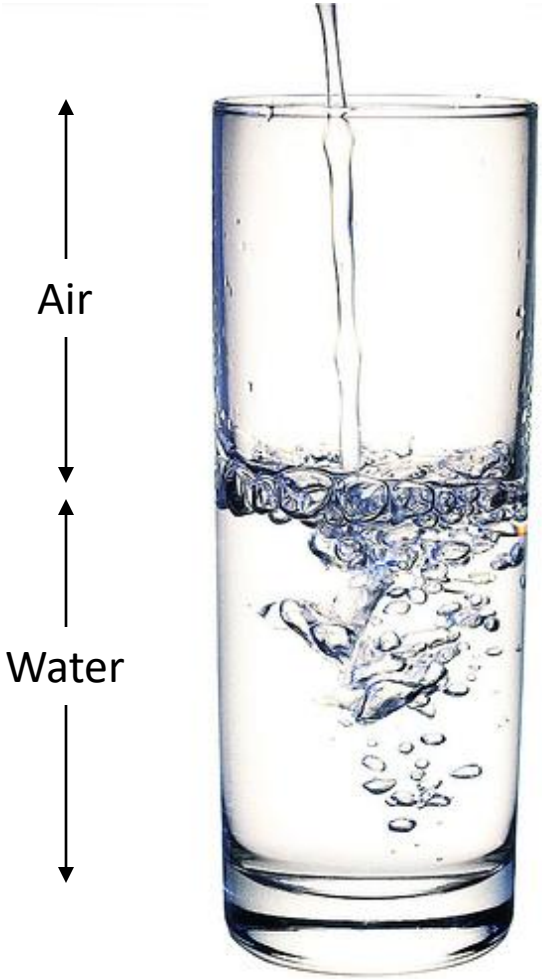
Comparing Chemistry to Outcome: Coupling a chemical distance metric, with clustering and hierarchal visualization.

Andrew E. Bruno, Amanda M. Ruby, Joseph R. Luft, Thomas D. Grant,
Jayaraman Seetharaman, Gaetano T. Montelione,
John F. Hunt, and Edward H. Snell

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)

Fantasy

High-throughput Crystallization Screening at the Hauptman-Woodward Medical Research institute

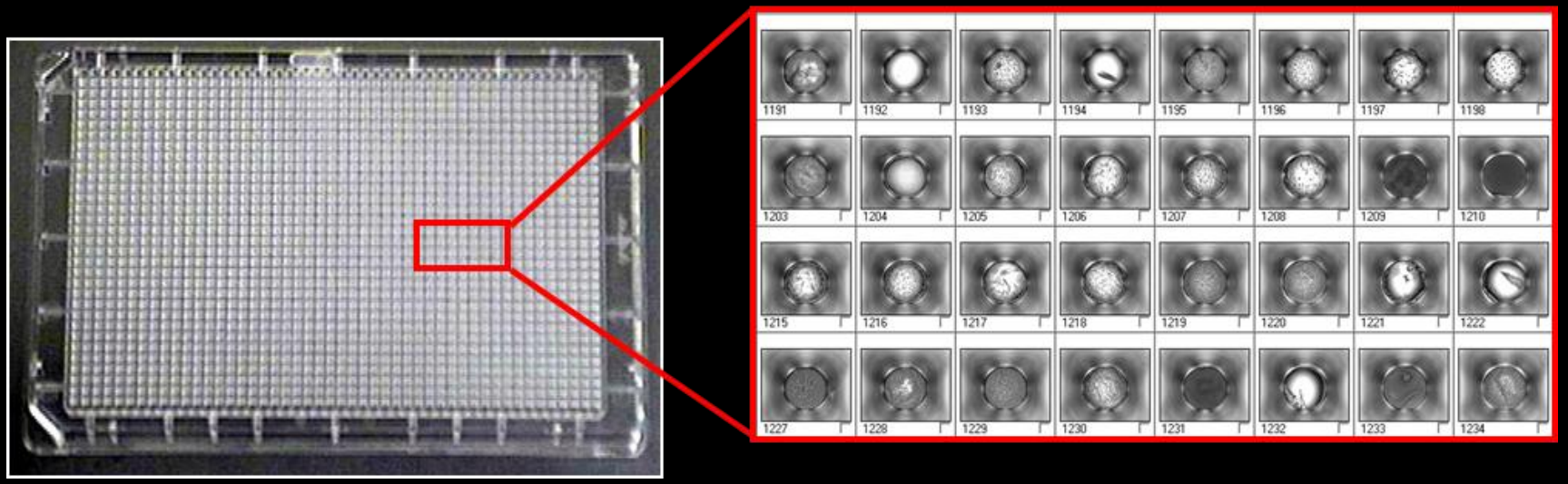
The Crystallization Screening laboratory at the Hauptman-Woodward Medical Research Institute

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.

