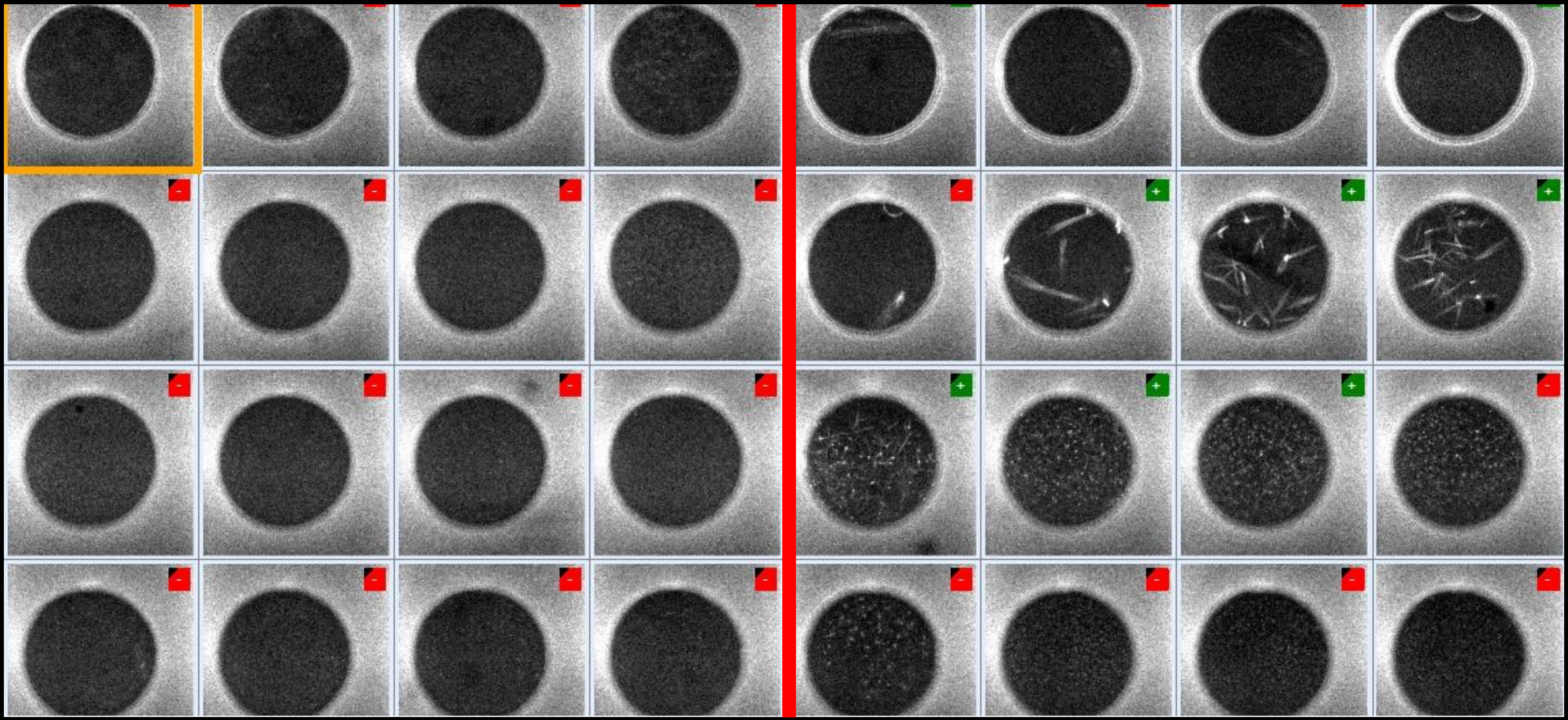
SHG at 4wk

Hampton Research Ionic Liquids
5%(w/v) 1-Butyl-3-methyimidazolium dicyanamide

Protein 2

X14164- Full[P]-10 mg/ml

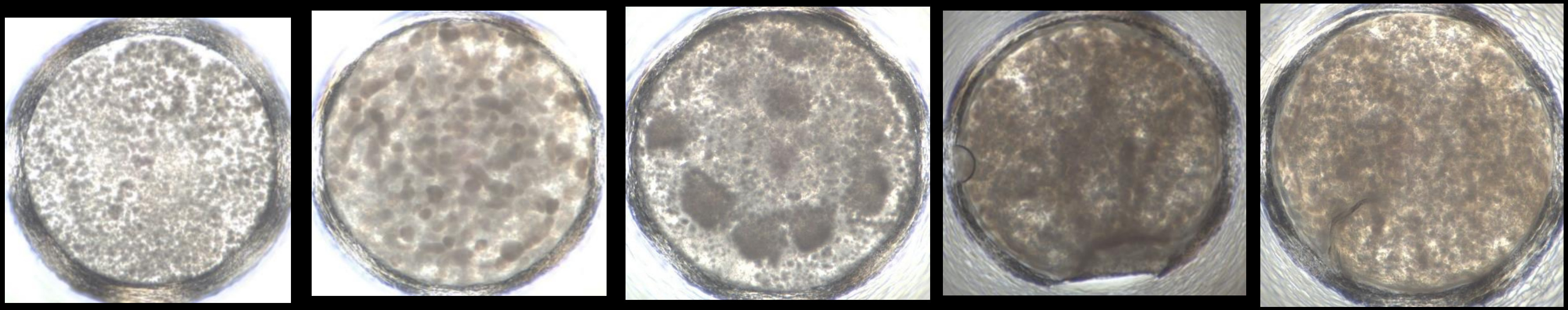
X14164- [P]/2- 5 mg/ml

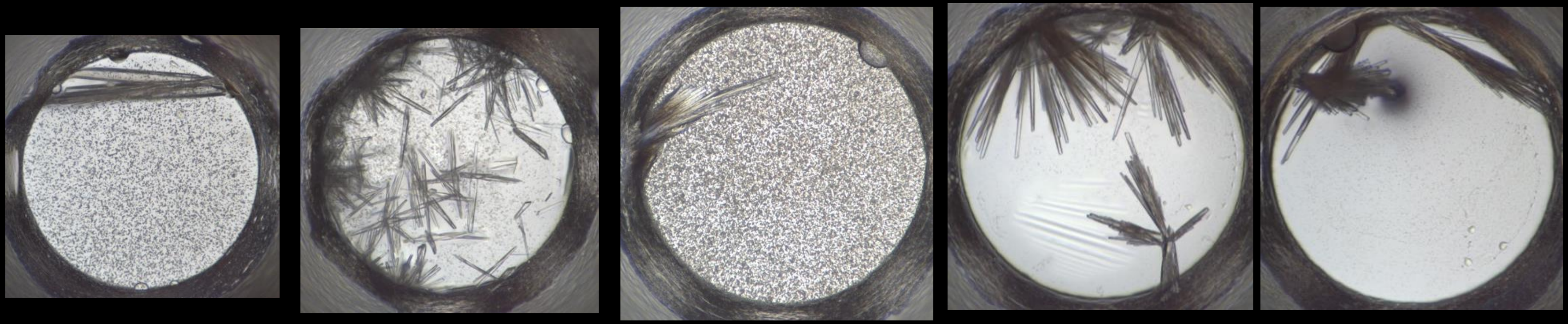


UV-TPEF at 4wk

Protein 2 (crystals identified visually in other conditions)

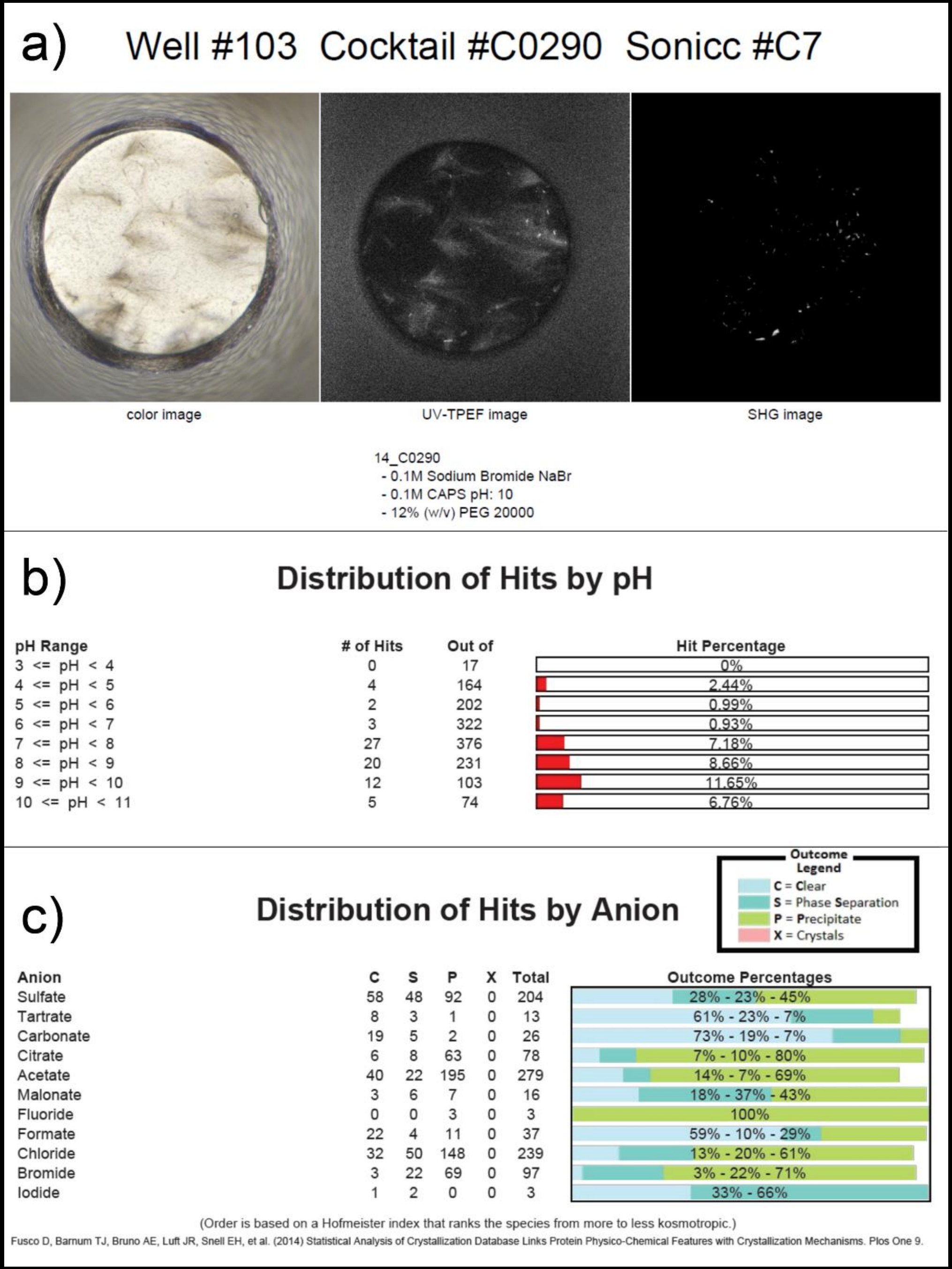
Visual image where SHG/UV-TEV signal detected





