University of BRISTOL

Department of Computer Science

Using Image Texture Analysis to Increase Protein Solubility and Detect Fluorescent Protein Crystals

Oliver Nicholas Fiske King

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering

September 2016 | CSMSC-16

Declaration:

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Oliver Nicholas Fiske King, September 2016

Using Image Texture Analysis to Increase Protein Solubility and Detect Fluorescent Protein Crystals

Executive Summary

Growing a protein crystal is the first step in determining the 3D atomic structure of that protein using X-ray crystallography. Knowing the structure, in turn, gives insight into how these molecules function, how changes (mutations) such as those associated with cancers, affect their function and how drugs can be made to target specific proteins in order to treat disease. Growing a crystal is difficult and unlikely. To improve their chances, a crystallographer may set up hundreds of microscopic experiments. In each experiment, the protein of interest is mixed with a different 'cocktail' of chemical components that has been selected from a library of these cocktails. The experiments are then imaged using a robotic system and the micrographs either inspected manually or analysed computationally.

TeXRank is software, developed by, and currently in use at, the Structural Genomics Consortium (SGC) at the University of Oxford. It uses texture analysis to score images for the likely presence of crystals and then displays these images to the experimenter in rank order. This reduces the number of images that have to be viewed before finding a crystal. This project builds upon this work with the aim of providing new tools to the experimenter that have the potential to improve their chances of successfully growing or identifying a crystal of a given protein.

In particular, the project includes:

- methodology to find a set of chemical cocktail conditions that can be used as a tool to improve the solubility of sparingly soluble proteins (a limiting factor in obtaining crystals)
- a 'virtual' experiment to evaluate the potential usefulness of the solubility tool
- adaptation of image processing and machine learning methods to allow TeXRank to use microscope images, captured under green light, of experiments containing protein labelled with a fluorescent dye
- an evaluation of the usefulness of the novel combination of texture analysis with fluorescence labelling, both in comparison to standard imaging and to studies in the literature.

The main outcomes of this project were:

- a set of 48 chemical cocktail conditions, selected through mining the results of image analysis on real experimental data and found to be significantly better as solubility tool than a random set of conditions ($p < 0.05$)
- adaptations to the TeXRank processing pipeline and the training of a new classifier that allows the ranking of images from experiments containing fluorescent labelled protein
- a comparison of ranking performance between pairs of images taken of the same experiments, proving that the use of fluorescence in combination with texture analysis is better than texture analysis alone
- the fluorescence classifier created in this project also has superior ranking characteristics (area under receiver operating characteristic curve) to the current classifier used by TeXRank
- comparison to the research literature determined that texton analysis was more accurate at detecting fluorescent protein crystals than a recent study that utilised image thresholding followed by edge detection

Contents

List of Figures

Acknowledgements

Firstly, I would like to thank my parents for their support throughout the last year of my studies.

Many thanks to Prof. Frank von Delft for allowing me do this work, for being enthusiastic, and for arranging a desk so that I could have a routine and people to eat lunch with...

Which brings me to everyone else at SGC, Oxford for being so warm and welcoming, particularly Beth MacLean and the rest of the Protein Crystallography group.

Dr. Marc L. Pusey for running from computer to computer to transfer all the images of labelled protein experiments across the pond and Dr. Jia-Tsing Ng for helpful discussions.

In Bristol, Prof. Julian Gough for advice and agreeing to supervise me and Dr. Tim Kovacs for pastoral support.

Finally, to Venetia, for your patience, understanding and cups of strong tea.

Chapter 1

Research Area Overview and Introduction

1.1 X-ray Crystallography

X-ray crystallography is a technique used to determine the three-dimensional position of the atoms in a molecule. A crystal of the molecule is grown and then rotated in an X-ray beam. The patterns produced by the X-rays being diffracted by the layers of atoms in the crystal are computationally combined and used to determine the position of the atoms relative to one another. This is the main technique used to determine protein structure. The limiting factor in this process is growing the crystals in the first place. Although growing crystals of small molecules, such as copper sulphate, is relatively easy, growing crystals of large molecules, such as proteins, is difficult. To form a crystal, thousands of these molecules, each of which, in turn, contain thousands of atoms, have to pack together in a regular way, purely because it is energetically favourable given the conditions that they are subjected to (McPherson & Gavira, 2014). This is statistically very improbable. This unlikelihood of getting a protein to crystallise is further amplified by other variables, for example:

- Purifying proteins to a purity high enough to enter crystallisation experiments is challenging and expensive.
- A high concentration of protein in solution is needed in order to be able to grow large, high-quality crystals (Izaac, Schall, *et al.,* 2006). Even knowing the sequence of the amino acids that make up the protein, it is hard to predict how soluble it will be and whether this solubility can be increased by adding components such as salts or adjusting the pH of the solution.

