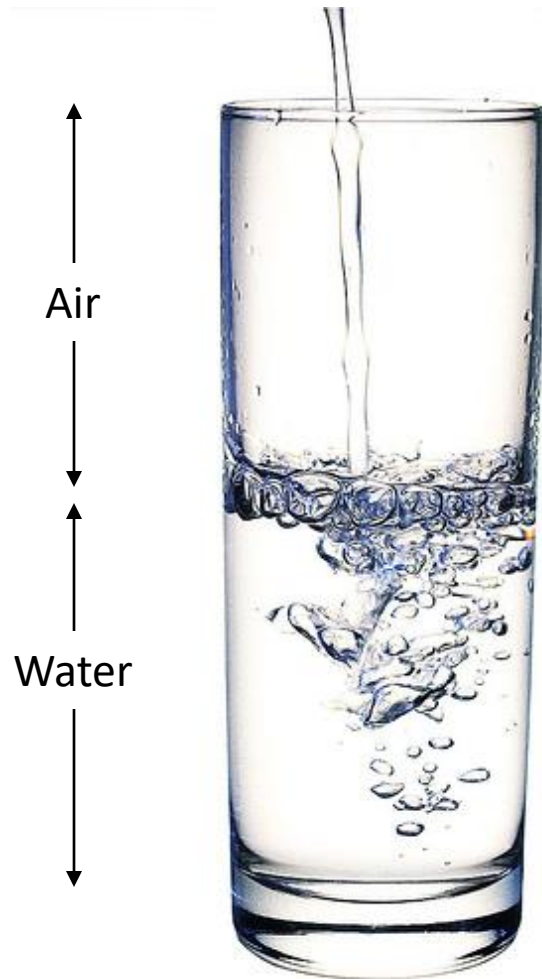


The truth is out there?
Collating, Visualizing and Using Information
Obtained from Crystallization Screening

Edward H. Snell,
Hauptman-Woodward
Medical Research Institute

Pessimists, Optimists, and Crystallographers



Consider a glass of water

Pessimist
(the glass is half empty)

Optimist
(the glass is half full)

Crystallographer
(the glass is completely full)



Only approximately 11% of the proteins we target for crystallography yield a crystallographic structure.

Acta Crystallographica Section F
Structural Biology
and Crystallization
Communications
ISSN 1744-3091

Janet Newman,^{a*} Evan E. Bolton,^b Jochen Müller-Dieckmann,^c Vincent J. Fazio,^a Travis Gallagher,^d David Lovell,^e Joseph R. Luft,^{f,g} Thomas S. Peat,^a David Ratcliffe,^e Roger A. Sayle,^h Edward H. Snell,^{f,g} Kerry Taylor,^e Pascal Vallotton,ⁱ Sameer Velanker^j and Frank von Delft^k

^aMaterials Science and Engineering, CSIRO, 343 Royal Parade, Parkville, VIC 3052, Australia, ^bNCBI, NLM, NIH, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, MD 20894, USA, ^cEMBL Hamburg Outstation c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany, ^dNational Institute for Standards and

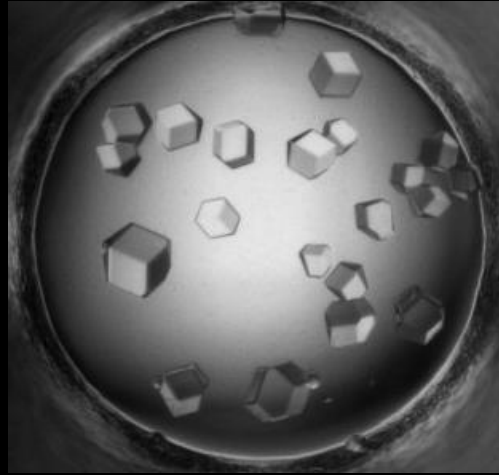
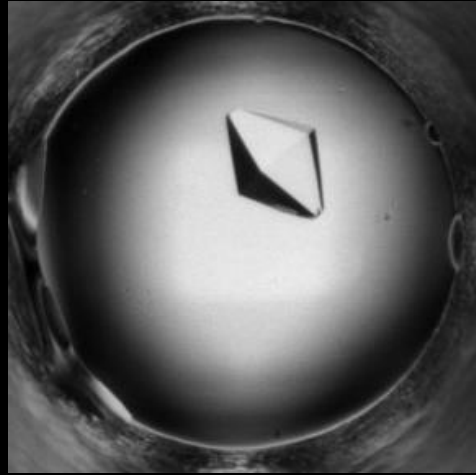
Acta Cryst. (2012). F68

On the need for an international effort to capture, share and use crystallization screening data

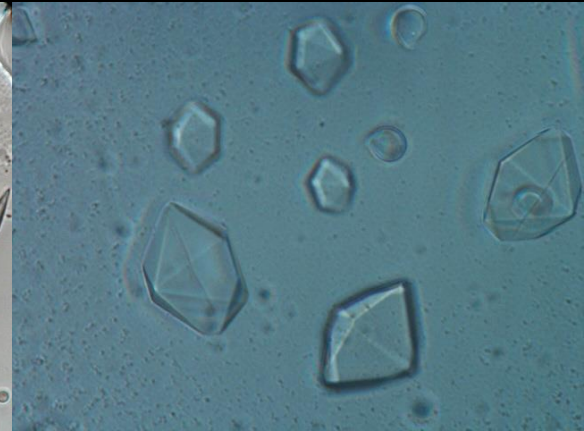
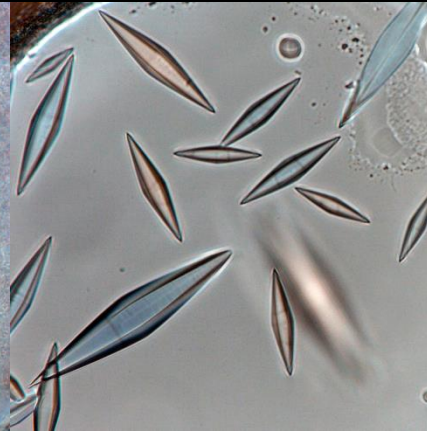
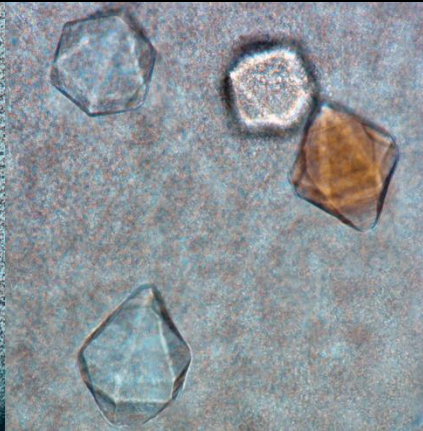
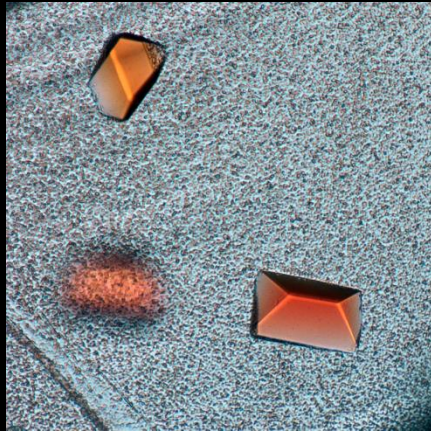
When crystallization screening is conducted many outcomes are observed but typically the only trial recorded in the literature is the condition that yielded the crystal(s) used for subsequent diffraction studies. The initial hit that was optimized and the results of all the other trials are lost. These missing results contain information that would be useful for an improved general understanding of crystallization. This paper provides a report of a crystallization data exchange (XDX) workshop organized by several international large-scale crystallization screening laboratories to discuss how this information may be captured and utilized. A group that administers a significant fraction of the world's crystallization screening results was convened, together with chemical and structural data informaticians and computational scientists who specialize in creating and analysing large disparate data sets. The development of a crystallization ontology for the crystallization community was proposed. This paper (by the attendees of the workshop) provides the thoughts and rationale leading to this conclusion. This is brought to the attention of the wider audience of crystallographers so that they are aware of these early efforts and can contribute to the process going forward.

At least 99.8% of crystallization experiments produce an outcome other than crystallization.





Crystallography Requires Crystals



No crystal ...

No crystallography

No crystallographer

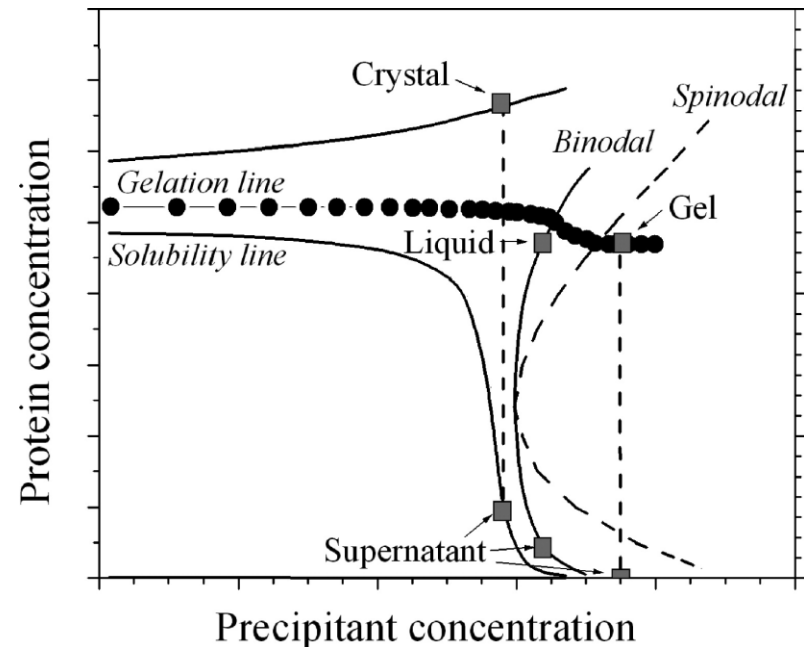
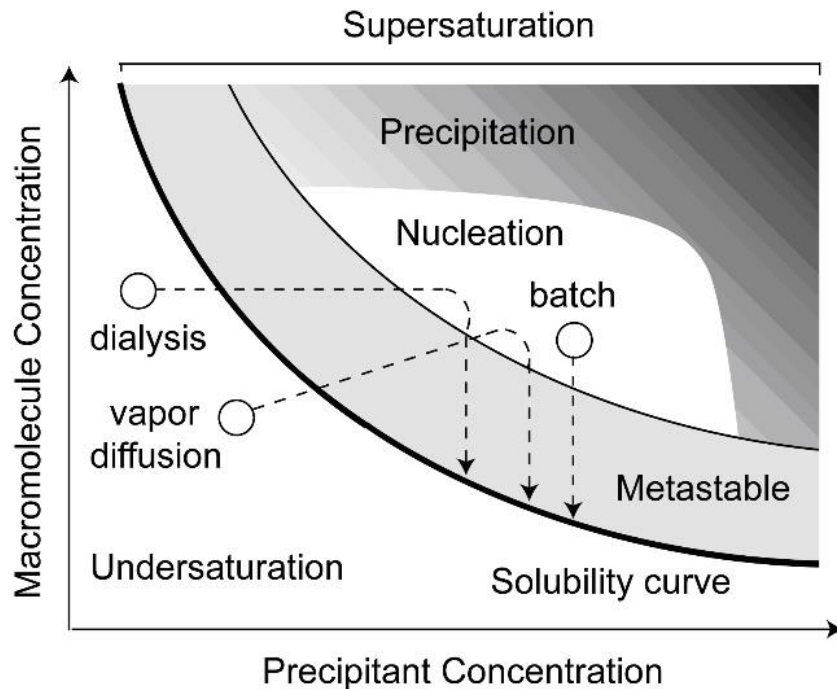
(Change careers to cryo electron microscopy
– a complementary technique)

Fantasy

Crystallize
Now

Crystallization theory is well established

The fundamental theory around the protein phase diagram is well understood.
The most efficient method to probe it is chemically screening different conditions



(Dumetz et al., 2009)

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

Since February of 2000 the High Throughput Crystallization Center has been screening potential crystallization conditions as a high-throughput service