Best optimized condition

Generate automated report



a) Well #103 Cocktail #C0290 Sonicc #C7

color image UV-TPEF image SHG image

14_C0290
- 0.1M Sodium Bromide NaBr
- 0.1M CAPS pH: 10
- 12% (w/v) PEG 20000

b) **Distribution of Hits by pH**

pH Range	# of Hits	Out of	Hit Percentage
3 <= pH < 4	0	17	0%
4 <= pH < 5	4	164	2.44%
5 <= pH < 6	2	202	0.99%
6 <= pH < 7	3	322	0.93%
7 <= pH < 8	27	376	7.18%
8 <= pH < 9	20	231	8.66%
9 <= pH < 10	12	103	11.65%
10 <= pH < 11	5	74	6.76%

c) **Distribution of Hits by Anion**

Outcome Legend
- C = Clear
- S = Phase Separation
- P = Precipitate
- X = Crystals

Anion	C	S	P	X	Total	Outcome Percentages
Sulfate	58	48	92	0	204	28% - 23% - 45%
Tartrate	8	3	1	0	13	61% - 23% - 7%
Carbonate	19	5	2	0	26	73% - 19% - 7%
Citrate	6	8	63	0	78	7% - 10% - 80%
Acetate	40	22	195	0	279	14% - 7% - 69%
Malonate	3	6	7	0	16	18% - 37% - 43%
Fluoride	0	0	3	0	3	100%
Formate	22	4	11	0	37	59% - 10% - 29%
Chloride	32	50	148	0	239	13% - 20% - 61%
Bromide	3	22	69	0	97	3% - 22% - 71%
Iodide	1	2	0	0	3	33% - 66%

(Order is based on a Hofmeister index that ranks the species from more to less kosmotropic.)

Fusco D, Barnum TJ, Bruno AE, Luft JR, Snell EH, et al. (2014) Statistical Analysis of Crystallization Database Links Protein Physico-Chemical Features with Crystallization Mechanisms. Plos One 9.

Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
    - The methodology must be equal or better to any other methods
    - The amount of sample used should be minimal
    - The amount of information obtained needs to be maximal and interpretable.
    - The results must be useable, reproducible and if necessary scalable.
    - Single point failures must be eliminated or minimized

Information management

- Capture the data and make it available to the user rapidly – realtime secure ftp account.
- Provide an easy way to image the data (MacroscopeJ, a program for the analysis and classification of images).
- Backup the data, in multiple places.
- Provide full experimental details (and keep experimental samples of cocktails).
- Publish details of analysis and and keep an extensive website with practical details (getacrystal.org).

Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized.

Identify single point failures

- Where possible duplicate instrumentation.
- Have multiple plates ready to receive protein.
- For expensive instrumentation, identify alternative pathways (which may be more time consuming).
- Have very clear experimental protocol and communication strategies.

Efficient High-Throughput Crystallization is hard

- Successful high-throughput crystallization approaches require efficiency
  - The methodology must be equal or better to any other methods
  - The amount of sample used should be minimal
  - The amount of information obtained needs to be maximal and interpretable.
  - The results must be useable, reproducible and if necessary scalable.
  - Single point failures must be eliminated or minimized

Going beyond efficient crystallization is harder

There is more information in crystallization screening results than where crystals occur

Molecular Fingerprints

Molecular fingerprints are representations of chemical structures designed to capture molecular activity.

We use atomic properties and a SMILES string to capture six components:

1. Atomic number
2. Number of directly-bonded neighbors
3. Number of attached hydrogens
4. The atomic charge
5. The atomic mass
6. If the atom is contained in a ring

These components are calculated for the whole molecule in an iterative manner starting from an arbitrary non-hydrogen.

Example:
Sodium chloride, NaCl

Sodium [11,0,0,1,22.99,0]
Chlorine [17,0,0,-1,35.45,0]

Starting from Na two, properties are associated with Na and encoded by: (3,855,292,234,1) and (3,737,048,253, 1)*

One property is associated with Cl and encoded by: (2,096,516,726,1)

This information is stored in single integer with bits 3,855,292,234, 3,737,048,253 and 2,096,516,726 set to on.

* Rodgers and Hahn, J. Chem. Inf. Model. 2010, 50, 742-754

Cocktail Fingerprints

Cocktail fingerprints combine the molecular fingerprints and account for the molarity of each in the crystallization cocktail.

For example, consider a very simple example: 0.1 M sodium chloride and 0.1 M ammonium sulfate

Molecular fingerprint:	Sodium chloride	[(3855292234, 1),(3737048253, 1),(2096516726, 1)]
	Ammonium chloride	[(847680145, 1), (3855292234, 1),(2214760707, 1)]

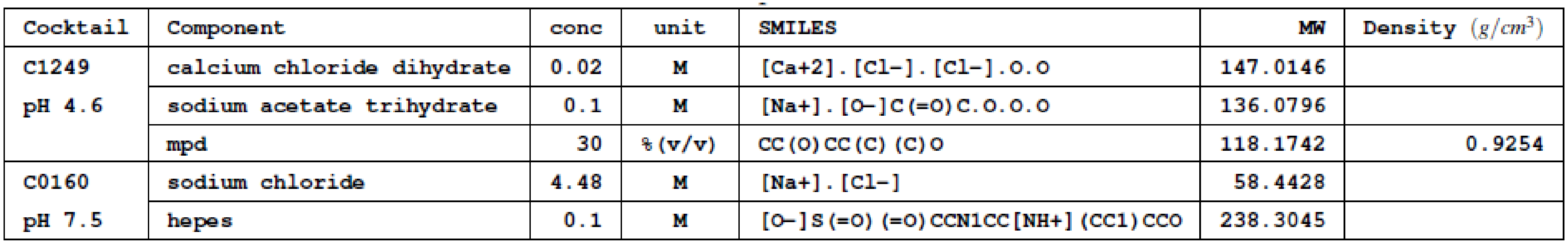
Bit (3855292234, 1) is common in both so we set the bit count to 2 and multiply by the molar concentration

Cocktail fingerprint: [(3855292234, 0.2),(3737048253, 0.1),(2096516726, 0.1)
(847680145, 1),(2214760707, 0.1)]

The bits are stored in a single 64 bit number with the bit counts stored in a sequential array

Comparing Cocktail Fingerprints

Take a real example of two crystallization screening cocktails as stored in our database



Cocktail	Component	conc	unit	SMILES	MW	Density (g/cm^3)
C1249	calcium chloride dihydrate	0.02	M	[Ca+2].[Cl-].[Cl-].O.O	147.0146	
pH 4.6	sodium acetate trihydrate	0.1	M	[Na+].[O-]C(=O)C.O.O.O	136.0796	
	mpd	30	%(v/v)	CC(O)CC(C)(C)O	118.1742	0.9254
C0160	sodium chloride	4.48	M	[Na+].[Cl-]	58.4428	
pH 7.5	hepes	0.1	M	[O-]S(=O)(=O)CCN1CC[NH+](CC1)CCO	238.3045	

First convert all concentrations to molarity

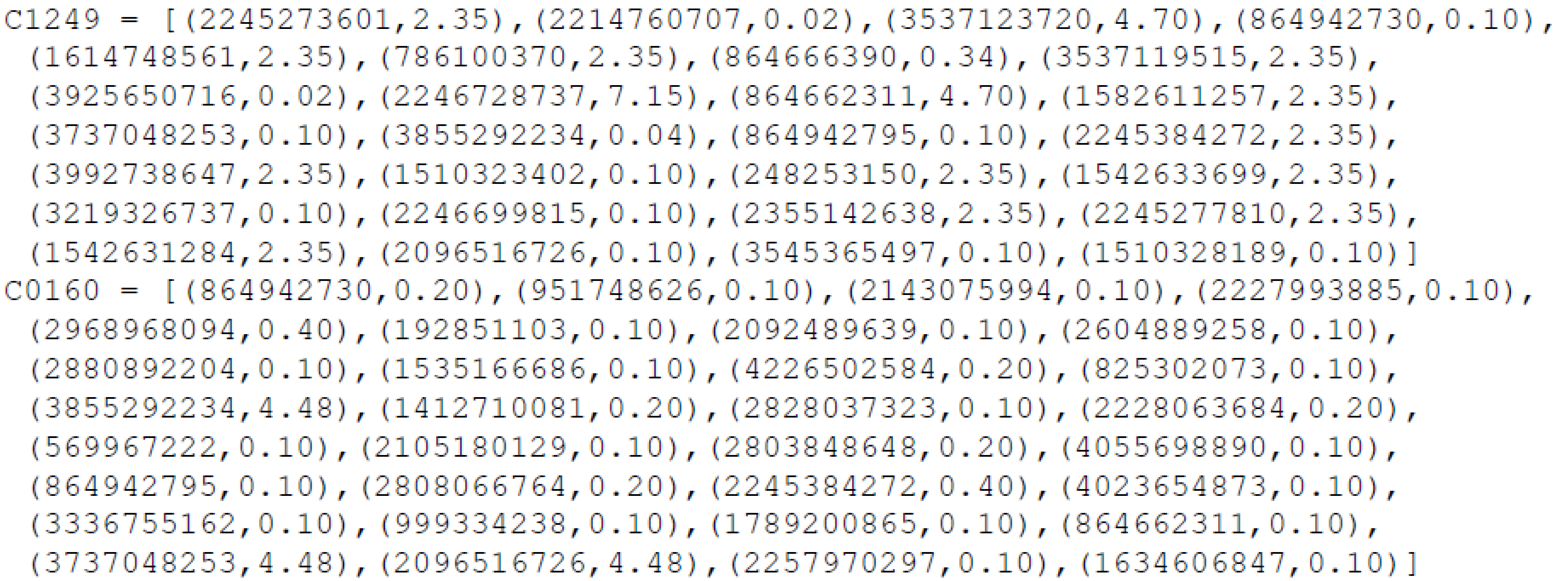
Cocktail C1249 contains 30% (v/v) MPD. This is converted to 2.349 M. PEGs are more problematic as they can be polydispersive in which case the average molecular weight is used.

