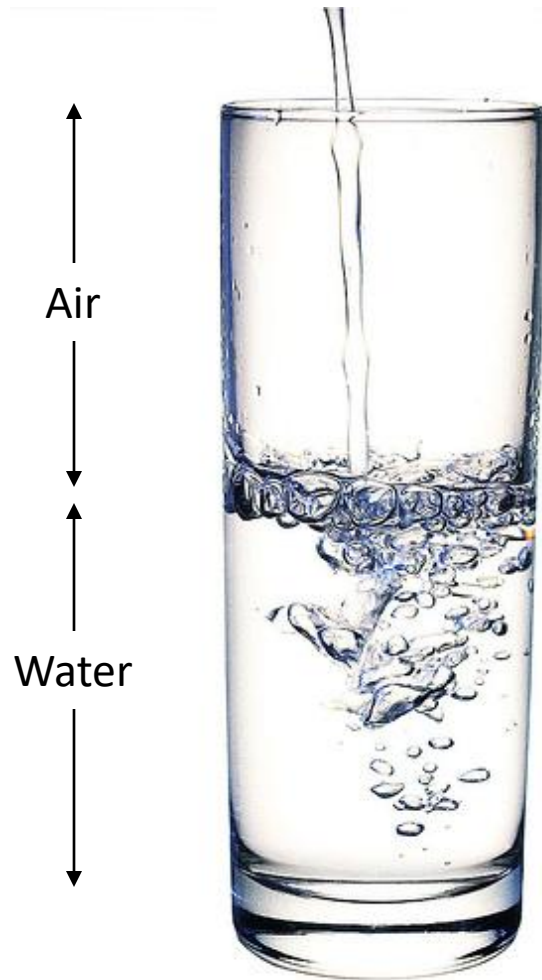


A model is not a structure: Using chemistry and physics to right wrongs and get useful biological information along the way.



Edward H. Snell CEO 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

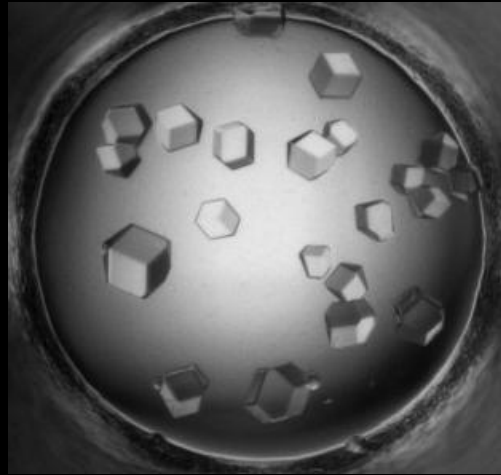
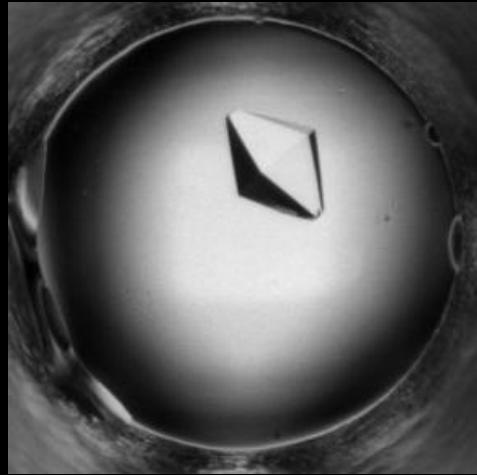
On the need for an international effort to capture, share and use crystallization screening data

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. *Acta Cryst.* (2012). **F68** 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.

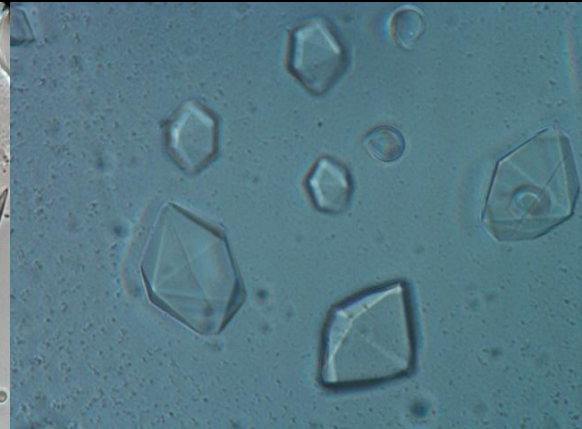
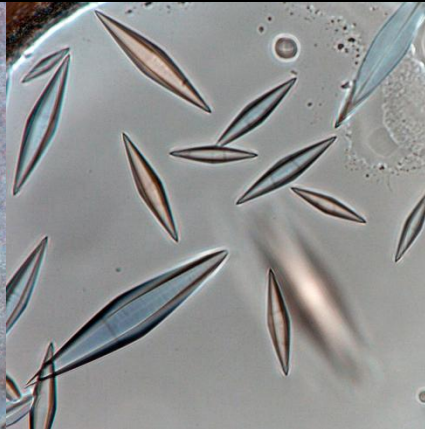
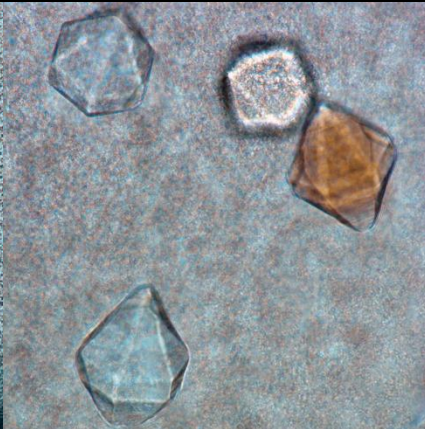
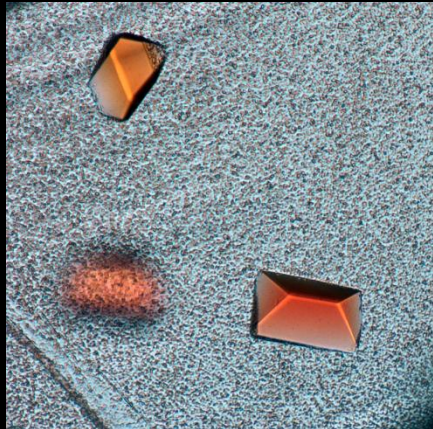
At least 99.8% of crystallization experiments produce an outcome other than crystallization.

Fantasy

Crystallize
Now



Crystallography Requires Crystals



No crystal ...

No crystallography

No crystallographer

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.

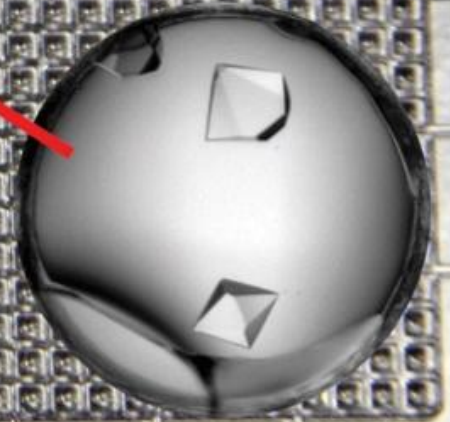
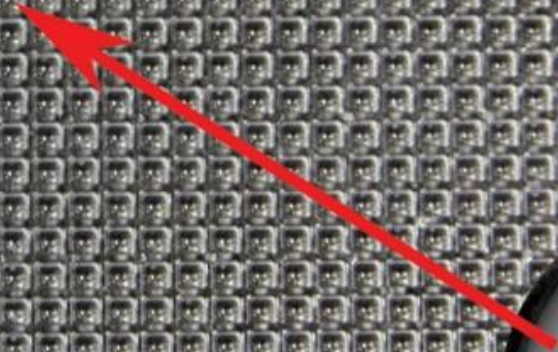
The HWI crystallization cocktail screen.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48																						
Bromide										Chloride										Nitrate										Monobasic and dibasic phosphate										Sulfate										Calcium																			
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96																						
Lithium										Chloride										Acetate										Magnesium										Manganese										Chloride																			
97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144																						
Acetate										Bromide										Carbonate										Chloride										Nitrate										Phosphate										Thiocyanate									
145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192																						
Rubidium										Chloride										Sulfate										Sodium										Iron																													
193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237																									
Zinc										Potassium										Lithium										Potassium										Ammonium										Iron																			
Acetate										Phosphate dibasic										Sulfide heptahydrate										Sulfide monohydrate										Phosphate tribasic										Thiocyanate																			
Highly soluble salt, cation and anion screen																																																																					
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288																						
PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)																													
289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336																						
Sodium										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)																													
337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384																						
Potassium										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)										PEG 20000 20% (v/v)																													
385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432																						
Ammonium										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)																													
433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480																						
Sodium										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)										PEG 8000 20% (v/v)																													
481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528																						
Ammonium										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)																													
529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576																						
Sodium										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)										PEG 8000 40% (v/v)																													
577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624																						
Magnesium										PEG 4000 20% (v/v)										PEG 4000 20% (v/v)										PEG 4000 20% (v/v)										PEG 4000 20% (v/v)																													
625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672																						
Sodium										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)																													
673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720																						
Sodium										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)										PEG 4000 40% (v/v)																													
721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768																						
Sodium										PEG 1000 20% (v/v)										PEG 1000 20% (v/v)										PEG 1000 20% (v/v)										PEG 1000 20% (v/v)																													
769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816																						
Ammonium										PEG 1000 40% (v/v)										PEG 1000 40% (v/v)										PEG 1000 40% (v/v)										PEG 1000 40% (v/v)																													
817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864																						
Sodium										PEG 400 20% (v/v)										PEG 400 20% (v/v)										PEG 400 20% (v/v)										PEG 400 20% (v/v)																													
865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912																						
Sodium										PEG 400 40% (v/v)										PEG 400 40% (v/v)										PEG 400 40% (v/v)										PEG 400 40% (v/v)																													
913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960																						
Magnesium										PEG 400 80% (v/v) 80%										PEG 400 80% (v/v) 80%										PEG 400 80% (v/v) 80%										PEG 400 80% (v/v) 80%																													
961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008																						
Sodium										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%																													
1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023	1024	1025	1026	1027	1028	1029	1030	1031	1032	1033	1034	1035	1036	1037	1038	1039	1040	1041	1042	1043	1044	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056																						
Sodium										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%										PEG 400 20% (v/v) 80%																													