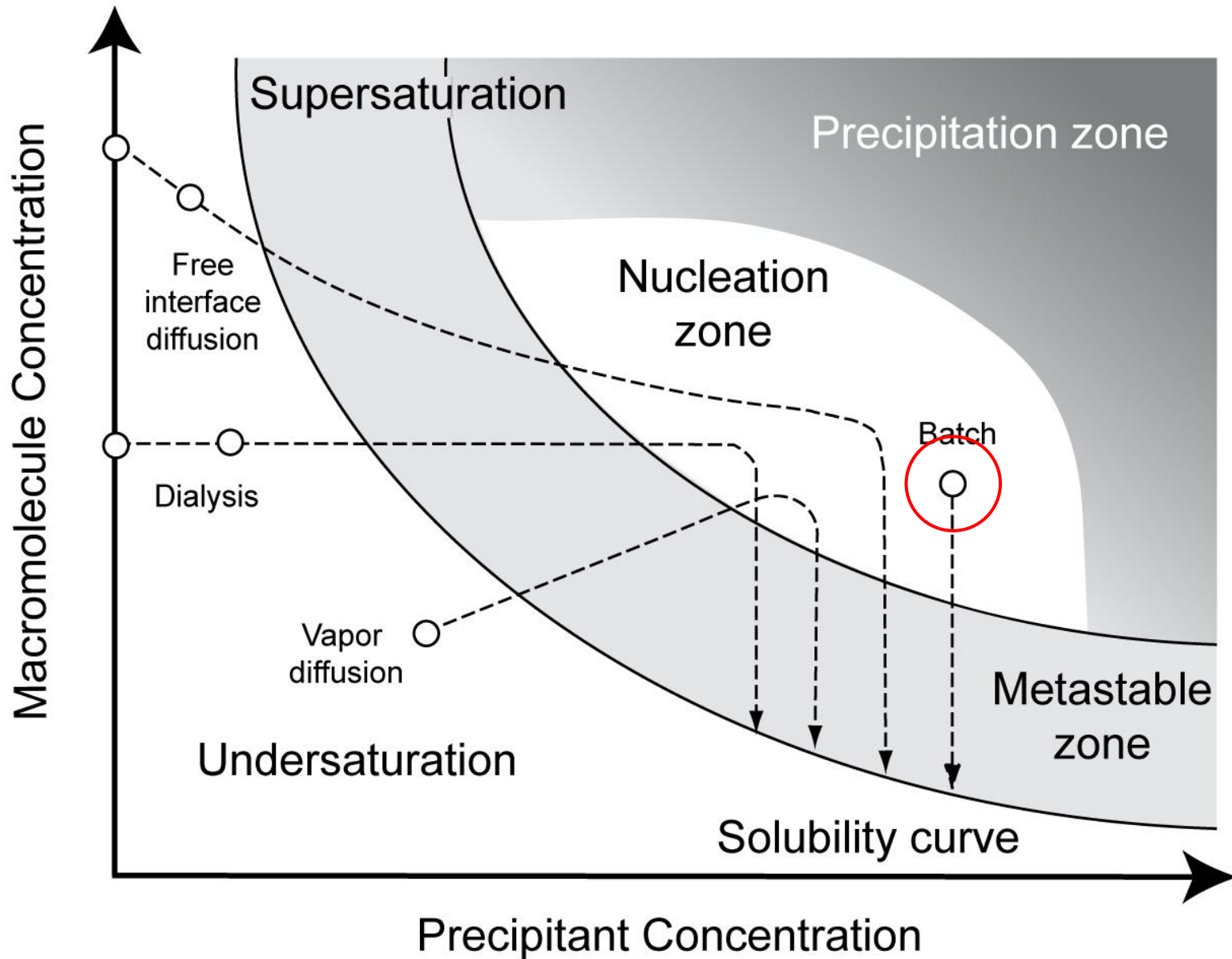
The HTS lab screens samples against three types of cocktails:

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.

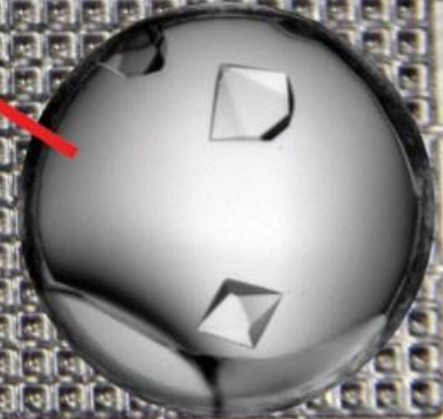
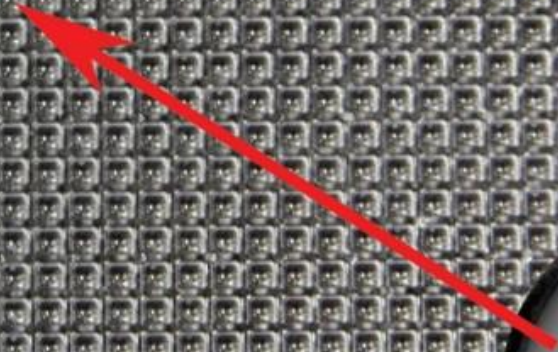
The HTSlab has investigated the crystallization properties of over 19,000 individual proteins archiving approximately 180 million images of crystallization experiments.

All data and in many cases, dead volume recovered samples are available

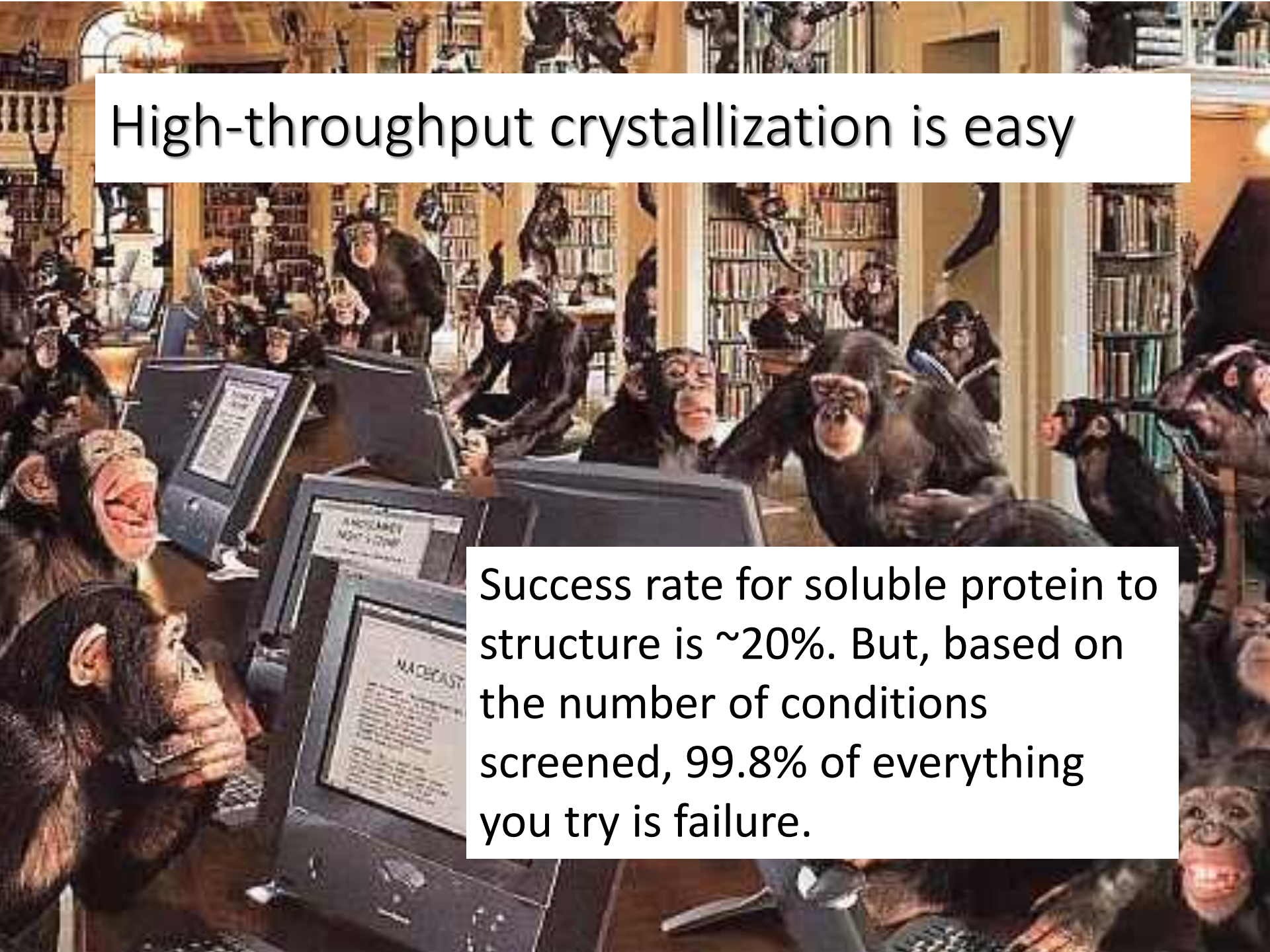
Simplified phase diagram for crystallization



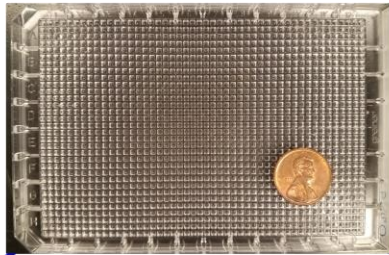
Minimize sample volume



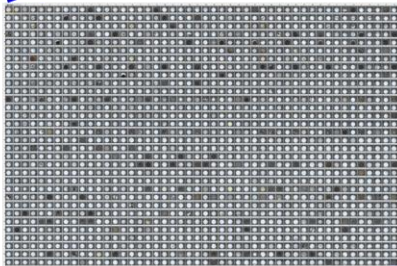
High-throughput crystallization is easy

A large group of chimpanzees is depicted in a library setting, each sitting at a computer workstation. The chimpanzees are shown in various states of activity, some looking at the screens, some typing, and some appearing to be in conversation. The background is filled with bookshelves and other library elements, creating a sense of a busy, high-throughput environment. The overall scene is a humorous representation of high-throughput screening in a laboratory context.

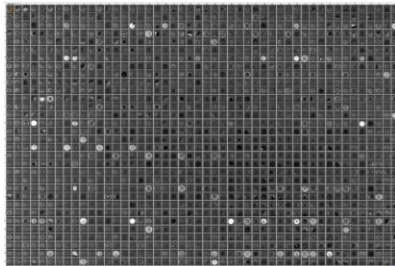
Success rate for soluble protein to structure is $\sim 20\%$. But, based on the number of conditions screened, 99.8% of everything you try is failure.



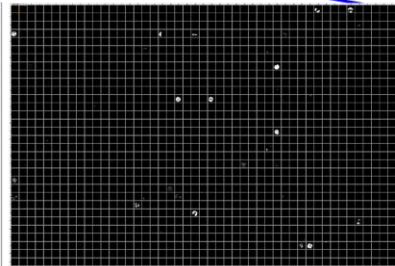
1,536 wells each imaged
at 9 timepoints



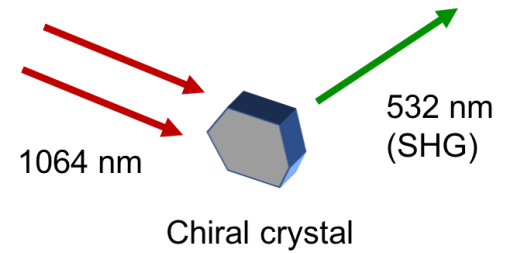
7x visible images



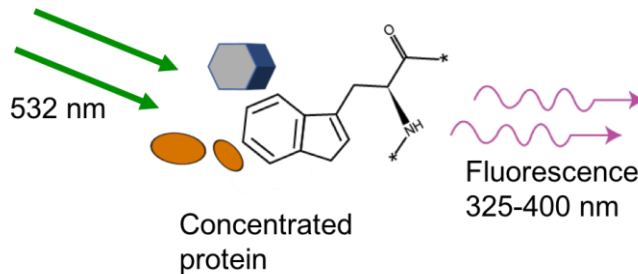
UV-TPEF images



SHG images



Chiral crystal



Concentrated
protein

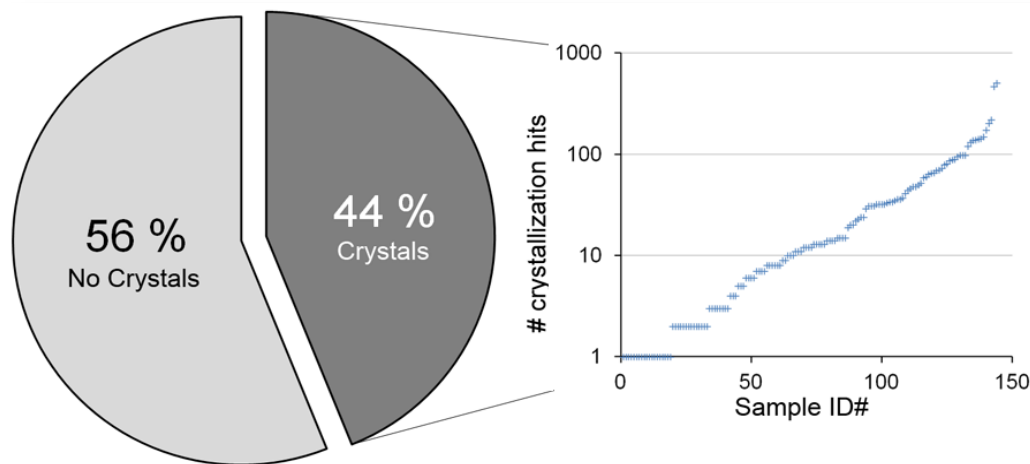
UV-TPEF and SHG images at 4 weeks

13,824 images generated for
each experimental screen

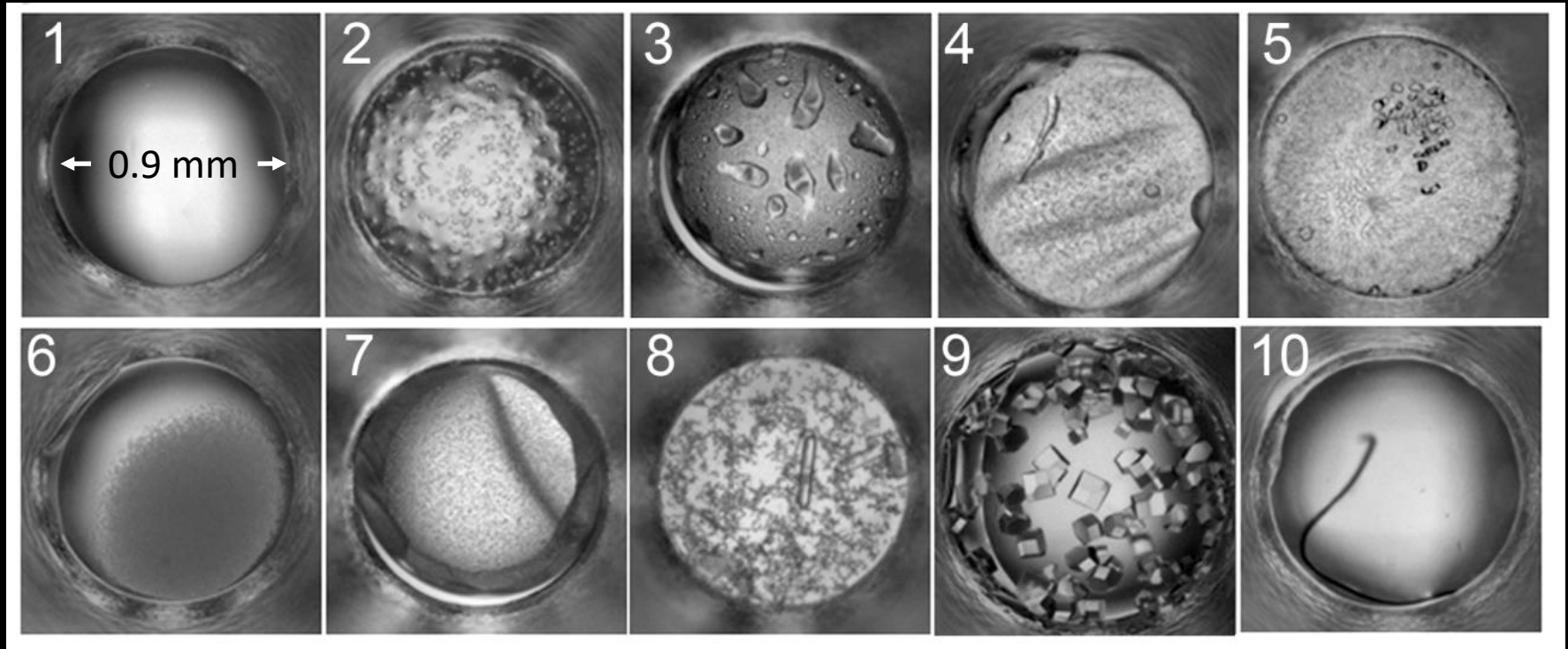


Good at finding crystals

Based on data from NESG from 144 out of 328 targets with one or more verified hits