The cocktail fingerprint is calculated using the molecular fingerprint for each component and its molar concentration

$$F_k = \sum_{i=1}^{n} f_{ik}[c_i]$$

Where F_k is the cocktail fingerprint, *i* is the number of components, *f* the molecular fingerprint and *c* the concentration

An example of two cocktail fingerprints



```
C1249 = [(2245273601,2.35),(2214760707,0.02),(3537123720,4.70),(864942730,0.10),
 (1614748561,2.35),(786100370,2.35),(864666390,0.34),(3537119515,2.35),
 (3925650716,0.02),(2246728737,7.15),(864662311,4.70),(1582611257,2.35),
 (3737048253,0.10),(3855292234,0.04),(864942795,0.10),(2245384272,2.35),
 (3992738647,2.35),(1510323402,0.10),(248253150,2.35),(1542633699,2.35),
 (3219326737,0.10),(2246699815,0.10),(2355142638,2.35),(2245277810,2.35),
 (1542631284,2.35),(2096516726,0.10),(3545365497,0.10),(1510328189,0.10)]
C0160 = [(864942730,0.20),(951748626,0.10),(2143075994,0.10),(2227993885,0.10),
 (2968968094,0.40),(192851103,0.10),(2092489639,0.10),(2604889258,0.10),
 (2880892204,0.10),(1535166686,0.10),(4226502584,0.20),(825302073,0.10),
 (3855292234,4.48),(1412710081,0.20),(2828037323,0.10),(2228063684,0.20),
 (569967222,0.10),(2105180129,0.10),(2803848648,0.20),(4055698890,0.10),
 (864942795,0.10),(2808066764,0.20),(2245384272,0.40),(4023654873,0.10),
 (3336755162,0.10),(999334238,0.10),(1789200865,0.10),(864662311,0.10),
 (3737048253,4.48),(2096516726,4.48),(2257970297,0.10),(1634606847,0.10)]
```

Each is encoded in a single hashed number.

The Bray-Curtis dissimilarity measure is used to compute the dissimilarity.

$$BC(F_i, F_j) = \sum_k | F_{ik} - F_{jk} | / \sum_k | F_{ik} + F_{jk} |$$

This pH is incorporated along with the ability to weight individual components and the Cocktail Dissimilarity coefficient calculated.

$$CD_{coeff} = \frac{1}{sum(w)} \left(\left(\frac{| E(pH_i) - E(pH_j))}{14} \right) w_1 + BC(F_i, F_j) w_2 \right)$$

The Cocktail Similarity coefficient given by:

$$CS_{coeff} = 1 - CD_{coeff}$$

The Dissimilarity Measure Over the Whole Screen

Aspects of the screen design are clearly seen

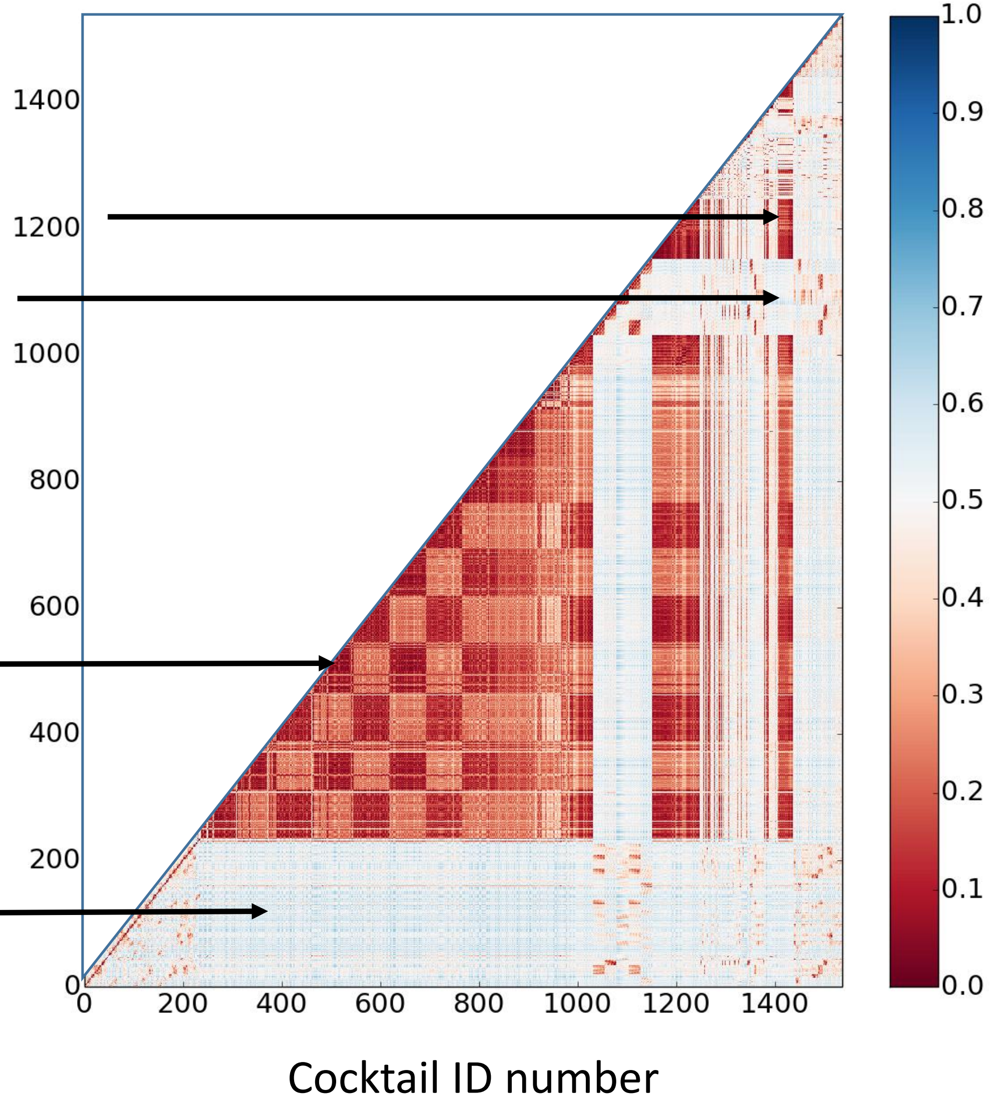
Hampton Research PEG/Ion screen

Hampton Research Silver Bullets

PEG based conditions sampling different molecular weight PEGS at two concentrations

Salt based screens

The scale is normalized to the most dissimilar chemical conditions

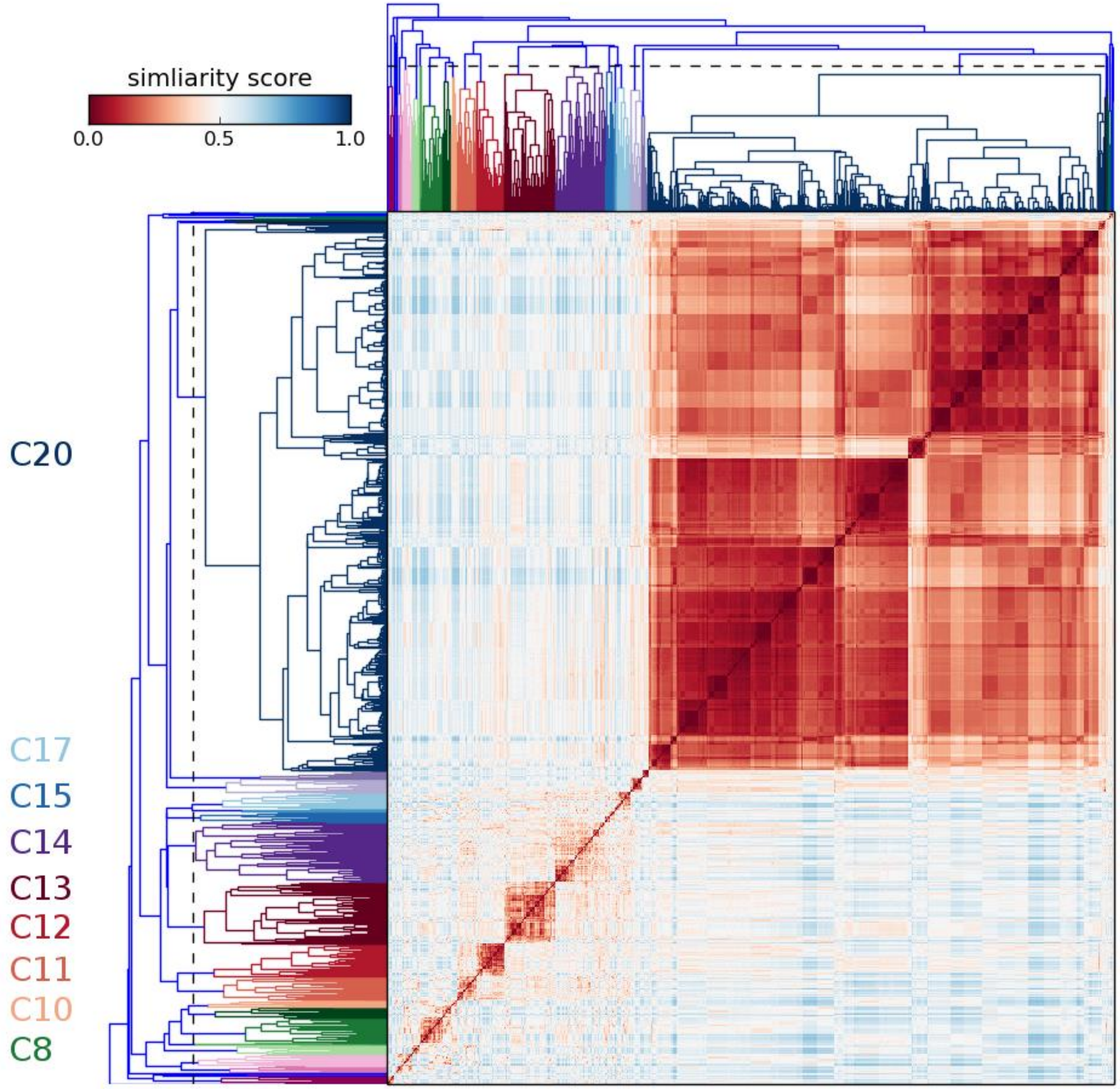


Automatic Clustering of the Results

Hierarchical clustering using a default max cophenetic distance cutoff of one standard deviation identified 28 clusters.

