# Precipitation Pattern as a Proxy for Protein Crystallization

Jia Tsing Ng

Lady Margaret Hall
University of Oxford

Supervisors:

Prof. Frank von Delft (University of Oxford)
Prof. Michael Osborne (University of Oxford)
Dr Carien Dekker (Novartis Institutes for Biomedical Research)
Dr Markus Kroemer (Novartis Institutes for Biomedical Research)

A thesis submitted for the degree of
Doctor of Philosophy in Systems Approaches to Biomedical Science

Trinity Term 2015

Precipitation Pattern as a Proxy for Protein Crystallization
Jia Tsing Ng, Lady Margaret Hall
Submitted for the degree of DPhil. Systems Approaches in Biomedical Science

# Abstract

X-ray crystallography has been the workhorse behind most 3D protein structures, which are crucial in the understanding of their biological function and interaction with other molecules. However, a major rate-limiting step in X-ray crystallography remains obtaining suitable protein crystals. The best approach to crystallise a protein can be summarized as a random-screen-and-wait procedure, with little to no readout from experiments that do not produce crystals. This project aims to make the most out of present screening practice, by establishing objective analyses of sparse-matrix screening experiments that extract informative readouts from this standard front-line experiment, independent of whether it yields visible crystals or not.

We have developed methods to objectively characterize crystallization outcome based on image analysis, enabling several things. Firstly, the ranking of droplets based on their likelihood of crystallinity to increase the efficiency and accuracy of human visual identification of crystals. Secondly, fingerprints of the collective precipitation behaviour of a protein across standard sparse-matrix can be generalised, and compared objectively to fingerprints of historical experiments, to assess crystallizability and infer optimization strategies based on past successes. Thirdly, clear drops can be automatically identified, and mapped to chemical components in a sparse-matrix screen to suggest alternative buffers for protein formulation. Fourthly, TeXRank, a user interface could be developed to present and make all algorithm output accessible for daily use. Fifthly, the associated data mining led us to evaluate the strategies for setting up screening experiments with limited protein samples, based on over ten years of crystallization data at the Structural Genomics Consortium, Oxford. Our methods capitalizes on present day standard screening procedure and hardware to extract useful information, bypassing laborious and subjective evaluation of each droplet.

# Author's Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in text.

Jia Tsing Ng

7 September 2015

# Acknowledgements

My deepest gratitude goes to Professor Frank von Delft, for entrusting this project to me, providing ample supervision and guidance (even on weekends!), and for being a great friend. I must also thank Professor Michael Osborne for all the machine learning/data science supervision; Dr Carien Dekker and Dr Markus Kroemer, for their hospitality, time and expertise, and making my visits to Novartis always productive and enjoyable. Thank you all for making 'managing your supervisors' such a walk in the park.

I have to thank all members of the Protein Crystallography group at the SGC, both past and present, for all the help I have received practically and in generating ideas; David Damerell, the database guru; all members of the SGC for making it a great place to work, for sharing all data used here, and feedback for improving TeXRank.

Special thanks to the Khazanah Foundation for the Merdeka Scholarship, and the funding from the Engineering and Physical Science Research Council (EPSRC), as well as the Novartis Institutes for Biomedical Research, through the Systems Approaches for Biomedical Science IDC.

Friends from the Doctoral Training Centre, St Aldates Church and the Tuesday Postgraduate group, 127 Magdalen Road-ers, who've all made my time in Oxford without a moment of boredom; Radek, Dave, Alex, and Anthony (best work-neighbour ever), whom I'm glad to have shared this DPhil phase-of-life with daily in the lab; my family, whose love and support have always been unconditional; and Jesus, for in Him all things were created, through Him, and for Him.

# Table of Contents

# Table of Figures

# List of Tables

# List of Abbreviations

| **Institution** | |
|---|---|
| SGC | Structural Genomics Consortium, Oxford |
| NIBR | Novartis Institutes for Biomedical Research |
| RCaH | Research Complex at Harwell |

| **Sparse-matrix screens** | |
|---|---|
| JCSG | JCSG+ Screen (from the Joint Centre for Structural Genomics) |
| LFS | Ligand Friendly Screen |
| HCS | Hampton Crystal Screen |
| HIN | Hampton Index Screen |

| **Algorithms, distance metric** | |
|---|---|
| AUROC | Area under the receiver operating characteristic curve |
| EM | Expectation-Maximization algorithm |
| GLCM | Grey level co-occurrence matrix |
| GMM | Gaussian mixture model |
| GPLVM | Gaussian processes latent variable model |
| KL Divergence | Kullback-Leibler divergence |
| LVQ | Learning vector quantization |
| PCA | Principal component analysis |
| SLIC | Simple linear iterative clustering |
| SVM | Support vector machine |

| **Protein** | |
|---|---|
| ATAD2A | Two AAA domain containing protein |
| BRD1A | Bromodomain containing protein 1 |
| BTBD12B | SLX4 structure-specific endonuclease subunit homolog (S. cerevisiae) |
| CAMK1DA | Calcium/calmodulin-dependent protein kinase ID |
| CECR2A | Cat eye syndrome chromosome region, candidate 2 |
| DACASAA | D-alanyl-D-alanine carboxypeptidase |
| DOPVA | Pup deamidase/depupylase |
| EP300A | E1A binding protein p300 |
| FAM83AA | Family with sequence similarity 83, member A |
| GADD45BA | Growth arrest and DNA-damage-inducible, beta |
| GYG2A | Glycogenin 2 isoform a |
| HCVPr | HCV NS3 protease |
| JARID1BA | Jumonji, AT rich interactive domain 1B (RBP2-like) |
| JMJD2AA | Jumonji domain containing 2A |
| JMJD2BA | Jumonji domain containing 2B |
| JMJD2CA | Jumonji domain containing 2C |
| JMJD2DA | Jumonji domain containing 2D |

| | |
|---|---|
| OTUB2A | OTU domain, ubiquitin aldehyde binding 2 |
| PAHA | Phenylalanine Hydroxylase |
| PHIPA | Pleckstrin homology domain interacting protein |
| PRKCBP1A | Protein kinase C binding protein 1, isoform a |
| PXKA | PX domain containing serine/threonine kinase |
| SPIN3A | Spindlin-3 |
| STK6A | Serine/threonine kinase 6 |
| VPS28A | Vacuolar protein sorting-associated protein 28 homolog isoform 1 |
| XX02RIPK2A | Complex of receptor-interacting serine-threonine kinase 2 with baculoviral IAP repeat-containing protein 4 |
| XX08LATS1A | Complex of LATS homolog 1 with Mps One Binder kinase activator-like 1A |
| ZAKA | Zipper containing kinase AZK |
| ZFYVE9C | Zinc finger FYVE domain-containing protein 9 isoform 3 |
| | |
| **Others** | |
| CIELAB | L*a*b colour space defined by the *Commission internationale de l'éclairage* |
| MCR | Matlab Compiler Runtime |
| PDB | Protein Data Bank |
| PEG | Polyethylene glycol |
| PFL | Precipitation Fingerprint Library |
| SEM | Scanning Electron Microscopy |
| SFX | Serial Femtosecond Crystallography |
| SONICC | Second-order nonlinear optical imaging of chiral crystals |
| TEM | Transmission Electron Microscopy |
| UAA | Uranyl-Acetate Alternative |
| XFEL | X-ray Free Electron Laser |

# 1. Introduction

## 1.0 Overview and thesis outline

Unsurprisingly, X-ray crystallography requires crystals. Although recent advancement in electron microscopy and free-electron lasers have made them viable alternative methods to obtain atomic level structural images of proteins without large crystals, X-ray crystallography will still be needed for high-resolution structures, and this will not change in the foreseeable future. The development in technologies required at each step of X-ray crystallography has enabled the exponential increase in the number of structures deposited the Protein Data Bank (PDB) (Shaw Stewart & Mueller-Dieckmann, 2014), but obtaining suitable protein crystals remains a major rate limiting step. While it is remarkable that proteins of odd shapes, sizes, surface charges and flexibility are able to pack together in an orderly manner to form crystals, they nevertheless do, though rarely and not naturally in most cases. Indeed, it has been over 150 years since the first protein crystals of haemoglobin from worms were observed, yet inducing protein crystallization remains empirical in nature, with no comprehensive theory to guide crystallization efforts (McPherson & Gavira, 2014).

The typical process for obtaining structures from X-ray crystallography is represented in Figure 1.1; the focus of this project is on the crystallization stage, with an overarching goal of extracting more information from standard protein crystallization screening experiments using sparse-matrix screens, which is now the default first-line experiment for most structural biology groups (Newman et al., 2012). While the ultimate aim of a screening experiment is to obtain protein crystals or identify potential follow-up strategies, this is however a rare event, and when crystallinity is not observed, one is generally left in the dark, no more informed of how to crystallise the protein than before.

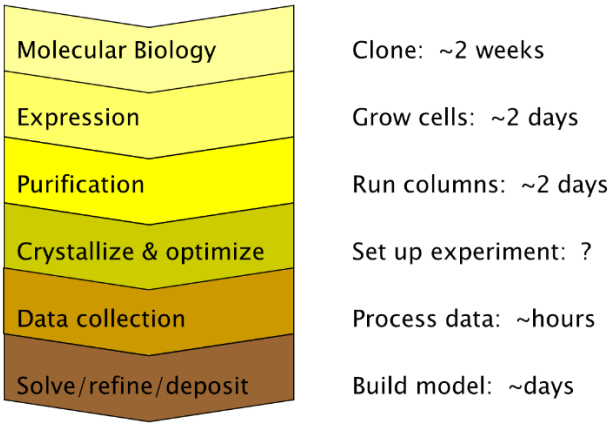



*Figure 1.1: Typical pipeline for obtaining models of protein structure, and the rough estimate of time for each process, assuming they were successful. The focus of this project is on crystallization, where the time taken is often unpredictable since crystallization may occur anytime between under 24 hours, 60 days, or not at all.*

Here, we present methods which combine image processing and machine learning to aid and reconsider how crystallization trials can be evaluated, as well as obtaining readout from screening experiments without relying on the presence of crystalline behaviour. We chose images as our primary source of data since they most reliably capture all outcomes of crystallization experiments across laboratories. Using robotically captured bright field images of crystallization droplets, our approach is to objectively characterize ALL precipitation patterns in a screening experiment through texture analysis, and assess how to use the resulting collective precipitation pattern as an assay of the protein. With the vast amount of data generated at the Structural Genomics Consortium, Oxford, these assays can answer the question of whether and where in crystallization space is the protein likely to crystallise, in a finite time and a finite number of experiments. Our method extracts previously unobtainable or disregarded information from experiments already carried out by an experimenter, requiring no further effort or sample consumption, nor any changes to standard operating procedures.

The following sections of this chapter introduce topics pertinent to the project. We first discuss the requirements for crystallization and the typical screening experiments central to this project, followed by a discussion on the commonly observed outcome in such experiments. We review past computer vision efforts in crystallization, given that image analysis of crystallization droplet images is central to our proposed methods, and conclude with overviews of the machine learning methods utilized in this project.

In the remaining thesis, definitions and our methods are presented in Chapter 2. Chapter 3 addresses our first goal of aiding experimenters in their evaluation of crystallization droplets by ranking crystallization droplets by their likelihood of crystallinity. Chapters 4 and 5 describe how the collective precipitation behaviour of a protein in a standard screen can lead to the identification of potential optimization conditions and alternative buffers for protein formulation respectively, irrespective of whether crystals appeared in the screen. Chapter 6 discusses TeXRank, the user interface developed for the presentation of algorithm outputs for everyday use. This is followed by a retrospective analysis of the crystallization strategies at the SGC over its first 10 years in Chapter 7 to serve as a guideline for the community in setting up screening experiments with limited protein samples. Chapter 8 describes preliminary experimental work on correlating sub-visible structures with textural patterns on crystallization droplets. We conclude by summarizing the main contributions of this thesis in Chapter 9.

## 1.1 The pursuit of a protein crystal

The ultimate aim in a classical crystallization experiment is to grow few, sufficiently large and high quality crystals from a homogenous, pure and stable protein, typically in solution, to provide X-ray diffraction patterns to allow for the visualization of the atomic structure of proteins. Two processes are involved in the formation of crystals: nucleation and growth.

Crystallization starts with nucleation, which is believed to be a two-step mechanism: first the formation of disordered protein-rich clusters, followed by the formation of crystal nucleus inside the cluster (Vekilov & Vorontsova, 2014). Crystal nuclei are thought to be completely ordered assemblies having gone through first-order phase transition from disordered to ordered state. Nucleation is stochastic in nature, and since it is the precursor to crystal growth, it is generally thought of as an important factor in crystallization reproducibility, even from identical protein batches and crystallization conditions (Newman et al., 2007). Crystal growth entails the incorporation of new molecules to the kink or lattice site of a crystal (McPherson & Kuznetsov, 2014). This generally involves the reorientation, dehydration, and conformational changes of a molecule to fit the bulk (growing crystal body), followed by an establishment of water network between the new molecule and the bulk. Furthermore, the system has to have a higher probability of new incorporation than the dissociation of molecules from the bulk for net growth (McPherson & Kuznetsov, 2014).

Bernhard Rupp (2015) summarizes the conditions that must be met for crystallization to happen:

(1) The protein must be inherently crystallisable: if no periodic intermolecular contacts can be formed, crystallization will never take place;
(2) Crystals must be thermodynamically possible, so that a protein-rich crystal can exist in equilibrium with its growth solution;
(3) Kinetic processes to reach the thermodynamically stable state must exist and be favourable.

(1) is a property of the protein, but can be adjusted by various strategies, including construct design (Savitsky et al., 2010) and residue mutation (Price et al., 2009). The chemical (also referred to as precipitant) and physical conditions for crystals to nucleate and grow, and how to

get there through kinetics, is often unknown. Furthermore, there are no clear correlations between crystallization conditions and the protein structure or family (Chayen & Saridakis, 2008), and thus, identifying (2) typically involves screening a protein against a large number of conditions, either systematically or randomly. Criterion (3) can be explored through different crystallization experiment formats.

The general workflow in the attempt to crystallize a new protein is to (1) 'screen' the protein against a large number of precipitants and physical conditions to identify those that result in crystalline behaviour, and (2) 'optimize' the identified conditions to improve crystal quality. In practice, these are carried out by mixing protein (in solution) with precipitant (or crystallization cocktail, see Section 1.1.3), and equilibrating the system in well-established experiment formats discussed next.

### 1.1.1 Experiment formats for crystallization experiments

Various experiment formats have been developed for protein crystallization, including vapour diffusion, batch method, and dialysis (Figure 1.2). Physical properties of these formats, together with the precipitant, induce the kinetic processes to achieve supersaturation of protein molecules required for the crystallization of macromolecules. Supersaturation occurs when some macromolecules remain in solution even though the system is beyond its solubility limit, thanks to specific chemical and physical conditions. To re-establish equilibrium, the macromolecules enter a solid state, which may take the form of crystals or precipitates. Supersaturation may be achieved by decreasing the solubility of undersaturated protein by modifying the properties of the solution and/or the protein, for example through the alteration of pH, temperature, and salt, or the removal of water by evaporation or using semipermeable membranes (McPherson & Gavira, 2014). The different experiment formats aim to induce such

modifications, each in different ways best described by the respective phase diagrams shown in Figure 1.2.

Sitting-drop (Figure 1.2(a)) and hanging-drop vapour diffusion are popular methods for crystallization. In vapour diffusion experiments, the protein/precipitant mixture equilibrates against a reservoir, either with the same precipitant or salt solutions (Newman, 2005). Because the reservoir has lower vapour pressure than the protein/precipitant droplet due to its higher solute concentration, water, in the form of vapour, leaves the droplet and preferentially readsorbs to the reservoir until there are equivalent vapour pressures. The dehydration of the droplet increases concentrations of the protein and precipitant, causing desolubilization of the protein, ideally in the form of nucleation and crystal growth. Factors affecting equilibration rate in vapour diffusion experiments include droplet size, its distance to the reservoir, experimental set-up (sitting or hanging-drop), concentration and types of precipitant used, droplet volume to reservoir volume ratio (Forsythe et al., 2002), as well as if a different reservoir solution was used to equilibrate the droplet (Newman, 2005).

In batch experiments, the protein/precipitant mixture is dispensed under paraffin oil, resulting in a much slower evaporation or dehydration rate. Thus, a fixed concentration of precipitant is generally assumed throughout the experiment (Figure 1.2(b)). Dehydration may be controlled using combinations of oils to obtain different water-permeability (Luft et al., 2014). In contrast, dialysis starts with the protein separated from the precipitant by a semipermeable membrane, and equilibration is achieved by the migration of water and precipitant into the protein sample. This results in a gentle increase of precipitant concentration, as shown in Figure 1.2(c) (Chayen & Saridakis, 2008). Variation in results is expected for different experiment formats, and hence, repeating experiments with different formats may be a worthwhile effort (Chayen, 1998).

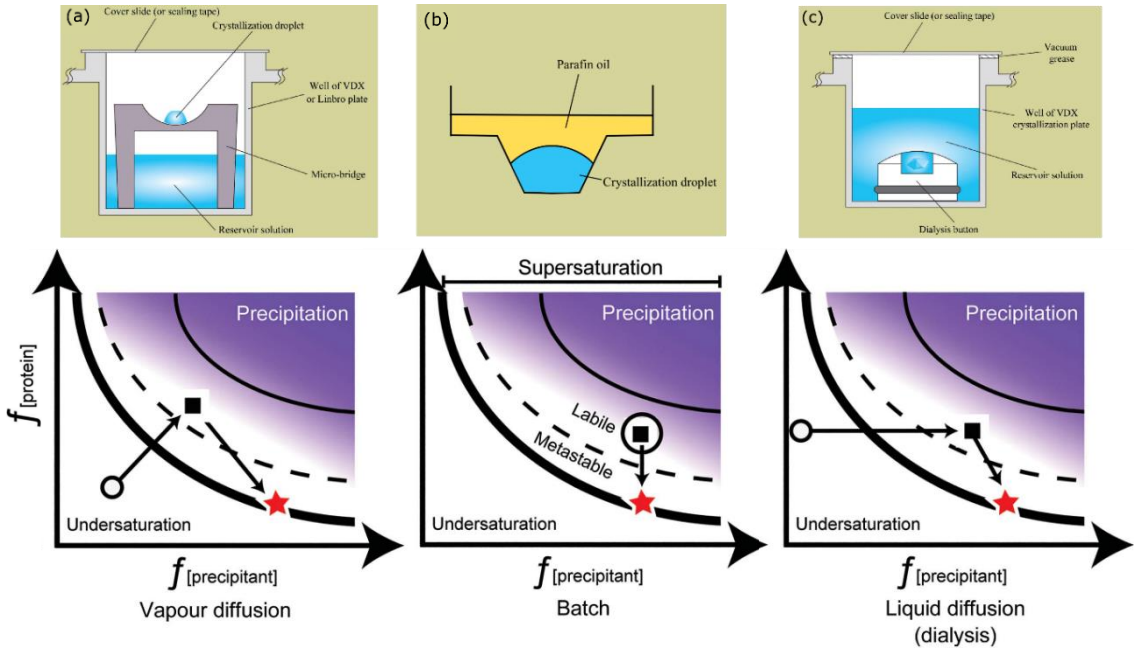



*Figure 1.2: Schematic representations of common crystallization methods and the respective trajectories through the phase diagram, including (a) vapour diffusion, (b) microbatch, and (c) dialysis. As the crystallization system equilibrates, each method induces supersaturation through different paths on the phase diagram, to induce nucleation and crystal growth. The x-axis of the phase diagram represents the precipitant, the y-axis the protein concentration. The circle indicates the starting point of the experiment, the black square the point of nucleation, and the red star the equilibrium point of the crystal. Figures adapted from McPherson & Gavira* (2014) *and Luft et al.* (2014).

### 1.1.2 The protein

The protein is the most important variable in crystallization (Dale et al., 2003). Not only is the inherent crystallizability of a protein mandatory for successful crystallization, sample purity, solubility and stability are crucial for producing high quality protein crystals. Development in biotechnology and genetic engineering has enabled large scale expression of protein through recombinant approaches, and sophisticated purification methods by tag, size or charge have enabled the production of pure and concentrated protein samples unobtainable naturally (McPherson & Gavira, 2014). Protein expression and purification methods are beyond the scope of this work, and hence will not be discussed in detail; instead we focus on methods developed

to improve crystallization success from an informatics (sequence/structure) point of view, since these are related to predicting crystallizability, which is relevant this project.

One way to increase success rates of crystallization efforts is to select the 'right' protein targets to work on. Protein sequence-based predictors have been primarily designed to identify the suitability of a potential target by predicting its crystallizability based on sequence alone. Most of these methods are currently available as web servers, including the more recent CRYSTALP2 (Kurgan et al., 2009), MCSG-Z score (Babnigg & Joachimiak, 2010), XANNPred (Overton et al., 2011), RFCRYS (Jahandideh & Mahdavi, 2012) and CRYSpred (Mizianty & Kurgan, 2012). Features such as the average hydrophobicity, content of mono-, di- and tripeptides, instability indices, isoelectric points, contents of selected residues etc. are derived from the sequence. These features are used as inputs into standard predictive models which then quantify the protein's crystallization propensity. The positive-labelled sequences used to train the above algorithms were largely obtained from the PDB, while the negative-labelled sequences were obtained from protein entries in TargetTrack (http://sbkb.org/tt/, previously TargetDB) that were soluble but recalcitrant for crystallization. Predictive models used include support vector machines (SVMs), artificial neural networks, naïve Bayes classifiers and random forest classifiers.

The limitation of such methods is that they can only consider intra-molecular factors encoded in the protein sequence, which don't necessarily capture the variables involved in protein production and crystallization. Although the highest reported accuracy is 80% (Jahandideh & Mahdavi, 2012), judged by the proportion of true positive and true negative predictions, such methods are only of limited use in target selection, and certainly do not predict crystallizing conditions for predicted favourable targets.

A better way to use protein sequence information is in the design of protein target constructs, which are fragments of the sequence to be cloned. This is currently considered best practice for large eukaryotic multi-domain proteins. The approach entails generating multiple constructs using bioinformatics prediction of structured domains, conserved domains, secondary structures and disordered region, from structure-informed sequence alignment against known structures (Savitsky et al., 2010).

Another use of protein sequence information is to analyse surface residues for sites to mutate, to induce crystal contacts, reduce surface entropy, delete disordered/hydrophobic loops, or disrupt unwanted crystal contacts (Price et al., 2009). Fusco *et al.* (2014) used solved structures and the corresponding crystallization conditions to find correlation of protein properties and crystallization propensity. Provided that surface residues can be predicted for unknown structures, this forms a guideline to avoid or enrich certain residues where targeted mutagenesis can be carried out. Ongoing work at the SGC on surface residue mutations resulted in new crystal forms as well as new crystallization conditions for a given protein target.

The upstream testing of multiple constructs and/or epitope mutations inevitably increases the number of crystallization experiments. While robotics help tremendously in the setting up of these experiments, these do not ease the burden of evaluating the outcome, which is usually by visual inspection, and deciding on follow-up experiments.

### 1.1.3 Screening to identify crystallizing conditions

Assuming that the protein of interest is inherently crystallisable and a pure, concentrated sample is available, crystallization of a protein typically begins with the screening for chemical, biochemical and physical conditions that result in crystalline behaviour. Variables to explore

include pH, temperature, ionic strength, as well as precipitant type and its concentration (Stevens, 2000; McPherson & Gavira, 2014) .

A systematic search through all the parameters is intractable. The common strategy is therefore to use sparse-matrix screens, which are collections of conditions that sample chemical space that have historically been successful at inducing crystallization. The notion of a sparse-matrix screen was first put forth by Jancarik and Kim (1991), who compiled 50 'sensible' conditions based on deposited structures. The commercialization of this screen in less than a year started a revolution in the screening process; the over 2200 citations to the paper attest to the success of the design and method. There are strong arguments for using past successes as a first attempt at crystallization: a study from the Joint Centre for Structural Genomics showed that 67 out of 480 conditions sampled produced crystal hits for 84% of 465 proteins (Page et al., 2003); similar results emerged at the University of Toronto where only 6 conditions were responsible for the crystals of 180 out of 338 proteins (Kimber et al., 2003). This principle of sampling previously-successful chemical space for crystallization forms the basis for the more than 200 commercially available screens currently available (Newman et al., 2013), including the Hampton Crystal Screen which was based on Jancarik and Kim's design, and membrane protein specific screens like MemGold (Ferrandon & Newstead, 2008). Additionally, past successes can also inform us on molecules that have positive effect in rigidifying proteins; screens like MORPHEUS (Gorrec, 2009) and Silver Bullets Screen (McPherson & Cudney, 2006) are designed to include molecules which are often found to be stabilizing proteins or mediating crystal contacts.

The major advantage of sparse-matrix screen is convenience: the commercial availability of such screening kits not only removes the effort required to formulate hundreds of conditions, but also lend themselves easily to automation since they are constant, thus lowering the barrier to crystallization by reducing time, protein sample and expertise required (Luft et al., 2014). It also captures domain knowledge and makes it broadly available. The simplicity of such screening kits

have made them the tool of choice for crystallizing proteins and other macromolecules. It has been estimated that sampling around 300 conditions is sufficient to identify crystallization conditions (Segelke, 2001), although high throughput laboratories tend to sample more conditions (up to 1536 at the Hauptman-Woodward Medical Research Institute) to increase the chance of success and overall information content of the experiment to guide subsequent efforts (Newman et al., 2012). However, selecting which commercial screen to maximise chemical space sampling may not be obvious, especially with high levels of duplication amongst vendors. Tools such as the C6 web tool (Newman et al., 2010) and Cockatoo (Bruno et al., 2014) have described objective distances between conditions to allow for numerical comparisons of screens, as well as characterize the internal diversity of chemical conditions in a screen.

In general, the crystals obtained from screening experiments are not necessarily of sufficient quality for X-ray diffraction analysis. Optimization of chemical components in the condition(s) identified from the coarse screen (polymer, salt, buffer, pH, temperature) may be required to improve the quality of crystals. This usually involves the systematic variation of some or all of the components; because effects may be coupled, the problem is complex and nonlinear (McPherson & Cudney, 2014). Other parameters to optimize include the protein to precipitant mixing ratio, the total mixture volume, protein sample concentration, introducing nucleants through different seeding methods (Stura & Wilson, 1991; Bergfors, 2003; D'Arcy et al., 2007; Khurshid et al., 2014), or by changing the crystallization experiment format. Because the ultimate test of crystal quality is not visual but diffraction, selection of which hit to optimize, in cases where many initial hits were identified, should be led by diffraction quality wherever possible.

## 1.2 Evaluation of crystallization experiment

### 1.2.1 Common outcome

Evaluation of crystallization experiments typically involves visual inspection of crystallization volume. While this was historically done under the microscope, the development and deployment of robotics not only enabled crystallization experiments to be set up easily, but also the automatic capture of crystallization droplet images at fixed time points. Software accompanying the robotic imaging system allow for the viewing of droplet images at the experimenter's convenience, from the comfort of the office desk or even remotely. Other basic features in most vendor software include the display of information associated with the experiment, as well as functions for the annotation or scoring of droplets, to keep track of any crystalline behaviour for subsequent action. Vendor software is generally integrated with hardware, creating a centralised environment for further experimental design and execution.

Crystals are undoubtedly what all experimenters hope for as the outcome of their crystallization experiments. Examples of droplets from sitting-drop vapour diffusion experiments with crystals are shown in Figure 1.3. In the absence of crystals, *i.e.* 99.8% of the time (Newman et al., 2012), other visual outcome may still point toward optimization strategies, although they are less direct and often difficult to describe. Luft *et al.* (2011) defined the following frequently observed outcome (either alone or in combination), with some examples shown in Figure 1.4:

1) Clear drops: When the protein sample remains in solution, this may indicate either undersaturation or metastable supersaturation. Differentiating between these states require observations of the overall screen: if similar conditions are clear, the system is likely to be undersaturated; if precipitation, phase separation or even crystals appear in related conditions, the system is likely to be in the metastable state, but the transition

into solid state will not take place. Seeding (adding nucleants to bypass nucleation) may be effective at inducing crystallization in metastable systems (Bergfors, 2003).

2) Liquid-liquid phase separation: Protein-rich and protein-poor liquid phases may exist when there are highly anisotropic interactions between protein molecules. Such observations indicate that the system is close to conditions for crystallization, and the protein-rich region may also be used for seeding.

3) Precipitation: This occurs when there is supersaturation beyond the level suitable for crystallization, and is widely described as either bad (amorphous in appearance and irreversible, likely to be denatured protein) or good (granular in appearance and reversible, protein is still in its native confirmation). Conditions that produce good precipitate may be worth optimizing, or used as seeds.

4) Skin: When proteins adsorp onto interfaces (solution/oil, solution/air, or solution/surface), they may undergo conformational changes and aggregate to form gel-like structures which are typically irreversible. They can be removed from crystallization droplets with tools like acupuncture needles, although often crystals nucleate on skin and become difficult to separate.

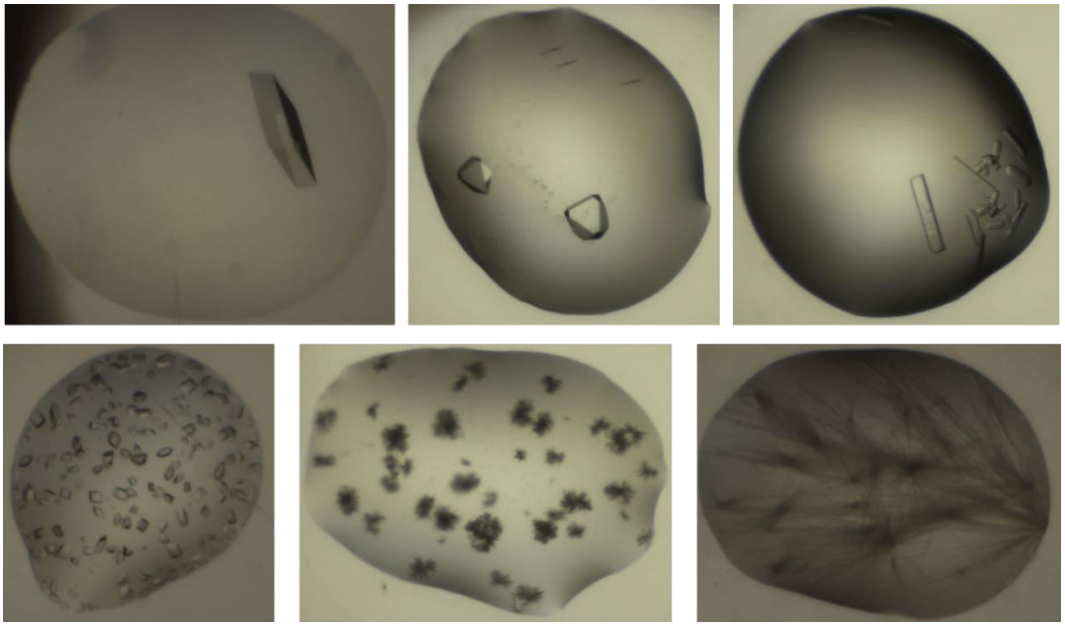


*Figure 1.3: Examples of protein crystals in sitting-drop vapour diffusion experiments, including single large crystals and microcrystals.*

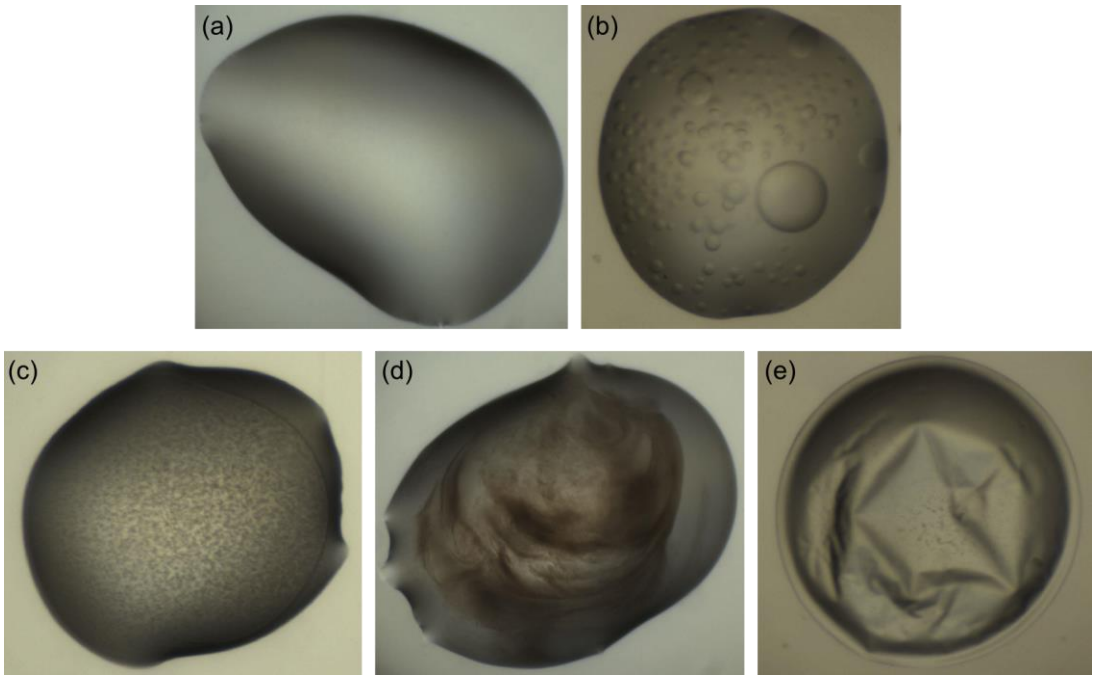



*Figure 1.4: Non-crystalline behaviour commonly observed in crystallization experiments, including clear drop (a), liquid-liquid phase separation (b), the loosely defined 'good' precipitate (c) and 'bad' precipitate (d), as well as skin (e).*

These very broad definitions of possible outcome and the possibility for any combination of outcome to occur simultaneously make the evaluation of crystallization experiments complicated and largely subjective. Furthermore, the large number of crystallization experiments, thanks to robotics, make detailed observation for all droplets impractical (Newman et al., 2012); even the fundamental task of only identifying crystals remains time consuming and prone to human error. Therefore, algorithms have been developed, predominantly on bright field images to, automate the task of evaluating crystallization outcome. Since crystals are rare and expensive to miss, other imaging modalities have also been developed to increase signal from protein crystals for easier identification.

### 1.2.2 Automating droplet evaluation with computer vision

A significant body of research exists on accurately identifying crystals from robotically captured images of crystallization droplets. Early work used edge detection solely to detect crystals (Ward et al., 1988; Zuk & Ward, 1991); this was later extended to the auto-classification of experimental outcome into various classes to better represent the diversity of crystallization experiments, with information derived from a diverse range of texture analysis methods and/or edge detection, paired with off-the-shelf machine learning algorithms. Wilson (2002) derived 13 features from edge objects in a droplet, and each droplet was classified into one of five categories ranging from “single crystals” to “unfavourable outcomes” with a Naïve Bayes classifier. Spraggon *et al.* (2002) added texture descriptors derived from grey level occurrence matrices (GLCM) to train a 5-class neural network classifier, and reported a high accuracy of 82% for crystal images, but a mere 55% and 47% for “homogenous” and “inhomogeneous” precipitation respectively.

Instead of deriving features from the whole droplet, local features from sub-sections of an image later became the method of choice. Bern *et al.*(2004) and Zhu *et al.*(2004) derived features from a 25 × 25 'most promising region' for each image as inputs to a C5.0 (adaptive boosting of decision tree) classifier, while Pan *et al.* (2006) derived features from overlapping 40 × 40 pixel blocks for support vector machines (SVM). The SVM algorithm was also used by Kawabata (2006) on a cropped image of 450 × 450 pixels without droplet segmentation to identify crystals suitable for diffraction analysis (dimensions greater than 50 µm × 10µm × 10µm) against microcrystals.

Most subsequent work returned to deriving global features from the entire droplet. Spectral methods for texture analysis were introduced, including Fourier transform (Walker et al., 2007) and wavelet modelling (Watts et al., 2008). These were both paired with a learning vector quantization (LVQ) neural network with 7 target classes, which could alternatively be pooled to three classes: images that "should be", "may be", and "need not be" inspected. The wavelet modelling method improved in the classification of microcrystals as 'should be inspected', namely an increase of 9.3%, when compared to using Fourier transform descriptors.

In an attempt to investigate the most effective classifier, as well as compare between the object (Wilson, 2002) and wavelet (Watts et al., 2008) methods, Buchala and Wilson (2008) trained 2 SVMs with different kernels and 2 neural networks (LVQ and self-organizing maps). It was found that the best single classifier was the SVM, but a combination of 2 SVM and 2 neural networks, with both object and texture methods, produced the best results.

A requirement in training the aforementioned classifiers is the availability of annotated data sets, and earlier studies used training data containing hundreds to less than 3000 images. With the availability of more crystallization trials, coupled with better annotation and database facilities, larger scale studies were carried out by Liu *et al.* (2008) and Cumbaa and Jurisica

(2010), using data from high-throughput centres (Hauptman-Woodward Medical Research Institute and Joint Centre for Structural Genomics respectively). Liu *et al.* focused on local regions of 127 × 127 pixels, deriving 466 features with Gabor wavelet and orientation histograms. 21,477 squares were used to train a decision-tree variant of boosting classifier. When the classifier was used to rank 319,112 images from 150 proteins that yielded solved structures, the top 22% of the ranked images contained crystal images for 145 of these proteins; here, the focus was only on harvestable crystals greater than 10µm. The highest number of features per image was derived by Cumbaa and Jurisica (2010), where 12,375 features from 165,351 microbatch-under-oil images were calculated using the computing resources of the World Community Grid. The resulting Random Forest classifier successfully identified 80% of crystal images, 89% of precipitates and 98% of clear drops. More recently, Lekamge *et al*. (2013) used time series information by calculating textural features from the difference image of consecutive inspections, and decision trees were used to classify experiment outcomes into 5 classes with an accuracy of 74.5%.

Identifying which pixels in the image belong to the droplet ("segmentation") is important to avoid interpreting noise from regions outside the droplet. Approaches to the problem include edge detection (Spraggon et al., 2002; Wilson, 2002; Bern et al., 2004), deriving masks from wavelet transforms (Watts et al., 2008), using plate-specific features (Pan et al., 2006; Cumbaa & Jurisica, 2010), or simply using the centre or regions of an image (Kawabata et al., 2006; Liu et al., 2008). A more robust algorithm was developed by Valloton *et al.*(2010), which transforms the image from Euclidean to polar coordinates, centred at a point in the droplet, allowing a shortest path algorithm to be used to trace the droplet boundary, resulting in a robust method with no constraints on the droplet shape, such as having to be close to a perfect circle.

Although most of the above work classify experimental outcomes into discrete classes, this is still an oversimplification of protein behaviour: overlapping categories are often ignored or not

supported, and greatly undermines the information content of the crystallization experiment it represents. Classification accuracy of non-crystalline classes are also hard to judge since there is a lack of objective definitions. While it is initially useful to automate image annotation, experimenters will ultimately visually re-evaluate droplets because missing crystals is unacceptable, and all algorithms are inaccurate. Furthermore, because all of the above work was developed and validated on specific imaging systems, a comparison of reported accuracies is meaningless.

Nevertheless, the value in the objective translation of images into some form for analysis is that these images are generally the only measured outcome for most crystallization experiments (Newman et al., 2012). It is not apparent that the features derived from images in the above work suitably capture and describe non-crystalline behaviour, since they were primarily developed and optimised to detect crystals, while all other outcomes received little to no attention.

### 1.2.3 Other imaging techniques

Identifying crystals in bright-field images relies on the difference of the refractive index of a protein crystal to its surrounding environment. It is the most ubiquitous form of imaging for crystallization droplet and comparatively, the cheapest and quickest. However, this approach is unreliable because such difference in refractive index is also true for salt crystals, and may be low for protein crystals due to high solvent content (Desbois et al., 2013), or if it is physically hidden behind precipitates. More sophisticated imaging techniques have emerged to increase signal strength from protein crystals for visual and automated identification, by relying on specific properties of proteins or protein crystals, namely fluorescent imaging and SONICC.

#### *1.2.3.1 Fluorescent imaging*

Fluorescent imaging of crystallization experiments either rely on the intrinsic fluorescence of proteins from residues like tryptophan, or trace labelling of proteins with fluorescent dye. For intrinsic tryptophan fluorescence, a crystallization droplet is excited at wavelengths of 290 ± 5 nm, and an image is formed by detecting fluorescent emission at 320-350 nm. Since crystals are regions with higher density of proteins, they are expected to produce more signal than the background. However, the level of fluorescence depends not only on the number of fluorescent residues, but also how buried they are within the protein. Tryptophan, the only residue with sufficiently high quantum efficiency to be used as a probe, is only estimated to form 1.09% of residues in proteins (Gilis et al., 2001), and is not a guaranteed residue in all proteins. Some crystallization conditions used in typical screens (conditions with nitrates, cobalt and/or iron) have also been found to quench fluorescence signal, while false positives may arise from protein adsorption layers at liquid-gas interfaces or phase separation, as well as proteins adsorbing onto the surface of inorganic crystals (Desbois et al., 2013).

Using fluorescent dye in contrast, while independent of florescent residues, requires the labelling of proteins, for which labelling protocols may have to be optimized for different proteins. Forsythe et al (2006) showed that covalently bonding < 1% of protein molecules with fluorescent probe at the N-terminal amine is sufficient to produce the signal strength required for visual detection, without adversely affecting crystallization and X-ray diffraction quality through the introduced impurity. The ability to change the fluorescent probe also allows components of complexes to be labelled differently, and thus allow the identification of successful complex formation.

Examples of intrinsic UV fluorescent or labelled-fluorescent crystal images are shown in Figure 1.5(a) and Figure 1.5 (b).

#### *1.2.3.2 Second-order nonlinear optical imaging of chiral crystals (SONICC).*

SONICC has been shown to be a sensitive and selective method to image protein crystals beyond the optical diffraction limit with an estimated lower bound of 90nm in all dimensions (Kissick et al., 2011), and especially good at imaging membrane protein crystals in lipidic mesophases (Calero et al., 2014). SONICC detects the frequency-doubling response from excitation with femtosecond infrared light source, an effect only produced by chiral crystals, i.e. lacking internal planes of symmetry. Examples are shown in Figure 1.5(c) and (d). Since most salt crystals are non-chiral, they give no SONICC signal, whereas up to 99% of known protein crystals are expected to be chiral (Wampler et al., 2008). Because SONICC is inherently based on crystallinity, precipitation gives little to no signal, and hence SONICC images are great for identifying a subset of crystals and microcrystals, including those with visual appearance similar to that of granular precipitate (Calero et al., 2014). However, this means they are not useful for characterizing precipitation. Indeed, it is a method that reflects the community's "fixation, almost a glorification, of crystals as the only useful result" (Newman et al., 2012). Depending on the number of image slices required, imaging time is between 1.5 to 2.5 hours for 288 droplets, significantly longer than conventional bright field images, namely ~15 minutes on the same system (Timm Maier, personal communication).

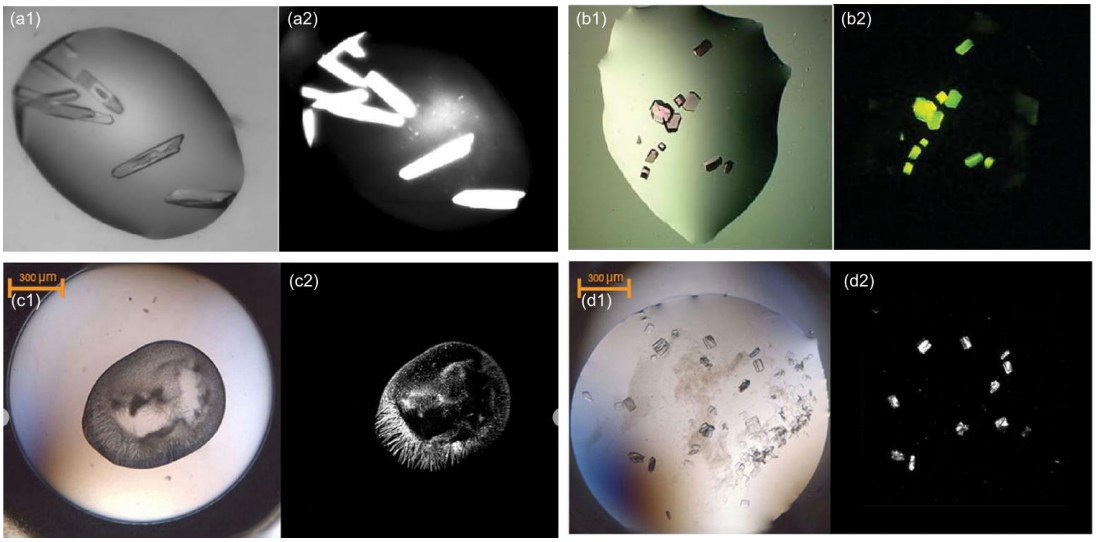



*Figure 1.5: Alternative imaging techniques for protein crystallization. (a1) and (a2) are bright field and the intrinsic UV fluorescence image respectively. (b2) is the florescent image (b1) where the protein was labelled at 0.25% with carboxyrhodamine succinimidyl ester. (c2) and (d2) are the SONICC images corresponding to (c1) and (d1), where only chiral protein crystals give signal, and precipitates are completely black. (a) is taken from (Desbois et al., 2013), (b) is taken from (Forsythe et al., 2006), (c) and (d) are taken from [www.formulatrix.com](http://www.formulatrix.com)*

### 1.2.4 Combining all conditions: using the whole screen

The body of work described in the previous section mainly treat crystallization droplets as single experiments, even when all droplets in a sparse-matrix screen typically share the same protein. This is in contrast to how experienced crystallographers analyse their screening outcome, and hence the notion of using crystallization screening experiments as a whole is certainly not new. For example, the neural network approach by DeLucas (2005) uses all scored experimental outcome of a 360-condition incomplete factorial screen, as inputs to a trained network to predict conditions likely to yield crystals.

Luft *et at*. (2011) demonstrated ways to use information derived from screening experiments to provide guidelines to a protein's solubility and identify conditions to optimize. Using manual evaluation of crystallization droplets, they were able to obtain an empirical picture of a protein's crystallization phase diagram by relating crystallization outcome and biochemical environment; it did however require "careful note taking and the representation of results in a manner that allows simple human interpretation". AutoSherlock (Nagel et al., 2008; Snell et al., 2008) requires manual annotations of crystallization experiments to produce a graphical map that links the annotation of crystallization outcome to chemical space covered. It was developed to present large quantities of screening data for rapid interpretation. Figure 1.6 shows a section of output of AutoSherlock which can be used to identify 'islands' of crystallization hits (red regions) as a possible optimization point.


*Figure 1.6: Snapshot of AutoSherlock, which is a chemical space map of human annotations of crystallization experiments. Red, green and turquoise indicate crystal hits, precipitate and clear drops respectively, with other colours indicating the cross of multiple categories. The 'island' of hotspots with PEG 4000 and phosphate/sulfate indicate an interesting region to explore. Figure taken from (Snell et al., 2008).*

The main limitation, and also the common theme of the methods mentioned is that they require the translation of visual observations into numerical scores or labels, which is a complex undertaking, since, to quote McPherson and Cudney, 'crystallization outcomes are enormously

varied, sample-dependent, often ambiguous and difficult to describe or assign scores that are physically meaningful: rubbish in, rubbish out' (2014).

## 1.3 Data analysis with machine learning

15 years after the widespread adoption of automation and funding of large-scale centres, there is now a vast amount of crystallization data. Newman *et al.* (2012) have discussed the need to capture crystallization data in a standardized and manner; machine learning methods would appear to be the best tool for utilizing and learning from such data. Furthermore, computer vision and machine learning tend to go hand in hand: computer vision generates the inputs required for machine learning algorithms, which may lead to the discovery of patterns and structure in the data that may have otherwise been missed due to the volume and high dimensionality of the data. Various machine learning methods were indeed deployed in the work on computer vision in crystallization data (Section 1.2.2). This project was no exception, and here, we introduce the specific algorithms used.

Machine learning algorithms include supervised and unsupervised methods, with the main difference being whether labelled datasets are needed: supervised learning algorithms are trained on completely labelled datasets, where every data point has an associated outcome value or class; unsupervised learning identifies the inherent patterns in a dataset without resorting to labels.

### 1.3.1 Supervised Learning

Supervised learning methods include regression and classification. A standard supervised learning problem takes training examples in the form of $\{(\boldsymbol{x_1}, y_1), \dots (\boldsymbol{x_m}, y_m)\}$ and outputs a hypothesis about the true function $y = f(\boldsymbol{x})$. The features $\boldsymbol{x_i}$ are typically vectors, while $y_i$ are real values for regression or drawn from a discrete set of classes $\{1, 2, \dots, K\}$ for classification; both $\boldsymbol{x}$ and $y$ are typically corrupted with noise in real-world data. Regression is less helpful for crystallization data since there is no continuous score for crystallization outcome (or rather no such dataset exists that is reliable and consistent). In contrast, classification has always been the method for analysing crystallization outcome; this includes labelling of images as described in Section 1.2.1. We thus limit this discussion to classification.

Ensemble classifiers, as the name suggests, determine class output with a set of classifiers whose individual decisions are combined in some way, typically by voting, to generate a consensus classification of a new example. Ensembles have been shown to be more accurate than individual classifiers, if the individual classifiers are independent of each other (diverse). This way, when one member of the ensemble is wrong, the others will not necessarily be wrong, thereby still biasing the vote in the right direction, providing that the error rate of each individual classifier is below 0.5 (Dietterich, 2000).

An important ensemble method is the random forest (Breiman, 2001), which is a collection of tree predictors. The trees break down complex decision making processes, and use the union of several simpler decisions to classify a data point (Safavian & Landgrebe, 1991). To decrease the correlation between decision trees in a random forest, a two-fold randomisation is carried out to build each decision tree: (1) a random subset of the training dataset is selected, i.e. bootstrapped samples, with the remaining used to validate the tree; and (2) only a random subset of features is used to build the tree instead of all features. When used for prediction, each tree casts a unit vote for the most popular class. The random forest is found to be relatively

robust to outliers and noise, faster than bagging (randomisation with only bootstrapped samples) or boosting (assigning different weights to each classifier), and also produces useful estimates of error and variable importance (Breiman, 2001). The output of the random forest classifier is also a soft classification based on the voting of the individual trees. The soft classification score thus informs the confidence of the classification, and will prove to be useful in our application of the algorithm later.

### 1.3.2 Unsupervised learning

An advantage of unsupervised learning is that target outputs are not required, which means the algorithm does not need to be told what the outcome should be. Classical examples of unsupervised learning are clustering, which seeks identify patterns in the data beyond what is unstructured and random noise, and dimensionality reduction, which is the representation of data in a more efficient manner (Ghahramani, 2004). Dimensionality reduction allows for the identification of important features or some form of combination of features to simplify the description of a dataset with minimal information loss. Clustering, on the other hand, is important in exploratory data analysis, allowing the study of the unknown nature of data with little to no ground truth. By partitioning data into discrete clusters, underlying structure can be observed to provide insights for further analyses. While no formal definitions of clusters exists, several operational definitions of a cluster can be formulated, as outlined by Xu & Wunsch (2010):

(1) A cluster is a set of data objects that are similar to each other,

(2) The distance between an object in the cluster to the centroid of the cluster is less than the distance between this object to the centroids of any other clusters,

(3) For perfect clustering, the distance between two objects in a cluster is less than the distance between any object in a cluster to any object not in it,

(4) A cluster is a continuous region of data objects with a relatively high density, separated from another such dense regions by low-density regions.

Of the vast variety of clustering algorithms available, the methods used in this project are discussed below.

#### *1.3.2.1 Gaussian mixture models*

Clustering of data can be achieved by identifying the mixture model distribution that best represents the data, since a mixture model is a distribution made up of the superposition of $\mathrm{K}$ simpler component distributions to represent more complex probability distributions that cannot be captured by the simpler components alone (Ghahramani, 2004). Clusters can thus be determined or defined by the simpler components, and the cluster identity of a data point is simply the *K*-th simpler component that best describes it. One of the most popular and useful mixture model, particularly for normally distributed real-world data is the Gaussian mixture model (GMM). The component distributions in GMMs are multivariate normal distributions $\mathcal{N}(x|\mu, T)$, where $\mu$ and $T$ are mean and inverse covariance parameters respectively. The model for $\mathrm{K}$ components is thus

$$p(\mathrm{X}|\pi, \mu, T) = \sum_{k=1}^{\mathrm{K}} \pi_k \, \mathcal{N}(x|\mu_k, T_k), \qquad 1$$

where $\pi_k$ is the mixing proportion for component $k$ and $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$. The probability of data set $\mathrm{X}$ with $N$ observations assumed to be drawn from the above mixture model is then

$$p(\mathrm{X}|\Theta) = \prod_{n=1}^{N} P(X|\Theta) = \mathcal{L}(\Theta|\mathrm{X}), \qquad 2$$

where we have bundled the parameters as $\Theta = (\pi, \mu, T)$, and $\mathcal{L}(\Theta|\mathrm{X})$ is the likelihood function of the parameters given the data. In a maximum likelihood framework, the goal is to find $\Theta$ which maximises $\mathcal{L}$. The iterative Expectation-Maximization (EM) algorithm is typically used to find the maximum likelihood estimate for GMMs, by introducing a latent variable $y$ for each data point which specifies the mixture component that generated the data point, or $P(y = k|\pi) = \pi_k$. The 'complete' likelihood function thus becomes $\mathcal{L}(\Theta|\mathrm{X}, \Upsilon) = p(\mathrm{X}, \Upsilon|\Theta)$.

The lower bound of the log likelihood for any data point can be written as

$$\log p(\mathrm{X}|\Theta) = \log \sum_{\Upsilon} q(\Upsilon) \frac{p(\mathrm{X}, \Upsilon|\Theta)}{q(\Upsilon)} \geq \sum_{\Upsilon} q(\Upsilon) \log \frac{p(\mathrm{X}, \Upsilon|\Theta)}{q(\Upsilon)}. \qquad 3$$

The EM algorithm iterates between optimizing the lower bound as a function of $q$ and $\Theta$, and $q(\Upsilon)$ is some arbitrary density over the latent variable. After initialization of $\Theta$, the E-step optimises the equation with respect to distribution $q$ with fixed parameters $\Theta_{i-1}$ at iteration $i$:

$$q_i(\Upsilon) = \arg\max_{q(\Upsilon)} \sum_{\Upsilon} q(\Upsilon) \log \frac{p(\mathrm{X}, \Upsilon|\Theta_{i-1})}{q(\Upsilon)}. \qquad 4$$

Since $p(\mathrm{X}, \Upsilon|\Theta_{i-1}) = p(\Upsilon|\mathrm{X}, \Theta_{i-1}) p(\mathrm{X}|\Theta_{i-1})$,

$$q_i(\Upsilon) = \arg\max_{q(\Upsilon)} \left[ \log p(\mathrm{X}|\Theta_{i-1}) + \sum_{\Upsilon} q(\Upsilon) \log \frac{p(\Upsilon|\mathrm{X}, \Theta_{i-1})}{q(\Upsilon)} \right]. \qquad 5$$

The second term is equivalent to the negative Kullback-Leibler divergence from $q(\Upsilon)$ to $p(\Upsilon|\mathrm{X}, \Theta_{i-1})$, and so we have:

$$q_i(\Upsilon) = \arg\max_{q(\Upsilon)} \left[ \log p(\mathrm{X}|\Theta_{i-1}) - KL(q(\Upsilon)||p(\Upsilon|\mathrm{X}, \Theta_{i-1})) \right]. \quad 6$$

Since the first term is constant with respect to $q(\Upsilon)$, the equation is maximised when $q(\Upsilon) = p(\Upsilon|\mathrm{X}, \Theta_{i-1})$ where $KL(q(\Upsilon)||p(\Upsilon|\mathrm{X}, \Theta_{i-1})$ becomes zero (KL divergence is positive otherwise). The E-step thus identifies the distribution of the latent variable given the observed data and current parameter settings.

The M-step optimises with respect to $\Theta$ while holding $q(\Upsilon)$ constant:

$$\Theta_i = \arg\max_{\Theta} \int q_i(\Upsilon) \log \frac{p(\mathrm{X}, \Upsilon|\Theta)}{q_i(\Upsilon)} \, d\Upsilon \quad 7$$

The E- and M-steps are repeated until convergence to a local maxima.

A major disadvantage of the maximum likelihood method is the sensitivity to initialization parameter, which often requires heuristics or prior knowledge to identify good solutions. It also does not provide any guidance on the number of components, $\mathrm{K}$. Larger $\mathrm{K}$ may lead to better fit but low generalization of the model. The problem of choosing $\mathrm{K}$ can be addressed using variational Bayesian model selection (Corduneanu & Bishop, 2001). The method starts with a mixture model initialised with an improbably large number of potential components (or the maximum estimate), $M$, and optimises unwanted mixing components to zero. Compared to the EM method (Equation 1), the joint distribution is defined over data set $\mathrm{X}$, component means $\mu$, inverse covariance $T$ and an additional latent variable $s$, conditioned on mixing coefficients $\pi$, or $p(\mathrm{X}, \mu, T, s \mid \pi)$. The goal is then to maximise the likelihood given by $p(\mathrm{X}|\pi)$ to optimise the values of mixing coefficients $\pi$.

The latent variable introduced takes the value of $s_{k,n} \in \{0, 1\}$ for data point $n$ and component $k$. If data point $x_n$ is generated from component $k$, then $s_{k,n} = 1$, otherwise, $s_{k,n} = 0$.

The likelihood function of the data drawn is given by

$$p(\mathrm{X}|\mu, T, s) = \prod_{k=1}^{M}\prod_{n=1}^{N} \mathcal{N}(x_n|k, T_k)^{s_{k,n}}, \quad 8$$

and the latent variable is in governed by $\pi$:

$$p(s|\pi) = \prod_{k=1}^{M}\prod_{n=1}^{N} \pi_k^{s_{k,n}}. \quad 9$$

To obtain $p(\mathrm{X}|\pi)$, $\mu, T, s$ has to be marginalised from $p(\mathrm{X}, \mu, T, s \mid \pi)$. For simplification, we will bundle parameters $\mu, T, s$ as $\Theta$. Thus, we have

$$p(\mathrm{X}|\pi) = \int p(\mathrm{X}, \Theta|\pi) d\Theta, \quad 10$$

which is in the same form as Equation 3. The lower bound can thus similarly be

$$\mathcal{L}(\mathrm{Q}) = \int Q(\Theta) \log \frac{p(\mathrm{X}, \Theta|\pi)}{Q(\Theta)}\, d\Theta.$$

And $\mathcal{L}(Q)$ is maximum when $Q(\Theta) = p(X, \Theta|\pi)$. The EM procedure can then be used to optimize $\mathcal{L}(Q)$ with respect to $\pi$ in the M-step, and $Q_\mu, Q_T, Q_s$ in the E-step, resulting in a mixture model with some mixing coefficient $\pi = 0$, and is only non-zero where there is evidence for the component to exist (i.e. the cluster is valid).