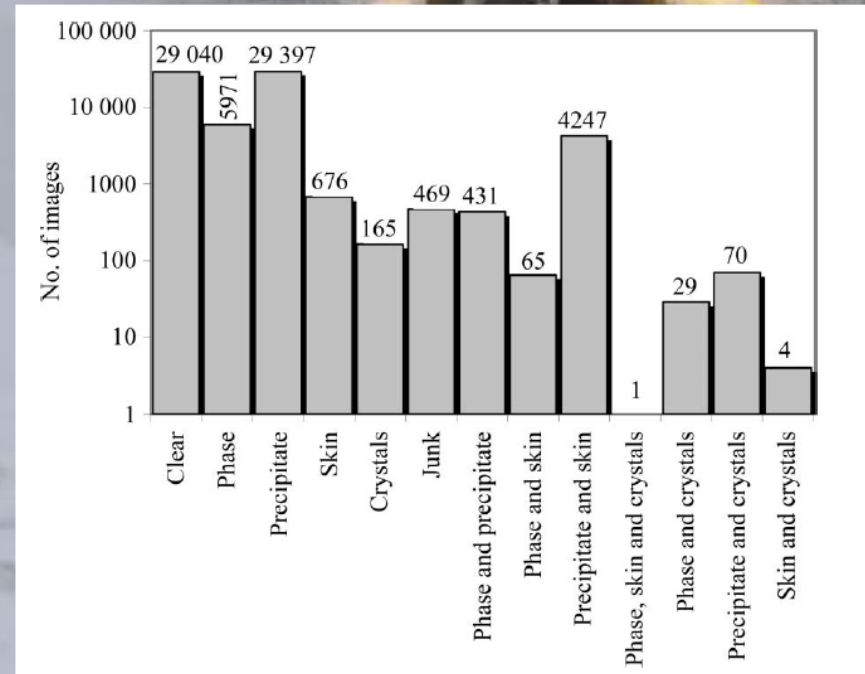
Many outcomes



Can we automate the classification of outcomes?

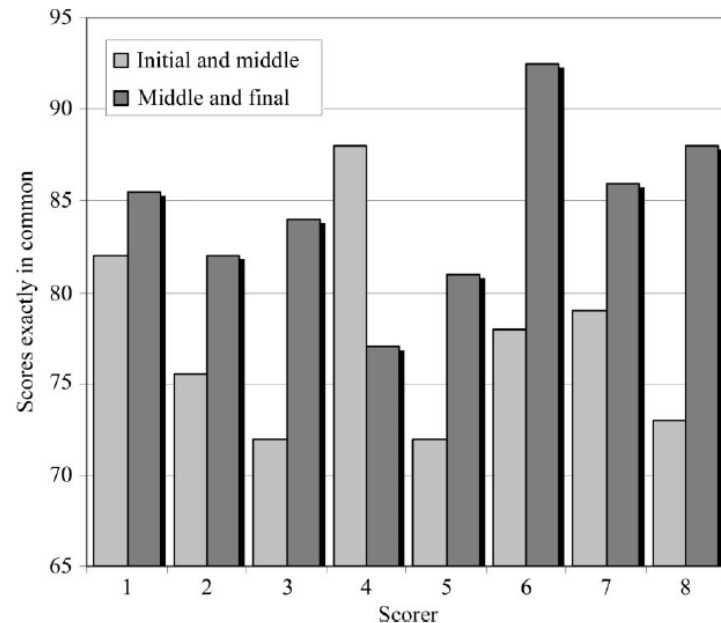
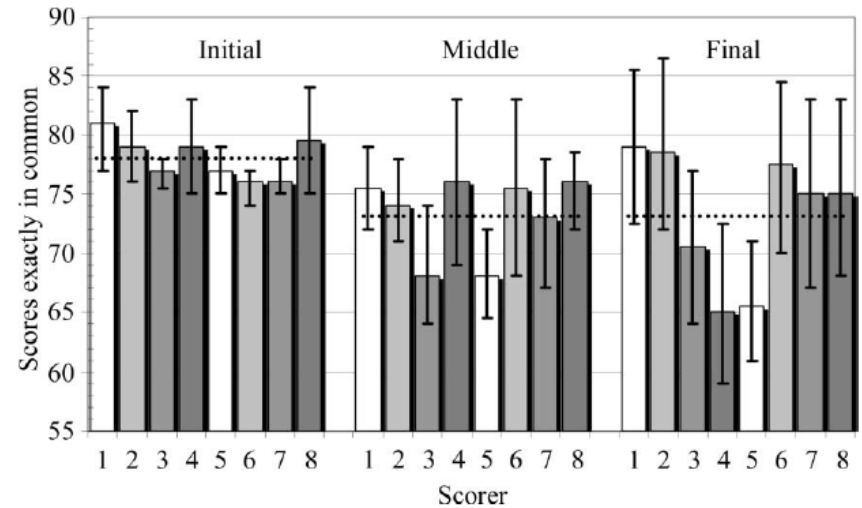
Training set:

- Set 1: ~150,000 images (96 proteins) classified by eight people in a set of categories (crystal, clear, precipitate, skin, phase separation, other, and combinations) with a minimum of three people looking at each image.
- Set 2: ~420,000 images (269 proteins) classified as crystal or no crystal by eight people, a single person looking at each image.



Testing the humans with the training set.

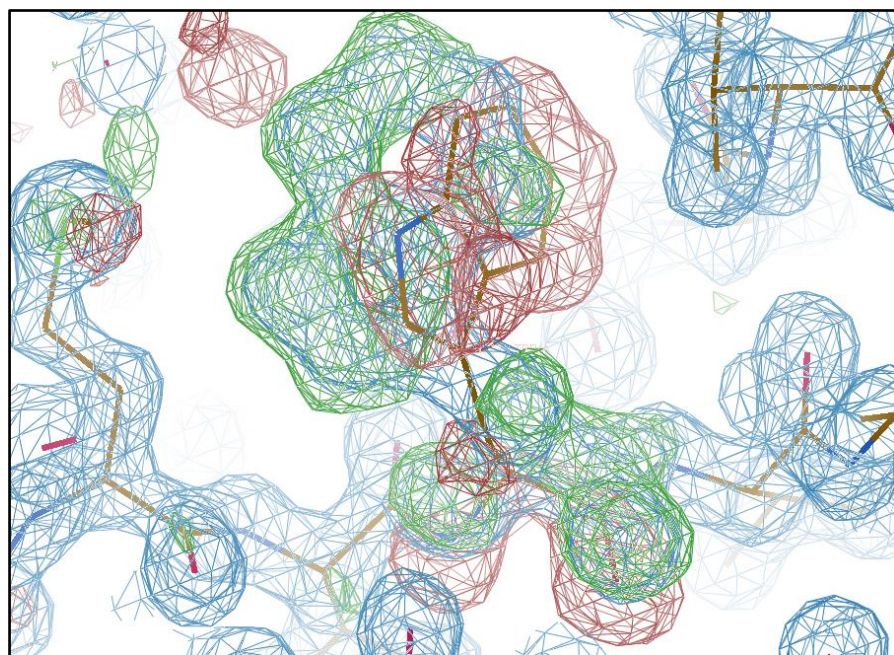
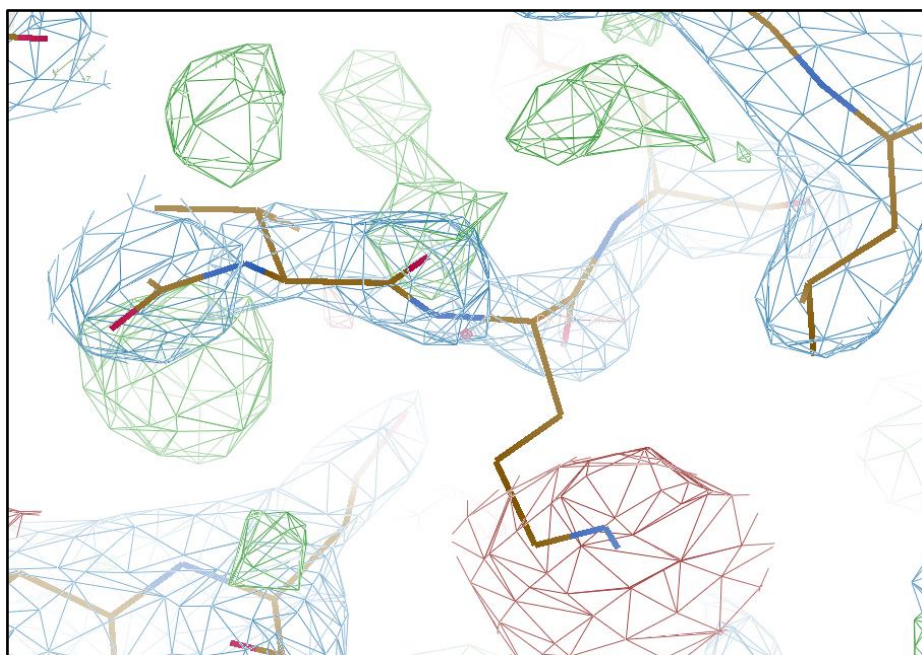
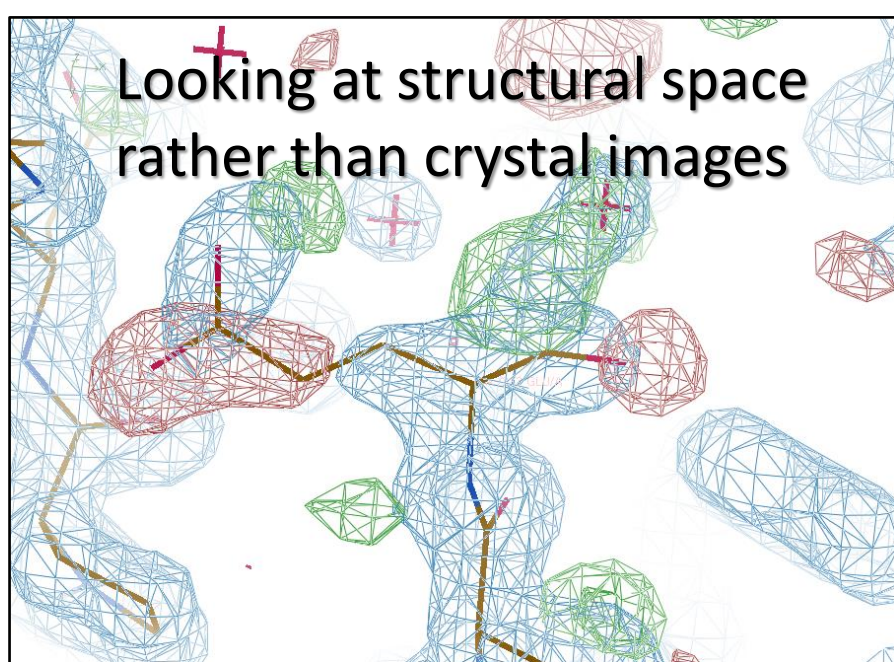
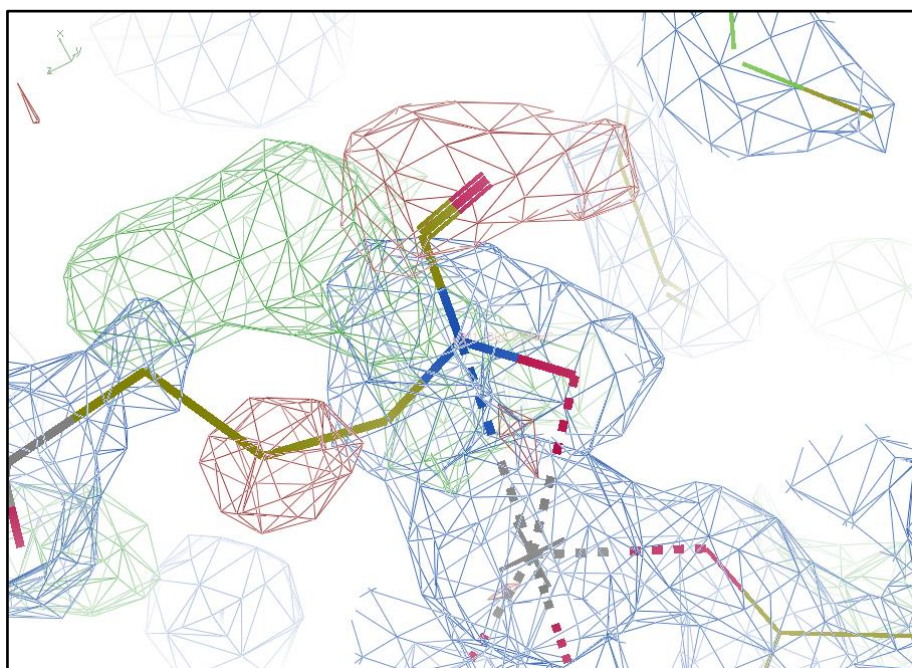
- Extra layer of experiment added without human classifier knowledge.
 - The same set of images was included at the start, half way through, and at the end of the study.
 - All human classifiers were given this data.
 - Agreement between same classifier and same image was at maximum 80% and often a lot less.
 - **Humans cannot agree even with themselves.**
 - Training sets were trimmed so that training images had an agreement with at least two classifiers for training set 1.