1.2 Precipitation and Crystallisation

Due to the nature of problems described above, crystallographers perform hundreds, often thousands, of experiments, in order to increase their chances of success. In each of these experiments, a solution containing the protein is mixed with a mixture of chemical components termed a *cocktail* in order to form a *drop* or *droplet*. Generally, the cocktails are selected from commercial collections of components that have been collated based upon what has been reported, in the literature, to have been successful for other proteins (Jancarik & Kim, 1991). These sets of (normally 96) different chemical cocktails are termed *sparse-matrix screens*. It is hoped that the components will push the protein out of solution in a gentle way that favours crystal growth rather than simply causing it to crash out of solution as an unstructured *amorphous precipitate* (McPherson & Gavira, 2014).

Although amorphous precipitate in a drop is a bad sign, there is such a thing as good precipitate. Sometimes a precipitate may be made up of, or contain, microscopic crystals (*crystalline precipitate*) which then may go on to grow large enough to be of use. Even if these *micro-crystals* do not grow, the cocktail of chemicals that led to their formation can be adjusted and optimised in further experiments, possibly leading to bigger, useable crystals. Therefore, each experiment in a set of experiments, contains information that is potentially useful, even if the goal of obtaining a crystal has not been reached. Further to this, the collective set of many such experiments contains information that defines a protein in terms of its precipitation behaviour in these chemical conditions. This is termed the *precipitation fingerprint* of a protein (Ng, 2015).

1.3 Experimental Setup

The protein is a precious commodity, and as a result, the experiments are set up on a small scale, often using a liquid-dispensing robot. A typical combined drop volume of the protein and cocktail mix in this case would be 150 nl (150×10^{-9} l). The drops are set up in transparent plastic containers called *crystal plates*. A standard crystal plate has 8 rows × 12 columns = 96 positions, each containing a different cocktail. In addition, at each position, there are three *subwells*, each of which actually contains the experimental drop made up from protein and cocktail solution mixed together in a certain ratio (typically protein:cocktail of 1:2, 1:1 and 2:1). This means that each crystal plate contains 96 × 3 = 288 drops. A layout diagram for a crystal plate can be seen in Figure 1.1.