### *1.3.2.2 DP-means clustering*

The DP-Means algorithm was developed as a Bayesian nonparametric approach of the k-means algorithm by Kulis and Jordan (2011). It is a hard clustering algorithm derived from Gibbs sampling for Dirichlet process mixture model. Let

- $k =$ number of clusters,
- $l_i =$ members of cluster $i$,
- $\mu_i =$ cluster centre of cluster $i$,

DP-means starts with $k = 1$, where all data points are assigned to one cluster and $\mu_1 =$ the global mean. The algorithm loops through all data points, and assigns each data point to a cluster given by its closest cluster centre; if the closest cluster centre is further away than the parameter $\lambda$, a new cluster is formed centred around this data point. At the end of the loop, the $\mu_i$'s are updated, and the process is repeated, where cluster assignment of data points are retained, reassigned or starting new clusters. Convergence is achieved when the following cost function ceases to decrease:

$$\min_{\{l_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{x \in l_i} \|x - \mu_i\|^2 + \lambda k, \qquad 11$$

where

$$\mu_i = \frac{1}{|l_i|} \sum_{x \in l_i} x. \qquad 12$$

The cost function is similar to that of the k-means algorithm, which partitions the data point into *k* clusters, with each data point belonging to the cluster with the nearest mean. In DP-means, the cost function has the additional term of $\lambda k$ (Equation 11) which penalises high number of clusters. The main differences of DP-means and k-means are (1) Unlike k-means where the

number of clusters has to be specified, DP-means requires a distance cut-off, $\lambda$, and automatically determines the best number of clusters. $\lambda$ can be estimated from physically known distances, and indirectly allows the user to define the span of the clusters. (2) Far-away/outlier data points naturally start new clusters in DP-means, whereas they get absorbed into an existing cluster in k-means if the number of clusters specified is insufficient to support them. This is especially advantageous in cases where there are no outliers and all data points are of interest; far-away data points should not be averaged down in their representation.

#### *1.3.2.3 Hierarchical Clustering*

In this project, only agglomerative hierarchical clustering by Ward's minimum variance criterion is used to achieve compact clusters with non-Euclidian distance metrics. The algorithm starts with $n$ singleton clusters, and at each stage, amalgamate the most similar pair of clusters until all subsets are in one group. The advantage of agglomerative over divisive hierarchical clustering is its lower complexity and thus computational efficiency. In the Lance-Williams dissimilarity update formula, Ward's method for the dissimilarity of clusters between a newly amalgamated class $P \cup Q$ and class $R$ takes the form of

$$d(P \cup Q, R) = \alpha_P d(P, R) + \alpha_Q d(Q, R) + \beta d(P, Q), \qquad 13$$

where

$$\alpha_P = \frac{n_P + n_R}{n_P + n_Q + n_R}, \qquad \alpha_Q = \frac{n_Q + n_R}{n_P + n_Q + n_R}, \qquad \beta = -\frac{n_R}{n_P + n_Q + n_R},$$

and $n_i =$ number of objects in cluster $i$. Although typically used with Euclidean distances, Ward's criteria for hierarchical clustering is also possible with other distance metric (Murtagh &

Legendre, 2011). For example, if we have decided to accept $P \cup Q$, and in the consideration of the amalgamation of $(P \cup Q) \cup S$, its distance to class $R$ would be

$$d((P \cup Q) \cup S\,, R) = \alpha_{PQ} d(P \cup Q, R) + \alpha_s d(S, R) + \beta d(P \cup Q, S). \quad 14$$

The first and third term can be calculated using Equation 13, while the second term is calculated using the desired distance metric for data points $S$ and $R$. Likewise, the distance between amalgamated $(P \cup Q)$ and formed $(R \cup S)$ would be

$$d(\,R \cup S, P \cup Q) = \alpha_R d(R, P \cup Q) + \alpha_s d(S, P \cup Q) + \beta d(R, S). \quad 15$$

With a symmetrical distance metric, $d(R, P \cup Q) = d(P \cup Q, R)$, and thus can be calculated with Equation 13. Hence, it is mathematically sound to use alternative distance metric that better suit the data object of interest, instead of the default Euclidean distance.

#### *1.3.2.4 Dissimilarity measures*

It will be apparent in Chapter 2 that our data objects take the form of histograms. Hence we present dissimilarity measures that compare discrete and nonparametric distributions. The following notations will be used: $\boldsymbol{x} = [x_1, x_2 \ldots, x_N]$ and $\boldsymbol{y} = [y_1, y_2 \ldots, y_N]$ are the normalised discrete histograms of data objects $\mathrm{X}$ and $\Upsilon$ with length $N$, and thus $\sum x_n = \sum y_n = 1$. The dissimilarity measures considered are summarized in Table 1.1, all of which are non-negative, and only equal zero if $\boldsymbol{x} = \boldsymbol{y}$. These will be used subsequently in Chapters 2 and 4.

*Table 1.1: Dissimilarity measures used in this project to calculate distances between discrete normalised histograms*

| Dissimilarity measure | Comments |
|---|---|
| $\chi^2$-distance<br>$$d_{\chi^2}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i - y_i)^2}{x_i + y_i}$$ | A variant to the Euclidean distance, the $\chi^2$-distance weights each bin-count difference by their sum, making the difference in each bin count proportional to their size. Large bins and small bins thus contribute to the distance equally.<br>Bound: $0 \leq d_{\chi^2}(\boldsymbol{x},\boldsymbol{y}) \leq 1$ |
| Symmetrised Kullback-Leibler divergence<br>$d_{KL}(\boldsymbol{x},\boldsymbol{y}) = KL(\boldsymbol{x}\|\|\boldsymbol{y}) + KL(\boldsymbol{y}\|\|\boldsymbol{x})$, where<br>$$KL(\boldsymbol{x}\|\|\boldsymbol{y}) = \sum_{i=1}^{N} x_i \ln\frac{x_i}{y_i}$$<br>and<br>$$KL(\boldsymbol{y}\|\|\boldsymbol{x}) = \sum_{i=1}^{N} y_i \ln\frac{y_i}{x_i}$$ | The KL divergence on its own is non-symmetric, but can be symmetrised using the sum of divergence from both ways. The KL divergence of $x$ to $y$ measures the information lost when $y$ is used to approximate $x$.<br>Bound: $d_{KL}(\boldsymbol{x},\boldsymbol{y}) \geq 0$ |
| Hellinger distance<br>$$d_H(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{N}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2}$$ | The Hellinger distance = 1 when $\boldsymbol{x}$ is non-zero where $\boldsymbol{y}$ is zero and vice versa.<br>Bound: $0 \leq d_H(\boldsymbol{x},\boldsymbol{y}) \leq 1$ |

## 1.4 Summary

In this chapter, we discussed the requirements for protein crystallization, and how typical experiments are set up to obtain protein crystals. Sparse-matrix screens have been the workhorse for the initial identification of chemical conditions suitable to induce crystallization. Advances in robotics and automation for crystallization have allowed large numbers of crystallization trials to be set up easily, but the analysis of the resulting data remains labour

intensive. We reviewed past efforts on using computer vision to automatically identify crystals with varying success, as well as alternative imaging technologies developed to increase signals from protein crystals for higher detection rate. We also introduced the machine learning algorithms used in this project, including the random forest classifier, Gaussian mixture models and hierarchical clustering paired with distance metric suitable for comparing discrete histograms, used in subsequent chapters to extract and exploit the information in our crystallization dataset.

The next chapter outlines the data mining and image processing pipeline crucial for the work described in the subsequent chapters.

## 1.5 References

Babnigg, G. & Joachimiak, A. (2010). *J. Struct. Funct. Genomics*. **11**, 71–80.

Bergfors, T. (2003). *J. Struct. Biol.* **142**, 66–76.

Bern, M., Goldberg, D., Stevens, R. C., & Kuhn, P. (2004). *J. Appl. Crystallogr.* **37**, 279–287.

Breiman, L. (2001). *Mach. Learn.* **45**, 5–32.

Bruno, A. E., Ruby, A. M., Luft, J. R., Grant, T. D., Seetharaman, J., Montelione, G. T., Hunt, J. F., & Snell, E. H. (2014). *PLoS One*. **9**,.

Buchala, S. & Wilson, J. C. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **D64**, 823–833.

Calero, G., Cohen, A. E., Luft, J. R., Newman, J., & Snell, E. H. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 993–1008.

Chayen, N. E. (1998). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 8–15.

Chayen, N. & Saridakis, E. (2008). *Nat. Methods*. **5**, 147–153.

Corduneanu, A. & Bishop, C. (2001). *Artif. Intell. Stat.* 27–34.

Cumbaa, C. a & Jurisica, I. (2010). *J. Struct. Funct. Genomics*. **11**, 61–69.

D'Arcy, A., Villard, F., & Marsh, M. (2007). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 550–554.

Dale, G. E., Oefner, C., & D'Arcy, A. (2003). *J. Struct. Biol.* **142**, 88–97.

DeLucas, L. J., Hamrick, D., Cosenza, L., Nagy, L., McCombs, D., Bray, T., Chait, A., Stoops, B., Belgovskiy, A., William Wilson, W., et al. (2005). *Prog. Biophys. Mol. Biol.* **88**, 285–309.

Desbois, S., Seabrook, S. a., & Newman, J. (2013). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 201–208.

Dietterich, T. (2000). *Mult. Classif. Syst.* 1–15.

Ferrandon, B. & Newstead, S. (2008). **1**, 466–472.

Forsythe, E., Achari, A., & Pusey, M. L. (2006). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**, 339–346.

Forsythe, E. L., Maxwell, D. L., & Pusey, M. (2002). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1601–1605.

Fusco, D., Barnum, T. J., Bruno, A. E., Luft, J. R., Snell, E. H., Mukherjee, S., & Charbonneau, P. (2014). *PLoS One*. **9**,.

Ghahramani, Z. (2004). *Adv. Lect. Mach. Learn.* **3176**, 72–112.

Gilis, D., Massar, S., Cerf, N. J., & Rooman, M. (2001). *Genome Biol.* **2**(11).

Gorrec, F. (2009). *J. Appl. Crystallogr.* **42**, 1035–1042.

Jahandideh, S. & Mahdavi, A. (2012). *J. Theor. Biol.* **306**, 115–119.

Jancarik, J. & Kim, S. H. (1991). *J. Appl. Crystallogr.* **24**, 409–411.

Kawabata, K., Takahashi, M., Saitoh, K., Asama, H., Mishima, T., Sugahara, M., & Miyano, M. (2006). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**, 239–245.

Khurshid, S., Saridakis, E., Govada, L., & Chayen, N. E. (2014). *Nat. Protoc.* **9**, 1621–1633.

Kimber, M. S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C. H., et al. (2003). *Proteins*. **51**, 562–568.

Kissick, D. J., Wanapun, D., & Simpson, G. J. (2011). *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **4**, 419–437.

Kulis, B. & Jordan, M. I. (2011). *arXiv*. 1111.0352.

Kurgan, L., Razib, A. a, Aghakhani, S., Dick, S., Mizianty, M., & Jahandideh, S. (2009). *BMC Struct. Biol.* **9**, 50.

Lekamge, B. M. T., Sowmya, A., Mele, K., Fazio, V. J., & Newman, J. (2013). *AIP Conf. Proc.* **1557**, 270–276.

Liu, R., Freund, Y., & Spraggon, G. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1187–1195.

Luft, J. R., Newman, J., & Snell, E. H. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 835–853.

Luft, J. R., Wolfley, J. R., & Snell, E. H. (2011). *Cryst. Growth Des.* **11**, 651–663.

McPherson, A. & Cudney, B. (2006). *J. Struct. Biol.* **156**, 387–406.

McPherson, A. & Cudney, B. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **70**, 1445–1467.

McPherson, A. & Gavira, J. a. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **70**, 2–20.

McPherson, A. & Kuznetsov, Y. G. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **70**, 384–403.

Mizianty, M. J. & Kurgan, L. (2012). *Protein Pept. Lett.* **19**, 40–49.

Murtagh, F. & Legendre, P. (2011). *arXiv. 1111.6285*:20

Nagel, R. M., Luft, J. R., & Snell, E. H. (2008). *J. Appl. Crystallogr.* **41**, 1173–1176.

Newman, J. (2005). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61**, 490–493.

Newman, J., Bolton, E. E., Müller-Dieckmann, J., Fazio, V. J., Gallagher, D. T., Lovell, D., Luft, J. R., Peat, T. S., Ratcliffe, D., Sayle, R. a, et al. (2012). *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **68**, 253–258.

Newman, J., Burton, D. R., Caria, S., Desbois, S., Gee, C. L., Fazio, V. J., Kvansakul, M., Marshall, B., Mills, G., Richter, V., et al. (2013). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 712–718.

Newman, J., Fazio, V. J., Lawson, B., & Peat, T. S. (2010). *Cryst. Growth Des.* **10**, 2785–2792.

Newman, J., Xu, J., & Willis, M. C. (2007). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 826–832.

Overton, I. M., van Niekerk, C. a J., & Barton, G. J. (2011). *Proteins*. **79**, 1027–1033.

Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C., & Lesley, S. a. (2003). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **59**, 1028–1037.

Pan, S., Shavit, G., Penas-Centeno, M., Xu, D. H., Shapiro, L., Ladner, R., Riskin, E., Hol, W., & Meldrum, D. (2006). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**, 271–279.

Price, W. N., Chen, Y., Handelman, S. K., Neely, H., Manor, P., Karlin, R., Nair, R., Liu, J., Baran, M., Everett, J., et al. (2009). *Nat. Biotechnol.* **27**, 51–57.

Rupp, B. (2015). *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **71**, 247–260.

Safavian, S. R. & Landgrebe, D. (1991). *Electr. Eng.* **21**, 660–674.

Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. a, & Gileadi, O. (2010). *J. Struct. Biol.* **172**, 3–13.

Segelke, B. W. (2001). *J. Cryst. Growth*. **232**, 553–562.

Shaw Stewart, P. & Mueller-Dieckmann, J. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 686–696.

Snell, E. H., Nagel, R. M., Wojtaszcyk, A., O'Neill, H., Wolfley, J. L., & Luft, J. R. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1240–1249.

Spraggon, G., Lesley, S. a., Kreusch, A., & Priestle, J. P. (2002). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1915–1923.

Stevens, R. C. (2000). *Curr. Opin. Struct. Biol.* **10**, 558–563.

Stura, E. a. & Wilson, I. a. (1991). *J. Cryst. Growth*. **110**, 270–282.

Vallotton, P., Sun, C., Lovell, D., Fazio, V. J., & Newman, J. (2010). *J. Appl. Crystallogr.* **43**, 1548–1552.

Vekilov, P. G. & Vorontsova, M. a. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 271–282.

Walker, C. G., Foadi, J., & Wilson, J. (2007). *J. Appl. Crystallogr.* **40**, 418–426.

Wampler, R. D., Kissick, D. J., Dehen, C. J., Gualtieri, E. J., Grey, J. L., Wang, H.-F., Thompson, D. H., Cheng, J.-X., & Simpson, G. J. (2008). *J. Am. Chem. Soc.* **130**, 14076–14077.

Ward, K. B. K., Perozzo, M. A., & Zuk, W. W. M. (1988). *J. Cryst. Growth*. **90**, 325–339.

Watts, D., Cowtan, K., & Wilson, J. (2008). *J. Appl. Crystallogr.* **41**, 8–17.

Wilson, J. (2002). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1907–1914.

Xu, R. & Wunsch, D. C. (2010). *IEEE Rev. Biomed. Eng.* **3**, 120–154.

Zhu, X., Sun, S., & Bern, M. (2004). *Proc. 26th Annu. Int. Conf. IEEE EMBS*. 1628–1631.

Zuk, W. M. & Ward, K. B. (1991). *J. Cryst. Growth*. **110**, 148–155.

# 2. Methods

This chapter defines the source and nomenclature of crystallization data at the Structural Genomics Consortium (SGC), Oxford used throughout the project. This is followed by a detailed description of the texton method and image processing pipeline developed to extract features from images of crystallization droplets. These aspects form the foundation for the proposed analyses described in Chapters 3 to 8: image analysis provides the objective framework to describe crystallization outcome, while information recorded in the database gives context, reference points and serve as labels for the training of machine learning algorithms.

## 2.1 Definitions and nomenclature

The following describes the source and nomenclature of crystallization data from the SGC analysed in this project, where a new term introduced is in **bold.**

The approach of the SGC is to test the crystallizability of many variations of a **target**, where a target is a particular recombinant domain of a protein of interest. Such variations may increase a target's crystallizability without significant effects on the final structure. This can be achieved by changing parts of a target's sequence, lengthening or shortening at either or both ends. A modified target is known as a **construct**. Each construct may be purified multiple times, with potentially different protocols and purification methods. Each purification is identified by a unique **purification ID**. The **sample state** of a protein used for crystallization can be **fresh** or **frozen**, whereby fresh samples have never gone through a freeze-thaw process, while frozen samples are thawed from storage at -80$^{o}$C. Crystallization of the protein sample is set up at a given **sample concentration**. Up to two small molecules may be added for co-crystallization;

these are known as **compound 1** and **compound 2**. **Incubation temperatures** of 4°C or 20°C are used to equilibrate crystallization experiments. For the purpose of this project, an **experiment** is the attempt of crystallization defined by a unique combination of purification ID, sample concentration, sample state, compound 1, compound 2, and incubation temperature.

The standard crystallization screening experiment at the SGC entails the setup of two to four sparse-matrix screens. **Screen-type** defines the specific sparse-matrix screen with 96 conditions used. The more popular screen-types at the SGC are summarized in Table 2.1. Three commonly used screens were in-house designs:

- Ligand Friendly Screen (LFS) is based on the PACT screen (Newman et al., 2005), and samples combinations of PEG 1k to 6k, salts and buffers in a semi-systematic way: the goal was that crystals obtained from here would be directly usable for studying compound-binding;

- Basic ChemSpace (BCS) consolidates a wide diversity of PEG precipitants into a single screen, by using four PEG mixtures or "smears", grouped by molecular-weight, thereby retaining the potential specific properties of a given polymer size without a combinatorial explosion (Chaikuad et al., 2015);

- JCSG+ is similar to the commercial version (Newman et al., 2005), but was rationalised with the author's help to reduce the number of stock solutions required.

Majority of analyses in this project are limited to the four most popular screen-types: JCSG+ (**JCSG**), Ligand Friendly Screen (**LFS**), Hampton Crystal Screen (**HCS**), and Hampton Index Screen (**HIN**) (see Table 2.1). These abbreviations of screen-types will be used subsequently in this thesis.

*Table 2.1: Most popular sparse-matrix screens at the SGC.*

| Screen | Abbrev. | Vendor | Product code |
|---|---|---|---|
| JCSG+ | JCSG | MD (modified) | MD1-40 |
| Ligand Friend Screen | LFS | In-house* | |
| Hampton Crystal Screen HT | HCS | HR | HR2-130 |
| Hampton Index Screen | HIN | HR | HR2-134 |
| Basic Chemspace | BCS | In-house* | |
| Modern Intelligent Dynamic Alternative Screen | MIDAS | MD | MD1-60 |
| (Emerald Bio) Precipitant Synergy Screen | EPS | JB | CS-EB-PS-B |
| Morpheus | MORPHEUS | MD | MD1-47 |
| SaltRX HT | Salt-Rx | HR | HR2-136 |
| MemGold | MemGold | MD | MD1-41 |
| MemGold2 | MemGold2 | MD | MD1-64 |

Vendor abbrev.: MD = Molecular Dimensions, HR = Hampton Research, JB = Jena Bioscience.
*In-house = SGC design, formulation by MD. Full list of conditions for LFS in Table S1 (Supplementary Materials, Section S.1), conditions of BCS in Chaikuad et al. (2015).

A crystallization screening experiment occupies a **plate** of 96 **wells**. A *plate* is thus defined by the combination of *experiment* and *screen-type*; each well in the plate contains different reservoir solutions defined by the screen-type, mixed with same experiment. Three 150nl droplets per well are prepared by mixing protein and precipitant in ratios of 2:1, 1:1 and 1:2, with Mosquito Crystal (www.ttplabtech.com) robots at room temperature. Before February 2009, the primary plate type used was CrystalQuick™ Plus 96 Well (Greiner) (flat bottom, 3 sub-wells, available from www.hamptonresearch.com), with 80µl of reservoir solution; after that, 3 Lens Crystallization Microplates (www.swissci.com) were exclusively used. These plates have 3 concaved sub-wells per well (Figure 2.1), and was developed by SwissCi in partnership with SGC, by modifying the MRC 2-Well Crystallization Plate to accommodate the SGC's preference for three droplets per condition, and to lower the reservoir volume to only 20µl.

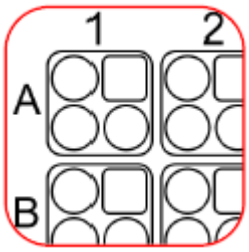



*Figure 2.1: Schematic of the 3 Lens Crystallization Microplate. The three sub-wells in a well (A1 in this figure) equilibrate with their mother liquor in a vapour diffusion experiment. Figure taken from www.swissci.com.*

Plates are incubated at 4°C or 20°C, and droplets are automatically imaged at fixed time intervals by Minstrel HT systems (www.rigaku.com). The inspection schedule is nominally 1, 2, 4, 7, 14, 28, 56 days from setup, with time tolerance tighter for early inspections; the actual inspection times varies with imaging load and backlogs.

Vendor software was set up to enable experimenters to view and score their own experiments from their desk. Whether images are actually viewed is optional, and typically only a few inspections are viewed, though viewing generally entailed seeing every drop of the inspection. Images could be labelled with a keystroke from 1 to 10 (the **Image Score**) according to the scheme outlined in Table 2.2, along with other non-crystalline labels not shown here. The score of crystallinity is broadly reliable, though inevitably carrying significant error bars and subjectivity. The sliding scale of Image Score was designed specifically to mitigate the effect of such subjectivity. However, the large number of images per plate (up to 8 inspections X 288 droplets = 2304) meant that not all images were scored, and typically only images with crystalline behaviour were labelled. Crystals that were tested for diffraction with our in-house rotating anode generator or at synchrotron sources, were tagged with a **Crystal Quality Score**, which although subjective, does however capture the fact that there was diffraction very reliably (*Crystal Quality Score* ≥ 1).

*Table 2.2: Scoring system at the SGC for interesting droplets. Other scores/classes are not shown. The tags and colour scheme are reproduced here exactly as they are configured in the software, since this is the only description of these categories that was ever seen by all users (though most were given personal training). This sparseness of information was somewhat deliberate, since it was always assumed that scoring would be largely subjective.*

| Score | Description |
|---|---|
| 1 | Granular precipitate/Phase separation/Spherolites |
| 2 | Crystalline precipitate |
| 3 | Microcrystals |
| 4 | Crystals (bit small) |
| 5 | Crystals (bit crap) |
| 6 | Mountable 6 (dubious) |
| 7 | Mountable 7 (okay) |
| 8 | Mountable 8 (nice) |
| 9 | Mountable 9 (spiffing!) |
| 10 | Mountable 10 (hell-fire!!) |

All information concerning the structure determination pipeline at the SGC, including details of the protein, its preparation and its co-crystallization compounds, along with downstream experiments up to deposition in the PDB are recorded in an in-house database, which functions in tandem with the database from the imaging system; user-assigned crystal drop scores are recorded in the imaging database, but are copied regularly to the in-house database. Figure 2.2 shows portions of the databases relevant to our analysis, which allows easy querying of all information along the pipeline from target to structure.

At the Structural Biophysics Group of the Novartis Institutes for Biomedical Research (**NIBR**), Basel, the crystallization strategy is predominantly similar to that of the SGC, Oxford. The definition of an experiment at the SGC can similarly be applied at NIBR, although with different naming conventions in the database. NIBR uses the 2-sub-well MRC plates for crystallization with similar concaved sub-wells. The main difference in crystallization strategy of the two centres is the sampling of mixing ratio: while the SGC uses 3 droplets with different mixing ratio,

NIBR samples only 1:1 mixing ratio of protein to precipitant, at larger crystallization droplet volumes of 400nl. Hence the number of droplets for each plate is typically 96, as opposed to 288 at the SGC.

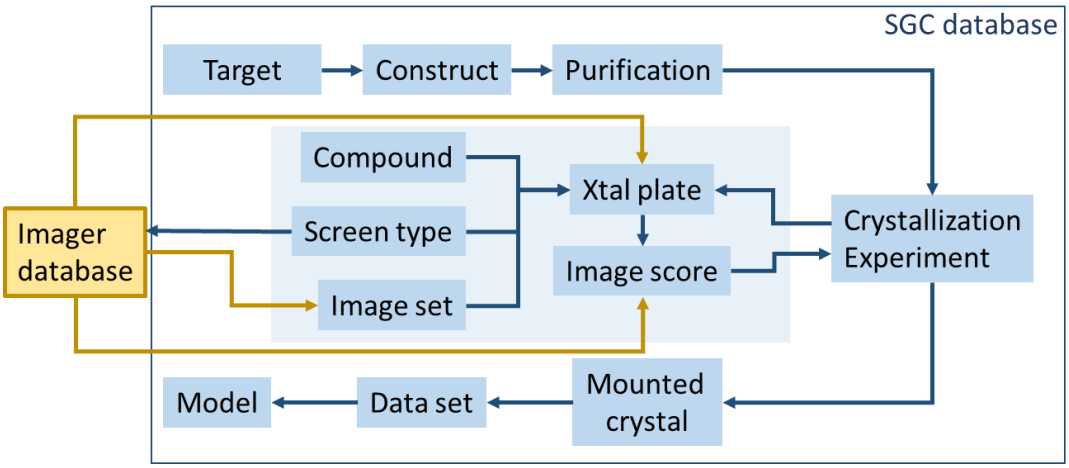



*Figure 2.2: Simplified and partial schematics of the databases used. The SGC database contains information of the entire structure determination pipeline, while the Imager database contains all information on imaging of crystallization droplets. At the SGC, both databases have been linked to parse necessary information automatically. Here, we show a very high level representation of the tables that mirror the pipeline, as well as the interaction of the Imager database and the SGC database.*

## 2.2 Characterizing precipitation patterns using Textons

### 2.2.1 Textons

Because our goal was to characterize precipitation patterns rather than identify crystals, we focused on the textural analysis of precipitation. While most computer vision efforts in crystallization previously depended on grey-level co-occurrence matrices or spectral methods (Fourier or wavelet transforms) to analyse textures (see Chapter 1.2.2), we used textons, a relatively recent method based on pre-attentive vision. The term 'textons' was first introduced by Julesz (Julesz, 1981) to describe local conspicuous features that allows for discrimination between texture pairs effortlessly. The features described by Julesz were qualitative and limited

to simple line segments, which included oriented segments, crossings and terminators. Malik *et al.* (1999) later formalised the definition of textons and generalised its use to grey-level images by mathematically describing textons as *cluster centres of filter responses*.

In image processing, filter response can be thought of as a numerical measure of the similarity between regions in the source image to the pattern of the filter. To illustrate this, Figure 2.3 shows the response of an image filtered by a black horizontal line filter. Pixels of the original image corresponding to horizontal lines give maximal response (dark red), while vertical lines and plain regions have minimal response. Hence, the more similar the pixel and its surrounding region is to the filter, the higher its filter response. If multiple filters are used, the vector of filter responses for each pixel forms a description of the pixel with respect to its surrounding, in the feature space defined by the filters used.

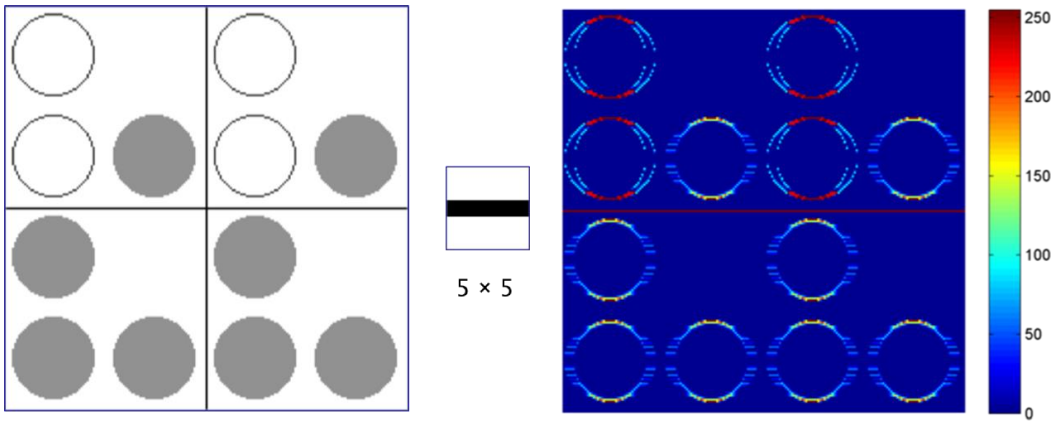



*Figure 2.3: Example of filter response. The image on the left is filtered with a 5×5 'black horizontal line' filter. The image on the right shows the heat map of filter response for each pixel. Pixels with black horizontal lines have maximal response, while vertical lines or solid background/foreground receive no response. The step portion at the top and bottom of the filled grey circles get mid-range response due to its partial similarity to the filter.*

The texton formalization by Leung and Malik (1999) is based on the argument that textures, by definition, contain repeating elements which correlate similarly to certain defined image motifs,

distributed over a region of an image. Thus, if pixels of a texture image are described individually by vectors of filter responses, the repetition or periodicity from the texture will allow these vectors to be reducible to only a few clusters. These cluster centres are thus the primitive description of textures, or textons; they can be thought of as prototypes of a texture, with other data points around them being noisy variations of these prototypes. Textons have since been used in many applications areas, including in the diagnosis Alzheimer's disease (Morgado et al., 2013), the identification of brain tumours (Islam et al., 2013), classification malignant mass regions in mammograms (Li et al., 2012) and evaluating effects of cancer treatment in animal models from ultrasound data (Gangeh et al., 2013). Outside of the medical field, textons have been used to identify skin regions for gesture detection (Medeiros et al., 2013), describe iris images for counterfeit detection (Wei et al., 2008) and ethnic classification (Qiu et al., 2007), as well as 3D face recognition (Zhong et al., 2007). Textons have also been used to, classify different real-world textures (Varma & Zisserman, 2005). However, it has not been used for analysing crystallization experiments prior to this project.

Since textons are cluster centres of filter responses, image filters are thus integral to the method. Varma and Zisserman (2005) proposed a filter bank that is rotationally invariant, which contains edge filters and bar filters (at 6 orientations and 3 scales each), a Gaussian, and a Laplacian of Gaussian filter (see Figure 2.4(a)). An image is filtered with all 38 filters in the filter bank, but only the maximum filter response for the edge and bar filters at each scale is recorded, resulting in vector with 8 (3 edge, 3 bar, 2 Gaussians) rotation-invariant filter responses per pixel. An example of the final filter response vector is shown in Figure 2.4(b). Additional advantages of this filter bank compared to that developed by Leung and Malik (Leung & Malik, 2001) include the lower dimensionality (8 vs 48), and was found to be superior due to enhanced feature detection and clustering. Hence, we chose this filter bank, but used it at half the scale

originally proposed, which was found to be suitable for the precipitation patterns within our droplet images (Ng et al., 2014).

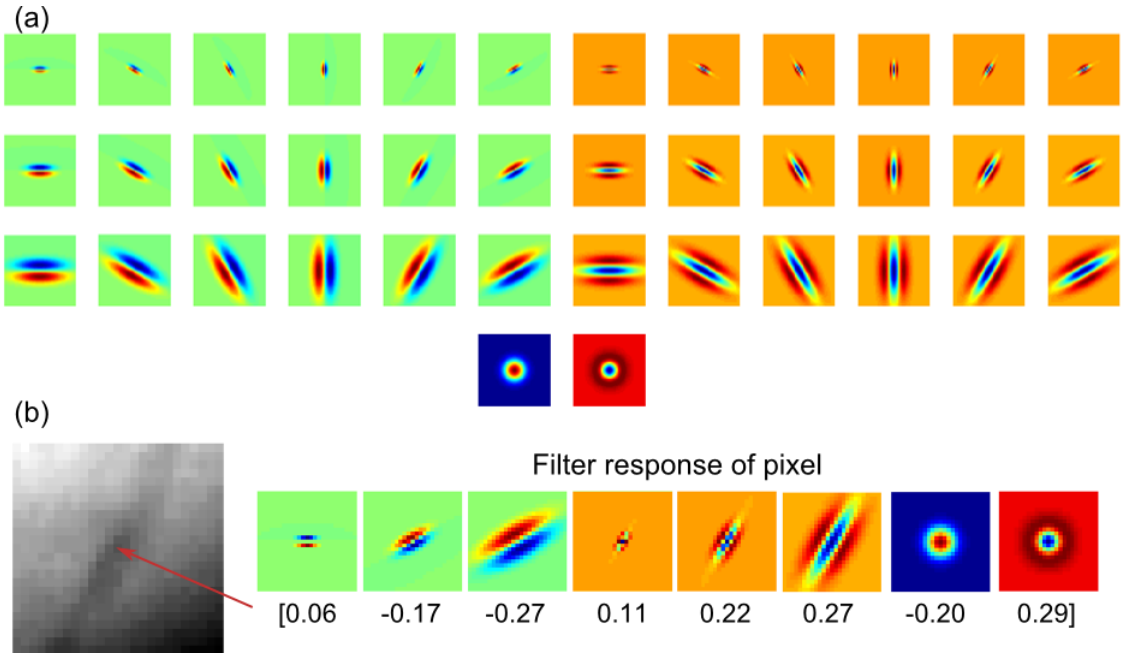



*Figure 2.4: The MR8 filter bank, as proposed by Varma and Zisserman (2005). (a) The filter bank consists of edge (left half, first three rows) and bar (right half, first three rows) filters at six orientations and three scales, and two rotationally symmetric filters (Gaussian and Laplacian of Gaussian, bottom row). (b) Example of the filter response of a pixel, where only the maximum response of the edge and bar filters at each scale are kept, resulting in a response vector with eight numbers. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

The texton method entails (1) building a texton dictionary containing unique texture prototypes, and (2) comparing the filter response of each pixel in a new image to this dictionary to find the closest texton, and label the pixel with the corresponding texton label. The conventional way of building this dictionary is by taking a fixed number of textons from each class of textures, and combining them to form the final dictionary. However, it is difficult, if not impossible to classify continuous precipitation patterns, and hence a labelled dataset was not available for generating

this dictionary. We thus took a different approach, outlined in the following section (Ng et al., 2014).

### 2.2.2 Building the Texton Dictionary

We arbitrarily selected 100 droplets with precipitates and no crystals to cover a wide range of precipitation patterns. We found these to be sufficient to produce the final dictionary and additional images did not result in new clusters. Examples of these cropped images are shown in Figure 2.5, with the complete set shown in Figure S1 (Section S.1 of Supplementary Materials).

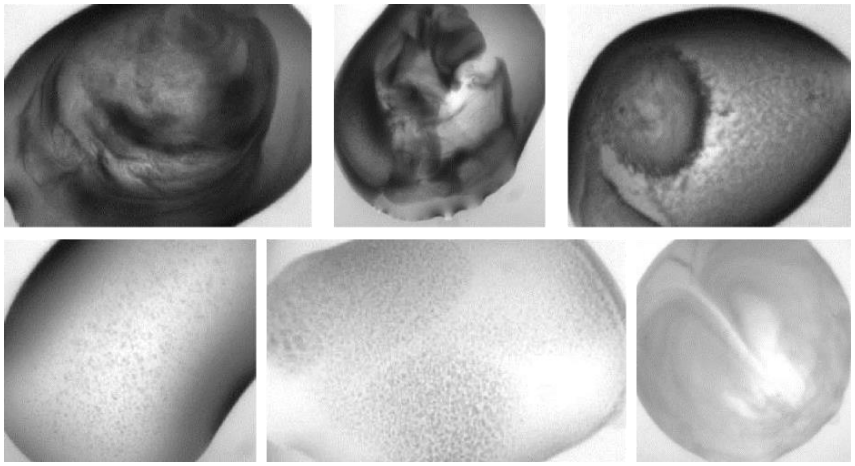


*Figure 2.5: Examples of precipitation patterns used to build the texton dictionary. All images selected at this stage had no crystals.*

In a typical texton generation scheme, K-means clustering is used to cluster filter response vectors, which requires prior knowledge or estimates of the number of clusters. Instead, we chose to cluster the filter response with Gaussian mixture models with variational Bayes model selection (Corduneanu & Bishop, 2001) to avoid restricting the number of clusters arising from one image. The variational Bayes model is typically initialised with an arbitrarily large number of clusters (Chapter 1.3.2.1), where unwanted mixture components are eventually collapsed to

zero. We started our models with 50 components, which were generally reduced to 8 to 15 clusters. Examples of the original image and its texton-labelled image are shown in Figure 2.6. The resulting textons from each image were concatenated into a library of 1319 textons.

These textons were subsequently clustered again using Dirichlet Process-means (DP-means, Chapter 1.3.2.2) (Kulis & Jordan, 2011), with $\lambda = 0.5$, where $\lambda$ is the Euclidean distance threshold to start a new cluster, resulting in a dictionary with 239 entries. This cluster-and-cluster-again approach was required because the first clustering method was not configured to remove inter-image redundancy of patterns found in the selected images. Clusters were formed independently for each image, thus similar patterns, for example, clear regions, in different images would produce similar or overlapping cluster centres (textons). These redundant textons were pruned by the second clustering method, DP-means, which clustered nearby points for a single representation, but allowed far-away, single data points to form a new cluster with some penalty, rather than grouping it into the nearest cluster. Such properties of DP-means were desirable because a far-away point in this case was not an outlier, but an actual texture found in the training data set, and hence should be kept (Ng et al., 2014).

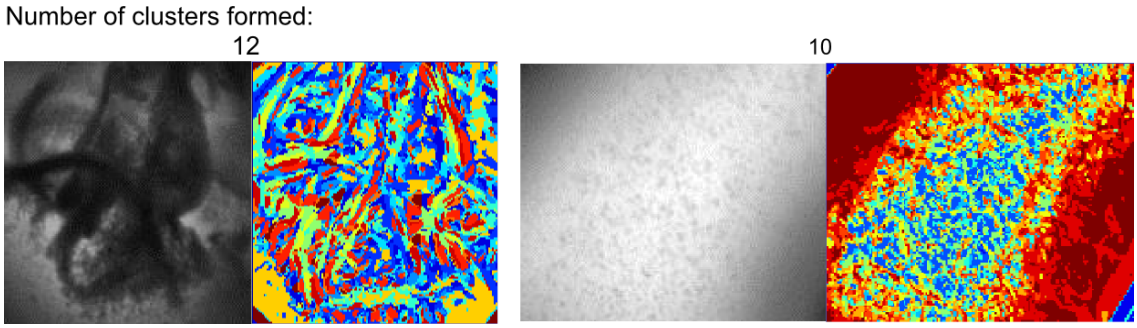



*Figure 2.6: Correlating resulting textons from a precipitation image to its original texture. Each precipitation image typically resulted in 8 to 15 textons, and here we map the textons back to the original image, with each texton label represented by a colour, in no particular colour order. The number of clusters or textons formed is stated above the corresponding images.*

To extend our image analysis pipeline for crystal detection, we added 52 crystal-containing images to our training set (see Figure S2, Section S.1 of Supplementary Materials). Similarly, these images were filtered with the filter bank and the filter response clustered with Gaussian mixture models with variational Bayes model selection. However, to ensure crystal-related textons were well represented, we only kept textons corresponding to crystals regions which were selected visually (Figure 2.7). The selected textons from the 52 images were then clustered again with DP-means with similar parameters as previously described. The resulting 61 entries were added to our previous library to give the final dictionary with 300 entries. Figure 2.8 shows a diagram of this process. For the purpose of visual representation, the dictionary was rearranged so that the first entry is the texton with the lowest magnitude, and all subsequent entries were sorted by their distance with respect to this texton. Graphically, examples of the resulting textons at $0^{o}$ orientation are shown in Figure 2.9.

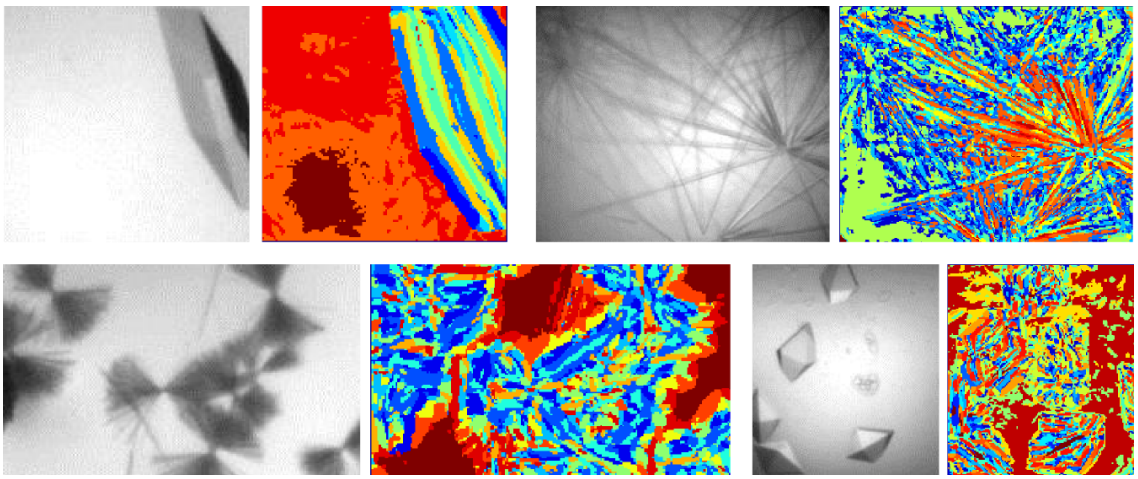


*Figure 2.7: Examples of textons generated from crystal-containing images. Instead of using all textons, we visually selected textons corresponding to crystals. Each colour corresponds to a resulting texton. For example, only the non-red textons were selected for the image at the top left. The selected textons from all images were combined and clustered again with DP-means, resulting in 61 crystal-related textons.*

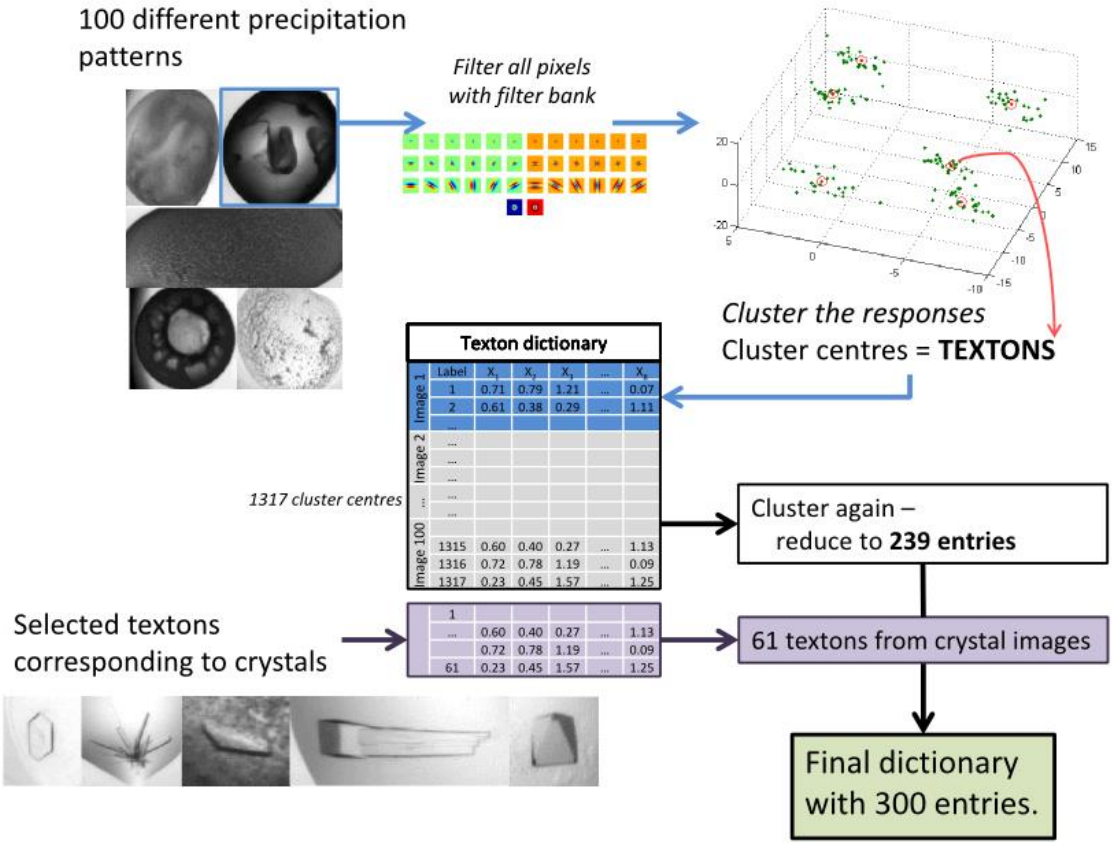



*Figure 2.8: Process of generating the texton dictionary. 100 droplets with a wide range of patterns were selected; clusters of filter responses (textons) in these images were combined to form a dictionary of 1317 entries; to remove redundancy, the dictionary entries were clustered again and reduced to 239 entries. Furthermore, selected textons corresponding to crystals were added to the dictionary to form the final 300-entry long dictionary. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

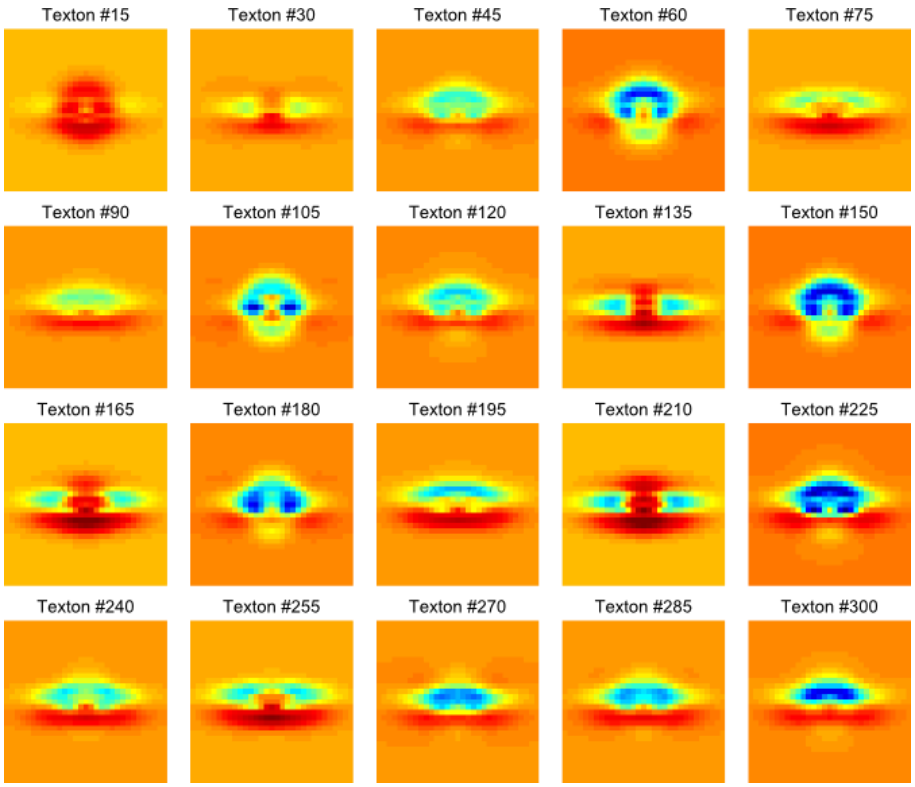



*Figure 2.9: Graphical representation of selected textons in the dictionary. Each texton shown here is a linear combination of the filters from the filter bank (edge and bar filters at $0^o$ and the circular filters) with magnitudes of each filter weighted by the texton vector. These are thus our texture prototypes for which filter responses of new images are compared to.*

### 2.2.3 Generating Features

To evaluate textures in an image, the image is similarly filtered with the filter bank. The filter response of each pixel is compared to entries in the texton dictionary, and the texton label of the closest match as calculated by Euclidean distance, is used to label the pixel. The frequency histogram of the 300 textons forms the final numerical descriptor of the image. This process is illustrated in Figure 2.10 (Ng et al., 2014).

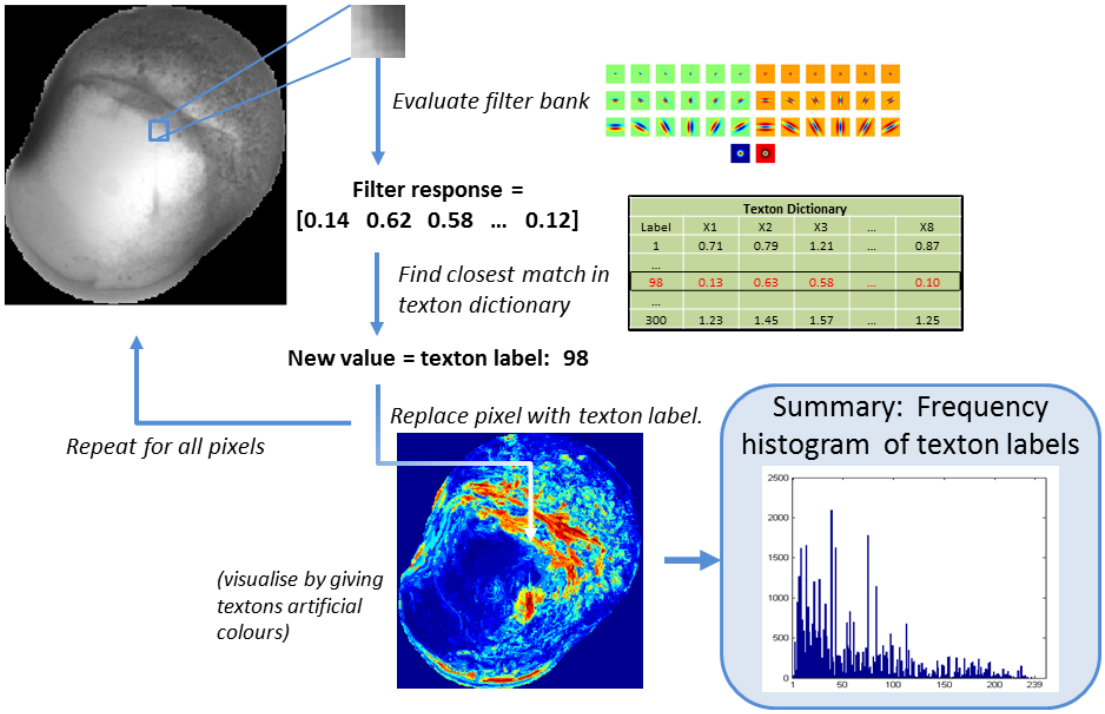



*Figure 2.10: Calculating the texton distribution from an image. Given a new image, each pixel is filtered with the filter bank and its filter response is compared with the texton dictionary generated previously. The label of the closest match in the dictionary is used to label the pixel. The final feature for each image is the frequency histogram of all texton labels in the dictionary. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

## 2.3 Image processing pipeline

Before image filtering and texton distribution can be calculated, images have to be pre-processed to filter out errors and be made compatible with the texton method. The pipeline developed to pre-process and extract texton features from a droplet has been described in our paper (Ng et al., 2014). Full analysis of droplets involves four main steps, as illustrated in Figure 2.11, each described in detail in the following sections: (1) fault detection, (2) droplet segmentation, (3) droplet pre-processing, and (4) calculation of texton distribution.

Each new droplet image is first converted to grey scale and contrast adjusted so that its grey levels cover the full spectrum of 0 to 255, and passed through a fault detection system, which identifies whether a droplet is acceptable or faulty (empty well, incomplete dispensing or camera faults). Faulty droplets are removed and not further processed. For acceptable droplets, the droplet boundary is determined automatically, and the segmented droplet is processed to correct for lighting and shadow effects around the boundary, and extra pixels are padded to extend the droplet radially. Finally, the texton distribution is calculated as described in the previous section. The pipeline was implemented in MATLAB with the Image Processing Toolbox and Statistics Toolbox (Ng et al., 2014).

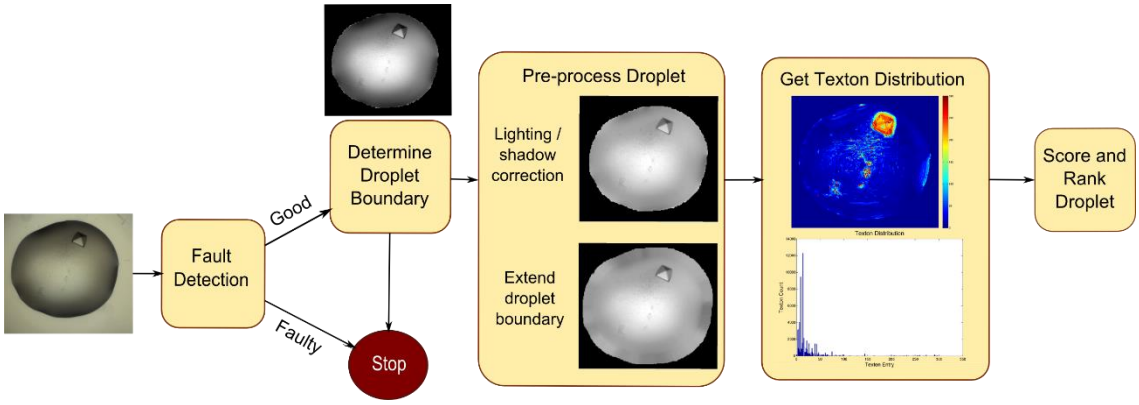



*Figure 2.11: Overview of image processing pipeline. An image is first passed through a fault detection system. If the droplet is not faulty, the droplet is segmented, corrected for lighting and shadow effects, and its boundaries are extended radially. Features are derived using the texton method (see Section 2.2 for description), which are used to score the image with a Random Forest classifier, and/or compared to predefined clusters to determine the type of precipitation. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 2.3.1 Well segmentation and fault detection

Errors due to inaccuracies of the droplet dispensing robots may result in empty sub-wells, unusually small droplets (where either the protein or reservoir solution was not added), or

droplets sitting at the edge of the sub-well. Errors in the imaging system may also produce images where the sub-wells are partially out of the field of view of the camera. Examples are shown in Figure 2.12.

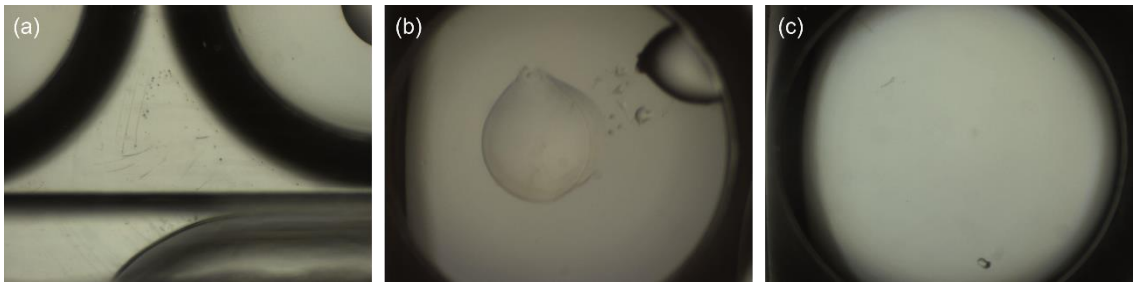



*Figure 2.12: Examples of faulty droplets. (a) Camera positioning error, and liquid dispensing error, where (b) the protein and precipitant were not mixed properly, or (c) nothing was dispensed, resulting in an empty well.*

To identify these failed droplets, background images were obtained for the each of the sub-wells by taking the average of that sub-well from all wells from an empty plate, and used to separate well region from area outside the well-frame. This should be done for every type of plate since it is plate-specific; in this case, we generated background images for the 3 Well Crystallization Microplate most commonly used at the SGC, Oxford. The background image is rigidly registered to a new image by searching for the x- and y-translation that minimises the pixel-to-pixel difference, giving the location of the well relative to the image. The area outside the well (well-frame) is masked and the remaining pixels are intensity-normalised to have a mean intensity of 0 and standard deviation of 1 (Ng et al., 2014).

The following statistical descriptors are calculated from the gradient image of the normalised image: mean, standard deviation, skewness, kurtosis, and the distribution of the absolute gradients at fixed 50 bin centres ranging from 0 to 50. These 54 features are used with the area, centroid, eccentricity (ratio of difference between foci and the major axis length), major and minor axis length of the segmented droplet (described next) as inputs to a Random Forest

classifier which predicts if a droplet image is good or faulty. The classifier was built with 500 decision trees, and trained with 11,326 images (5225 faulty and 6101 non-faulty images). The accuracy of the classifier was over 94%, with majority of false classification occurring in experiments set up with detergent. The much lower contrast of droplet boundary in such droplets (Figure 2.14(d)) makes droplet segmentation difficult and gives little contrast, and hence low differences in the difference image (Ng et al., 2014).

Since background images are central to our fault detection method (and droplet segmentation, as will be discussed in the following section), as part of deploying our code and allowing for generalization, a simple application for users to generate such background images compatible with the image processing pipeline was developed. A snapshot of the user interface is shown in Figure 2.13, where upon receiving multiple background images, the application outputs the average background image and its corresponding binary mask.

### 2.3.2 Droplet segmentation

To segment the droplet, the well-frame location as identified in the rigid registration step is used in a modified version of DroplIT (Vallotton et al., 2010). DroplIT identifies closed contours around a point where the average pixel intensity along the contour of its gradient image is extremal. Images are transformed to the polar coordinates and the circular shortest path is computed. In contrast to the original method which uses thresholds to identify the well-frame, the well-frame location is used instead to more reliably remove strong edges from the well-frame, and thus improve droplet segmentation. Comparisons of segmentation with the original and modified method are shown Figure 2.14 (Ng et al., 2014).