PEG based conditions

Salts with different anions and cations

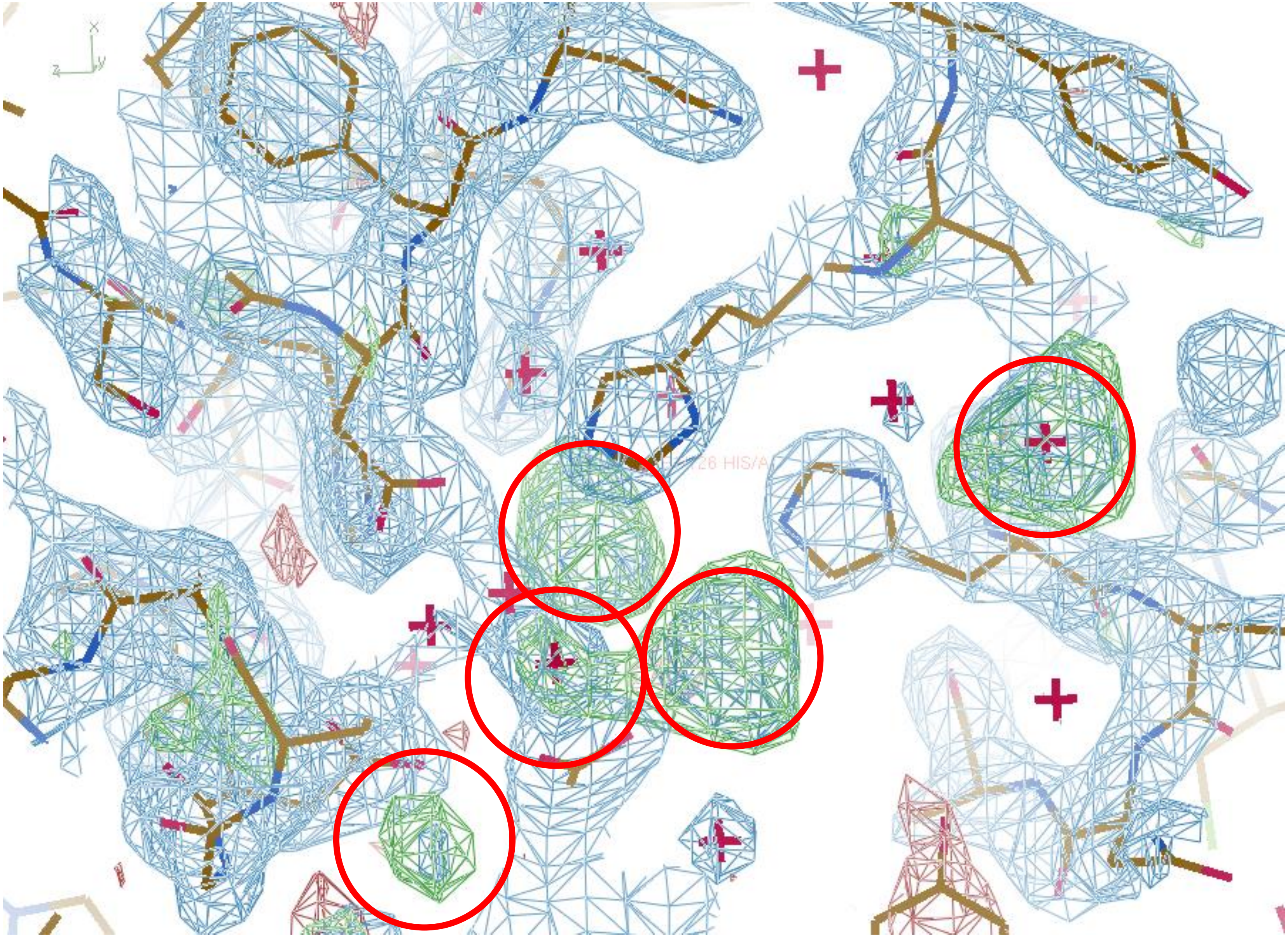


A structural genomics target.

BfR192, is a 343 residue protein with a molecular weight of 39.77 kDa. For crystallization screening the protein was prepared at 7.4 mg/ml in a 5 mM DTT, 100 mM NaCl, 10 mM Tris-HCl, pH 7.5, 0.02% NaN_3 buffer.

Several potential crystallization conditions for BfR192 SelMet labeled protein were identified

The optimized conditions for crystallization combined 5µl of the protein at 7.4 mg/ml concentration was mixed with the precipitant containing 320mM potassium acetate, 100 mM sodium acetate, pH 6.5 in 1:1 ratio. Crystals appeared in one week.

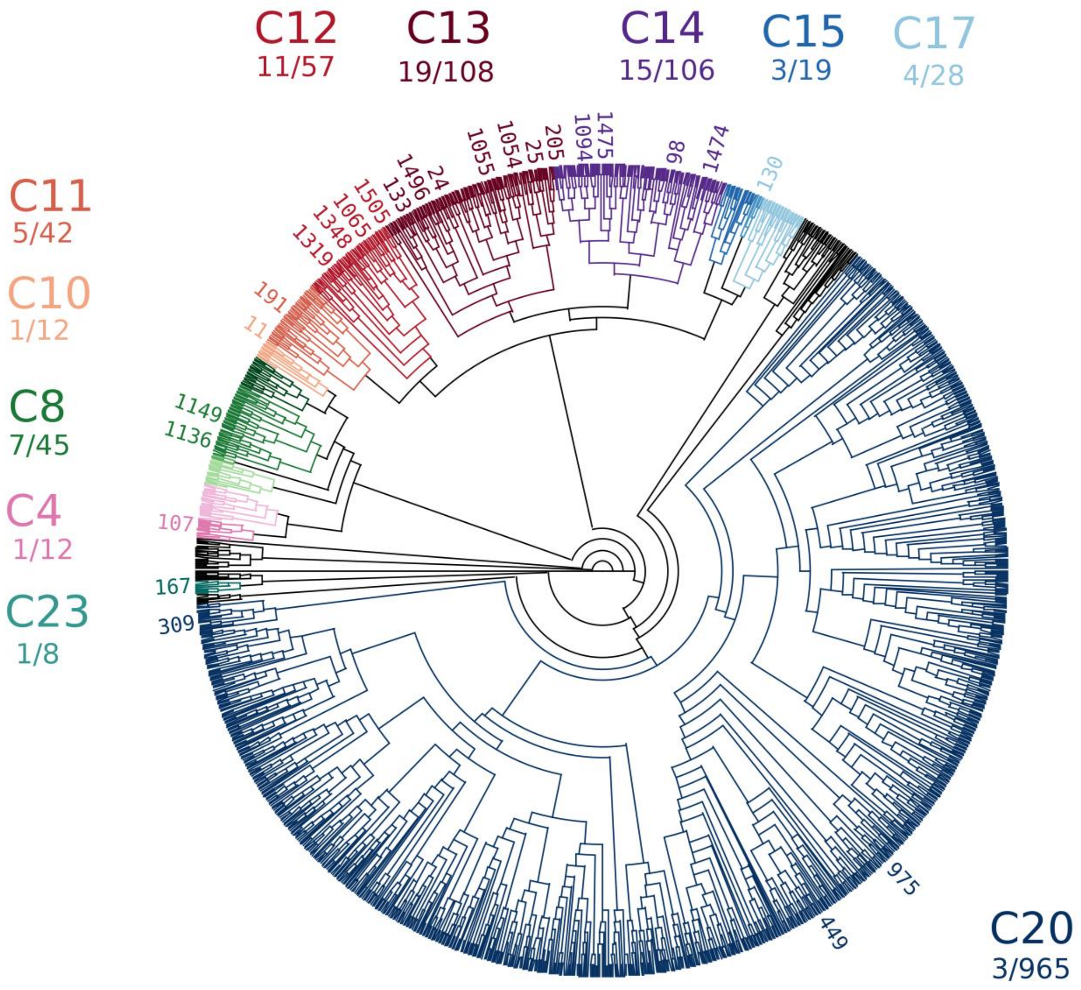


PDB ID 3DMA as deposited in the PDB

Overlaying Crystal Hits on the Cocktail Clustering

Conditions showing crystal hits are given for each cluster along with the total number of cocktails in that cluster.

A selection of cocktails that showed hits are listed on the outside of the dendogram. For clarity not all hits are shown