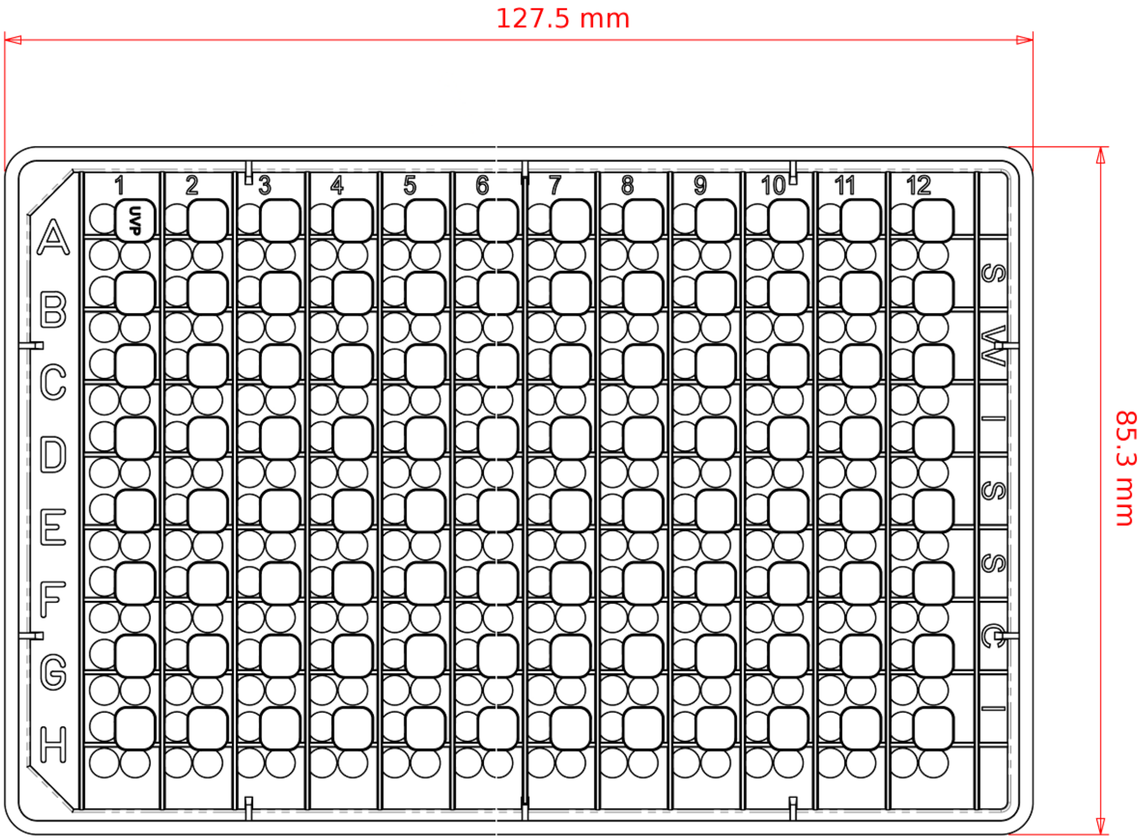


Figure 1.1: Layout diagram for a 3 drop crystallisation plate made by the company Swissci. The plate contains 96 wells which hold the crystallisation cocktail (rounded squares). Each well is surrounded by 3 subwells (circles), where the crystal drops sit. Adapted from information on the *Hampton Research Website* 2016.

The tiny experimental droplets can only be viewed under a microscope. However, since it is not known how long it will take for the protein to crystallise, if at all, the drops need to be inspected multiple times over a period that can stretch to several months. As a result, the crystal plates are often stored in an incubator and retrieved periodically by a robotic arm that transfers them to a digital imaging microscope which records a micrograph for each drop in turn. The crystallographer can then view the images of the experiments in sequence on a computer and annotate any image in which crystalline behaviour is observed with a number (usually an integer in the range 1 to 10) which is a key to a category such as ‘crystalline precipitate’, ‘dubious’ crystal or ‘hell-fire’ crystal (Ng, Dekker, Reardon, *et al.*, 2016). In this way, a set of data can be built up that links the image to (i) the experimental conditions, (ii) the protein information and (iii) the experimental outcome.

For every protein that enters crystallisation trials, the experimenter may set up several crystal plates worth of experiments. This number is often multiplied across different incubation temperatures; normally one set of experiments are stored at 4 °C and an identical set at 20 °C. To monitor the experiments properly, the crystallographer can therefore be faced with thousands of viewings of images, taken for each drop, in each plate, at each temperature and at each time point.

It is understandable, given the circumstances, that there has been much interest in automating the analysis of crystal drop micrographs to aid the crystallographer and to reduce the chances of not identifying a crystal that has grown.

1.4 Crystallisation Images

The main outcomes of experiments are (i) crystals, (ii) protein precipitation and (iii) clear drops. Both crystals and precipitation can take many forms, in addition, anomalies such as dust and fibres or improperly formed drops can occur. Automated image analysis algorithms need to take all of these factors into account. Example images will be given here followed by an overview of the analysis methods that have been described in the literature.

1.4.1 Crystal forms

Examples of some of the forms that protein crystals can take are shown in Figure 1.2. It can be seen that as well as separate, single crystals, some can be overlapping in the image or joined together physically. The images in Figure 1.2 serve to show the variety of larger 'pretty' crystals that may be found. In reality, crystals and the images taken of them are often not so beautiful. The size of usable crystals varies greatly, from approximately 20 µm across, to over a millimetre.

The top half of Figure 1.3 shows some examples of images that contain microcrystals. The diversity of these images demonstrates that microcrystals is a broad term and that an analysis method able to classify these images may not be straightforward to implement.