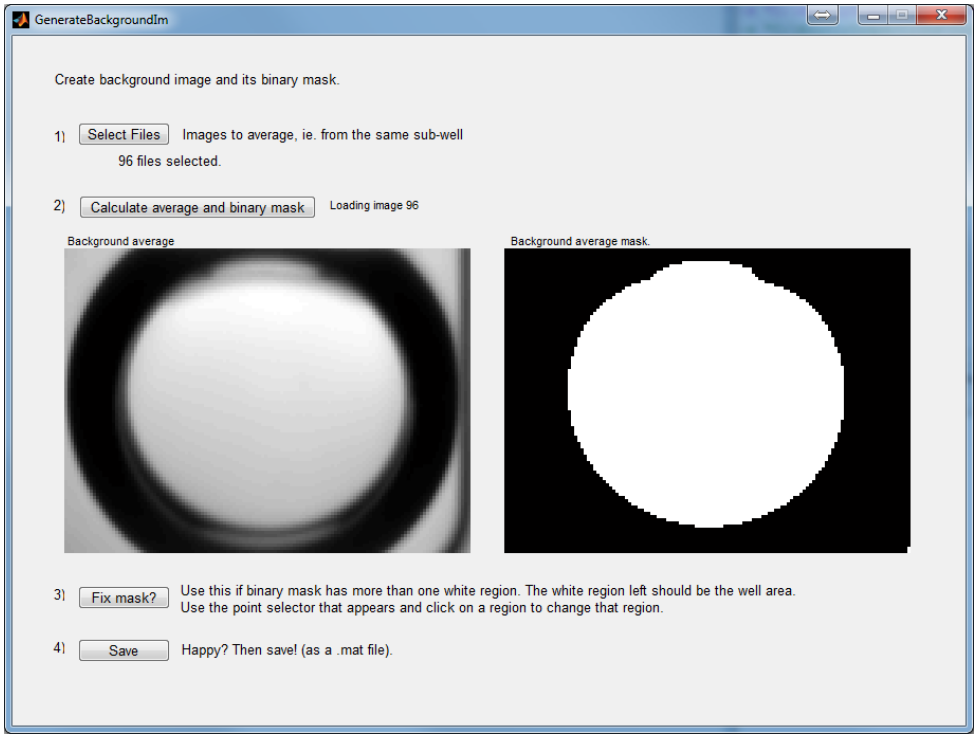



*Figure 2.13: Application for simple generation of background image file. Users will need to select empty-well images from similar sub-wells. An average image and the binary mask will be automatically generated. If there are errors with the mask, users can invert (black to white and vice versa) regions using the 'Fix Mask?' button. The generated mask will be saved as a Matlab data file (.mat), compatible with the image processing pipeline.*

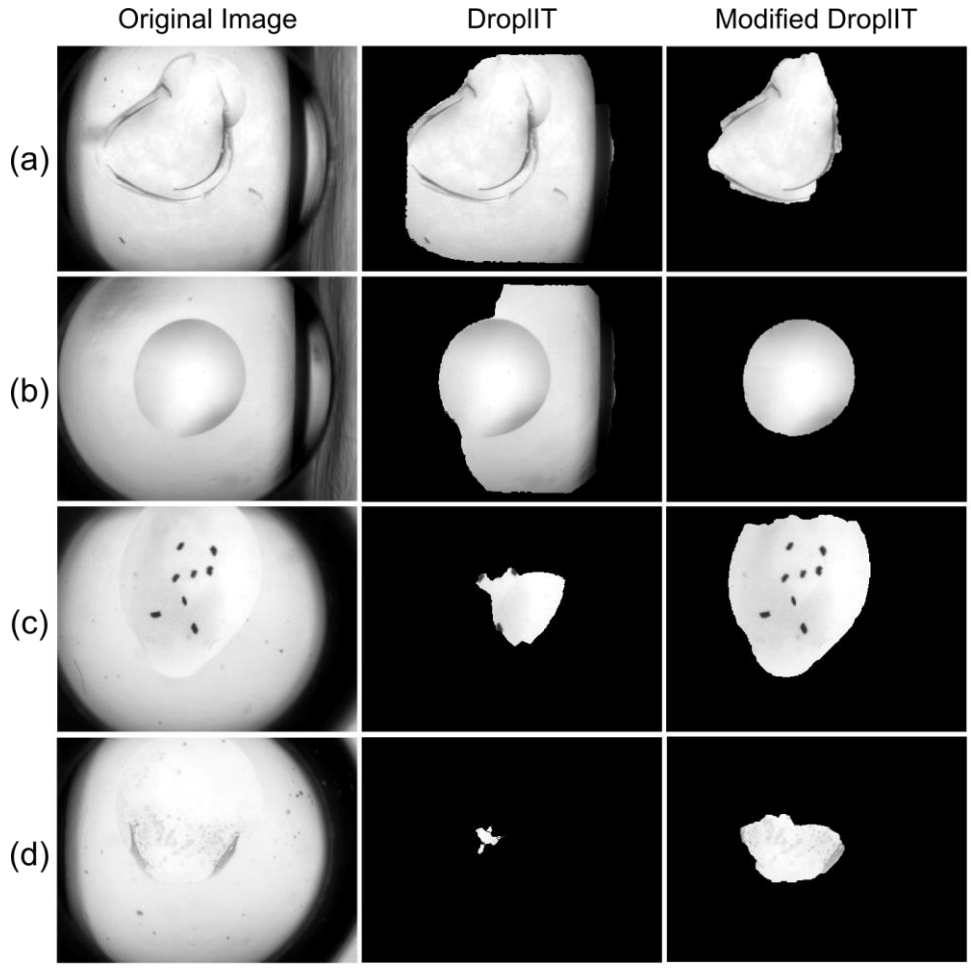



*Figure 2.14: Comparison of droplet segmentation with DropIIT (Vallotton et al., 2010) (middle column) and the modified version (right column). DropIIT may over-segment the droplet, especially if the well-frame has the characteristics in (a) and (b). (c) and (d) show more difficult examples, where the droplet boundaries are almost invisible at certain parts. However, droplets are still segmented, and useful information can still be found within the segmentation. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 2.3.3 Pre-processing droplet

#### *2.3.3.1 Lighting correction*

The segmented droplet is corrected for shadows (if any) around its edge to reduce artificial textures due to lighting conditions or droplet morphology. Gamma correction is normally applied to brighten or darken images in a non-linear scale. Typically, a single $\gamma$ value is used

uniformly across an image. However, in our case, we only wanted to boost intensities of dark pixels by the edge of the droplet, and have minimal changes to dark pixels (normally associated with precipitates) in the centre of the droplet. Hence, we introduce a non-uniform Gamma correction method, which selectively increases the intensity of darker pixels around the edge of a droplet. Corrected pixel intensities, $I_{ij_{new}}$ are calculated as:

$$I_{ij_{new}} = 255 \times \left(\frac{I_{ij}}{255}\right)^{\gamma_{ij}}$$

where 255 is the maximum intensity in a grey-scale image, $I_{ij}$ is the original pixel value at row $i$ and column $j$, and

$$\gamma_{ij} = \frac{|imLP_{ij}|}{\max|imLP|}$$

where $imLP$ is the low-pass filtered image (with Gaussian low pass filter, $\sigma = 1$) of the droplet. The low-pass filtered image of the droplet can be thought of as a blurred out image of the droplet. Lower $\gamma$ values result in higher intensity boost for darker pixels, hence with the mask (black) on non-droplet region, the dark regions outside of the droplet naturally blurs into the edge, but not the centre of the drop, allowing the edge pixels to be increased more than the centre pixels. However, the original intensities of the pixels around the edge still weights the $\gamma$ value, and hence if the image does not have strong shadows, the resulting $\gamma$ will be higher. Figure 2.15 shows the process and outcome of Gamma correction (Ng et al., 2014).

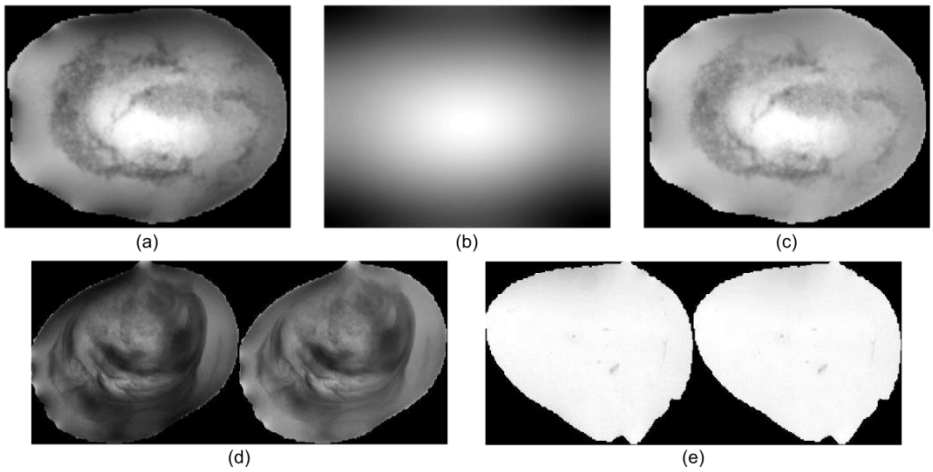



*Figure 2.15: Gamma correction to correct for shadows around droplets. The droplet (a) is filtered with a low-pass filter, resulting in a blurred image as shown in (b). The normalised values of the low-pass filtered image are then used as γ values, resulting in the corrected image in (c). (d) and (e) show more examples of droplets before (left) and after (right) Gamma correction. Shadow pixels along the edges of the droplet have been `boosted', while centre details remain relatively similar. Droplets with no dark edges should have no changes. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

#### *2.3.3.2 Droplet boundary extension*

Filter responses at the edge of droplets will be dominated by the droplet boundary. One way to avoid this is to exclude filter responses from pixels around the droplet edge when generating our histogram, but this is less ideal since crystals quite often grown near the edge of the drop. Instead, droplets are artificially extended radially, and only the filter response of this extended region is ignored in the final frequency count. This operation first requires the conversion of the cropped and segmented droplet its polar form centred at the droplet centroid. The polar image was chosen to have 360 columns, representing 360 degrees, and R rows for its radius, where R is half of the largest dimension of the cropped image. Figure 2.16(a) illustrates this operation, along with colour-wheel as an example. Extension of the droplet is done by replacing 13 pixels – equivalent to half the filter size plus 1 pixel – outside the radius of the droplet with the median

of intensities of the 10 pixels closest to the droplet boundary. The median is used instead of the mean to avoid outliers from possible segmentation artefacts. The new extended polar image is then converted back into rectangular space, resulting in a padded droplet. Due to interpolation of pixel values when converting to and from polar coordinates, analysis on the resulting padded droplet was avoided. Instead, the ring of added pixel is copied to the original droplet image for subsequent filtering. The ring provides sufficient neighbouring pixels for filter operations on the pixels in the original edge. However, since they are artificially introduced, the filter responses for pixels in the extended ring are ignored in the final histogram count. Figure 2.16(b, c, d) shows a comparison of texton labels with and without droplet extension (Ng et al., 2014).

### 2.3.4 Calculation of texton distribution

To achieve intensity invariance, the Gamma-corrected and extended image is normalised to have $\mu = 0$ and $\delta = 1$. The texton distribution is then calculated as described in 2.2.3. Computation time from reading an image to generating the texton distribution histogram is on average under 1.5s on a Windows 7 machine with 8GB of RAM and Intel® Core™ i5-2500.

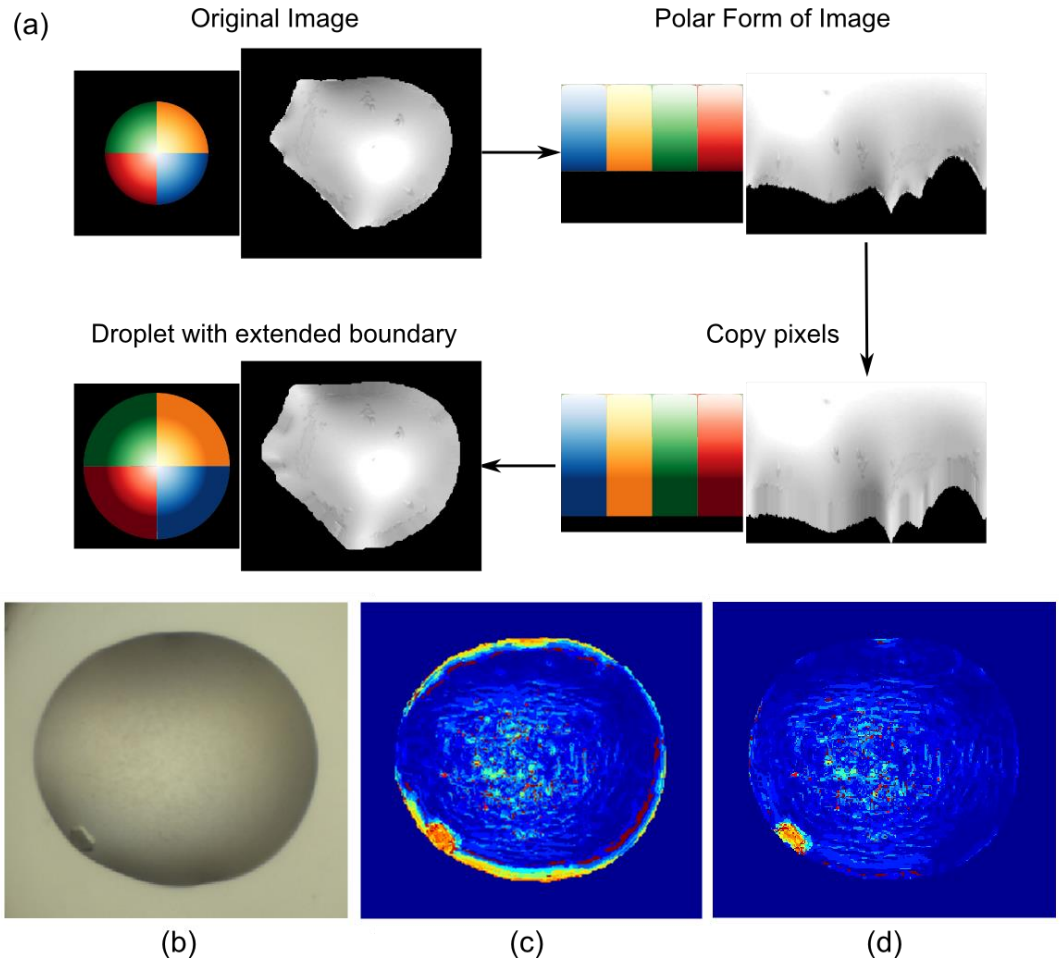



*Figure 2.16: Extending droplet boundary. (a) A colour wheel is shown beside the droplet for illustration. The segmented droplet is cropped and converted to its polar form. The droplet boundary is extended replacing 13 pixels beyond the boundary in each column with the median of intensities in the 10 pixels closest to the boundary in the corresponding column. The image is finally converted back to Cartesian space. The padded pixels should be similar to the boundary it was derived from, with some variations due to interpolation errors. Effects of droplet boundary. (b) A typical example crystal growth at the edge of the droplet. (c) The strong edge of droplet boundaries gives strong signal that often correspond to crystal edges, either masking the presence of crystals by the edge or creating false positive signals. (d) By extending the droplet boundary to avoid the strong edge, noise can be supressed for better crystal detection. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 2.3.5 Deploying the pipeline

All crystallization images captured at the SGC, Oxford and Novartis Institute for Biomedical Research (NIBR, Basel) are automatically analysed through the image processing pipeline. Using Matlab's deploytool, the script and functions for the above pipeline were compiled into an executable that only requires the freely available Matlab Compiler Runtime (Mathworks) to run on 64-bit Windows 7 and Windows 8.1 machines. An hourly task is scheduled to analyse all images captured in the previous hour. The resulting feature vectors are written into tables created in an existent database for future queries and further applications discussed in subsequent chapters. At the SGC, the average processing time per plate of 288 images is under 6 min on an Intel Core i5-2500 CPU with 8 GB RAM, which is well below the image acquisition time of ~10 min per plate with the Minstrel HT, and causes little to no backlog in processing even though two imagers are running concurrently. At Novartis, a plate of 96 images is processed in less than 4 min on an Intel Xeon X5550 CPU with 8 GB RAM, which also easily accommodates the typical image-acquisition time of less than 6 min for 96 wells with the Rock Imager (Ng et al., 2014). The code has been optimized for Matlab operations, although not for parallelized execution, which was found to be unnecessary at the moment.

At the SGC Oxford, we have added a second random forest classifier to the fault detection stage, to further identify faulty images as error-droplets (liquid handling error, camera error) or empty wells. A similar 500-tree random forest classifier was trained with 5764 images (3195 faulty droplets and 2569 empty wells). This additional information is used in a daily report of the crystallization infrastructure: an automatic email is sent out at midnight detailing the proportion of faulty droplets in new plates set up, to provide a quick overview for consistent liquid handling errors or imaging system problems.

## 2.4 Validation of textons as a quantifier of droplet content

### 2.4.1 Comparing droplets

Since droplets are objectively described by their texton distribution histograms, they can be compared using distance metrics for comparing discrete distributions, including the $\chi^2$-distance, symmetrised KL-divergence, and Hellinger distance (Chapter 1.3.2.4). As a representative example, the Hellinger distances between histograms of droplets are shown in Figure 2.17.

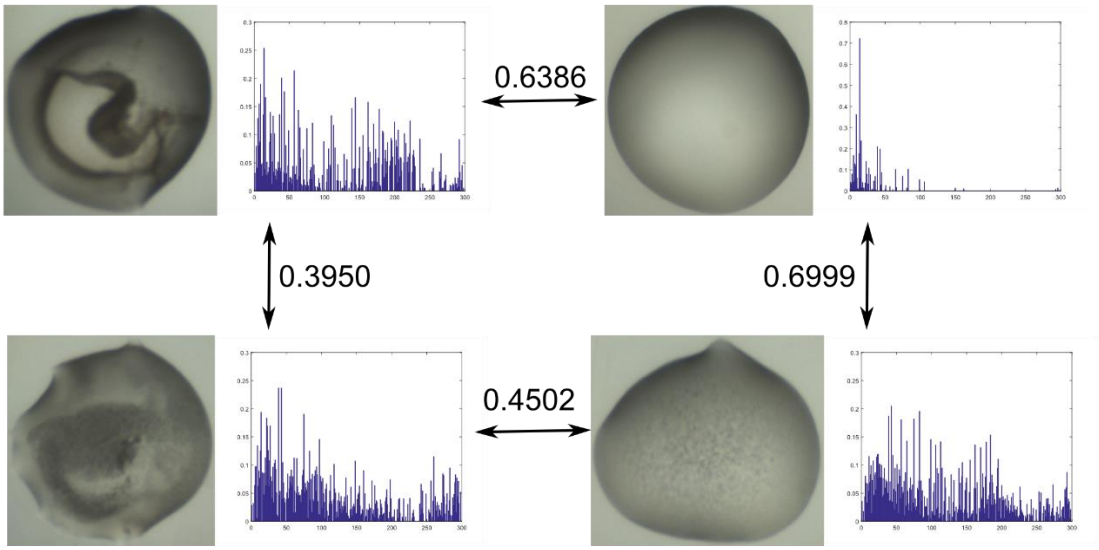



*Figure 2.17: Distance between two droplets defined by the Hellinger distance of the respective histograms. This allows the quantification of how 'different' two droplets are from each other on a scale of 0 (identical) to 1 (different).*

To validate our method, precipitation patterns of 1400 images randomly selected from 28 crystallization plates were clustered using hierarchical clustering with Ward's minimum variance method (Chapter 1.3.2.3). This follows the argument that if the texton method is able to describe and discriminate droplets, droplets with visually similar behaviour should cluster together, and was necessary since there were no labelled datasets of precipitation by classes or

types available The resulting 50 clusters were manually inspected and found to have visual consistency; examples of these clusters are shown in Figure 2.18 (Ng et al., 2014).

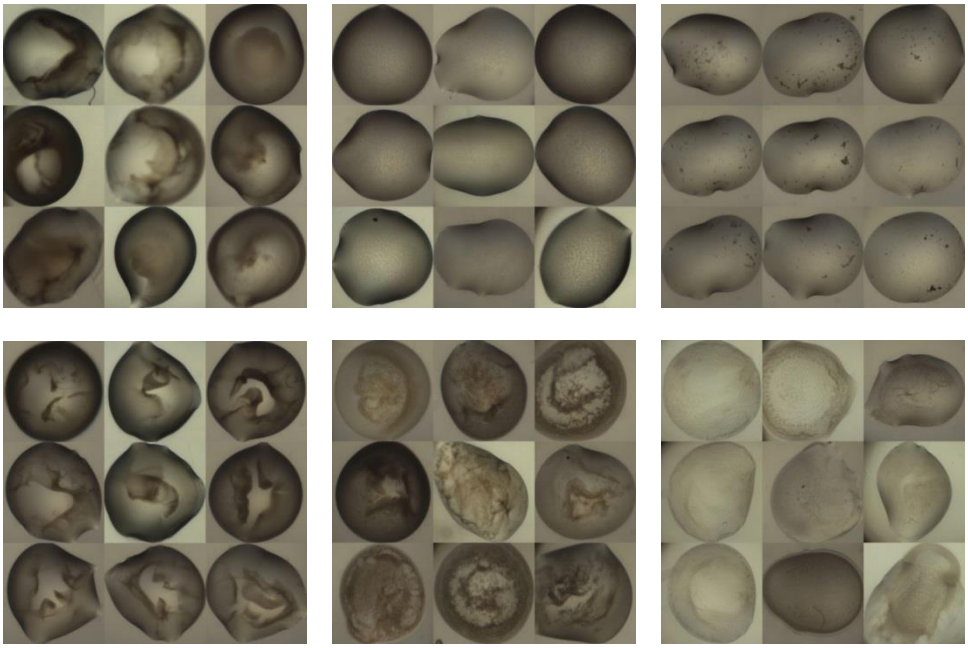


*Figure 2.18: Examples of clusters of precipitation patterns generated from hierarchical clustering with Ward's minimum variance and $\chi^2$-distance. The clusters had visually consistent precipitation behaviour, indicating that the features produced can group similar precipitation patterns together. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 2.4.2 Comparing plates

While the normalised histograms for each droplet can be compared robustly with established distance metrics, constructing a global plate-to-plate distance was less clear. For the plate-to-plate distance to be useful, the final distance measure should allow for simple comparison and ranking. We chose a bottom-up approach: the arithmetic mean of distances between pairs of droplets in the plates was used, ignoring distances that cannot be calculated due to faulty droplets (Figure 2.19). It should be noted that plate-to-plate comparison is only meaningful when the precipitant in corresponding wells are identical; this can easily be achieved by only

comparing plates of the same screen-type. The distance thus reflects the average difference in the response of two experiments (or protein sample) in a common chemical space.

The mean of distances, as simple as it may be, captures the global difference between plates, and does allow for the direct ranking and comparison of distances. This, however, may be too simplistic a summary, and the assumption of a normal distribution of distances between corresponding droplets may be inaccurate. However, complex modelling of the distribution of drop-to-drop distances on a plate, for example, using Gaussian mixture models, do not lend themselves to direct comparisons or ranking, which is central to our method in Chapter 4. Further work should be carried out to identify better methods to calculate distances between plates that allows the ordering by similarity.

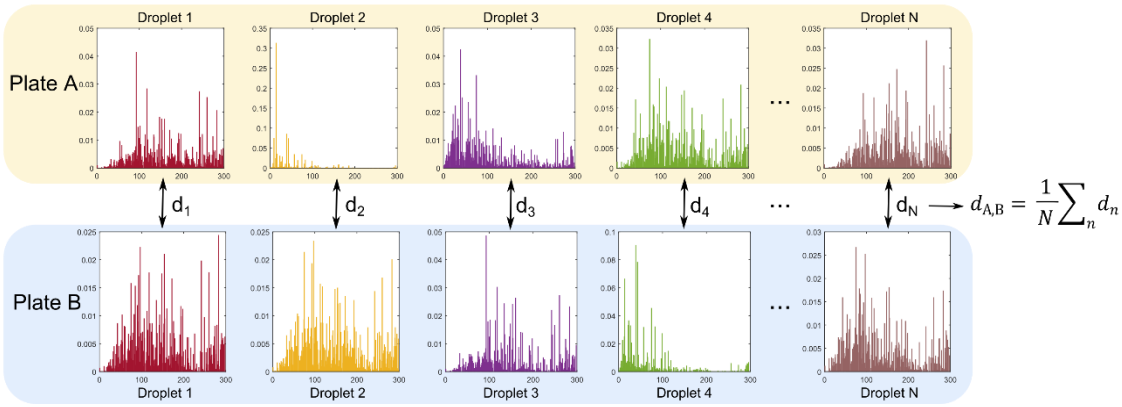



*Figure 2.19: Plate-to-plate distance is calculated by taking the mean of distances between corresponding droplets. The crystallization condition in each corresponding well should be the same, and hence the distance is a measure of the difference in response of two different experiments defined by the same chemical space.*

We evaluated the suitability of the distance metrics tested by identifying if they consistently produced minimal distance for repeated experiments. 4 plates were set up consecutively with different protein samples and different sets of precipitate. A 9-fold repetition on each plate was achieved by (1) repeating the set of conditions in columns 1 to 4, 5 to 8, and 9 to 12, and (2)

each of the three sub-wells had identical 1:1 mixture of protein to precipitant. We thus have 9 similar sets (3 blocks × 3 sub-wells) of 32 crystallization droplets (8 rows × 4 columns) in each plate. Details of the protein sample used and conditions used are described in Table 2.3.

Each set of 32 droplets was treated as an experiment, and the mean of pairwise distances between experiments were calculated. Figure 2.20 shows a heat-map of distances calculated with $\chi^2$-distance, symmetrised KL-divergence and Hellinger distance. We found that all metrics gave relatively lower distances for repeats (blue blocks across the diagonal). The symmetrised KL-divergence resulted in the lowest variance for intra-repeat distances (0.0017 vs. 0.0034 ($\chi^2$-distance) and 0.0032 (Hellinger distance) when all three were normalised by their respective maximum). However, the lack of an upper-bound for symmetrised KL-divergence makes it less intuitive for selecting cut-offs, and hence we chose Hellinger distance for its boundedness between 0 and 1, and the lower variance compared to the $\chi^2$-distance.

It should be noted that the distances between different proteins, including the distances between experiments of JMJD2BA-p049 and other experiments, were not meaningful in this analysis. This is because different conditions were purposely used for related proteins to maximise the contrast between the protein samples, and thus increasing the signal strength of similarity between repeats. Because the conditions were different, comparing the response of proteins in different chemical space is thus meaningless.

*Table 2.3: Plates set up as control dataset. Each plate was divided into three repeated blocks (4 columns each) with conditions from 4 columns of a screen-type. Further information on the protein samples can be found in Table S2 in Supplementary Materials*

| Barcode | Protein | Concentration | Screen-type | Columns of screen used |
|---|---|---|---|---|
| CI038734 | JMJD2AA-p086 | 11.5 | HCS | 1, 2, 3, 4 |
| CI038735 | PHIPA-p022 | 18 | HCS | 1, 2, 3, 4 |
| CI038736 | JMJD2BA-p049 | 8 | HIN | 1, 3, 5, 7 |
| CI038737 | PHIPA-p023 | 11 | HIN | 1, 3, 5, 7 |

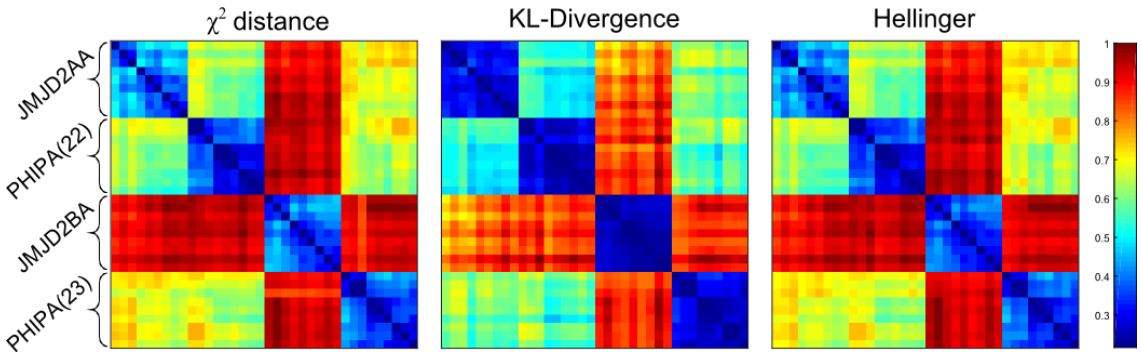



*Figure 2.20: Distance matrix of repeated control experiments. Each element represents the normalised distance between two experiments, where each experiment consists of a set of 32 droplets. The blue squares along the diagonal shows the lower distances calculated between repetitions, which was expected since the precipitation patterns among repeats should be similar to each other.*

Extending the analysis further as a sanity check, we clustered 1505 plates of the four popular screen-types (JCSG, LFS, HCS and HIN). The resulting clusters showed that different screen-types tend to cluster together (Figure 2.21). This was expected since the chemical conditions are constant for a screen-type, and hence proteins would behave more similarly in the same screen when compared to other screen-types. This agreement with expectation further validates our texton method. However, there is also more than one 'behaviour' for each screen-type, shown by their appearance in different branches of the dendrogram in Figure 2.21, indicating different groups of responses that a screen can generate.

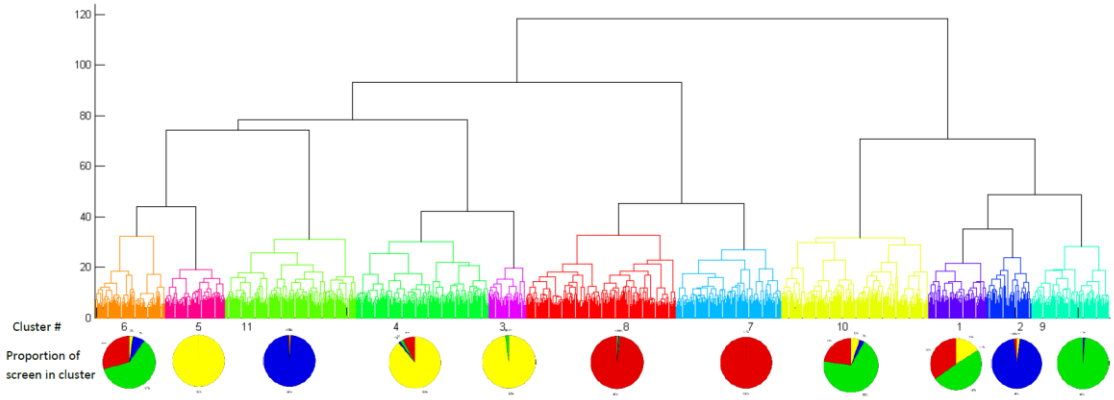



*Figure 2.21: Hierarchical clustering of 1505 plates at 4°C. Each node in the dendrogram represents a plate. The pie charts at the bottom shows the fraction of screens in the cluster with the following key: red – JCSG, green – LFS, blue – HCS, and yellow – HIN. There is no correlation between colours in the dendrogram and pie chart keys.*

## 2.5 Discussion and further work

The false colour representation (examples in Figure 2.22) of textons corresponds well with human perception of droplet images. Compared to features from spectral methods, which often requires simplification via statistical summary, further reducing the information content, each feature in the texton method is meaningful and represents a complex behaviour.

In our calculation of texton distribution, the closest match of a pixel's filter response to the dictionary entries was calculated with Euclidean distance, which is the commonly used distance measure in texton research, and hence was our natural choice. At the same time, distance measures remain an unresolved question in the field, and there is no *a priori* reason to believe the Euclidean metric is meaningful for this application; thus, exploring other distance measures,

and identifying those that are consistent with human perception, may significantly improve the algorithm's performance; this should be addressed in future work.

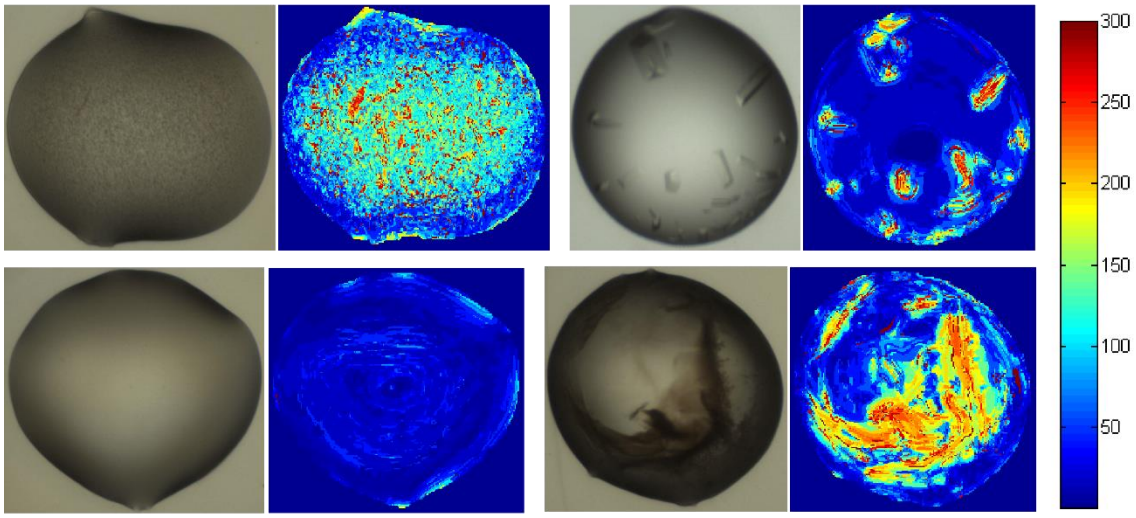



*Figure 2.22: Examples of droplets and the corresponding texton labels for each pixel, colour coded accorded to the arrangement of textons in the dictionary. The artificial colouring of droplets by their texton labels often correspond to human perception.*

The image processing pipeline has also been developed and tailored for vapour-diffusion experiments. Improvements for each step of the image processing pipeline will undoubtedly lead to increased performance in all subsequent application areas. Currently, fault detection and droplet segmentation are less accurate on droplets with detergent due to low contrast from droplet boundary; our algorithm output are thus less accurate for membrane protein targets. While we expect the texton dictionary to be effective and applicable for other experiment formats, it is almost certain that modifications to the image processing pipeline, especially at the pre-processing stage, will be required to accommodate for the differences.

Information content from each droplet may be increased by including the following: (1) Change of droplet behaviour over time, which requires time-point comparison; (2) Extension of the

algorithm to other imaging modality, for example UV-fluorescent images or SONICC images of crystallization droplets. We have applied the pipeline and texton dictionary to neither of these imaging modalities. Challenges anticipated in the analysis of UV-fluorescent images include the normalisation of UV intensity to account for its dependence on the intrinsic fluorescence of proteins or labelling concentrations. For SONICC images, segmentation of droplet will be difficult since no boundary signal is observed; cross-imaging modality registration will be required and non-trivial, especially with different camera resolutions, which is the case in current systems.

## 2.6 Concluding remarks

The above sections described the terminologies used throughout this thesis, as well as the texton method central to our analysis of crystallization experiments. The well-captured and curated metadata attached to crystallization experiments at the SGC and NIBR in structured databases, together with the objective description of crystallization outcome from our image processing pipeline, provide a framework for the understanding, learning, and training of algorithms. The following chapters build on this framework, starting with the ranking of crystallization droplets based on their probability crystallinity to address the multitude of images to view and make judgement calls.

## 2.7 References

Chaikuad, A., Knapp, S., & von Delft, F. (2015). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 1627–1639.

Corduneanu, A. & Bishop, C. (2001). *Artif. Intell. Stat.* 27–34.

Gangeh, M. J., Sadeghi-Naini, A., Kamel, M. S., & Czarnota, G. J. (2013). *Proc. - Int. Symp. Biomed. Imaging*. 1372–1375.

Islam, A., Reza, S. M. S., & Iftekharuddin, K. M. (2013). *IEEE Trans. Biomed. Eng.* **60**, 3204–3215.

Julesz, B. (1981). *Nature*. **290**, 91–97.

Kulis, B. & Jordan, M. (2011). *Proc. 29th Int. Conf. Mach. Learn.*

Leung, T. & Malik, J. (2001). *Int. J. Comput. Vis.* **43**, 29–44.

Li, X., Williams, S., Lee, G., & Deng, M. (2012). *2012 12th Int. Conf. Control. Autom. Robot. Vis.* **2012**, 5–7.

Malik, J., Belongie, S., Shi, J., & Leung, T. (1999). *Proc. Seventh IEEE Int. Conf. Comput. Vis.* **2**,.

Medeiros, R., Scharcanski, J., & Wong, A. (2013). *Multimed. Expo Work. (ICMEW), 2013 IEEE Int. Conf.*

Morgado, P., Silveira, M., & Costa, D. (2013). *2013 IEEE Int. Work. Mach. Learn. Signal Process.* 0–5.

Ng, J. T., Dekker, C., Kroemer, M., Osborne, M., & von Delft, F. (2014). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 2702–2718.

Qiu, X., Sun, Z., & Tan, T. (2007). *IEEE Int. Conf. Image Process. 2007. ICIP 2007.* **2**, 405–408.

Vallotton, P., Sun, C., Lovell, D., Fazio, V. J., & Newman, J. (2010). *J. Appl. Crystallogr.* **43**, 1548–1552.

Varma, M. & Zisserman, A. (2005). *Int. J. Comput. Vis.* **62**, 61–81.

Wei, Z., Qiu, X., Sun, Z., & Tan, T. (2008). *2008 19th Int. Conf. Pattern Recognit.* 1–4.

Zhong, C., Sun, Z., & Tan, T. (2007). *2007 IEEE Conf. Comput. Vis. Pattern Recognit.* 1–6.

# 3. Ranking Crystallization Droplets by the Likely Presence of Crystals

The need for sampling large chemical space to identify crystallization conditions, and the ease of doing so with the development and widespread deployment of robotics for high-throughput crystallization have led to experimenters being typically confronted with evaluating very large numbers of crystallization trials. The most straight forward analysis of screening experiments is the visual inspection of images to identify crystals. This task is challenging, time-consuming and subjective even though the images are electronically captured (McPherson & Gavira, 2014). A quick yet robust assessment of the presence of crystallinity can be hugely beneficial in identifying the gene construct or protein preparation most likely to crystallize, or reducing the time require for identifying crystals in parallel co-crystallization experiments and soaking experiments. Robotic imaging of droplets offers an opportunity to do this, and this chapter presents an algorithm that seeks to exploit it (Ng et al., 2014).

We revisit the problem of crystal recognition, both using recent techniques of texture analysis and drop identification, and reconsidering how the results can be most usefully presented to allow quick identification of a protein preparation's propensity to crystallize. Previous published studies focused on automatically classifying crystallization droplets into distinct but ultimately arbitrary experiment outcomes (see Chapter 1.2.2). Rather than classifying drops into discrete categories of experimental outcomes, which has been consistently insufficiently accurate in all reports, we circumvent the question by instead ranking drops based on their precipitation behaviour, thereby rearranging the viewing order in an experimentally meaningful way, as opposed to the common top-left to bottom-right viewing sequence of droplets that is merely an artefact of the image numbering. This is in the spirit of Liu *et al.* (2008), who sought to identify harvestable crystals greater than 10µm; but our algorithm additionally identifies droplets with

microcrystals and showers of crystals, which are equally interesting to the question of crystallizability. We show the enrichment of crystal-containing images early in the viewing order, and how crystals and microcrystals are found more efficiently and accurately when viewed in rank order.

Majority of the following sections have been published (Ng et al., 2014), including the accompanying figures and tables, reproduced with permission of the International Union of Crystallography. We will avoid repeated referencing of this paper henceforth. We first describe the datasets used to train our classifier and its validation in cross-platform imaging, followed by analysis of the algorithm's performance as well as its effectiveness in practical usage.

## 3.1 Data sets and training algorithm

Our approach entails first objectively characterizing crystallization droplets with the texton method described in the previous chapter. The normalised frequency histogram of textons is used as input to a random forest classifier, though not to classify the drops, which would entail selecting a classification threshold, but instead to obtain the posterior probability that is output by the classifier, which is used as a score to rank the droplets.

### 3.1.1 Dataset

Crystallization experiment droplets images automatically captured with the Minstrel HT system (Rigaku) at the SGC formed our training dataset, along with the associated Image Scores (Table 2.2) given by experimenters. For this chapter, we simplify the Image Scores to non-crystallinity (1 to 2), microcrystals (3 to 5), and mountable crystals (6 to 10). It should be noted that labelling is optional, and it is usually only the interesting droplets that are labelled.

Two sets of training images were selected: (1) *'interesting'* images: random subset of 2501 images of droplets that were given image scores of 3 and above captured between April 2013 and July 2013; (2) '*uninteresting*' images: random subset of 3553 images from the same period with image scores < 3 or no scores recorded.

### 3.1.2 Training algorithm

The set of 2501 '*interesting*' droplets and 3553 '*uninteresting*' droplets were used to train a 2-class Random Forest classifier (Breiman, 2001) with 500 decision trees using Matlab's Treebagger function. The Random Forest algorithm was chosen for its speed and ability to deal with both the large numbers of instances and features, which rule out competing algorithms. Given a set of features from new test images, the Random Forest produces the posterior probability for a particular class, which is used directly as scores for the images. In this case, for the '*interesting*' class, a score of 1 indicates that the droplet is likely to contain crystals or crystalline behaviour, while a score of 0 indicates it is '*uninteresting*'. In a typical classification exercise, a threshold is set to determine which class the data point belongs to. 10-fold cross validation of the classifier at a cut-off of 0.5 gave an average area under ROC curve of 0.9418 ± 0.0027, indicating good separation of scores for both classes, and an average accuracy of 0.8930 ± 0.0044. However, as the intention is to rank and not classify droplets, no threshold was selected; instead the scores were used to rank droplets directly.

### 3.1.3 Validation and algorithm performance

The aim of the proposed ranking method is for wells containing crystals to be scored high and consequently be moved to the top of the rearranged viewing order. To test the algorithm, a

separate set of images was selected from 196 plates set up at SGC Oxford over a different date range (July to September 2013), with at least one recorded crystal (label ≥ 3, scored by SGC crystallographers). Of these, 101 plates were sparse matrix screens, while the remaining 95 were optimization experiments. Each plate was set up with the standard 2:1, 1:1, and 1:2 protein to precipitant mixing ratio. Droplets in these plates were ranked, and the highest rank of crystals marked by crystallographers was determined, either by sub-well or by well, where the maximum score of its three sub-wells was used.

### 3.1.4 Cross-imaging-platform application

A common issue in machine vision algorithms is transferability to different imaging platforms or data sources. For validation of the robustness of the algorithm across imaging platforms, we analysed images acquired by the Structural Biophysics group of the Novartis Institute for Biomedical Research (NIBR, Basel), using the Rock-Imager (Formulatrix). Images of 134 plates with at least one recorded crystal were selected. A comparison of this dataset and that of the SGC is summarized in Table 3.1. These images were captured in the Extended Focus Imaging mode, where multiple focal depths are combined to form the final image, hence producing sharper images across the droplet. A combination of seal materials and polarizing optics also produces more colourful images. Figure 3.1 shows a comparison of SGC images and Novartis images. Background images for the type of plate used were obtained similarly as described in Chapter 2.3.1.

Droplet segmentation proved to be more difficult due to the sharp edges of precipitates within the droplets, for example, Figure 3.1(c). When DroplIT fails on the original image, it is instead applied to a gamma-corrected image, as described in Chapter 2.3.3.1 (with a Gaussian low pass filter, $\sigma = 10$), which blurs large, dark regions often associated with precipitates, but the high

frequency change of droplet boundary is not affected; or a blurred version of the image. This reduces the failure of droplet segmentation rate from 12% to 2%. The same analysis of determining the highest rank of crystals marked by crystallographers was carried out.

*Table 3.1: Comparison of validation dataset acquired from SGC, Oxford and NIBR, Basel. Table taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

| | **SGC** | **NIBR** |
|---|---|---|
| **Total number of plates:** | 196 | 134 |
| **Sparse-matrix screens** | 101 | 114 |
| **Optimization screens** | 95 | 19 |
| **Scoring Criteria:** | | |
| **Interesting** | Label ≥ 3 | Label = {Phase Separation, Salt, Microcrystals, Needles (1D), Plates (2D), Crystals (3D), Interesting} |
| **Uninteresting** | Label < 3 | Label = {Clear, Precipitate} |
| **Additional differences** | | |
| **Num. sub-wells used** | 3 | 1 |
| **Imaging mode** | Single focus depth | Extended focus imaging |
| **Optics** | No polarization effect | Visible polarization effect |

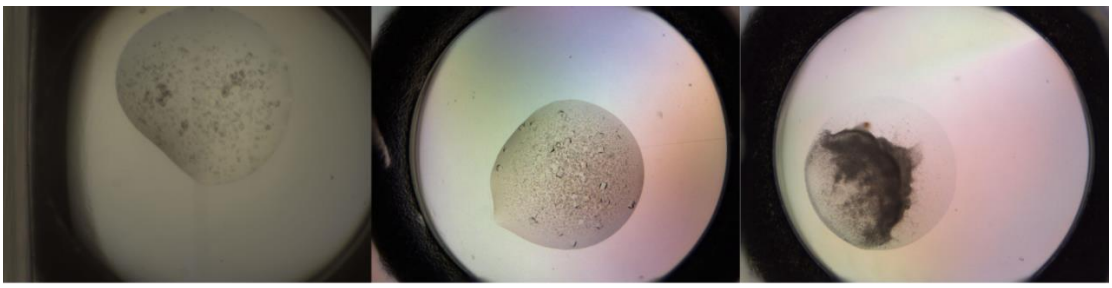



*Figure 3.1: Comparison of images captured at (a) SGC, Oxford, with the Minstrel HT and (b, c) Novartis, Basel, with the Rock Imager. SGC droplets are captured at one focal depth, resulting in out-of-focus regions, whereas Novartis images are a combination of 7 focal depths, and*

*hence sharper across the droplet. Plate material, seals and optics used at Novartis also produces colour gradients in the images. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 3.1.5 Human evaluation of the crystallization outcome with the ranking system

To show that such ranked viewing of images result in drops with microcrystals being more carefully evaluated, 10 plates containing (micro)crystals were randomly selected from the SGC, and divided into two sets (A and B). Each set consisted of three sparse-matrix screens and two optimization screens of different protein targets. Two groups of five crystallographers from the SGC with varying experience were given 2 minutes to evaluate each plate, which was generally insufficient to view all 288 droplets in a plate, to standardize and constrain the time taken by each crystallographer. The first group viewed set A in a ranked order, set B in an un-ranked order, while the second group viewed set B in a ranked order and set A in an unranked order; in other words, both groups performed both ranked and unranked viewing. The majority vote (microcrystals or crystals) of each group of crystallographers is used to compare against the other group for missed (micro)crystal annotations to reduce human error.

## 3.2 Results and discussion

### 3.2.1 Ranking *versus* classification or filtering

Although previous studies set out to classify crystallization drops into discrete human-assigned categories, it is not clear that they demonstrated this is achievable, or indeed that the underlying premise of one-drop-one-score is even useful: individual drops routinely exhibit multiple precipitation behaviours that may inform one another but which are nevertheless only very loosely defined by the community (Newman et al., 2012). Unsurprisingly, even human

classification of droplets yields poor agreement rates (Buchala & Wilson, 2008), and using such variable opinions as ground truths severely undermines the training of learning algorithms and not only reduces accuracy, especially for multi-class classifiers, but makes it unmeasurable. The increased rate of false negatives is particularly pernicious since the formation of crystals is in general a rare event yet experimentally crucial to detect.

We have targeted a more realistic goal, namely to prioritise droplets for viewing: we judge this to be more useful because it does not pre-empt decisions, but helps them to be made more accurately as well as more rapidly. One version of this is filtering, the strategy chosen by Mele *et al.* (2014), who filter out images from further examination based on the lack of change (differences) in a droplet over time, on the basis that changes over time may indicate the formation, growth, or disappearance of crystals. Vendors also employ such filtering, e.g. in Rigaku Automation's viewing software for images captured on a UV-enabled instrument. However, filtering is merely an extended case of classification, where instead of a single classification cut-off, a tuneable cut-off or criterion is still required to hide a subset of data.

In contrast, ranking circumvents the problem of selecting filtering cut-offs or criteria, which tend to be arbitrary; instead it rearranges the data in a more meaningful way. In our case, we rank droplets on a continuous scale (0 to 1) of their probability of being *'interesting'*: this means crystals or microcrystals are likely to be viewed first, while the '*uninteresting'* droplets (precipitates, clear drops) are viewed later. In the real-world environment where time is limited and attention level decreases with time, the ranking system focuses resources on what is most likely to matter, namely the likely presence of crystallinity.

The ranking score is obtained by repurposing a two-class classifier trained on images assigned manually into only two categories (interesting or uninteresting). When applied to new images, the classifier generates a score (0 to 1) which ordinarily would be compared against a threshold

to assign images to either category; instead, here the score is directly employed for sorting a given set of images, typically from a crystallization plate. Multi-class classifiers can in principle also be used to rank images, most trivially by assigning each class to a rank (e.g. 'crystal' over 'microcrystal' over 'spherulite' and so on); however, this ranking is algorithmically arbitrary (for example, should 'clear' drops be viewed before 'precipitate' drops?), and moreover it is non-obvious how to rank images within classes, since classifier scores are no longer uni-directional. On the other hand, Buchala and Wilson (2008) have shown that reducing the number of classes increases the human agreement rate and hence accuracy of the classifier, suggesting that the fewest possible class (two) would support the highest reliability.

### 3.2.2 Algorithm performance

Our algorithm appears to rank images effectively, if judged by criteria that are reasonable in routine laboratory usage: for plates where wells were viewed with our new ranking order, the first well was a crystal-containing image for 128 out of a total of 196 plates (65.31%). The number of such "successful" plates increases if the criterion is relaxed to expect at least one crystal in the top 10 or 32 ranked wells (Table 3.2), also illustrated by the curve in Figure 3.2(a) which shows that for most plates, the first human-annotated crystal is very high in the ranking order. We show both the results for viewing by well and sub-well in Table 3.2: while viewing by sub-well (single droplets) is the common practice, when all drops in a well have the same chemical composition and differ only in mixing ratio, there is added value in viewing all the well's sub-wells side-by-side, since precipitation trends can be directly observed.

It is not only the top drop, but all drops in a plate that seem to be effectively ranked: Figure 3.2(b) and Figure S3 (Section S.2 in Supplementary Materials) illustrate how the top of the viewing order is generally enriched with images with crystals and microcrystals. Moreover,

where the ranking failed to move the human-annotated crystal to the top, it was usually questionable whether those images had been correctly marked (Figure 3.2(c)), or else the images themselves were problematic: they were either out of focus (Figure 3.3(a)), questionable crystals (Figure 3.3(b and c)), crystal embedded in precipitation (Figure 3.3(d)), or inaccurate segmentation of the droplet (Figure 3.3(e and f)). As with all learning approaches, the performance is expected to improve as more images are added to the training set; but we conclude that the method does push at least one crystal to the top with high likelihood.

*Table 3.2: A comparison of performance of the algorithm before and after review of image annotations. The top and bottom half of the table shows the number and percentage of plates where the first marked crystal was ranked in the plate, in terms of ranking by well and ranking by droplet respectively. Table taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

| Rank (well) | No. of plates | % |
|---|---|---|
| **1** | 128 | 65.31 |
| **<10** | 185 | 94.39 |
| **<32** | 192 | 97.96 |
| **Rank (droplet)** | | |
| **1** | 118 | 60.20 |
| **<29 (10%)** | 189 | 96.43 |
| **<72 (25%)** | 191 | 97.45 |
| **Total No. Plates** | 196 | |

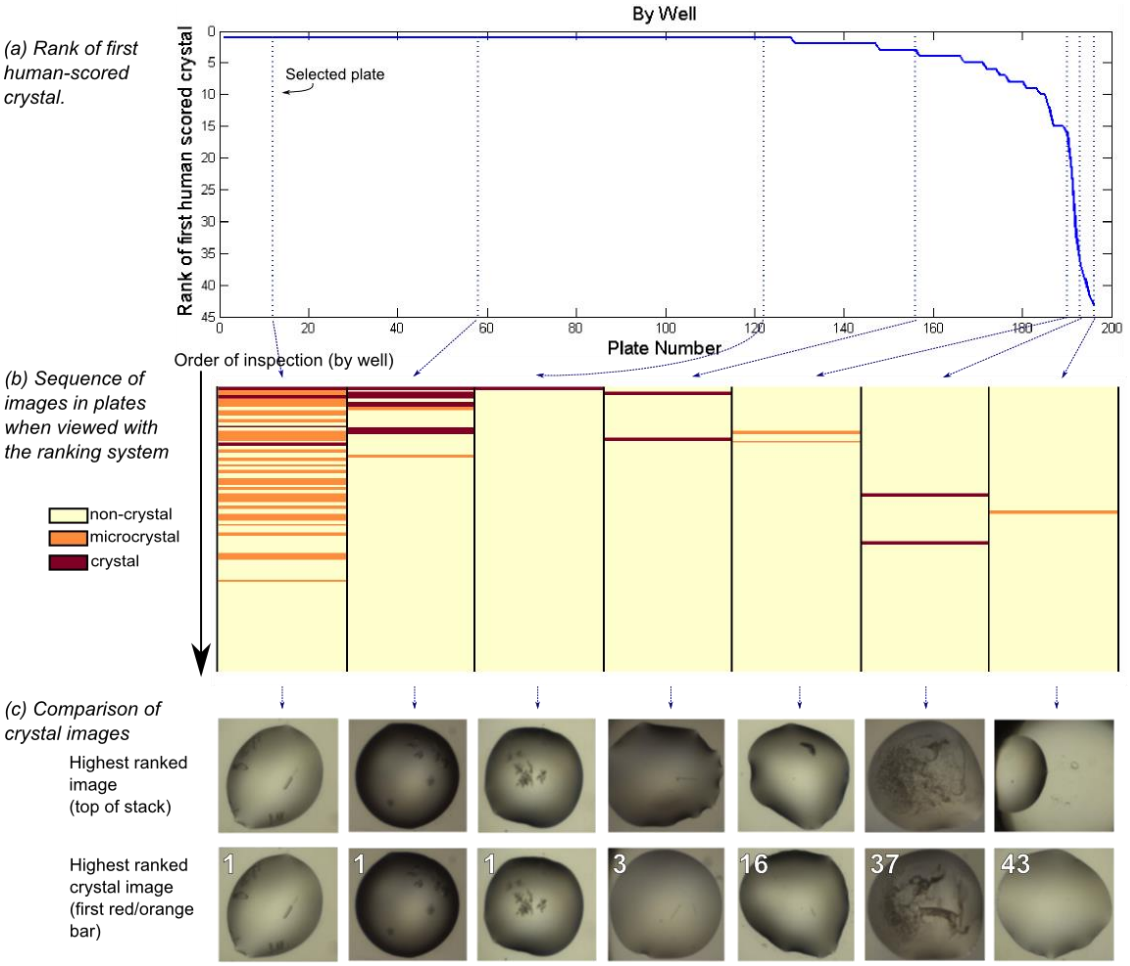



*Figure 3.2: Effectiveness of the algorithm at ranking crystal containing images. (a) Rank of first human-scored crystal image in the plate. The ideal curve is a horizontal line with y-axis = 1. In (b), each column represents a plate, and shows the position of mountable crystals (red), microcrystals (orange), and uninteresting droplets (yellow) as scored by a crystallographer when viewing the plate with ranking algorithm. It is clearly beneficial to inspect by the ranking algorithm since Interesting images are generally moved to the top of the viewing order. In cases where the algorithm did not perform well, the marked crystals were usually questionable, as shown in (c). More examples are shown in Figure 3.3. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

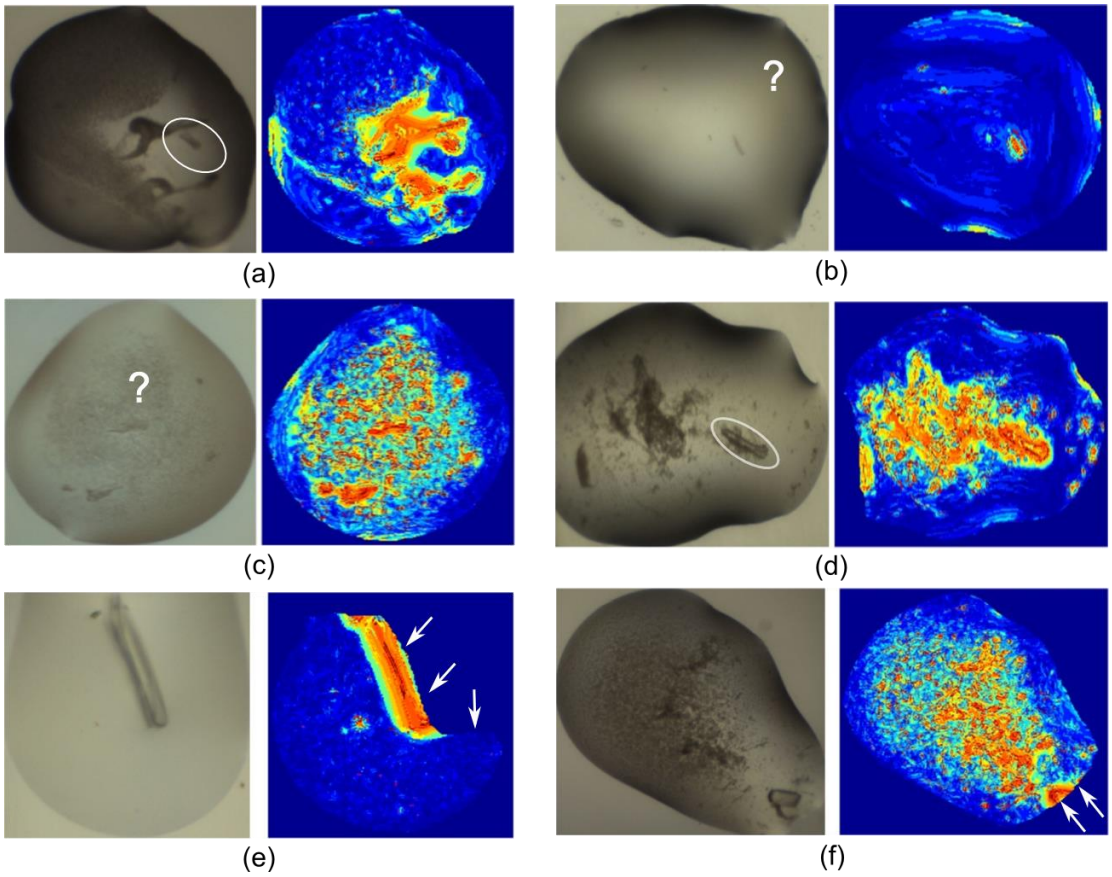



*Figure 3.3: Example of images and the corresponding texton map of annotated crystals and microcrystals that received low ranks. Such crystal images were usually out of focus (a), questionable or generously labelled (b and c), surrounded by precipitates (d), or incomplete droplet segmentation (e and f). Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 3.2.3 Performance for different imaging systems

The texton approach is robust enough to work well on different imaging systems, without additional training. Using the learning parameters obtained with images acquired by the Minstrel HT system at the SGC Oxford, the method was transferred to Novartis, Basel, which has a Rock Imager. The transfer was generally straight-forward, apart from the modifications in droplet segmentation as mentioned previously and a different scale factor (0.5 instead of 0.25 at the SGC) to approximately match the resolution at the filtering stage. Table 3.3 and Figure 3.4 show the performance of direct application of the system, as well as the same system with

additional training images (15 *interesting* and 150 *uninteresting* randomly selected images), compared to the SGC Oxford datasets; and as expected the results improved with more training images. Note that a direct comparison of the SGC and Novartis numbers is not meaningful for this number of plates, in addition to the ratios of sparse-matrix screens and optimization screens being different.

This result indicates that the texton approach does not need a very high level of detail in the images: SGC and Novartis images have similar field of view (the entire sub-well of a typical SBS format crystallization plate, Figure 3.1), but different pixel resolutions, yet the final scaling is at lower resolution than both for the texton analysis. The more important factor presumably is that the depth-of-field encompasses the majority of the precipitation behaviour, based on Figure 3.3(e, f), where this is necessary for good droplet segmentation, which leads to accurate calculation of texton distribution.

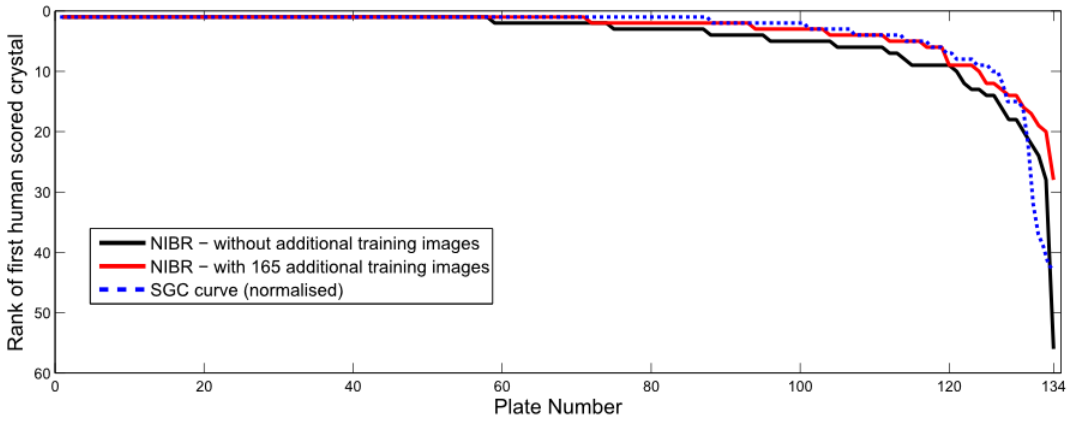



*Figure 3.4: Rank of first human-scored crystal image in the plate, for the NIBR dataset of 134 plates. The black curve shows the performance of direct application of the ranking system without additional training images. By adding 165 (15 crystal and 150 non-crystal) images acquired with the Rock-Imager, more crystal images were ranked higher, as shown by the red curve. The SGC performance curve (blue) is normalised to equal length for qualitative comparison. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

*Table 3.3: Percentage of plates according to the rank of the first human-scored crystal for the SGC and Novartis. Table taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

| Rank | % Plates (SGC) | % Plates (Novartis, without additional training images) | % Plates (Novartis, with 165 additional training images) |
|---|---|---|---|
| 1 | 65.31 | 43.28 | 52.99 |
| <10 | 94.39 | 90.30 | 92.54 |
| <32 | 97.96 | 99.25 | 100.00 |

### 3.2.4 Effectiveness of drop ranking for human scoring

Figure 3.5 shows the annotations of the two groups of crystallographers for the 10 selected plates, and the labels for individual crystallographers are in Figure S4 (Section S.2 in Supplementary Materials). As expected, more (micro)crystals are annotated within the short allocated time frame when images were viewed in the ranked order, supporting our view that a ranked view is more efficient in identifying successes. Furthermore, where crystallinity was rare (plates 9 and 10), unranked viewing resulted in (micro)crystals being missed, even those located early in the viewing order, in contrast to the ranked viewing: evidently images are considered more carefully if they are known to be likely to be interesting. At the same time, ranked viewing does not appear to introduce more generous scoring at the top of the rank through confirmation bias, because microcrystals were also missed in ranked viewing (Figure 3.5b, blue rows in plates 1, 4, and 8, black arrows); on the contrary, annotation appears to become more stringent if larger (micro)crystals have already been observed.