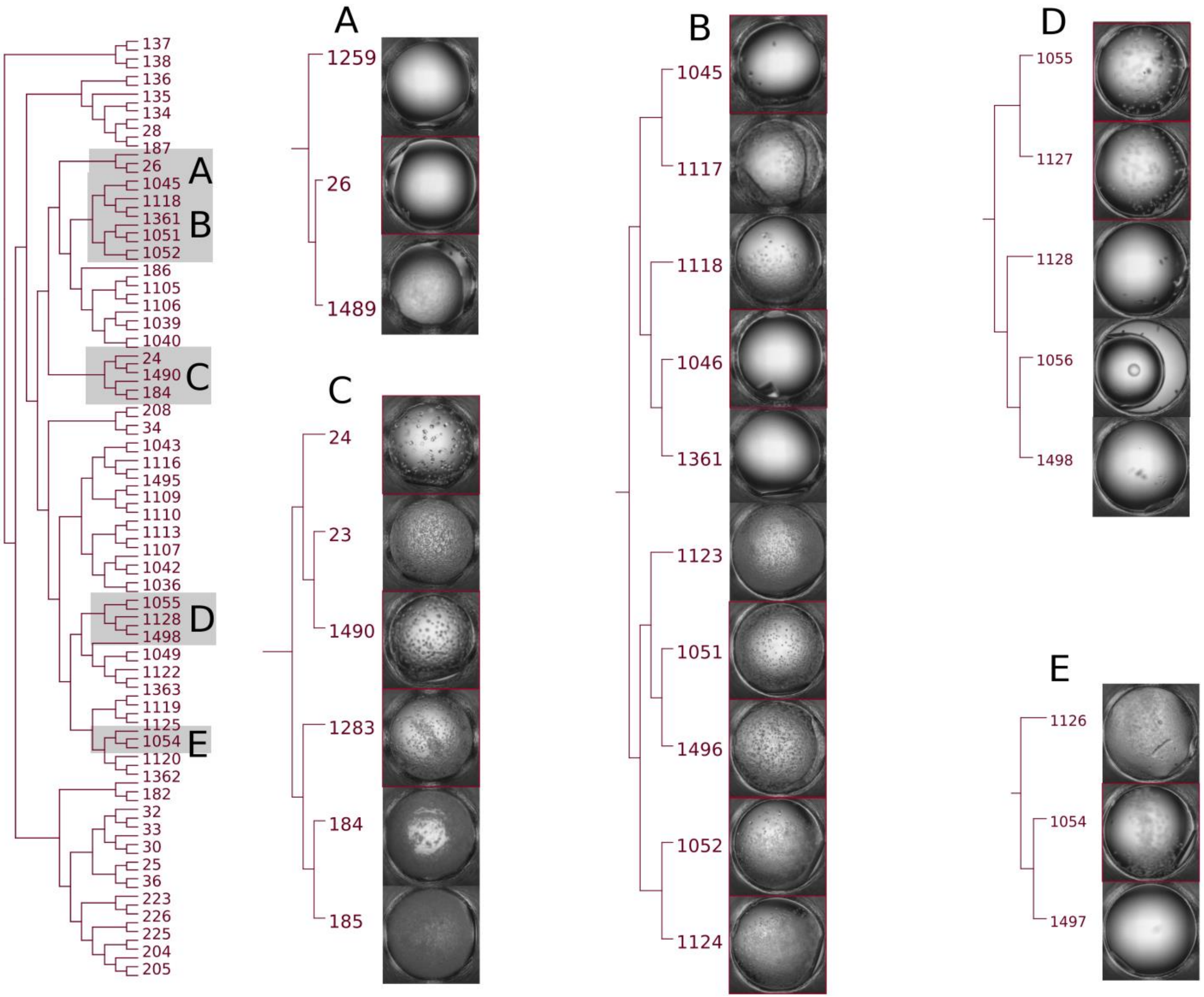


Cluster 20, PEG based, only 3 hits

Cluster	Total	Hits	% hits	Sodium %	Potassium %	Phosphate %
All cocktails						
	1536	70	4.5	47	24	16
All crystal						
	70	70	100	70	27	30
Clusters with crystals						
C13	108	19	17.6	73	72	100
C14	106	15	14.2	65	21	0
C12	57	11	19.3	16	2	0
C8	45	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C11	42	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C17	28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C20	965	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C15	19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C23	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
C4	12	1	8.3	83	25	0
C10	12	1	8.3	75	25	0

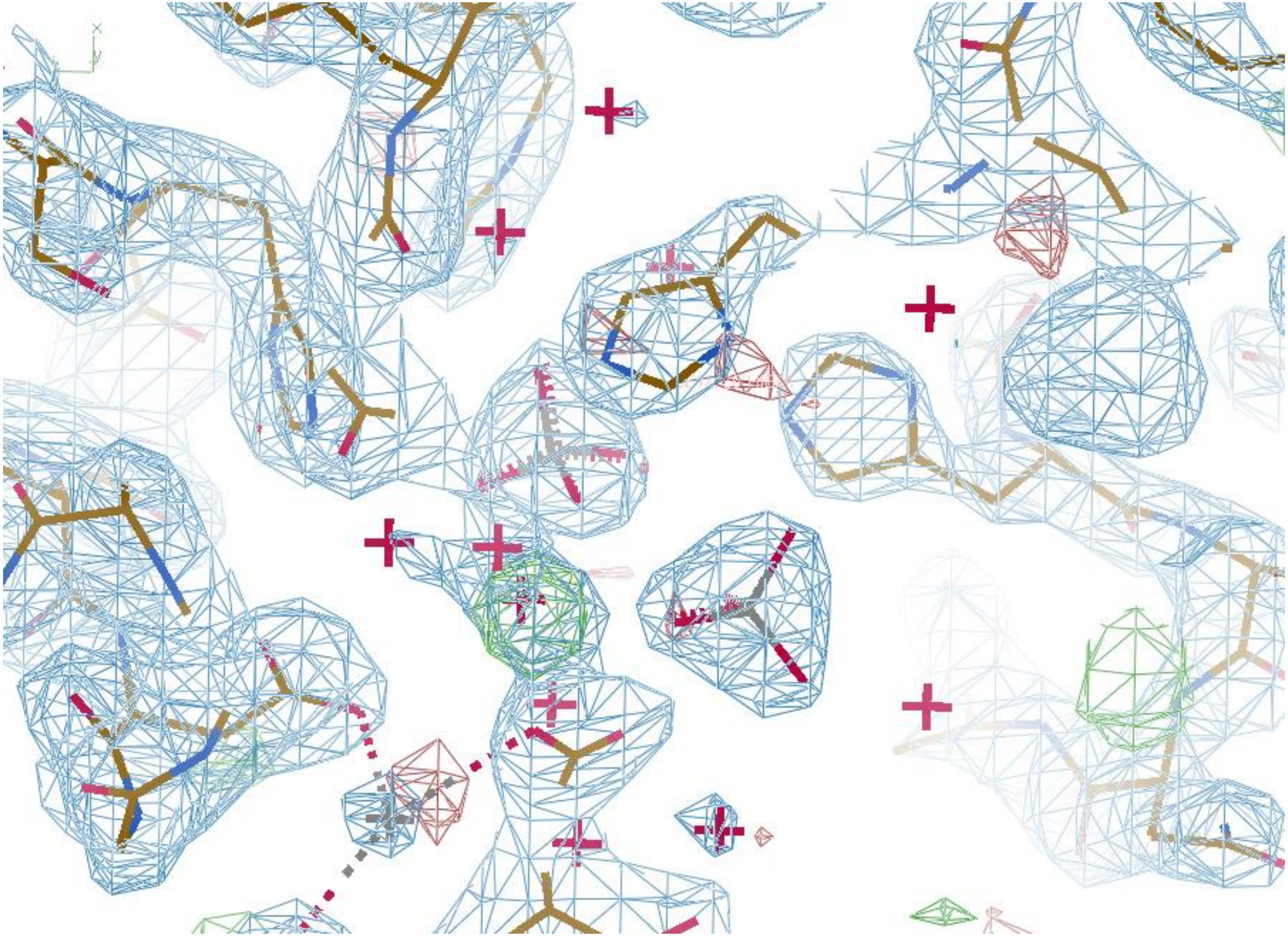
Cluster 13 proved interesting in that sodium is present in 73% of the conditions versus 47% for the 1536 condition screen overall, potassium is present in 72% of the conditions verses 24% overall and finally phosphate is present in 100% of the conditions versus 16% overall. This suggests a strong influence of these components in crystallization in this cluster.

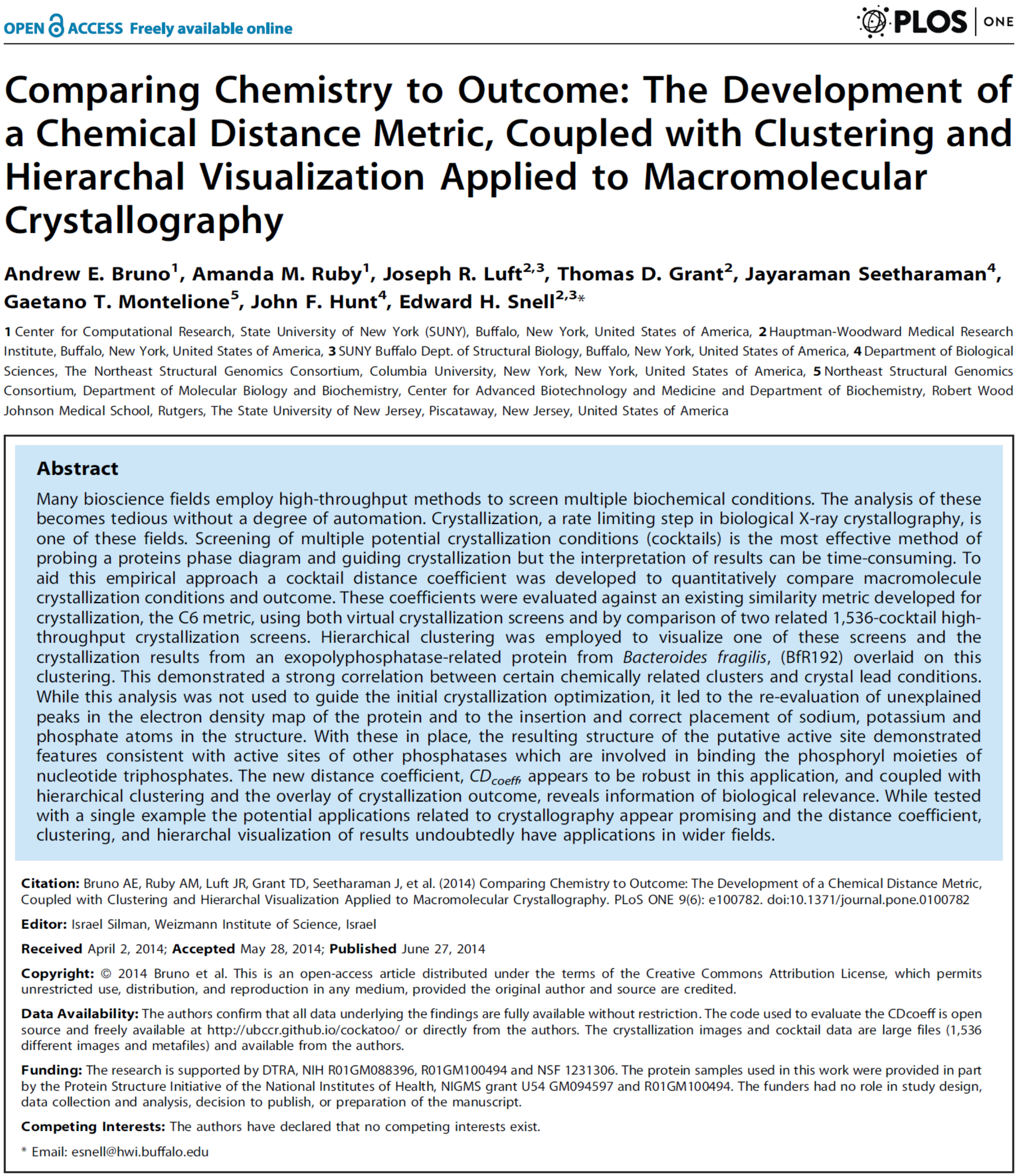
Zoom in on Cluster 13



Identifies
a pipette
error

Clustering samples the phase diagram





OPEN ACCESS Freely available online

PLOS | ONE

Comparing Chemistry to Outcome: The Development of a Chemical Distance Metric, Coupled with Clustering and Hierarchal Visualization Applied to Macromolecular Crystallography

Andrew E. Bruno[1], Amanda M. Ruby[1], Joseph R. Luft[2,3], Thomas D. Grant[2], Jayaraman Seetharaman[4], Gaetano T. Montelione[5], John F. Hunt[4], Edward H. Snell[2,3]*