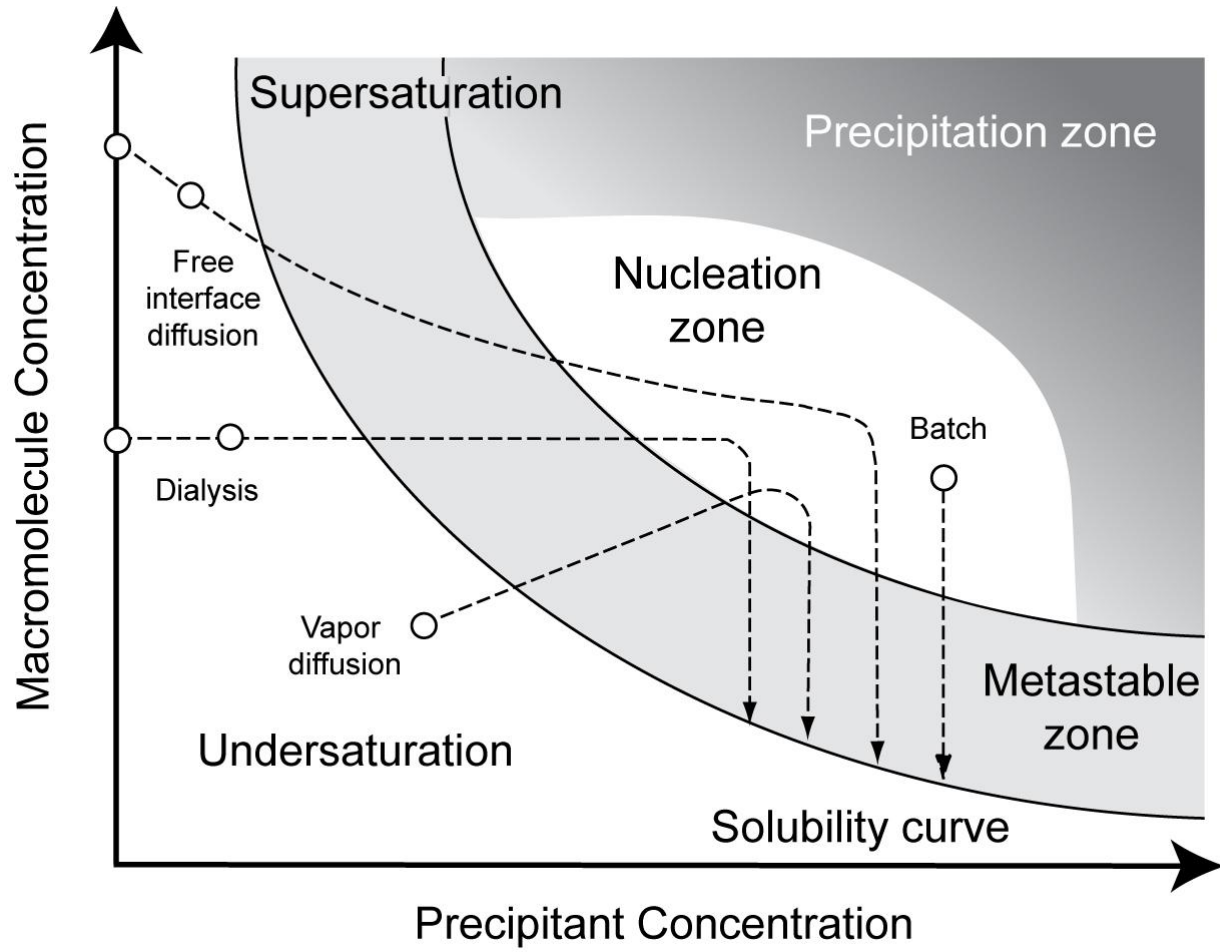
The 1536 diverse chemical cocktails (Luft et al., 2003). The 984 in-house conditions comprise a incomplete factorial sampling of 36 salts, eight buffers, and 5 different PEGs.

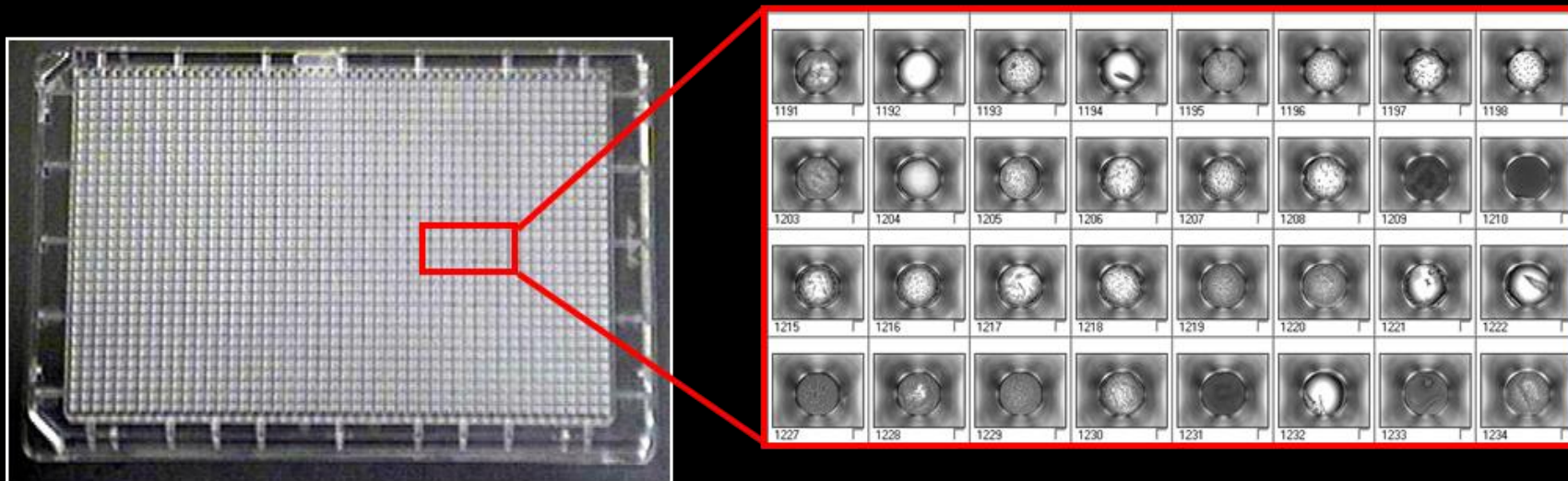
The remainder of 1536 cocktails are comprised of commercial screens available from Hampton Research. Specifically, in order of use; the Matrix Screen, Quick Screen, Nucleic Acid Screen, Sodium Malonate Grid, PEG/Ion, PEG 6000 Grid, Ammonium Sulfate Grid, Sodium Chloride Grid, HT Screen, Index and the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

Minimize sample volume



Simplified phase diagram for crystallization



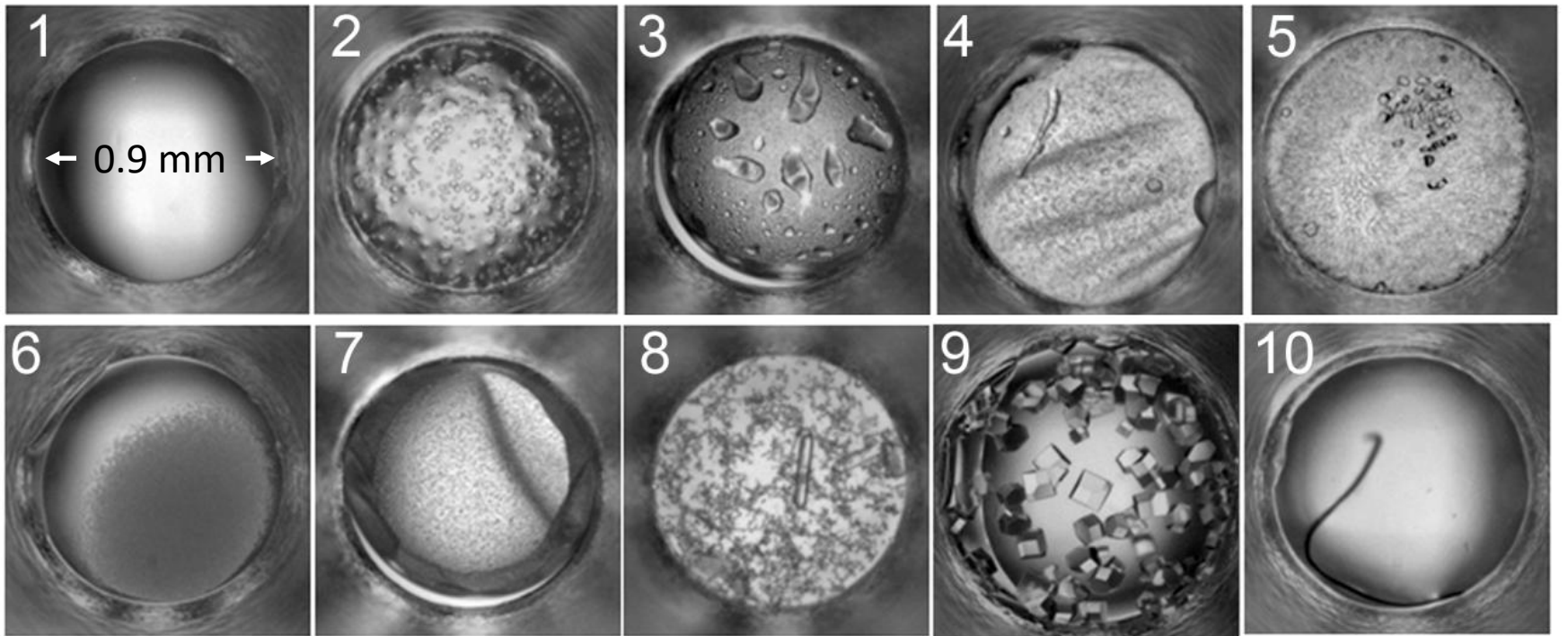


The crystallization method used is micro-batch under oil with 200 nl of protein solution being added to 200 nl of precipitant cocktail in each well of a 1536 well plate.

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

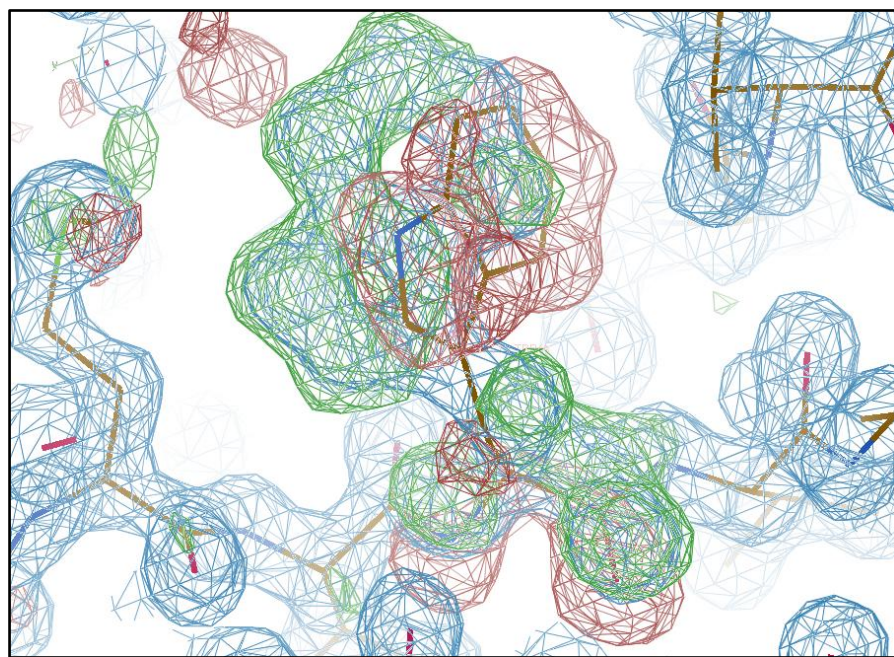
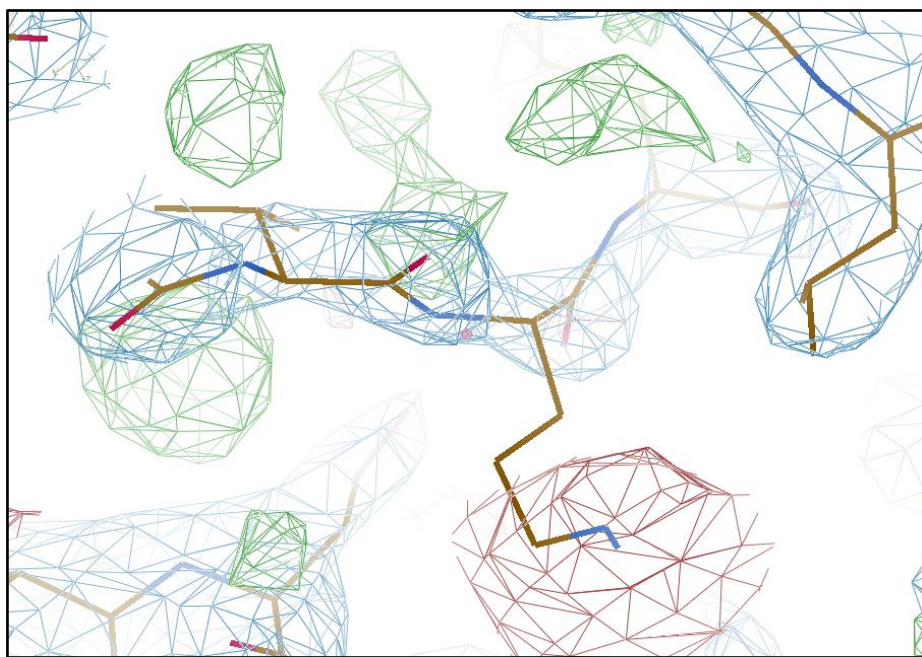
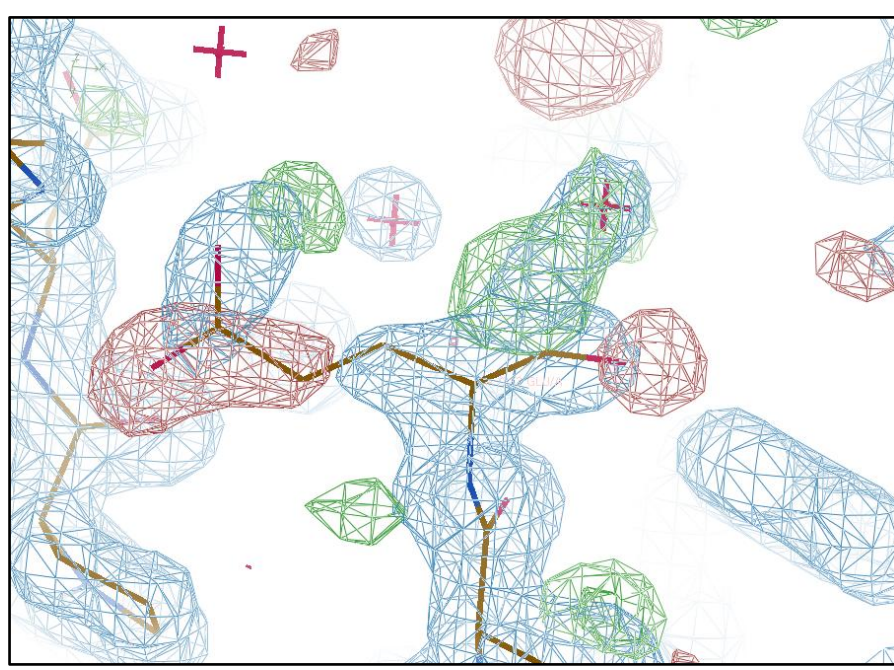
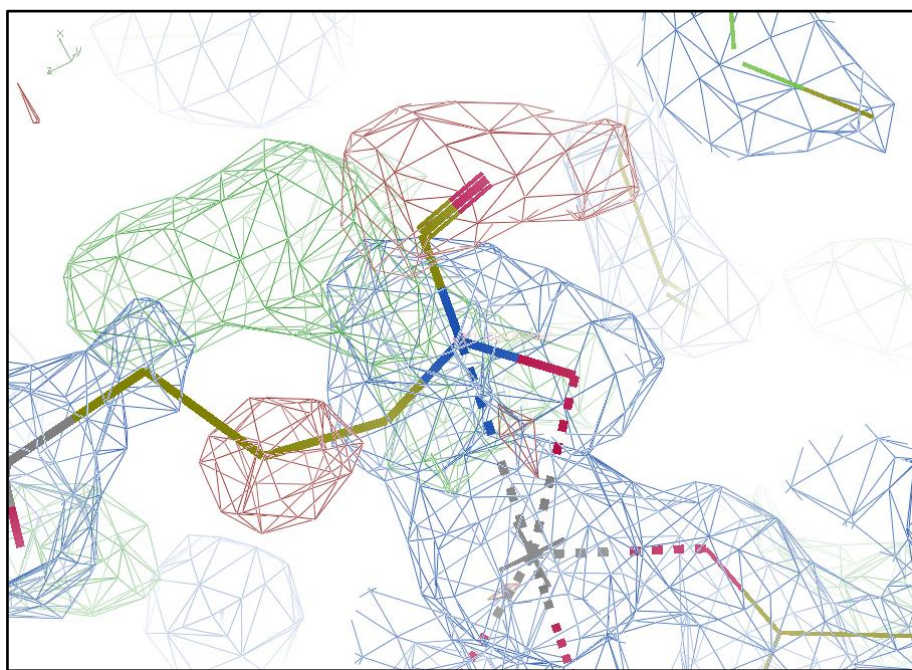