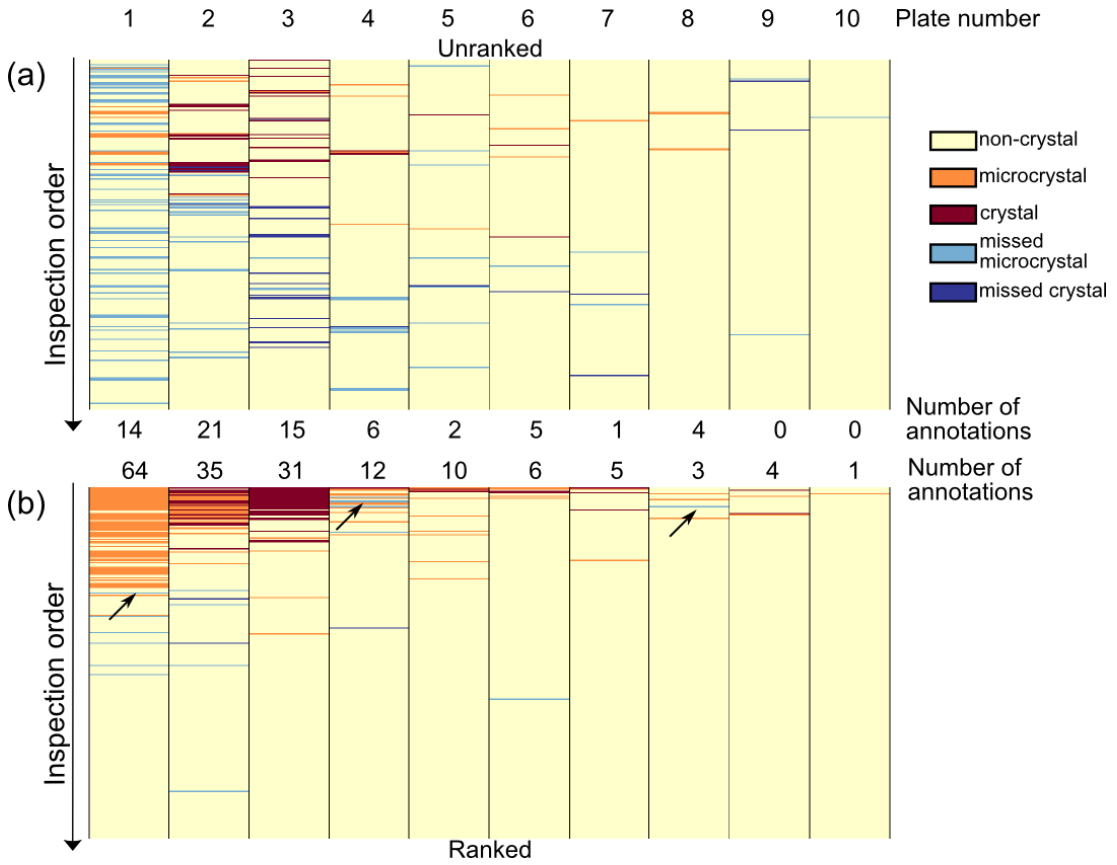



*Figure 3.5: Comparison of annotations between 2 groups of crystallographers. Each column represents a plate of 288 droplets inspected in the unranked order (a) and ranked order (b), and the rows are coloured according to the majority vote for these images: yellow – non crystals, orange – microcrystals, red – mountable crystals. Blue rows indicate missed (micro)crystal, where the corresponding images were annotated in one viewing order but not the other. The total numbers of annotations (orange and red rows) are shown in between. The black arrows show the location of missed (micro)crystals in ranked order. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

### 3.2.5 Scores as a profile of the plate

The collective scores of droplets across a plate form a profile of the plate and can be used to make quick judgements. Since each droplet score is the posterior probability of the droplet being "*interesting*", a plate with many wells containing crystal will have a different profile from

one with none. Figure 3.6 shows the different profiles of a plate with 244 crystals (green) and one with no crystals (brown). The profile can also be used to determine a cut-off for images to inspect. Across the 196 plates analysed, over 88% of plates would have at least 1 crystal found by just inspecting images with scores greater than 0.5. This also corresponds to not inspecting an average of 90% of '*uninteresting*' droplets in the plate. Table 3.4 shows the trade-off of percentage of plates with at least one crystal found and the amount of uninteresting droplets to view, which can be used as a confidence indicator for the various cut-off values.

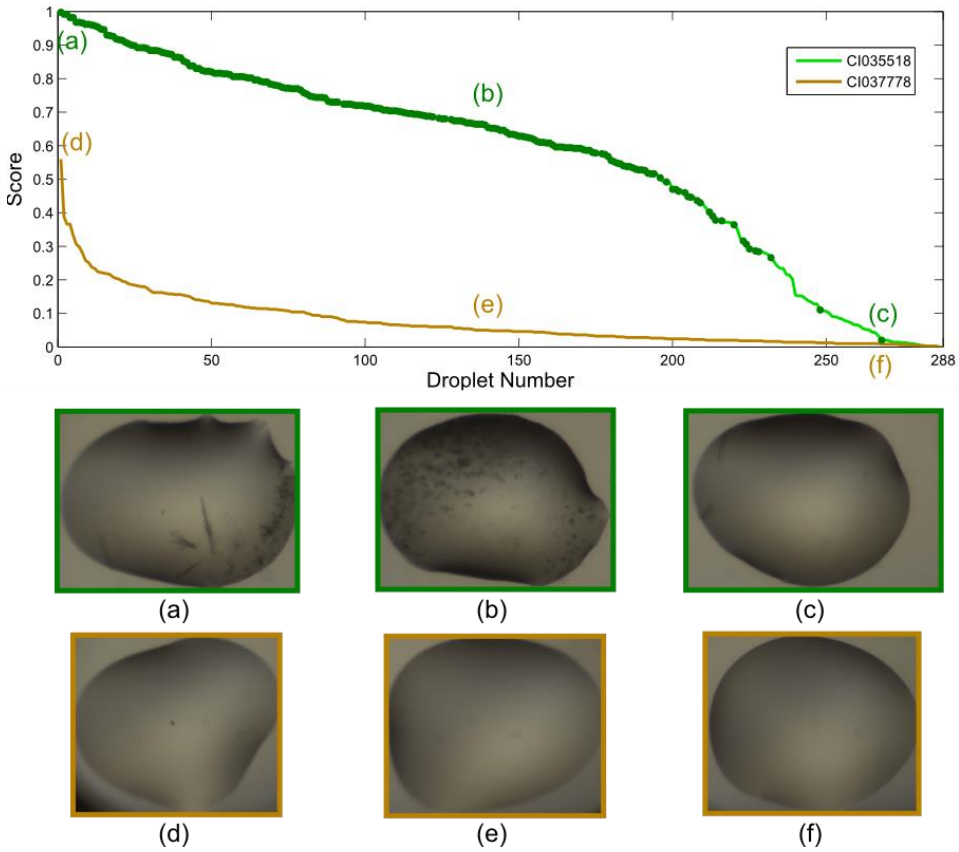



*Figure 3.6: Comparison of scores of droplets in 2 plates. The scores are the posterior of a droplet being "interesting", and the profile of the curve gives an overview of a plate. Plate CI035518 was an optimization screen, where 213 droplets contained crystals or microcrystals, marked with dark green dots on the curve. It thus received generally higher scores. In contrast, CI037778 had no crystals and were mainly clear drops. Scores for this plate tailed off rapidly. The highest scored image here was a droplet with dust speckles. The embedded images are the corresponding droplets that that were ranked 1, 100 and 268. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

*Table 3.4: Percentages of plates with at least 1 crystal found for different cut-off of scores for viewing, and the corresponding average of uninteresting droplets not inspected. Figure taken from Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

| Cut-off | % Plates with ≥ 1 crystal found | mean % uninteresting droplets unseen in a plate |
|---|---|---|
| **0.8** | 60.20 | 97.88 ± 6.66 |
| **0.5** | 88.78 | 90.12 ± 12.17 |
| **0.2** | 98.98 | 58.87 ± 23.59 |
| **0.1** | 100.00 | 35.92 ± 22.78 |
| **0.05** | 100.00 | 20.24 ± 17.15 |
| **0.01** | 100.00 | 3.61 ± 6.03 |

## 3.3 Concluding remarks

The ability to rank droplets by their probability of containing crystals or microcrystals, using the descriptors derived with the texton method served as further validation of our image processing algorithm. The transferability of the classifier between two widely used commercial imaging systems for sitting-drop vapour diffusion experiments also attests to the robustness and generalization of our texton dictionary and trained classifier. The practical application of a ranking system have indeed shown that enriching the number of crystal-containing images to inspect first does result in closer attention apparently given to interesting droplets, increasing the efficiency and accuracy of visual identification of crystals. The collection of scores across a plate also forms a useful profile for a quick overview of the suitability of the protein for structural studies.

Improvement in ranking quality would be the main focus for further work. As with most classifiers, the addition of training examples will increase the accuracy of the classifier and thus ranking scores. Presently, the texton distribution does not take into account the spatial relationship between textons: crystal-related textons clustered together spatially (from a large

crystal, for example) receive equal representation as those scattered throughout the droplet from noise, dirt, or smaller crystals. We thus expect features that incorporate such spatial relationship of textons to increase ranking quality, as would information from time-change differences and other imaging modality. A weighted-sum model can be trained to incorporate all of these information for a better ranking system, although we expect the law of diminishing returns to be apparent, and thus may not be worth the additional computation time and input data.

While our ranking approach aids in the ultimate goal of identifying crystals, the majority of experimental outcomes are non-crystalline droplets. Such 'failed' experiments form the subject of interest in the next chapters, where the aim is to extract information to provide guidance for subsequent crystallization effort for a protein target from initial trials, without relying on the presence of crystalline behaviour.

*Special thanks to Alexandra MacKenzie, Chitra Shintre, Fiona Sorrell, Jit Ang, Joseph Newman, Marbella Fonseca, Mariana Grieben, Radoslaw Nowak, Yin Dong, Jolanta Kopec, Carline Sanvitale, Aleks Szykowska, and Hazel Aitkenhead for volunteering for the experiment in Section 3.1.5, and all crystallographers at the SGC and NIBR for their scoring and labelling of images over the years, generating the datasets used in this chapter.*

## 3.4 References

Breiman, L. (2001). *Mach. Learn.* **45**, 5–32.

Buchala, S. & Wilson, J. C. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **D64**, 823–833.

Liu, R., Freund, Y., & Spraggon, G. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1187–1195.

McPherson, A. & Gavira, J. a (2014). *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **70**, 2–20.

Mele, K., Lekamge, B. M. T., Fazio, V. J., & Newman, J. (2014). *Cryst. Growth Des.* **14**, 261–269.

Newman, J., Bolton, E. E., Müller-Dieckmann, J., Fazio, V. J., Gallagher, D. T., Lovell, D., Luft, J. R., Peat, T. S., Ratcliffe, D., Sayle, R. a, et al. (2012). *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **68**, 253–258.

Ng, J. T., Dekker, C., Kroemer, M., Osborne, M., & von Delft, F. (2014). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 2702–2718.

# 4. Collective Precipitation Pattern as a Proxy for Optimization Strategy

Efforts in practical crystallization thus far have mainly focused on how to sample more with less protein sample (robotics, automation, commercial screens), and how to best identify crystals. While these have led to state of the art methodologies and high-throughput facilities, the number of 'failed' experiments has not decreased; on the contrary, the ease of setting up experiments and the lower cost may have even encouraged more brute-force experiments.

The typical evaluation of crystallization trials from any crystallization experiment involves the inspection of each trial in the screen to identify crystals or crystalline behaviour. The previous chapter addressed this aspect by ranking droplets according to their likelihood of crystallinity. However, this process is fundamentally treating crystallization trials as singular binary experiments (crystal/no crystal), independent of all other droplets in the screen even though they contain the same protein sample. While effective for identifying hits, such treatment of a crystallization trials is limited in the absence of crystallinity or precipitates judged by some experienced experimenters to be worth following up on. When no crystal or crystalline behaviour can be identified, no readout is obtained, and the experimenter is no more informed that before the experiment was carried out.

We therefore set out to develop methods to move the analysis of screening experiments beyond a crystal-spotting exercise, and instead extract a reliable readout even in the absence of crystallinity. While the presence of crystallinity is the most robust readout of a protein's suitability for structural studies with X-ray crystallography, these are often rare and unreliable; Newman (2005) found that only half of the expected crystals were observed in a four-fold redundancy setup. In contrast, precipitation is very common in screening experiments, and

here, we propose that it is also consistent: in other words, the same precipitate (at least judging by visual appearance) will form reliably for equivalent experiments.

While it is difficult to extract conclusions from a single observation of precipitate, our hypothesis is that, thanks to precipitation being consistent, the collective precipitation behaviour of a protein across a standard sparse-matrix screen – where and how the experiment precipitates – should also be a reliable descriptor of the protein. All droplets in a screening experiment should thus be treated as one experiment, and this collection of precipitation patterns, which characterizes the experiment, is known as its *precipitation fingerprint*.

One way to use such fingerprints is through a knowledge-based approach of comparing it to a collection of historic fingerprints, and thence infer potential crystallizing conditions, by proxy of its precipitation fingerprint similarity to those that also produced crystals. This also implies a low likelihood of success if no experiments with similar fingerprints have been crystallized in the past, thus providing the framework for assessing crystallizability of the experiment. Moreover, the principle should be extendable to already crystallizing systems, to identify alternative crystallization conditions for increased crystal reproducibility, or to generate different crystal forms suitable for specific purposes.

Comparing experiments using precipitation fingerprints is advantageous over sequence based methods: (1) surface mutations dramatically alter crystallization, which is not robustly captured in sequence similarity, and (2) the precipitation fingerprint encapsulates all protein sample treatment prior to crystallization, as well as the inter- and intra-molecular interactions during the screening experiment, which makes it much closer to experimental reality.

More than 11 years of high-throughput operation on medically relevant human protein targets have allowed the SGC to accumulate a vast dataset of diverse, well-curated and reliably annotated crystallization experiments, which can be tapped to form libraries of such *fingerprints*. This chapter presents first the development of the *precipitation fingerprint libraries (PFL)*, to which new screening experiments can be compared to, followed by the follow-up experiment design based on conditions suggested from such comparisons. We show how our strategy fares on human protein samples, producing new crystal forms and better diffraction resolution.

## 4.1 Methods

Our analysis is limited to the following screen-types, chosen for the high number of datasets available: JCSG, LFS, HCS and HIN (see Table 2.1). Given a new plate, the workflow of our method is summarized in Figure 4.1. This entails first extracting the texton distribution for all images of a plate using the image processing pipeline described in Chapter 2.3. The normalised texton distribution (discrete histogram) from each image in a plate occupies a row in a 288-by-300 matrix, arranged by column, row and sub-well (*i.e.* A1a, A1b, A1c, B1a, B1b, B1c, … H12a, H12b, H12c, where sub-well a, b, and c correspond to 2:1, 1:1 and 1:2 protein to precipitant mixing ratio). This matrix forms our mathematical definition of precipitation fingerprint for a plate.

Plate-to-plate distances to entries in the precipitation fingerprint library of the same screen-type are calculated using the methods outlined in Chapter 2.4.2. These distances are used to identify nearest-plates in the library to the new plate, and the ID's of these nearest-plates are written to a database. The Nearest-plate Curve and Crystallization Conditions Profile are generated on-the-fly, depending on the number of nearest-plates to include. The following sections outlines the development of the precipitation fingerprint library and how this is used in practice to design follow-up experiments.

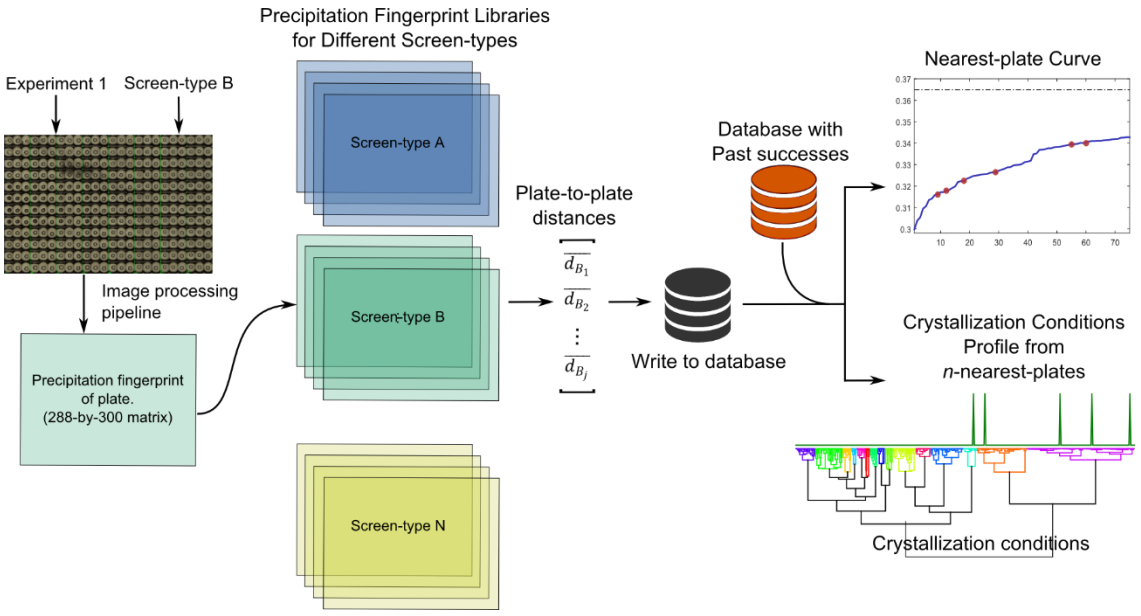



*Figure 4.1: Workflow for identifying potential optimization conditions for a new screening experiment. The precipitation fingerprint of a new plate is generated using the image processing pipeline (Chapter 2.3), and compared to entries in the Precipitation Fingerprint Library of the same screen-type. Plate-to-plate distances from the new plate to all plates in the library are calculated and written to the database. The Nearest-plate Curve and Crystallization Conditions Profile are generated, and potential optimization conditions may be inferred from the successes of nearest-plates to the new plate, which should have similar precipitation behaviour.*

### 4.1.1 Compiling the Precipitation Fingerprint Library (PFL)

The premise of our approach is to compare new screening experiments to historical data captured in the Precipitation Fingerprint Libraries (PFLs). Each screen-type (JCSG, LFS, HCS and HIN) at each incubation temperature forms its own PFL, where each entry of the dictionary is the precipitation fingerprint from a standard screening experiment (three 150nl droplets with 2:1, 1:1 and 1:2 mixing ratio of protein to precipitant per condition) set up at the SGC from January 2005 to May 2014. There are two parts to the library: (1) the collection of precipitation fingerprints for a given screen-type at an incubation temperature, and (2) the corresponding conditions(s), if any, that resulted in diffracting protein crystals, irrespective of screen-type.

#### *4.1.1.1 Extraction of precipitation fingerprints from historical data for the library*

The first part of the library consists of fingerprints in the form of 288-by-300 matrices. These fingerprints were derived from images of plates captured at t = +4 days, selected to balance waiting time and equilibration of droplets. Eight PFLs corresponding to each supported screen type and incubation temperature were compiled, as detailed in Table 4.1.

There were two considerations in using the historic data generated at the SGC: (1) there was a change in plate type in early 2009, from the CrystalQuick™ 96 Well Greiner plates (flat bottom) to SwissCi 3-Lens Crystallization Microplate (concaved bottom), which may generate different precipitation behaviour for the same experiment; (2) images older than a year that were available on the network drive were compressed versions (downsized from 2560-by-1920 to 800-by-600 pixels); original high-resolution images were archived in magnetic tapes, but retrieval was non-trivial and limited to authorized IT personnel.

To investigate if different plate type significantly affects precipitation behaviour, and if our feature extraction accommodates for the different droplet morphologies, a control experiment comparing the distances between corresponding droplets of two proteins (JMJD2AA-p085 and PHIPA-p022) in JCSG and HCS, set up with SwissCi and Greiner plates was carried out. The distribution of distances in these repeat experiments, where the only difference was the plate type, is shown as the blue curve in Figure 4.2(a). In comparison with repeat experiments set up in SwissCi plates (five plates of PHIPA-p022 in JCSG, red curve), there is a shift in the distribution of distances, presumably caused by the different plate types. Figure 4.2(b) shows examples of corresponding droplets where the distance was >0.4, with visibly different precipitation behaviour. These were however, few compared to others with visually similar behaviour and hence lower distance. Furthermore, the slight shift in distance from plate type difference was judged to be acceptable when compared to the distribution of droplet-to-droplet distances of

two randomly selected proteins (black curve in Figure 4.2(a), SwissCi plates), especially since the added information content from proteins in Greiner plate experiments was invaluable.

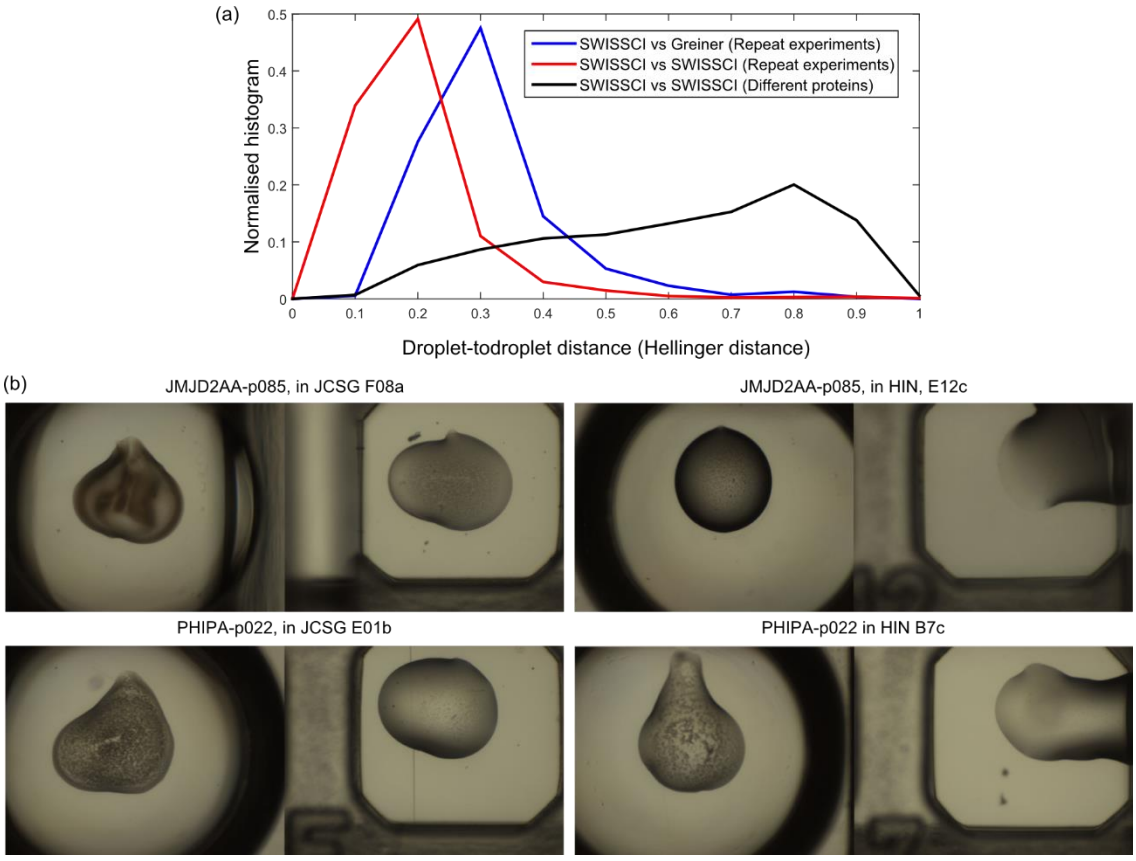



*Figure 4.2: Comparing precipitation in SwissCi and Greiner plates. (a) shows the distribution of droplet-to-droplet distances in repeat experiments with different plate types (blue) or similar plate types (red). Distances are generally larger when using different plate types, although this resulting noise should not be significant when comparing to non-repeat experiments (black curve). (b) Examples of difference in precipitation pattern for repeat experiments due to plate type difference and droplet morphology.*

To investigate if the downsized images were sufficient for our analysis, we determined if (1) there were significant differences in the histograms derived from downsized images and high-resolution images, and (2) if the distances of two droplets calculated using downsized to high-resolution and high-resolution to high-resolution were similar. We used the downsized and original images of 2 randomly selected plates (CI045940 and CI045810).

For (1), Figure 4.3(a) shows that the distance between histograms derived from downsized and its corresponding high-resolution image were low, with a maximum of 0.16, and 85% below 0.10. Thus, unsurprisingly, for (2), the distances calculated when using downsized images against high-resolution image were comparable to using high-resolution images for both plates (Figure 4.3(b)). Together, these show that the same readout is obtained with downsized images, and that the texton method was sufficiently robust for compressed images. The contrast of distances in Figure 4.3(a) and Figure 4.3(b) also shows that there is a high signal-to-noise ratio. Thus, we concluded that the downsized images were sufficient for the analysis, and used them where high resolution images were no longer available on disk.

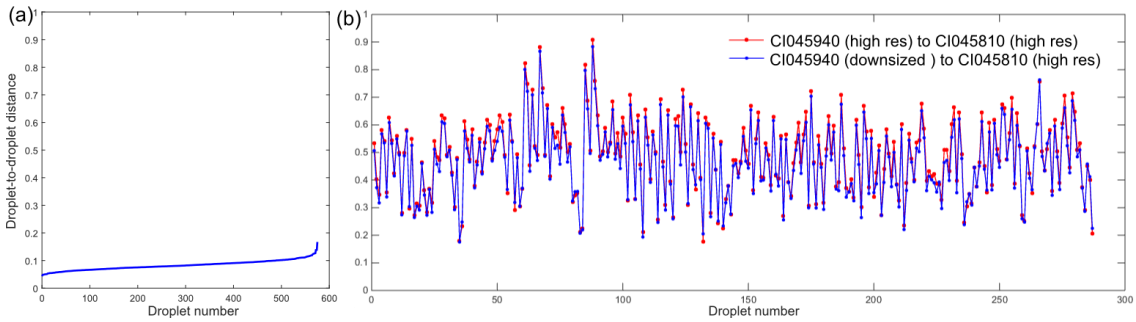



*Figure 4.3: Comparing features derived from full-sized and compressed images. (a) Sorted distances between texton distributions generated from high resolution images and the corresponding compressed (downsized) images. (b) Droplet-to-droplet distances for two example plates, calculated using high resolution images for both plates (red) and low resolution vs high resolution images (blue). The correlation of distances was 0.995, indicating little to no noise from using compressed images.*

Using the computing power of 25 Windows machine (with a mixture of Intel Core i7 and i5 processors, all with at least 8GB of RAM) over two weekends, a total of 24967 plates were analysed, although only 15699 plates were included in our libraries. We removed duplicate plates (those with identical purification ID, crystallization concentration, sample state, co-crystallization compounds, and incubation temperature) to avoid over-representation of any

experiment in the library, and only kept plates with < 20% of faulty droplets and < 80% of clear drops (detailed description of automatic clear drop identification in Chapter 5) to ensure high information content: faulty droplets are essentially non-experiments, and a generally clear plate will be similarly observed for all under-concentrated protein samples, thus giving low discriminatory power.

#### *4.1.1.2 Identification of successful conditions for experiments in the PFL.*

The second part of the library contains all conditions that successfully induced diffracting protein crystals for the experiments associated with each fingerprint in the PFL. Diffracting protein crystals were identified regardless of its diffraction quality, incubation temperature, and screen-type, which included subsequent optimization screens. We chose to use diffracting crystal as a measure of success instead of image annotation to avoid the inaccuracies in image labelling, potential salt crystals, and increase certainty of success.

Figure 4.4 shows an example of how these conditions are identified for an experiment. In this case, six screens were set up at two temperatures, including a custom screen (knowledge-based, either from publications or past experience). The successful conditions associated with this experiment were conditions A4 and B12 of LFS, C5 of HIN from the optimized screen, and the condition among the four supported screen-types most similar to condition F9 in the custom screen, if and only if the minimum distance as determined by Cockatoo (Bruno et al., 2014) was below 0.01. Determination of success was hence not screen-type dependent, but spanned across the chemical space sampled by the four supported screen-types at two temperatures; the only limitation was if a screen was not set up in the past. In the example of Figure 4.4, there is no data on the outcome of conditions sampled in HCS.

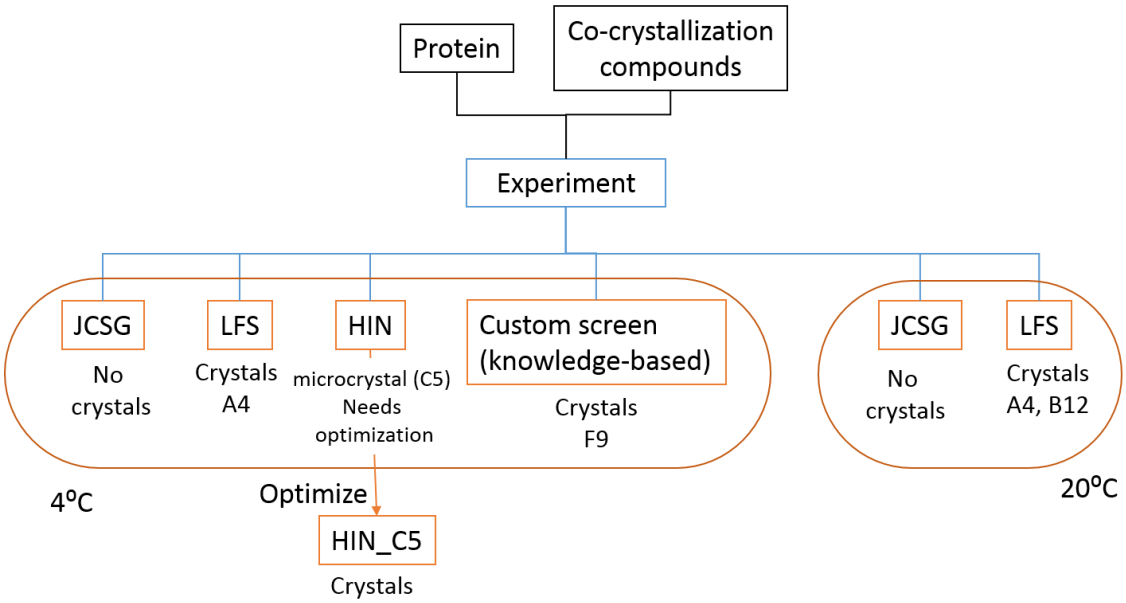



*Figure 4.4: Example of successful conditions for a hypothetical experiment. At the SGC, the database records allows for the tracking of all conditions that resulted in diffracting crystal for any given experiment, including those derived from optimized screen (C5 of HIN in this example). Conditions from custom screens can be compared objectively to other conditions using Cockatoo (Bruno et al., 2014) because they follow fixed naming convention. "Crystals" in this figure is taken to mean diffracting protein crystals.*

Such identification of successful conditions is possible at the SGC because all crystals are tracked along each step of the structure determination process. Names of optimization screens designed from a hit in sparse-matrix screens follow standard naming conventions, allowing its direct mapping to the origin condition. Furthermore, conditions in non-standard (custom) screens follow standard naming conventions, allowing them to be compared to other supported conditions using the metric defined by Bruno et al. (2014) to identify its closest match.

In practice, every fingerprint in the PFL has an associated 384-element condition vector, where each element corresponds to a condition, and it is set to TRUE if the condition was directly or indirectly responsible for a diffracting crystal, and FALSE otherwise. The joint 384 conditions from these screen-types can be further reduced to only 340 unique conditions, with 27 shared

conditions between JCSG and HIN, and 11 between JCSG and HCS, a further 3 between HIN and HCS, and 1 between JCSG, HIN and HCS. For simplicity, we combined the successes from such identical conditions, but treated non-identical conditions separately, even if they were very similar (for example, 3M sodium chloride, 0.1M bis-tris pH 5.5 (JCSG, G12) and 3M sodium chloride, 0.1M bis-tris pH 6.5 (HIN, A10) were treated separately). Table 4.1 shows the number of experiments in the PFL with at least one associated successful condition.

The data from NIBR could not be included in this analysis because there is no direct linking of diffracting crystals to the crystallizing condition. Furthermore, information of a crystal is obfuscated if sent for data collection, making manual mapping of crystals to crystallizing conditions impractical.

*Table 4.1: Precipitation Fingerprint Libraries compiled from data at the SGC, Oxford. 2 libraries were compiled for each screen type based on incubation temperatures (4 and 20°C).*

| **Screen** | **JCSG** | | **LFS** | | **HCS** | | **HIN** | |
|---|---|---|---|---|---|---|---|---|
| Inc. temperature | 4 | 20 | 4 | 20 | 4 | 20 | 4 | 20 |
| Num. plates | 2410 | 2793 | 1281 | 2274 | 1292 | 1782 | 1448 | 2419 |
| Num. with successful condition(s) | 273 | 396 | 170 | 353 | 178 | 282 | 193 | 396 |

### 4.1.2 Using the PFL to identify optimization conditions from nearest-plates

The fingerprint of a new screening plate is compared to all plates in the corresponding PFL to obtain plate-to-plate distances (Chapter 2.4.2). Starting with the nearest plate, a list of potential conditions and its frequency of success can be generated by appending successful conditions associated with the *n*-th nearest-plates to the list. We propose that this list holds conditions that may induce crystallization of the new experiment, by proxy of its precipitation fingerprint similarity to those where these conditions have also induced crystallization; because they share similar precipitation behaviour, they may also share the same crystallizing condition.

Since the crystallization conditions in the supported screens are random (from sparse-matrix screens), we rearranged them by similarity for a more meaningful presentation. A dendrogram of the 384 conditions was generated using Ward's method with distances between conditions given by Cockatoo (Bruno et al., 2014). A condition profile can thus be formed by mapping the list of potential conditions onto an x-axis defined by this dendrogram, and a y-axis that captures the frequency of success for the condition. As the number of nearest-plates to include increases, the profile will also increase in its overall count of successful conditions, if the new plate to include had associated successes. The profile can be used to further identify conditions in the vicinity of those in the list of potential conditions for follow-ups; entire clusters of conditions with high density of marked potentials may be worth pursuing. Figure 4.5 shows an example of generating such profiles when more nearest-plates are included.

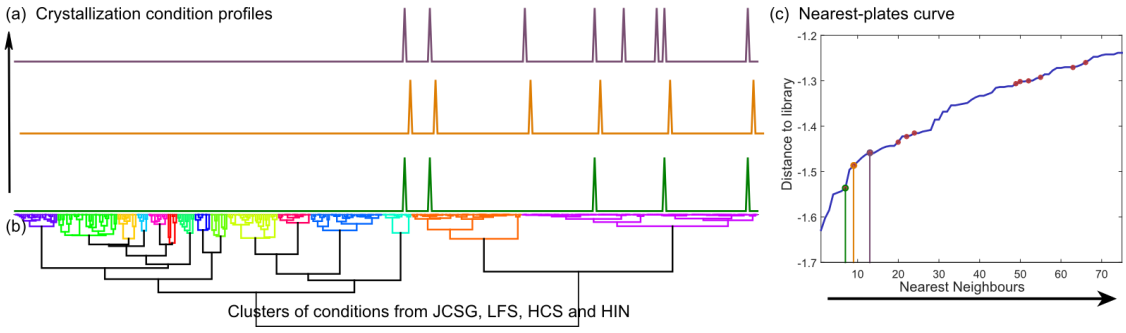



*Figure 4.5: Generating condition profiles with nearest neighbours of a screen. The dendrogram (b) shows clusters from conditions in JCSG, LFS, HCS and HIN. Each node represents a condition and the arrangement forms the x-axis of the condition profiles in (a). The sorted normalised-distances of a given plate to the nearest-plates in the precipitation fingerprint library (PFL) are shown in (c). Plates of experiments with successful conditions are marked with red dots on the curve. Crystallization condition profiles are formed by incrementally incorporating the successful conditions corresponding to these experiments. In this example, as the number of nearest neighbours to include increases (green to orange to purple in (a), with corresponding profiles in (c)), more conditions are found that may be used in an optimization screen.*

## 4.2 Controls and validation of methods

The following sections discuss control experiments and the testing of our hypothesis, followed a description of the experimental method as proof of concept.

### 4.2.1 Protein coverage in the library

As a representative example, Figure 4.6 shows the range of proteins represented in the PFL for JCSG at 4$^{o}$C. The biological diversity of the PFL is limited by the targets worked on at the SGC. Kinases and Oxidoreductases form the majority of entries in our library, although there is further diversity due to the different constructs, domains and co-crystallization compounds not captured on a target-level representation alone. The expected mass of the proteins of this PFL ranges from 6 kDA to over 150 kDa with a mean of just under 40 kDa.

### 4.2.2 Baseline plate-to-plate distance and internal pairwise distances of the PFLs.

To establish the baseline noise of plate-to-plate comparison arising from experimental setup and image analysis artefacts, plate-to-plate distances were calculated for the five identical plates with the same experiment used previously in Section 4.1.1 (PHIPA-p022, no co-crystallization compounds in JCSG). Ideally, any two of these five plates should have low pairwise distances since the precipitation behaviour should be similar. However, the different resulting droplet morphology and final mixture volumes due to robotic errors will lead to different equilibration kinetics. The 10 resulting pairwise distances from these five plates allow us to quantify the noise, which was found to be just under 0.2 (Table 4.2). As expected, the baseline distance increases with time due to additional variations from equilibration kinetics over time.

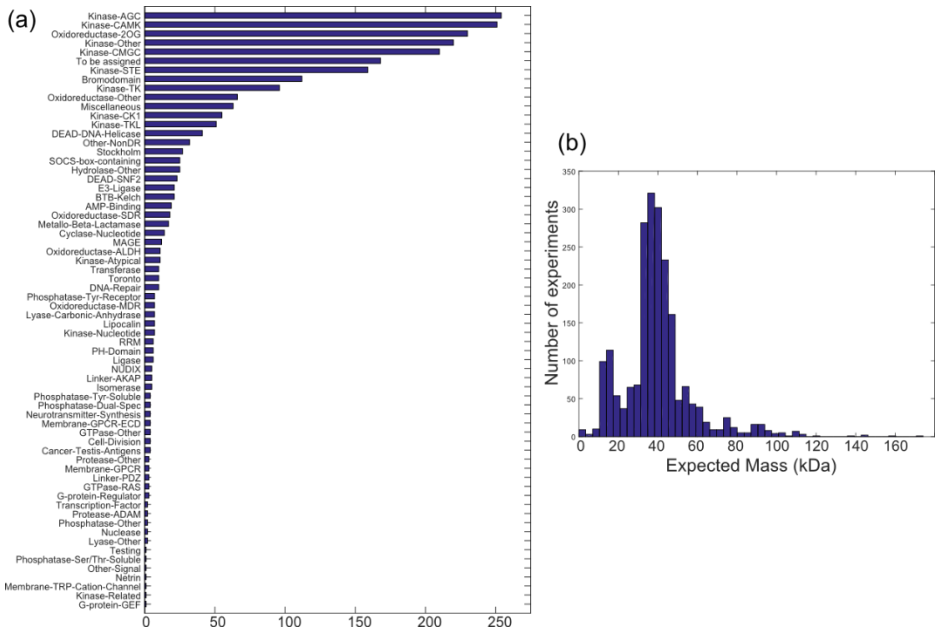



*Figure 4.6: Protein samples used to form the PFL for JCSG at 4°C, summarized by their (a) classification of protein family, and (b) expected mass. These are crude summary of 2410 unique experiments, and shows the variety of protein families and sizes of protein covered.*

*Table 4.2: The pairwise distances between five plates containing identical protein in JCSG screen, imaged after 1 day, 2 days and 4 days.*

| Plate | 1-2 | 1-3 | 1-4 | 1-5 | 2-3 | 2-4 | 2-5 | 3-4 | 3-5 | 4-5 | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| +1 d | 0.185 | 0.204 | 0.205 | 0.189 | 0.174 | 0.182 | 0.165 | 0.193 | 0.180 | 0.175 | **0.185** |
| +2 d | 0.197 | 0.212 | 0.214 | 0.196 | 0.180 | 0.188 | 0.171 | 0.202 | 0.191 | 0.179 | **0.193** |
| +4 d | 0.202 | 0.215 | 0.216 | 0.198 | 0.189 | 0.192 | 0.179 | 0.205 | 0.196 | 0.182 | **0.197** |

The internal plate-to-plate pairwise distances between experiments in our libraries were found to be Gaussian distributed (Figure 4.7, $R^2 \approx 1$). In the context of these distributions, the baseline plate-to-plate distance is at least 2.5 standard deviations away from the mean of all libraries, indicating little to no duplicate experiments, affirming the effective removal of duplicate experiments. The Gaussian distribution for all libraries allows us to define a common reference

point for all PFLs, by converting the distances to z-scores with respect to the library's internal mean and standard deviation.

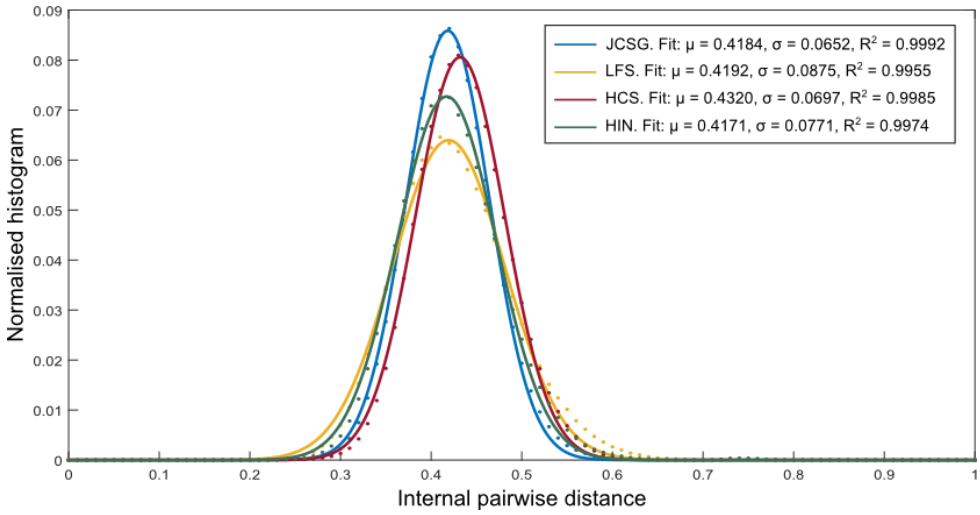



*Figure 4.7: Gaussian fits of the distributions of pairwise distances of experiments in the PFLs. Shown here are the PFLs for 4°C.*

### 4.2.3 Consistent separation of experiments in PFL-space.

For our hypothesis on the collective precipitation behaviour across a fixed set of conditions being a reliable descriptor of an experiment to be true, experiments should be discriminated from each other consistently, regardless of the set of conditions (screen-type) used; since it is the protein that is the subject of characterization, the comparison of the precipitation patterns of two different proteins, be it in JCSG or HIN for example, should reflect them as different.

To test this hypothesis, a data set of 22 experiments (Table 4.3, with further details in Table S2 in Supplementary Materials) screened with both JCSG and HIN was selected, where each plate had less than 5% of faulty droplets to minimise noise from non-experiments. The goals of this exercise were: (1) to determine if the pairwise distances of the experiments also reflect similarities of the experiments, and (2) if the experiments can be discriminated similarly with

either screen-type. The pairwise distances between experiments in the same screen are shown as heat maps in Figure 4.8.

Some of the selected experiments are notably more similar to each other, for example, Experiments 13 to 16 (Table 4.3), which are of different purification batches of the same target from constructs with sequence similarity of > 98% (see Figure S5 in Supplementary Materials). They were further set up with only slight differences in the sample concentrations for crystallization. This is reflected by a region of high similarity (blue patch on rows 13-16 and columns 13-16) in the distance matrices in Figure 4.8. The high sequence similarity between Experiments 1, 2, and 7 (Figure S5) was however not reflected in Figure 4.8, which was expected since the sample concentration of Experiment 7 is double that of Experiments 1 and 2 Together, these form subjective, yet reasonable achievement of goal (1). Sequence similarity thus does not necessarily result in precipitation pattern similarity (comparing Figure 4.8 and Figure S5), furthering our case that sequence and protein properties alone are insufficient to characterize and reflect experimental reality; additional information can instead be derived from precipitation patterns.

Notably, the strong positive correlation between the two resulting distance matrices (correlation coefficient = 0.9650) indicates that both screen-types characterize the experiments consistently and have similar discriminatory power, *i.e.* when two experiments are different, they are described as different by either screen, likewise, when two experiments are similar, they are described as similar by either screen. Furthermore, this correlation coefficient only decreased to 0.9409 when droplets from the 27 overlapping conditions of JCSG and HIN were excluded, indicating that the strong correlation was not biased by identical experiments. Two experiments are thus consistently identified as relatively similar or relatively different, regardless of the screen-type or set of conditions used, in line with goal (2).

Hence, while the selection of screen-type affects the sampling of chemical space to directly obtain crystal hits, it does not limit its use as an assay of protein precipitation. The vast sampling of chemical space of a screen is sufficient to characterize the experiment through its precipitation pattern and discriminate it from others, or identify them as similar. Figure 4.9 further shows that a correlation coefficient of > 0.9 can be achieved with just 15 wells or conditions, indicating fewer droplets may be sufficient to discriminate experiments. The minimum number of droplets for consistent characterization in a generalized dataset however requires further investigation, especially for its consistency in suggesting conditions.

*Table 4.3: Experiments screened with both JCSG and HIN. An experiment comprises of the protein batch (identified by its purification ID), co-crystallized with up to two compounds at a given protein concentration. All experiments selected here also had less than 5% of bad drops in their screens. Further details of protein samples can be found in Table S2 in Section S.3 in Supplementary Materials.*

| Label | Purification ID | Conc (mg/ml) | Compound 1 | Compound 2 |
|---|---|---|---|---|
| 1 | JARID1BA-p082 | 8 | S39130a | Z00052 |
| 2 | JARID1BA-p092 | 7.7 | S40263a | Z00052 |
| 3 | SPIN3A-p013 | 14.78 | none | none |
| 4 | ZAKA-p026 | 20 | Z00113 | none |
| 5 | ZFYVE9C-p011 | 10 | none | none |
| 6 | CECR2A-p030 | 6.5 | none | none |
| 7 | JARID1BA-p070 | 16.8 | SP0372a | S00218 |
| 8 | PAHA-p006 | 17.43 | Z00061 | none |
| 9 | PRKCBP1A-p005 | 12 | none | none |
| 10 | STK6A-p009 | 26 | K03093a | none |
| 11 | XX02RIPK2A-p001 | 1 | Z00094 | none |
| 12 | XX08LATS1A-p001 | 10 | Z00113 | none |
| 13 | BTBD12B-p003 | 8.2 | none | none |
| 14 | BTBD12B-p004 | 9.8 | none | none |
| 15 | BTBD12B-p007 | 7.5 | none | none |
| 16 | BTBD12B-p008 | 6.5 | none | none |
| 17 | CECR2A-p029 | 9.56 | E47242a | E47242a |
| 18 | EP300A-p031 | 13 | E27907a | none |
| 19 | GADD45BA-p001 | 9.5 | none | none |
| 20 | JARID2A-p024 | 9.7 | S00218 | Z00045 |
| 21 | JARID2A-p026 | 12.5 | S00218 | Z00045 |
| 22 | PXKA-p011 | 23 | none | none |

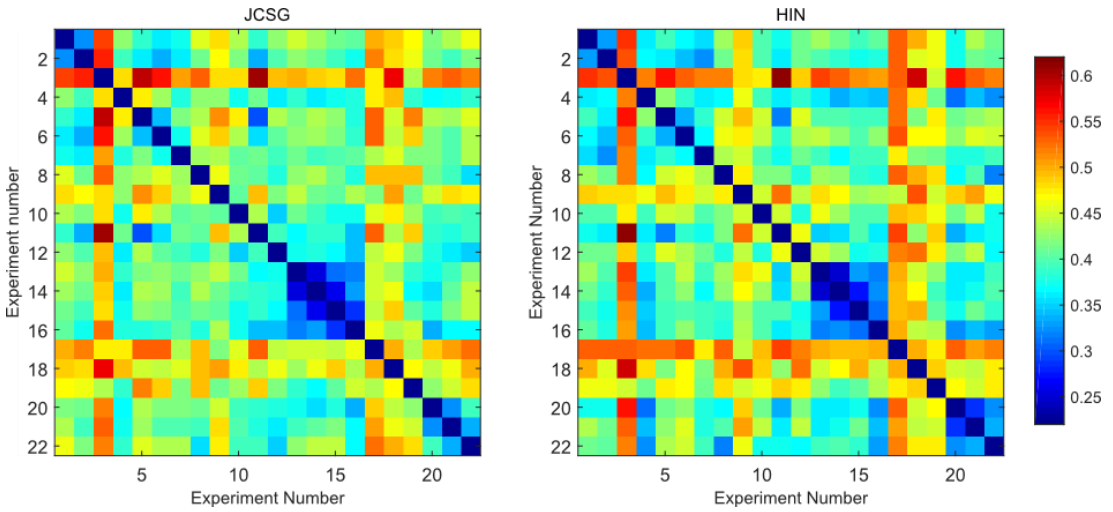



*Figure 4.8: Distances between pairs of experiments. Each row and column represents the corresponding experiment in Table 4.2, and element (i, j) is the distance between experiment i and experiment j, calculated by the precipitation patterns in JCSG (left panel) or HIN (right panel). The correlation coefficient of the two distance matrices is 0.9650, showing that both screens describe the selected experiments in a similar manner.*

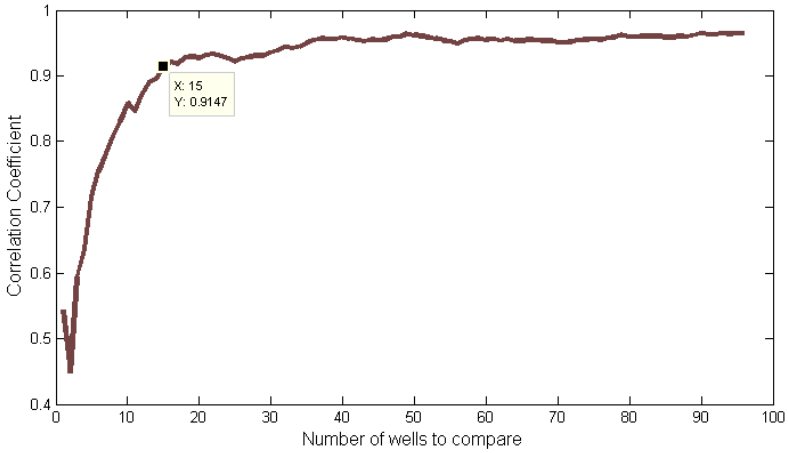



*Figure 4.9: Correlation coefficient of distance matrices of the same set of experiments in JCSG and HIN, calculated using incremental number of wells. A correlation > 0.9 is achieved with just 15 wells.*

### 4.2.4 Experimental validation: Follow-up screen design from suggested conditions

#### *4.2.4.1 Protein target selection*

As a proof of concept, and to evaluate the effectiveness of the predicted conditions, we designed follow-up screens for 17 human protein samples, 14 of which have been crystallized before. The majority of the test proteins had known crystallization conditions to ensure that failure of our approach at producing crystals, can be confidently attributed to the unsuitability of the suggested conditions, instead of the possibility of the protein being inherently uncrystallisable. The crystallisable proteins were further classified as 'easy' or 'difficult' samples, where easy samples crystallize in a broad range of conditions (> 3 conditions in sparse-matrix screens) within 3 days. Details of the selected targets are in Table 4.4, with further information in Table S2 in Supplementary Materials. Although three JMJD2AA samples were used, two were different epitope mutations constructs (p092 and p097) while the other contained just the tudor domain (p100), and hence were considered different samples. The follow-up screen for each experiment was designed based on the suggested optimization conditions derived from one or more sparse-matrix screens set up previously. For all 16 proteins, a follow-up screen was designed, and for two selected proteins (BRD1A-p040 and VPS28A-p003), a second follow-up screen was created.

#### *4.2.4.2 Follow-up screen design*

The principle of our follow-up experiment strategy is to sample a suggested condition on a 4-by-4 grid, allowing six **base conditions** to be tested on a typical 96-well plate, minimising protein sample consumption while allowing for some fine-tuning of conditions. We generated crystallization profiles using the number of nearest-plates that gave ~6 base conditions, combined from multiple screen-types if available. However, a subsequent successful plate was not included if it was more than 10 plates apart from the previous successful plate, since the unsuccessful experiments in between indicate the low probability of its usefulness. The

complete set of base conditions used in our follow-up screen design for each target are found in Table S5 (Section S.3 in Supplementary Materials).

*Table 4.4: Protein samples used to evaluate the follow-up strategy, and the sparse-matrix screens previously set up which were compared to the respective precipitation fingerprint libraries. If crystals of the protein sample (the specific purification ID) have been previously identified, they were deemed crystallisable. Easy samples were those that crystallize in > 3 conditions in sparse matrix screens. Further details of protein samples can be found in Table S2 in Section S.3 in Supplementary Materials.*

| **Protein purification ID** | **Crystalli -sable?** | **Easy/ Difficult** | **Sparse-matrix screens available** | **Protein prep purified by** |
|---|---|---|---|---|
| ATAD2A-p033 | Yes | Easy | HIN | Michael Fairhead |
| BRD1A-p040 | Yes | Easy | HIN | Michael Fairhead |
| GYG2A-p033 | Yes | Easy | JCSG, HIN | Fiona Fitzpatrick |
| JMJD2AA-p092 | Yes | Easy | HIN | Michael Fairhead |
| JMJD2AA-p097 | Yes | Easy | HIN | Michael Fairhead |
| JMJD2DA-p037 | Yes | Easy | HIN | Michael Fairhead |
| PRKCBP1A-p023 | Yes | Easy | HIN | Michael Fairhead |
| CAMK1DA-p014 | Yes | Easy | HCS, HIN | Fiona Sorrell |
| GYG2A-p032 | Yes | Difficult | JCSG, HIN | Fiona Fitzpatrick |
| CECR2A-p031 | Yes | Difficult | HIN | Michael Fairhead |
| JMJD2AA-p100 (tudor) | Yes | Difficult | HIN | Michael Fairhead |
| JMJD2CA-p067 | Yes | Difficult | HIN | Michael Fairhead |
| VPS28A-p003 | Yes | Difficult | JCSG, LFS, HIN | Jolanta Kopec |
| FAM83AA-p007 | Yes | Difficult | HIN | Michael Fairhead |
| DOPVA-p002 | No | Difficult | JCSG, LFS, HCS, HIN | Ritika Sethi |
| LACTB2A-p118 | No | Difficult | JCSG, LFS, HCS, HIN | Hazel Aitkenhead |
| OTUB2A-p006 | No | Difficult | JCSG, LFS, HCS, HIN | Ritika Sethi |

Random grids were designed for samples where only 5 base conditions were identified. To generate random grids, the components of the suggested base conditions were classified as (1) major precipitant, (2) minor precipitant, (3) buffer, and (4) organic. Each base condition must have only one major precipitant, while other components were optional. A random condition was formed by independently drawing from each class with replacement, resulting in only one major precipitant in each new condition, randomly combined with other components. In cases

where there were less than 5 conditions identified, we swapped components in the identified conditions to create new base conditions. Where there were more than 6 conditions identified, we consolidated them to 6 base conditions by combining conditions that share majority of components.

Sampling of concentration levels in a 4-by-4 grid for a typical base condition was as such: the major precipitant was always sampled at 4 concentration levels, while other precipitants may be sampled at 2 or 4 levels depending on the number of components. Figure 4.10(a) and Figure 4.10(b) summarizes the general guideline used to design sampling grids for base-conditions. The design for consolidated base conditions however, were tailored to sample as equally as possible for each of the grouped conditions. Examples of how consolidation was carried out is shown in Figure 4.10c.

#### *4.2.4.3 Positive control experiments*

To ensure that the crystallisable proteins used were viable, wells H9, H10, H11 and H12 were designated as positive controls, where a previously crystallising condition was used to reproduce the crystals. The multiplicity reduces the probability of missed crystallization due to stochastic nucleation events, and concentration of components of the previously found condition were slightly varied (± 1% for polymers, ± 0.01 M for salts and ± 0.1 pH) to account for possible liquid dispensing errors.

#### *4.2.4.4 Experimental setup*

An in-house Excel-based screen designer was used to design the screens with standard commercial stock solutions. Given the concentrations for each chemical component in a well,

the screen designer generates the worklist file required by the MPII liquid handler used at the SGC for screen formulation, and writes information of the screen into Scarab, SGC's main database for storing experimental information. Each follow-up screen was formulated with the MPII liquid handler as 400µl blocks. With the exception of one target (DOPVA, due to limited protein sample), two plates were set up for each follow-up design, incubated at 4 and 20°C. All protein samples were rapid-thawed from -80°C and centrifuged at 14,000 rpm for 3 minutes to remove precipitates. Crystallization experiments were set up according to the standard configuration described previously with the Mosquito (TTP Labtech) and imaged with the Minstrel HT system (Rigaku).