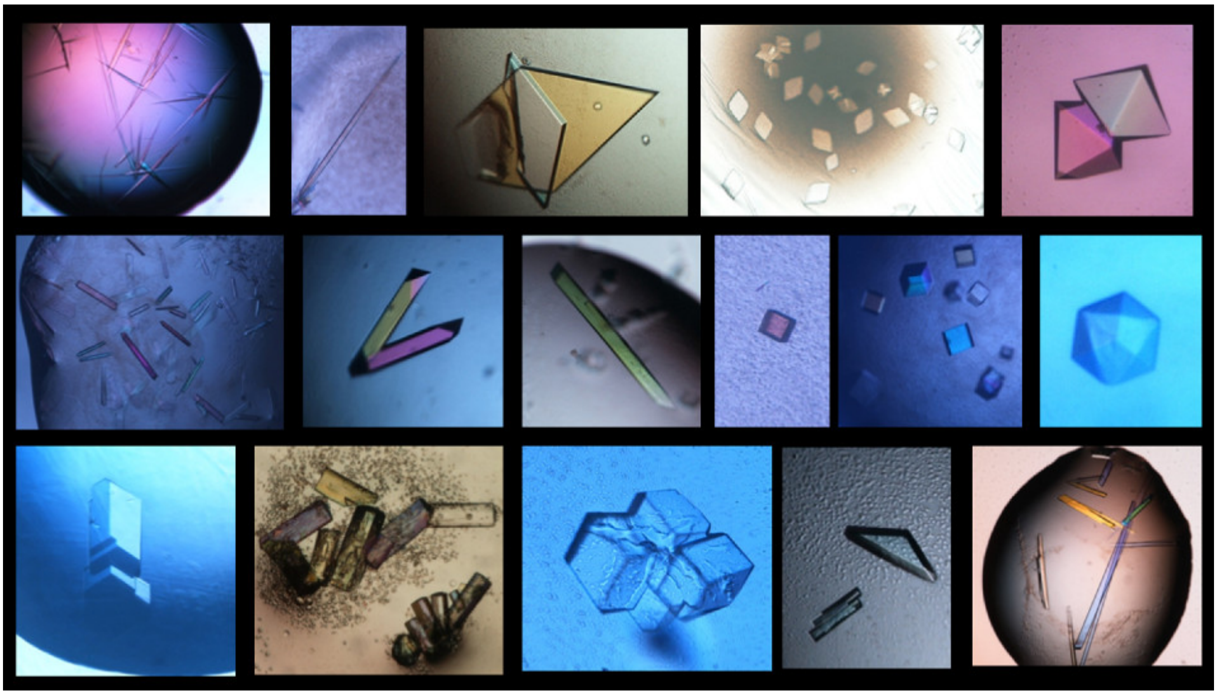


Figure 1.2: Example images of different forms of protein crystals. Adapted from Russo Krauss, Merlino, *et al.*, 2013.

1.4.2 Precipitation types

As mentioned previously (Section 1.2), not all precipitate is undesirable. The lower half of Figure 1.3 shows images of drops containing a variety of precipitate types. Crystals sometimes grow in drops also containing precipitates, in those cases the drop would preferably be classified as a crystal-containing drop rather than a precipitate drop.