The protein data bank

- The Protein Data Bank contains depositions for 108,607 biological macromolecules.
- Some 90,506 of those are from data derived by X-ray crystallography.
- Simple validation tests are available but a deposition is still accepted even if a test is failed.
- How accurate are the 'structures' in the PDB?

What are the errors, if any?

- Residues have well defined geometries.
- Sequence information is well known.
- Potential problems are:
 - Structural perturbation due to radiation damage
 - Incorrect ligand identification
 - Missing ligands
 - Just generally bad refinement
 - Crystallographic oligomer



How common is the problem?

- More common than you may think
- The examples presented on the previous slide are in the PDB and all come from here ...
- Despite care and diligence, errors still get through
- There are serious problems in many models yet the non-crystallographic community use these as 'structures' on the assumption that the model accurately represents the structure

How can we overcome these problems?

- Structural perturbation due to radiation damage
 - Radiation damage studies, knowledge of the chemical processes and signatures
- Incorrect ligand identification
 - Better ligand treatment during refinement
 - Careful analysis of the crystallization conditions
 - Analysis of the sample pre or post crystallization
- Missing ligands
 - Similar approaches to the above
- Just generally bad refinement
 - To paraphrase Bernard Rupp, sometimes it is worthwhile to look at the map!
- Crystallographic oligomer
 - Solution scattering

How can we overcome these problems?

- Structural perturbation due to radiation damage
 - Radiation damage studies, knowledge of the chemical processes and signatures
- Incorrect ligand identification
 - Better ligand treatment during refinement
 - Careful analysis of the crystallization conditions
 - Analysis of the sample pre or post crystallization
- Missing ligands
 - Similar approaches to the above
- Just generally bad refinement
 - To paraphrase Bernard Rupp, sometimes it is worthwhile to look at the map!
- Crystallographic oligomer
 - Solution scattering

Careful analysis of
crystallization conditions

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 (worked)

Take a real example of two crystallization screening cocktails

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	

1. Convert all component concentrations to molarity. Cocktail C1249 contains 30 % (v/v) of MPD which we must first convert to molarity using the following equation: $molarity = \%v/v * ((density/mw) * 1000)$. Plugging in the values for MPD we get: $2.349 = 0.30 * ((0.9254/118.1742) * 1000)$
2. Compute cocktail fingerprints using the molecular fingerprints for each component and it's molar concentration, as described in the previous section and equation (1). Cocktail fingerprints for C1249 and C0160 are listed below (each component fingerprint was computed using RDKit):

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

3. Compute the Bray-Curtis dissimilarity measure as described in equation (2) from the paper. Using the cocktail fingerprints in step 2 we obtain: $0.97 = \frac{|.1-.2|+|.04-4.48|+|2.349-.4|+|4.698-.1|+|.1-4.48|+|.1-4.48|+46}{|.1+.2|+|.04+4.48|+|2.349+.4|+|4.698+.1|+|.1+4.48|+|.1+4.48|+46}$
4. Compute the pH distance: $0.207 = \frac{|4.6-7.5|}{14}$
5. The final cocktail distance coefficient using $w = \{1, 1\}$ is: $CD_{coeff} = 0.589 = \frac{1}{2}(0.207 + 0.97)$

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 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:

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

Clustering then using
a hierarchal display

The Dissimilarity Measure Over the Whole Screen

Aspects of the screen design are clearly seen

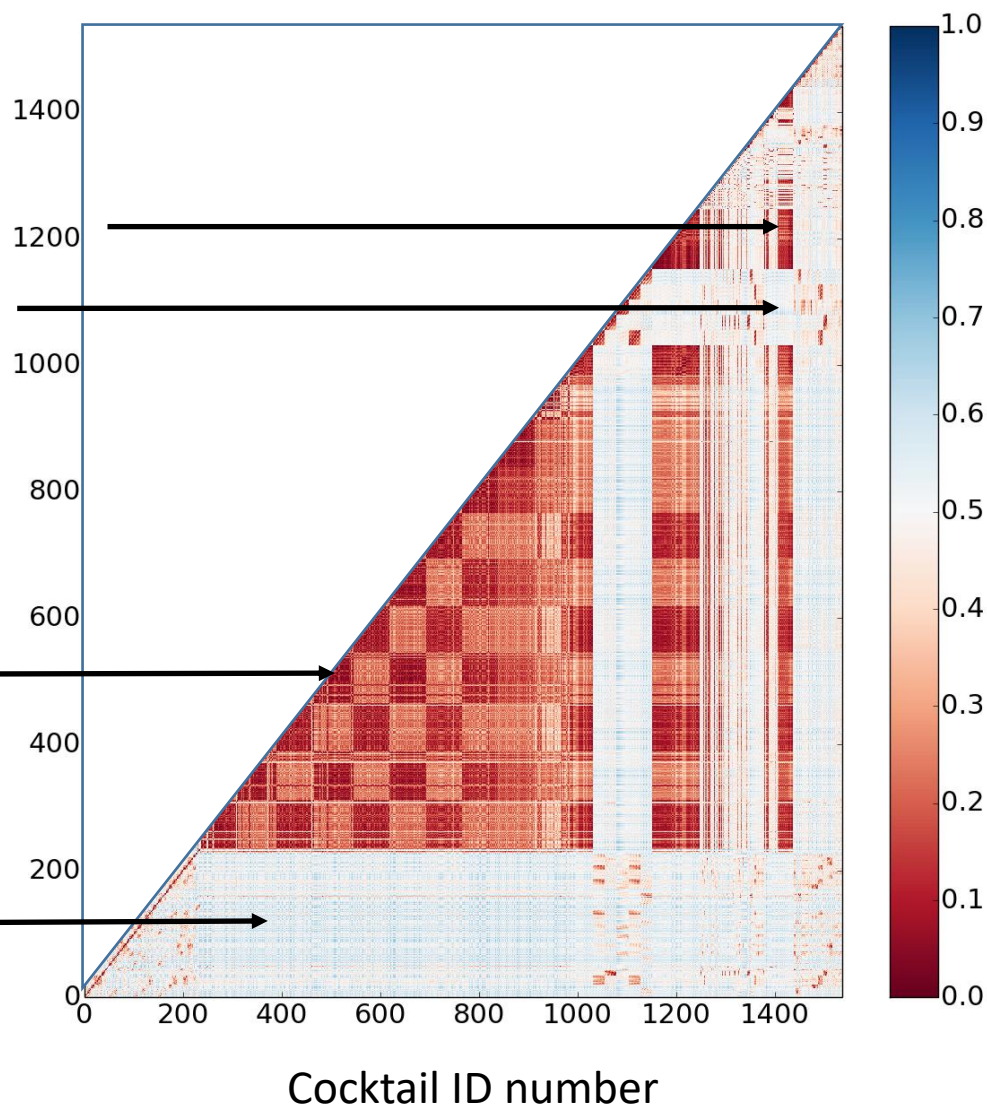
Hampton Research PEG/Ion screen

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

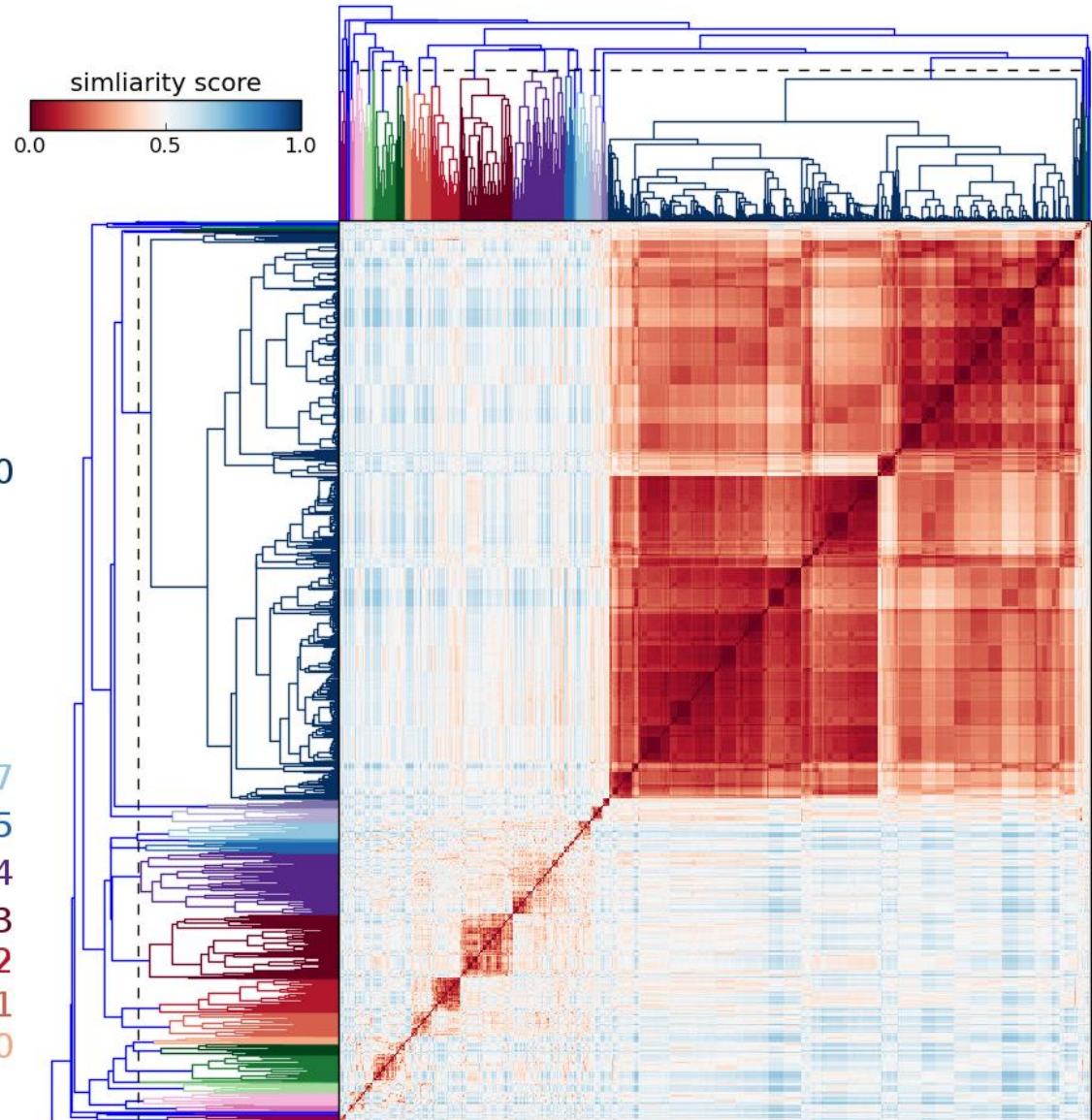


C20

Salts with different anions and cations



C17
C15
C14
C13
C12
C11
C10
C8

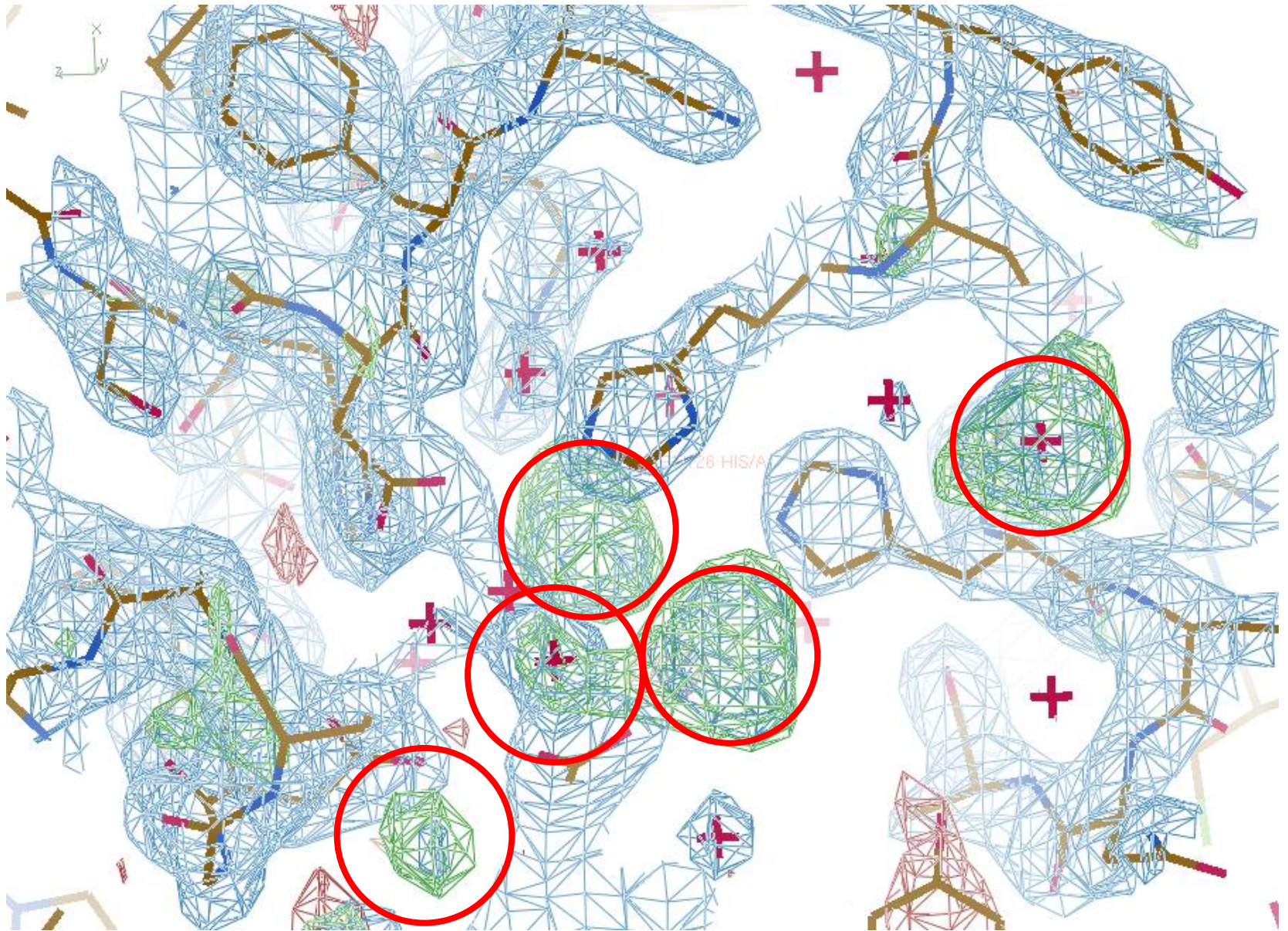


A structural genomics target.

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% NaN₃ 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.



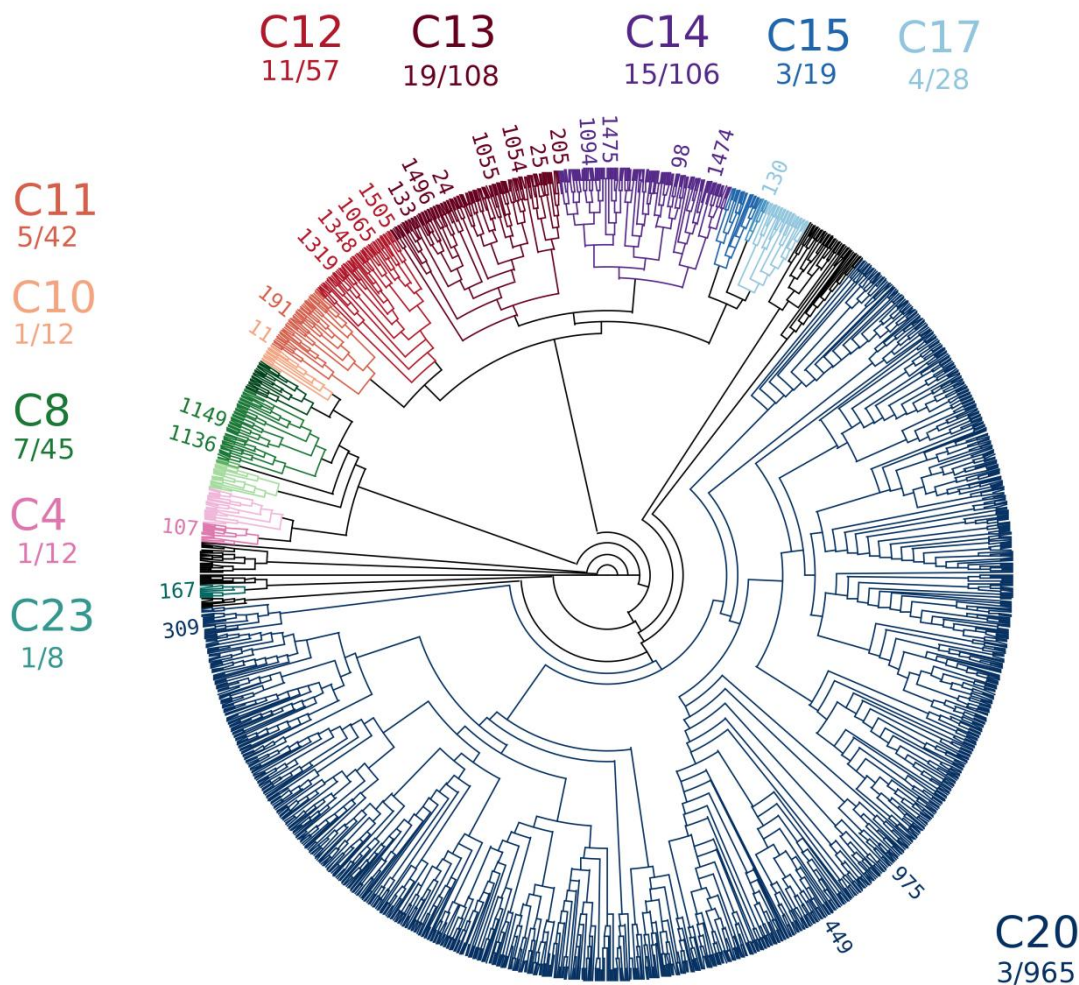
PDB ID 3DMA as deposited in the PDB

Overlaying crystallization data

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 dendrogram. For clarity not all hits are shown