Using the training set for classification of other images

- Worldwide community grid with random tree classifier used, a minimum of five year run time to process all available images.
- Very good results within the training and test set, but less so outside of that. Results were not useable.
- Training set also provided to other groups, several still working on automated analysis. Test set kept for validation.
- Several recoding efforts succeeded in taking the five year run time down to one month.
- Strategy was a comprehensive feature extractor followed by supervised learning. Classifier was unique to our images.
- Results in a word, **unsuccessful**.





Big data in X-ray crystallization is old data ...

The screenshot shows the xtuition website interface. The browser address bar displays the URL: `xtuition.org/well/1305?classified=true&crystals=true&class=crystal`. The page header includes the xtuition logo and navigation links for Search, Screens, Compounds, Samples, API, and About. Below the header, a breadcrumb trail reads "Samples / X000009233 / Wells". A filter bar contains buttons for "Skin", "Phase", "Precipitant", "Clear", "Crystal", and "Verified" (which is selected). The main content area displays a grid of six X-ray diffraction images, each with associated data:

Image ID	Read Date	Cocktail	Crystal Status
X0000092331536	2007-10-12 15:23:00	7_C1536	Crystal
X0000092331505	2007-10-12 15:23:00	7_C0569	Crystal
X0000092331525	2007-10-12 15:23:00	7_C0574	Crystal
X0000092331504			
X0000092331509			
X0000092331533			



Old data ported
functionality. S

Switch gears and look at the chemistry instead of pictures

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.

This information is stored in single integer with bits set depending on the properties.

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



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



The Dissimilarity Measure Over the Whole Screen

Adapted from Newman J, Fazio VJ, Lawson B, Peat TS (2010) *Crystal Growth & Design* 10: 2785-2792