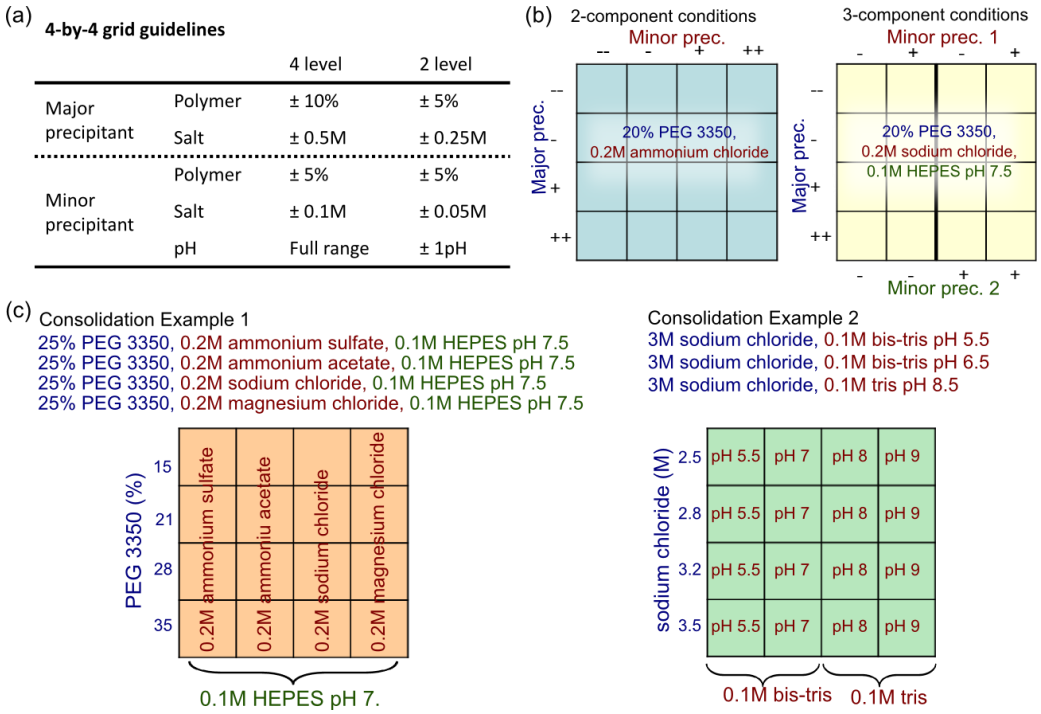



*Figure 4.10: General guidelines adopted in the follow-up screen design. A 96-well plate is divided into six 4-by-4 blocks, each hosting a base condition. Only one major precipitant is identified in each base condition, while the other components are classified as minor precipitants. (a) shows the range to sample for major and minor precipitants of each type, while (b) shows the arrangement in a block for 2-component or 3-component conditions. Examples of consolidating similar conditions into a block is shown in (c), where the guidelines in (a) and (b) do not strictly apply.*

## 4.3 Experimental results and discussion

The primary goal of the experimental validation was to evaluate the effectiveness of our suggested conditions in crystallizing a given protein, as well as our follow-up screen design. We also evaluated the diversity of suggested conditions for each selected target to determine whether only similar conditions were suggested, and attempted to infer crystallizability from an experiment's Nearest-plates Curve. For crystallizing systems, we further sought to identify if our approach gave alternative crystallizing conditions that increased diffraction quality, crystal reproducibility, or result in alternative crystal forms.

### 4.3.1 Effectiveness of suggested conditions

The outcome of each suggested base condition is summarized in Figure 4.11. At least one predicted condition for 11 of 14 crystallisable proteins successfully produced crystals (green ticks in Figure 4.11). These included crystals that were suitable for harvesting and microcrystals (dark and light blue disks respectively in Figure 4.11), with varying results at different incubation temperatures. Although some of the resulting crystals were unsuitable for data collection, they were nevertheless starting points for further optimization. Two of three crystallisable targets (JMJD2AA-p100(tudor), CAMK1DA-p014) that failed to produce any crystals in our follow-up screen design also did not crystallize in the positive control wells (last column of Figure 4.11). Efforts to reproduce these crystals may have failed due to the overly-narrow variations of the previously successful conditions, missed nucleation in all four wells, or degradation of the protein samples. The only confirmed negative result of our method was for FAM83AA-p007, which crystallized reproducibly in the positive control, but not in any of our suggested conditions.

The remaining test samples (DOPVA-p002, LACTB2A-p118 and OTUB2A-p006) have so far failed to crystallize, even after substantial efforts by the respective scientists involved. Our approach did not fare any better; but we are now more confident in classifying these proteins as uncrystallisable.

### 4.3.2 Effectiveness of follow-up screen design

#### *4.3.2.1 4-by-4 grid sampling of untested and the re-sampling of base conditions*

We found high success rates for conditions previously untested (screen-type containing the condition was not set up), and an increased success rate for conditions previously sampled. The background colours in Figure 4.11 indicate if a base condition was previously sampled, and if so, were crystals formed in these conditions. 33% of our base conditions have not been tested (white background), since setting up multiple screen-types requires more sample, which may not always be available. 23 of these 35 untested base-conditions (66%) resulted in crystals in our follow-up design. As for the 70 previously sampled conditions, only 3 produced crystals in the previous screen set up (orange background); these were successfully reproduced here. In contrast, 21 of our 4-by-4 coarse grids of these 70 base conditions produced crystals, giving an increased success rate of 30%. The higher success rate may be due to the increased fine-tuning of concentrations of the condition across the 4-by-4 grids, instead of a single experiment. The 16-fold multiplicity, although at different concentration levels, also increases the likelihood of crystallization by increasing the probability of nucleation.

Moreover, our follow-up design also allows the inference of crystal reproducibility for a base condition from the number of wells in the 4-by-4 grid that crystallized (proportional to radius of the circles in Figure 4.11). A high number indicates reproducibility of crystals in a broad range of concentrations sampled by the grid; a low proportion of crystals in a grid, especially if

crystallization only occurs in a sub-region of the block, indicates a narrow range of concentrations for crystallization, and hence more challenging to reproduce.

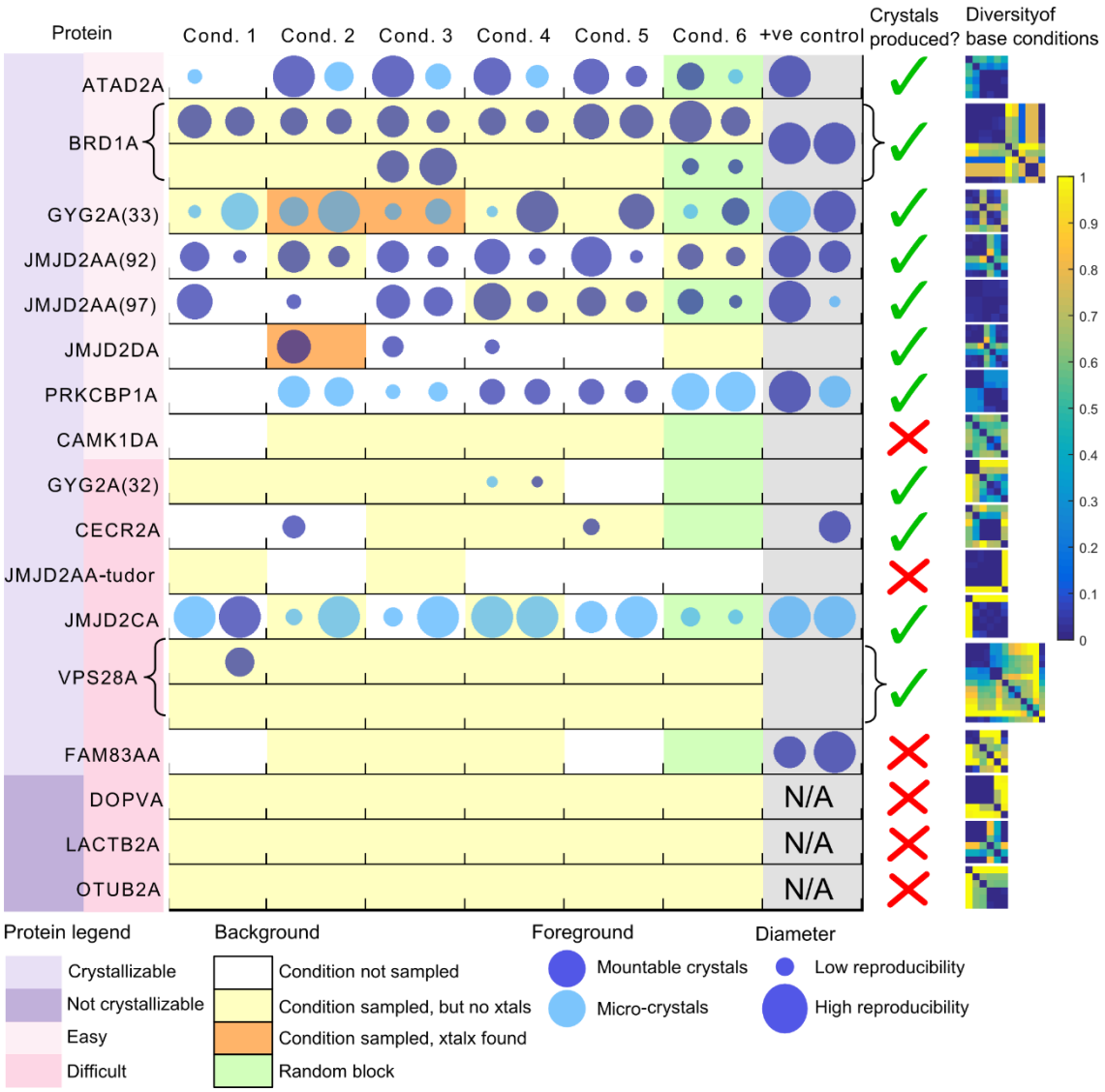



*Figure 4.11: Crystallization outcome for the suggested conditions for the test samples. The protein samples have been colour-coded to indicate if they are crystallisable (light purple) or not (dark purple), and if they were easy (light pink) or difficult (dark pink). Each row represents a follow-up design with 6 base conditions, with each condition occupying a column. Background colours indicate if a condition was previously sampled, and if so, whether it resulted in crystals. Dark and light blue disks mark if mountable crystals or micro-crystals were found in that condition respectively. The first and second circle in every column correspond to (micro)crystals at 4°C and 20°C. The radius of the circle is proportional to the number of wells in the 4-by-4 grid that resulted in crystalline behaviour, an indirect measure of the reproducibility of crystals in that condition. The grey column indicates the outcome of the four*

*positive control wells where applicable. Crystals were obtained for 11 of the targets tested. The left-most panel shows the pairwise distances between base conditions in the designed screen, excluding conditions in the random block. The last column/row of each distance matrix reflects the distance of the positive control condition (if available) to the other base conditions.*

#### *4.3.2.2 Random block*

The random block design (green background) was based on the work of Klei *et al.* (Klei et al., 2011). In their work, a new and untested condition successfully co-crystallized HCV NS3 protease (HCVPr) with a compound that was resistant to other previously successful conditions. The new condition was identified through a screen designed by randomly combining components of proven conditions from other co-crystallization structures of HCVPr. Here, we extend this to a random combination of potential conditions, but found that the random blocks were only successful when other base conditions were also successful. While our sample size may be too small to make far reaching conclusions, our observation was that the crystals generated were not superior to those from the base conditions.

### 4.3.3 Diversity of predicted condition

To investigate if the suggested base conditions were chemically diverse, the distance matrices calculated with Cockatoo (Bruno et al., 2014) for each set of suggested base conditions is shown in the far-right panel of Figure 4.11. For simplicity, one representative condition was selected for blocks with consolidated base conditions, and random blocks were left out of this analysis. The last column/row of each matrix shows the distance of the positive-control condition (if available) to the suggested conditions. Based on these distances alone, the internal diversity varies for each target: there is low diversity in the predicted conditions for JMJD2AA-p097 and JMJD2AA-p100 (tudor domain), but high diversity for VPS28A-p003 and FAM83AA-p007.

This is further shown in Figure 4.12, where the sorted distances of the all predicted base conditions (including those in random blocks) for each protein sample to its respective positive control condition are plotted. The predicted conditions had different curves for each target, and were different from a baseline distribution generated by 10,000 repetitions of randomly selecting a base condition (analogous to a positive control), and comparing it to 10 other randomly selected conditions (median in blue dotted lines, bounded by the 25$^{th}$ and 75$^{th}$ percentile in Figure 4.12). The diversity of chemical space of our predicted conditions are thus case dependent, in line with the general observations that some proteins crystallize promiscuously while others will only crystallize in a narrow range of chemical conditions.

Moreover, since JMJD2AA-p092 and JMJD2AA-p097 were epitope mutations of the wild-type protein, they differed only marginally in molecular weight (see Table S2). However, a different set of conditions were suggested for these experiments, all of which successfully generated crystals. This indicates that even though samples may share similar properties or characteristics, the suggested conditions may differ if the precipitation behaviour were different, since mutations may drastically alter the behaviour of a protein and it's interaction with other molecules.

It should be noted that the Cockatoo distances are biased by the size of the molecular weight of the components in the condition. For example, the distance between conditions '2M ammonium sulfate, 0.1M acetate pH 4.5' and '2M ammonium sulfate' is 0.0244, but the distance between conditions '25% PEG 3350, 0.1M acetate pH 4.5' and '25% PEG 3350' is 0.0011, even though both pairs of conditions only differ by the presence or absence of acetate buffer. Hence, PEG conditions generally result in lower distances, and reflects lower diversity in our distance matrices. However, the Cockatoo distances nevertheless provide a quick overview of diversity, especially in differentiating PEG and non-PEG conditions, and allows for a meaningful

arrangement of the supported conditions for the Crystallization Conditions Profile (as in Figure 4.5).

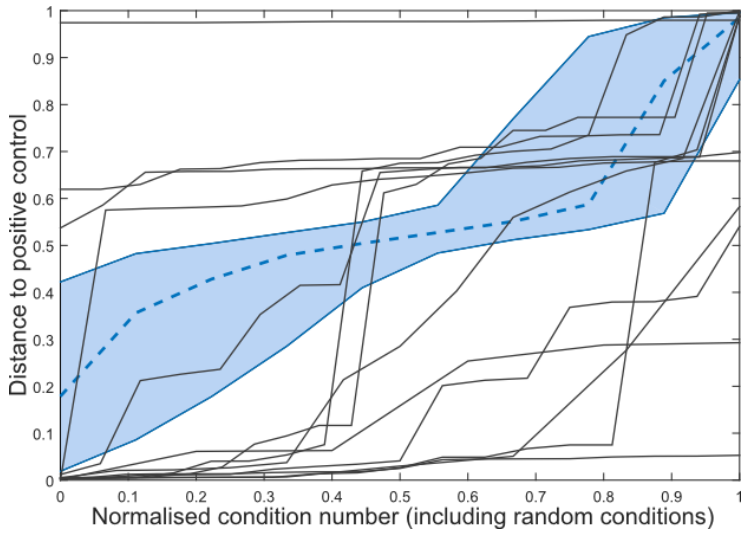



*Figure 4.12: Sorted distances of predicted base conditions to their respective positive controls. The differences in curve shapes show that each set is unique to its protein sample. They were also different from a random distribution generated by randomly selecting a condition and comparing it to 10 other random conditions (blue dotted line marks the mean of this distribution, bounded by the 25$^{th}$ and 75$^{th}$ percentile), indicating that our predicted condition lists were not random.*

### 4.3.4 Inferring crystallizability from nearest-plate curve

A retrospective analysis of the Nearest-plates Curves for the selected targets and its follow-up outcome indicates that they may form a framework for inferring crystallizability, particularly in its overall z-scores and the rank of successful nearest-plates. We found that the successful follow-up experiments consistently had (1) at least one successful plate among its first 25 nearest-plates, and (2) at least three successful plates below the first standard deviation threshold. If these criteria were used to determine if follow-up effort should be carried out, the strategy would have been employed for 9 of the 11 targets where our strategy successfully produced crystals, as well as for the crystallisable samples JMJD2AA-p100 and CAMK1DA-p014,

even though our method failed to crystalize these. The two cases where our follow-up strategy worked outside these criteria (ATAD2A-p033 and JMJD2CA-p067) had less than 3 nearest-plates below the one-standard deviation threshold, indicating the limited information from experiments with similar behaviour in the PFLs, and hence low predictability. Some example profiles from the 75 nearest-plates are shown in Figure 4.13.

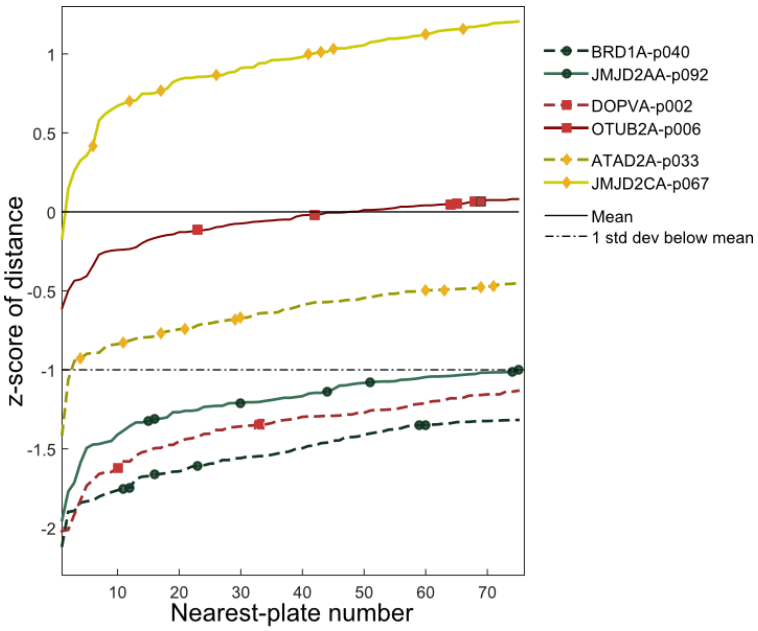



*Figure 4.13: Examples of nearest-plate curves used to identify potential conditions in our follow-up experiments. The y-axis of the plot is the z-score of nearest-plate distances with respect to the mean of the internal distances of the respective libraries. Plates with associated crystallizing conditions are marked with coloured markers on the curve. The normalised mean and first standard deviation below the mean are marked by the black solid line dotted line respectively. Our follow-up strategy for BRD1A-p040 and JMJD2AA-p092 (green lines) resulted in crystals. Both curves were, and had many successes below the one-standard-deviation threshold. The red curves (DOPVA-p002 and OTUB2A-p006) were profiles from uncrystallisable proteins. We note that the few successes in DOPVA-p002's curve, and the lack of success early in the curve of OTUB2A-p006 may indicate the difficult of these targets. Though follow-up design for ATAD2A-p033 and JMJD2CA-p067 (yellow curves) also gave crystals, the distances of these two plates, along with that of OTUB2A-p006 from their respective libraries indicate the lack of similar fingerprints in the library for accurate inference. These curves, along with those from the other protein samples led us to define the criteria for inferring crystallizability described in the text.*

### 4.3.5 Beyond identifying crystallization conditions

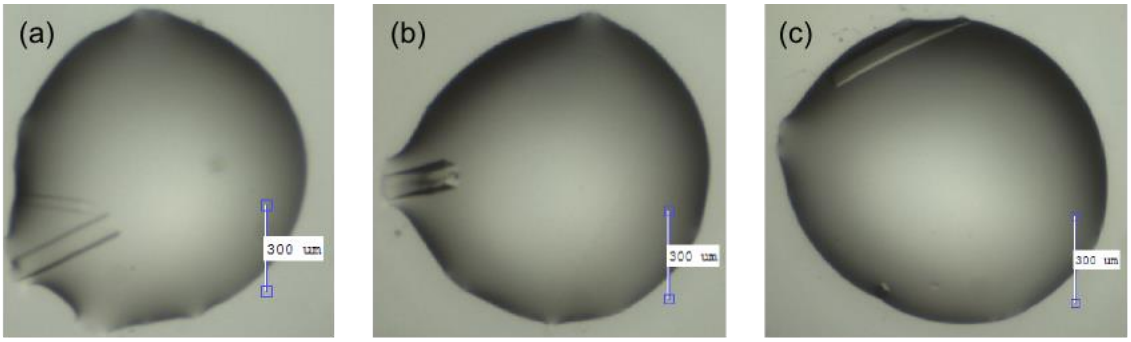
As a spot-check exercise, datasets for selected crystals (Table 4.5) were collected at I04-1, Diamond Light Source, under the supervision of Dr Romain Talon. We focused on crystals with different morphology than those observed previously, and those for which no high resolution datasets were available. From the datasets collected, we found that the experimental design for our test proteins also resulted in (a) new crystal forms, (b) improved diffraction resolution, (c) larger crystals, as well as (d) faster and more reproducible crystal growth.

*Table 4.5: Datasets collected for selected crystals.*

| CrystalID | Condition | Res. (Å) |
|---|---|---|
| BRD1A-x787 | 35% PEG3350, 0.25M sodium chloride, 0.1M bis-tris pH 7.5 | 2.16 |
| BRD1A-x788 | 30% PEG3350, 0.16M sodium malonate | 1.33 |
| JMJD2CA-x198 | 0.975M succinic acid | 3.43 |
| JMJD2AA-x866 | 20% PEG5KMME, 0.25M ammonium sulfate, 0.1M MES pH 6.5 | 1.35 |
| JMJD2AA-x867 | 35% PEG3350, 0.2M ammonium sulfate, 0.1M tris pH 7.5 | 1.83 |
| JMJD2AA-x868 | 21% PEG3350, 0.2M ammonium acetate, 0.1M tris pH 7.5 | 2.41 |
| JMJD2DA-x1066 | 20% PEG3350, 0.25M proline (L-), 0.1M HEPES pH 7.0 | 3.10 |
| VPS28A-x053 | 2.65M sodium acetate | 2.80 |
| PRKCBP1A-x087 | 35% PEG3350, 0.15M lithium sulfate, 0.1M bis-tris pH 5.5 | Not phased |

#### *4.3.5.1 New crystal form*

A new crystal form was identified for JMJD2AA. Constructs for both JMJD2AA-p092 and JMJD2AA-p097 were specifically designed to increase crystallizability through epitope/surface mutations. Datasets of three different crystal forms were collected (Figure 4.14). While two of these had been previously observed, our crystals diffracted to marginally higher resolution. The packing of protein molecules in the crystal with space group P2221 was novel. Identifying different crystal forms is especially important for soaking experiments, where the active site may or may not be accessible depending on the crystal packing or crystal contacts. All JMJD2AA structures reported here were modelled by Dr Michael Fairhead.




20% PEG 5KMME,
0.25M ammonium sulfate,
0.1M MES pH 6.5

35% PEG 3350,
0.25M ammonium sulfate,
0.1M Tris pH 7.5

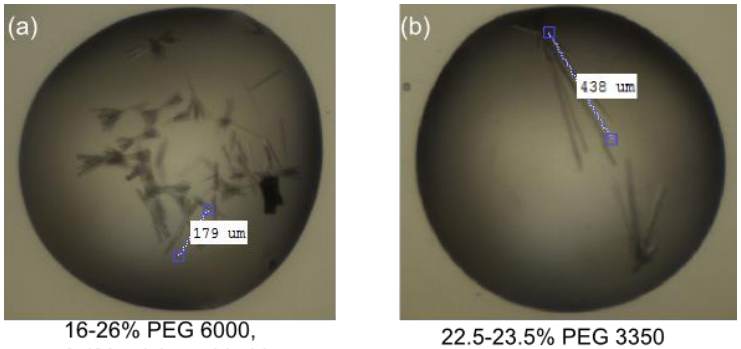
21% PEG 3350,
0.25M ammonium acetate,
0.1M Tris pH 7.5

| | Res.(Å) | Previous best | Space group | a (Å) | b (Å) | c (Å) | α(°) | β(°) | γ(°) |
|---|---|---|---|---|---|---|---|---|---|
| (a) | 1.4 | 2.2 | P 1 21 1 | 57.03 | 101.6 | 141.43 | 90 | 99.38 | 90 |
| (b) | 1.8 | 1.9 | C 1 2 1 | 117.7 | 90.78 | 90.67 | 90 | 107.19 | 90 |
| (c) | 2.4 | -- | P 2 2 21 | 56.79 | 100.47 | 148.05 | 90 | 90 | 90 |

*Figure 4.14: Datasets collected for JMJD2AA. Three datasets were collected, each with different crystal form. (a) and (b) were previously identified, while (c) though with a previously observed space group, had new packing, and thus crystal form.*

#### *4.3.5.2 Larger, reproducible crystals*

As part of an ongoing fragment-soaking experiment, reproducible crystals for ATAD2A-p033 were required. The clustered rod-shaped crystals at 20°C were known to be the better diffracting crystals. The best rod-shaped crystals in our follow-up design for ATAD2A were found in 23% PEG 3350, 0.16M magnesium formate, a condition not previously sampled. We reproduced crystals for this condition in a new plate, and compared it with a separate optimization screen designed by the owner of the protein sample, Dr Romain Talon (16% to 26% PEG 6000, 0.1M calcium chloride, 10% ethylene glycol, 0.1M tris pH 7.5). After 7 days, we found over 150 droplets with at least one crystal that was >400µm along the largest dimension in our condition, compared to under 50 droplets with such crystals in the screen set up by the owner. This may be due to the fewer nucleation points in our conditions, thus allowing better crystal growth. Figure 4.15 shows the comparison of the crystals that typically appear in both of these conditions.




16-26% PEG 6000,
0.1M calcium chloride
10% ethylene glycol
0.1M tris pH7.5

22.5-23.5% PEG 3350
0.16M magnesium formate

*Figure 4.15: Comparison of crystals from a previous optimization screen and crystals from our follow-up design. The previous optimization screen had just under 50 droplets with >400µm crystals, and most droplets containing showers of small rods as in (a). Larger crystals are more reproducible in our condition (23% PEG 3350, 0.16M magnesium formate), where over 150 droplets had >400µm crystals (b), while other droplets remained clear.*

#### *4.3.5.3 Faster crystallization, higher resolution crystal*

The crystal previously identified for VPS28A-p003 appeared only after 15 days, and diffracted to ~7Å. Since there was an abundant supply of the protein, we designed two follow-up screens, exploring 12 base conditions obtained from the JCSG, LFS and HIN screens set up previously. Figure 4.16a shows that the JCSG and LFS plates were generally distant from plates in their respective PFLs, above the one standard deviation cut-off. The HIN screen however had more comparable experiments, with the first success found in its nearest plate. Interestingly, it was the condition from this nearest-plate (2.8M sodium acetate) that resulted in our crystals, which appeared within 24 hours and gave an improved diffraction resolution of ~3Å. Interestingly, this condition resulted in clear drop in the initial screen and also failed to crystallize the protein at 2.8M in our grid; crystals were only produced at 2.55M to 2.75M (Figure 4.16b), thus emphasizing the importance of exploring potential conditions with more than just a single experiment. Exploring all 96 conditions from a sparse-matrix screen beyond a single experiment

is costly; our analysis allows the narrowing of this to an informed selection that can be practically achieved. All VPS28A datasets were processed by Dr Jolanta Kopec.

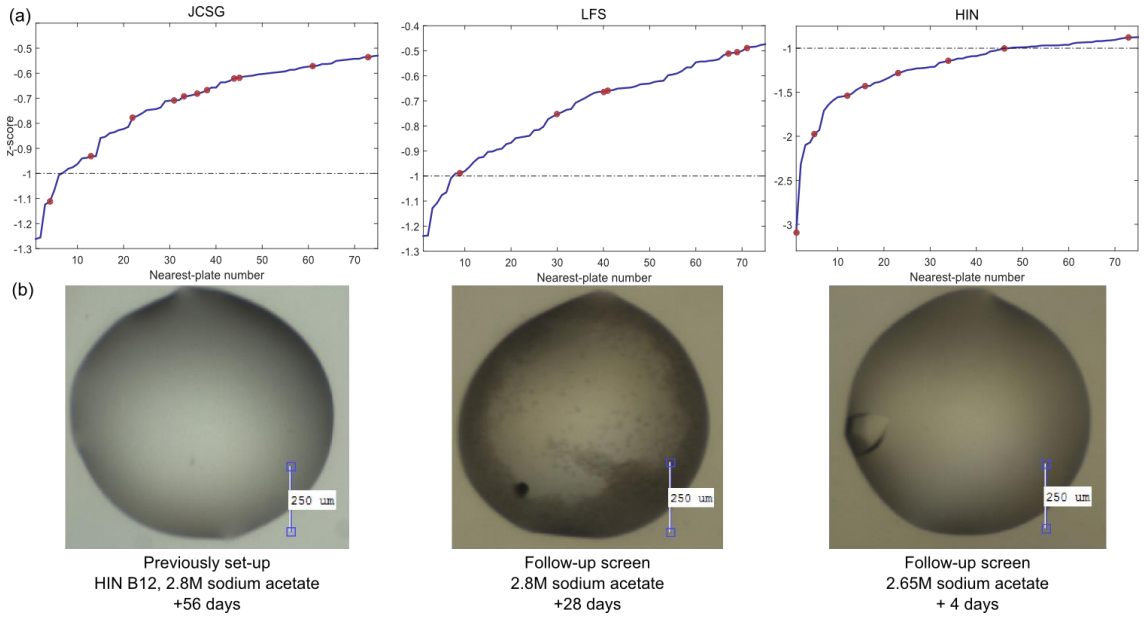



*Figure 4.16: Nearest-plate curve and crystallization outcome for VPS28A-p003. (a) The HIN plate provided the most informative nearest-plate curve, and the nearest neighbour in this screen gave the only base condition that was successful: 2.8M sodium acetate. (b) A look back at the initial screen, the droplet with 2.8M sodium acetate remained clear after 56 days (left). In our 4-by-4 grid, VPS28A-p003 also did not crystallize at 2.8M sodium acetate (centre); instead crystals were found between 2.55M and 2.75M within 24 hours. The crystal shown in the right-most figure diffracted to ~3Å.*

## 4.4 From classification to clustering to nearest-plates

This section briefly recounts the progression of strategies explored to use precipitation fingerprints from sparse-matrix screens to infer crystallisability and identify potential follow-up conditions for a given protein. The input for all methods described was the normalised texton distribution of droplets in a screen.

### 4.4.1 Classification

The first attempt at predicting crystallizability was to classify a given protein as crystallisable or not, based on its precipitation fingerprint in a given sparse-matrix screen. The labels of our training set (crystallisable and non-crystallisable) were defined by the presence or absence of diffracting crystal for the experiment in the plate, regardless of where the crystal appeared (any other screen-type or optimization experiments).

Most classification algorithm requires input features to be 1D (vectors). Hence, with 288 droplets per plate and each droplet described by a 300-long vector, a plate has a total of 86,4000 features, $d$. This gives $d \gg n$, where $n$ is the number of training examples, which were in the range of low thousands for any given screen-type at an incubation temperature. Dimensionality reduction was required to reduce the number of features prior to training. A key requirement for the dimensionality reduction method is the mapping of a new data point from high to low dimensional space. This ruled out two of the methods explored: Isomap (Tenenbaum et al., 2000), and Gaussian process latent variable model (GPLVM) (Lawrence, 2004), where re-optimization of the model was required for every new data point. We thus chose principle component analysis (PCA) (Maaten, 2009) to reduce the number of features from 300 to 3 per image, while maintaining maximum variance in the data. This resulted in $d = 864$.

The reduced features were used as inputs to Random Forest classifiers, chosen for its ability to deal with high dimensionality and missing inputs, which was prevalent in the dataset due to empty wells or faulty droplets. Each forest had 500 decision trees, and five classifiers were trained for each of the 20 random splits of our data set with a 9:1 ratio for training and validation. The mean area under the receiver operating curve (AUROC) for all 100 classifiers trained per screen (20 splits × 5 classifiers each) is shown in Figure 4.17(a), and the AUROC for each random split for JCSG only is shown in Figure 4.17(b).

The overall AUROC for each screen was less than satisfactory. This could be due to the imbalanced number of positives and negatives (approximately 1:9 in our data set), making generalization of crystallisable protein behaviour difficult. The large variations in AUROC between random splits (Figure 4.17(b)) show the wide variety of protein behaviour, some of which were not well represented in the training set and hence had low AUROC. Furthermore, while there is high confidence in the positive labels (crystallisable protein), there may be many false negatives, since there are no conclusive ways of determining if a protein is un-crystallisable; nucleation may have been missed, or the right conditions may not have been sampled thus far.

A perfect classification was thus concluded to be impossible, especially with added layers of noise from imaging artefacts and data annotation errors. A hard or fuzzy classification of crystallizability also does not address the issue of crystallizing chemistry. However, with a general AUROC of > 0.5, it can be concluded that there is a non-uniform distribution of positives and negatives, although there is significant overlap between the two populations.

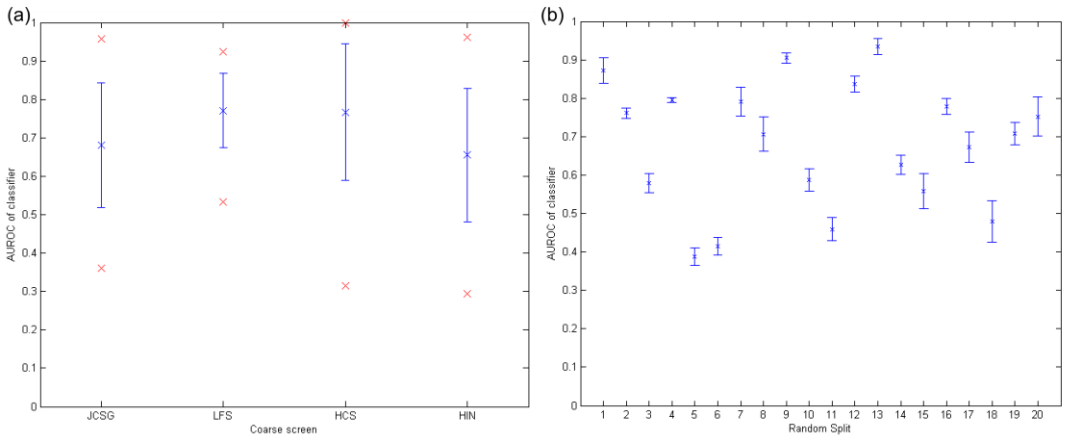



*Figure 4.17: Classification AUROCs. (a) Average AUROC for the four main screen-types. Shown here as red crosses are the maximum and minimum AUROC values obtained. (b) Average AUROC for each random split of the data set for JCSG at 4°C. The inconsistent AUROCs show that the training set heavily influences the classifier's performance.*

### 4.4.2 Clustering without labels

We subsequently chose to identify clusters of protein precipitation behaviour, with the hypothesis of different success rates for different clusters, thus allowing the inference of crystallizability of a protein based on which cluster it falls into. Furthermore, the conditions that produced diffracting crystals for experiments in a cluster can also be used to form condition profiles to suggest optimization conditions, similar to that described in our nearest-plates method.

Hierarchical clustering with Ward's minimum variance method was used to cluster 3259 JCSG plates at 4$^{o}$C, with plate-to-plate distances defined by the Hellinger distance. The resulting dendrogram is shown in Figure 4.18(a). To ensure that our clustering was sensible, we added three control data points (repeated plates), where we expect these to appear side-by-side in the dendrogram since they are most similar to each other. Figure 4.18(a) shows that this is indeed the case. At a cut-off of 25 clusters, the distribution of success rates in each cluster (success = experiments that has associated diffracting crystal(s)), shown in Figure 4.18(b) is non-uniform, supporting our initial hypothesis.

The intended practical use of the clustering approach was this: given a new precipitation fingerprint, the cluster it most likely belongs to can be identified by the majority vote of its *n*-th nearest-neighbours. We chose $n = 15$ which is just under half the size of the smallest cluster (34). From this cluster, the crystallizability is inferred and optimization conditions can be obtained from the profile.

One advantage the clustering method has over our chosen nearest-plates approach is the fixed condition profile for each cluster. With fixed profiles, follow-up screens can be pre-designed and pre-formulated for each cluster, lowering the energy barrier of adoption. However, it was not obvious if our selection of 25 clusters was appropriate. Using Gaussian mixture models with

variational Bayes model selection resulted in a single cluster, indicating no clear partitions of the dataset. Furthermore, the clustering outcome is highly dependent on the clustering algorithm and distance metric, for which the best of these was not obvious. We also often found that the confidence of cluster assignment, determined by the fraction of majority vote from its nearest-neighbours was typically low (<50%).

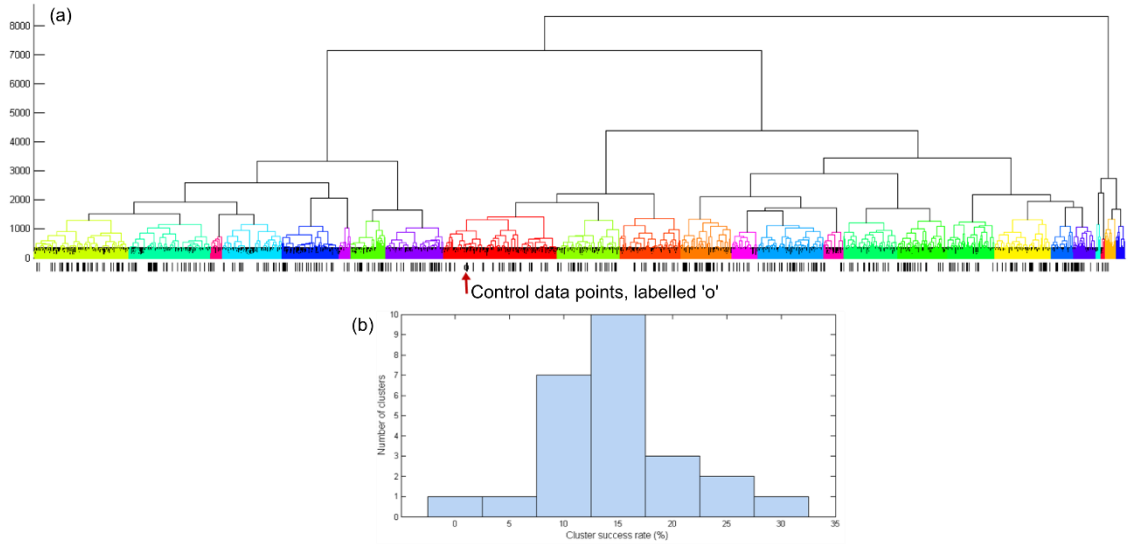



*Figure 4.18: Clusters of JCSG plates and success rate distribution. (a) Hierarchical clustering of 3259 JCSG plates at 4°C. Each node in the dendrogram represents a JCSG plate, and if the experiment of the plate has associated diffracting crystals, it is labelled with '|'. The control data points ('o') were inserted as a validation of the clustering, and as expected, appear side-by-side. The cluster sizes ranged from 34 to 176 plates per cluster. (b) The distribution of success rates of the 25 clusters (colour blocks in (a)). The non-uniform distribution shows that the precipitation fingerprints have different associated crystallisation propensity.*

### 4.4.3 Nearest-plates

Given the lack of understanding and validation of clustering outcome, we simplified the approach, generating the condition profiles directly from the *n*-th nearest-plates to the experiment of interest, instead of a group of experiments from an arbitrary cluster cut-off. This requires the profile to be generated on-the-fly, and will likely be unique to each experiment,

thus requiring the formulation of customised screens for every case. However, the conditions selected are most closely related the present experiment of interest, and we have shown them to be effective in our experiments described earlier.

## 4.5 Further Work

While the limited set of 17 test proteins have shown the potential of our approach, more data is required, especially to form an objective and continuous crystallizability score, in place of the binary inference from the Nearest-plates Curve meeting the present criteria. Ideally, to generate data, screening for new protein targets should always be followed up with at least six suggested conditions (one plate) from the analysis. Features beyond what is currently used, for example the shape of the curve or rank-distribution of successful neighbours, can be extracted from the resulting curves and the effectiveness of the suggested conditions to train sophisticated models. These models should output scores to determine the experiment's crystallizability, and if follow-up effort is worth pursuing.

As mentioned in Chapter 2, the plate-to-plate distance is presently calculated by taking the mean of droplet-to-droplet distances. A better distance may improve plate-to-plate comparison, and thus the nearest-plates curve. The idea of clusters of precipitation behaviour should also be revisited if/when there is a better handle of the experiments being clustered, *i.e.* there exists ground truths to validate the clustering outcome and selection of parameters.

The PFLs for which precipitation fingerprints are compared to are currently stored locally, or on network servers that analyses the images. While this allows fast loading of libraries during execution, updating the library across sites becomes less convenient due to the large and ever increasing file sizes (~200 Mbytes per library). To overcome this, the nearest-plate identification

process can be deployed as a web-server, where users outside the organization may generate the texton features for a plate, and upload the resulting features to the server; this involves the transfer of <150 Kbytes. The server identifies nearest-plates and returns the nearest-plate IDs and distances, which can be used to form the profiles on the user's local computer.

A further extension of this web-server may include public contributions of precipitation fingerprints and its associated successful conditions (or lack of success). This entails the community uploading images from screening experiments (or the texton features from these images), and its corresponding crystallization conditions with verified diffraction. Because our method only requires images and not protein nor ligand information, we foresee no data sensitivity issues, and thus should encourage participation from both academia and industry. Furthermore, partnerships with institutions or commercial entities can be formed to formulate the suggested optimization screens at a low cost for laboratories without such facilities.

## 4.6 Concluding remarks

A typical crystallization screening experiment has high information content that can be extracted through image analysis methods as described. This data can subsequently be used to form our newly defined precipitation fingerprints, which objectively characterize the precipitation behaviour of a protein across a typical sparse-matrix screen. The precipitation fingerprint is arguably the most comprehensive description of the experiment for crystallization, incorporating variables from sample treatment and interactions before and during crystallization.

Potential optimization conditions can be inferred from successful conditions of past experiments with similar fingerprints in a library. While the libraries used here were limited to

the protein targets of the SGC, we believe there is sufficient data for useful inference, and these libraries will only increase in protein coverage over time as more data is incorporated. Our approach successfully produced crystals for 11 of 14 crystallisable proteins, which further resulted in new crystal forms, better diffracting or reproducible crystals. The strategy did not lead to crystals for the three selected proteins that had never yielded crystals, and hence these proteins could be deemed uncrystallisable.

We found similar discriminatory power for different screen-types in characterizing proteins, indicating that the choice of screen-type is not important in this context, as long as there is sufficient prior data for comparison. This has limited our analysis and support to JCSG, LFS, HCS and HIN, but we believe that the high uptake of at least one of these screens in most laboratories leads to general applicability. We are thus not suggesting any change of screen-types, nor requiring additional experiments to be carried out, but rather capitalizing on present day standard screening procedure to extract useful information, independent of crystallinity and without laborious and subjective evaluation of each droplet. Decisions still have to be made on the number of nearest-plates to include when deciding on follow-up experiments. It should be possible to make it more objective with increased usage and the subsequent inclusion of more data.

Previous efforts, including the work of Luft (2011), Nagel (2008) and Snell (2008) (see Chapter 1.4) have also emphasized the need to move from treating crystallization trials as singular experiments to one that considers its global outcome. Our work differs from theirs in three ways: (1) While only subsets of outcomes are used to either derive the empirical phase diagram or crystal-hits islands in their work, we use all droplets, regardless of their outcome; (2) historical data is central as points of reference in our method, but plays little to no role in previous efforts; (3) no subjective labelling and interpretation of crystallization trials are required in our method.

We chose bright-field images as our source data due to its availability in most laboratories. UV-fluorescent imaging of crystallization droplets would be the immediate extension of the algorithm to compliment the signal from bright-field images, if/when bulk data is available. Issues on intensity normalization, which is dependent on the intrinsic fluorescence or the concentration of dye will have to be addressed, and characterization of UV signal from precipitates need to be established. SONICC images, while powerful for detecting crystals, were found to be unsuitable since precipitates emit no detectable signal.

The next chapter explores an orthogonal readout from sparse-matrix screens, where the focus changes from crystal-driven to protein-sample-stability driven.

*Special thanks to Dr Michael Fairhead, Dr Jolanta Kopec, Dr Ritika Sethi, Fiona Fitzpatrick, Dr Fiona Sorrell, and Hazel Aitkenhead for the protein samples tested (Table 4.4); Dr Romain Talon for supervising data collection at Diamond Light Source, Beamline I-04 1; Dr Michael Fairhead and Dr Jolanta Kopec for processing the data collected, and Beth McLean*

## 4.7 References

Bruno, A. E., Ruby, A. M., Luft, J. R., Grant, T. D., Seetharaman, J., Montelione, G. T., Hunt, J. F., & Snell, E. H. (2014). *PLoS One*. **9**,.

Klei, H. E., Kish, K., Russo, M. F., Michalczyk, S. J., Cahn, M. H., Tredup, J., Chang, C., Khan, J., & Baldwin, E. T. (2011). *Cryst. Growth Des.* **11**, 1143–1151.

Lawrence, N. (2004). *Adv. Neural Inf. Process.*

Luft, J. R., Wolfley, J. R., & Snell, E. H. (2011). *Cryst. Growth Des.* **11**, 651–663.

van der Maate, L., Postma, E., & van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review.

Nagel, R. M., Luft, J. R., & Snell, E. H. (2008). *J. Appl. Crystallogr.* **41**, 1173–1176.

Newman, J. (2005). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61**, 490–493.

Snell, E. H., Nagel, R. M., Wojtaszcyk, A., O'Neill, H., Wolfley, J. L., & Luft, J. R. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1240–1249.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). *Science*. **290**, 2319–2323.

# 5. Using Clear Drops to Identify Alternative Formulation Buffer to Increase Protein Stability

Supersaturation of protein sample has to be achieved prior to crystallization; this typically requires the protein sample to be suitably concentrated. Increased stability of protein in its formulation buffer increases its solubility, thus allowing for higher sample concentration to be achieved without the unfolding of protein, aggregation, or precipitation. Furthermore, higher apparent melting temperature ($T_m$) of protein samples was also found to increase crystallization success rates, indicating the importance of sample stability for successful crystallization (Dupeux et al., 2011). Although the stability and solubility of a protein sample are dependent on its formulation buffer, purification protocols or standard buffers typically dictate what the final formulation buffer is, which may not be optimal for the protein of interest. Various methods have been described to identify alternative buffers, including specially designed solubility and stability screens (Jancarik et al., 2004) and thermoflour methods (Ericsson et al., 2006). These are both time and sample consuming, so uptake, at least at the SGC is apparently low; instead, most experimenters set up sparse-matrix screens in the hope of crystalline behaviour.

Thus, we aim to exploit sparse-matrix screening outcome as a solubility screen, based on the work of Collins *et al.* (2005), to identify chemical components that appear to stabilize specific protein samples, so that they can be used to design alternative buffers for the reformulation of the protein prior to subsequent crystallization experiments. This is done by identifying a very common outcome of crystallization trials: clear drops. The implications behind clear drops are either (1) under-saturation of the protein sample in the mixture solution, or (2) the protein is in the metastable super-saturated state if the conditions are suitable for crystallization (see Chapter 1.2.1). Similar to the strategy of Collins *et al.*, we assume that the clear drops observed are under-saturated protein samples. In their work, Collins and colleagues identified an

alternative buffer from an initial trial of 192 crystallization droplets of a test protein. Only 10 droplets remained clear after a day, and six of these shared a common buffer: 10mM CHES, pH 9.25 to pH 9.5. The protein sample was re-purified with a similar protocol, but was immediately buffer-exchanged after the final purification step (anion-exchange). The new buffer allowed protein concentration to be increased from 8mg/mL to 16mg/mL without precipitation, and subsequently led to crystallization and structure determination.

Despite the elegance and simplicity of the approach, to our knowledge, no further reports on the success and/or extension of the work were published.

The principle underlying the above work, and that of the solubility screens (Jancarik et al., 2004) is that conditions that increase protein stability, and thus its solubility will result in clear drop. Thus the goal is to identify conditions that consistently produce clear drops when other conditions precipitate the protein. This requires the sample to be just below its solubility limit prior to such screening, to exhibit both precipitation and clear drops: no useful information can be gleaned if all drops precipitated or all drops remained clear. This chapter extends the work of Collins *et al.* (2005) by identifying clear drops automatically, using the texton distribution of each droplet as inputs to classifier algorithms. Thanks to the automation, we are further able to estimate the increase, decrease, or consistency of 'clear-ness' across droplets with different protein to precipitant mixing ratio for increased information content beyond binary classification of clear vs not-clear. Whereas Collins *et al.* used the method to crystallize a difficult protein, we present an example where a stabilizing buffer increased the size and diffraction resolution of a crystallisable system.

## 5.1 Methods

### 5.1.1 Clear drop identification

A random forest classifier with 500 decision trees was trained using Matlab's Treebagger function, based on a training set of 2556 images, 968 of which were labelled manually as clear drops, and the remaining 1588 non clear drops. A ten-fold cross validation of our classifier, at a cut-off of 0.5 gave an average area under the receiver operating characteristic (AUROC) curve values of 0.9881 ± 0.0044, indicating highly reliable separation of clear and not-clear populations. Misclassifications can be attributed to clear drops with dirt in the well Figure 5.1(a), or barely visible precipitates Figure 5.1(b). Unlike crystal detection, clear drop classification is binary in nature with little variations; it is thus unsurprisingly the class with the highest agreement rate in human labelling of crystallization images (Buchala & Wilson, 2008). Hence we concluded that automated classification is well suited for the problem, compared to ranking described in Chapter 3.

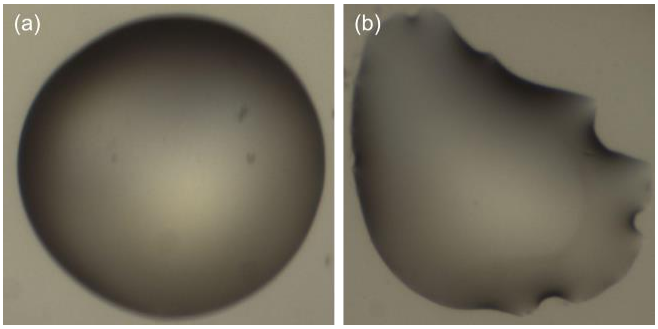



*Figure 5.1: Typical misclassification of clear drops. (a) was classified as unclear due to dirt in the droplet, while (b) was classified as clear even though there is very light precipitate.*

### 5.1.2 Mapping clear drops to chemical components in a screen

With classification at hand and knowledge of conditions behind each well, the goal of our interface design was to highlight correlations of clear drops and chemical component in sparse-matrix screens automatically, and present this information to the experimenter. We treated

each component in a condition separately and independently for simplicity, and assigned these components into the following classes: buffer, anion, cation, polymer, and organic, as well as a 'pH' category. The classes were further divided into sub-categories (for example, PEGs and non-PEGs in the polymer class), and finally each component (e.g. PEG 3350) was broken down to the concentrations sampled, or pH in the case of buffers. An example of how this zooming-in process for magnesium ion is shown in Figure 5.2. The clear drops identified in the crystallization screening experiment were mapped to each of these chemical components, and we visually present the number of clear drops vs. non-clear drops for each component in a given category for each sub-well. If different screen-types were set up, they may be combined across all screens to produce a single analysis.

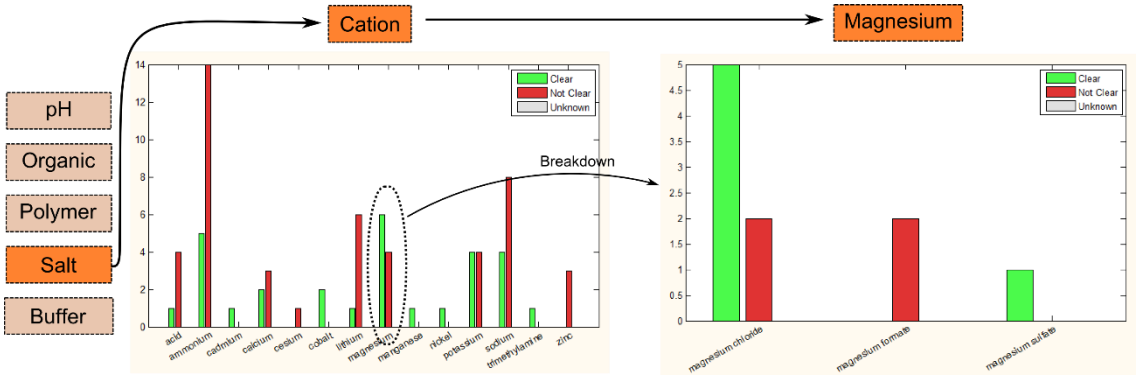



*Figure 5.2: Clear drop analysis of a given component. The user interface shows the comparison of clear vs. non-clear drops at different levels of analysis. The example here shows the process starting from Salt, focusing on cations. All cations sampled will be displayed. The user can further analyse the pairings for each cation. In this case, we show all magnesium salts. Large clear-to-not-clear ratio are immediately observable through the green and red bar heights.*

Furthermore, we can also estimate the trend of 'clearness' across the three sub-wells in our typical screen setup. Sub-wells A to B to C corresponds to decreasing protein concentration and increasing precipitant concentration. With such progression of protein and precipitant

concentrations, it is expected that stabilizing conditions should result in decreasing quantities of precipitate and increasing clear regions in the droplets, since there is less protein in sub-well C compared to sub-well A. Components that are counter to this trend should be avoided because they are effectively precipitating or destabilizing the protein.

The quantity of clear-ness is estimated by the percentage of textons associated with visually clear regions (first 50 entries in in our texton dictionary, visually estimated). Using ±10% of clear textons proportion as a threshold of change, the trend of precipitation across sub-wells can be classified as constantly clear, constantly precipitated, increase in precipitation, decrease in precipitation, kinks (inconsistent increase/decrease), or unknown if there were faulty droplets.

## 5.2 Results

### 5.2.1 DACASAA

While Collins *et al.* (2005) used clear drops to identify stabilizing buffers for a previously uncrystallised protein, we used the output of our analysis to improve the crystal quality of the protein DACASAA. The original formulation buffer of DACASAA was 20mM tris pH 8.0 with 300mM sodium chloride. The maximum concentration achieved was 3.8 mg/ml. The protein nevertheless crystallized, and diffracted to 4 to 3Å. However, high resolution diffraction was required since the protein was intended for fragment soaking experiments to resolve the binding of small fragment molecules. We combined the analysis from the three screens previously set up for DACASAA (JCSG, LFS and HIN) to identify conditions that consistently resulted in clear drops, or decrease in precipitate across sub-wells, factoring the imbalanced sampling of conditions when deciding which components to use. We excluded polymers from the analysis since the primary function of polymers in crystallization conditions are as major precipitants.

The trend across sub-wells are shown in Figure 5.3. The aim was to identify components that resulted in constantly clear (yellow) or a decrease in precipitate (green) across the sub-wells. Higher pH was found to be more favourable (Figure 5.3(a)). While sodium (in the original formulation buffer) did not seem detrimental, we chose magnesium instead for its higher ratio of constant-clear to increasing-precipitate (Figure 5.3(b)). For the same reason, chloride, formate, malonate and tartrate were viable anions (Figure 5.3(c)). Further investigation of the magnesium salts sampled showed that magnesium chloride at 0.2M had 10 wells that were constantly clear, increasing the confidence of its stabilizing effect compared to the lower numbers for magnesium formate (Figure 5.4(a) and Figure 5.4(b)). There were no sampling of magnesium malonate and magnesium tartrate in our screens, and we thus chose not to use these. Finally, for buffer choice, bis-tris-propane was the clear winner (Figure 5.3(d)), and since pH 8.5 was well within its buffering range, our recommended formulation buffer was thus 0.1M bis-tris-propane pH 8.5 with 0.2M magnesium chloride. The experimenter decided to add 5% glycerol as cryoprotectant; the low number of samples Figure 5.4(c) made it difficult to draw conclusions on the effects of glycerol, especially since half of the droplets with glycerol were constantly clear while the other half precipitated.

Analysis of the graphs took no more than ten minutes. Buffer-exchange was carried out from thawed protein sample through three repetitions of dialysis by incubating protein sample with 50 times of volume of the new buffer solution. The protein sample in the new buffer could be concentrated beyond 25mg/mL without precipitating. Crystallization in the same condition was set up, which resulted in larger crystals with increased diffraction resolution of 2Å. Figure 5.5 compares the resulting crystals of the protein in the original and new formulation buffer.

Purification, formulation, buffer-exchange and the modelling of DACASAA structures reported here were all carried out by Dr Helton Wiggers. The clear drop analysis was done together with Dr Helton Wiggers.