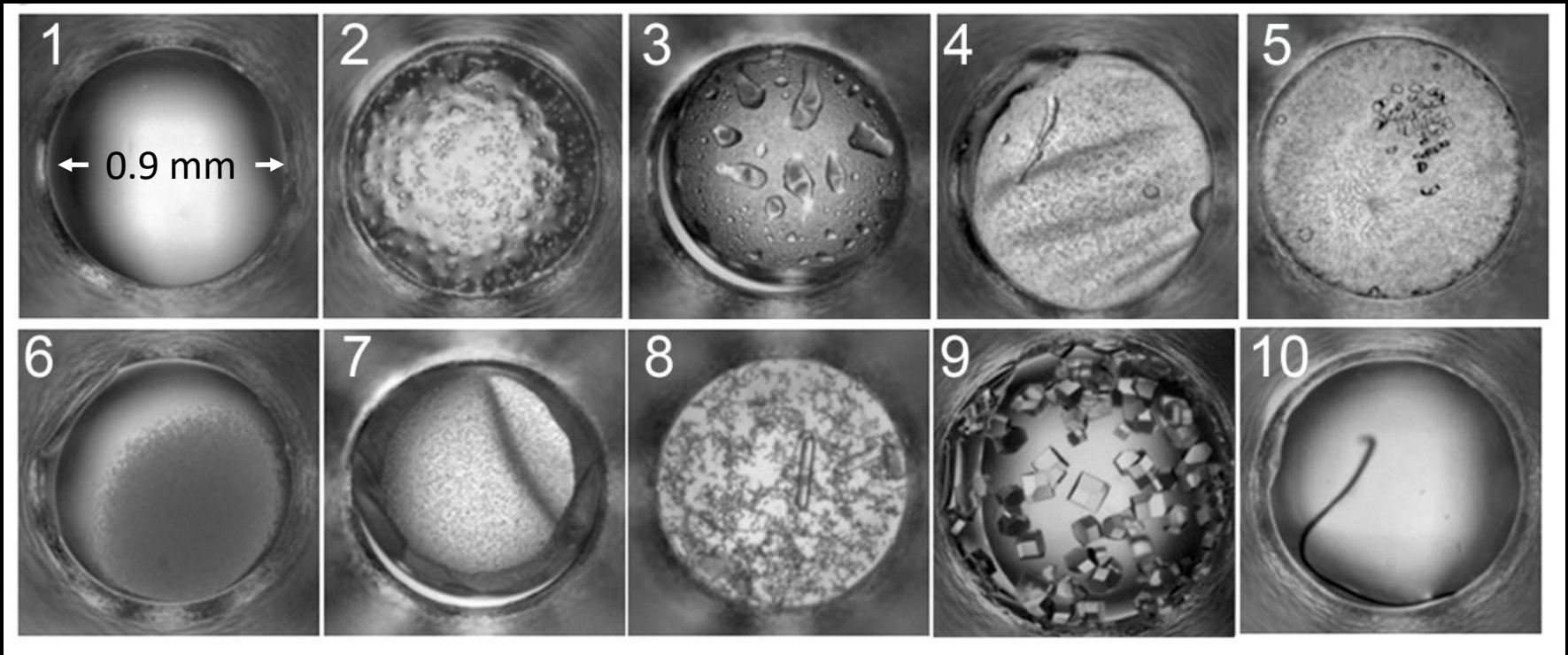


The crystallization method used is micro-batch under oil with 200 nl of protein solution being added to 200 nl of precipitant cocktail in each well of a 1536 well plate.

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



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***

Chemical/Molecular Fingerprints

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 pH 4.6	calcium chloride dihydrate	0.02	M	[Ca+2].[Cl-].[Cl-].O.O	147.0146	
	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 pH 7.5	sodium chloride	4.48	M	[Na+].[Cl-]	58.4428	
	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.

Comparing Cocktail Fingerprints

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:

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

A real example with our
1,536 condition screen

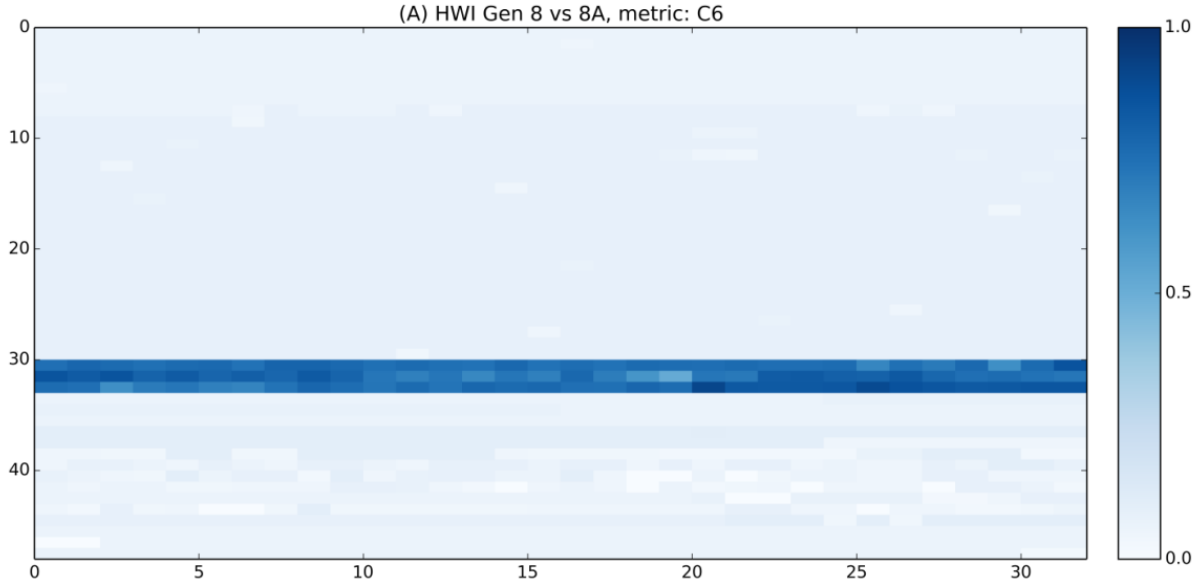
Cocktail similarity measures are not new.

We build on the original work by Janet Newman's in Melbourne, Australia who originated the concept of a similarity measure (termed C6) within crystallization to compare individual cocktails and different screening kits. (Newman J, Fazio VJ, Lawson B, Peat TS (2010) The C6 Web Tool: A Resource for the Rational Selection of Crystallization Conditions. Crystal Growth & Design 10: 2785-2792).

Our internal 1,536 screens are reformatted on a yearly basis to remove any conditions that produce salt crystals, to incorporate the latest screening developments, and building on internal research into crystallization processes.