Aspects of the screen design
are clearly seen

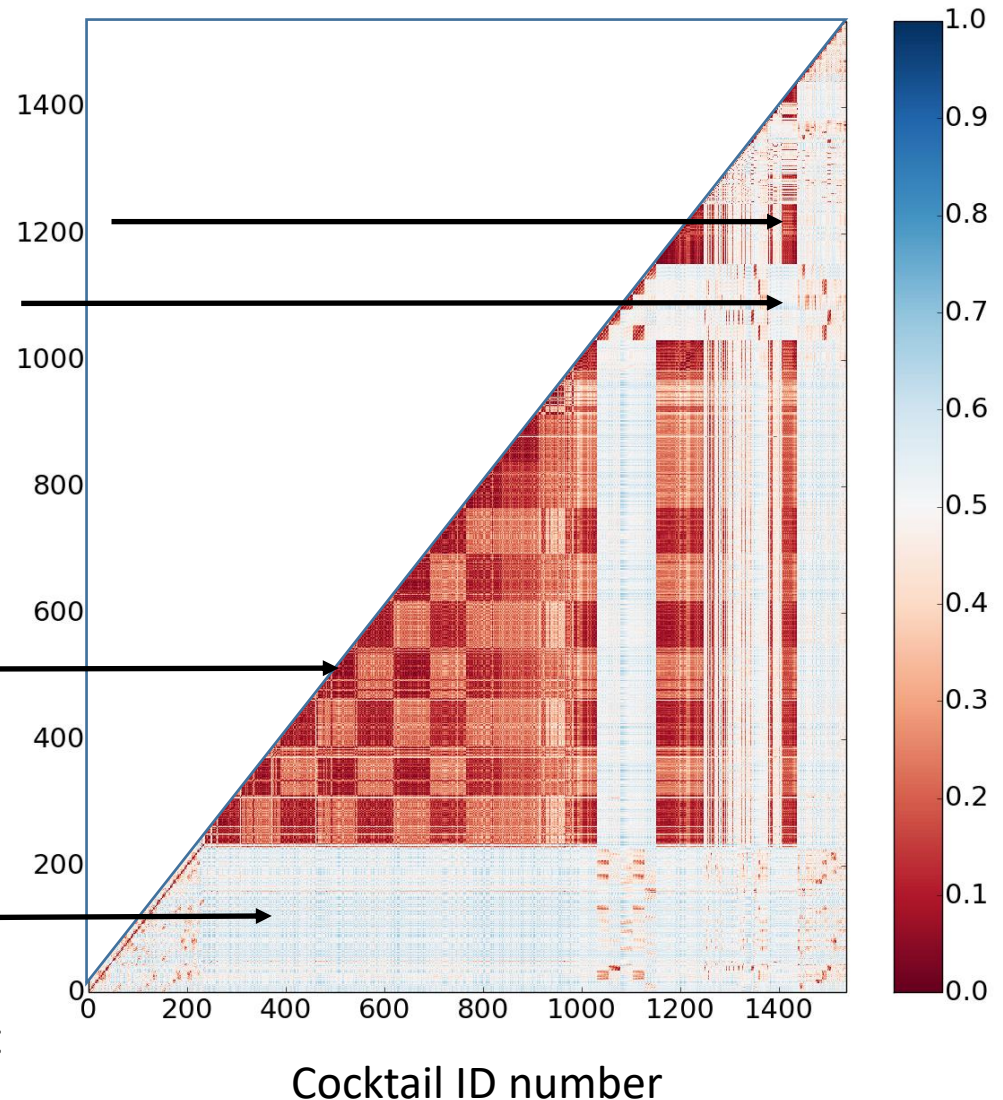
Hampton Research PEG/Ion screen

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

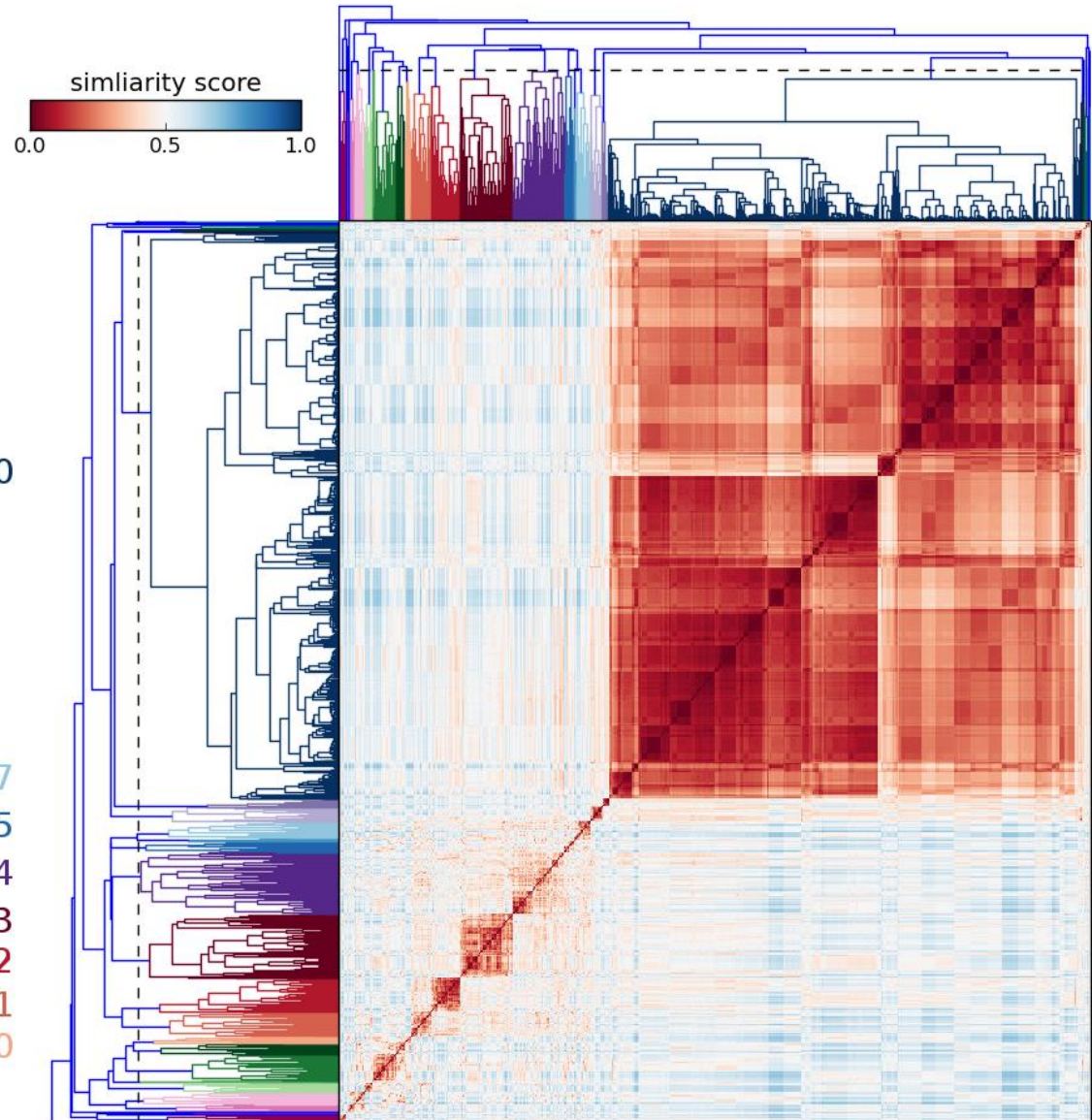


C20

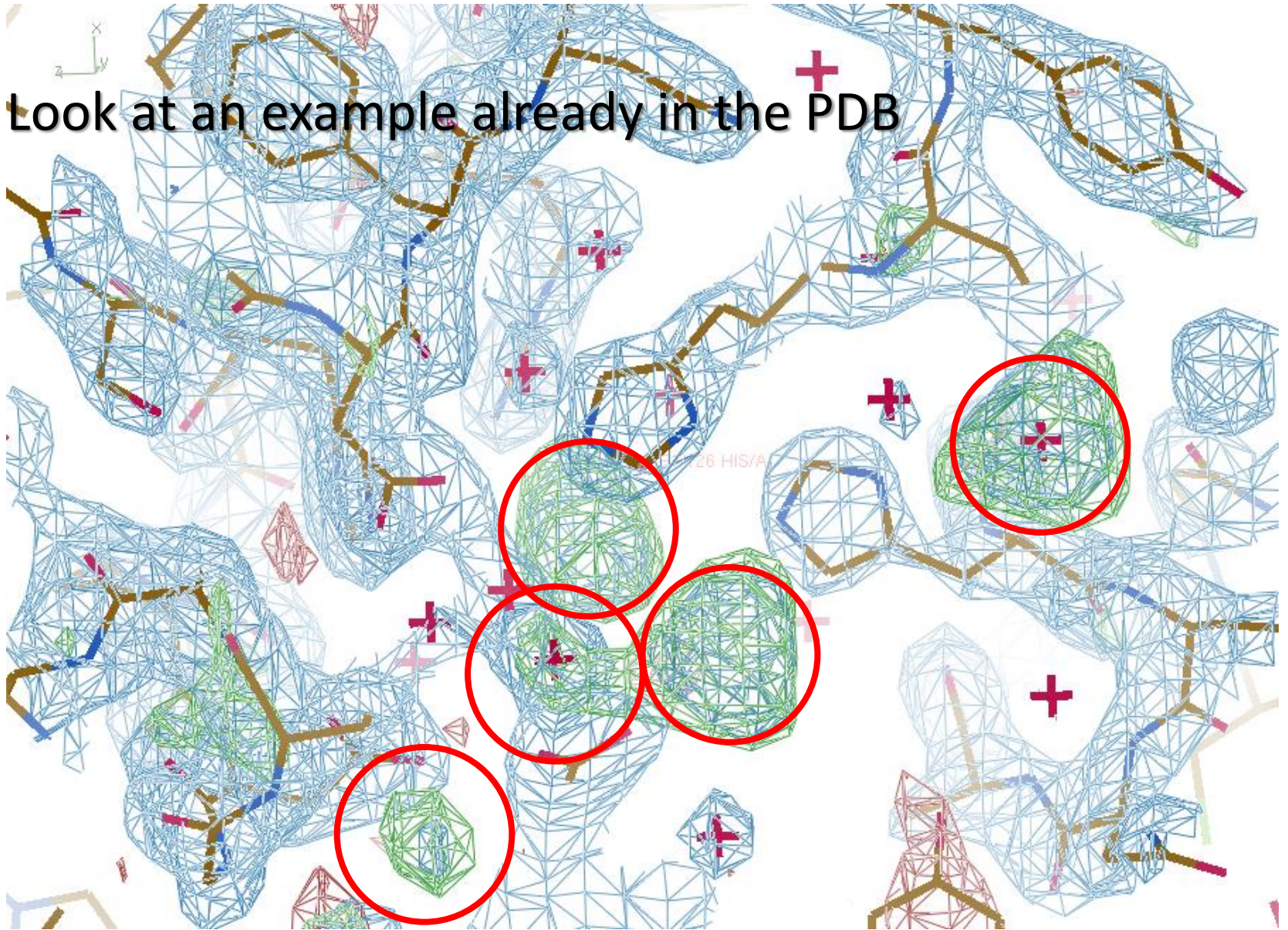
Salts with different anions and cations



C17
C15
C14
C13
C12
C11
C10
C8



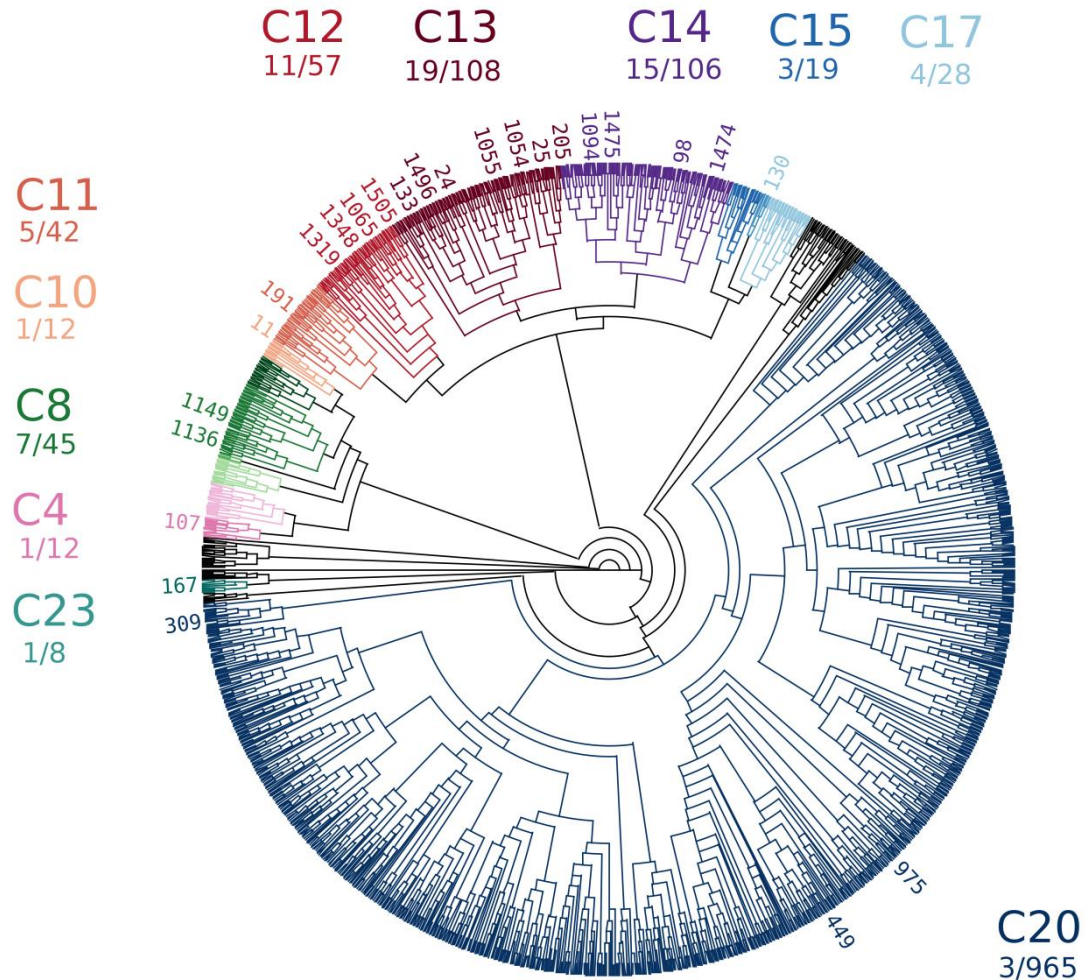
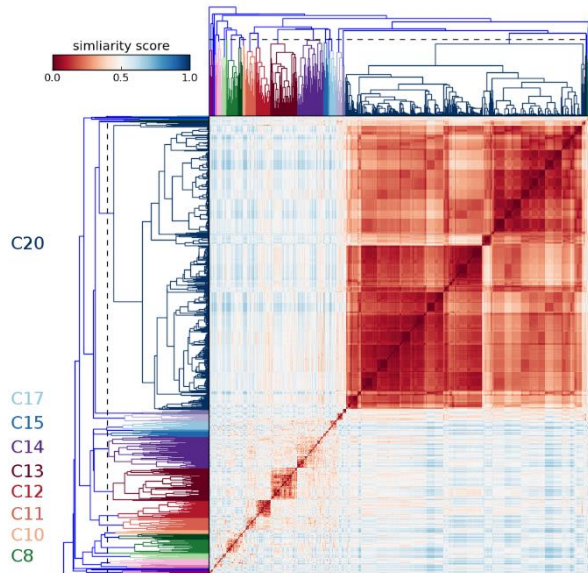
Look at an example already in the PDB



PDB ID 3DMA as deposited in the PDB

New information representations

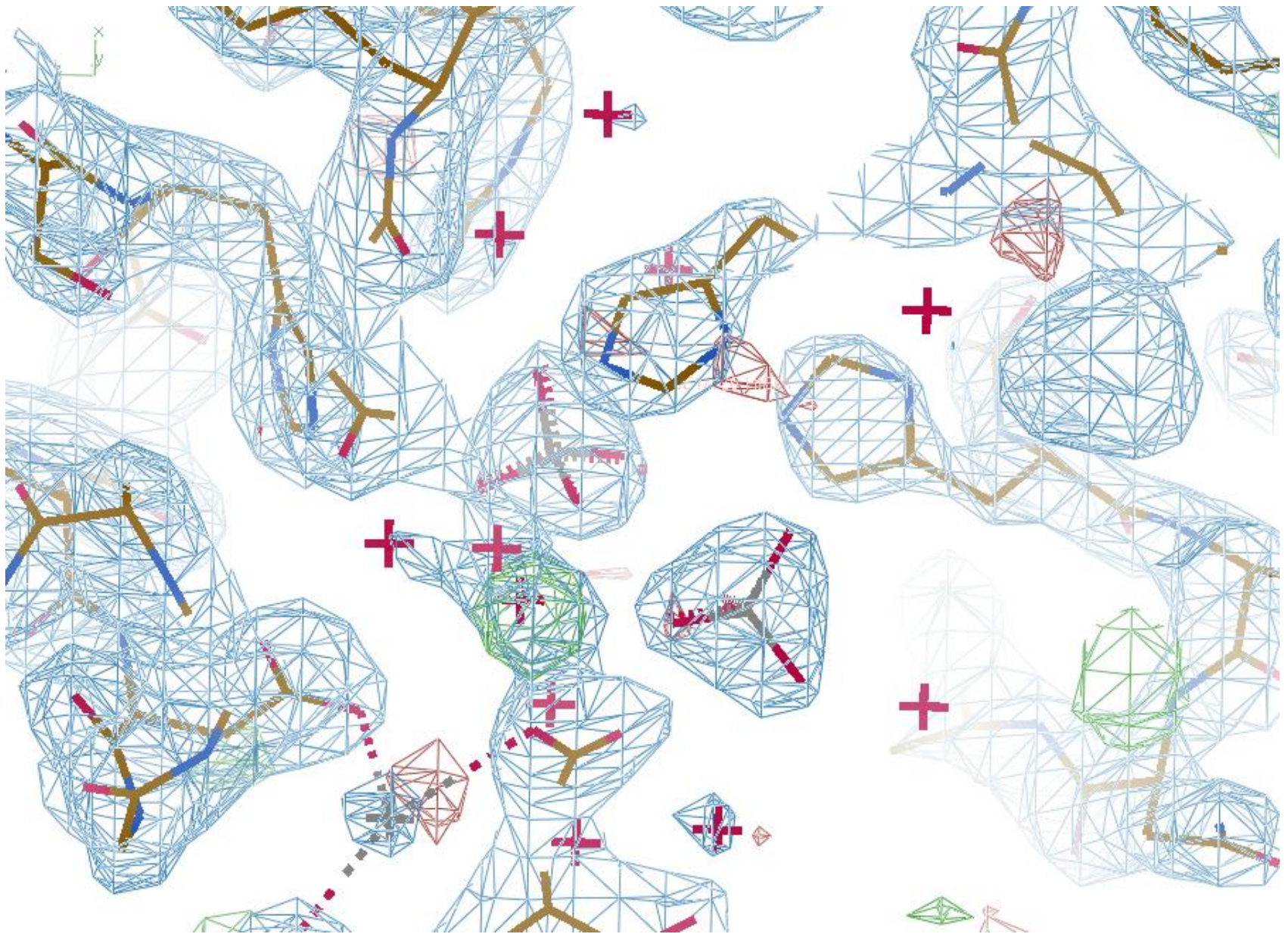
Conditions showing crystal hits are given for each cluster along with the total number of cocktails in that cluster.



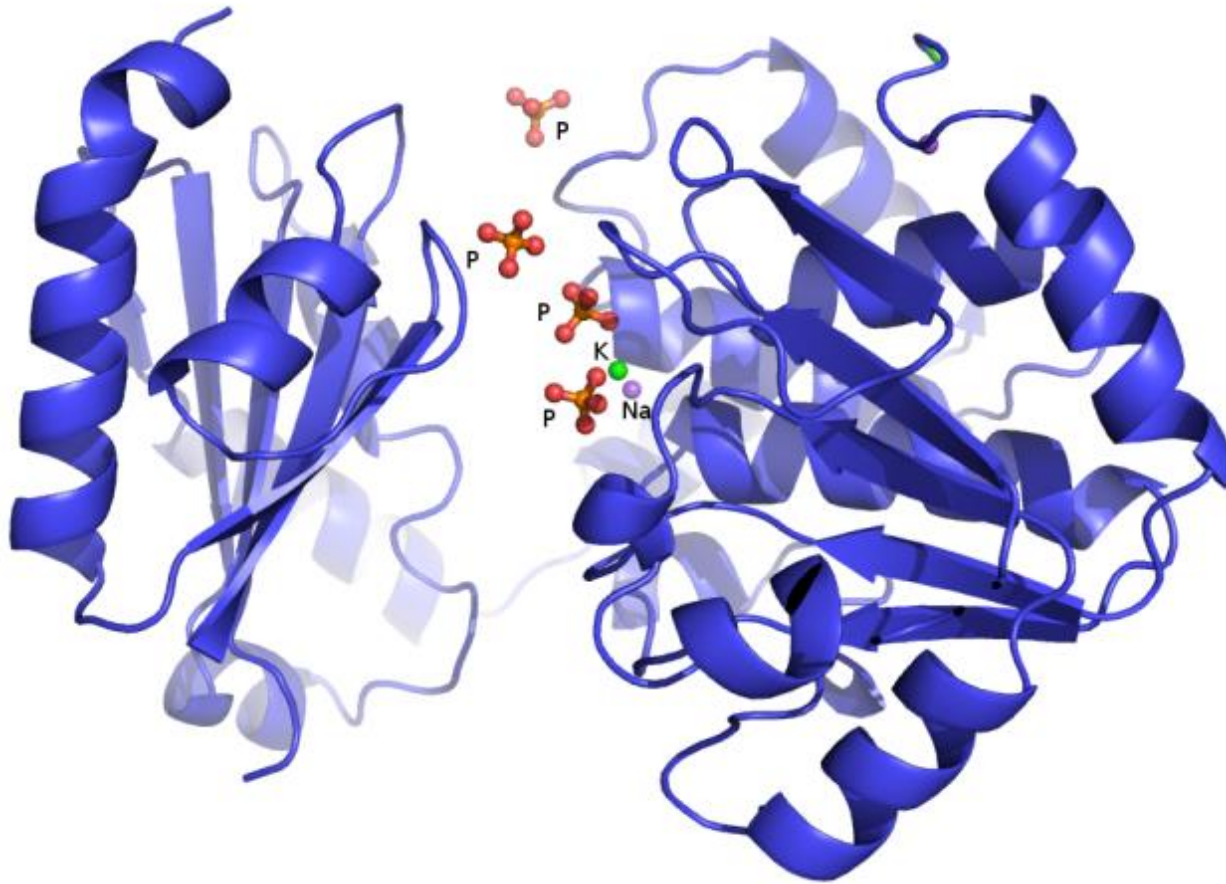
Cluster 20, PEG based, only 3 hits

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					
C11	42					
C17	28					
C20	965					
C15	19					
C23	8					
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 versus 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.