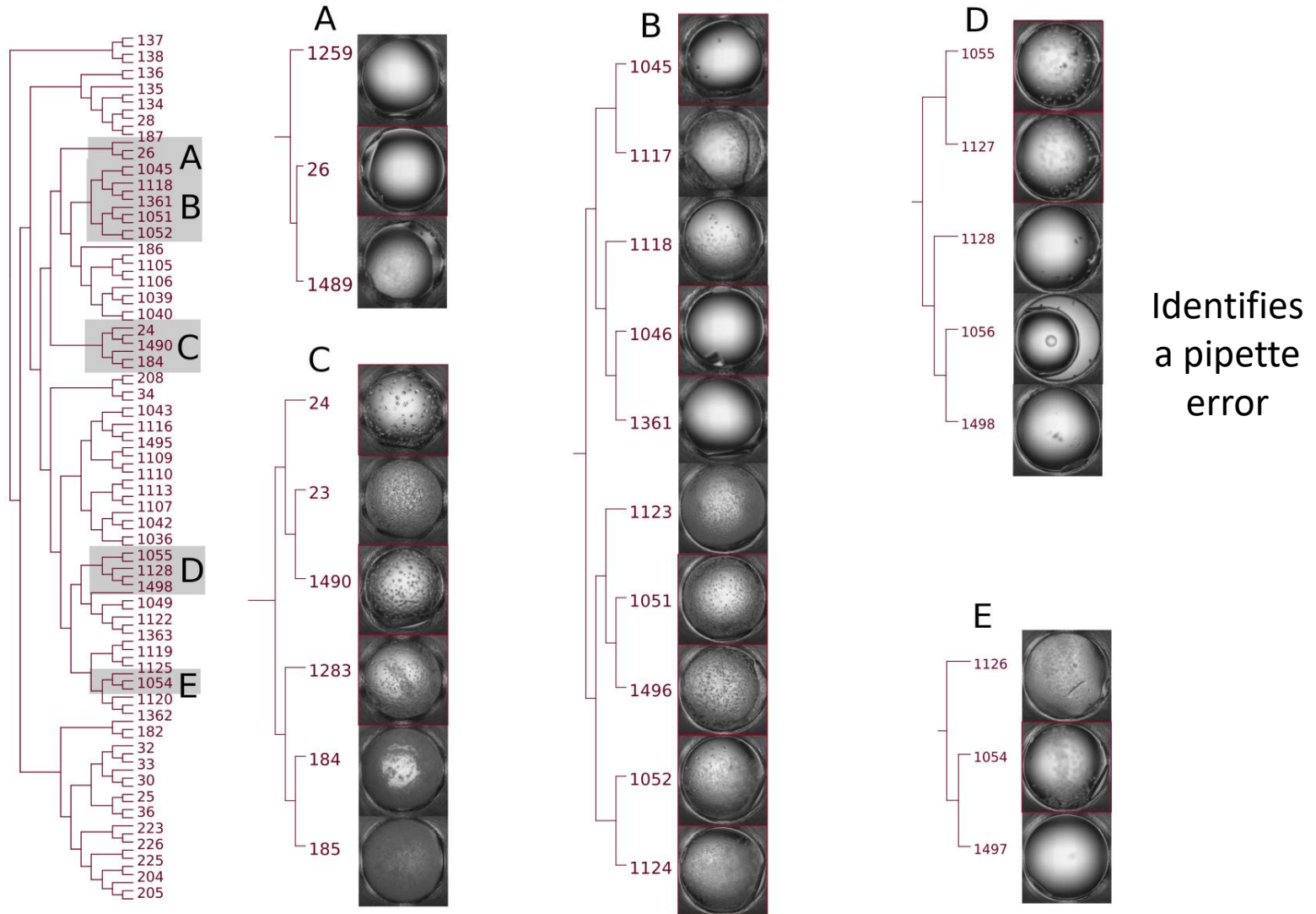


Cluster 20, PEG based, only 3 hits

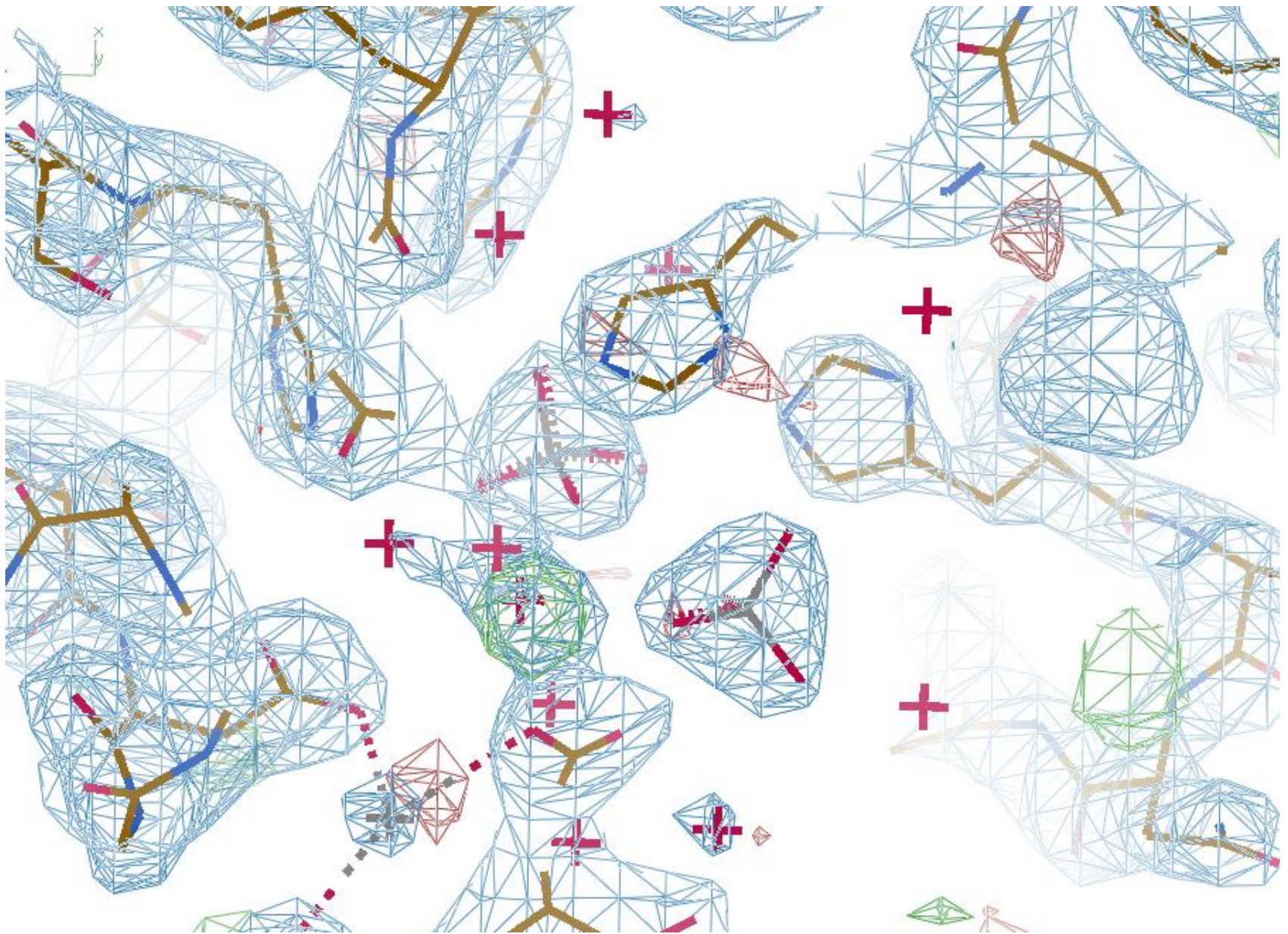
Cluster	Total	Hits	% hits	Sodium %	Potassium %	Phosphate %
All cocktails						
	1536	70	4.5	47	24	16
All crystal						
	70	70	100	70	27	30
Clusters with crystals						
C13	108	19	17.6	73	72	100
C14	106	15	14.2	65	21	0
C12	57	11	19.3	16	2	0
C8	45					
C11	42					
C17	28					
C20	965					
C15	19					
C23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

Zoom in on Cluster 13



Clustering samples the phase diagram



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

Andrew E. Bruno¹, Amanda M. Ruby¹, Joseph R. Luft^{2,3}, Thomas D. Grant², Jayaraman Seetharaman⁴, Gaetano T. Montelione⁵, John F. Hunt⁴, 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 hierarchical 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 Hierarchical 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

Copyright: © 2014 Bruno et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The code used to evaluate the CD_{coeff} is open source and freely available at <http://ubccr.github.io/cocktail/> 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.

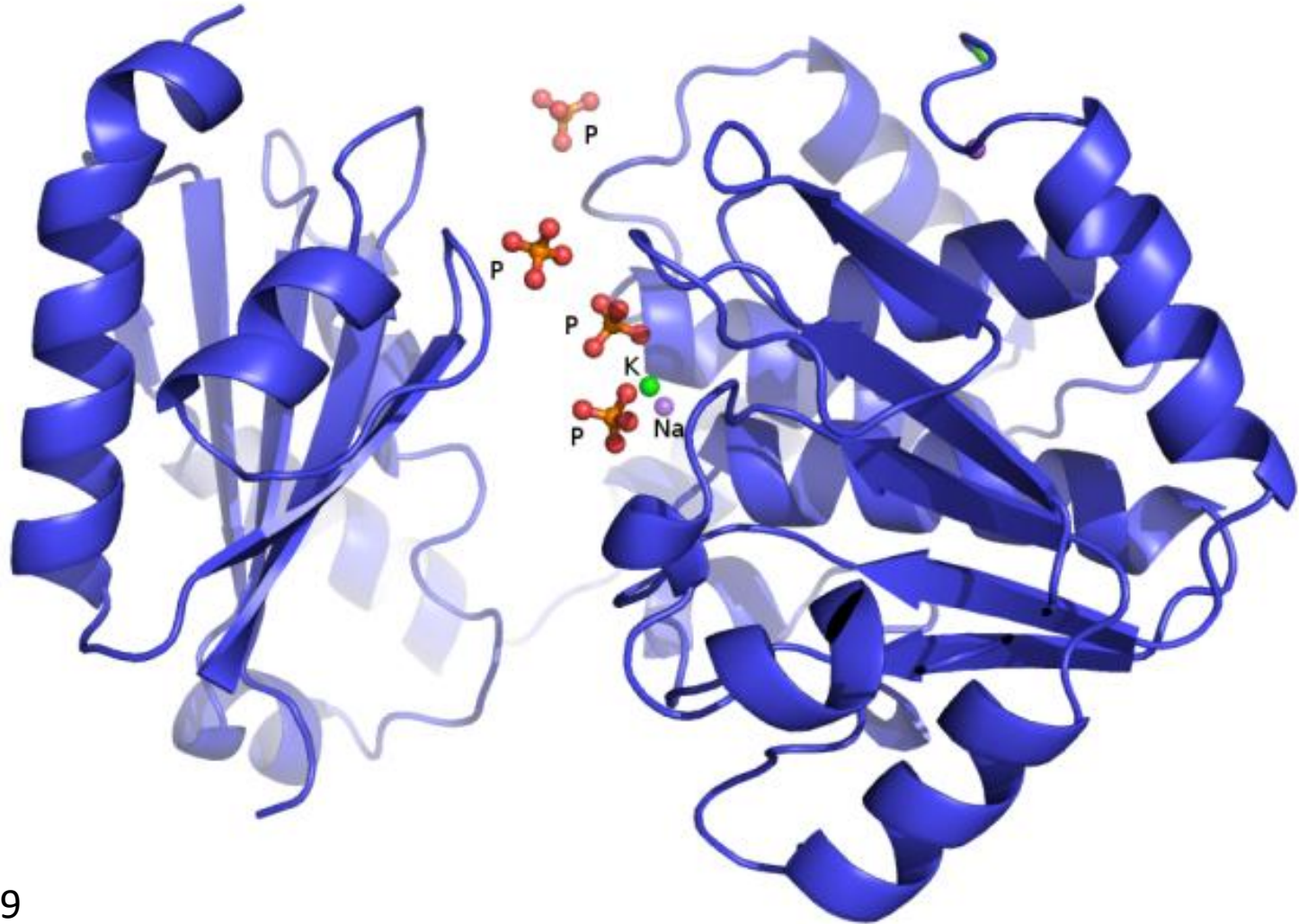
Competing Interests: The authors have declared that no competing interests exist.

* 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

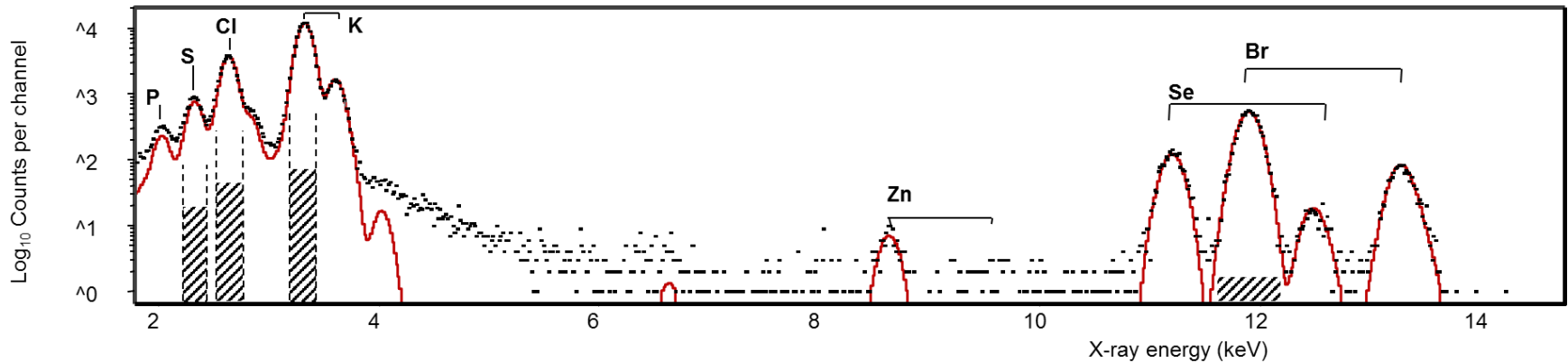
- The structure consists of two domains (N-terminal domain; residues 2 -212 and C-terminal domain residues 217-343) which are connected by a short loop – seen in the initial structure
- The N-terminal domain contains the DHH (Asp224-His225-His226) motif and the C-terminal domain contains a glycine-rich (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 cleft between the two domains
- The phosphate ions interact with the protein backbone
- The location of the phosphate ions might anchor in this pocket.
- The putative active site has features which are involved in binding to the substrate
- The possible roles of the active site residues and polarization of the phosphate for nucleophilic attack.
- The space around the phosphate ions

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 one 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.

Elemental Analysis

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.