1 Center for Computational Research, State University of New York (SUNY), Buffalo, New York, United States of America, 2 Hauptman-Woodward Medical Research Institute, Buffalo, New York, United States of America, 3 SUNY Buffalo Dept. of Structural Biology, Buffalo, New York, United States of America, 4 Department of Biological Sciences, The Northeast Structural Genomics Consortium, Columbia University, New York, New York, United States of America, 5 Northeast Structural Genomics Consortium, Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine and Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States of America

Abstract

Many bioscience fields employ high-throughput methods to screen multiple biochemical conditions. The analysis of these becomes tedious without a degree of automation. Crystallization, a rate limiting step in biological X-ray crystallography, is one of these fields. Screening of multiple potential crystallization conditions (cocktails) is the most effective method of probing a proteins phase diagram and guiding crystallization but the interpretation of results can be time-consuming. To aid this empirical approach a cocktail distance coefficient was developed to quantitatively compare macromolecule crystallization conditions and outcome. These coefficients were evaluated against an existing similarity metric developed for crystallization, the C6 metric, using both virtual crystallization screens and by comparison of two related 1,536-cocktail high-throughput crystallization screens. Hierarchical clustering was employed to visualize one of these screens and the crystallization results from an exopolyphosphatase-related protein from *Bacteroides fragilis*, (BfR192) overlaid on this clustering. This demonstrated a strong correlation between certain chemically related clusters and crystal lead conditions. While this analysis was not used to guide the initial crystallization optimization, it led to the re-evaluation of unexplained peaks in the electron density map of the protein and to the insertion and correct placement of sodium, potassium and phosphate atoms in the structure. With these in place, the resulting structure of the putative active site demonstrated features consistent with active sites of other phosphatases which are involved in binding the phosphoryl moieties of nucleotide triphosphates. The new distance coefficient, CD_{coeff}, appears to be robust in this application, and coupled with hierarchical clustering and the overlay of crystallization outcome, reveals information of biological relevance. While tested with a single example the potential applications related to crystallography appear promising and the distance coefficient, clustering, and hierarchal visualization of results undoubtedly have applications in wider fields.

Citation: Bruno AE, Ruby AM, Luft JR, Grant TD, Seetharaman J, et al. (2014) Comparing Chemistry to Outcome: The Development of a Chemical Distance Metric, Coupled with Clustering and Hierarchal Visualization Applied to Macromolecular Crystallography. PLoS ONE 9(6): e100782. doi:10.1371/journal.pone.0100782

Editor: Israel Silman, Weizmann Institute of Science, Israel

Received April 2, 2014; **Accepted** May 28, 2014; **Published** June 27, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The code used to evaluate the CDcoeff is open source and freely available at http://ubccr.github.io/cockatoo/ or directly from the authors. The crystallization images and cocktail data are large files (1,536 different images and metafiles) and available from the authors.

Funding: The research is supported by DTRA, NIH R01GM088396, R01GM100494 and NSF 1231306. The protein samples used in this work were provided in part by the Protein Structure Initiative of the National Institutes of Health, NIGMS grant U54 GM094597 and R01GM100494. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

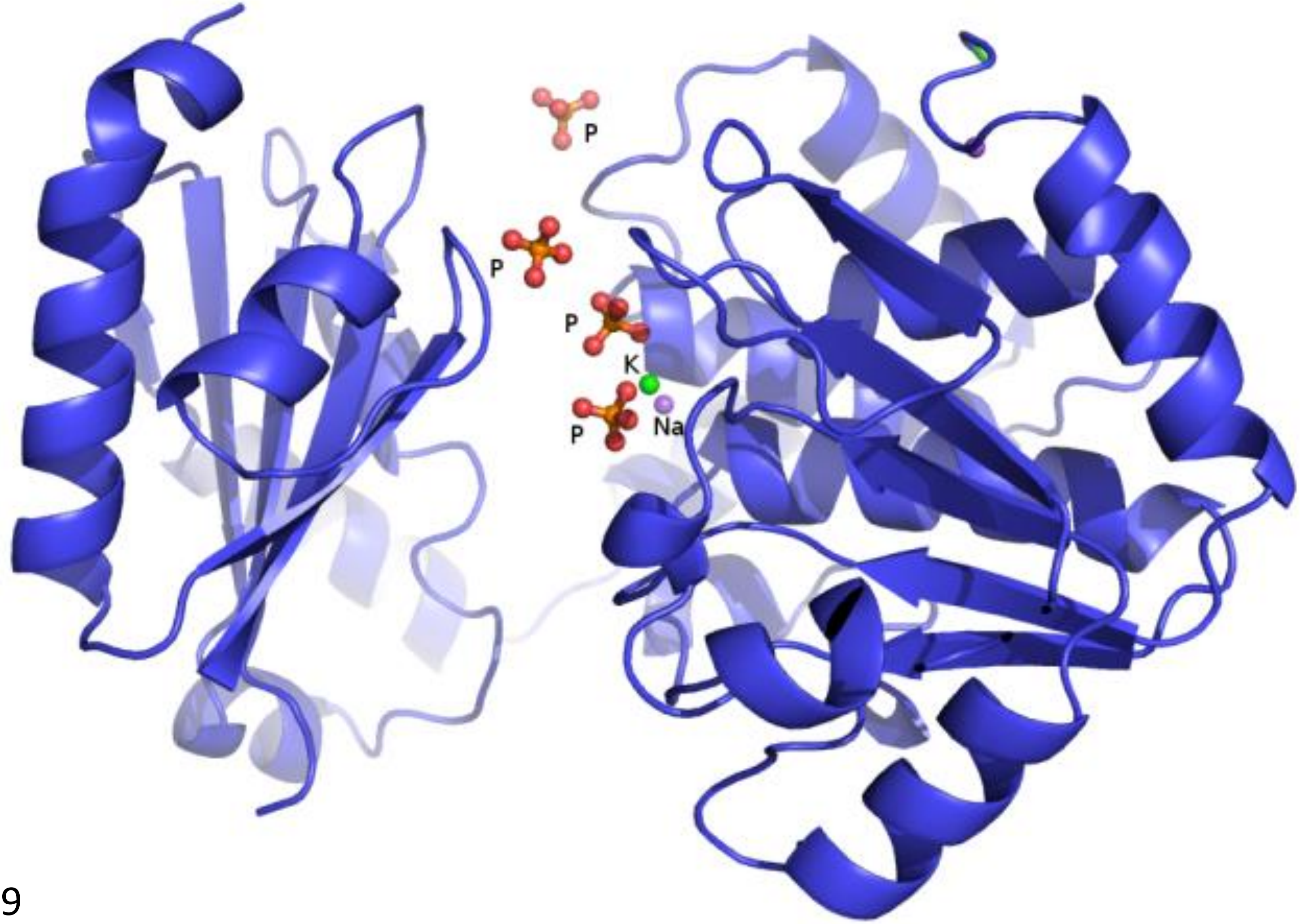
Competing Interests: The authors have declared that no competing interests exist.

* Email: esnell@hwi.buffalo.edu

Incorporating the correct ligands reduced the R and R_{free} from to 23.5% and 26.4% to 20.7% and 24.3% respectively.

The software is publically available and while it takes some time to run for each generation of screen it only has to be run once.

A Revised Structure Illustrating Mechanism



PDB 4PY9

Biological implication of the phosphates identified

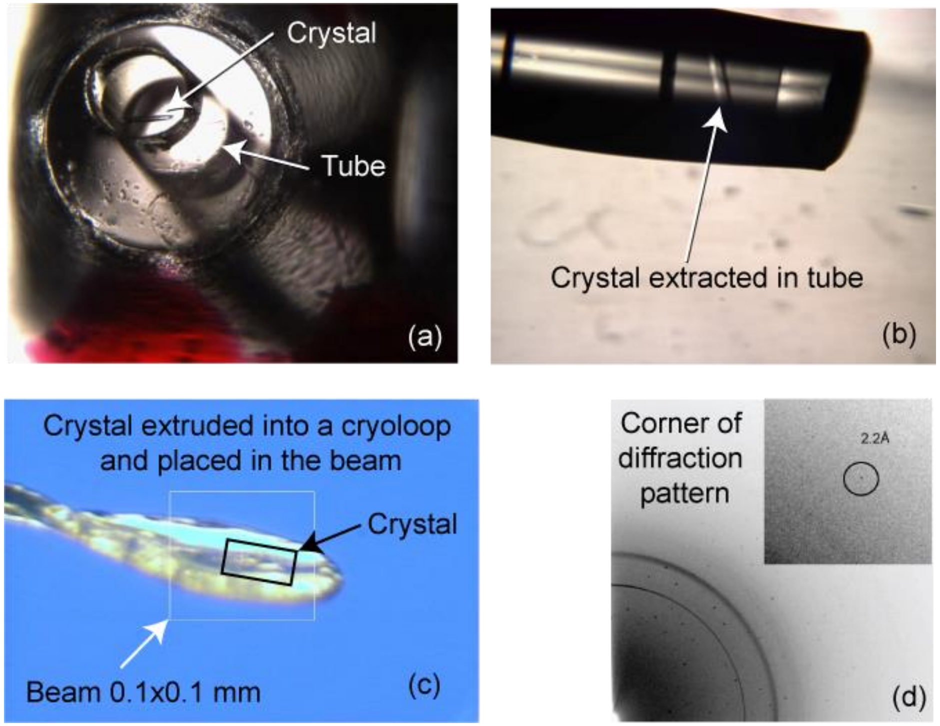
- The structure consists of two domains (N-terminal domain; residues 2 -212 and C-terminal domain residues 217-343) which are connected by a short loop – seen in the initial structure

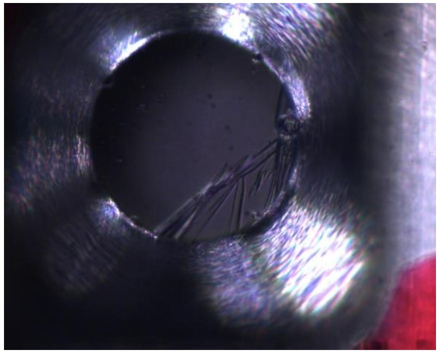
- The N-terminal domain contains the DHH (Asp224-His225-His226) motif and the C-terminal domain contains a glycine-rich (GGGH-Gly308-Gly309-Gly310-His311) phosphate binding motif – seen but not identified in the initial structure.
- Three of the phosphates (presumably carried with the protein), and the potassium and the sodium ion are bound in the cl
- The phosphate ions interact wi 7.
- The location of the phosphate
might anchor in this pocket.
- The putative active site has fea es
which are involved in binding t
- The possible roles of the active on
and polarization of the phosph
nucleophilic attack.
- The space around the phospha s.