A Revised Structure Illustrating Mechanism

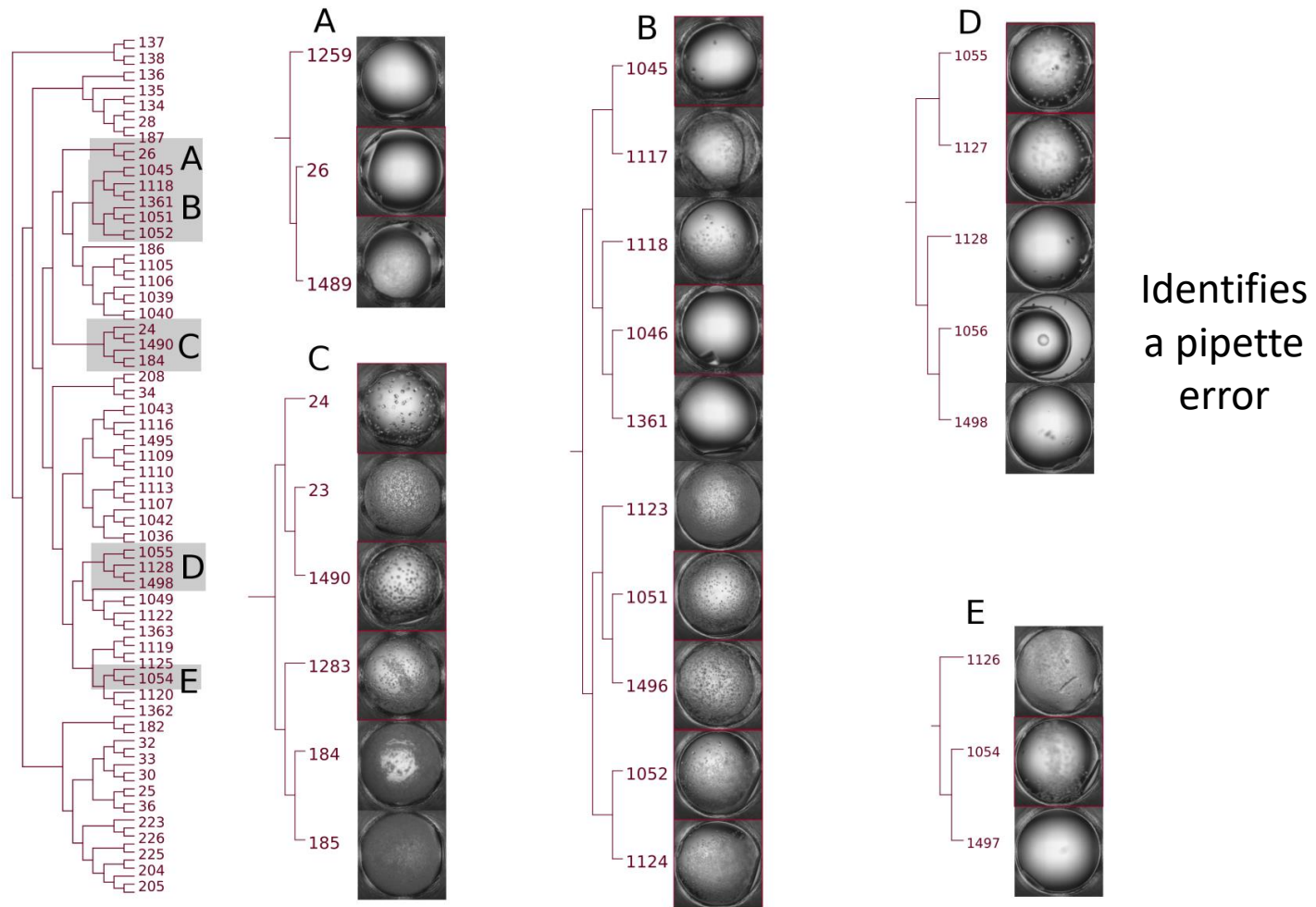


The putative active site has features that are consistent with active sites of other phosphatases which are involved in binding the phosphoryl moieties of nucleotide triphosphates

PDB 4PY9

Reduced the R and R_{free} from 22.3% and 25.9% to 20.7% and 24.3%

Potential to understand phase diagram in terms of X-ray diffraction properties



Clustering samples the phase diagram

Using historic data to identify patterns.

OPEN ACCESS Freely available online

PLOS ONE

Statistical Analysis of Crystallization Database Links Protein Physico-Chemical Features with Crystallization Mechanisms



Diana Fusco^{1,2}, Timothy J. Barnum^{2,3}, Andrew E. Bruno⁴, Joseph R. Luft⁵, Edward H. Snell^{5,6}, Sayan Mukherjee⁷, Patrick Charbonneau^{2,8*}

1 Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina, United States of America, **2** Department of Chemistry, Duke University, Durham, North Carolina, United States of America, **3** Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **4** Center for Computational Research, State University of New York, Buffalo, New York, United States of America, **5** Hauptman-Woodward Medical Research Institute, Buffalo, New York, United States of America, **6** Department of Structural Biology, State University of New York, Buffalo, New York, United States of America, **7** Department of Statistical Science, Department of Computer Science and Department of Mathematics, Duke University, Durham, North Carolina, United States of America, **8** Department of Physics, Duke University, Durham, North Carolina, United States of America

Abstract

X-ray crystallography is the predominant method for obtaining atomic-scale information about biological macromolecules. Despite the success of the technique, obtaining well diffracting crystals still critically limits going from protein to structure. In practice, the crystallization process proceeds through knowledge-informed empiricism. Better physico-chemical understanding remains elusive because of the large number of variables involved, hence little guidance is available to systematically identify solution conditions that promote crystallization. To help determine relationships between macromolecular properties and their crystallization propensity, we have trained statistical models on samples for 182 proteins supplied by the Northeast Structural Genomics consortium. Gaussian processes, which capture trends beyond the reach of linear statistical models, distinguish between two main physico-chemical mechanisms driving crystallization. One is characterized by low levels of side chain entropy and has been extensively reported in the literature. The other identifies specific electrostatic interactions not previously described in the crystallization context. Because evidence for two distinct mechanisms can be gleaned both from crystal contacts and from solution conditions leading to successful crystallization, the model offers future avenues for optimizing crystallization screens based on partial structural information. The availability of crystallization data coupled with structural outcomes analyzed through state-of-the-art statistical models may thus guide macromolecular crystallization toward a more rational basis.

Three classes of protein in data set:

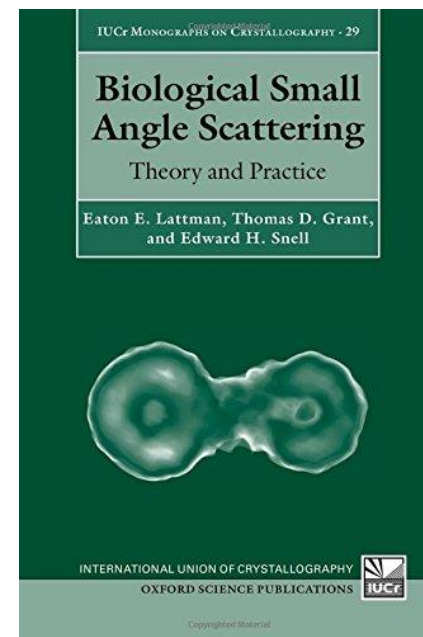
- Crystallizable by PEG
- Crystallizable by salt
- Crystallizable by very high salt concentration

The last class corresponded to salt crystals!

Simple patterns can be found from a sub-set of historical data. In this case samples likely to be crystallized via charge (salts) versus crowding (e.g. PEGs) could be identified.

Complementary analysis (after the fact)

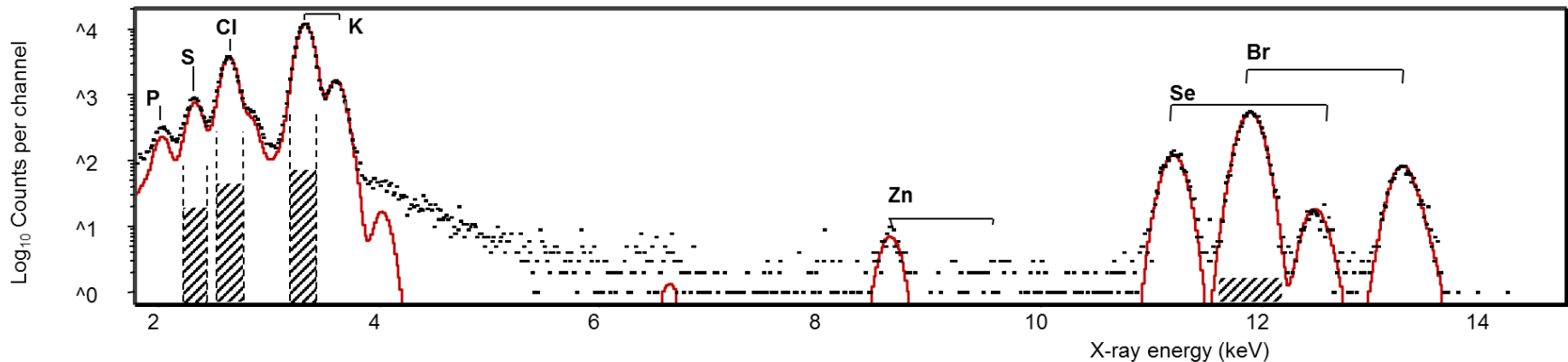
- Small Angle X-ray Scattering
 - Possible in a high-throughput manner but not as high-throughput as crystallization screening.
 - Characterizes sample in term of globularity, flexibility, and dynamics.
 - Provides oligomer information.
- Particle induced emission of X-rays
 - Collaboration with Elspeth Garman and Geoff Grimes in the UK who developed high-throughput processes.
 - Atomic technique – can use ‘dead’ protein.
 - Quantitative technique to accurately identify and measure heavy atoms.



Shameless plug

Particle induced X-ray emission

The energy of an X-ray emitted when an atomic electron undergoes an energy transition between its shell and a vacant electron site in a lower energy shell (e.g. for an M to L shell transition, sulphur gives a 2.3 keV X-ray) gives an unambiguous identification of atoms.



Emission of the characteristic X-rays from a sample can be induced by an incident beam of high energy protons (Particle Induced X-ray Emission: PIXE).

Collaboration with Elspeth Garman and Geoffrey Grime


The experiment


- 34 metalloprotein samples chosen from a set of samples successfully crystallized in the High-Throughput Crystallization Screening Center.
- All were SeMet samples.
- All produced crystals and a had structural model deposited in the PDB.
- PIXE analysis was carried out on each sample.

- The samples used were split into four groups based on PIXE analysis
 - Those where the PDB was inconsistent with the PIXE data
 - Those where extra metals were seen in the PIXE data (but not present in the PDB)
 - Those that were consistent with the PIXE data.
 - Those that produced no signal.

	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
PDB inconsistent with PIXE							
1		BiR14	456	Ca	-	Ca, Mn	18% PEG 3350, 0.2M Ca acetate, 0.1M MES, pH 6.15
2		NsR437I	106	Mn	-	-	20% PEG 4000, 0.1M Mn chloride, 0.1M MES, pH 6.0
3		SnR135D	161	Ca	-	Ca	20% PEG 8000, 0.1M Ca acetate, 0.1M MES, pH 6.0
4		Protein A	283	Fe/Zn	Ca (3.3), Mn (0.5), Fe (1.2), Co (1.2)	Zn	15% PEG 8000, 0.17 M sodium acetate, 0.01 M L-cysteine, 0.1 M MES pH 6.2
5		NsR236	119	K	-	Ca	8.64 M K acetate, 0.1 M TAPS, pH 9.0
6		NsR437H	141	Mn	-	Fe, Co	20% PEG 1000, 0.1M Mn chloride, 0.1M MES, pH 6.0
7		SoR237	137	Na	Co (0.7), Zn (0.7)	Fe, Ni	NaCl 200 mM, MES PH6, PEG 3350 20%, pH 6.15
8		BtR324A	169	Zn	-	Ca, Mn, Fe*	0.75M Mg Formate, 0.1M Bis-Tris, pH 7.0
9		GR157	262	Zn	-	Co	100 mM Na Acetate (pH 4.6), 30% MPD, and 200 mM NaCl.

 Model in the PDB containing a metal from the crystallization cocktail and not protein


 Model in the PDB containing an incorrect metal

	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
PDB inconsistent with PIXE							
1		BiR14	456	Ca	-	Ca, Mn	18% PEG 3350, 0.2M Ca acetate, 0.1M MES, pH 6.15
2		NsR437I	106	Mn	-	-	20% PEG 4000, 0.1M Mn chloride, 0.1M MES, pH 6.0
3		SnR135D	161	Ca	-	Ca	20% PEG 8000, 0.1M Ca acetate, 0.1M MES, pH 6.0
4		Protein A	283	Fe/Zn	Ca (3.3), Mn (0.5), Fe (1.2), Co (1.2)	Zn	15% PEG 8000, 0.17 M sodium acetate, 0.01 M L-cysteine, 0.1 M MES pH 6.2
5		NsR236	119	K	-	Ca	8.64 M K acetate, 0.1 M TAPS, pH 9.0
6		NsR437H	141	Mn	-	Fe, Co	20% PEG 1000, 0.1M Mn chloride, 0.1M MES, pH 6.0
7		SoR237	137	Na	Co (0.7), Zn (0.7)	Fe, Ni	NaCl 200 mM, MES PH6, PEG 3350 20%, pH 6.15
8		BtR324A	169	Zn	-	Ca, Mn, Fe*	0.75M Mg Formate, 0.1M Bis-Tris, pH 7.0
9		GR157	262	Zn	-	Co	100 mM Na Acetate (pH 4.6), 30% MPD, and 200 mM NaCl.