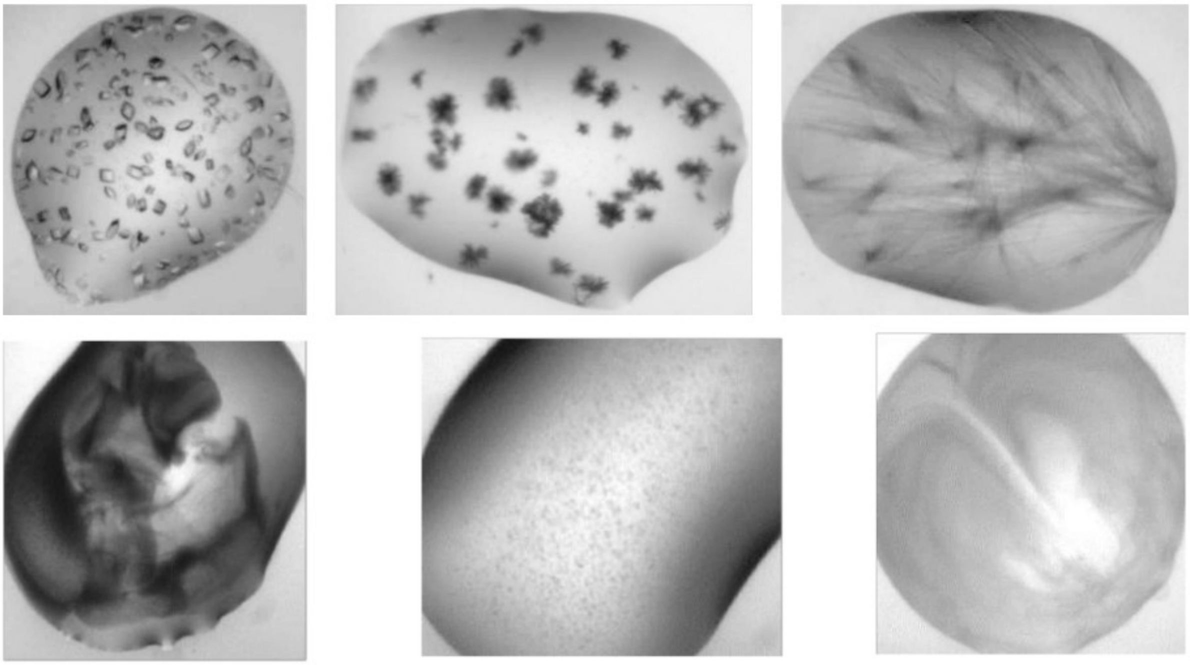


Figure 1.3: Example images of microcrystals and precipitates. Pictures on the top row show different forms of microcrystals. Pictures on the bottom row show precipitates. The bottom central image is a granular precipitate thought to be more promising as a lead for future experiments. Adapted from Ng, 2015.

1.4.3 Analysis Methods

Initial efforts to analyse crystallisation micrographs date back to the 1980s. Ward, Perozzo, *et al.,* 1988 used algorithms written in BASIC and Fortran to process images using Sobel vertical and horizontal edge detection filters. Since crystals generally have straight edges, their presence was later determined by using a method (Hough transform) to find straight lines in the enhanced images (Zuk & Ward, 1991). More recently, in order to classify images, machine learning algorithms have been used with features derived from edge detection (Wilson, 2002; Bern, Goldberg, *et al.,* 2004). However, as presented in Figures 1.2 and 1.3, it can be seen that the micrographs that are of interest to the crystallographer cover a wide range of types. As such, various forms of texture analysis have been utilised, for example Spraggon, Lesley, *et al.,* 2002; Liu, Freund, *et al.,* 2008; Watts, Cowtan, *et al.,* 2008; Cumbaa & Jurisica, 2010; Ng, Dekker, Kroemer, *et al.,* 2014. Texture analysis involves characterising an image in terms of its texture content and will be described in greater detail in Chapter 2.

1.5 Project Introduction

This project follows on from work on texture analysis carried out previously by the Structural Genomics Consortium at the University of Oxford (Ng, Dekker, Kroemer, *et al.,* 2014; Ng, 2015). In that work, software called TeXRank was developed that had a number of features to help the crystallographer. Texture analysis and TeXRank will be described in Chapter 2.

The first section of project work involves creating a protein solubility tool based on the identification of clear drops in crystallisation plates, this is detailed in Chapter 3.

Identification of crystals in droplets can potentially be improved by labelling protein with fluorescent dye (Forsythe, Achari, *et al.,* 2006; Sigdel, Pusey, *et al.,* 2015). When illuminated with the correct wavelength of light, any protein crystals 'light up'. The second section of this project combines texture analysis and fluorescent labelling, something that has not been done before and that has the potential to be even more powerful. This is described in Chapter 4.

1.5.1 About the SGC and the Database

The SGC is a public-private partnership; it is funded partly by industry and has sites at the University of Toronto, the University of Oxford and the State University of Campinas in Brazil. It has a mandate to determine the protein structures of medically-relevant proteins on a large scale. Another part of this mandate is open access science, which results in all information produced by the SGC being released to the public immediately without restriction with the aim of accelerating scientific progress, particularly drug discovery (*SGC | Structural Genomics Consortium Website* 2016).

From 2004 to 2014, the Oxford site produced 700 protein structures (deposited in the Protein Data Bank, to view see *RCSB Protein Data Bank Website* 2016), during this period over 60,000 crystal plates were set up containing over 2,300 different proteins (Ng, Dekker, Reardon, *et al.,* 2016). This data, in addition to details of over 24,000 protein purification experiments is stored in the SGC database and will be used as a resource particularly in the part of this project related to creating a solubility tool (Chapter 3).

Chapter 2

Description of Textons and TeXRank

This project builds on previous work on texture analysis using textons. This chapter describes the texture analysis method and reviews the most relevant work carried out so far. In this case it is work contained in a PhD Thesis (Ng, 2015) and a research paper which resulted from the same body of work (Ng, Dekker, Kroemer, *et al.,* 2014).

2.1 Texture and the Analysis of Texture

There is no singular, succinct definition of texture. One way to define it incorporates a notion of randomness, such as in the definition given by Cross & Jain, 1983 of a “stochastic, possibly periodic, two-dimensional image field” but it can also be defined in a structural way in that can be decomposed into “primitives” that make up the texture in one dimension with a second dimension that describes how these primitives are organised in space (Haralick, 1979).