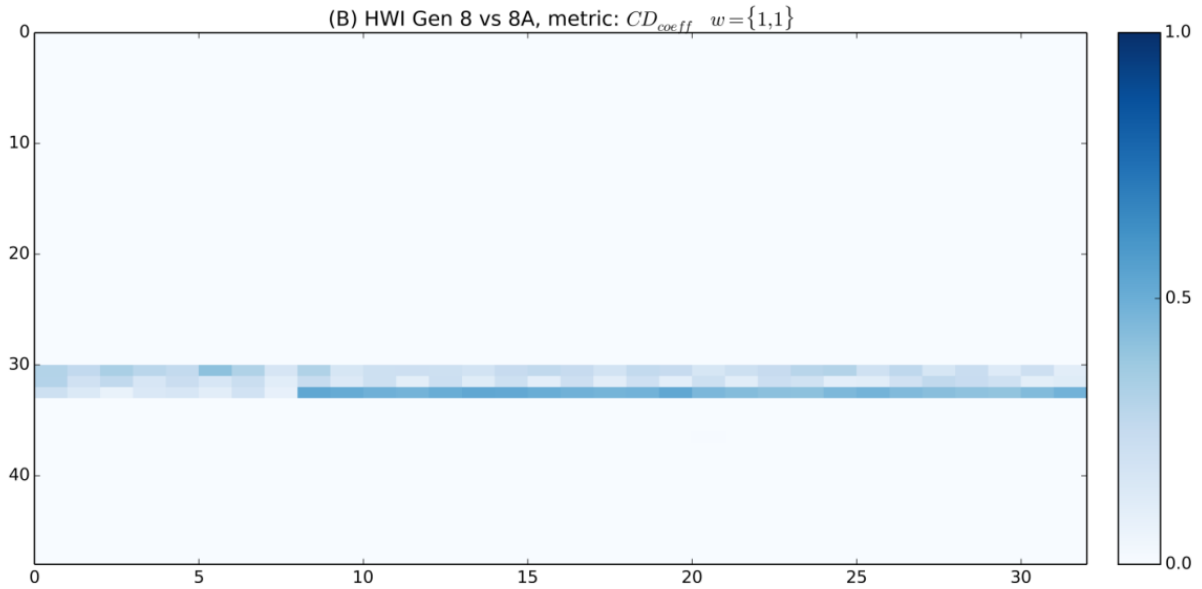
In this example we apply both the C6 and our new similarity measure to two generations of screen where 96 conditions have been replaced with a new commercially available screen/

The C6 metric color coded according to dissimilarity (0 is identical, 1 is most dissimilar)



The new dissimilarity metric.