 Model in the PDB containing a metal from the crystallization cocktail and not protein

 Model in the PDB containing an incorrect metal

	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
Extra metals present in PIXE							
1		MuR16	210	Fe/Zn	Fe (0.6), Co (0.9), Ni (0.4), Zn (0.7)	-	0.1 M Na ₂ MoO ₄ *2H ₂ O, 0.1 M Bis-Tris propane, 12% PEG 20000
2		MqR88	420	Na♦	Ca (7.1)	Fe	0.1 M Na ₂ MoO ₄ , 0.1 M Tris, pH 8.0, 20% PEG 8000
3		SR677	222	Mg♦	Ca (0.7), Fe (0.05)	K/Br	0.1 M KH ₂ PO ₄ , 0.1 M NaC ₂ H ₃ O ₂ , pH 5.0, 12% PEG 20000
4		DrR130	296	Mg♦	Ca*	-	0.1 M NaCl, 0.1 M TAPS, pH 9.0, 18% PEG 3350, MgCl ₂
5		BtR319D	172	Mg♦	Ca (1.74)	-	None given
6		ShR87	320	Mg♦	Mn (0.8), Fe (0.7)	-	0.1 M Na citrate, pH 5.2, 1.25 M Li ₂ SO ₄ , 0.5 M (NH ₄) ₂ SO ₄
7		SmR83	218	Mg♦	Ca (0.5), Fe (0.1)	Ti, Co, Cu	0.1 M LiCl ₂ , 0.1 M Bis-Tris, pH 5.5, 18% PEG 3350
8		NsR141	225	Mg♦	Mn (0.2), Fe (0.4), Ni (0.4)	Co	0.1 M citric acid, pH 5.0, 1.6 M (NH ₄) ₂ SO ₄
9		ZR319	289	Mg♦	-	Ca, Fe, Cu	0.1 M Tris-HCl, pH 9.1, 18% PEG 3350, 0.1 M MgSO ₄

 Model in the PDB containing an extra misidentified metal

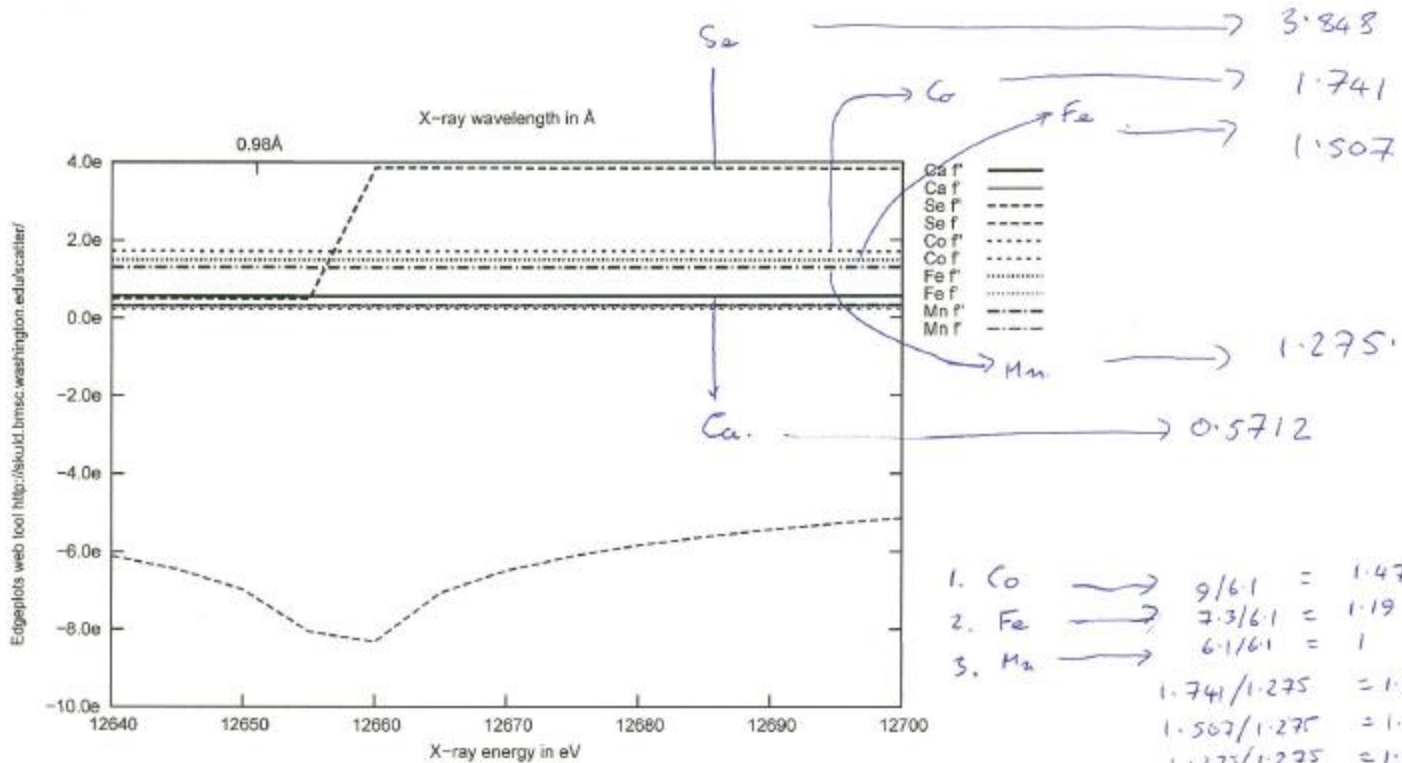
- Of the 34 samples analyzed, 9 were inconsistent with the PDB results, 9 had extra metals present, 18 were consistent, and 2 were unsuitable for analysis due to low protein concentration on the sample.
- In total, 18 of the 32 analyzable samples (56%) were not correctly or fully described in the PDB deposition.

Se 3.848
 Co 1.741
 Fe 1.507
 Mn 1.275
 Ca 0.5712

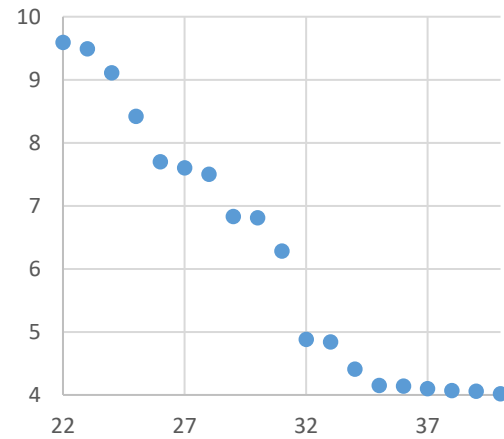
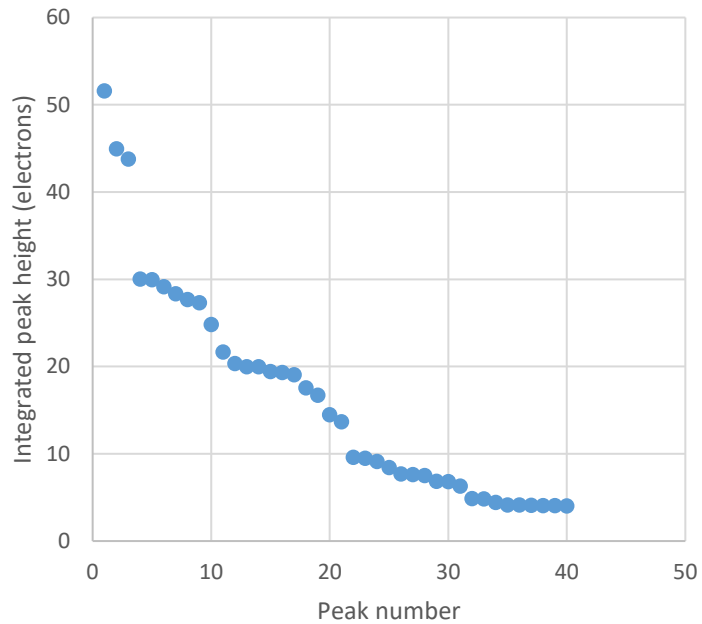
A curve

→ (A) 1.
 2.
 3.

ED.
 25.79 → Next level ② 7.3 min
 17.19 → Lowest ED ③ 6.1 min.
 13.95 → Highest ① 9.0e-5



Wavelength		0.97931				
	f'	f''	f'' n_Se	f'' n_Zn	f'' n_Co	f'' n_Fe
Se	-8.6571	3.843	1.000			
Zn	-0.3843	2.477	0.645	1.000		
Co	0.1697	1.715	0.446	0.692	1.000	
Fe	0.2421	1.500	0.390	0.606	0.875	1.000
Mn	0.2905	1.303	0.339	0.526	0.760	0.869
Ca	0.2938	0.565	0.147	0.228	0.329	0.376
O	0.0163	0.012	0.003	0.005	0.007	0.008



21 Se atoms, 7 in each chain A,B and C

Wavelength

0.97931

	f'	f''	f'' n_Se	f'' n_Zn	f'' n_Co	f'' n_Fe
Se	-8.6571	3.843	1.000			
Zn	-0.3843	2.477	0.645	1.000		
Co	0.1697	1.715	0.446	0.692	1.000	
Fe	0.2421	1.500	0.390	0.606	0.875	1.000
Mn	0.2905	1.303	0.339	0.526	0.760	0.869
Ca	0.2938	0.565	0.147	0.228	0.329	0.376
O	0.0163	0.012	0.003	0.005	0.007	0.008

Chain A

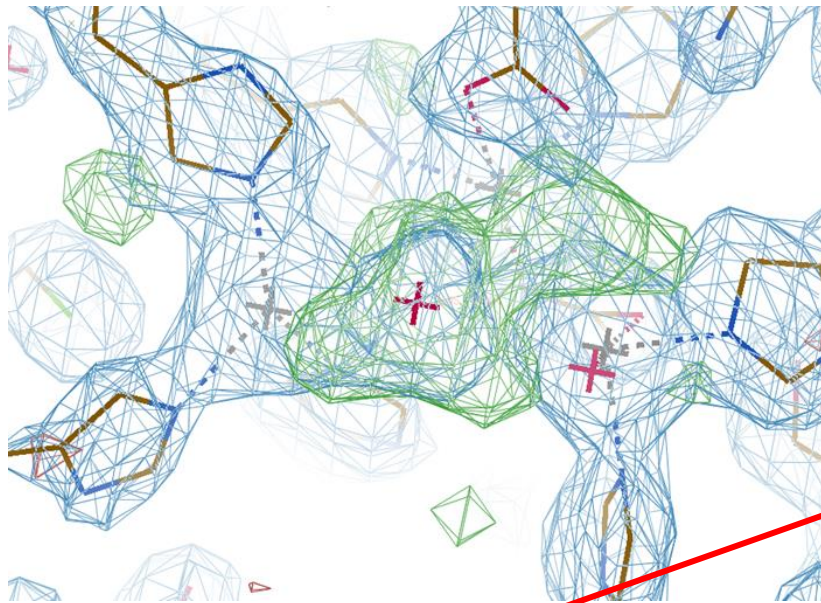
9.59 1.00
6.83 0.71
6.81 0.71

Chain B

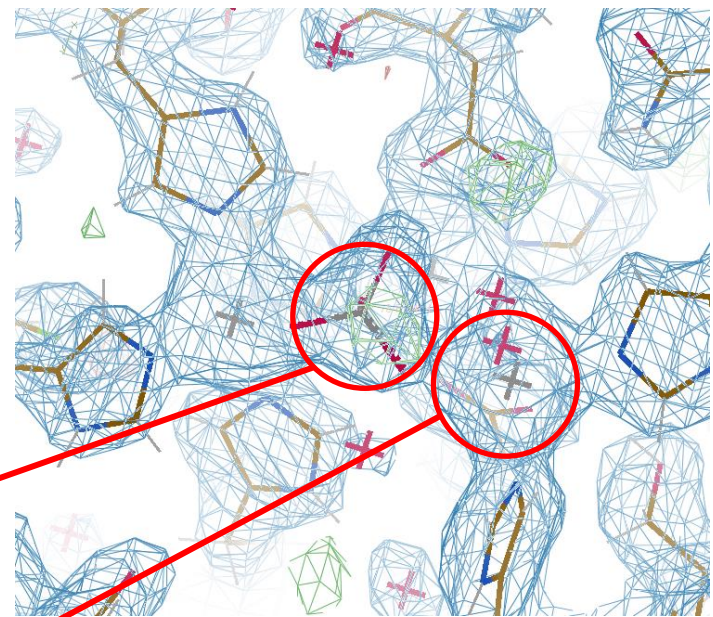
8.42 1.00
7.60 0.90
7.50 0.89

Chain C

9.11 1.00
7.70 0.85
6.28 0.69



PO₄



Fe

Metals in new structure, Fe, Mn, Co cluster

	R _{work}	R _{free}	RMS(bonds)	RMS(angles)	Clash	Ram-fav	Ram-out	Rot-out
PDB	0.193	0.212	0.008	1.2	11.97	96.07	0.61	
Re-refined	0.1847	0.2143	0.0031	0.744	1.9	96.81	0.61	2.82
Metal	Metals replaced with Co, Fe and Mn, PO ₄ added in active site. Ca added in places							
	15.60	18.50						



RESEARCH ARTICLE

Classification of crystallization outcomes using deep convolutional neural networks

Andrew E. Bruno¹, Patrick Charbonneau^{2,3}, Janet Newman⁴, Edward H. Snell^{5,6}, David R. So⁷, Vincent Vanhoucke^{7*}, Christopher J. Watkins⁸, Shawn Williams⁹, Julie Wilson¹⁰

1 Center for Computational Research, University at Buffalo, Buffalo, New York, United States of America, **2** Department of Chemistry, Duke University, Durham, North Carolina, United States of America, **3** Department of Physics, Duke University, Durham, North Carolina, United States of America, **4** Collaborative Crystallisation Centre, CSIRO, Parkville, Victoria, Australia, **5** Hauptman-Woodward Medical Research Institute, Buffalo, New York, United States of America, **6** SUNY Buffalo, Department of Materials, Design, and Innovation, Buffalo, New York, United States of America, **7** Google Brain, Google Inc., Mountain View, California, United States of America, **8** IM&T Scientific Computing, CSIRO, Clayton South, Victoria, Australia, **9** Platform Technology and Sciences, GlaxoSmithKline Inc., Collegeville, Pennsylvania, United States of America, **10** Department of Mathematics, University of York, York, United Kingdom

* vanhoucke@google.com



June 20th 2018

Revisiting image analysis



Machine Automated Recognition of Crystallization Outcome

- A collaboration with Duke University, GalaxoSmithKline Inc, Google Brain, the University of Buffalo, CSIRO, University of York, Bristol-Myers Squibb, Merck, Abbvie and others (growing effort).
- Current training set of 493,214 human classified images limited to crystal, clear, precipitate, and other.
- Random set of 50,284 used for testing.
- Multiple image types
 - Different growth geometries – microbatch under oil and vapor diffusion
 - In house designed imagers, Rigaku, and Formulatrix systems
 - A human can interpret images from any imager, why can't an automatic procedure.