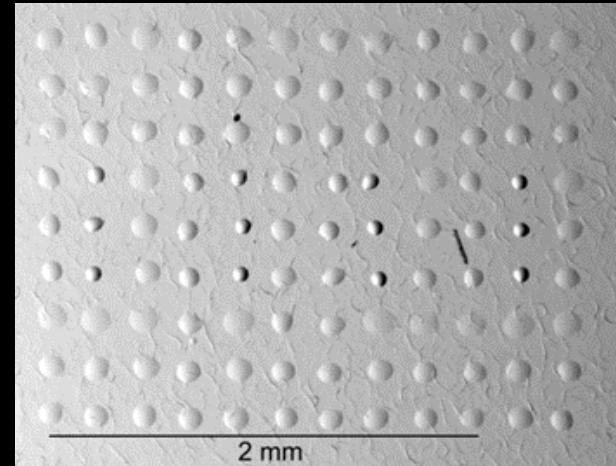


Emission of the characteristic X-rays from a sample can be induced by an incident beam of high energy protons (Particle Induced X-ray Emission: PIXE).

High-throughput Sample Preparation



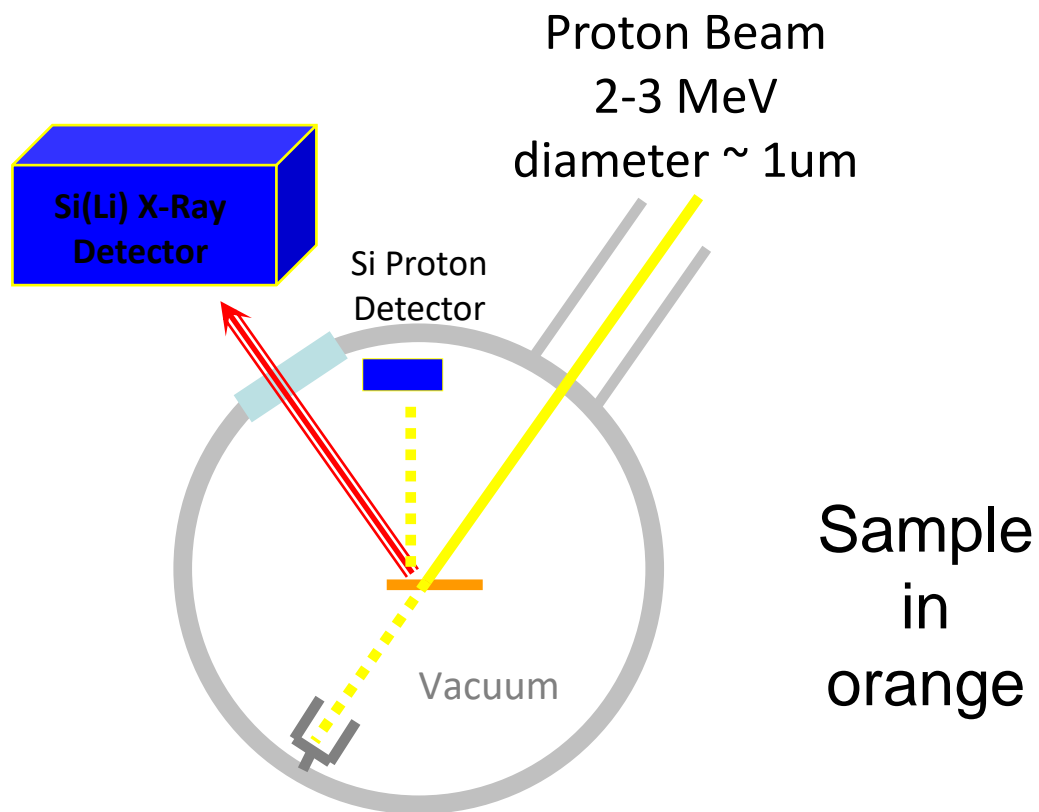
Dispense samples with a non-contact microarray printer

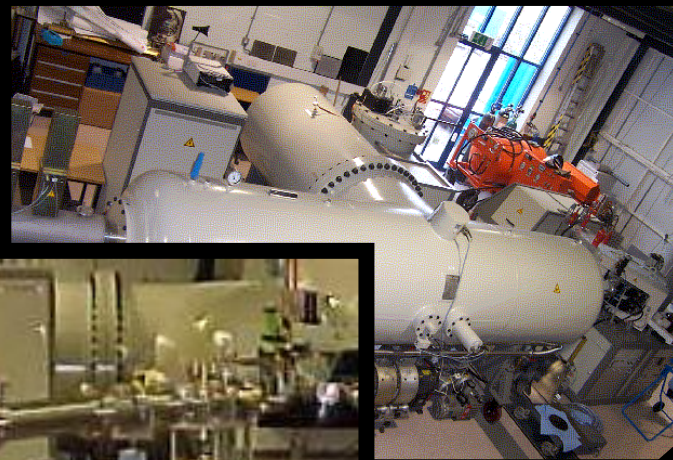


Up to 144 samples dispensed into a 384 well plate and printed into a 12x12 array of 60 μm drops with 200 μm spacing.

Up to five arrays can be mounted into a single sample holder giving a total of 720 samples per slide.

Scanning Proton Microprobe for PIXE analysis. 2-3 MeV protons emerge from the van de Graaff accelerator and are focussed by high precision magnets onto the sample. The whole beamline is kept under vacuum.

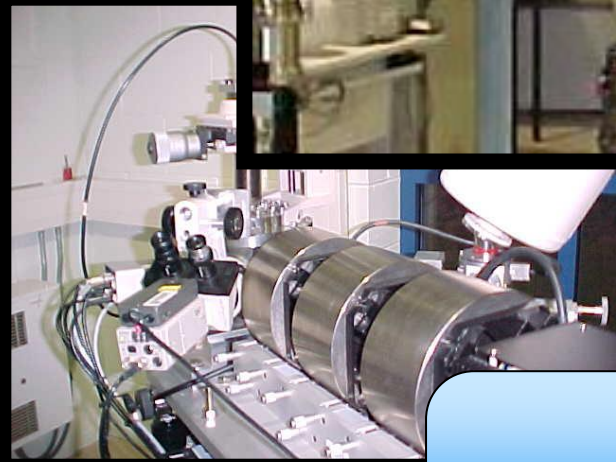




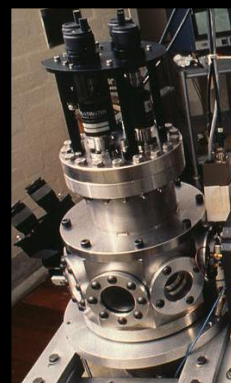
Source & Accelerator



Scanning System



Focusing System



~1 μm diameter beam on target

High-Throughput PIXE

- 34 samples analyzed chosen from NESG samples submitted to the high-throughput crystallization screening laboratory on the basis of a PDB model available and that the model in the PDB contained at least one metal ion.
- 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.

Re-refinement

- 34 samples analyzed chosen on the basis of a PDB structure available and that structure containing at least one metal ion.
- 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.

High-Throughput PIXE

- MicroPIXE can be used to determine the proportion of methionine substitution where no sulfur is present in the buffer.
- The concentration of an element is determined by fitting the area of the X-ray peak corresponding to the element.
- If the total number of Se atoms per protein molecule is α_{Se} , the total number of S atoms left per protein molecule is α_S , and the original number of S atoms (cysteines + methionines) in the sequence was α then $\alpha = \alpha_S + \alpha_{Se}$ and we can write:
$$\frac{\alpha_S}{\alpha_{Se}} = \frac{c_S A_{Se} (\alpha - \alpha_{Se})}{c_{Se} A_S \alpha_{Se}}$$
- Where A_S and A_{Se} are the atomic masses of S and Se respectively and c_S and c_{Se} are the mass concentrations.

High-Throughput PIXE

- In our case the NESG buffer has Sulfur.
- However, all the proteins studied were expressed with SeMet for phasing purposes.
- The number of atoms of element Z per protein can be determined by

$$\alpha_Z = \frac{c_Z}{c_{Se}} \frac{A_{Se}}{A_Z} \alpha_{Se}$$


- Where A_Z and A_{Se} are the atomic masses of element Z and Se respectively and c_Z and c_{Se} are the mass concentrations determined from the PIXE spectrum.

	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
PDB inconsistent with PIXE							
1	3LV4	BiR14	456	Ca	-	Ca, Mn	18% PEG 3350, 0.2M Ca acetate, 0.1M MES, pH 6.15
2	3HIX	NsR437I	106	Mn	-	-	20% PEG 4000, 0.1M Mn chloride, 0.1M MES, pH 6.0
3	3HLY	SnR135D	161	Ca	-	Ca	20% PEG 8000, 0.1M Ca acetate, 0.1M MES, pH 6.0
4	3DCP	LmR141	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	3JSR	NsR236	119	K	-	Ca	8.64 M K acetate, 0.1 M TAPS, pH 9.0
6	3ILM	NsR437H	141	Mn	-	Fe, Co	20% PEG 1000, 0.1M Mn chloride, 0.1M MES, pH 6.0
7	3I24	SoR237	137	Na	Co (0.7), Zn (0.7)	Fe, Ni	NaCl 200 mM, MES PH6, PEG 3350 20%, pH 6.15
8	3GGL	BtR324A	169	Zn	-	Ca, Mn, Fe*	0.75M Mg Formate, 0.1M Bis-Tris, pH 7.0
9	3KB1	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	3LMC	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	3K2Q	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	3LM8	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	3E5Z	DrR130	296	Mg♦	Ca*	-	0.1 M NaCl, 0.1 M TAPS (listed as "TOPS"-no such thing), pH 9.0, 18% PEG 3350, MgCl ₂ (listed as "MgL2") – no concentration given
5	3HNM	BtR319D	172	Mg♦	Ca (1.74)	-	None given
6	3DEV	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	3IHK	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	3KB4	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	3E48	ZR319	289	Mg♦	-	Ca, Fe, Cu	0.1 M Tris-HCl, pH 9.1, 18% PEG 3350, 0.1 M MgSO ₄