Note that the only change in the screen was replacing 96 conditions



Clustering then using a hierarchal display

The Dissimilarity Measure Over the Whole Screen

Aspects of the screen design are clearly seen

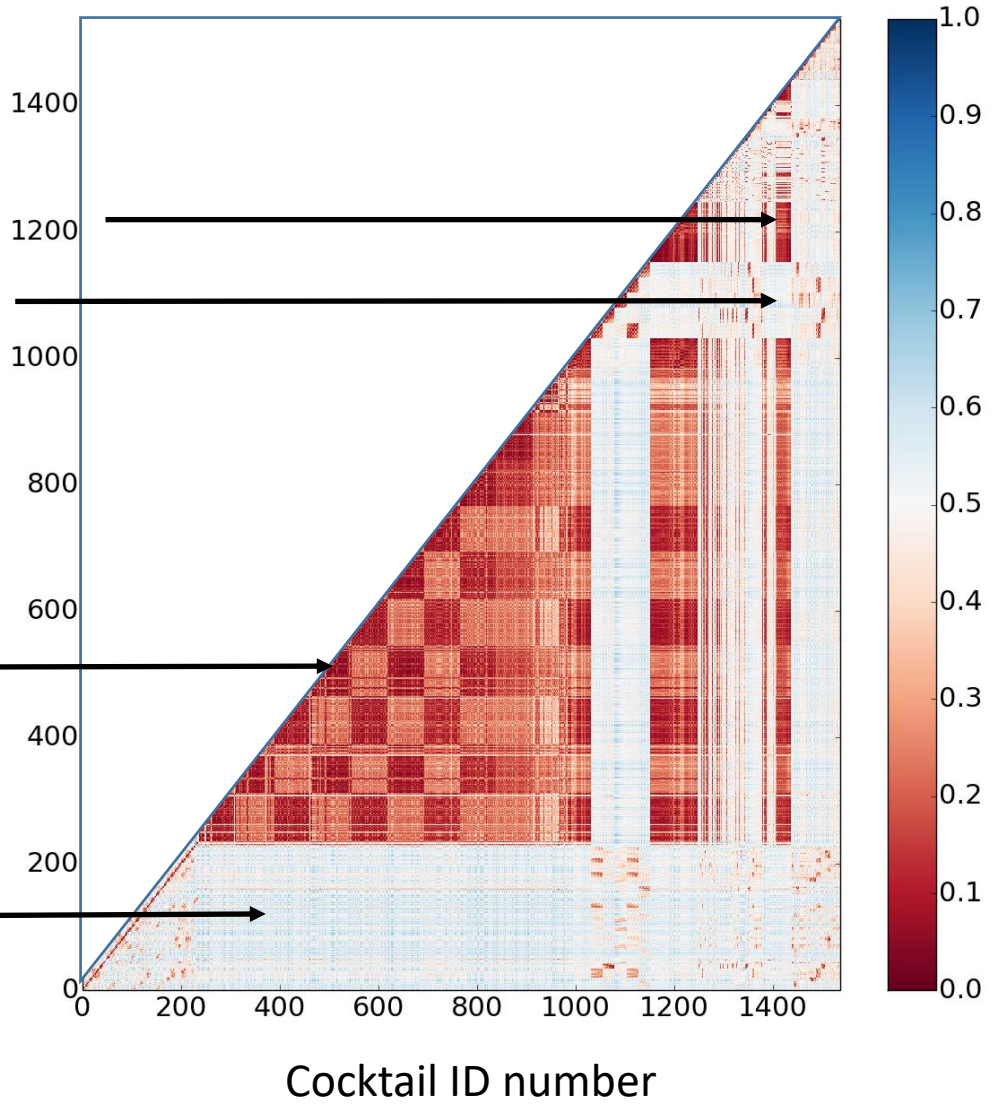
Hampton Research PEG/Ion screen

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

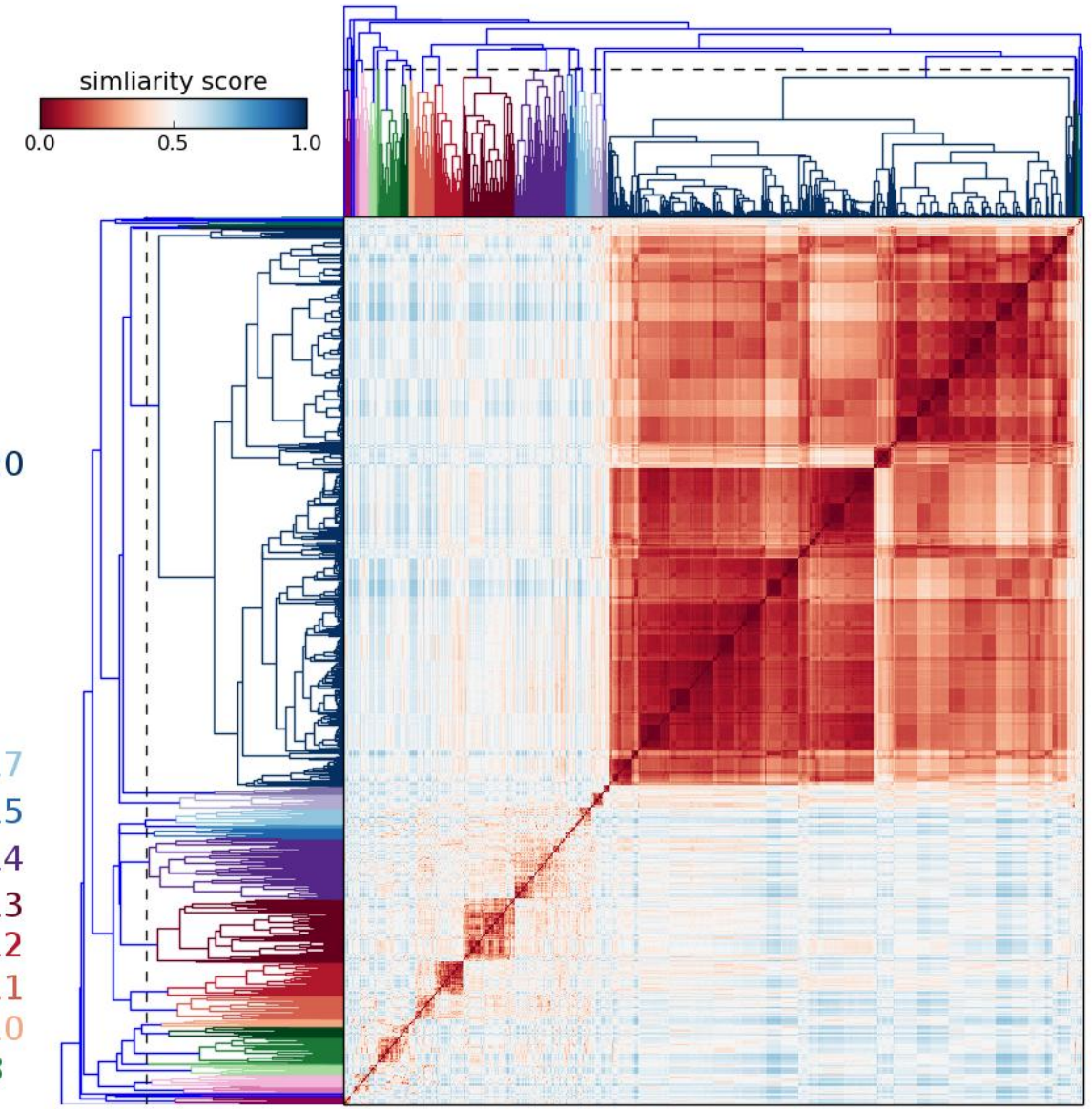


Automatic Clustering of the Results

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 →



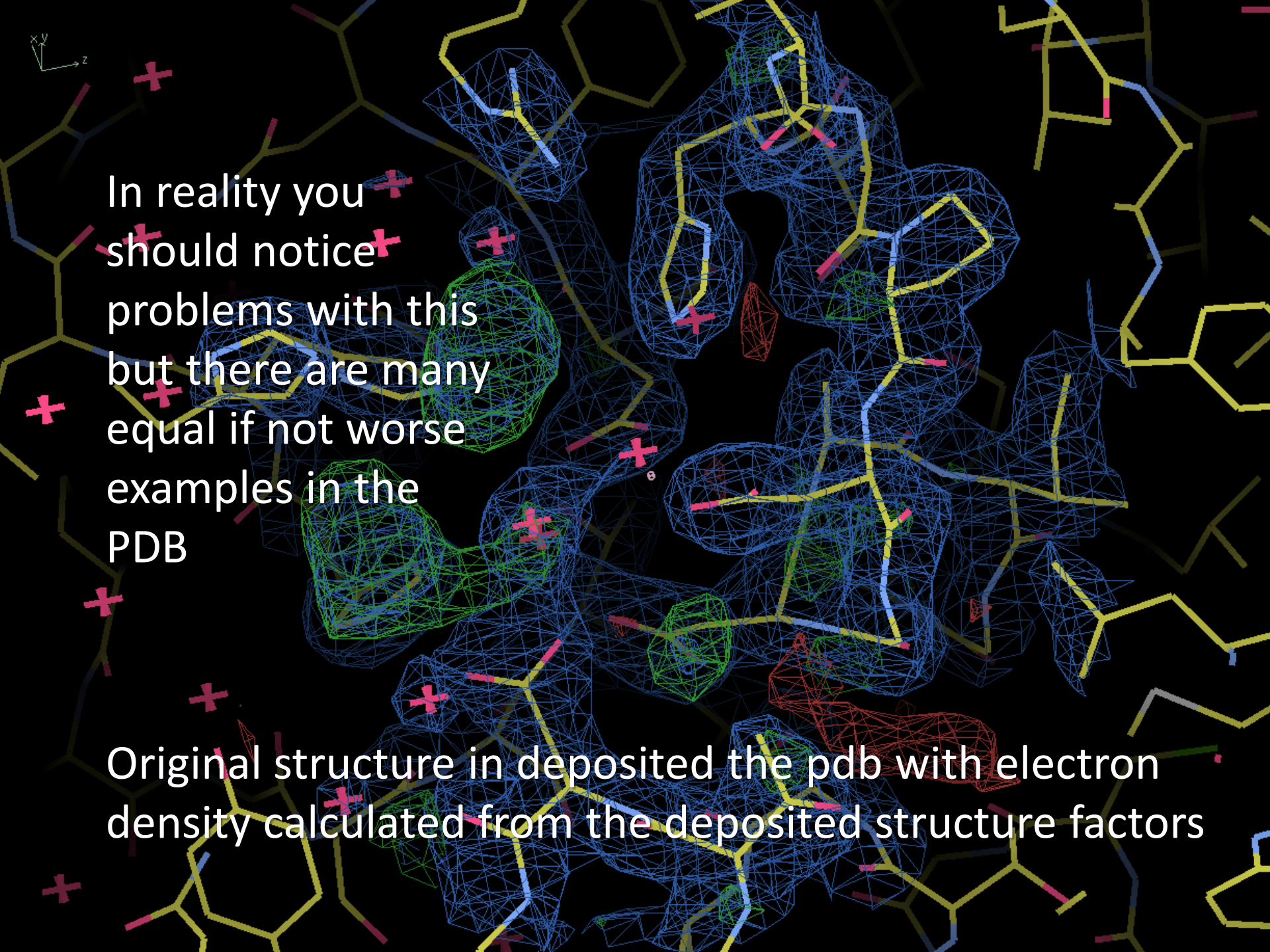
So how do we make use of it?