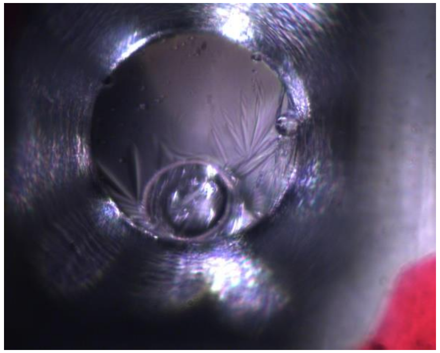
The important point here is not the details of the new information but that this information was obtained after the correct ligands were identified. Potential function and mechanism was revealed. While on could argue that these could have been identified earlier many examples in the PDB have ambiguous atoms – we have explored only a small sample of structures and seen problems in many of them.

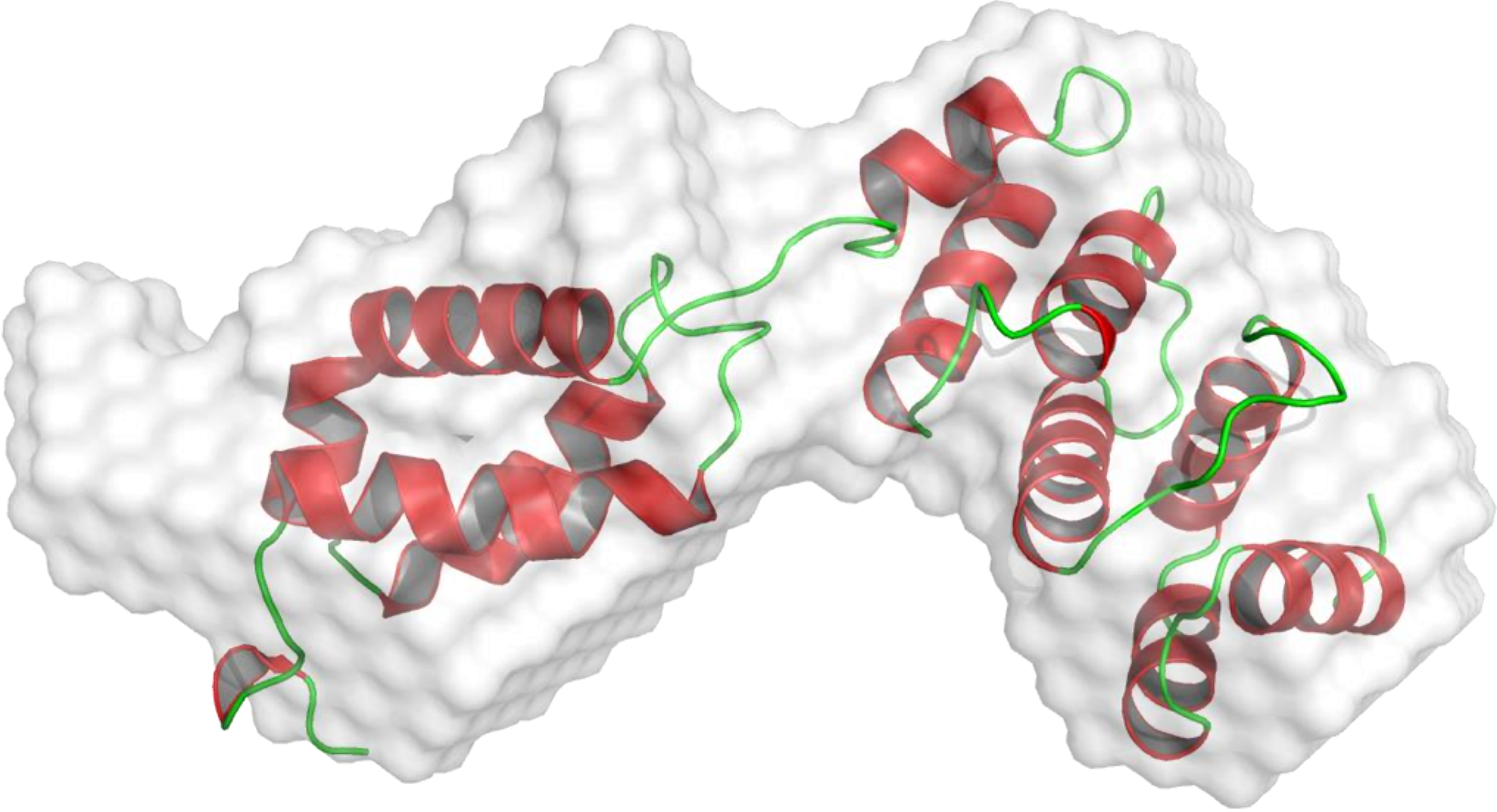
Going from crystals to diffraction properties

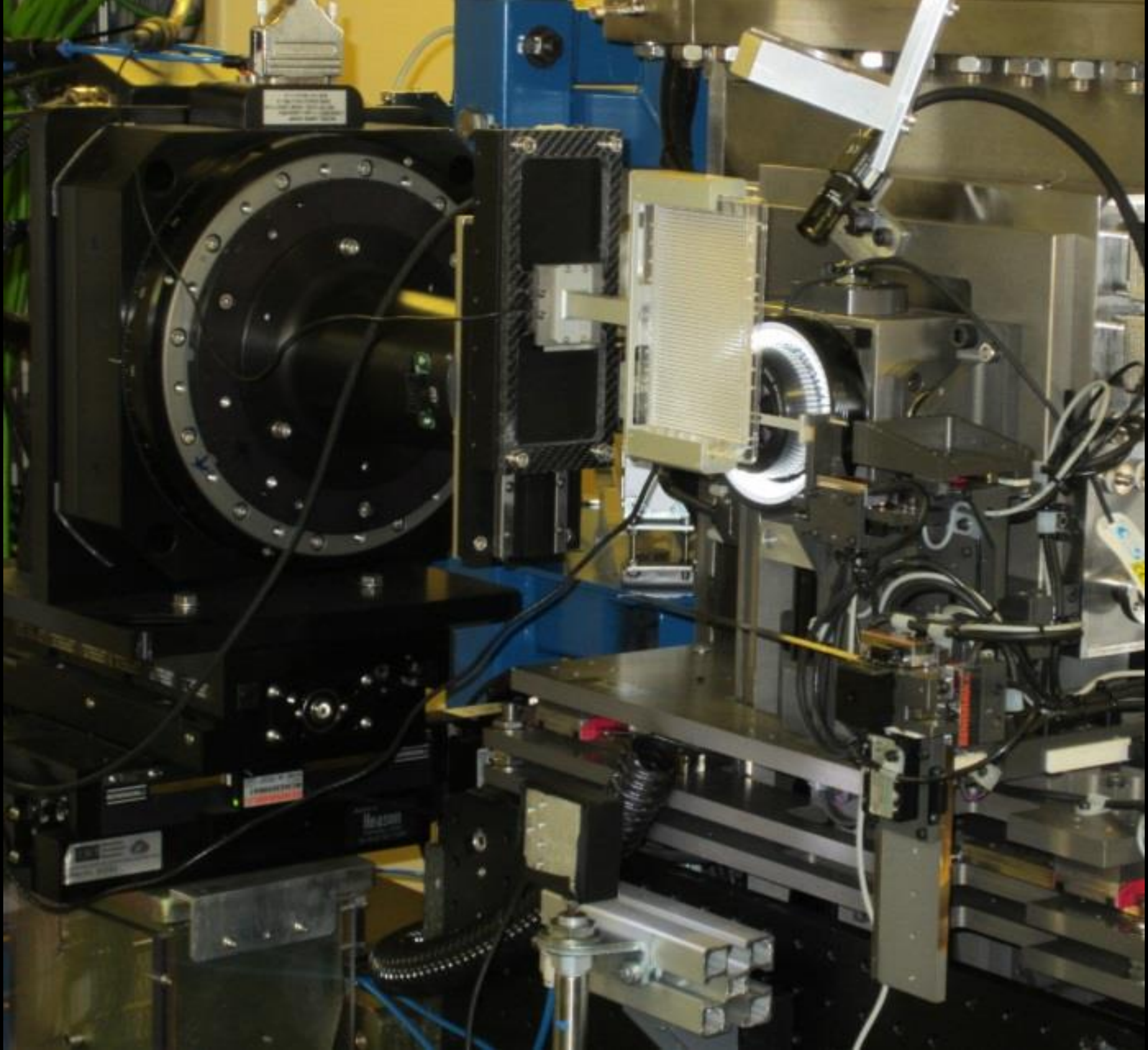
Does it diffract? Screening before the synchrotron