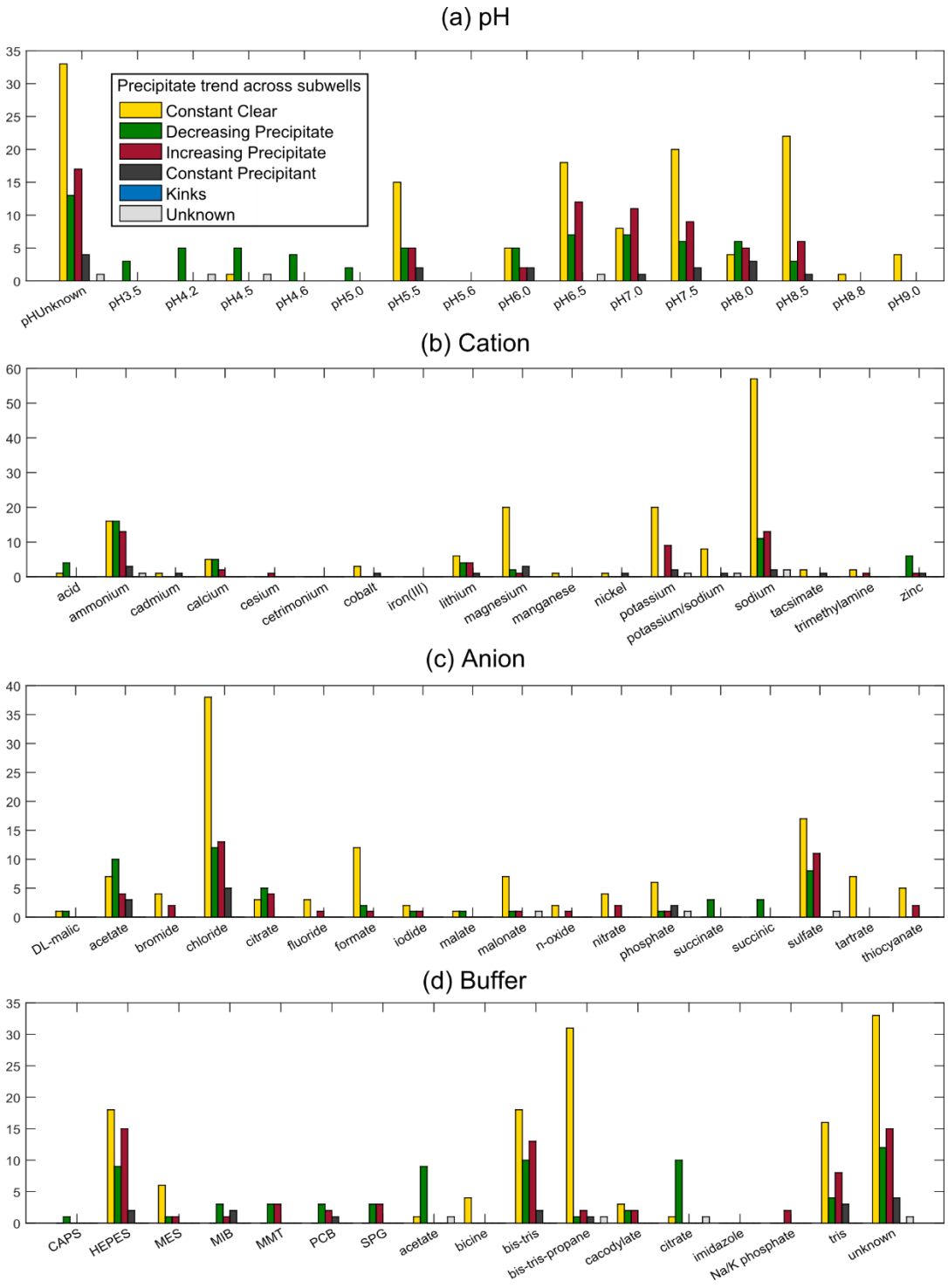



*Figure 5.3: Clear drop analysis for DACASAA from a combination of trends across sub-wells from JCSG, LFS and HIN screens. Each class of component was analysed to identify stabilizing components. From these charts, we chose (a) pH 8.5, (b) magnesium$^{2+}$, (c) chloride$^{-}$, and (d) bis-tris-propane as the buffer. The selection of cation and anion was influenced by the magnesium salts sampled, shown in Figure 5.4.*

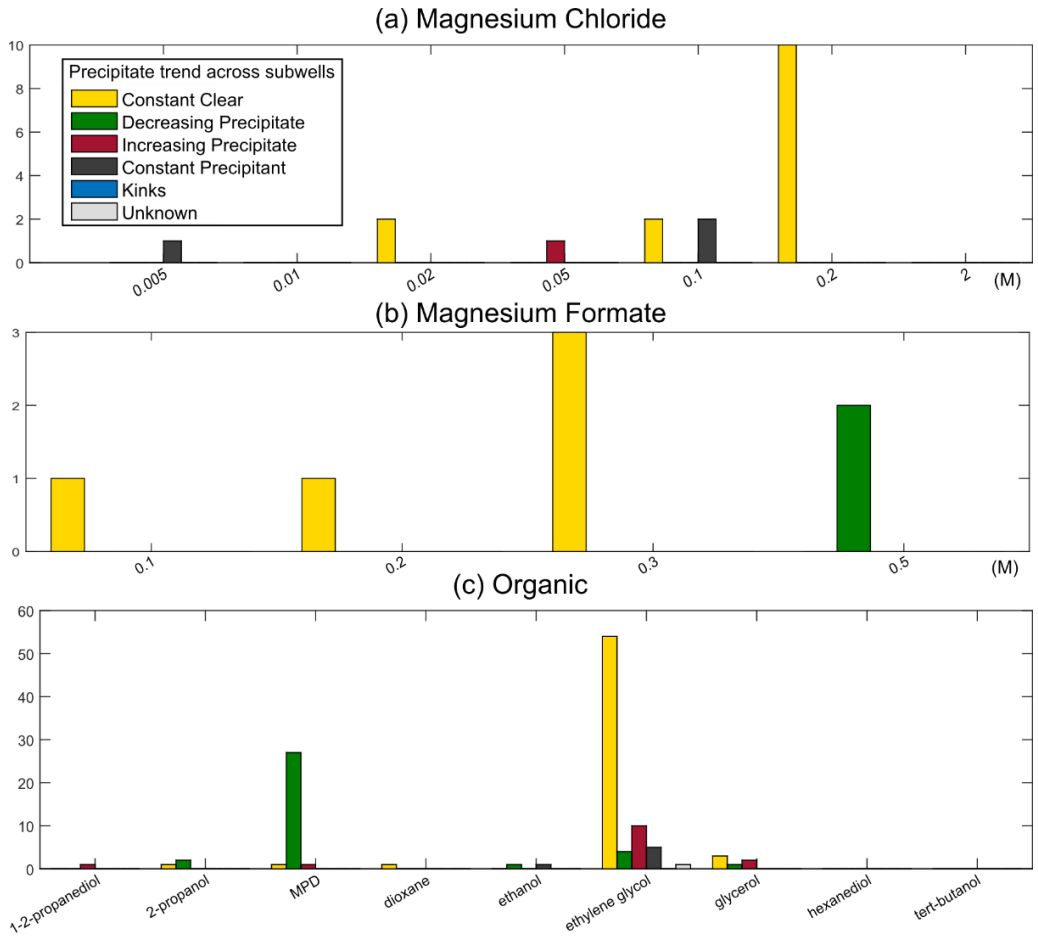



*Figure 5.4: Clear drop analysis for magnesium salts (a and b) and organics (c). The high number of wells with constantly clear drops for magnesium at 0.2M (10) made this the salt of choice, versus magnesium formate (3). (c) The effects of glycerol is indeterminate with the low number of samples, and was added to the recommended buffer as cryo-protectant.*

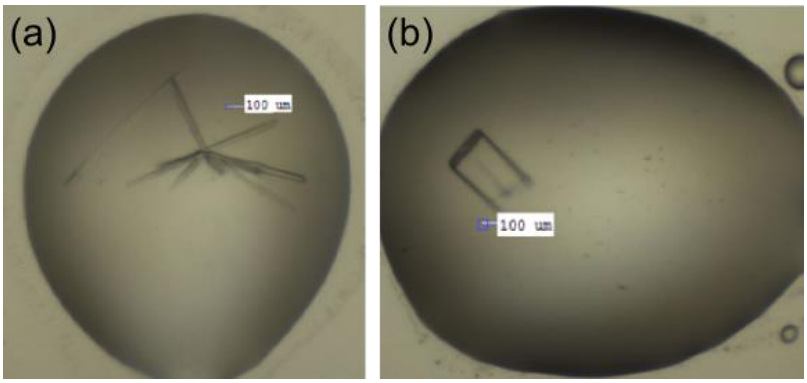



*Figure 5.5: Crystals of DACASAA. (a) Crystals grown from DACASAA in the original purification buffer (20mM tris pH8.0, 0.3M sodium chloride), concentrated to 3.8mg/ml. (b) Crystal after buffer exchange to 0.1M bis-tris-propane pH 8.5, 0.2M magnesium chloride and 5% glycerol, concentrated to 25mg/ml.*

## 5.3 Discussion and concluding remarks

They key factor in crystallization has always been the protein sample; a stable and concentrated sample increases crystallization success. Here, we have extended the work of Collins *et al* (2005) by automating the process of classifying droplets as clear or non-clear using ensemble classifiers, and mapping clear drops to the chemical components sampled in a sparse-matrix screen or combination of screens. This exploits the outcome of screening experiments that have already been set up, requiring no additional protein sample. The case presented by Collins was where precipitate predominate; our tool allows for more subtle trends to be observed, since there were a high number of clear drops in the screens of DACASAA. We have also shown that the method is able to improve the crystal quality for a crystallizing system by allowing higher concentrations to be achieved.

While the analysis may be carried out at no additional cost, the barriers to its practical adoption include (1) sample limitation, especially if the experimenter prioritises other strategies, for example screening with other screen-types, or changing experiment format; and (2) trends cannot be identified in the clear drop analysis, when the plate is generally clear, or if very few to no clear drops were produced with nothing in common.

Implementing automated and objective suggestions of stabilizing components, *e.g.* based on the number of clear drops observed proved to be non-trivial, especially due to the imbalanced sampling of conditions in sparse-matrix screens. For example, centrimonium bromide is only sampled once, but sodium chloride appears in 31 conditions, at concentrations between 0.1 M and 4.2 M across the four supported screens. Hence, we judged it more productive to build an interface for exploration, and leave the judgement of which components to use to the experimenter, whose prior knowledge on the compatibility of the potential components and the protein may also be very useful. In any case, black-box decision generally face adoption problems with users as suspicious as crystallographers.

Further work can be carried out to separate metastable from undersaturated clear drops, not only for more accurate suggestions of stabilizing components, but also to determine or suggest conditions where seeding may be effective. This will require a thorough understanding of condition similarity, and the guidelines given by Luft et al. (2011) on differentiating between undersaturated and metastable clear drops should serve as a good starting point.

The next chapter addresses an issue intentionally left out since Chapter 3: the presentation of results to experimenters, and making algorithm output accessible and practically useful in analysing crystallization experiment outcome and designing subsequent experiments. We will introduce TeXRank, an image viewer developed to incorporate the ranking of droplets, analysis of precipitation patterns across sparse-matrix screens, and the clear drop analysis presented in Chapters 3, 4 and this chapter respectively.

*Special thanks to Dr Helton Wiggers for re-formulating, crystallizing, and processing the datasets for DACASAA reported in this chapter.*

## 5.4 References

Buchala, S. & Wilson, J. C. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **D64**, 823–833.

Collins, B., Stevens, R. C., & Page, R. (2005). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **61**, 1035–1038.

Dupeux, F., Röwer, M., Seroul, G., Blot, D., & Márquez, J. a. (2011). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 915–919.

Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N., & Nordlund, P. (2006). *Anal. Biochem.* **357**, 289–298.

Jancarik, J., Pufan, R., Hong, C., Kim, S. H., & Kim, R. (2004). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 1670–1673.

Luft, J. R., Wolfley, J. R., & Snell, E. H. (2011). *Cryst. Growth Des.* **11**, 651–663.

# 6. TeXRank: Effective Presentation and Deployment of Algorithm Output

From the point of view of experimenters, accessibility and presentation of results are as important as the analyses themselves for practical use and impact. This is an aspect that has not been explored much in the existing literature on computer vision and data analysis for crystallization, be it the computed classifications of crystallization droplets (Chapter 1.2.2), or for that matter, the features obtained from alternative imaging techniques (SONICC or UV) (Ng et al., 2014). Vendor software are essentially image viewers displaying droplet images and its associated information (protein sample, precipitant, image properties *etc.*), with additional features to allow for the labelling of droplets and storing these annotations in databases for future references. However, being proprietary software, there is limited (if any) flexibility to introduce new features or customisation to incorporate the analyses described in the preceding chapters into existing vendor software. It therefore quickly became unavoidable to develop a custom software to centralise all algorithm output, and provide an image viewer to enable the most fundamental tasks in analysing crystallization experiment data.

TeXRank was first written as an image viewer that displays droplets of a plate in the ranked order described in Chapter 3, but was found to be a great vehicle for additional functionality, especially for the presentation of results from the precipitation pattern and clear drop analyses discussed in Chapters 4 and 5 respectively. Although developed in Matlab as a prototype, TeXRank has since been compiled and deployed at both the SGC, NIBR, Basel and Diamond Light Source as standalone executables, requiring only the freely available Matlab Compiler Runtime (MCR). The following sections describe the development and features of TeXRank, which has also become a cornerstone of the medium throughput fragment screening pipeline at Diamond. Figure 6.1 shows the relationship between TeXRank and the image analysis pipeline described

in Chapter 2, as well as the output of the ranking, clear drop identification, and precipitation pattern matching algorithms. These are all carried out by a separate executable which analyses new images as they are captured, and does not involve any user-interactions.

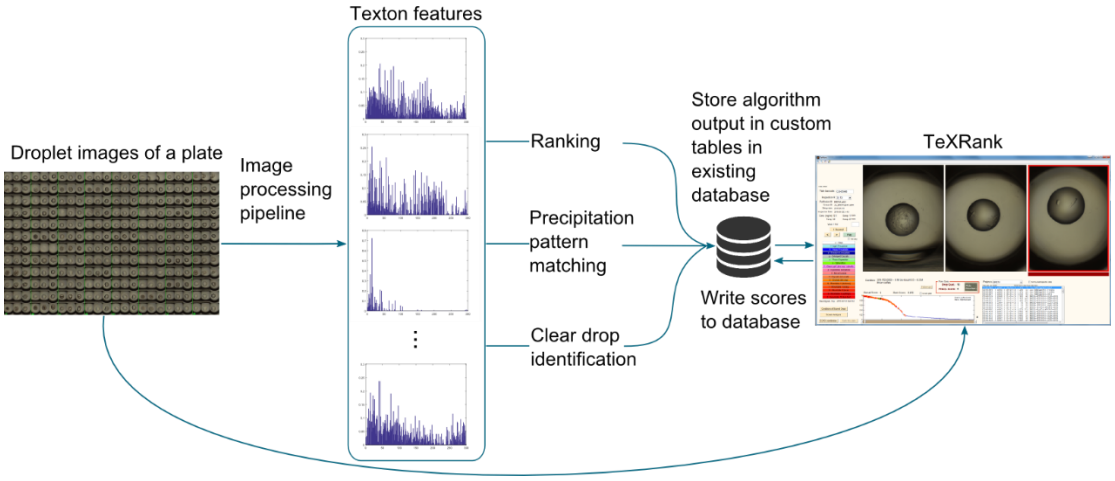



*Figure 6.1: Schematics of the relationship between the image processing pipeline, algorithm output for ranking, precipitation pattern, and clear drop analyses, as well as TeXRank. TeXRank reads the output of the analyses stored in custom tables in existing database, and presents the images accordingly. Labels given by users also recorded in the database.*

## 6.1 Presentation of ranked images

For the presentation of manual or automated scoring of droplet images, the common approach in vendor software is to label the images with tags and different colour schemes (Mele et al., 2014), and sophistication is introduced to allow users to sort by labels or hide subsets of data. Nevertheless, images are still arranged based on their physical location on the crystallization plate, which is effectively random and intuitively not the most efficient way of viewing images (Ng et al., 2014).

It was the lack of support in vendor software at both the SGC and NIBR for the rearranging of image order, fundamental to our approach in ranking droplet images, and for integrating the

informative profile of ranking scores (Chapter 3.5), that triggered the development of TeXRank. TeXRank was designed to link images with the ranking scores and score profiles, applying the algorithm output directly, instead of existing merely as a separate collection of numbers. It is integrated with the existing database and data-storage infrastructure (CrystalTrak (Rigaku) and Scarab, SGC's in-house database system, and Rockmaker database (Formulatrix) at NIBR) to mirror the existing image annotation or labelling system and experiment information display for familiarity. A screenshot of TeXRank is shown in Figure 6.2.

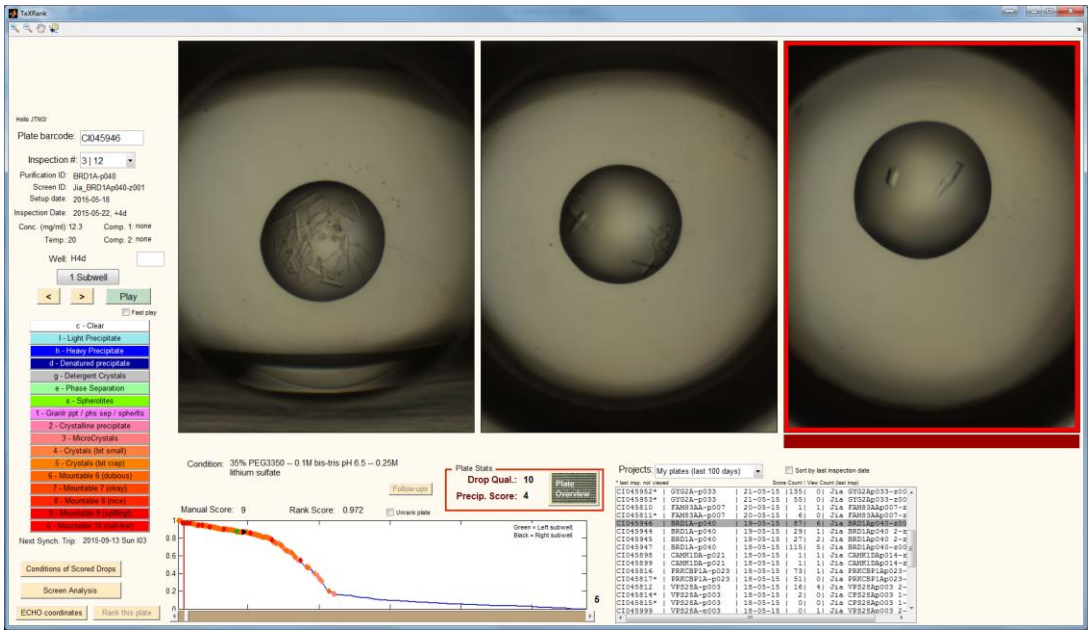



*Figure 6.2: Snapshot of TeXRank, the custom viewer written to display images of a given plate in the ranked order. The plot at the bottom shows the ranking scores of droplets, giving a quick overview of the plate. We also display three sub-wells of a well together, with each typically at 2:1, 1:1 and 1:2 protein to precipitant mixing ratio from left to right. The image to be scored is highlighted with a red bar at the bottom, and the scoring system and experiment information mirror the software that users are accustomed to.*

Various features of known usefulness have also been added to ensure that the software is fully functional for routine use, and being an in-house software, customisation could be carried out easily to best suit the local workflow. These features include:

1) The option of viewing sub-wells of a well side by side: this is very informative for SGC's typical set-up of variations in mixing ratio, where the juxtaposition allows the immediate evaluation of precipitation trends across the different protein concentrations;

2) Plate thumbnail overview with colour-coded labels corresponding to manual scores for quick overview of scoring trend (Figure 6.3);

3) Tools for measuring crystal size on high-resolution images (Figure 6.4);

4) Various navigation methods from sub-well to sub-well (click on the score profile, text field to jump to a desired sub-well, or graphical selection from the plate thumbnail);

5) Plate-to-plate navigation by project or barcode, with a default display upon login, of all plates set up by the user in the last 100 days for convenience. Any inspections not viewed are also highlighted as a reminder;

6) Display of basic calculated plate statistics, including Drop Quality (0 = all faulty droplets, 10 = all good drops) and Precipitate Score (0 = all drops were clear, 10 = all drops precipitated);

7) Typical keyboard shortcuts for scoring and sub-well to sub-well navigation.

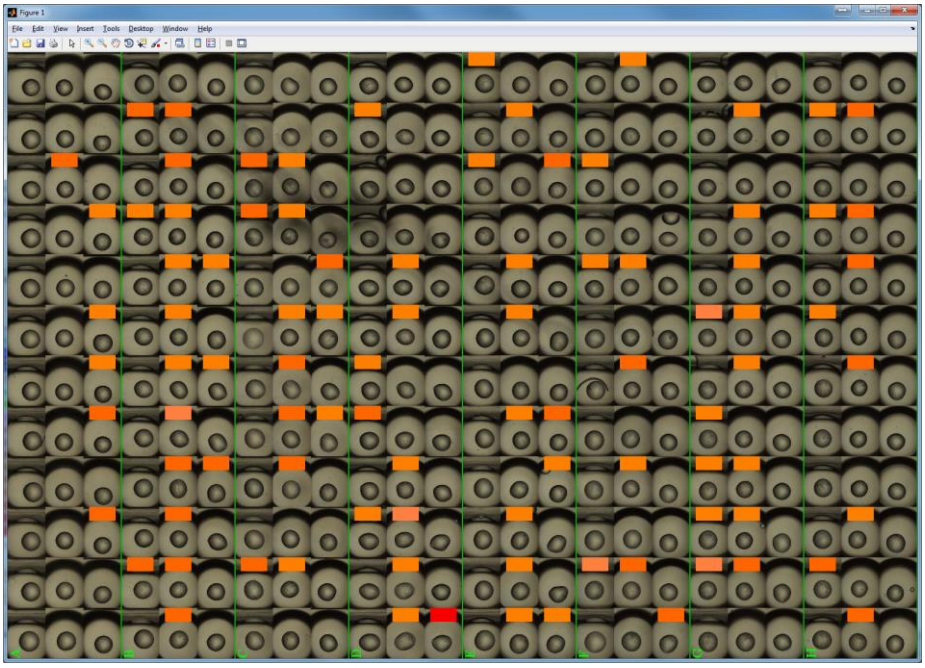


*Figure 6.3: Plate thumbnail, with coloured boxes corresponding to manual scores for the sub-well. Users can also navigate to a desired sub-well by double clicking a sub-well on the thumbnail.*

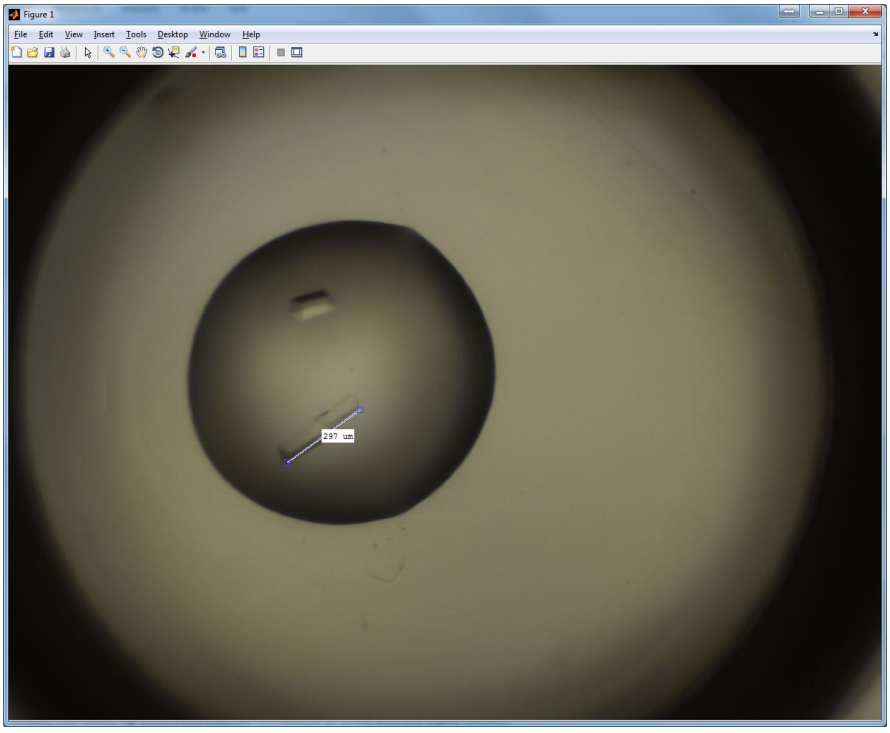



*Figure 6.4: High resolution of sub-well images are accessible by double clicking on an image in TeXRank, with a default ruler which can be moved and stretched to measure crystal sizes.*

## 6.2 Deployment of TeXRank

Being imager-independent, TeXRank works with both Rigaku's MinstrelHT system and Formulatrix's Rock Imager system. Full integration of TeXRank with existing database and imaging systems will invariably require bespoke effort. Due to the different database systems and architecture (Oracle for the Minstrel HT system vs Microsoft SQL Server for the Formulatrix system), changes in queries were required to fit the infrastructure of both sites. However, we have shown that integration is possible for both systems, and it could take as little as 3 days to integrate fully for someone familiar with either system and with general database knowledge (Ng et al., 2014).

Prior to the general release at the SGC, introduction of a pre-release version of TeXRank was done with the volunteers for the exercise described in Chapter 3.1.5, where we compared the annotations from 10 crystallographers who viewed images the ranked or unranked order. This exercise was not only useful for gathering the data for Chapter 3.2.4, but crucial for the development and uptake of TeXRank in the following ways:

1) Major bugs were identified and addressed prior to the general release, resulting in better first impressions when other members of the SGC started using TeXRank;

2) Crucial features for routine use were identified, including those discussed in the last section;

3) Tutorials were given to the volunteer crystallographers prior to this exercise, who then passed their knowledge to other members of SGC after the general release;

4) Personal recommendation of TeXRank by these volunteer crystallographers to other members of their group.

Since its deployment in June 2014 at the SGC, TeXRank has been well-received, being the default image viewer for most experimenters. A handful still do prefer vendor software, presumably out of habit and familiarity. Users have continued to contribute with further feature requests and bug reports, improving TeXRank and its usability. It is now a noticeable issue if and when the image processing pipeline and ranking algorithm fail to run (*e.g.* power shut down, problems with the database or machines analysing the images), indicating users' reliance and preference for ranked images. In such instances, we have also enabled users to rank specific plates with a click of a button using the computing resources of their local machine.

TeXRank runs on both Windows 7 and Windows 8.1 machines. For updating and maintenance purposes, a desktop shortcut is installed on each experimenter's computer, which all point to the same executable on the network drive. At every run, TeXRank first copies itself to a local folder (C:\TeXRank) before launching the viewer. This means updates and bug fixes need only to be applied to the network drive version, and will be automatically rolled out to users when they launch the software.

The current version of TeXRank has also been deployed similarly at NIBR, Basel, since February 2015 with some minor differences to fit the workflow and database structure. Uptake was found to be lower, most likely since scientists at NIBR use TexRank in parallel with RockMaker, the vendor software from Formulatrix, that has various viewing and analysis features and which screen design and formulation options have been fully integrated in the local workflows. In contrast, the competing vendor software at the SGC (CrystalTrak) is used solely as an image viewer, where TeXRank was better positioned to be a direct replacement. Further integration with the workflow at NIBR will be necessary to increase uptake, for example by scripting the launch of RockMaker to also start TeXRank, or allow better communication between both software, where a button on TeXRank will direct the user to RockMaker and vice versa, depending on the limitations of RockMaker.

## 6.3 Precipitation pattern analysis and automatic screen design

With the development of the precipitation pattern analysis described in Chapter 4, TeXRank became the natural deployment route for this new algorithm output. The nearest-plate profile and suggested conditions for JCSG, LFS, HCS and HIN plates can easily be accessed from a button on TeXRank (Figure 6.5).

In the Screen Analysis window, the user may select to generate the nearest-plate profile from any inspection that have been analysed, although the inspection corresponding to images captured after 4 days is recommended, since it was this inspection that was used to form the PFLs. The list of successful conditions from the selected number of nearest-plates is displayed (default = 10 nearest plates), along with the nearest-plates curve. If a base-condition has been sampled for the experiment, regardless of screen-type, a quick visual check for crystallinity in the well can be done with just one click; TeXRank automatically retrieves the specific sub-well images of the plate and displays it.

At the SGC, users may also select up to six of the suggested base conditions from this list, or any condition represented by the nodes on the conditions dendrogram (bottom panel in Figure 6.5) to be followed up. 4-by-4 grids for the selected base conditions are automatically generated using pre-defined grids to populate a typical 96-well plate. These pre-defined grids were designed based on the guidelines in Figure 4.10. TeXRank generates the corresponding worklist required by the liquid handler (MultiPROBE II plus HT/EX Robotic Liquid Handling System (Perkin Elmer, www.perkinelmer.co.uk)) to formulate the screen, and writes the necessary information into the database. Using this function to design a standard follow-up screen based on our suggested conditions, the experimenter thus bypasses the in-house screen designer, which is a separate Excel-based software, and is only required to name the new screen using the form shown in Figure 6.6; this translates to a savings of up to 30 minutes per screen design, including the CPU time require by the in-house screen designer.

This present link of TeXRank with the liquid handler is confined to the workflow and robots at the SGC. Extending this beyond the SGC will require robot-specific worklists to be generated for each 4-by-4 grid based on standardized stock solutions, which while tedious, would be a one-off effort. Furthermore, a screen designer may be written as part of TeXRank to allow non-standard follow-up screens or modifications to those generated automatically. This will also require a database of standardized and commercially available stock solution for general use across labs.

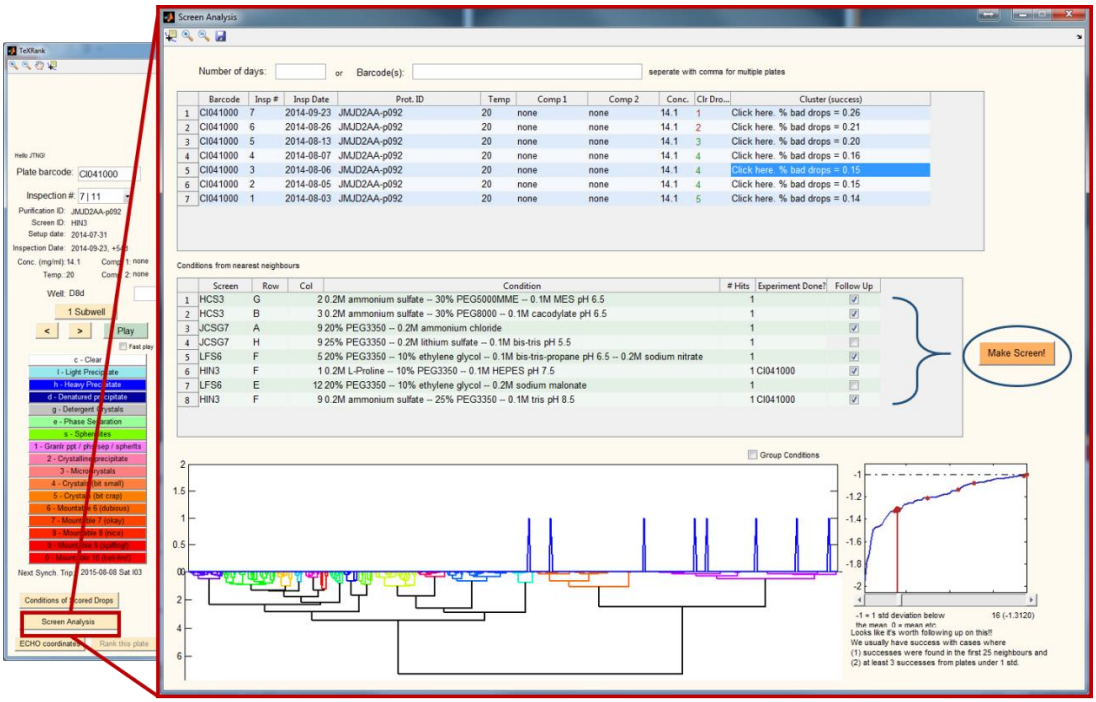



*Figure 6.5: Screenshot of the Screen Analysis window, accessible through a button in TeXRank (red outline). The Screen Analysis window allows users to select the inspection to generate the nearest-plate profile, and presents potential crystallization conditions based on the number of nearest-plates to include, as described in Figure 4.5. These conditions can be selected using the checkboxes (blue curly bracket), and a Make Screen button (blue circle) will automatically design a follow-up screen using predefine 4-by-4 grids for the selected conditions, but requires the user to name the screen through the form shown in Figure 6.6.*

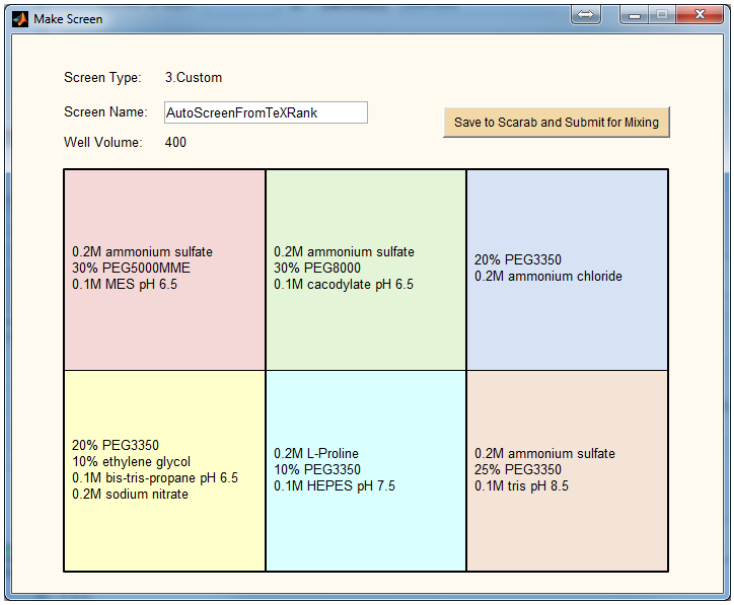



*Figure 6.6: The Make Screen window, for users to name the screen designed. A record of the screen will written to Scarab (SGC database), along with the conditions for each of the 96 wells. The figure in this window can also be saved for future references, and the worklist (csv file) for the liquid handler is automatically generated and saved in specific directories readable by the liquid handler.*

## 6.4 Presentation of clear drop analysis

The mapping of clear drops to chemical components in a screen or across multiple screens as presented in Chapter 5 can be accessed from the Screen Analysis window, as shown in Figure 6.7. In this window, users can either view the break-down of precipitation trend (increase or decrease of clear proportions across sub-wells) for each chemical component (Figure 6.7), or a more simplistic "clear vs. not clear" histogram (Figure 6.8). Visual validation of trend is immediate by evaluating the images displayed at the bottom panel, which are accessible through the plate layout plot at the top right corner. This plot is also colour-coded similarly to the bar charts, with each square's dimensions proportional to the magnitude of change. Analysis across multiple plates is also supported, where users may select up to four plates of supported screen-types for a combined analysis.

Based on user request at NIBR, we have also added a feature for time-change, where a selected inspection is compared to corresponding droplets of the first set of images captured. This is based on the observation that change in droplet content are often observed for crystallisable systems, or at the very least the appearance or growth of crystals; Mele *et al.* (2014) also used this principle to flag images for viewing. The Hellinger distance of histograms from corresponding droplets are calculated and used to define if change occurred. We have set a cut-off of $d_H > 0.5$ to indicate change. The lower imaging load and default imaging schedule at NIBR allows for reliable images captured at t = 0, which we expect to have most difference when compared to later inspections. In contrast, the earliest set of images at the SGC for a plate is captured sometime between 12 to 36 hours of incubation, depending on the imaging load. The concurrent entry of 20 plates is not uncommon, and with ~10 minutes to image a plate, the first and last droplet would be imaged 200 minutes apart. Furthermore, the imager may be down due robotic errors or communication errors with the database, which causes further backlogs in the imaging schedule. We thus expect this feature to have limited use at the SGC, but may be more effective at NIBR.

## 6.5 TeXRank in medium throughput fragment soaking protocols.

An orthogonal application of TeXRank was found in fragment soaking protocols with the ECHO liquid dispenser (Labcyte). Historically, fragment soaking at the SGC involves manual pipetting of soaking compounds onto droplets containing crystals. In an attempt to automate soaking to increase throughput, the Mosquito (TTP Labtech) was used to transfer soaking compounds but required all droplets to contain crystals so that no compounds were missed or wasted; thus, where crystals failed to grow, they were transferred manually from other plates.

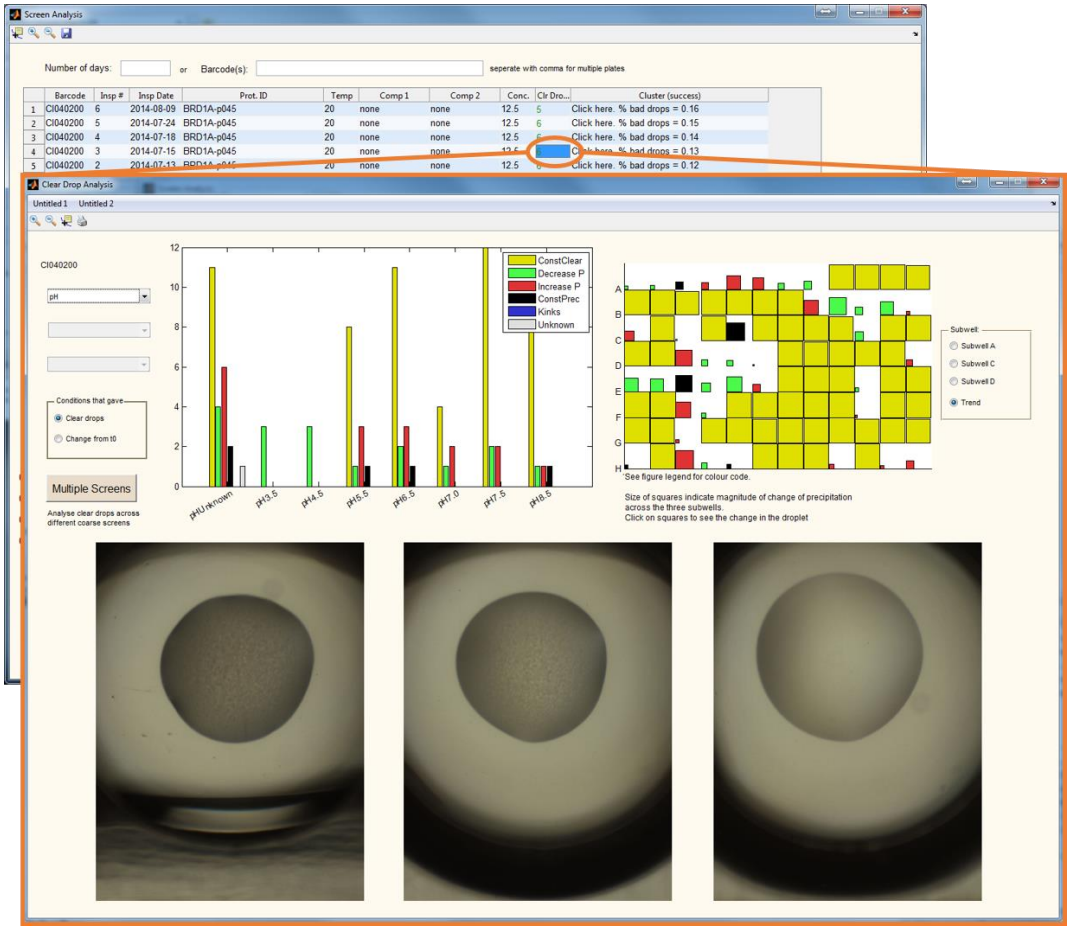



*Figure 6.7: Screen shot of the Clear Drop Analysis window, accessible from the Screen Analysis window of TeXRank. Users select the class of components they would like to analyse, and the bar chart shows the number of clear/not-clear drops for selected sub-wells, or the trend observed across all three sub-wells (constant clear, decrease in precipitate, increase in precipitate, constant precipitation, kinks, or unknown). The plate-layout plot on the top right corner is colour-coded similarly to the bar-chart, and allows immediate visual validation (by clicking on a selected well, images of the three sub-wells shown in the bottom panel) for increased confidence.*

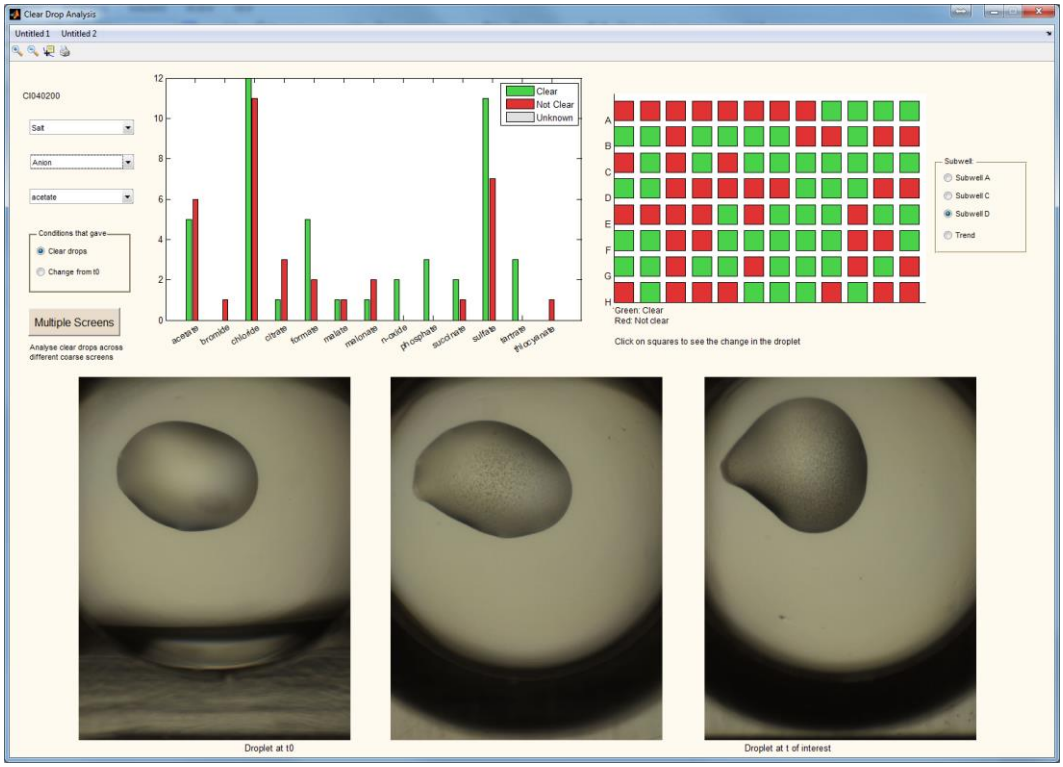



*Figure 6.8: The more simplistic view of "clear vs. not clear" for a selected sub-well. This analysis treats sub-wells independently.*

The power of the ECHO liquid dispenser is that it supports cherry-picking of source-wells and destination wells, removing the need for the laborious transfer of crystals; TeXRank was where we picked the cherries, by allowing quick identification of crystal-containing droplets. Furthermore, the ranking score profile was also found to be an extremely useful indication of whether all droplets should be evaluated. The profile of a typical plate for soaking experiments, without perfect crystal reproducibility is shown in Figure 6.9(a), where a fall-off in ranking score is normally observed. Because the conditions in such plates are narrow (usually one condition with minimal variations in concentrations), the crystallization outcome is typically limited to droplets with crystals (high ranking score) or clear drops (low ranking score). This contrast of scores informs if continued evaluation of droplets is necessary; in the example in Figure 6.9(a), droplets beyond those ranked 153 need not be evaluated.

An issue with using the ECHO for soaking experiments was the potential damages to the protein crystal from high local concentration of solvent, as well as the mechanical force from the impact of nano-droplets as they are dispense by the ECHO. To avoid such damages, the x- and y-coordinates with respect to the centre of the sub-well and away from the crystal(s) can be defined for each sub-well for accurate spatial deposition of compounds. However, the average time to estimate these coordinates by eye was approximately 2 hours for 288 droplets. A new feature was added to TeXRank to allow users to select a specific point (ctrl + left click) on the image where the nano-droplets should be deposited. TeXRank converts this image-space coordinate into the x- and y-coordinates with respect to the well centre (Figure 6.9(b)), because of the pre-processing it does for locating and masking the well-frame. This takes less than a second per image, or under 5 minutes for 288 droplets. For cocktail-soaking, users may select a point in the droplet (shift + left click) to generate a user-determined number of points along the droplet edge, furthest away from the selected point within a fixed spanning angle ($120^{o}$ by default). These points are converted to real-space x- and y- coordinates required by ECHO to dispense cocktails (Figure 6.9(c)).

An offline and database independent version (TeXRankO) with all major functionalities of TeXRank has been developed, and is currently actively used in Diamond Light Source, especially for its ECHO targeting functionality. TeXRank has been used for thousands of soaking experiments to date, including those from external collaborators, and with images captured with other imaging systems.

The code for TeXRankO is available from https://github.com/thesgc/TeXRank.

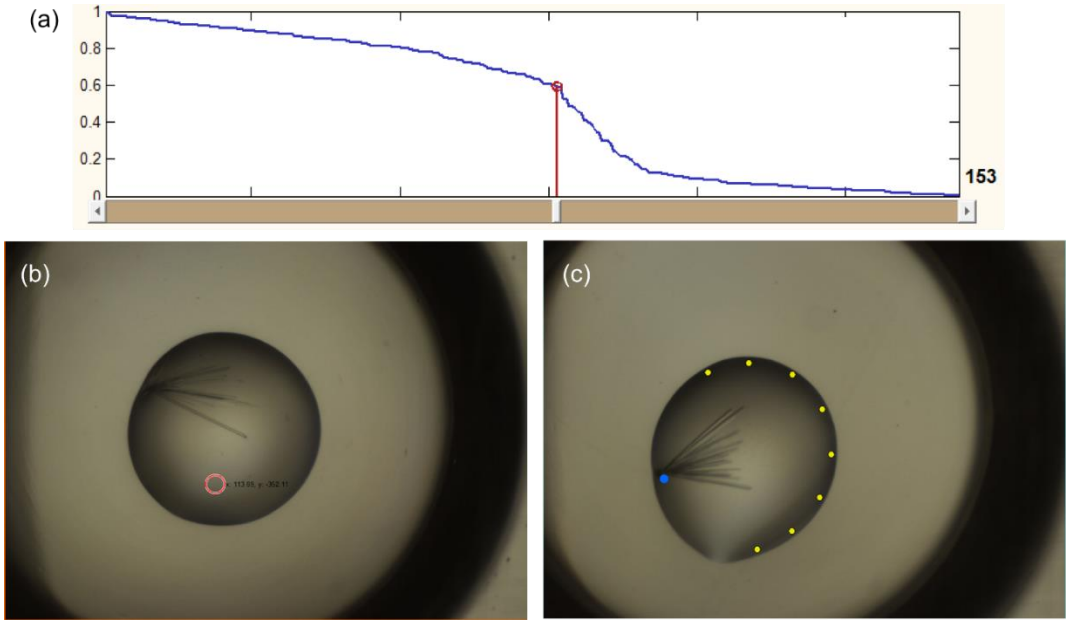



*Figure 6.9: Coupling TeXRank with the ECHO liquid dispenser for fragment soaking experiments. (a) The ranking-score profile of a typical plate for soaking experiments has a sharp drop in ranking scores due to contrasting scores of droplets with crystals and clear drops, allowing experimenters to target only droplets with crystals and ignore the rest. Using TeXRank with the ECHO liquid dispenser for soaking experiments: (b) single-shot mode, where users select where (red circle) the ECHO should deposit compounds, and (c) multi-shot mode, where users select a point (blue dot in this example), and TeXRank identifies a pre-determined number of locations (yellow dots) along the droplet boundary, furthest away from the selected point within a fixed spanning angle for the deposition of different compounds in a cocktail. 8 components and a spanning angle of 120$^o$ were selected in this example.*

## 6.6 Further work

TeXRank was written in Matlab, a language meant for quick prototyping instead of writing optimized and stable software or application. It has however, been surprisingly robust and functional at present. For scalability and future maintenance, TeXRank may need to be converted into more suitable and faster languages like C# or java. A web version of TeXRank will be useful for crystallization facilities that serve proxy users, without requiring any local installations of the software.

TeXRank as an image viewer will undoubtedly face competition from vendor software, which have the added advantage of being integrated with the imager itself for hardware controls like manipulating the imager (focal depths, lighting, imaging schedule), or other systems also supplied by the vendor. Thus, for long term sustainability of the algorithms and ideas introduced, incorporation with existing vendor software would seem most sensible; the synergy of the capabilities and established user-base of vendor software, with the analyses and proven advantages of TeXRank should make it a worthwhile effort.

## 6.7 Concluding remarks

TeXRank has been crucial in making our algorithms practically useful and accessible to experimenters. As an in-house developed software, the full flexibility allows trivial addition of features or modifications to fit the local workflow, limited only by our expertise as user interface developers. While its development was initially feared to be a time sink, TeXRank proved to be an excellent investment, at the very least for increasing efficiency in the workflow of the SGC, NIBR and Diamond, and providing a vehicle for the deployment of future algorithms.

This chapter concludes our proposed methods for the exploitation of sparse-matrix screening outcome. We have shown that with a handle on the characterization of crystallization experiment outcome, a screening experiment can be much more than a crystal-generator exercise with high failure-rate. We have also introduced TeXRank to centrally present the results of all analyses to experimenters, making all readout accessible. In the next chapter, we switch the focus from the analysis to the execution of screening experiments, by conducting a retrospective analysis on the crystallization screening practices at the SGC, Oxford to identify other factors in its setup that will increase success rates.

## 6.8 References

Mele, K., Lekamge, B. M. T., Fazio, V. J., & Newman, J. (2014). *Cryst. Growth Des.* **14**, 261–269.

Ng, J. T., Dekker, C., Kroemer, M., Osborne, M., & von Delft, F. (2014). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 2702–2718.

# 7. Lessons from 10 Years of Crystallization at the SGC

The focus of the previous chapters have been on the analysis of crystallization screening experiments, which first require the screens to be set up. Keeping in line with the theme of learning from historical data, but shifting focus instead to the execution of screening experiments, we present some observations from retrospective analyses of the crystallization efforts at the SGC, Oxford over its first 10 years, to draw guidelines and evaluate the effectiveness of various strategies employed at the organization.

The SGC Oxford deposited its first structure in the Protein Data Bank (PDB) on 20 April 2004 (PDB ID: 1T2A). At the cut-off point for this analysis (31 March 2014), SGC Oxford has contributed 700 deposited structures determined through X-ray crystallography, 462 of which were novel at their time of release. The methods and practices developed over the years for solving these structures have been outlined in past SGC publications, including the strategy for cloning, expression and purification methods used in the lab (Savitsky et al., 2010), as well as the techniques and procedures for data collection at synchrotron sources (Krojer et al., 2013). This chapter bridges these publications by analysing the effectiveness of the nano-droplet crystallization strategies employed at the SGC.

As a high-throughput structural genomics centre, robotics and automation play crucial roles in the crystallization facility. To ensure that crystallization was not a procedural bottleneck, the crystallization infrastructure was configured, at considerable effort and expense, to be accessible to anyone in the organization, with an explicit goal being that use should pre-suppose only minimal training in protein crystallization. Over the 10-year period covered for our analyses, more than 200 scientists have used the facility, generating over 60,000 crystallization plates, from 2339 targets and over 24,000 protein preps. Standard crystallization practice was

guided largely by ensuring the recommended experiment was the most convenient, e.g. predefined liquid handling protocols, or setting up defaults in the software and databases. Standard crystallization practice was implemented largely by ensuring the recommended experiment was also the most convenient, including: ensuring that primary sparse-matrix screens and plate types were always adequately stocked; predefining liquid handling protocols with obvious names; and setting up relevant defaults in the software and databases.

The discussions in this chapter is limited to the first phase of crystallization, namely the coarse screen, *i.e.* screening of conditions for crystallization (McPherson & Gavira, 2014). Shaw Stewart and Mueller-Dieckmann (2014) highlight that there are four main parameters for initial screening experiments: (1) ratio of sample to crystallization cocktail; (2) incubation temperature; (3) number and choice of initial screens; (4) experiment format. Our data and analyses inform the first three, but provide no insight into the fourth, because sitting-drop vapour diffusion was the only format compatible with high throughput in our particular hardware configuration.

We reviewed different aspects of the crystallization strategies employed and report the conclusions from these, particularly focusing on (1) repeating screening conditions with different protein to precipitant mixing ratio, (2) repeating screening conditions at different incubation temperatures (4°C and 20°C), (3) the choice of sparse-matrix screens, (4) protein sample concentration for crystallization, and (5) storage time of crystallization plates. Given the retrospective nature of the analysis, all observations pertain to the strategies actually used to accumulate these data, and say little about orthogonal strategies. Nevertheless, the conclusions drawn should be helpful whenever setting up crystallization screening experiments.

## 7.1 Data set

The SGC has focused on human proteins from multiple families. Figure 7.1(a) shows the families of deposited novel structures over the period analysed, along with the distribution of mass (Figure 7.1(b)) and construct length (Figure 7.1(c)). Protein classes not explored at the SGC, and hence not informed by this analysis, include viruses and large complexes. A few protein-DNA or protein-RNA complexes are present, as are a very small number of integral membrane proteins; however, in view of their negligible representation, these classes too cannot be considered to be represented for practical purposes. The structures deposited comprise of apo structures, protein-ligand structures and protein-protein complexes (Figure 7.1(d)).

Preparation of protein samples was performed by almost 200 scientists, and purification protocols were varied as deemed necessary, depending on sample, at the discretion of the respective scientist. However, the dominant strategy was affinity chromatography, primarily by His-tag/Nickel but also Glutathione S-Transferase tag/Glutathione, followed by size-exclusion chromatography; a final ion exchange chromatography step was occasionally used. The typical aim is to concentrate the protein to 10mg/mL or higher (dependent on protein stability, analysed subjectively); Millipore centrifugal parallel membrane concentrators are the standard approach. The distribution of concentrations of proteins for crystallization is shown in Figure 7.1(e), where we find that majority of experiments, including those that led to PDB depositions were concentrated to < 20 mg/ml.

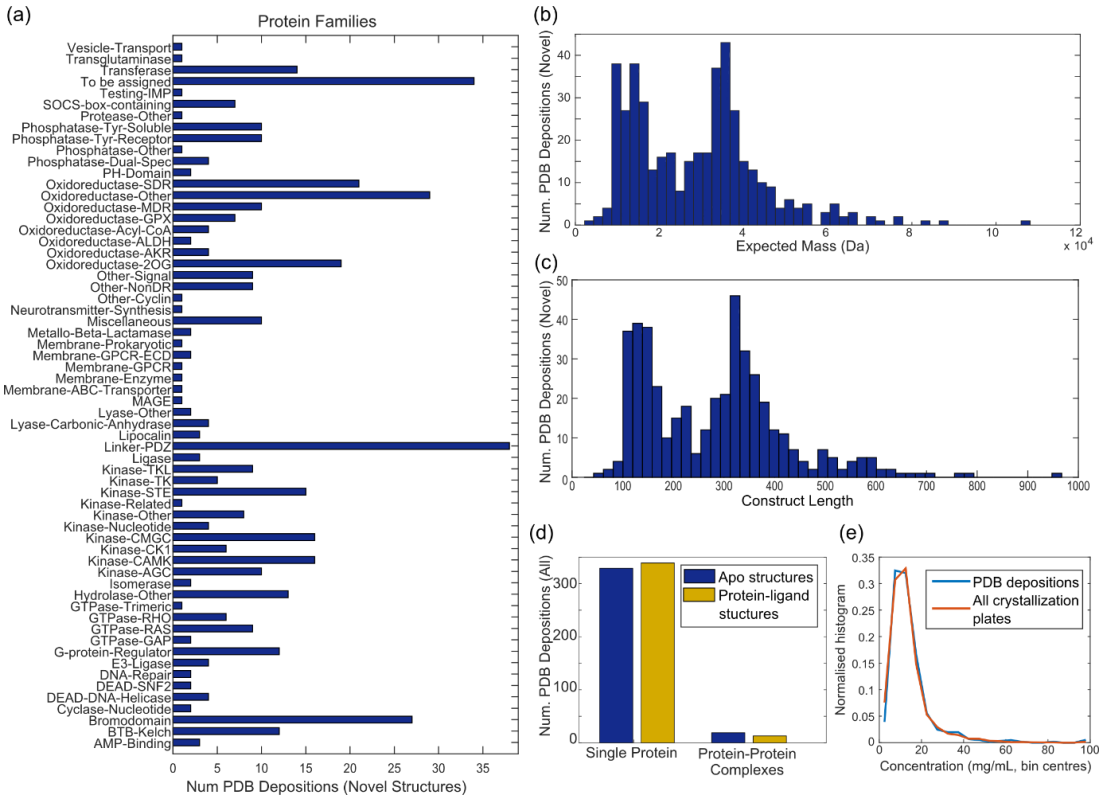



*Figure 7.1: Summary of novel structures deposited by the SGC, Oxford over 10 years. (a) 61 protein families are represented, as annotated a priori bioinformatically. (b,c) Masses range between 2.5kDa and 108 kDA, with chains up to 967 amino acids in length. (d) The majority of structures were single protein structures, either apo or ligand-bound; a small proportion were protein-protein complexes. (e) Comparison of protein concentrations in crystallization experiments for those that led to PDB depositions (blue) and all crystallization plates (red). Both populations have very similar distribution.*

As described in Chapter 2.1 and Table 2.1, standard sparse-matrix screens are used for coarse screening of protein targets. All screens were mixed in-house with the MultiPROBE II plus HT/EX Robotic Liquid Handling System (Perkin Elmer, www.perkinelmer.co.uk) until 2013, and since then have been commercially sourced from Molecular Dimensions (www.moleculardimensions.com). Over the period analysed, if and when required, optimization or fine-grid screens could be requested or self-designed, and a lab technician was available to formulate the requested screens robotically using commercial stock solutions.

As a reminder from Chapter 2.1, 150nl droplets are set up using the Mosquito Crystal (TTP Labtech) with 2:1, 1:1, and 1:2 mixing ratio of protein to precipitant. Plates are incubated at 4°C or 20°C, and automatically imaged at fixed time intervals (1, 2, 4, 7, 14, 28, 56 days from setup) with the Minstrel HT system (Rigaku). Images of droplets may be visually inspected and may receive an *Image Score* of 1 to 10 according to the scheme outlined in Table 2.2. Crystals tested for diffraction are given a *Crystal Quality Score*, where a score of ≥ 1 captures diffraction reliably. For the purpose of this analysis, the following definitions of crystallization success are used:

- **Hits**: *Images Score* ≥ 3, corresponding to microcrystals or larger, where there is at least potential for further optimization.
- **Diffracting crystal**: *Crystal Quality Score* ≥ 1, where the crystal was subjected to an X-ray beam and produced protein diffraction patterns.

## 7.2 Protein to precipitant mixing ratio

Sparse-matrix screens are used to quickly identify if a protein construct is suitable for structural studies; if hits are identified, the protein is likely to be suitable, and thus hit identification is crucial and often points to subsequent follow-up strategies. We analysed whether the practice of multiple mixing ratios in sparse-matrix screening experiments increased hit identification, *i.e.* if the three-droplet strategy resulted in hits that would have otherwise not been found if only one droplet was employed instead.

The human recorded scores for the 19,817 sparse-matrix plates set up since December 2001 with at least one annotated hit were retrieved from our database. We considered all scores indicating crystallinity (3 to 10) to be hits, and the last recorded score was taken as conclusive. For each plate, the number of wells or conditions that gave hits in any sub-well was identified:

plates with no more than 5 crystallizing conditions were classed as “difficult” experiments and the protein as rare crystallizer, since they only crystallized in a narrow range of conditions; plates with more were classed as “easy” and the protein considered promiscuous. For each crystallizing well, the sub-well(s) that gave the hit(s) was identified. All plates were treated as independent experiments, even though some would have shared the same protein: this analysis was meant to reflect a hit identification exercise, which is the simplest form of analysing crystallization screening results.

Figure 7.2 shows that, unsurprisingly, if a protein is promiscuous and thus less sensitive to the change of chemical conditions for crystallization, it is also less sensitive to the change in starting concentrations: hits occur more frequently in multiple sub-wells. More significant is that for the rare crystallizers (77% of the plates), each sub-well accounts for ~30% of hits identified: omitting any sub-well would reduce by almost 1/3$^{rd}$ the likelihood of identifying hits for precisely those targets where one could least afford this.

Sub-well A (2:1 protein to precipitant) consistently produced more hits when sub-wells B and C did not, indicating that the starting concentration of many experiments may have been low, or alternatively the screens were formulated to be too diluted. Screen formulations are historic, based on protein concentrations achievable in that context, assuming 1:1 ratio for crystallization. The fact that there is a bias towards a different ratio underlines how we cannot assume that the protein concentration is correct; instead exploration is required, and this is easily achieved by varying the protein-precipitant mixing ratio.

The dataset also contains an internal control, namely the incidence of hits in sub-wells A and C but not B; this situation is not normally expected, because if both A and C are in the right concentration range, so should B. These outliers may arise from incorrect scoring, faulty droplets, or stochastic nucleation.

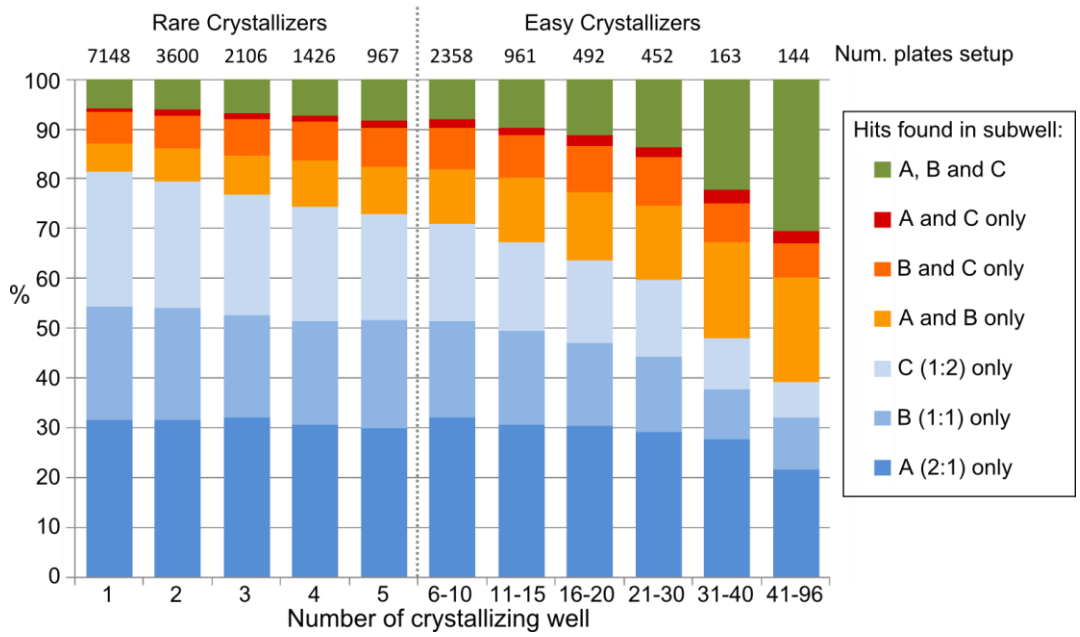



*Figure 7.2: Use of multiple drops increases likelihood of hit identification for rare crystallizers. The frequency of finding hits in any sub-well or combination of sub-wells, was plotted for crystallization plates of various levels of promiscuity, i.e. number of crystallizing wells. For rare crystallizers (number of crystallizing well ≤ 5) most hits were only found in only a single sub-well of a well (blue shades). The number of wells with hits in multiple sub-wells (green or orange) increases if the protein also crystallizes in many conditions. Hence, omitting any one sub-well would result in up to 30% of hits being missed. The total number of plates for each bar is given at the top of the graph.*

Since majority of hits were found only in one sub-well, but not consistently the same one across the entire plate, we conclude that it is important to set up multiple droplets per condition at different mixing ratios. Replications of experiments have been shown to be important in increasing crystallization success (Newman et al., 2007), and doing this with variations in mixing ratios additionally provides three sampling points on the phase diagram. The variation is likely to be considerable: Figure 7.3 shows that the 2:1, 1:1 and 1:2 mixing ratio corresponds to a three-fold, two-fold and 1.5-fold increase of both protein and precipitant concentration from the onset of experiment to equilibration, assuming complete equilibration with only the transfer of water vapour. Moreover, this introduces variations in kinetics beyond simple concentration

effects: minor changes in salt concentrations in PEG-mediated vapour diffusion result in major effects in equilibration rate (Luft & DeTitta, 1995); and the three droplets may end up in different sizes and volumes, thereby affecting the kinetics of crystallization (Forsythe et al., 2002). Importantly, trends across the concentration gradient are immediately observable, especially for non-crystallizing wells (Figure 7.4), providing additional information on the interaction of the protein and the crystallization cocktail, as well as an estimate if the concentrations were too high/low. Unfortunately our dataset does not capture how or whether such information was actually used.