Cocktail similarity measures are not new.

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

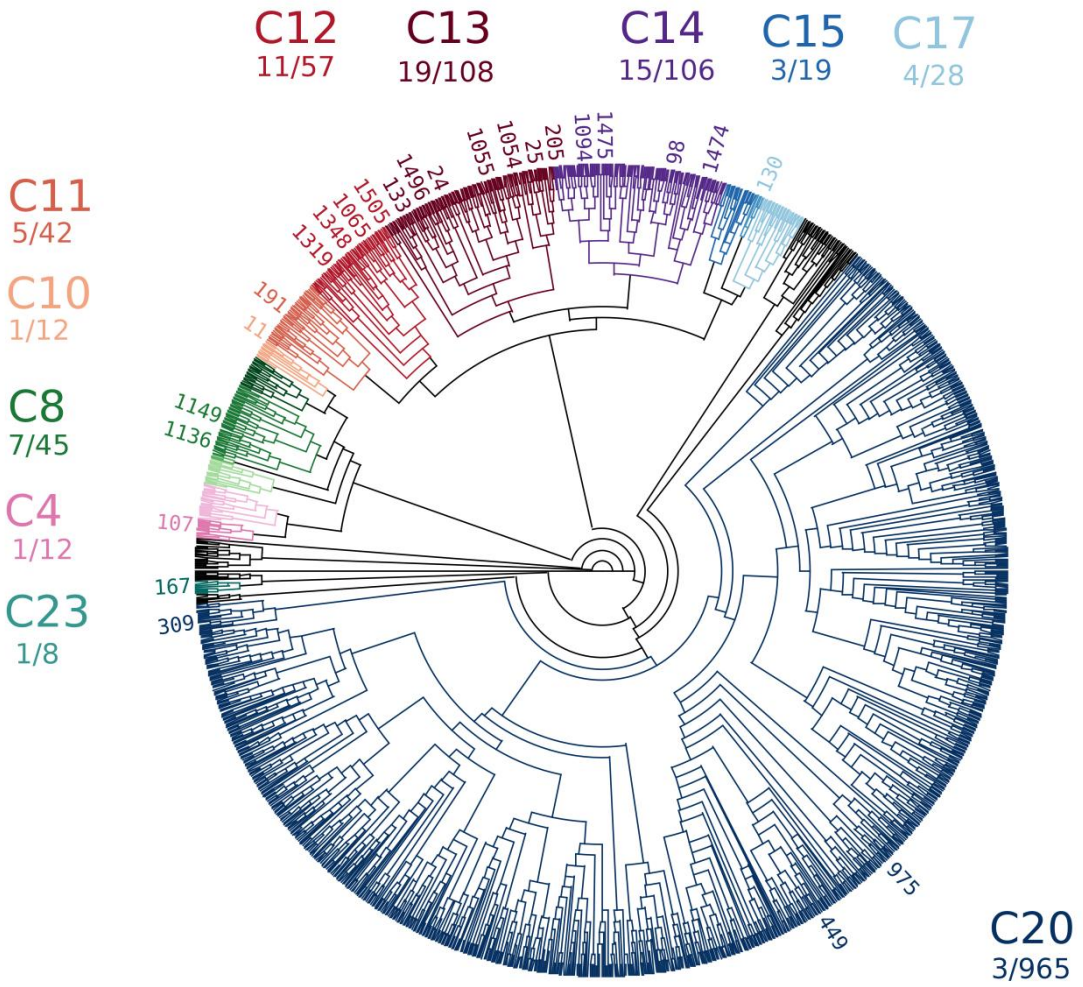
The optimized conditions for crystallization combined 5µl of the protein at 7.4 mg/ml concentration was mixed with the precipitant containing 320mM potassium acetate, 100 mM sodium acetate, pH 6.5 in 1:1 ratio. Crystals appeared in one week.