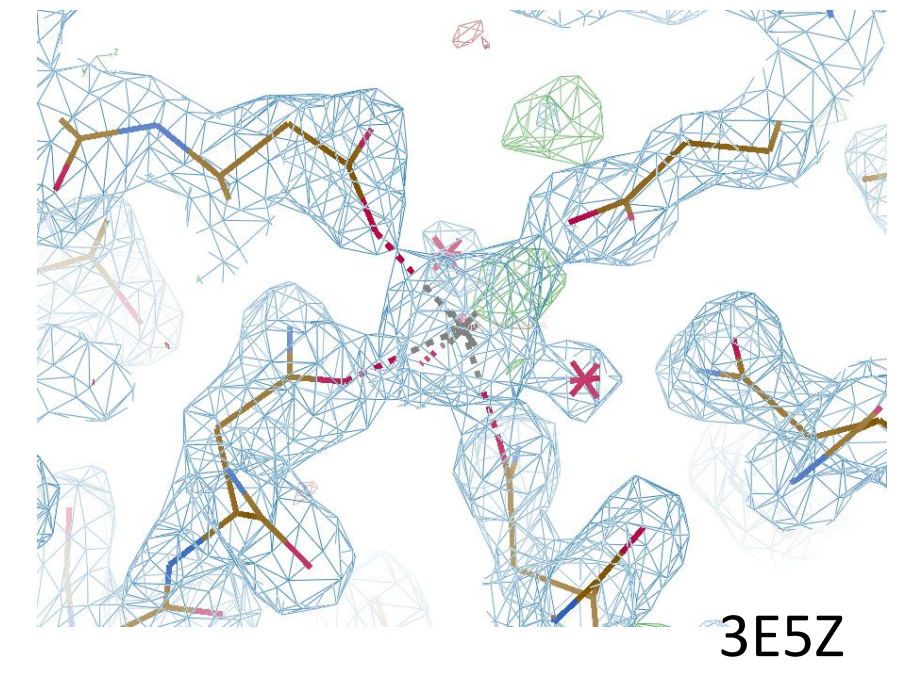
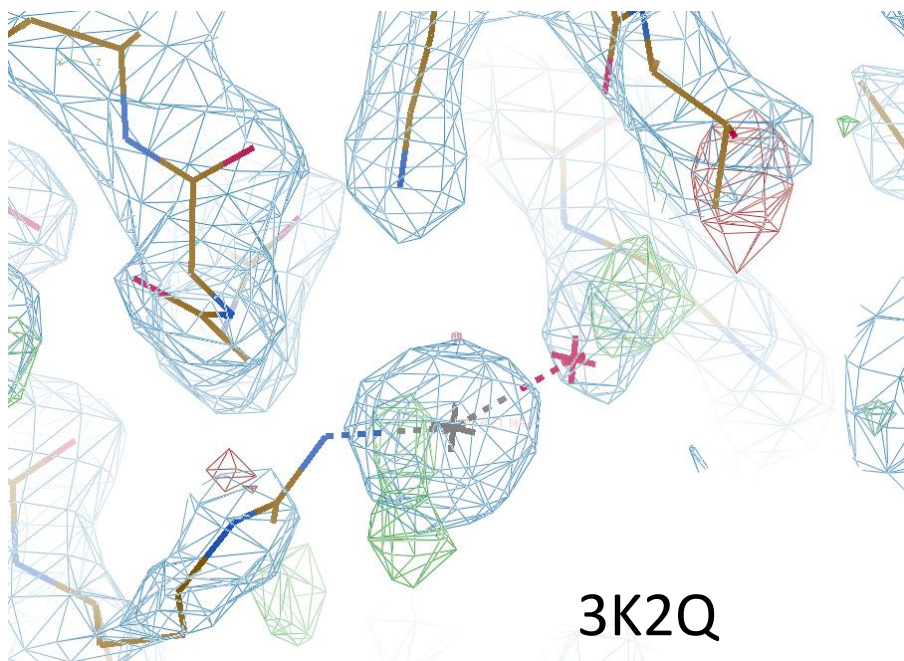
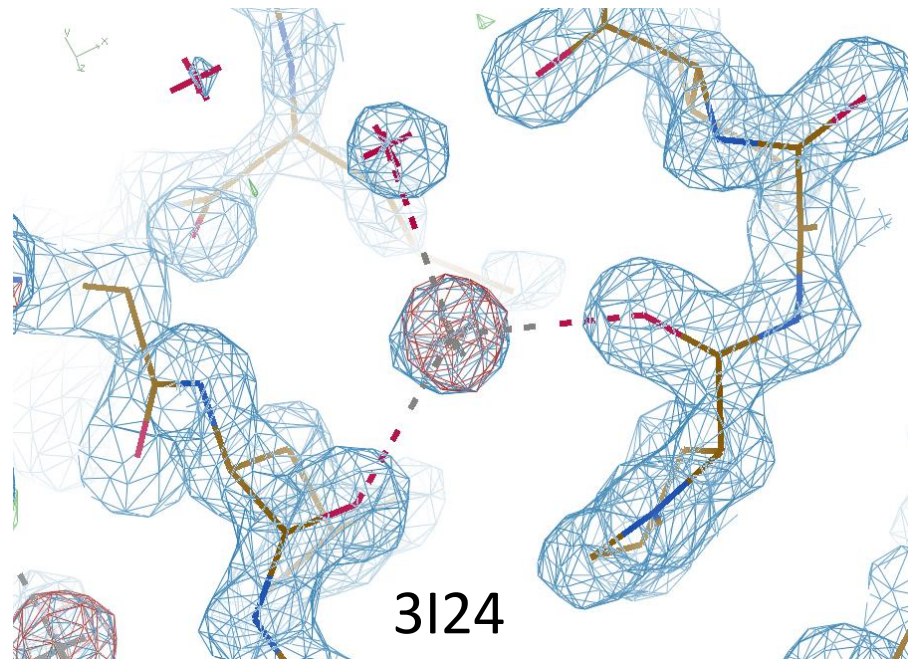
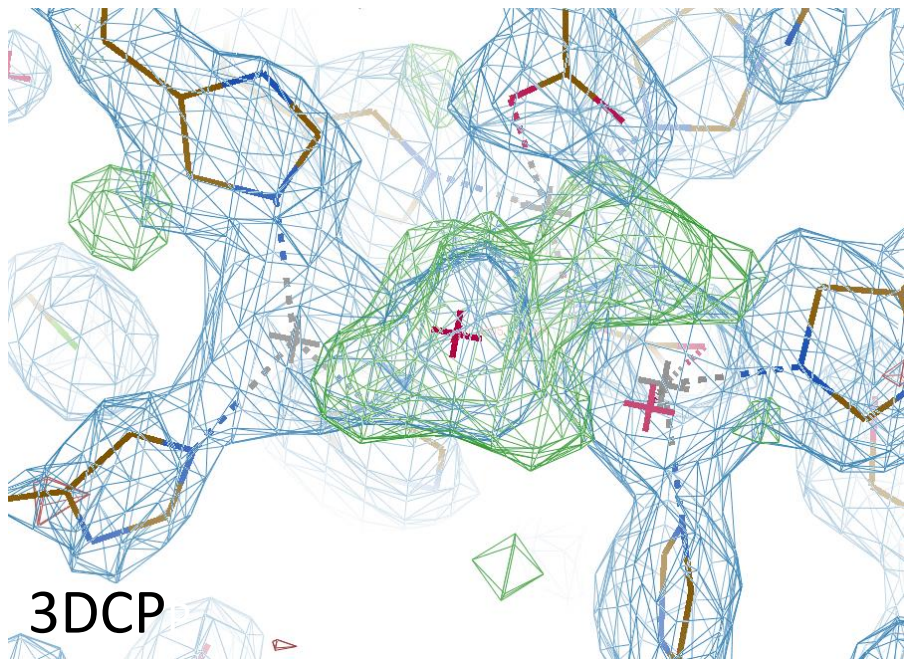
 Model in the PDB containing an extra misidentified metal

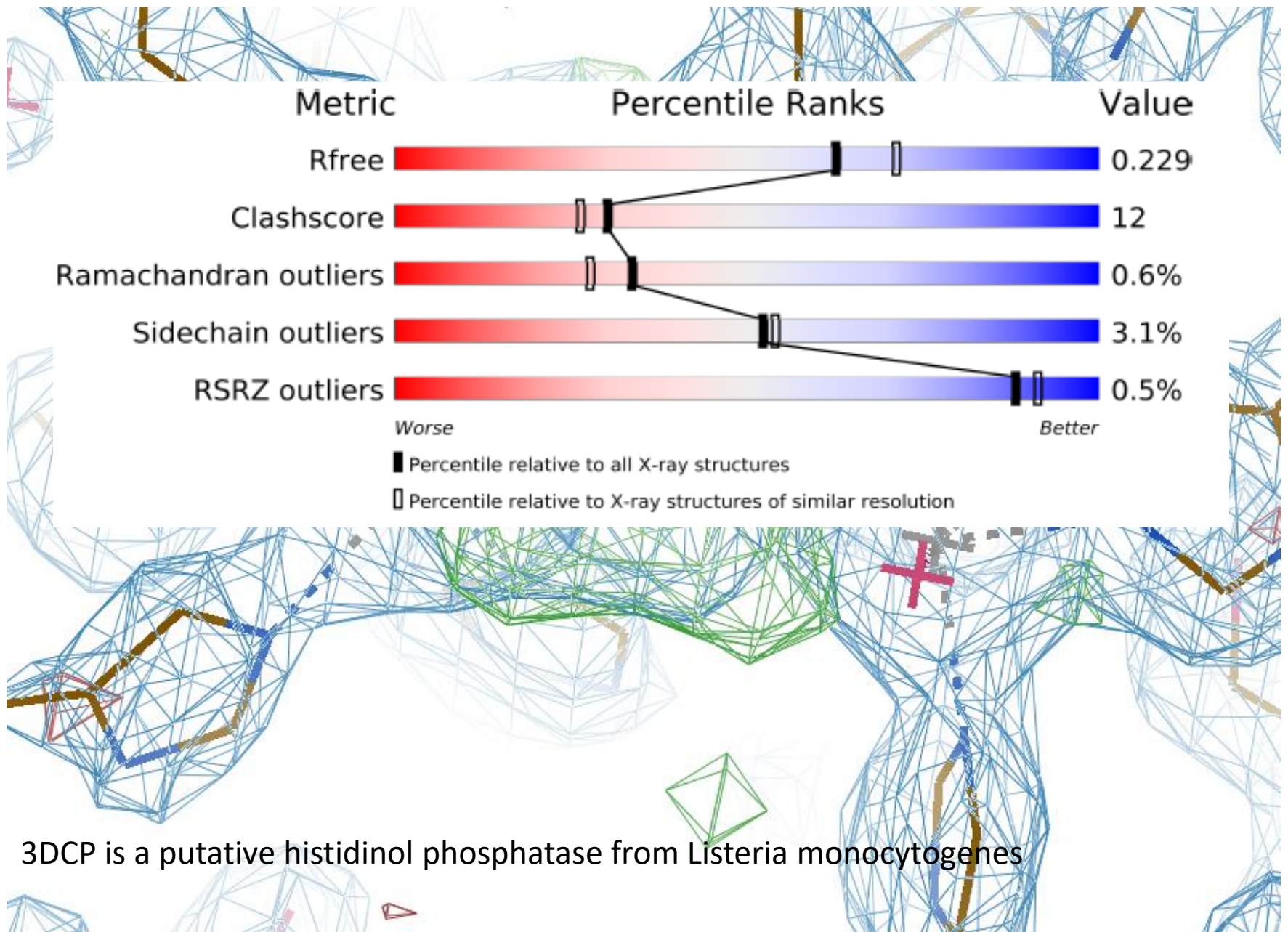
	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
PIXE data consistent with PDB							
1	3NNG	BfR258E	168	Ca	Ca (1.7)	Fe	40% PEG 4000, 0.1 M CaCl ₂ , 0.1 M Bis-Tris Propane, pH 7.0
2	2KPN	BcR147A	103	Ca	Ca (0.8)		NMR
3	3LRQ	HR4604 D	100	Zn	Zn (2.5), Fe (0.3)	Ca, Co, Cu	0.1% (w/v) MPD, 0.1% (w/v) 1,2,3-heptanetriol, 0.1% (w/v) diethylenetriaminepentakis (methylphosphonic acid), 0.1% (w/v) D-sorbitol, 0.1% (w/v) glycerol, 0.06 M HEPES, 12.5% PEG 3350
4	3NNQ	OR3	114	Zn	Ca, Zn*	Fe, Ni*	2.0 M Na ₂ C ₃ H ₂ O ₄ , 0.1 M NaC ₂ H ₃ O ₂ , pH 5.0, 0.05% Anapoe X-305
5	N/A	LkR105	290	-	Fe (0.04)	Ca, Cu	N/A
6	2K52	MjR117 B	80	-	Ca (0.2)	Fe	NMR
7	3ESI	EwR179	129	-	-	Ca, Fe	PEG 4000 (no concentration given), 0.2 M NH ₄ C ₂ H ₃ O ₂ , 0.1 M Na citrate, pH 5.6
8	3DM3	MjR118E	105	Na [◆]	-	-	0.1 M Na Citrate, 0.1 M NaCl, pH 5.0
9	3I24	VfR176	149	Na [◆]	-	Co	NaCl (no concentration given), 0.2 M MES, pH 6.0, 20% PEG 3350, pH 6.15
10	3L8M	SyR86	212	Na [◆]	-	Fe	RbCl (no concentration given), 0.1 M NaCitrate, pH 4.2
11	3FOJ	SyR101A	100	Na [◆]	-	Ca, Fe, Cu	0.15 M MgSO ₄ , 0.1 M Na Citrate, 20% PEG 3350
12	4EVW	VcR193	255	Mg [◆]	-	-	40-44% MPD, 0.1 M HEPES, pH 7.5
13	2KW4	DhR1A	147	Mg [◆]	-	Ca, Fe*	NMR
14	3DJB	BuR114	223	Mg [◆]	-	Fe, Ni	0.1 M HEPES, pH 7.5, 40% PEG 1000, 0.1 M KNO ₃

	PDB ID	Gene	Residues	Metal in PDB	Metals in PIXE (>3xLOD)	Potential metals in PIXE (1-3xLOD)	Crystallization conditions
Sample too dilute for PIXE (no Se signal)							
1	3D3N	LpR108	284	Ca	-	K, Mn	0.1 M HEPES, pH 7.5, 5% PEG 8000, 0.1 M Ca(C ₂ H ₃ O ₂) ₂
2	3DC7	LpR109	232	Mg/Na [◆]	-	-	0.1 M MgSO ₄ , 0.1 M Bis-Tris, pH 5.5, 16% PEG 8000

◆ Presence of sodium and magnesium could not be confirmed at the proton energies used in these experiments. *Selenium signal was below 3 times the limit of detection, so accurate stoichiometries could not be established.