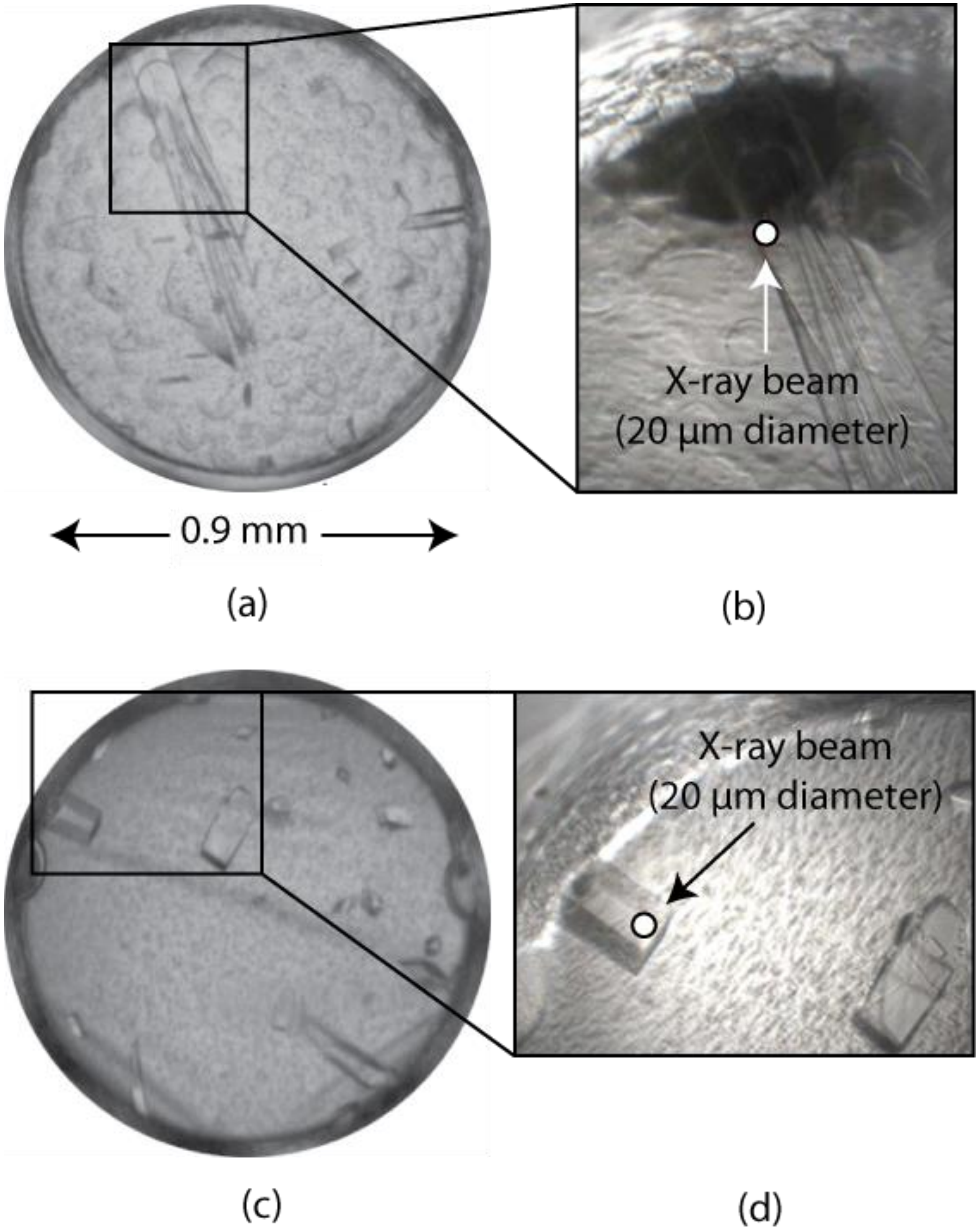

Crystal
Tube
(a)
Crystal extracted in tube
(b)
Crystal extruded into a cryoloop and placed in the beam
Crystal
Beam 0.1x0.1 mm
(c)
Corner of diffraction pattern
2.2A
(d)





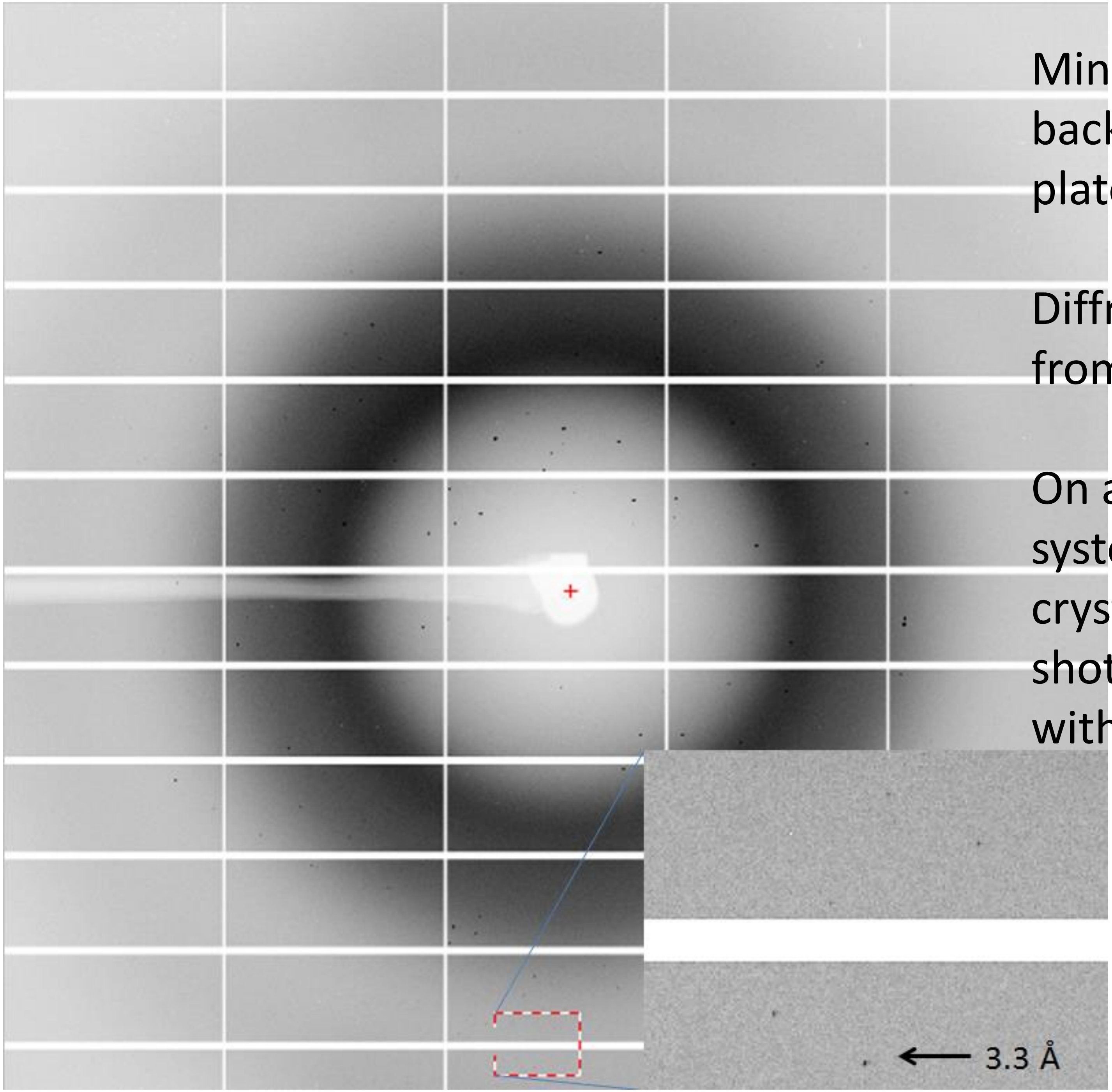






Crystal plates shipped by FedEx (Diamond and NSLS) and suitcase (Diamond)

Crystals remained in place and diffracted.



Minimal background from plate and oil

Diffraction to 2.3A from plate

On a microfocus system, multiple crystals can be shot individually within each well.

Not talked about.

Automated image analysis – been worked on for many years, often talked about, commercially very lucrative.

Tools for in-situ analysis – identifying crystals to X-ray characterize.

Analysis of multiple conditions to generally characterize the protein rather that where it crystallizes.

Other techniques to probe crystallization conditions.

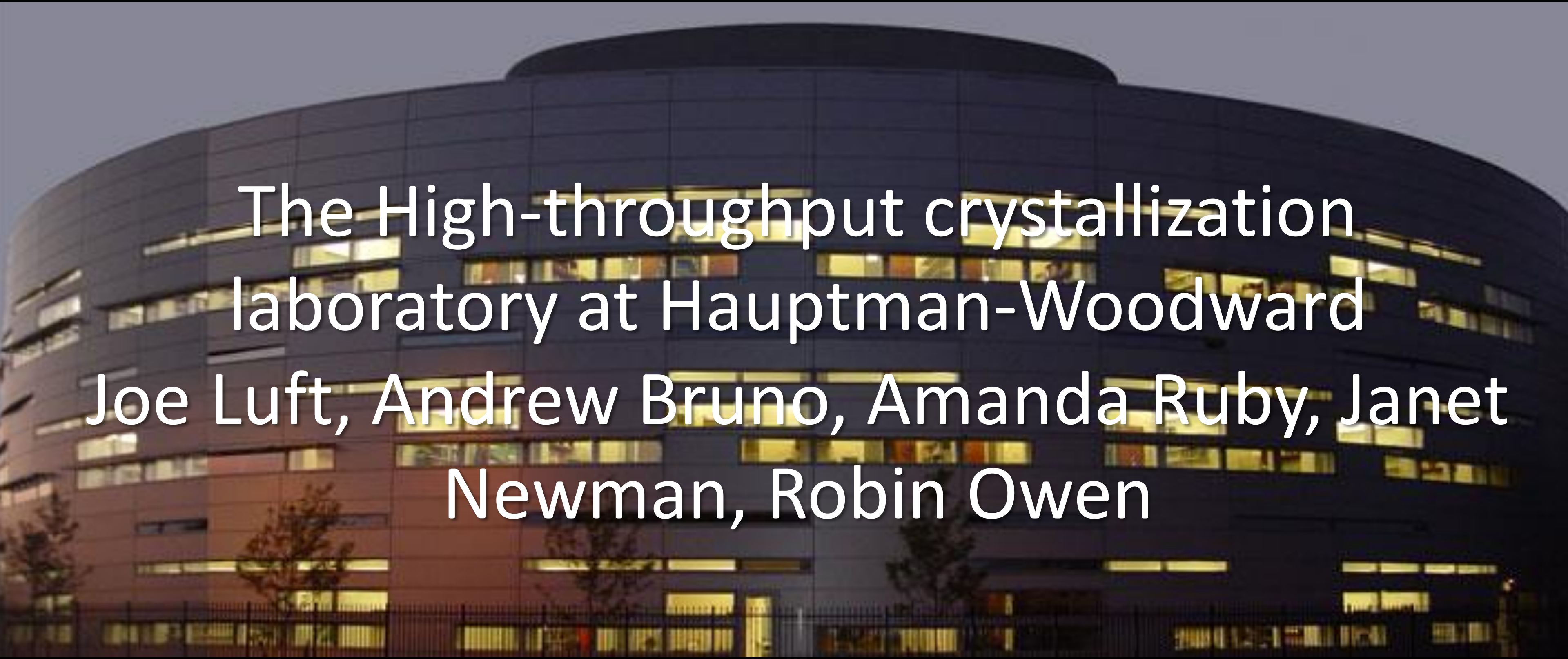
Got a protein?

Get a crystal™

500 µl protein at a ~10 mg/ml, setup against almost every Hampton screen and an incomplete factorial sampling of chemical space, visual images weekly over 6 weeks, SONICC and UV verification, remote data access. Automated optimization also available.

Details at: ***GetACrystal.org***

Acknowledgements



The High-throughput crystallization laboratory at Hauptman-Woodward
Joe Luft, Andrew Bruno, Amanda Ruby, Janet Newman, Robin Owen

Support and Funding: NIH, DoD and NSF

Thank you and questions?

↑

esnell@hwi.buffalo.edu