In reality you should notice problems with this but there are many equal if not worse examples in the PDB
Original structure in deposited the pdb with electron density calculated from the deposited structure factors

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

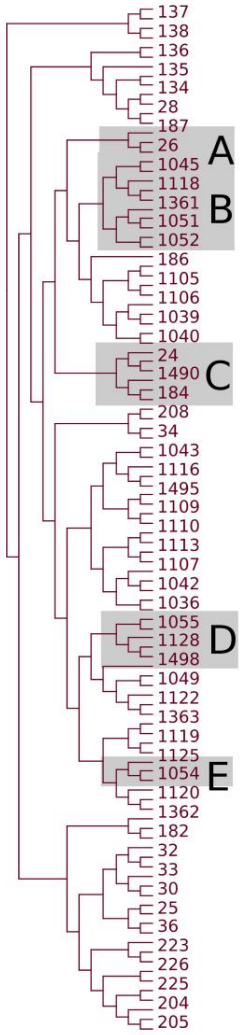


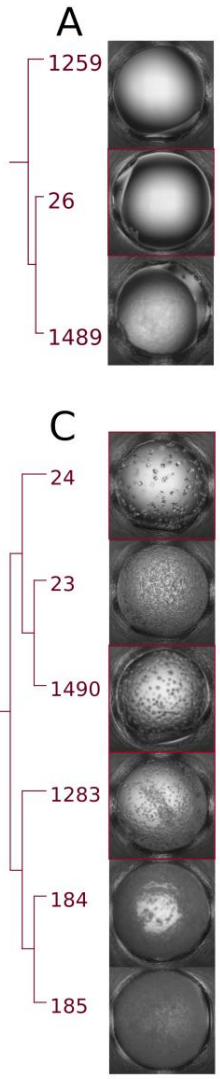
Cluster 20, PEG based, only 3 hits

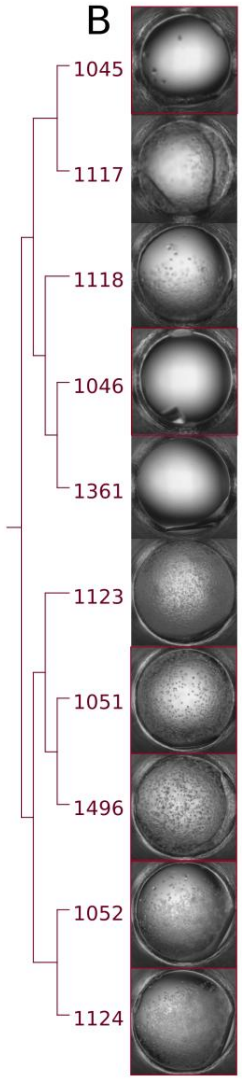
Cluster	Total	Hits	% hits	Sodium %	Potassium %	Phosphate %
All cocktails						
	1536	70	4.5	47	24	16
All crystal						
	70	70	100	70	27	30
Clusters with crystals						
C13	108	19	17.6	73	72	100
C14	106	15	14.2	65	21	0
C12	57	11	19.3	16	2	0
C8	45					
C11	42					
C17	28					
C20	965					
C15	19					
C23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

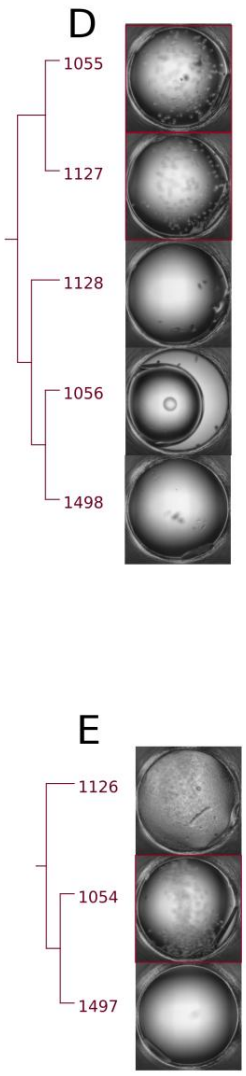
Cluster 13 proved interesting in that sodium is present in 73% of the conditions versus 47% for the 1536 condition screen overall, potassium is present in 72% of the conditions verses 24% overall and finally phosphate is present in 100% of the conditions versus 16% overall. This suggests a strong influence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