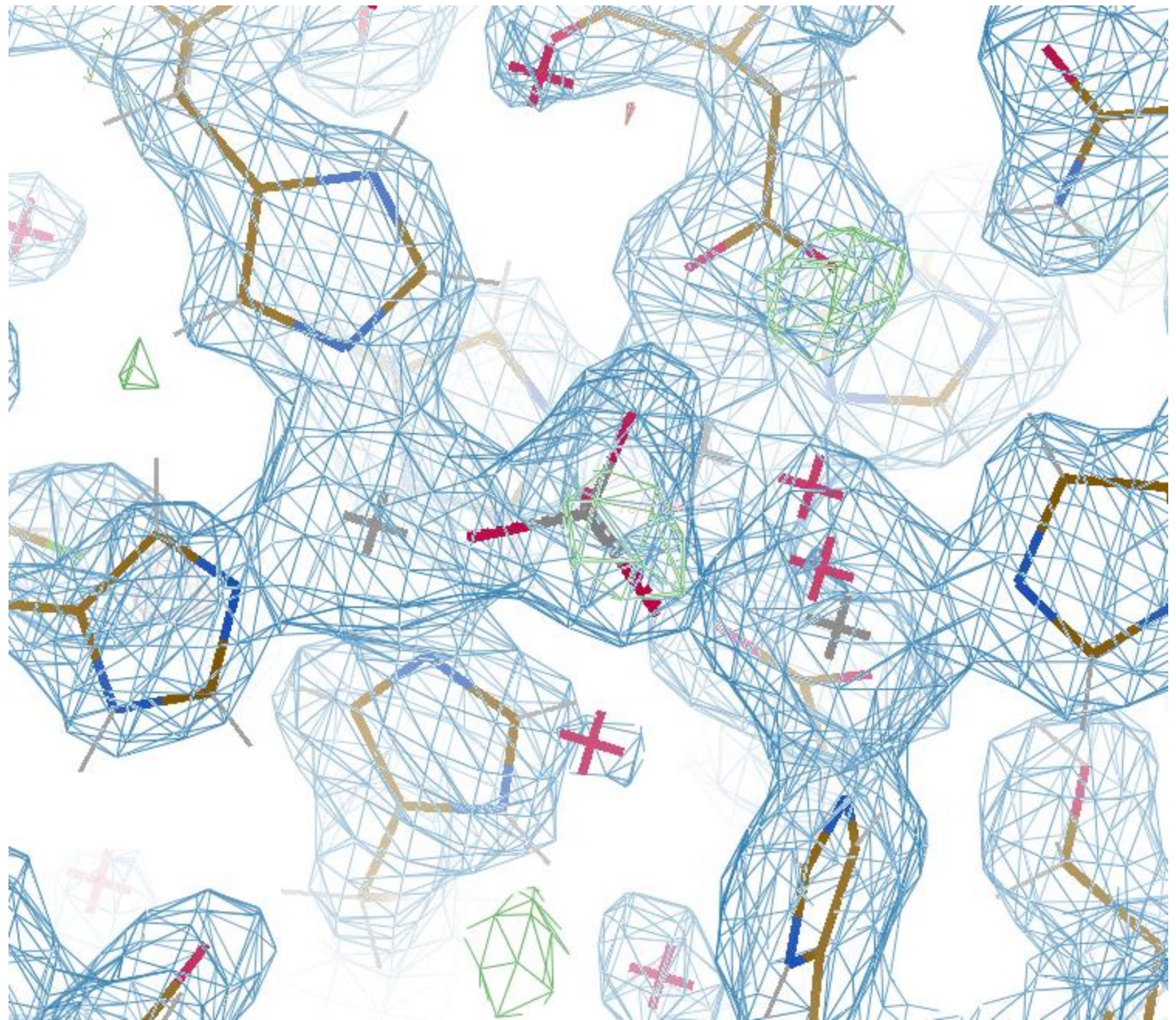
- Of the 34 samples analyzed, 9 were inconsistent with the PDB results, 9 had extra metals present, 18 were consistent, and 2 were unsuitable for analysis due to low protein concentration on the sample.
- In total, 18 of the 32 analyzable samples (56%) were not correctly or fully described in the PDB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

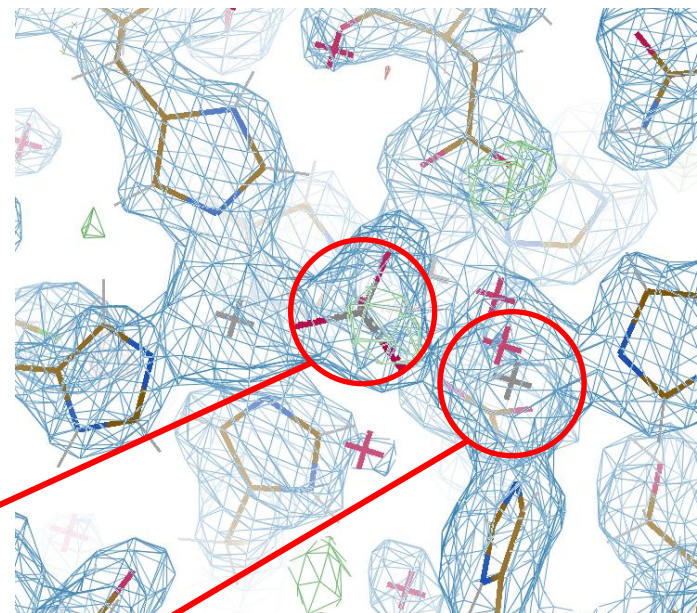
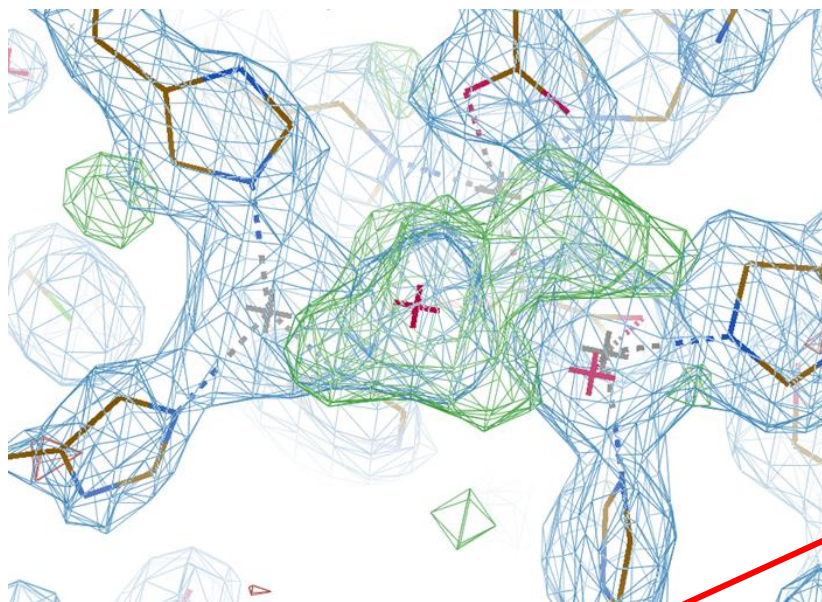




Re-Refining 3DCP

	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							
	18.08	21.111	0.003	0.707	1.1	97.3	0	0





PO₄

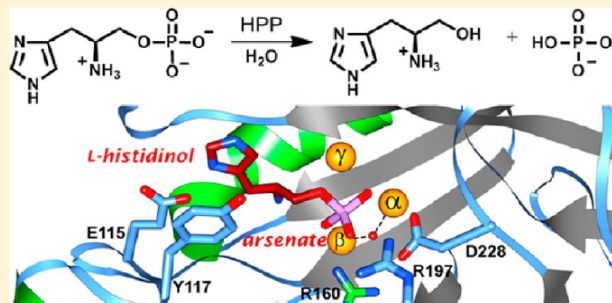
Fe

Metals in new structure, Fe, Mn, Co cluster

... site and had a catalytic efficiency of $\sim 10^3 \text{ M}^{-1} \text{ s}^{-1}$. Expression of the protein under iron-free conditions resulted in the production of an enzyme with a 2 order of magnitude improvement in catalytic efficiency and a mixture of zinc and manganese in the active site. Solvent isotope and viscosity effects demonstrated that proton transfer steps and product dissociation steps are not rate-limiting. X-ray structures of HPP were determined with sulfate, L-histidinol phosphate, and a complex of L-histidinol and arsenate bound in the active site. These crystal structures and the catalytic properties of variants were used to identify the structural elements required for catalysis and substrate recognition by the HPP family of enzymes within the amidohydrolase superfamily.

Supporting Information

ABSTRACT: L-Histidinol phosphate phosphatase (HPP) catalyzes the hydrolysis of L-histidinol phosphate to L-histidinol and inorganic phosphate, the penultimate step in the biosynthesis of L-histidine. HPP from the polymerase and histidinol phosphatase (PHP) family of proteins possesses a trinuclear active site and a distorted $(\beta/\alpha)_7$ -barrel protein fold. This group of enzymes is closely related to the amidohydrolase superfamily of enzymes. The mechanism of phosphomonoester bond hydrolysis by the PHP family of HPP enzymes was addressed. Recombinant HPP from *Lactococcus lactis* subsp. *lactis* that was expressed in *Escherichia coli* contained a mixture of iron and zinc in the active site and had a catalytic efficiency of $\sim 10^3 \text{ M}^{-1} \text{ s}^{-1}$. Expression of the protein under iron-free conditions resulted in the production of an enzyme with a 2 order of magnitude improvement in catalytic efficiency and a mixture of zinc and manganese in the active site. Solvent isotope and viscosity effects demonstrated that proton transfer steps and product dissociation steps are not rate-limiting. X-ray structures of HPP were determined with sulfate, L-histidinol phosphate, and a complex of L-histidinol and arsenate bound in the active site. These crystal structures and the catalytic properties of variants were used to identify the structural elements required for catalysis and substrate recognition by the HPP family of enzymes within the amidohydrolase superfamily.



Metal content
measured with
an inductively
coupled mass
spectrometer

Accurate Metal identification is important

- The original structure contained Fe and Zn.
- The revised structure shows the phosphate and Co
- The phosphate and tri-nuclear metal center are important for mechanism

Work in progress

- All the structures in the table are currently being re-refined
- Each is improved with the correct metal placed
- All will be revisited once completed to determine if there are any 'clues' to mechanism with the correct metal in place.

Important notes about the technique

- Because PIXE is an elemental analysis the sample does not have to be in any preserved state.
- Samples from years ago can be used to collect experimental data.
- The number and ratio of different metals (or other atoms) per protein molecule can be determined.
- Not discussed today, but the data reveals clear signatures in protein models that identify suspect metals.

Summary


- Crystallization analysis and elemental analysis have great potential in improving structural models.
- This improvement is needed as our limited study shows a greater than 50% error rate.
- Experimentally identifying errors defines signatures of those same errors in other structural models.
- The work leads to a potential quality control mechanism to identify suspect structural models.
- It also allows native metals (at least from expression) to be distinguished from opportune ones.

The Team

Andrew Bruno, Elspeth Garman, Geoffrey Grime,
Joseph Luft, Amanda Ruby, Edward Snell,
Elizabeth Snell, and Oliver Zeldin



Special thanks to the 'Pixie'



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Progress in Biophysics and Molecular Biology 89 (2005) 173–205
www.elsevier.com/locate/pbiomolbio

Progress in
Biophysics
& Molecular
Biology

Review

Elemental analysis of proteins by microPIXE

Elsbeth F. Garman^{a,*}, Geoffrey W. Grime^b

^aLaboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK
^bDepartment of Physics, University of Surrey, Guildford GU2 7XH, Surrey, UK

Available online 5 November 2004

Thank you and questions?



esnell@hwi.buffalo.edu