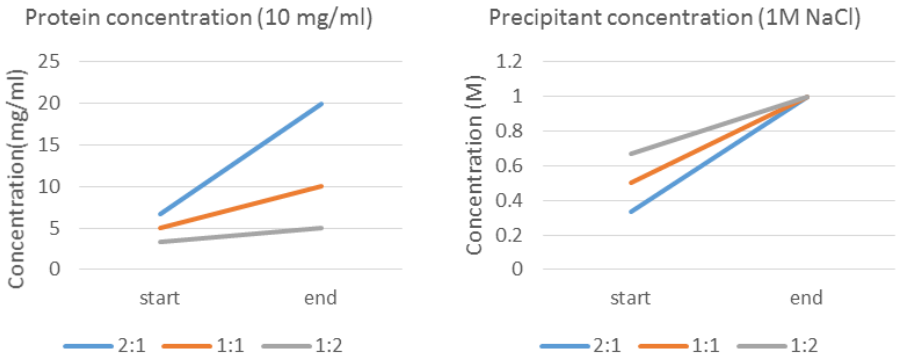



*Figure 7.3: Comparison of concentrations from the start of a vapour diffusion experiment to the equilibrated droplet of a hypothetical protein at 10 mg/ml mixed with an example precipitant (sodium chloride) at 1M, at different mixing ratio. Concentrations for both protein and precipitant increases 3-fold, 2-fold and 1.5-fold for 2:1, 1:1 and 1:2 mixing ratios respectively upon equilibration. These plots represent the case that equilibration runs to completion, which in practice varies considerably in the time required depending on the chemical components* (Luft & DeTitta, 1995)*, and assumes no phase transition of protein molecules.*

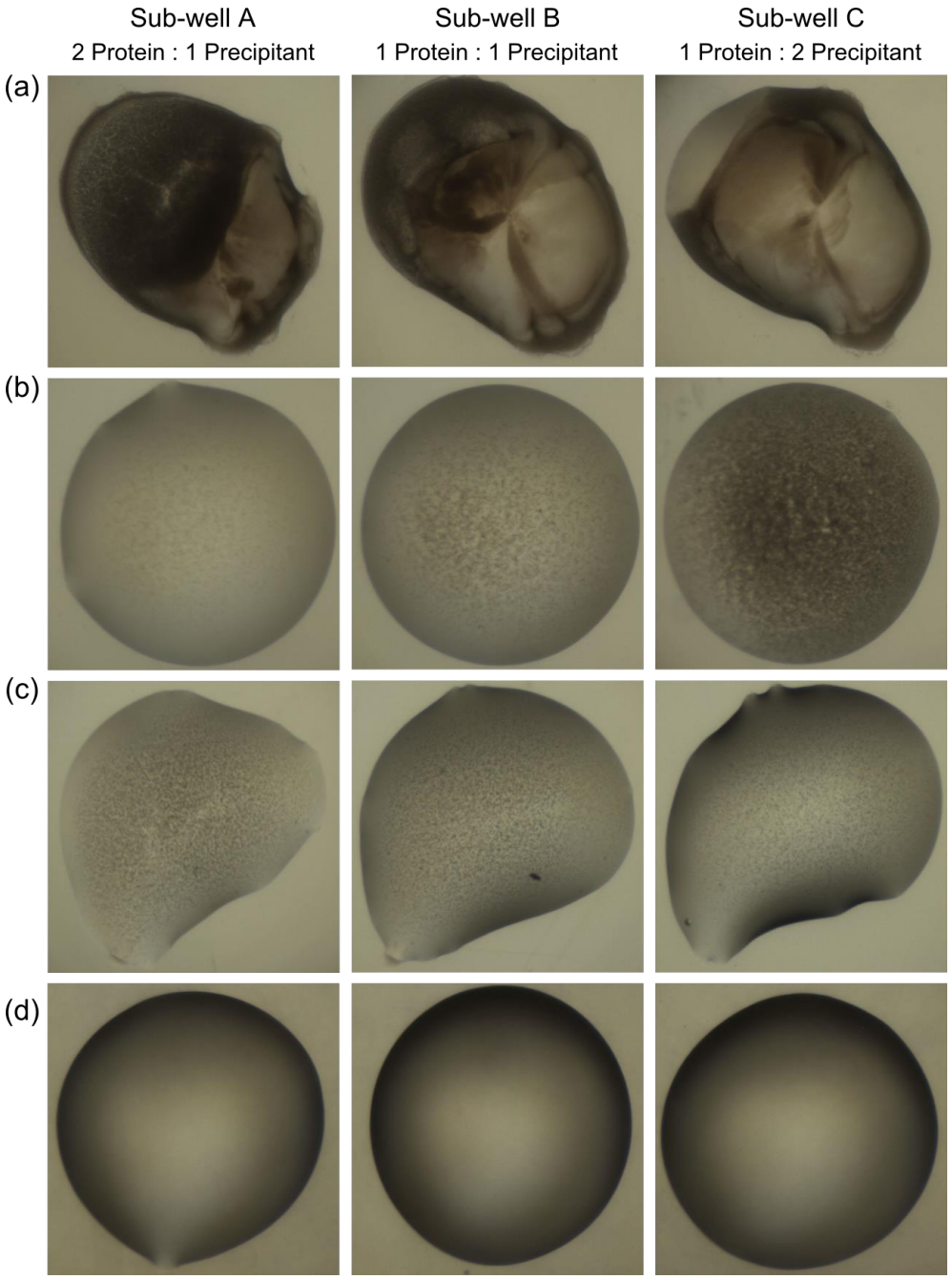



*Figure 7.4: Examples of precipitation trend observed across sub-wells with different protein – precipitant mixing ratio (2:1, 1:1, and 1:2). A decrease (a) or increase (b) of precipitate quantity across sub-wells may indicate suitable starting protein sample concentrations, where interaction with the precipitants are generating response that are not saturated. When no changes in precipitation is observed across sub-wells, protein concentration may be too high (all precipitated, (c)) or too low (all clear drops (d)).*

## 7.3 Incubation temperature.

We analysed the outcome of pairs of experiments set up at both 4°C and 20°C to identify if an incubation temperature was favourable in producing hits. Experiments were defined to be a pair if they contained identical sparse-matrix screens, the same protein sample (from the same purification batch, both either fresh or frozen), were set up on the same day with identical concentration for crystallization and compounds for co-crystallization (if any), incubated at 4°C and 20°C respectively. Such qualifiers increase our confidence that the only difference between a pair of experiment is the incubation temperature. We further limited our analysis to 587 pairs with associated diffracting crystals, where at least one diffracting protein crystal was produced in either one or both plates of each pair. For each pair, we identified if at least one hit was found in only 4°C, only 20°C, or both temperatures.

We conclude (Figure 7.5, left) that using multiple incubation temperatures was particularly important for rare crystallizers. Overall, h*its* were found at both temperatures for over 70% of the pairs, but when experiments were divided into easy and rare crystallizers (Section 7.2), 45% of *hits* were found only in one temperature for rare crystallizers. Furthermore, both temperatures yielded an almost equal share of single-temperature hits, each being over 20% for rare crystallizers: neither temperature is superior nor preferable, and omitting one risks missing many hits.

Although this effect is apparently far more marked when considering diffracting crystals (Figure 7.5, right), on further examination it transpired that by this criterion the analysis is less meaningful. In particular, experimenters tend to select crystals for harvest crystals from only one temperature, as crystals are pre-selected from images and it is then inconvenient to move between temperatures while harvesting: if those crystals then diffract well, there is then little reason to harvest at the other temperature as well.

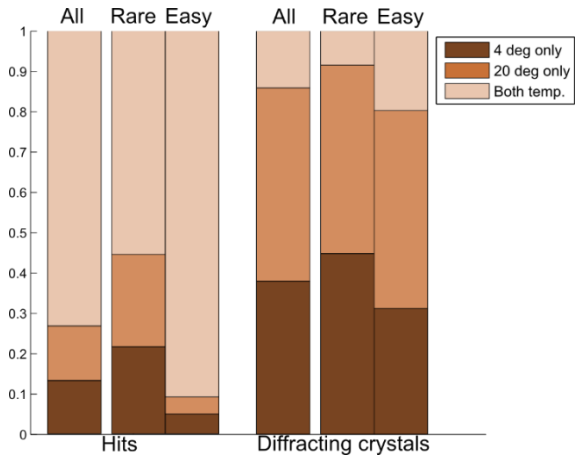



*Figure 7.5: Percentage of experiments that crystallized at only 4°C, or only 20°C, or at both temperatures for a total of 587 pairs of sparse-matrix screens. Rare crystallizers are those that exhibit crystalline behaviour in 5 or fewer conditions. Although majority of hits were found in both temperatures, up to 45% of hits were identified only in one temperature for rare crystallizers. Experimenters were even more selective when testing the crystals for diffraction, often picking only one temperature.*

## 7.4 Sparse-matrix Screens

We analysed the use of various sparse-matrix screen and its associated success by identifying the screens responsible for all PDB depositions from SGC Oxford. The breakdown of screens used and the associated success rates are shown in Figure 7.6. Of the 743 structures deposited, 52% were solved with crystals directly from a sparse-matrix screen, while the rest required further optimization, and for most of those (41% of total) the screen that yielded the original hit could be identified from the database. The JCSG, LFS, HCS and HIN screens make up 93% of sparse-matrix screen experiments at the SGC, as they were most reliably stocked; however, no fixed protocols were provided as to which should be prioritised.

These four screens were responsible for over 94% of the deposited structures. The success rates for these four screens were calculated as the percentage of structures originating from the screen over the total number of screens set up, and was found to be between 1.1% and 1.6%

(green line in Figure 7.6(a)). We also show the proportion of structures solved from crystals directly obtained from the sparse-matrix screen (blue) or an optimized condition of the screen (yellow) in Figure 7.6. As expected, the more systematic sampling of different molecular weight PEGs and ions in LFS resulted in a higher proportion crystals that did not require further optimization. However, the other screens have the advantage of sampling a wider range of chemical space, as indicated by the higher internal diversity scores, calculated by taking the mean of cocktail distances for all conditions in the screen defined by Bruno *et al.* (2014), as well as the higher number of distinct chemicals.

Other screens that also contributed to PDB depositions are shown in Figure 7.6(b), though here success rates were not calculated, since the number of samples is too low for that to be meaningful. The fact that these screens were so rarely used, even though they were always available, leads us to conclude cautiously that it is adequately informative to set up only a few screens (four in our case) – though admittedly training bias cannot be discounted.

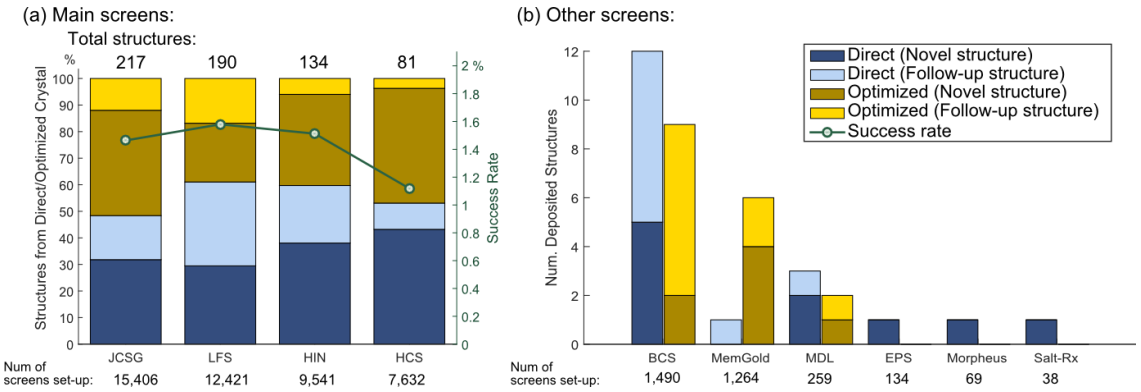



*Figure 7.6: Comparable success rates observed for the four most popular screens used at the SGC. (a) shows the number and proportion of structures solved with crystals directly harvested from sparse-matrix screens (blue) or those that required optimization (yellow) for these screens. Novel structures and follow-up structures have also been separated as they do not necessarily have the same crystallizing conditions, and are hence often re-screened. The green line indicates the success rate (percentage of structures derived from the screen/total number of screens set up). (b) shows the breakdown of the remaining deposited structures (6%) from other screens.*

*Table 7.1: Diversity and the number of distinct chemicals sampled in the four most popular screens at the SGC.*

| Screen | Internal diversity score | Num. distinct chemical components |
|---|---|---|
| **JCSG** | 0.5650 | 59 |
| **LFS** | 0.2728 | 35 |
| **HIN** | 0.5754 | 46 |
| **HCS** | 0.6196 | 54 |

## 7.5 Recommendation for 200ul of protein sample

As a first order approximation, protein purification yield is in the order of 200µl, corresponding to the volume given by Shaw Stewart and Mueller-Dieckmann (2014). Our workflow of three 150nl droplets at 2:1, 1:1 and 1:2 mixing ratio allows for the setup of 8 plates, namely 4 screens at 2 incubation temperatures; this requires 172.8µl. Each aspect of redundancy is of value: if we omit one droplet or one temperature, up to 30% and 20% of hits of rare crystallizers would have been missed respectively. The four popular screens also accounted for 94% of deposited structures.

Although some reports advocate larger drop volumes (Newman et al., 2007; Chayen & Saridakis, 2008) in our operation 150nl has been very effective: most of our structures came directly from sparse-matrix screens set up at 150nl, as were many optimization screens (though exact numbers were not recorded). This corroborates reports that nano-droplets in pore strips gave comparable results with larger volume hanging drops, implying that the pros and cons of small drop volumes balance each other (Dekker et al., 2004). We therefore do not see a reason to change this drop volume, and although we lack the direct comparative data to prove it conclusively, we submit that smaller drops are effective, especially for the added exploration of concentrations, temperature and chemical space.

## 7.6 Recommendations for Smaller Volumes:

Producing 200µl of protein sample is non-trivial. In cases where protein sample quantity is the limiting factor, we determined which of these would be the more effective use of proteins: (1) to set up more screens with fewer droplets, or to set up fewer screens with more droplets; and (2) to set up screens at multiple temperatures with fewer droplets, or to set up screens at single temperatures with more droplets. In both of these comparisons, limited by the dataset at hand, we selected subsets of the dataset to simulate outcomes of experiments by keeping the required protein volume constant. These comparison were also based on proteins that were well-behaving enough to produce sufficient samples for at least three plates to be set up.

### 7.6.1 Multiple screens vs multiple droplets

To compare if setting up more screens was advantageous over multiple droplets at different mixing ratios, we determined the best way of setting up 288 150nl droplets with the same volume of protein (21.6µl): three droplets in one screen (at 2:1, 1:1, and 1:2 mixing ratio) or three screens with one droplet each (at 1:1 mixing ratio).

166 groups of experiments were identified for this comparison. Each group consisted of three to four plates (any combinations of JCSG, LFS, HCS, and HIN), containing the same protein sample (from the same purification batch and same state (fresh or frozen)) at a similar concentration for crystallization, with identical compounds for co-crystallization (if any), incubated at the same temperature, and set up on the same day. Each group also had at least one diffracting crystal amongst the plates in the group, to ensure that protein quality was not limiting in this analysis. For each 3-screen combination in a group, we identified whether at least one hit was found if (1) all three screens were set up, but only with 1:1 mixing ratio droplets (dark blue bar, Figure 7.7), (2) only one of the screens was set up with the standard 3 droplet

protocol (green bars, Figure 7.7), or (3) no hits were found in both methods, *i.e*. hits would have only been found if the full three-drops in three-screens experiment was set up (red/orange bars, Figure 7.7).

Figure 7.7 shows that setting up 3 screens with one drop was marginally more effective at finding *hits* than 1 screen with 3 droplets for all screens (blue bars consistently higher than green bars). The exception is LFS, which also consistently had the highest number of groups where no *hits* were found (red bars with label 'L'), indicating that it is the least preferable screen among the four. This is in contrast to the data in Figure 7.6, where LFS had the highest overall success rate (structures produced from the screen), with the most structures obtained directly without further optimization. A possible explanation is that the proteins in this set of selected groups do not crystallize well in the narrow chemical space sampled by LFS, indicating that it is a screen that although works well when it does, directly producing crystals sufficient for diffraction analysis, fails miserably (no hits at all) when it does not.

This also further emphasizes that 'success rate' of screens is highly dependent on the proteins sampled for its measurement; unless all proteins were equally tested in all screens and followed up with equal rigour, reports of such screen success rates should not be taken seriously. Likewise, while the data here show that JCSG (green bar with label 'J') was consistently the most successful single screen, it is limited to the proteins included in these comparisons and should not be taken conclusively as the better screen; what is more important is that the strategy to sample more conditions, unsurprisingly, does better, although only marginally so.

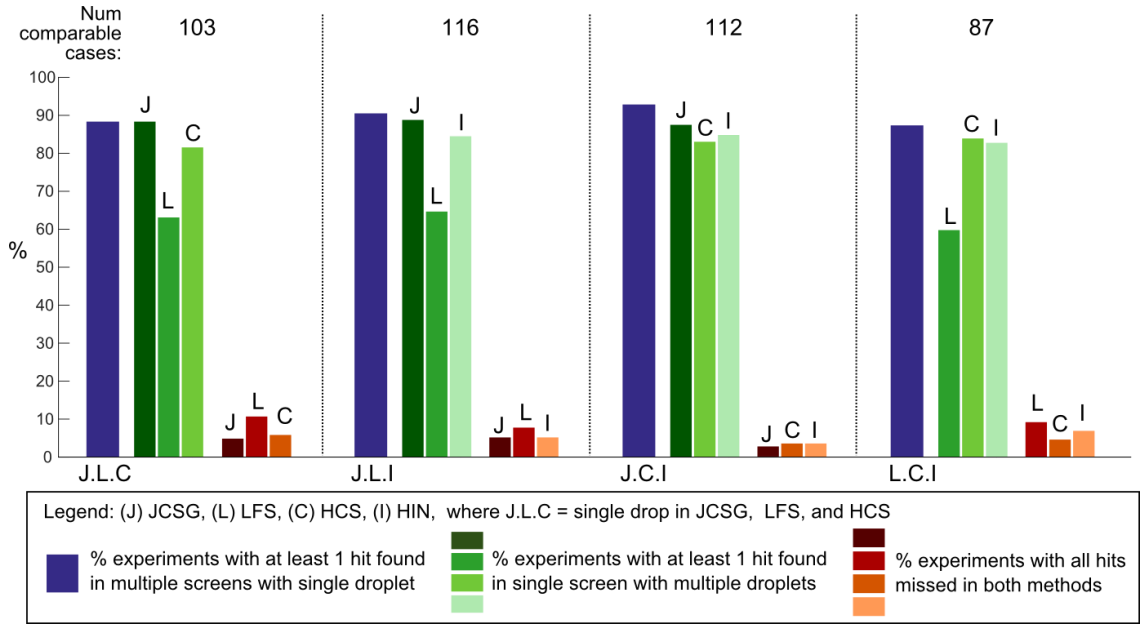



*Figure 7.7: Setting up multiple screens with single droplets lead to better identification of hits, when compared to setting up single screens with multiple droplets. Here, we compared hit identification rate if experiments were set up using one drop at 1:1 mixing ratio in three screens (blue bars), or three drops at 2:1, 1:1 and 1:2 mixing ratio in a single screen (green bars), to keep the required protein sample constant. The screens analysed were JCSG (J), LFS (L), HCS (C) and HIN (I). Each screen is only compared to permutations that include itself. Red bars indicate if no hits were found in either method.*

On the question of what ratio to use in the 1-drop-3-screen approach, it was found that three screens with 1:1 mixing ratio had consistently higher success rates (Figure 7.8). The intrinsic bias in commercial screens based on a historically strong preference for 1:1 ratio crystallization experiments may have resulted in the higher success rate. This seems especially prominent for data sets with confirmed diffracting crystals, where the proteins have also presumably been concentrated to levels suitable for the kinetics inherent in commercial screens. Comparing Figure 7.2 with Figure 7.8, a 2:1 ratio is more successful for initial hit finding, whereas for proteins falling in the required concentration range, a 1:1 ratio is more successful, particularly for finding diffracting crystals.

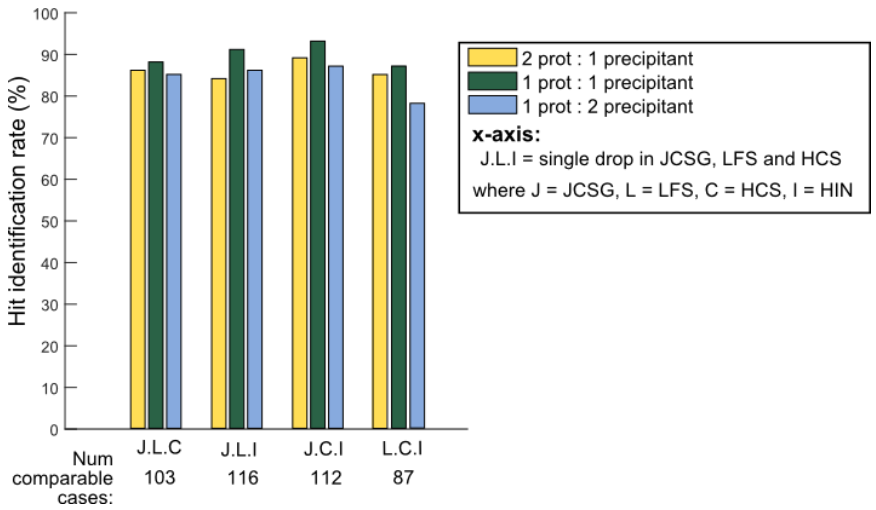



*Figure 7.8: For a one-drop-multiple-screen approach, 1:1 mixing ratio (green) consistently resulted in better hit identification rate, compared to 2:1 (yellow) and 1:2 mixing ratio (blue). The screens analysed here were JCSG (J), LFS (L), HCS (C) and HIN (I), and follows the same three-screen combinations of in Figure 7.7.*

### 7.6.2 Multiple temperatures vs multiple droplets

Similarly, we compared if setting up screens at multiple temperatures was advantageous over setting up multiple droplets at single temperatures. Once again, to keep the protein sample volume required constant, we chose the best way to set up 192 droplets (with 28.8 µl): two droplets in one temperature (at 2:1 and 1:2 mixing ratio) or two temperatures with one droplet each (at 1:1 mixing ratio).

We used the 587 pairs of experiment in Section 7.3, and similarly identified if at least one *hit* was found if (1) 2 droplets at 2:1 and 1:2 mixing ratio were set up at only 4°C, or only 20°C, and (2) single droplets at 1:1 mixing ratio were set up at both these temperatures. Table 7.2 shows a breakdown of *hits* found in both methods. The conclusion is similar to that of our previous analysis: sampling multiple temperature should be prioritised over multiple droplets at different mixing ratio, since hit identification rate was consistently higher in the 1-drop-2-temperature setup.

*Table 7.2: Comparison of hits (score ≥ 3) identified when comparing if experiments were set up only using 2 drops at 1 temperature, or 1 drop in 2 temperatures*

| | **Hits found in 2 drops, 1 temp.** | **Hits found in 1 drop, 2 temp.** | **No hits found** |
|---|---|---|---|
| **4°C** | 479 | 511 | 17 |
| **20°C** | 482 | 511 | 32 |

## 7.7 Protein concentration for crystallization

With regards to protein sample concentration for crystallization, we found no correlation between sample concentration and molecular weight. A breakdown of the protein sample concentration, as measured by experimenters prior to crystallization, vs. the molecular weight of the protein for deposited structures is shown in Figure 7.9, according to the sub-well where these crystals were obtained. We have only included depositions from crystals directly obtained from sparse-matrix screens for certainty of the protein:precipitant ratio. The high densities of data points at ~13kDa and ~35kDa were from bromodomains and kinases respectively.

It is apparent that most experimenters concentrated the proteins to ~10mg/mL, regardless of the protein's molecular weight, most probably because this has historically been the guideline. Furthermore, for practical purposes, the target when concentrating protein may also be the volume (200 µl, for example) instead of protein concentration to set up a desired number of screening experiments. The lack of correlation of molecular weight and starting concentrations (~-0.04 for all three sub-wells) shows that the guideline is worth following.

Our observations does not agree with those of Wilson and DeLucas (2014), whereby larger proteins were found to require higher protein concentrations for phase separation or crystallization to occur; larger proteins have a generally less negative second virial coefficient (B values), which corresponds to weaker protein-protein attractive forces. However, 'larger' proteins were defined as those with molecular weight > 60kDa, which we have very few data

points for. Wilson and DeLucas also showed that the narrow and moderately negative range of *B* values for all proteins translates to a large range of concentration (1 to 50 mg/ml, Figure 7.10) although the skewed distribution of B values (Figure 7.10(a)) favours protein concentrations of <5 mg/ml (green line, Figure 7.10(b)), which is in the ballpark of the effective starting protein concentrations in our vapour diffusion experiments.

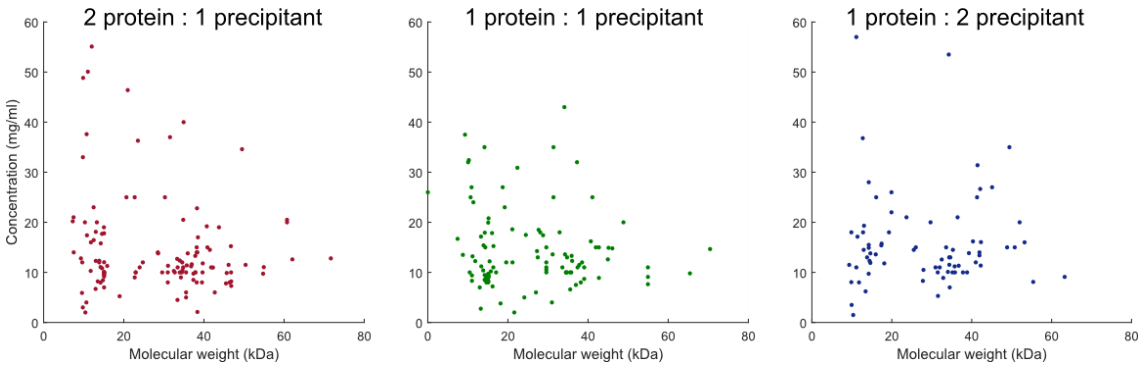



*Figure 7.9: Protein sample concentration vs molecular weights of PDB depositions from crystals directly obtained in sparse-matrix screens at the SGC Oxford, according to the protein:precipitant mixing ratio that produced the crystal responsible for the structure.*

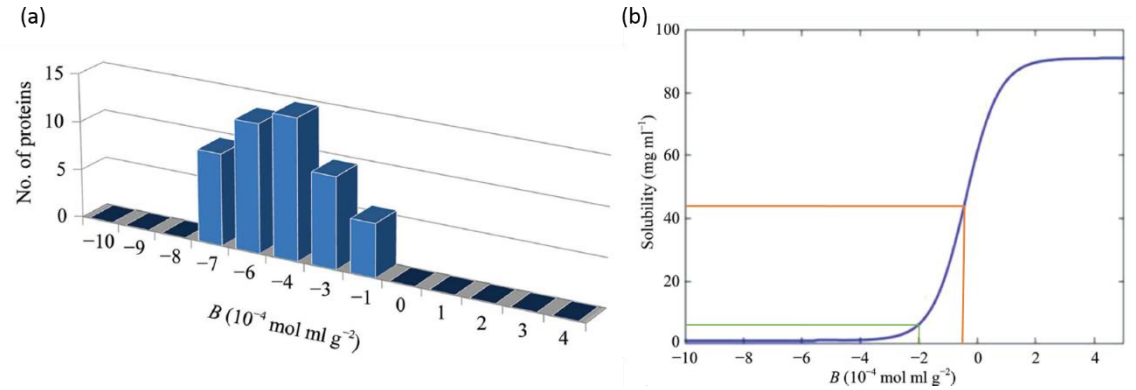



*Figure 7.10: Relationship between second virial coefficient (B values) and solubility. (a) Measured* B *values for 50 different proteins dissolved in crystallizing solvents. (b) Typical crystallizing proteins were soluble to under 50 mg/ml (*B *= -0.5 × $10^{-4}$ mol ml $g^{-2}$, orange lines), while the skewness of the distribution of* B *towards more negative values indicate that protein concentrations of < 5 mg/ml are favoured (*B *< -2 × $10^{-4}$ mol ml $g^{-2}$, green lines). (a) and (b) reproduced and modified from Wilson & DeLucas(2014) respectively*

## 7.8 Late formation of crystals

The length of time to keep a crystallization plate in the incubator is a balance between the likelihood of crystals appearing at some point in time, the limited space available, and one's hope or desperation. While our data cannot address the latter two, we examined how long it took crystals that led to PDB depositions to appear. At the SGC, plates are officially removed from the incubators after 60 days, but are available in practice for six to nine months without scheduled imaging. The latest time of the appearance for crystals responsible for our deposited structures was estimated using the timestamp of the images with crystal annotation. Furthermore, crystals where annotations were only found on images captured after 30 days were manually curated for accuracy; if the crystal appeared in earlier annotations, but was not annotated in the earlier inspection (false positive), we corrected the time of appearance to that of the earliest image with the crystal. We also removed crystals with missing inspection data or images for visual verification.

The cumulative histogram of the latest estimate of crystal appearance for the remaining 586 PDB depositions is shown in Figure 7.11(a). 11 structures (2%) were from late-forming crystals, most likely only forming after 30 days. The exact times when these crystals appeared were indeterminate since automated imaging is sparse after 7 days. However, the lower and upper bound of time can be determined using the inspection before and when the crystal was observed, as shown in Figure 7.11(b). Since 2% of structures formed after 30 days, we conclude it is necessary to keep plates in the incubators for 60 days.

We also found that almost 1% of hits scored by crystallographers were for late-appearing crystals, in an extended analysis on 15,781 plates with track-able scoring dates. Of these, 7.6% had hit annotations after 30 days of experiment setup. Since there were too many images to curate visually, plates were filtered using TeXRank scores. We only visually validated 154 plates with an increase of ranking score of > 0.2 from images captured before and after 30 days. Of

these, 128 were verified to produce crystals, phase separation or substantial crystal growth after 30 days; 4 targets in these plates only had annotations after 30 days, furthering our case for a 60-day incubation period.

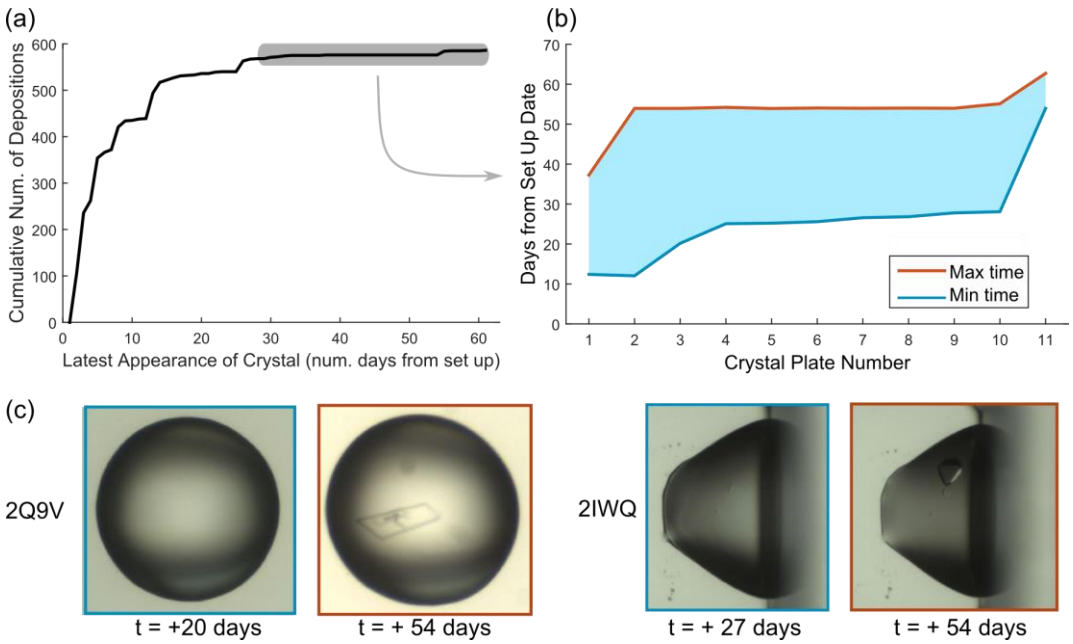



*Figure 7.11: Late-appearing crystals that led to PDB depositions. (a) Cumulative histogram of latest estimate (in days) for the formation of these crystals. The steps on the curve correspond to typical imaging times. The earliest and latest times for crystals appearing after 30 days (grey shaded area) are shown in (b), where the earliest time was taken from the image before the crystals can be visually identified. (c) shows examples of two late-appearing crystals.*

## 7.9 Concluding remarks

The data show that it is worth setting up variations of the same crystallization cocktail. This was a reassurance of the effectiveness and necessity of the initial crystallization strategy at the SGC, Oxford. If 200µl of protein sample is available, 8 plates should be set up using 4 screen-types (JCSG, LFS, HCS and HIN are generally used at the SGC) at 2 temperatures (4 and 20°C). Each well of a plate should consist of three 150nl droplets at different protein to precipitant mixing ratio (2:1, 1:1, and 1:2). Where protein sample is limited, priority should be given to diversifying

screening conditions and temperature, before repetitions with variable mixing ratio. Specifically, our data has shown that setting up single droplets at 1:1 ratio with JCSG, HCS and HIN, at 4°C would maximise success rates. It is also worth storing plates for automatic imaging for up to 60 days. The outcome of the current practice at the SGC is a target success rate (success defined by at least one diffracting crystal of the target) of 43.75%.

*Special thanks to Anna Carpenter, for her database curation efforts, especially for filling up missing crystallization conditions for deposited structures by the SGC in the database. Her work enabled the tracking of many more origin conditions in sparse-matrix for these deposited structures (yellow bars in Figure 7.6).*

## 7.10 References

Chayen, N. & Saridakis, E. (2008). *Nat. Methods*. **5**, 147–153.

Dekker, C., Haire, L., & Dodson, G. (2004). *J. Appl. Crystallogr.* **37**, 862–866.

Forsythe, E. L., Maxwell, D. L., & Pusey, M. (2002). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1601–1605.

Jancarik, J. & Kim, S. H. (1991). *J. Appl. Crystallogr.* **24**, 409–411.

Krojer, T., Pike, A. C. W., & von Delft, F. (2013). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1303–1313.

Luft, J. R. & DeTitta, G. T. (1995). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **51**, 780–785.

Newman, J., Xu, J., & Willis, M. C. (2007). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 826–832.

Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. a, & Gileadi, O. (2010). *J. Struct. Biol.* **172**, 3–13.

Shaw Stewart, P. & Mueller-Dieckmann, J. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 686–696.

Wilson, W. W. & Delucas, L. J. (2014). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **70**, 543–554.

# 8. Correlating Visible Textures in Droplets with Sub-visible Crystallinity

It is common to hear precipitates referred to as "good" and "bad" (Luft et al., 2011). "Good" precipitate presumably implies presence of crystalline material not resolvable by light microscopes, but are nevertheless worth pursing and optimising; while "bad" precipitate implies denatured protein. These concepts are however not formally defined, and there are no established and objective method to differentiate the two. To detect nanocrystals, methods involving dynamic light scattering, birefringence microscopy, UV fluorescence imaging and SONICC have been reported, although these methods are not reliable for crystals under 5µm, while dynamic light scattering does not even differentiate crystalline from amorphous aggregates (Stevenson et al., 2014). From the limited experience and data available, SONICC imaging of droplets with what would typically be considered as good precipitate usually give no signal (Timm Maier, personal communication). The guidelines available pertain to visually distinguishing these precipitates, though they are subjective and dependent on the experimenter's experience, or involves the addition of dye or determining if the precipitate can be redissolved. For the purpose of this chapter, we will define nano-crystallinity as “good” precipitate, and all other non-crystalline precipitate as bad precipitate.

The occurrence of multiple outcome in a droplet is not unusual. Figure 8.1 shows an example of a droplet with what is commonly termed denatured precipitate (brown-ish, amorphous), granular precipitate, crystals, and clear regions. It would appear that the ability to objectively and confidently classify precipitate as good or bad precipitate, and thus identify the presence of good precipitates in a mixed-behaviour droplet like that in Figure 8.1 will undoubtedly be useful to guide and validate optimization strategies. Beyond that, the advancement of Serial Femtosecond Crystallography (SFX) techniques and X-ray Free Electron Laser (XFEL) sources

have raised the need for large quantities of nanocrystals and microcrystals; if good precipitates can be proven to be crystalline or nanocrystals, this would be a powerful tool for identifying material for XFELs and SFX, since producing precipitate is more reliable and scalable.

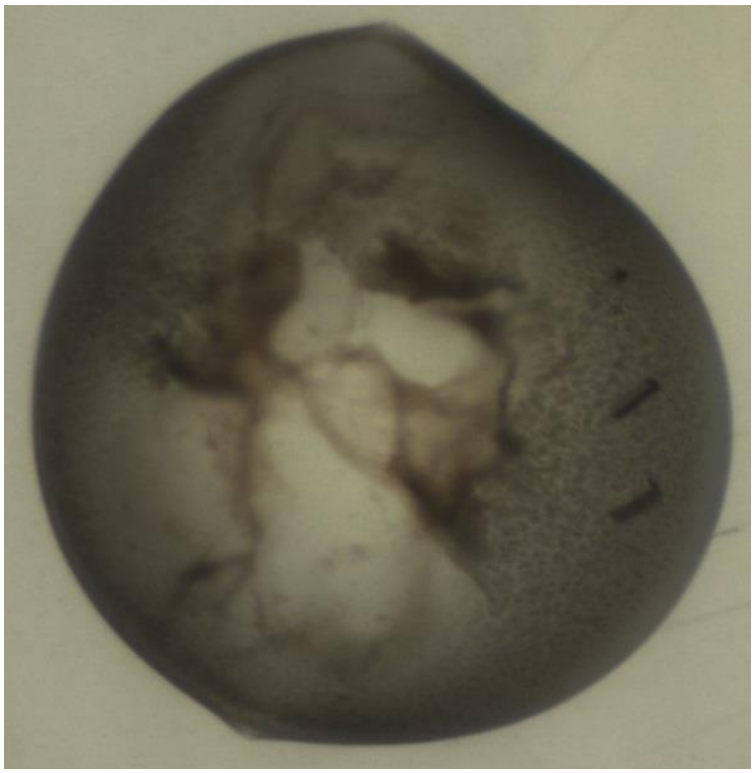


*Figure 8.1: Example of a droplet with multiple precipitation behaviour. Based on typical visual guidelines, the centre of the droplet contains clear regions and denatured precipitate, which changes to granular (good) precipitate towards the right portion of the droplet, where crystals are also found growing. However, no formal definitions exist to properly describe such droplets.*

Image analysis may be exploited as a cheap and objective method to classify precipitate. We thus extend the texton method to segment droplets by the type of precipitate commonly believed to be good or bad. The ultimate aim of such segmentation of a droplet was to correlate regions with underlying protein behaviour, or the proven presence of nanocrystals. We explored the use of electron microscopy as the experimental validation of sub-visible structures for different types of precipitate. The ability to correlate localised texture with underlying structures would allow for cheap and fast identification of nanocrystals from bright-field

imaging alone, without the need for further sample-heavy experiments with sophisticated equipment.

## 8.1 Segmenting precipitate in droplets with superpixels

Superpixels, as its name suggests, groups pixels with similar local features together to reduce redundancy, and subsequently its complexity for image processing tasks like segmentation. There are a number of established superpixel algorithms, typically divided into graph-based or gradient-ascend based methods. However, for superpixels to be useful in our application, the algorithm should be (1) computationally efficient due to the large number and continuous processing of new images, as well as (2) result in good segmentation. In a comparative study of superpixel algorithms, the Simple Linear Iterative Clustering (SLIC) outperformed eight other methods in both these criteria (Achanta et al., 2012). SLIC was thus our method of choice, and we used the implementation of the algorithm from the VLFeat library (Vedaldi & Fulkerson, 2008). The proposed SLIC approach (Achanta et al., 2010) is an adapted k-means clustering of pixels in a 5-D space defined by CIELAB colour space (L,a,b values) and spatial coordinates (x and y pixel coordinates). Additionally, two parameters control the superpixels generated by the algorithm: (1) number of superpixels in the image desired, $K$ and (2) regularizer, which controls the compactness of the superpixels.

With $K$ desired superpixels for an image with $N$ pixels, cluster centres are defined at grid interval $S = \sqrt{N/K}$ for approximately equal sized superpixels. The spatial extent for each superpixel is thus approximately $S^2$, and each pixel will be associated with one of the cluster centres within its $2S \times 2S$ reach. The search is thus limited to this area, which contributes to the increased computational efficiency of the algorithm. Because the Euclidean distance space for CIELAB

colours ($d_{lab}$) and spatial distance ($d_{xy}$) are not comparable, to give equal weighting to both aspects, the method uses the following distance measure $D_s$:

$$D_s = d_{lab} + \frac{m}{S} d_{xy},$$

where

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2},$$
$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}.$$

The term $\frac{m}{S}$ is the regularizer, where a larger value emphasizes spatial proximity, and thus produces more compact clusters.

In our application, the filter response vector for each pixel was used instead of CIELAB colour vectors, since texture is of interest and colour is not relevant in crystallization droplet images. As described in Chapter 2, the resulting filter response has eight dimensions (from 3 bar and 3 edge filters at different scales, as well as 2 circular filters). Here, we use this filter response vector for each pixel to calculate $d_{FilterResp} = \sqrt{\sum_{j=1}^{8}(FR_{kj} - FR_{ij})^2}$ in place of $d_{lab}$. We chose $S = 30$ as the step size of grids and $m = 0$, since precipitates are often not compact and may take odd and elongated shapes. Examples of superpixel segmentation are shown in the left side for each pair of droplets in Figure 8.2(a).

To simplify and give meaningful segmentation, the underlying distribution of textons in each superpixel from a set of images (those used to form the texton dictionary in Chapter 2.2.2) were clustered with DP-means (with λ = 0.75). In this clustering, we used Hellinger distance instead of Euclidean distance since we were clustering histograms. Clusters were labelled via visual inspection as (1) clear, (2) good (granular) precipitate, (3) bad precipitate, or (4) crystal. Analogous to the texton method, cluster centres of the superpixel-regions were recorded in a

dictionary, so that superpixels of a new droplet can be compared to the dictionary, and labelled accordingly to one of the four classes. Examples of such segmentation are shown in right side of each pair of droplets in Figure 8.2(a), with further examples in Figure 8.2(b), where the boundaries of each class have been colour coded accordingly. In general, the segmentations tend to match human perception; the main exceptions are dark-shadow regions at the edge of the droplets, which may be falsely classified as denatured precipitate (blue outline).

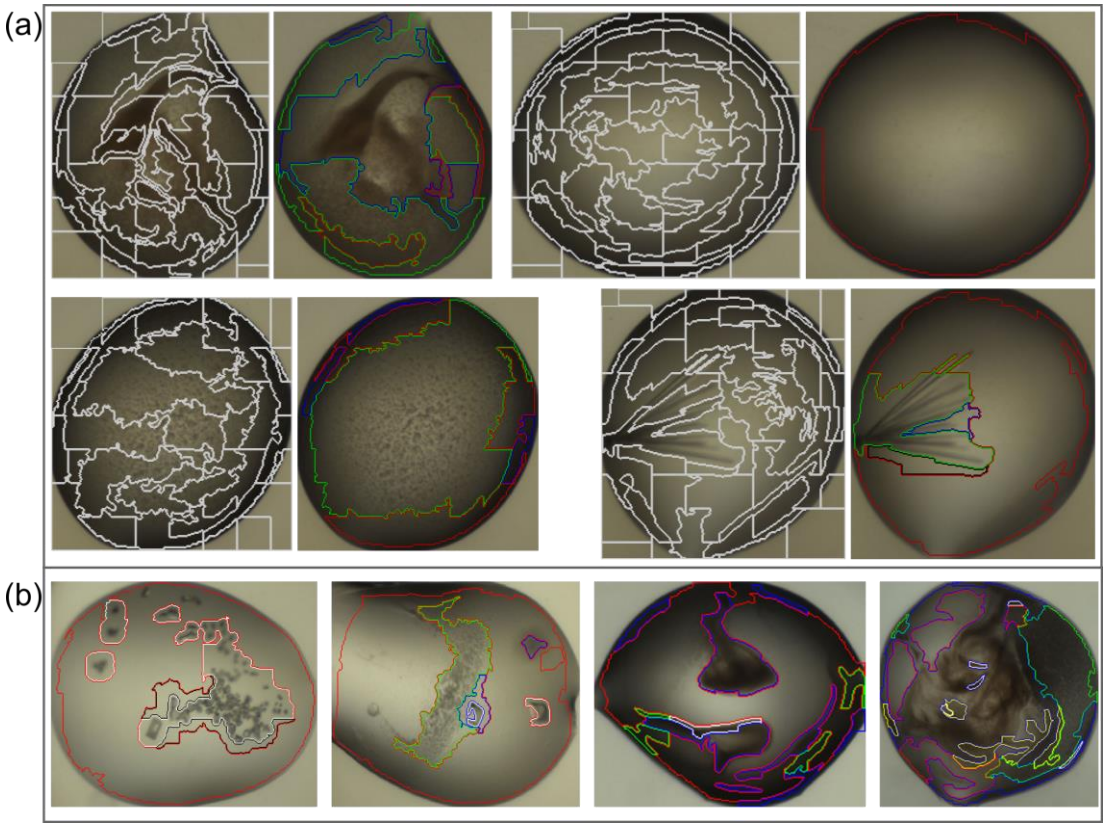



*Figure 8.2: Superpixel segmentation of crystallization droplets. (a) The left side of each image pair shows the SLIC superpixel boundaries. With regularizer = 0, the superpixels may take on any form or shape. Matching the underlying texton distribution of each superpixel region to a dictionary allows the classification of the following regions: clear (red), denatured precipitate (blue), granular precipitate (green) and crystalline region (white). These are shown in the corresponding right image of each pair, with more examples in (b).*

With these segmentations, we sought to experimentally identify the underlying sub-visible structures that gave rise to the textures observed. The next section describes our preliminary work using electron microscopy techniques to observe nano-crystallinity.

## 8.2 Identifying nanocrystals with electron microscopy

### 8.2.1 Transmission electron microscopy

Our first attempt at correlating droplet texture to nano-crystallinity was to use negative stain transmission electron microscopy (TEM), since it was recently reported as a means to observe nanocrystals from crystallization drops with granular precipitates not typically classified as hits (Stevenson et al., 2014). TEM has also been shown to be a viable method for solving the structure of lysozyme at cryo-temperature (Shi et al., 2013). Negative-stain TEM with uranyl acetate was used by Stevenson *et al.* with 400 square mesh copper grids. However, preparation was non-trivial: 5 to 8µl of sample was accumulated from multiple droplets for grid preparation, and glass beads were used to crush thick nanocrystals. The significant manipulation of droplets prior to grid preparation was not ideal for our application since the direct mapping of droplet behaviour to sub-visible structures can no longer be carried out.

However, due to its simplicity and published success, we nevertheless decided to use negative-stain TEM to identify the presence of nano-crystalline material. Unlike the work of Stevenson, we worked with 150nl droplets, in line with our crystallisation set up. As a positive control, we used lysozyme (Sigma Aldrich, 100mg/ml in 0.1M sodium acetate pH4.6) as the test protein, with 3.5M sodium chloride, 15% PEG 5000, 50 mM sodium acetate pH 4.5 as the precipitant to generate nanocrystals, as reported in the literature (Shi et al., 2013). Crystallization droplets were set up on flat-bottom, single-well Greiner plates, which had dimensions (well depth) suitable for the direct application of copper grids on crystallization droplets for grid preparation.

In our crystallization experiments, granular precipitates were visible within 20 minutes, which turned into visible protein crystals by 24 hours, indicating that the early precipitates observed were likely to be crystalline. Grid samples were thus prepared within 20 to 30 minutes from crystallization set up, by placing the copper grid on the crystallization droplet, and incubating for 1 minute. Excess liquid were blotted using filter paper, and the grid was incubated with the negative stain for 1 minute. The blot-and-dye step was repeated three times. Finally, the grid was left to dry and stored in typical copper grid holders prior to data collection.

Two dyes were explored as a workaround for the authorization and disposal requirements involved when working with uranyl acetate: Nano-W (www.nanoprobes.com) and Uranyl-Acetate Alternative (UAA), which is a gadolinium triacetate based stain (www.tedpella.com). As control experiments, a grid was prepared with no staining procedure, but left to dry after blotting off excess liquid from the crystallization droplet; another grid was prepared with a droplet with only precipitant solution and no protein, stained with Nano-W. The latter grid was to ensure that any crystalline material identified were not precipitant crystals, nor a product of the interaction of precipitant and stain.

Figure 8.3 shows images from grid preparation. For the unstained grid (Figure 8.3(a)), dehydration caused a thick layer of material on the grid, and hence no TEM data was collected. Data for other grids were collected at the Research Complex at Harwell (RCaH) with a JEM-2100 Transmission Electron Microscope, under the supervision of Dr Kyle Dent. Objects up to ~50nm with sharp edges, were identified on grids stained with UAA or Nano-W (grids in Figure 8.3(b) and Figure 8.3(c), objects in Figure 8.4). The edge and shape of the objects indicate that these were crystalline material, but could not be conclusively verified. However, the density of material in the crystallisation droplet was not reflected by the low number of objects on the grid. The triple staining procedure may have washed off samples from the grid. Furthermore, we did not glow-discharge the copper grids prior to sample preparation due to the lack of access

to a glow-discharger, and thus the copper grids may have been too hydrophobic for good adherence of material to the grid. However, since such objects were not observed in the control experiment with no protein (Figure 8.3(d)), we believe that the objects were indeed from the protein sample.

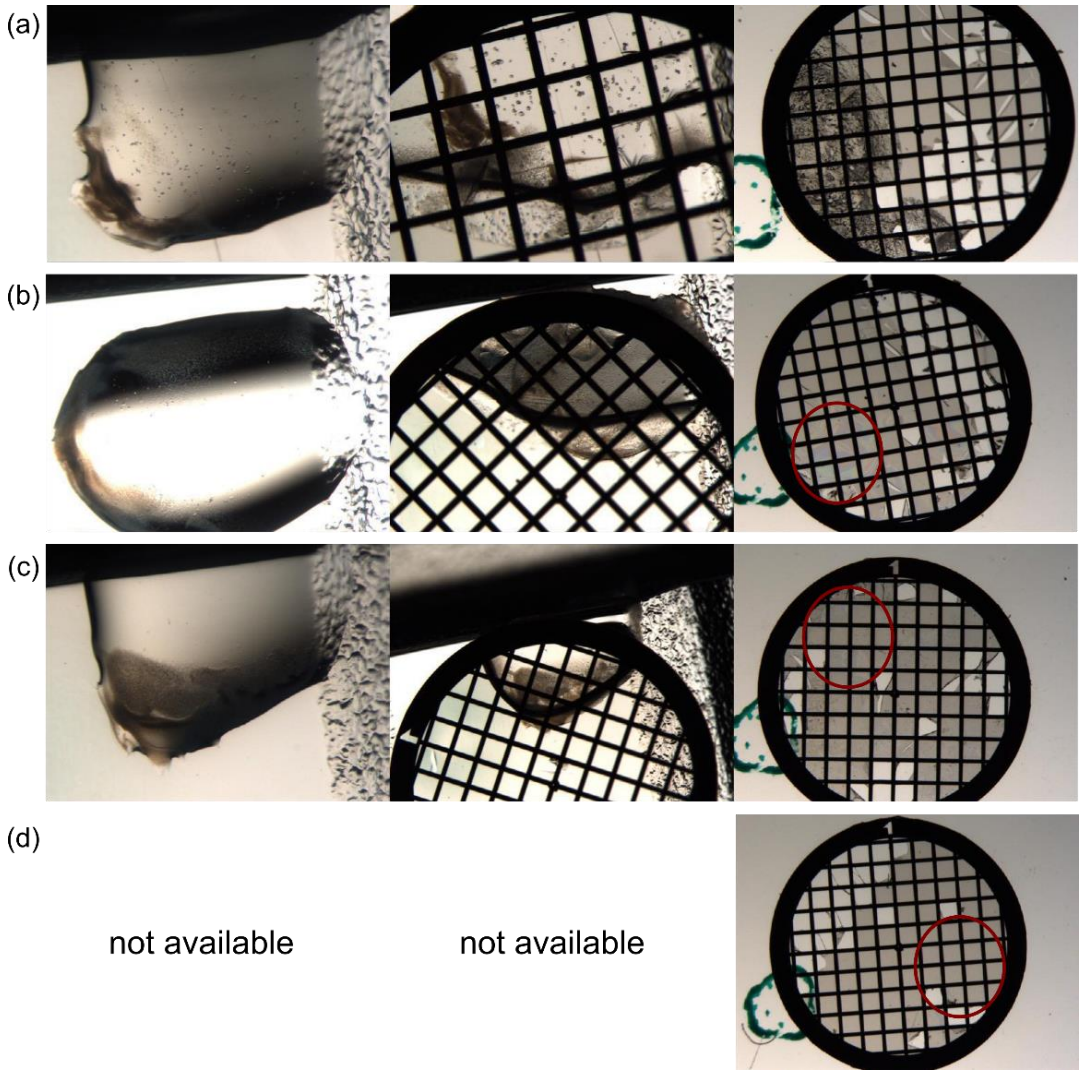



*Figure 8.3: TEM grid preparation. 100-mesh carbon-coated copper grids were prepared with no staining (a), staining with Uranyl Acetate Alternative (UAA) (b), Nano-W (c), and from droplets with no protein, but only the precipitant, stained with Nano-W (d). The left panel shows the 150nl droplet before grid-preparation set up on the flat-bottom CrystalQuick 96-well sitting drop plate (Greiner). The middle panel shows the incubation of grid with the droplet (carbon side down), while the right panel shows the resulting grid, with the estimated location of sample circled in red. The torn carbon film on the grids were due to handling, but were often away from the sample.*

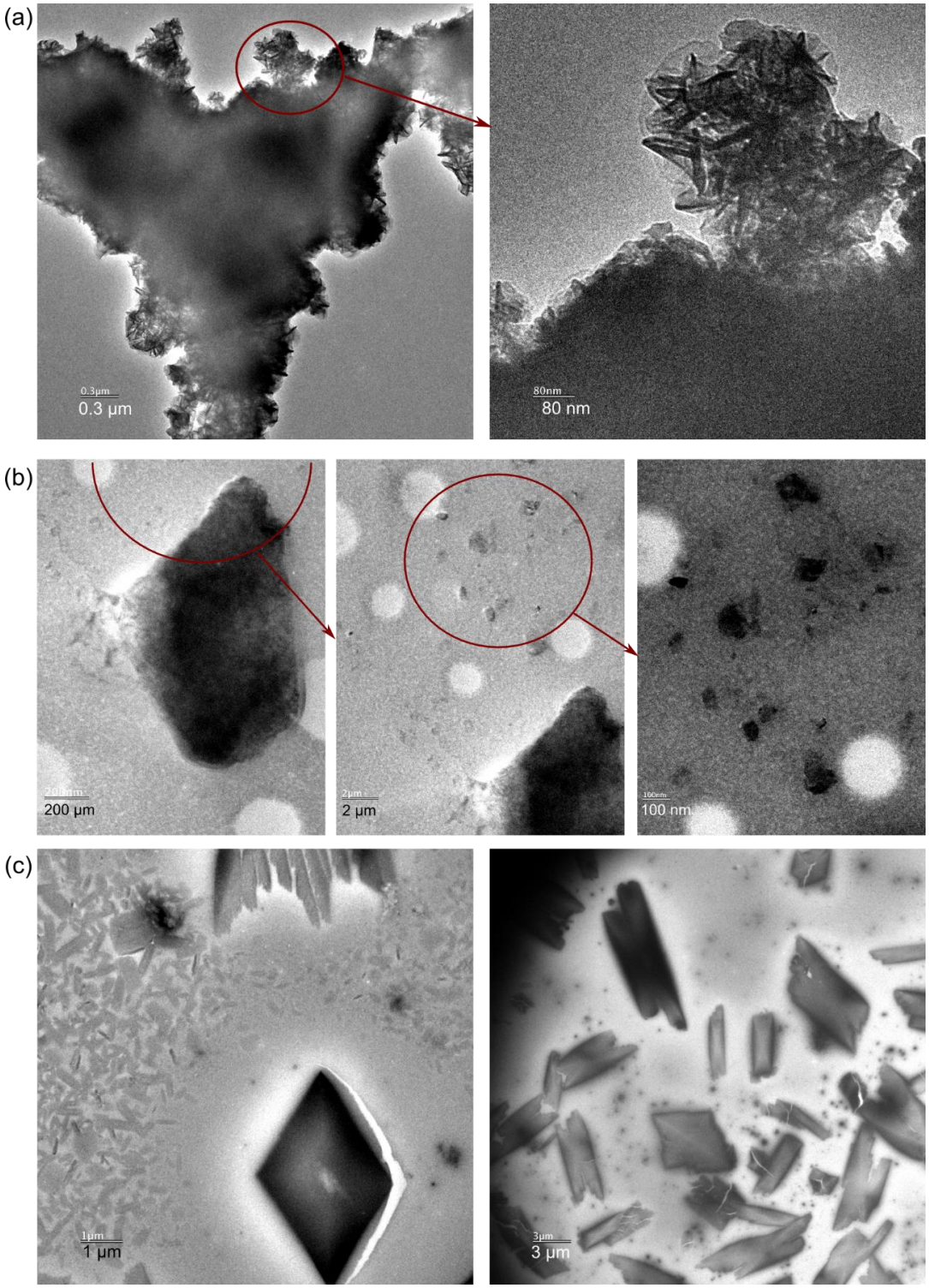



*Figure 8.4: TEM images from grids stained with Uranyl Acetate Alternative (a) or Nano-W (b and c). (a) and (b) shows a series of zoomed in or translated portions of the grid with decreasing scale, while examples are shown in (c).*

It was unclear what the effects of the dehydration of droplets during sample preparation were in preserving or inducing crystallinity. Furthermore, present grid preparation methods does not allow the direct mapping of superpixels to sub-visible behaviour of proteins, since the morphology and location of the droplet completely changes during grid preparation. We thus found negative-stain TEM to be less than ideal, especially when the conclusive presence of protein nano-crystals could not be established.