Zoom in on Cluster 13



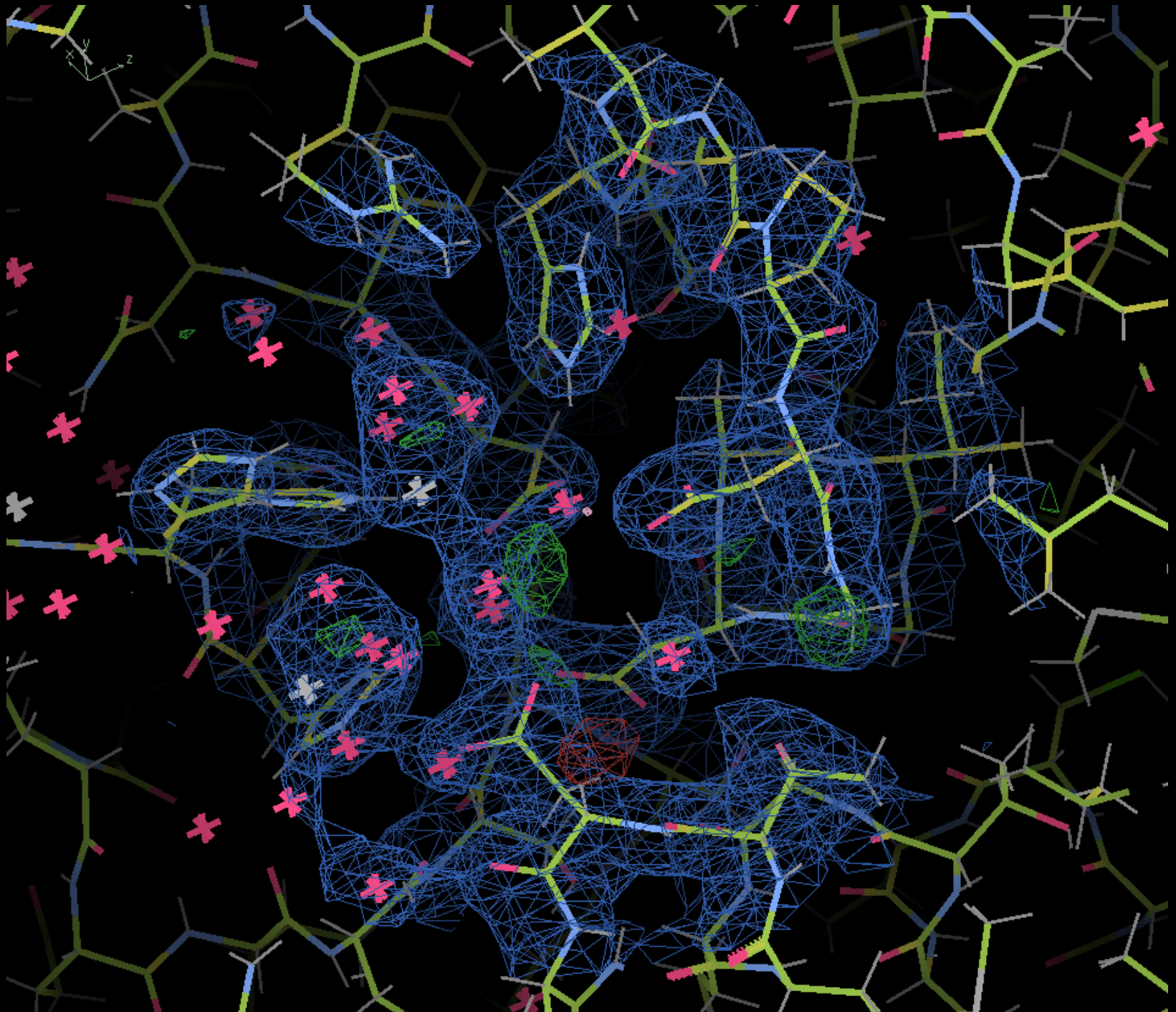




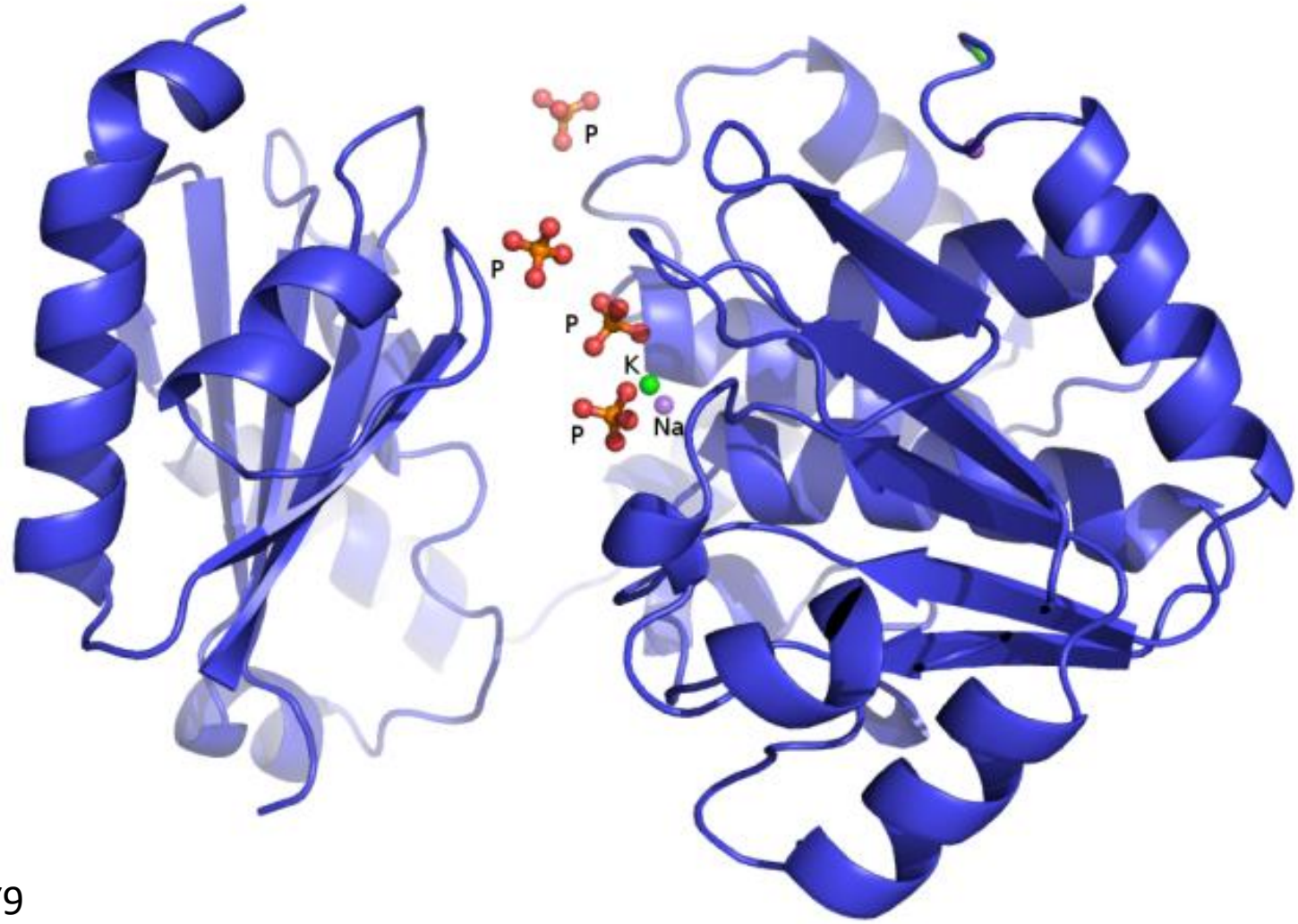


Identifies a pipette error

Clustering samples the phase diagram



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 cleft between the two domains
- 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.