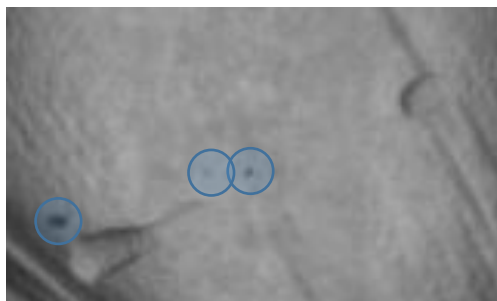
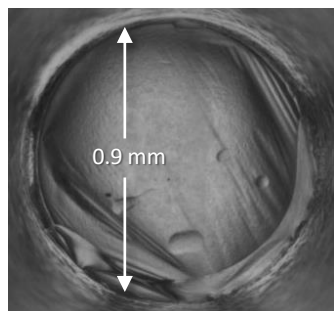
Some details

- Images classified by separate groups in multiple categories.
- Reclassified to four categories, crystal, precipitate, clear or other
- Classifier is a deep Convolutional Neural Network with an Inception-v3 architecture.
- Images are reduced to 599x599 pixel images which are further compressed to 299x299 pixels.
- Training data taken from images with random 599x599 patches treated to randomize brightness, saturation, hue, and contrast with random flipping left or right.
- Model was implemented in TensorFlow running across 50 Nvidia K80 GPUs.
- Training took 19 hours on 100 million images.
- Analysis for a new image is almost real time.

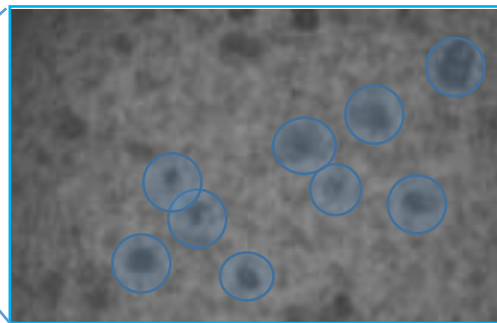
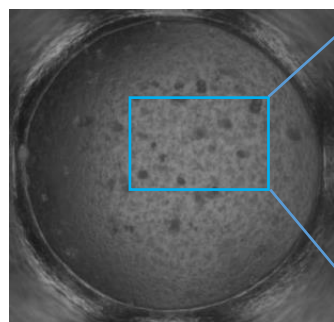
True label	Predictions			
	Crystals	Precipitate	Clear	Other
Crystals	91.0%	5.8%	1.7%	1.5%
Precipitate	0.8%	96.1%	2.3%	0.7%
Clear	0.2%	1.8%	97.9%	0.2%
Other	4.8%	19.7%	5.9%	69.6%

Remember, humans **at best** have a 80% success rate.

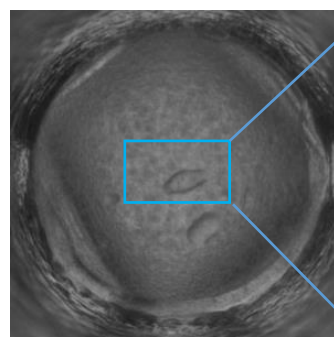
Sample X09664 - Reading 2/19/2008 – Week 2



Cocktail 1510, 0.93 probability of a crystal.



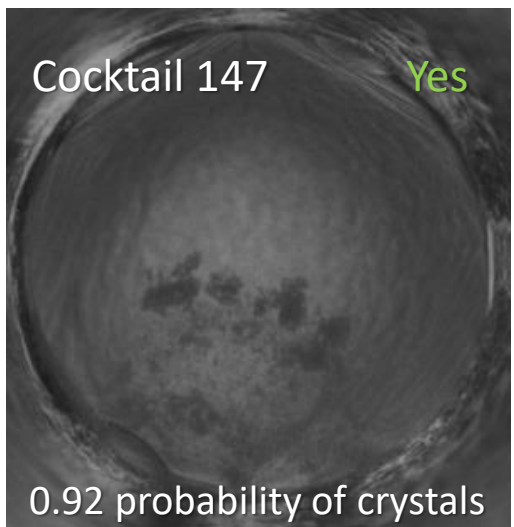
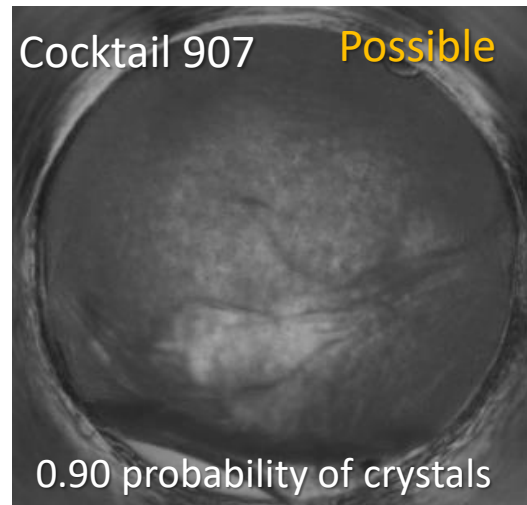
Cocktail 349, 0.93 probability of a crystal.



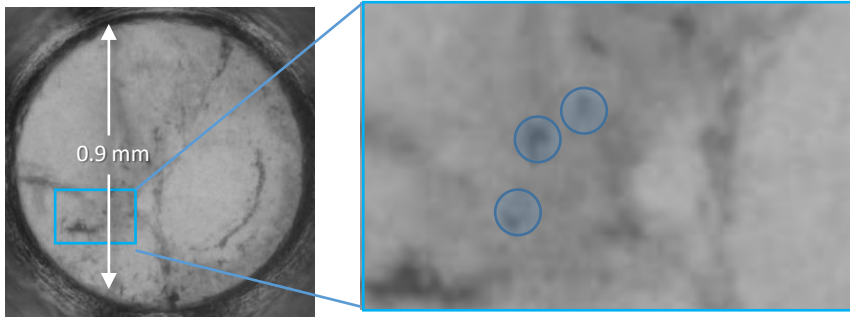
Cocktail 1492, 0.93 probability of a crystal (presence not clear by eye, questionable identification).

Crystals clearly identified (shown enlarged)

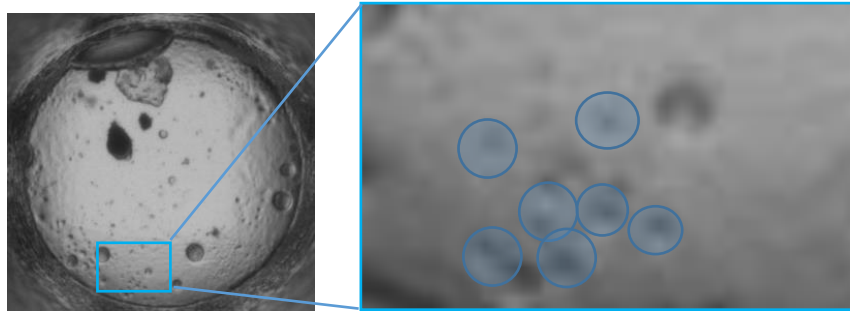
Sample X09664 - Reading 2/19/2008 – Week 2



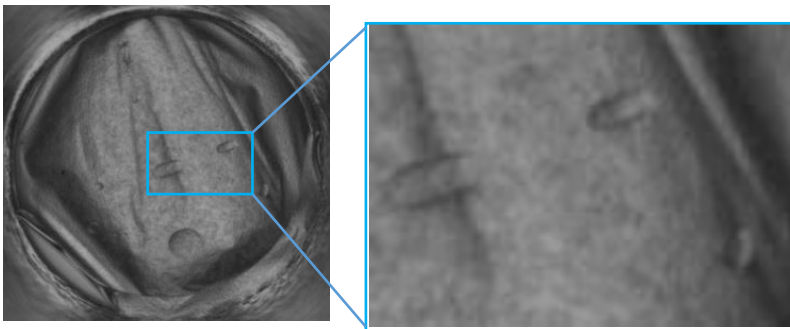
Sample X09664 - Reading 2/12/2008 – Week 1



Cocktail 1314, 0.93 probability of a crystal.



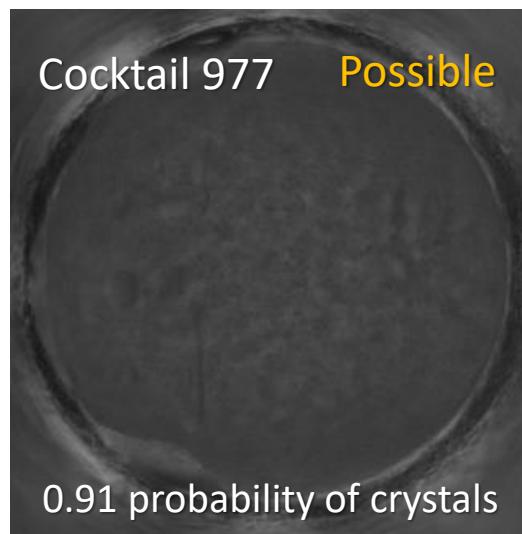
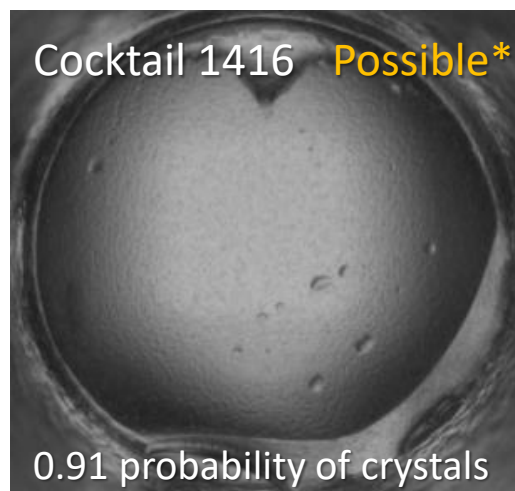
Cocktail 1255, 0.92 probability of a crystal (larger features in drop but also small crystals).



Cocktail 1332, 0.92 probability of a crystal.

Crystals clearly identified (shown enlarged)

Sample X09664 - Reading 2/12/2008 – Week 1



For 1416 – Autoscoring has possibly detected small air bubbles but small crystals are also present

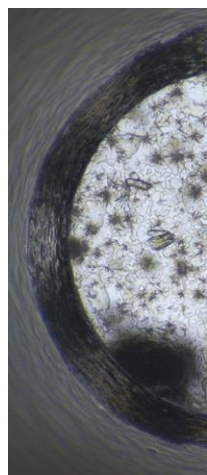
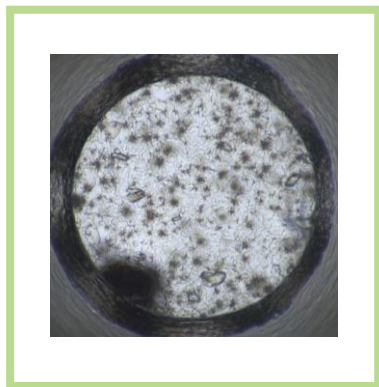
Big data needed

- Our data, 19,000 proteins, visual images at day 1 and weeks 1 through 6 (7 images total). UV two photon fluorescence and SONICC for many samples. Over 200 million data points – both success and failure.
- Biological Macromolecule Crystallization Database – crystallization conditions for ~100,000 proteins in PDB. No data on failure.
- Other large scale crystallization centers.
- The pharmaceutical industry.
- A robust image classification capability.

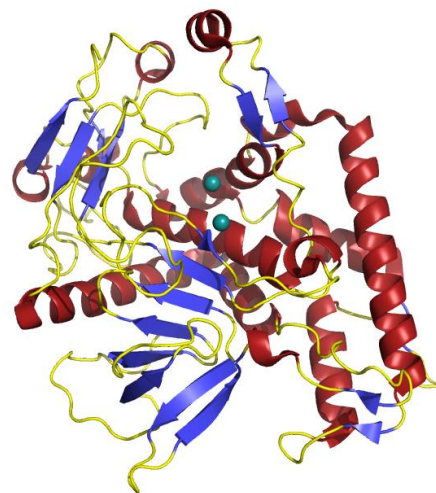
What if the visible images are ambiguous?

Do any of the

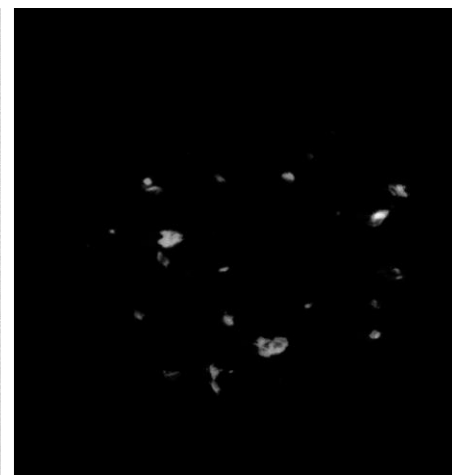
ystals?



Vis



Initial crystallization conditions for
structure of botulinum neurotoxin
Allen Lab BU and Janda Lab Scripps
JACS 2017; 139



SHG

Can we learn anything?

- Our experience with the manual analysis of images indicate that there is information available that is pertinent to structure.
- We don't know how much or if predictive patterns may exist.
- If they do, they may be complex.
- Do we have enough information to find signals in the noise?
- How can we get more?

Our goal is to link data representation tools with automatic image classification to obtain new information

Conclusion

T H E -ray F I L E S

Some truth may be out there



Thank you and questions?



esnell@hwi.buffalo.edu