### 8.2.2 Scanning electron microscopy with WETSEM capsules

To investigate precipitate in its native (wet) state, we used WETSEM QX-102 capsules (Quantomix), shown in Figure 8.5, which were designed for imaging wet biological samples with scanning electron microscopy (SEM). The capsules also allow for crystallization droplets to be set up directly in the capsule and imaged with a microscope prior to SEM data collection; this was ideal for the correlation of texton superpixel regions to the corresponding underlying physical behaviour. The recommended working volume was however 15 µl. We chose a different crystallisable protein (JMJD2AA-p085) in which granular precipitate can be reliably produced, instead of using lysozyme, where granular precipitate often turned into crystals overnight.

10 µl droplets were set up in the liquid dish of the capsules, at 1:1 mixing ratio of protein and the following precipitants: 20% PEG 3350, 0.2M ammonium nitrate (for granular precipitate) and 2.4M sodium malonate (for amorphous precipitate). The capsules were equilibrated overnight with the precipitant in the MP-10 Multi-well Plate (Quantomix), where the precipitant was placed in troughs along the sides of the plate to mimic the reservoir well in a typical vapour diffusion experiment (Figure 8.5). Prior to data collection, the capsules were mildly centrifuged (500g, 5 minutes) to bring precipitates to the membrane, before sealing with the sealing stub.

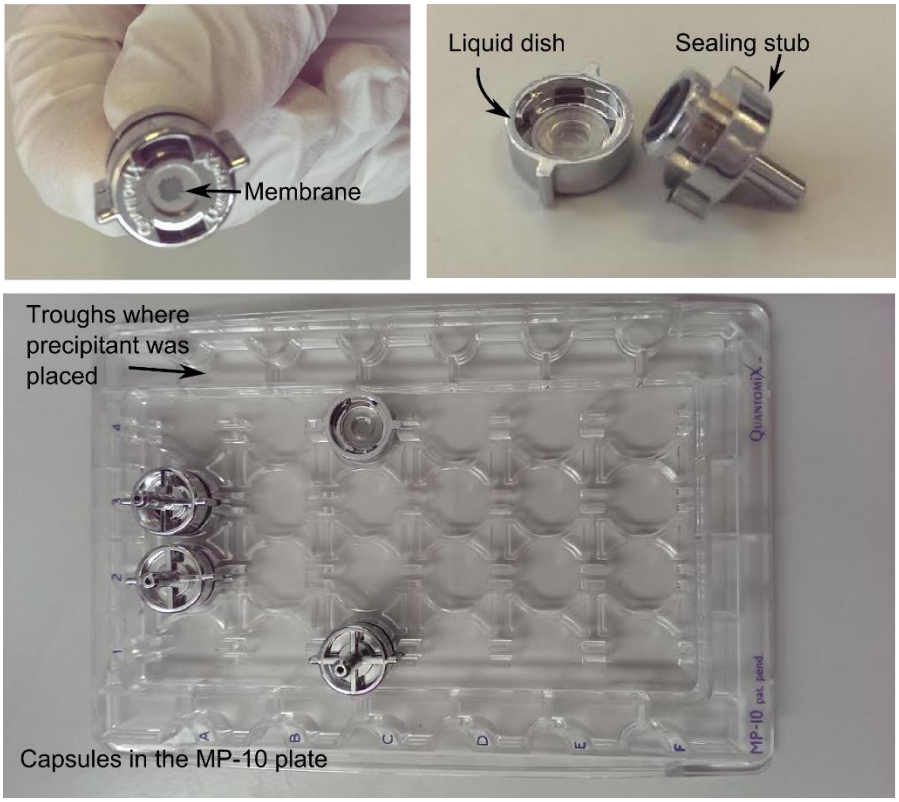



*Figure 8.5: WETSEM QX-102 capsules. Crystallization droplets were set up in the liquid dish, and equilibrated in the MP-10 plate without the sealing stub overnight. The MP-10 plate has troughs along the side where precipitant can be placed. The MP-10 plate also allows for easy imaging of the capsules with normal light microscopes (before sealing with the sealing stub).*

Separately, we also set up 450nl crystallization droplets with the above conditions to establish methods for the local transfer of precipitate onto a WETSEM capsule. This will theoretically allow us to selectively transfer precipitate from a droplet, thus enabling pixel-to-structure mapping, and maximise the capacity of each capsule, since multiple transfers may occupy a single capsule; the disadvantage of the method would be the manipulation of precipitate from its native droplet. We used nylon crystal mounting loops to 'scoop' precipitates from a desired portion of the droplet, and dotted it gently on the WETSEM capsule without puncturing the membrane. To maintain humidity in the capsule chamber, we added 1.5µl of water to one side of the liquid dish and 1.5µl of water to the base of the sealing stub. Furthermore, we found that the process could only be carried out in the cold room (4$^{o}$C) for lower evaporation rate.

Data was collected with a FEI Quanta 600 FEG at the Clarendon Laboratory, Oxford University, under the supervision of Dr Jason Brown. Pressure was set at ~$1\times10^{-1}$ mbar to prevent leakage from the capsules, and the Back Scattered Electron Detector was used to form images. Unfortunately, nothing deemed crystalline were observed in all capsules. Adherence of precipitate to the membrane of the capsule may have been poor; suspended precipitates were likely beyond the penetration depth of the electron beam. Membrane coating to improve adherence of biological samples with Poly-L-Lysine or Poly (sodium-4-styrenesulfonate) coating have been suggested by the developers of WETSEM, depending on the charge of molecule, although these could not be tested due to time constraints. Figure 8.6 shows the light-microscope image of the capsule and its corresponding SEM image where the crystallization droplet was set up directly in the capsule and Figure 8.7 show the results of the loop-transfer experiment.

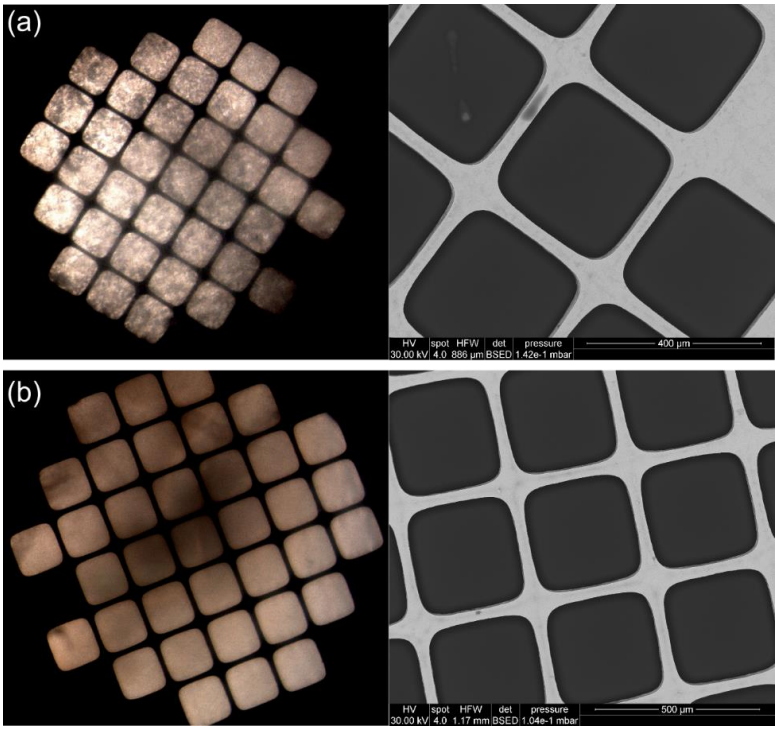



*Figure 8.6: Crystallization outcome in WETSEM capsules. (a) Protein and 20% PEG 3350, 0.2M ammonium nitrate for granular precipitate, and (b) protein and 2.4M sodium malonate to produce amorphous precipitate. The SEM image however shows no signal even at higher magnification.*

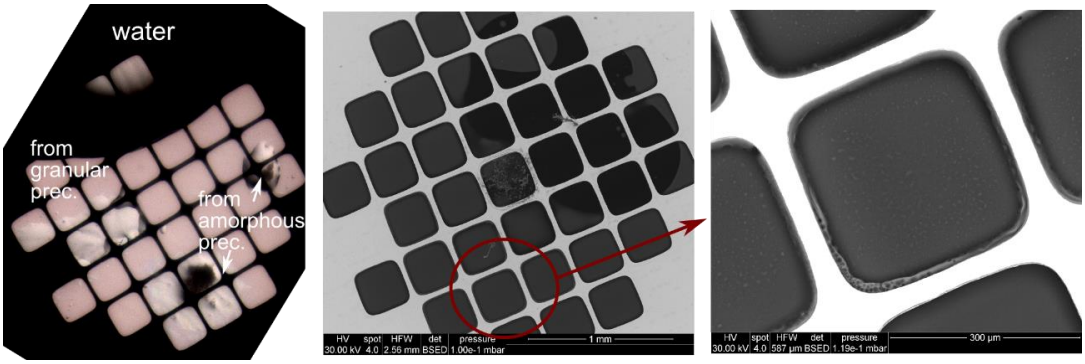



*Figure 8.7: WETSEM capsule with dotted precipitate from crystallization droplets. The bright-field image has been rotated to match the orientation of the SEM image. The dark region in the SEM image in the middle portion is likely to be from water movement due to transportation or handling of the capsule. Some undefined objects can be seen in certain grids, but is inconclusive due to the resolution.*

## 8.3 Concluding remarks

The experimental results with negative-stain TEM and WETSEM capsules were inconclusive and requires further optimization of protocols. A straight-forward and conclusive method to identify nanocrystals or crystalline behaviour in precipitate was not established, although both TEM and SEM methods have shown potential. Definite nanocrystals, crucial as a positive experiment in the development of such methods, were non-trivial to produce due to the very lack of identification methods. Published efforts have usually crushed larger crystals with glass beads to achieve this, although the effects of such treatment on fragile protein crystal have not been investigated in depth. Cryo-EM is likely to better maintain sample viability compared to negative-stain EM, although more expertise and infrastructure will be required for such sample preparation.

Higher resolution microscopy, including super-resolution microscopy may be useful tools, although the typical resolution limit of 500nm may not be sufficient. These techniques also often

do not support direct imaging of droplets in usual crystallization plates due to the short focal depth of the microscopes. These will again, require the perturbation of crystallization droplets similar to that of EM techniques, on top of the need to label proteins with fluorescent tags.

Even if these techniques become established for identifying nanocrystals, they are far from high-throughput, and access to an electron microscope or super-resolution microscope remains limited. This further underlines the importance of using image analysis on easily captured droplet images as a proxy for physical behaviour. Such automatic and objective determination of 'good' and 'bad' precipitate will definitely be useful for the community.

*Special thanks to Dr Kyle Dent for his advice and supervision in TEM experiments; Dr Opher Gileadi who brought the WETSEM capsules to our attention; Dr Jason Brown for his advice and supervision in SEM experiments.*

## 8.4 References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2010). *EPFL Tech. Rep. 149300*. 15.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2281.

Luft, J. R., Wolfley, J. R., & Snell, E. H. (2011). *Cryst. Growth Des.* **11**, 651–663.

Shi, D., Nannenga, B. L., Iadanza, M. G., & Gonen, T. (2013). *Elife*. **2013**, 1–17.

Stevenson, H. P., Makhov, A. M., Calero, M., Edwards, A. L., Zeldin, O. B., Mathews, I. I., Lin, G., Barnes, C. O., Santamaria, H., Ross, T. M., et al. (2014). *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8470–8475.

Vedaldi, A. & Fulkerson, B. (2008).

# 9. Conclusions

The effectiveness of sparse-matrix screens for the identification of crystallization conditions is undeniable. Whilst there may be differences in the preferences of screen-type, and the availability of special in-house screens designed from years of experience and observations, the usual first-step in crystallization, upon obtaining pure protein sample, is still a trial and error experiment with a feasible number of conditions. In this project, we sought to re-think screening experiments, particularly to move away from the present dependency on crystallinity as the only meaningful readout. We sought to extract as much as possible from the experiment format that has become standard and has been historically effective; no modifications nor disruption to established workflow is required.

Central to the analysis of screening experiments is the need for objective description of crystallization droplets. We have developed image processing algorithms and a texton dictionary that are robust across imaging systems, to numerically describe droplets. Such descriptors remove the need for manual and subjective labelling, and allow for crystallization screening data to be used with established machine learning methods, including, but not limited to those described in this thesis. By using crystallization droplets to characterize an experiment, all aspects of the protein, including its preparation and storage, the interactions with ligands and/or precipitants are taken into account; what is being described is arguably closest to experimental reality.

We have introduced new methods to analyse screening experiments. First, by ranking and prioritising interesting droplets, presenting them in a more meaningful order for visual evaluation. We have shown that using computational tools to aid rather than replace the task of annotating images through classification algorithms, results in more accurate and efficient

evaluation of crystallization droplets without introducing the mistrust of black-box approaches. Secondly, we have formally defined precipitation fingerprints for a protein, derived from its behaviour across a standard sparse-matrix screen. Such fingerprints were found to be a consistent and reliable characterization of the protein, regardless of the presence of crystallinity. In pattern matching context, the fingerprints could be used to identify similar experiments, and exploit past successful conditions as potential optimization strategy for a new protein with similar behaviour, or form a framework for identifying its crystallizability. Thirdly, we have extended the idea of using use screening experiments as an optimum-buffer screen, by automatically identifying clear drops and mapping them to the sampled chemical space.

We have also retrospectively analysed the crystallization screening practices at the SGC over its first 10 years to evaluate strategies for the setup of screening experiments. Specifically, increasing redundancy of screening conditions, both by additional protein-precipitant mixing ratio and incubation temperatures were found to increase success rates. However, we could conclude that when sample is limited, priority should be given to the sampling of more conditions (*e.g.* using more screens for example), above resampling conditions.

In summary, this thesis has contributed to the community in the following ways:

1. The development of the texton dictionary with 300 unique entries, and the image processing algorithms to use it as a robust and effective description of crystallization droplets, independent of imaging system.

2. A new presentation method of droplets for visual evaluation through the ranking of crystallization droplets by their likelihood of crystallinity, in contrast to previous efforts' focus on replacing the evaluation task by the classification of crystallisation outcome.

3. A novel optimization strategy using the collective precipitation behaviour of a protein across standard sparse-matrix screens, independent of crystalline behaviour. Our method relies on the comparison of the precipitation fingerprint of a given protein objectively to historical experiments with similar fingerprints, and using past successful conditions as an optimization option and inference for crystallizability. Such comparison can also be used as a framework for determining the crystallizability of an experiment.

4. Automatic clear drop identification and mapping of clear drops to chemical components in sparse-matrix screens, to highlight components that may have stabilizing effect on the protein, suitable as ingredients for alternative protein formulation buffer.

5. The development and deployment of TeXRank, a vendor-independent image viewer that incorporates the new features and algorithm output developed directly for practical and everyday use. It is now the viewer-of-choice for most scientist at the SGC, has been deployed at NIBR, and is a cornerstone of the medium throughput fragment soaking pipeline at Diamond Light Source. It is freely downloadable, and will be part of a university-wide effort for the commercialization of academic software.

6. Suggestions on effective crystallization set-up strategies with limited protein samples, based on the crystallization practices at the SGC over a decade.

7. The development of algorithms combining superpixels and texton for the segmentation of different outcome occurring in a crystallization droplet.

It is our hope that the above contributions will enable the crystallization of more proteins in a more effective manner, and cast more attention on precipitates, the less-loved and often overlooked outcome in crystallization.

# Supplementary Materials

## S.1 Supplementary materials for Chapter 2.

*Table S 1: Crystallization conditions for Ligand Friendly Screen (LFS)*

| Row | Column | Condition |
|---|---|---|
| **A** | 1 | 30% PEG1000 -- 0.1M SPG pH 6.0 |
| **A** | 2 | 30% PEG1000 -- 0.1M SPG pH 7.0 |
| **A** | 3 | 30% PEG1000 -- 0.1M SPG pH 8.0 |
| **A** | 4 | 60% MPD -- 0.1M SPG pH 6.0 |
| **A** | 5 | 60% MPD -- 0.1M SPG pH 7.0 |
| **A** | 6 | 60% MPD -- 0.1M SPG pH 8.0 |
| **A** | 7 | 20% PEG6000 -- 10% ethylene glycol -- 0.2M sodium chloride |
| **A** | 8 | 20% PEG6000 -- 10% ethylene glycol -- 0.2M ammonium chloride |
| **A** | 9 | 20% PEG6000 -- 10% ethylene glycol -- 0.2M lithium chloride |
| **A** | 10 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M magnesium chloride |
| **A** | 11 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M calcium chloride |
| **A** | 12 | 20% PEG6000 -- 10% ethylene glycol -- 0.01M zinc chloride |
| **B** | 1 | 30% PEG1000 -- 0.1M MIB pH 6.0 |
| **B** | 2 | 30% PEG1000 -- 0.1M MIB pH 7.0 |
| **B** | 3 | 30% PEG1000 -- 0.1M MIB pH 8.0 |
| **B** | 4 | 60% MPD -- 0.1M MIB pH 6.0 |
| **B** | 5 | 60% MPD -- 0.1M MIB pH 7.0 |
| **B** | 6 | 60% MPD -- 0.1M MIB pH 8.0 |
| **B** | 7 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.2M sodium chloride |
| **B** | 8 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.2M ammonium chloride |
| **B** | 9 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.2M lithium chloride |
| **B** | 10 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.1M magnesium chloride |
| **B** | 11 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.1M calcium chloride |
| **B** | 12 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M MES pH 6.0 -- 0.01M zinc chloride |
| **C** | 1 | 30% PEG1000 -- 0.1M PCB pH 6.0 |

| | | |
|---|---|---|
| **C** | 2 | 30% PEG1000 -- 0.1M PCB pH 7.0 |
| **C** | 3 | 30% PEG1000 -- 0.1M PCB pH 8.0 |
| **C** | 4 | 60% MPD -- 0.1M PCB pH 6.0 |
| **C** | 5 | 60% MPD -- 0.1M PCB pH 7.0 |
| **C** | 6 | 60% MPD -- 0.1M PCB pH 8.0 |
| **C** | 7 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.2M sodium chloride |
| **C** | 8 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.2M ammonium chloride |
| **C** | 9 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.2M lithium chloride |
| **C** | 10 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.1M magnesium chloride |
| **C** | 11 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.1M calcium chloride |
| **C** | 12 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.01M zinc chloride |
| **D** | 1 | 30% PEG1000 -- 0.1M MMT pH 6.0 |
| **D** | 2 | 30% PEG1000 -- 0.1M MMT pH 7.0 |
| **D** | 3 | 30% PEG1000 -- 0.1M MMT pH 8.0 |
| **D** | 4 | 60% MPD -- 0.1M MMT pH 6.0 |
| **D** | 5 | 60% MPD -- 0.1M MMT pH 7.0 |
| **D** | 6 | 60% MPD -- 0.1M MMT pH 8.0 |
| **D** | 7 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.2M sodium chloride |
| **D** | 8 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.2M ammonium chloride |
| **D** | 9 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.2M lithium chloride |
| **D** | 10 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.1M magnesium chloride |
| **D** | 11 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.1M calcium chloride |
| **D** | 12 | 20% PEG6000 -- 10% ethylene glycol -- 0.1M tris pH 7.5 -- 0.01M zinc chloride |
| **E** | 1 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium fluoride |
| **E** | 2 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium bromide |
| **E** | 3 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium iodide |
| **E** | 4 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M potassium thiocyanate |
| **E** | 5 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium nitrate |
| **E** | 6 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium formate |
| **E** | 7 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium acetate |
| **E** | 8 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium sulfate |
| **E** | 9 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium/potassium tartrate |
| **E** | 10 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium/potassium phosphate |
| **E** | 11 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M potassium citrate tribasic |
| **E** | 12 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium malonate |
| **F** | 1 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium fluoride |
| **F** | 2 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium bromide |
| **F** | 3 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium iodide |

| | | |
|---|---|---|
| **F** | 4 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M potassium thiocyanate |
| **F** | 5 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium nitrate |
| **F** | 6 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium formate |
| **F** | 7 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium acetate |
| **F** | 8 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium sulfate |
| **F** | 9 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium/potassium tartrate |
| **F** | 10 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.02M sodium/potassium phosphate |
| **F** | 11 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M potassium citrate tribasic |
| **F** | 12 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium malonate |
| **G** | 1 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium fluoride |
| **G** | 2 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium bromide |
| **G** | 3 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium iodide |
| **G** | 4 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M potassium thiocyanate |
| **G** | 5 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium nitrate |
| **G** | 6 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium formate |
| **G** | 7 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium acetate |
| **G** | 8 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium sulfate |
| **G** | 9 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium/potassium tartrate |
| **G** | 10 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.02M sodium/potassium phosphate |
| **G** | 11 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M potassium citrate tribasic |
| **G** | 12 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium malonate |
| **H** | 1 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium fluoride |
| **H** | 2 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium bromide |
| **H** | 3 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium iodide |
| **H** | 4 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M potassium thiocyanate |
| **H** | 5 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium nitrate |
| **H** | 6 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium formate |
| **H** | 7 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium acetate |
| **H** | 8 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium sulfate |
| **H** | 9 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M sodium/potassium tartrate |
| **H** | 10 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.02M sodium/potassium phosphate |
| **H** | 11 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 8.5 -- 0.2M potassium citrate tribasic |

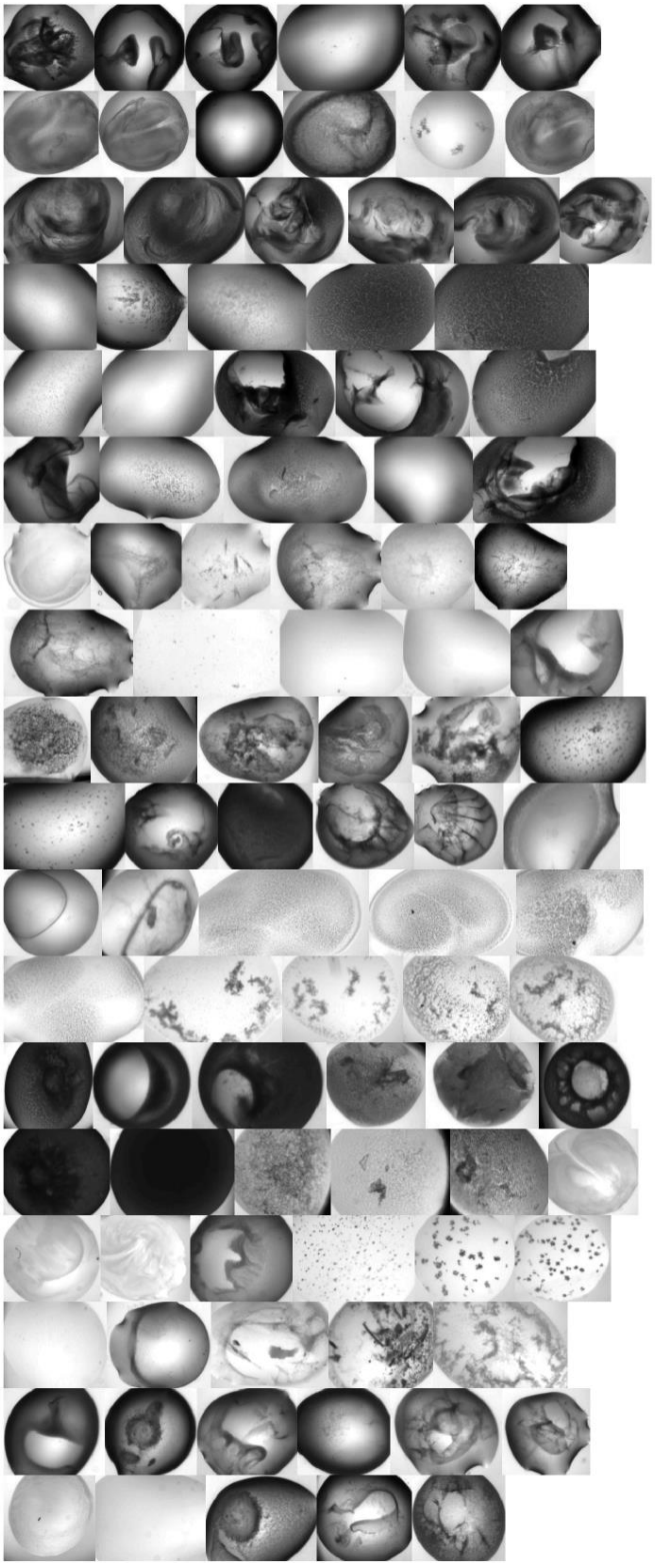


*Figure S 1: The 100 precipitation images used to develop the texton dictionary. Figure taken from Supplementary Materials of Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

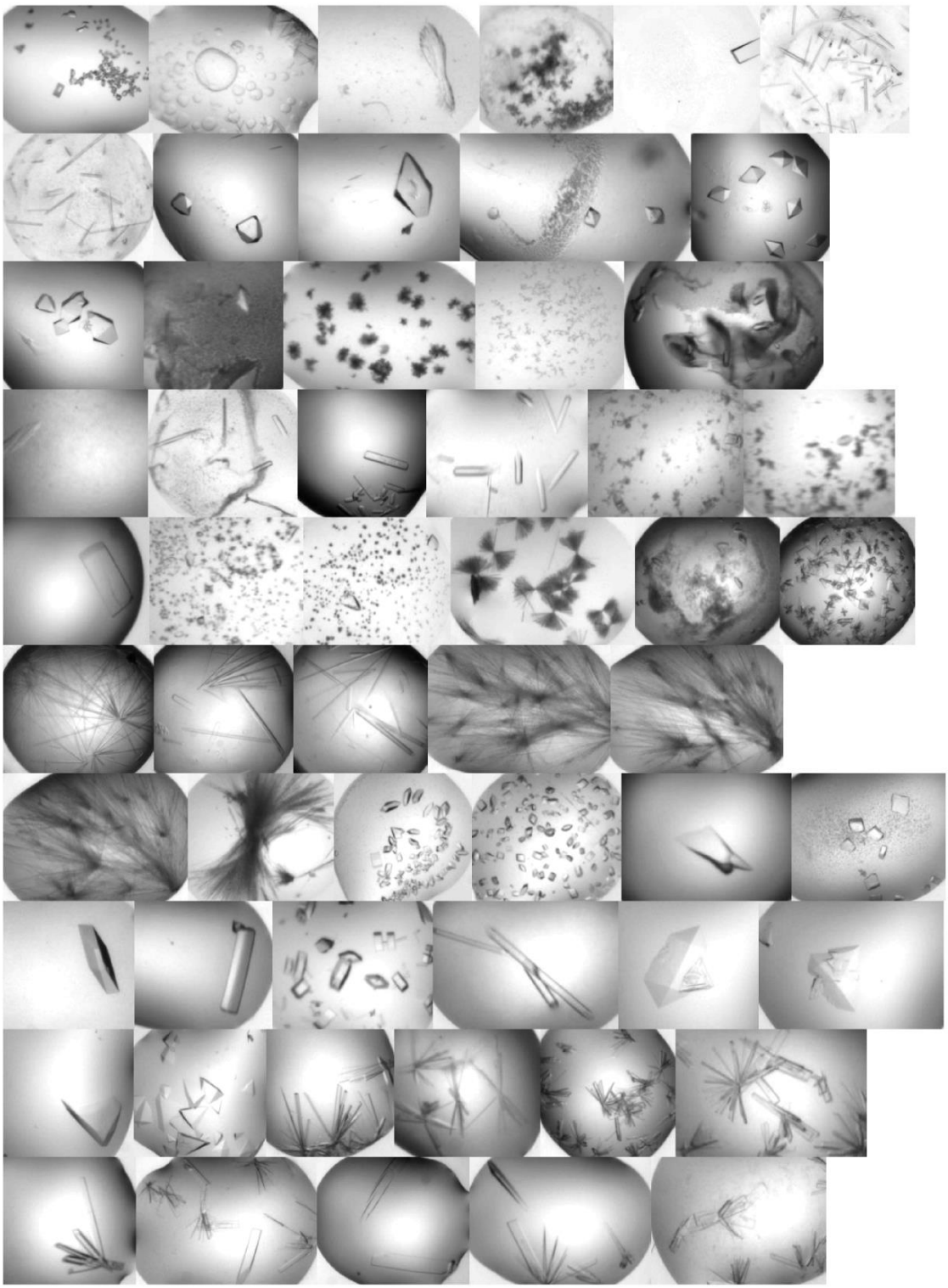


*Figure S 2: The 52-crystal-containing images used to develop the texton dictionary. 61 textons corresponding to crystal regions were derived from the filter responses of these images. Figure taken from Supplementary Materials of Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

## S.2 Supplementary materials for Chapter 3

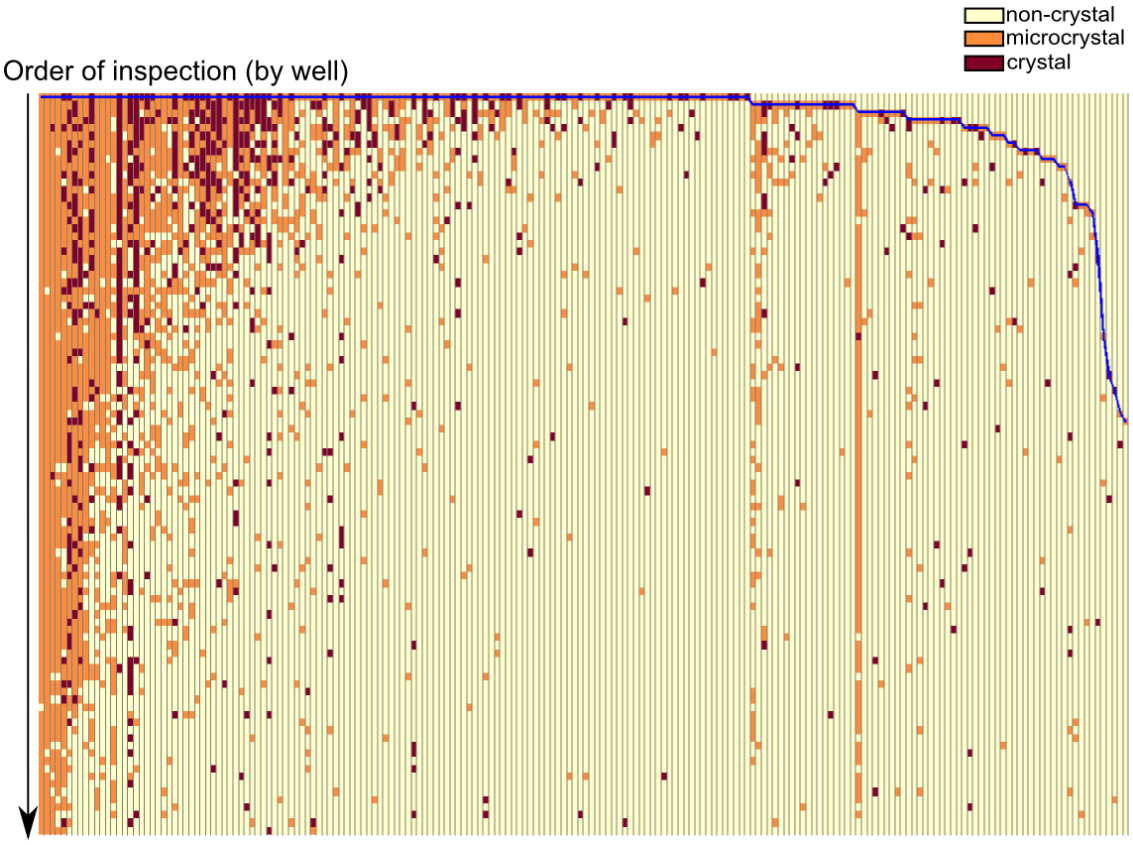



*Figure S 3: Position of all human-scored crystal images in the test set of 196 plates. Each column represents a plate and each row represents a well to be viewed. The blue curve marks the highest rank of human-scored crystal image for each plate. Figure taken from Supplementary Materials of Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

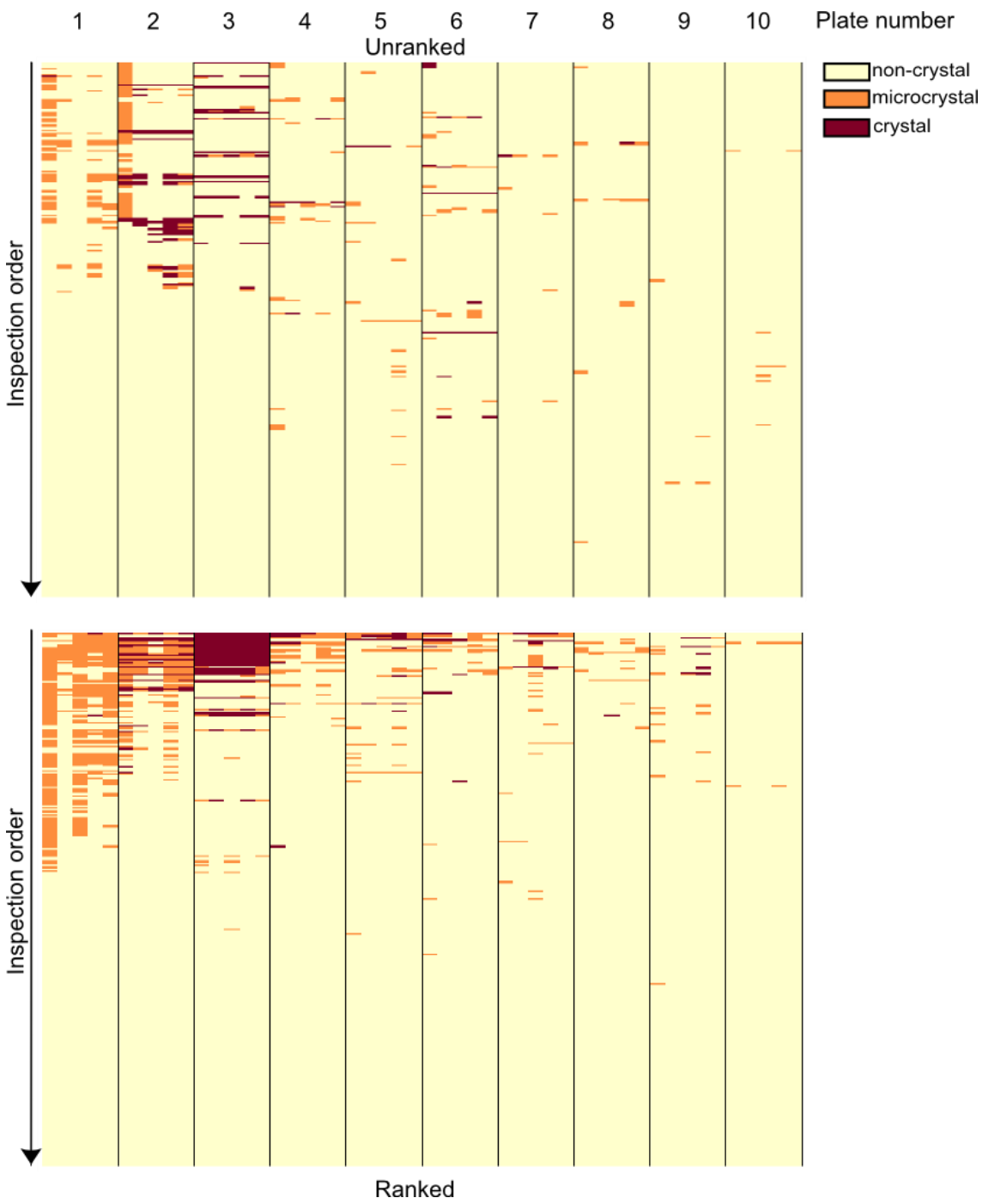



*Figure S 4: Individual annotations by the ten crystallographers, as represented by the coloured sub-columns. The top figure shows the annotations of the group of 5 crystallographers when viewing the plates in an unranked order, while the bottom figure shows the annotations of the crystallographers when they viewed the plates in the ranked order. The majority vote was used for Figure 3.5, and if there was a tie, the following priority was used: crystal, micro-crystal, non-crystal. Figure taken from Supplementary Materials of Ng et al. (2014), reproduced with permission of the International Union of Crystallography.*

## S.3 Supplementary materials for Chapter 4

*Table S 2: Further information on protein samples used in Chapter 4, listed in Table 4.3 and Table 4.4. See Table S3 for description of expression vectors and Table S4 for sequence information and construct positions.*

| Purification ID | Target Description | MW (kDa) | Expected pI | Buffer pH | Tag cleaved? | Expression system | Expression Cell Line | Expression vector |
|---|---|---|---|---|---|---|---|---|
| ATAD2A-p033 | Two AAA domain containing protein | 15.49 | 4.92 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| BRD1A-p040 | Bromodomain containing protein 1 | 15.73 | 5.45 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| BTBD12B-p003 | BTB/POZ Domain-containing protein | 16.69 | 4.95 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| BTBD12B-p004 | BTB/POZ Domain-containing protein | 16.74 | 5.25 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| BTBD12B-p007 | BTB/POZ Domain-containing protein | 16.69 | 5.15 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| BTBD12B-p008 | BTB/POZ Domain-containing protein | 16.77 | 5.45 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| CAMK1DA-p014 | Calcium/calmodulin-dependent protein kinase ID | 36.65 | 6.25 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| CECR2A-p029 | Cat eye syndrome chromosome region, candidate 2 | 13.71 | 4.95 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| CECR2A-p030 | Cat eye syndrome chromosome region, candidate 2 | 16.17 | 5.55 | 7.5 | no | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| CECR2A-p031 | Cat eye syndrome chromosome region, candidate 2 | 13.71 | 4.95 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| DACASAA-p001 | Diadenylate cyclase found in Staphylococus aureus | 33.75 | 6.15 | 8 | no | Bacteria | BL21(DE3)-R3 | pET-SUMO1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DOPVA-p002 | Deamidase-depupylase Dop of the Prokaryotic Ubiquitin-like Modification Pathway | 54.56 | 5.19 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | Custom |
| EP300A-p031 | EP300: E1A binding protein p300 | 13.81 | 4.55 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| FAM83AA-p007 | Family with sequence similarity 83, member A | 21.14 | 9.75 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| GADD45BA-p001 | Growth arrest and DNA-damage-inducible, beta | 17.43 | 4.15 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| GYG2A-p032 | Glycogenin 2 isoform a | 31.15 | 6.75 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC-CTHF |
| GYG2A-p033 | Glycogenin 2 isoform a | 31.15 | 6.75 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC-CTHF |
| JARID1BA-p070 | Jumonji, AT rich interactive domain 1B (RBP2-like) | 57.21 | 5.05 | 7.5 | yes | Baculo | DH10Bac | pFB-LIC-Bse |
| JARID1BA-p082 | Jumonji, AT rich interactive domain 1B (RBP2-like) | 55.15 | 5.05 | 7.5 | yes | Baculo | DH10Bac | pFB-LIC-Bse |
| JARID1BA-p092 | Jumonji, AT rich interactive domain 1B (RBP2-like) | 55.40 | 5.15 | 7.5 | yes | Baculo | DH10Bac | pFB-LIC-Bse |
| JARID2A-p024 | Jumonji, AT rich interactive domain 2 protein | 87.52 | 9.45 | 8.5 | no | Baculo | DH10Bac | pFB-CT10HF-LIC |
| JARID2A-p026 | Jumonji, AT rich interactive domain 2 protein | 83.24 | 9.45 | 8 | no | Baculo | DH10Bac | pFB-LIC-Bse |
| JMJD2AA-p086 | Jumonji domain containing 2A | 15.80 | 4.55 | 7.4 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| JMJD2AA-p092 | Jumonji domain containing 2A | 41.89 | 7.65 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| JMJD2AA-p097 | Jumonji domain containing 2A | 41.82 | 8.45 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| JMJD2AA-p100 | Jumonji domain containing 2A | 13.34 | 4.05 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| JMJD2BA-p049 | Jumonji domain containing 2B | 20.63 | 6.05 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| JMJD2CA-p067 | Jumonji domain containing 2C | 13.54 | 4.45 | 7.5 | yes | Bacteria | BL21(DE3)-R3 | pNIC28-Bsa4 |
| JMJD2DA-p037 | Jumonji domain containing 2D | 39.55 | 9.15 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| LACTB2A-p118 | Lactamase, beta 2 | 35.30 | 6.45 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| OTUB2A-p006 | OTU domain, ubiquitin aldehyde binding 2 | 28.19 | 5.75 | 8 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | Mock-receptable-vector |
| PAHA-p006 | Phenylalanine hydroxylase | 10.87 | 5.65 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| PHIPA-p022 | Pleckstrin homology domain interacting protein | 16.11 | 7.65 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| PHIPA-p023 | Pleckstrin homology domain interacting protein | 17.60 | 6.05 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| PRKCBP1A-p005 | Protein kinase C binding protein 1 [isoform a] | 37.78 | 8.45 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC-ZB |
| PRKCBP1A-p023 | Protein kinase C binding protein 1 [isoform a] | 37.88 | 8.01 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC-ZB |
| PXKA-p011 | PX domain containing serine/threonine kinase | 54.82 | 9.85 | 7.5 | yes | Bacteria | Mach1 | pNIC-CTH0 |
| SPIN3A-p013 | Spindlin-3 | 24.20 | 4.45 | 7.2 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| STK6A-p009 | Serine/threonine kinase 6 [isoform 1; tv2] | 32.90 | 9.55 | 7.5 | yes | Bacteria | BL21(DE3)-R3-lambda-PPase | pETM-11 |
| VPS28A-p003 | Vacuolar protein sorting-associated protein 28 homolog isoform 1 | 12.06 | 4.35 | 7.5 | yes | Bacteria | BL21(DE3)-R3-pRARE2 | pNIC28-Bsa4 |
| ZAKA-p026 | Sterile-alpha motif and leucine zipper containing kinase AZK | 35.11 | 5.75 | 7.5 | yes | Baculo | DH10Bac | pFB-LIC-Bse |
| ZFYVE9C-p011 | Zinc finger FYVE domain-containing protein 9 isoform 3 | 8.48 | 6.15 | 7.5 | no | Bacteria | BL21(DE3)-R3-pRARE2 | pCDF-LIC |

*Table S 3: Description of expression vectors used in the production of protein samples for Chapter 4.*

| Expression vector | Comments |
|---|---|
| Mock-receptable-vector | For Mock cloning of imported constructs |
| pCDF-LIC | Coexpression vector compatible with pET and pACYC, Bsa-LIC cassette in forst ORF, second ORF NdeI/XhoI as in pCDF-duet-CDS2; Spectinomycin-resistant. |
| pET-SUMO1 | Mock vector (Owner: David Damerell) |
| pETM-11 | Custom (Owner: Sarah Picaud) |
| pFB-CT10HF-LIC | Baculo vector with C-terminal 10His+flag, TEV-cleavable, for LIC cloning. |
| pFB-LIC-Bse | Baculovirus transfer vector (Bac-to-bac), N-terminal His-tag, TEV cleavable. |
| pNIC-CTH0 | C-terminal His, TEV-cleavable, for LIC cloning. EXP-10-AE0544 |
| pNIC-CTHF | C-terminal His+flag, TEV-cleavable, for LIC cloning. EXP-06-AA9343 |
| pNIC-ZB | Z-Basic solubility-enhancing tag, normal LIC cloning, TEV cleavage. Pavel EXP-09-AC5269 |
| pNIC28-Bsa4 | T7/lac regulated, N-terminal His-tag, TEV, LIC cloning using BsaI cleavage/T4 polymerase, SacB stuffer fragment, pET28 backbone. |

*Table S 4: Sequence information and construct position for protein samples used.*

| Purification ID | RefSeq NP (Protein) ID | Construct Position | |
|---|---|---|---|
| | | Start | Stop |
| **ATAD2A-p033** | NP_054828.2 | Q981 | R1108 |
| **BRD1A-p040** | NP_055392.1 | E556 | A688 |
| **BTBD12B-p003** | NP_115820.2 | G668 | E796 |
| **BTBD12B-p004** | NP_115820.2 | G668 | E796 |
| **BTBD12B-p007** | NP_115820.2 | G668 | E796 |
| **BTBD12B-p008** | NP_115820.2 | G668 | E796 |
| **CAMK1DA-p014** | NP_705718.1 | S10 | D329 |
| **CECR2A-p029** | NP_113601.2 | T430 | D543 |
| **CECR2A-p030** | NP_113601.2 | T430 | D543 |
| **CECR2A-p031** | NP_113601.2 | T430 | D543 |
| **DACASAA-p001** | WP_031765219.1 | E88 | K269 |
| **DOPVA-p002** | WP_003895348.1 | M50 | T550 |
| **EP300A-p031** | NP_001420.2 | I1048 | G1161 |
| **FAM83AA-p007** | NP_996889.1 | A122 | L304 |
| **GADD45BA-p001** | NP_056490.2 | L6 | R160 |
| **GYG2A-p032** | NP_003909.2 | T4 | A269 |
| **GYG2A-p033** | NP_003909.2 | T4 | A269 |
| **JARID1BA-p070** | NP_006609.3 | M138 | I907 |
| **JARID1BA-p082** | NP_006609.3 | F163 | I907 |
| **JARID1BA-p092** | NP_006609.3 | F163 | I907 |
| **JARID2A-p024** | NP_004964.2 | A502 | S1246 |
| **JARID2A-p026** | NP_004964.2 | Q537 | S1246 |
| **JMJD2AA-p086** | NP_055478.2 | Q897 | P1011 |
| **JMJD2AA-p092** | NP_055478.2 | M1 | L359 |
| **JMJD2AA-p097** | NP_055478.2 | M1 | L359 |
| **JMJD2AA-p100** | NP_055478.2 | Q897 | P1011 |
| **JMJD2BA-p049** | NP_055830.1 | S908 | E1070 |
| **JMJD2CA-p067** | NP_055876.2 | C875 | P991 |
| **JMJD2DA-p037** | NP_060509.2 | M1 | R342 |
| **LACTB2A-p118** | NP_057111.1 | M1 | L288 |
| **OTUB2A-p006** | NP_075601.1 | M1 | H234 |
| **PAHA-p006** | NP_000268.1 | G19 | R111 |
| **PHIPA-p022** | NP_060404.3 | D1302 | R1434 |
| **PHIPA-p023** | NP_060404.3 | S1315 | R1440 |
| **PRKCBP1A-p005** | NP_898869.1 | Q103 | S426 |
| **PRKCBP1A-p023** | NP_898869.1 | Q103 | S426 |
| **PXKA-p011** | NP_060241.2 | M4 | A471 |
| **SPIN3A-p013** | NP_001010862.2 | V50 | S258 |
| **STK6A-p009** | NP_940839.1 | E122 | S403 |
| **VPS28A-p003** | NP_898880.1 | D120 | A221 |
| **ZAKA-p026** | NP_598407.1 | G5 | K309 |
| **ZFYVE9C-p011** | NP_015563.2 | M767 | E824 |

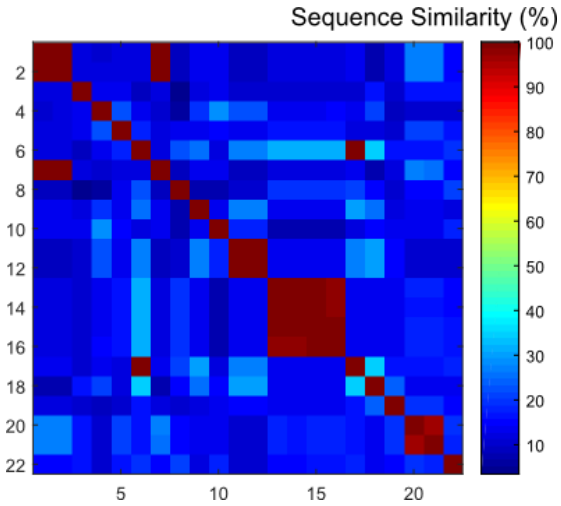



*Figure S 5: Sequence similarity between protein samples in Table 4.3 and Figure 4.8. Sequence similarity were calculated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/), using sequences of protein with or without tags as specified in Table S2.*

*Table S 5: Base conditions used in follow-up screens for each protein sample. Italicised conditions were generated by swapping components between base conditions. Components separated by '/' indicate consolidated base conditions, where the wells in the grid-design contained either one of the components in the list. Conditions marked with * were the positive control conditions, whereby these conditions have been previously identified to produce crystals for the respective targets*

| **Protein Sample** | **Base conditions and positive control in Follow-Up Screen.** |
|---|---|
| ATAD2A-p033 | 30% PEG1000 -- 0.1M SPG pH 8.0<br>20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.2M sodium chloride<br>20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium fluoride -- 0.1M bis-tris-propane pH 7.5/no buffer<br>25% PEG3350 -- 0.2M sodium chloride -- 0.1M HEPES pH 7.5<br>20% PEG3350 -- 0.2M magnesium formate<br>*Random*<br>0.1M bis-tris pH 6.7 -- 0.2M magnesium chloride -- 28% PEG3350 * |
| BRD1A-p040 | 20% PEG3350 -- 0.2M sodium malonate<br>20% PEG3350 -- 0.2M ammonium chloride<br>25% PEG3350 -- 0.2M ammonium sulfate -- 0.1M bis-tris pH 5.5 |

| | |
|---|---|
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 25% PEG3350 -- 0.2M sodium chloride -- 0.1M bis-tris pH 6.5 |
| | 25% PEG3350 -- 0.2M sodium chloride -- 0.1M HEPES pH 7.5 |
| | 30% jeffamine ED-2003 -- 0.1M HEPES pH 7.0 |
| | 30% PEG500MME -- 0.05M calcium chloride -- 0.1M bis-tris pH 6.5 |
| | 25% PEG4000 -- 0.15M ammonium sulfate -- 15% glycerol |
| | 30% PEG2000MME -- 0.1M potassium thiocyanate |
| | 30% PEG2000MME -- 0.1M potassium bromide |
| | *Random* |
| | 25% PEG3350 -- 0.1M bis-tris pH 6.5 * |
| GYG2A-p033 | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 25% PEG3350 -- 0.2M sodium chloride -- 0.1M HEPES pH 7.5 |
| | 1.6M magnesium sulfate -- 0.1M MES pH 6.5 |
| | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 7.5 -- 0.2M sodium bromide/iodide/malonate/chloride |
| | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5/7.5/8.5 -- 0.2M sodium bromide |
| | *Random* |
| | 0.5M magnesium formate -- 0.1M HEPES pH 7.5 * |
| JMJD2AA-p092 | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 10% PEG3350 -- 0.2M L-Proline -- 0.1M HEPES pH 7.5 |
| | 20% PEG3350 -- 0.2M sodium malonate/ammonium chloride |
| | 30% PEG5000MME -- 0.2M ammonium sulfate --0.1M MES pH 6.5 |
| | 30% PEG8000 -- 0.2M ammonium sulfate -- 0.1M cacodylate pH 6.5 |
| | 25% PEG3350 -- 0.2M ammonium sulfate -- 0.1M tris pH 8.5 |
| | 25% PEG3350 -- 0.2M ammonium acetate -- 0.1M bis-tris pH 6.5 * |
| JMJD2AA-p097 | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium nitrate |
| | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium malonate |
| | 20% PEG3350 -- 0.2M ammonium chloride |
| | 25% PEG3350 -- 0.1M tris pH 9.5-- 0.2M ammonium acetate/lithium sulfate/ammonium sulfate/sodium nitrate |
| | 25% PEG3350 -- 0.1M tris pH 7.5 -- 0.2M ammonium acetate/lithium sulfate/ammonium sulfate/sodium nitrate |
| | *Random* |
| | 0.2M lithium sulfate -- 25% PEG3350 -- 0.1M bis-tris pH 6.5 * |
| JMJD2DA-p037 | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 10% PEG3350 -- 0.2M L-Proline -- 0.1M HEPES pH 7.5 |
| | 20% PEG3350 -- 0.2M sodium malonate/ammonium chloride |
| | 30% PEG5000MME -- 0.2M ammonium sulfate --0.1M MES pH 6.5 |
| | 30% PEG8000 -- 0.2M ammonium sulfate -- 0.1M cacodylate pH 6.5 |
| | 25% PEG3350 -- 0.2M ammonium sulfate -- 0.1M tris pH 8.5 |
| | 25% PEG3350 -- 0.2M ammonium acetate -- 0.1M HEPES pH 7.5 * |
| PRKCBP1A-p023 | 10% PEG6000 -- 2M sodium chloride |
| | *10% PEG6000 -- 2M lithium sulfate* |
| | *20% PEG6000 -- 0.2M sodium chloride -- 0.1M bis-tris pH 5.5* |

| | |
|---|---|
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | *25% PEG3350 -- 0.2M sodium chloride -- 0.1M bis-tris pH 5.5* |
| | *25% PEG3350 -- 0.2M lithium sulfate* |
| | 0.2M lithium sulfate -- 25% PEG3350 -- 0.1M tris pH 8.5 * |
| CAMK1DA-p014 | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 20% PEG3000 -- 0.1M citrate pH 5.5 |
| | 1M 1,6-hexanediol -- 0.01M cobalt chloride -- 0.1M acetate pH 4.5 |
| | 30% MPD -- 0.02M calcium chloride -- 0.1M acetate pH 4.5 |
| | 45% MPD -- 0.2M calcium chloride -- 0.1M bis-tris pH 5.5 |
| | *Random* |
| | 0.2M lithium sulfate -- 25% PEG3350 -- 0.1M bis-tris pH 6.5 * |
| GYG2A-p032 | 20% PEG3350 -- 0.2M magnesium formate |
| | 25% PEG3350 -- 0.2M magnesium chloride -- 0.1M tris pH 8.5 |
| | 0.5M magnesium formate -- 0.1M bis-tris pH 6.5 |
| | 15% tacsimate -- 2% PEG3350 -- 0.1M HEPES pH 7.0 |
| | *15% tacsimate -- 2% PEG3350 -- 0.1M HEPES pH 7.0 -- 0.2M magnesium formate/chloride* |
| | *Random* |
| | 0.5M magnesium formate -- 0.1M HEPES pH 7.5 * |
| CECR2A-p031 | 50% PEG200 -- 0.2M magnesium chloride -- 0.1M cacodylate pH 6.5 |
| | 20% PEG8000 -- 0.2M magnesium chloride -- 0.1M tris pH 8.5 |
| | 20% PEG3350 -- 0.2M sodium formate |
| | 25% PEG3350 -- 0.1M bis-tris pH 6.5 |
| | 25% PEG3350 -- 0.2M magnesium chloride -- 0.1M bis-tris pH 6.5/HEPES pH 7.5/tris pH 8.5/cacodylate pH 6.5 |
| | *Random* |
| | 3M sodium chloride -- 0.1M bis-tris pH 5.5 * |
| JMJD2AA-p100 | 25% PEG3350 -- 0.2M sodium formate -- 0.1M bis-tris pH 5.5 |
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M bis-tris pH 5.5 |
| | 25% PEG3350 -- 0.2M magnesium chloride -- 0.1M bis-tris pH 6.5 |
| | *20% PEG3350 -- 0.2M sodium formate* |
| | *20% PEG3350 -- 0.2M magnesium chloride* |
| | *20% PEG3350 -- 0.2M lithum sulfate* |
| | 60% tacsimate * |
| JMJD2CA-p067 | 0.8M succinic acid |
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M tris pH 8.5 |
| | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium fluoride/sodium formate |
| | 15% PEG3350 -- 0.1M magnesium formate |
| | *15% PEG3350 -- 0.2M sodium fluoride/litium sulfate/sodium formate* |
| | *Random* |
| | 0.2M lithium sulfate -- 25% PEG3350 -- 0.1M bis-tris pH 5.5 * |
| VPS28A-p003 | 2.8M sodium acetate |
| | 20% PEG3350 -- 0.2M ammonium citrate dibasic |
| | 25% PEG3350 -- 0.2M sodium chloride -- 0.1M HEPES pH 7.5 |

| | |
|---|---|
| | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium fluoride -- (0.1M bis-tris-propane pH 7.5) |
| | 20% PEG6000 -- 10% ethylene glycol -- 0.1M HEPES pH 7.0 -- 0.2M sodium chloride |
| | 10% PEG6000 -- 2M sodium chloride |
| | 30% PEG1000 -- 0.1M SPG pH 8.0 |
| | 20% PEG2000MME -- 0.2M trimethylamine N-oxide -- 0.1M tris pH 8.5 |
| | 0.5% PEG5000MME -- 0.8M sodium/potassium tartrate -- 0.1M tris pH 8.5 |
| | 22% polyacrylic acid 5100 -- 0.02M magnesium chloride -- 0.1M HEPES pH 7.5 |
| | 0.5% jeffamine ED-2003 -- 1.2M sodium malonate -- 0.1M HEPES pH 7.0 |
| | 20% PEG3350 -- 0.2M magnesium formate |
| | 20% PEG3350 -- 10% ethylene glycol -- 0.1M bis-tris-propane pH 6.5 -- 0.2M sodium iodide * |
| FAM83AA-p007 | 20% PEG3350 -- 10% ethylene glycol -- 0.2M sodium iodide |
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M tris pH 8.5 |
| | 15% tacsimate -- 2% PEG3350 -- 0.1M HEPES pH 7.0 |
| | 3M sodium chloride -- 0.1M bis-tris pH 5.5/tris pH 8.5 |
| | 0.8M succinic acid |
| | *Random* |
| | 0.2M ammonium acetate -- 25% PEG3350 -- 0.1M bis-tris pH 5.5 * |
| DOPVA-p002 | 20% PEG3350 -- 0.1M succinic acid |
| | 20% PEG3350 -- 0.2M potassium nitrate |
| | 25% PEG3350 -- 0.2M magnesium chloride -- 0.1M tris pH 8.5 |
| | 25% PEG3350 -- 0.2M lithium sulfate -- 0.1M tris pH 8.5 |
| | 3M sodium chloride -- 0.1M tris pH 8.5 |
| | 60% tacsimate (not shown on graph); 0.8M succinic acid |
| LACTB2A-p118 | 15% PEG3350 -- 0.1M magnesium formate |
| | 20% PEG3350 -- 0.1M succinic acid |
| | 25% PEG3350 -- 0.2M sodium chloride -- 0.1M tris pH 8.5 |
| | 10% PEG5000MME -- 5% tacsimate -- 0.1M HEPES pH 7.0 |
| | 10% PEG8000 -- 0.2M magnesium chloride -- 0.1M HEPES pH 7.0 |
| | 20% PEG3350 -- 0.2M ammonium citrate dibasic; 60% tacsimate |
| OTUB2A-p006 | 3.5M sodium formate |
| | 12% PEG20000MME -- 0.1M MES pH 6.5 |
| | 30% PEG5000MME -- 0.2M ammonium sulfate -- 0.1M MES pH 6.5 |
| | 25% PEG3350 -- 0.2M ammonium sulfate -- 0.1M HEPES pH 7.5 |
| | 25% PEG3350 -- 0.2M sodium chloride/magnesium chloride/ammonium acetate/ammonium sulfate -- 0.1M HEPES pH 7.5 |
| | 20% PEG3350 -- 0.2M sodium malonate (10% ethy glycol -- 0.1M bis-tris-propane pH 7.5) |