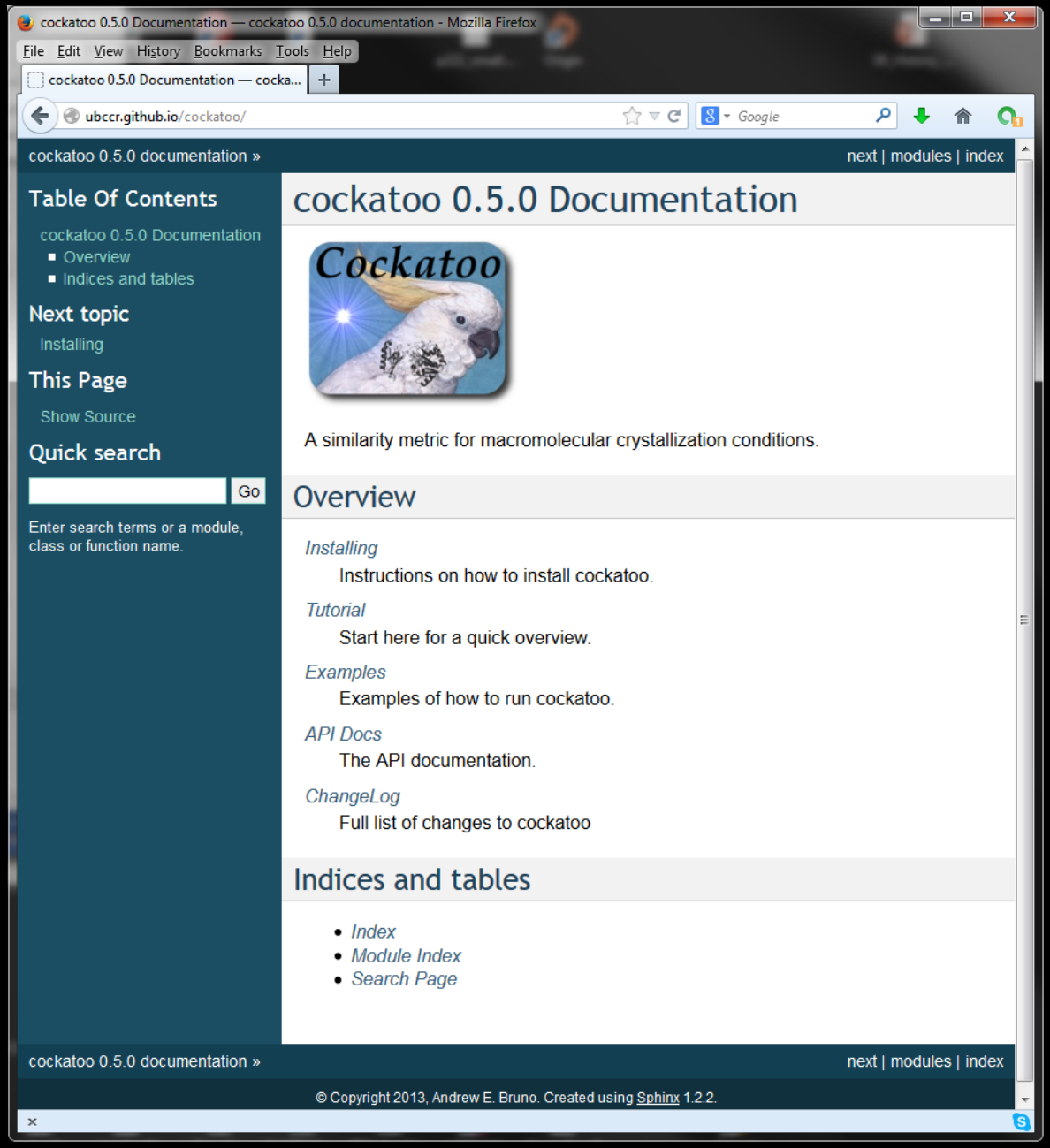
The important point here is not the details of the new information but that this information was obtained after the correct ligands were identified. Potential function and mechanism was revealed. While on could argue that these could have been identified earlier many examples in the PDB have ambiguous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

Other applications

- The code used to evaluate the CD_{coeff} is open source and freely available at http://ubccr.github.io/cockatoo/ or directly from the authors.
- Common chemical trends can be identified for optimization.
- The method can be applied with any crystallization screen, not just ours.
- It can also be used to design a screen where the clustering is equally spaced sampling the widest amount of chemical space with the minimum number of experiments.
- Other fingerprint definitions are available, e.g. activity. The fingerprints can be refined against outcome to determine how chemistry influences crystallization
- Comparing chemistry to outcome: The development of a chemical distance metric, coupled with clustering and hierarchal visualization applied to macromolecular crystallography. Bruno, Ruby, Luft, Grant, Seetharaman, Montelione, Hunt and Snell. PLOS One in press.

Summary

- By building on an existing chemical similarity metric and extending it to include all the components of the cocktail and the additional parameters of stoichiometry and chemical structure cocktails used for crystallization can automatically be clustered.
- The clustering can then be displayed as a hierarchal tree or dendogram.
- Overlying crystallization screening outcome on the dendogram can reveal details in an easily interpretable visible manner that drive further optimization
- The same overlay can also provide biological information that is otherwise missed.
- It can correct information that was missed or provide new information ‘fingerprinting’ the protein.
- It is quick – this analysis can be rapidly run on any result from the HWI screening laboratory.



Documentation, source code, modules and test data is available online at:

http://ubccr.github.io/cockatoo/

The workers and acknowledgements

Hauptman-Woodward Medical Research Institute
Edward Snell, Thomas Grant and Joseph Luft

University of Buffalo Center for Computational Resources
Andrew Bruno, Amanda Bruno and Steve Gallo

North East Structural Genomics (NESG) Rutgers
Guy Montelione and his group for protein samples

NESG at Columbia University
Jayaraman Seetharaman and John Hunt

Support and Funding
NIH, NSF and DoD

Thank you and questions?

esnell@hwi.buffalo.edu