In the world of computer vision, there are a large number of ways to analyse and categorise textures. The first step in analysis is to extract texture features. Generally these features are divided into four categories: (i) Structural, (ii) statistical, (iii) model-based and (iv) signal processing-based. Further information on these can be found in Xie & Mirmehdi, 2008, but for the purposes of this project, the structure-based features termed *textons* will be described further, as these are the features used for texture analysis in the TeXRank software (Ng, Dekker, Kroemer, *et al.,* 2014; Ng, 2015).

2.2 Textons

Textons were initially defined by Béla Julesz of Bell Laboratories when describing the pre-attentive (rapid and subconscious) discrimination of textures by humans (Julesz, 1981). In this context, a texton is a hypothetical element describing a local, conspicuous feature (a basic element) that allows two textures to be distinguished from each other by subconscious processing. Textons can be thought of as similar to *phonemes*, the smallest units of sound that allow words to be distinguished from one another. Malik, Belongie, *et al.,* 1999 went on to formalise Julesz’s definition of textons into an operational mathematical definition for grayscale images. Here, textons were defined in terms of outputs (also known as responses) to various image filters of differing scale and orientation. If these outputs are plotted in higher-dimensional space, **each cluster centre of points is a texton.** To understand this further, the concept of responses to image filters will now be explained.

2.2.1 Responses to Image Filters

When a filter is applied to an image, the response of a particular region of that image is related to how similar that region is to the filter. Figure 2.1 shows the response of an image to a horizontal line filter. The areas of the image with greatest similarity to the line filter (the top and bottom of the closed circles and the horizontal central line) give the greatest response (shown in red). If the original image is subsequently filtered by other filters, each pixel can then be described by a vector (one dimensional matrix) that contains the response of that pixel to each of the filters. This vector describes the pixel with respect to its surroundings in a feature space defined by the filters that have been applied (Ng, 2015).

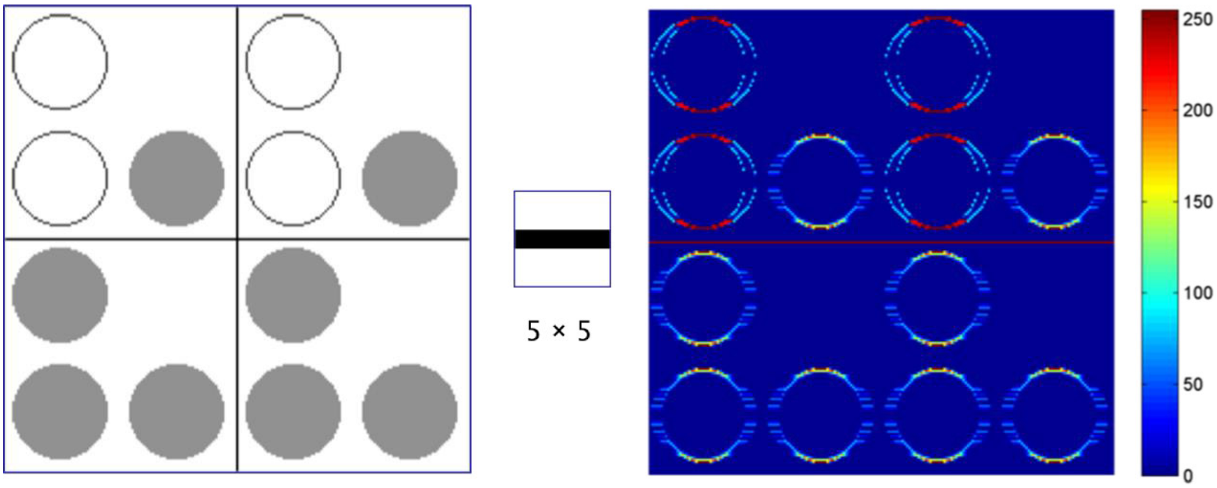


Figure 2.1: Diagram showing filter response. The heat map on the right shows the response when the image on the left is filtered using the 5 × 5 horizontal line filter (centre). Areas of the image which show similarity to the filter give the maximal (red) response. Taken from Ng, 2015.

Textures have properties that repeat; the method of describing pixels with vectors of filter responses will therefore result in some of the pixels over the area of the texture having similar vectors to one another. These similar response vectors will form a cluster in feature space and each of these clusters in the set of "prototype response vectors" is what is termed a texton (Malik, Belongie, *et al.,* 1999).

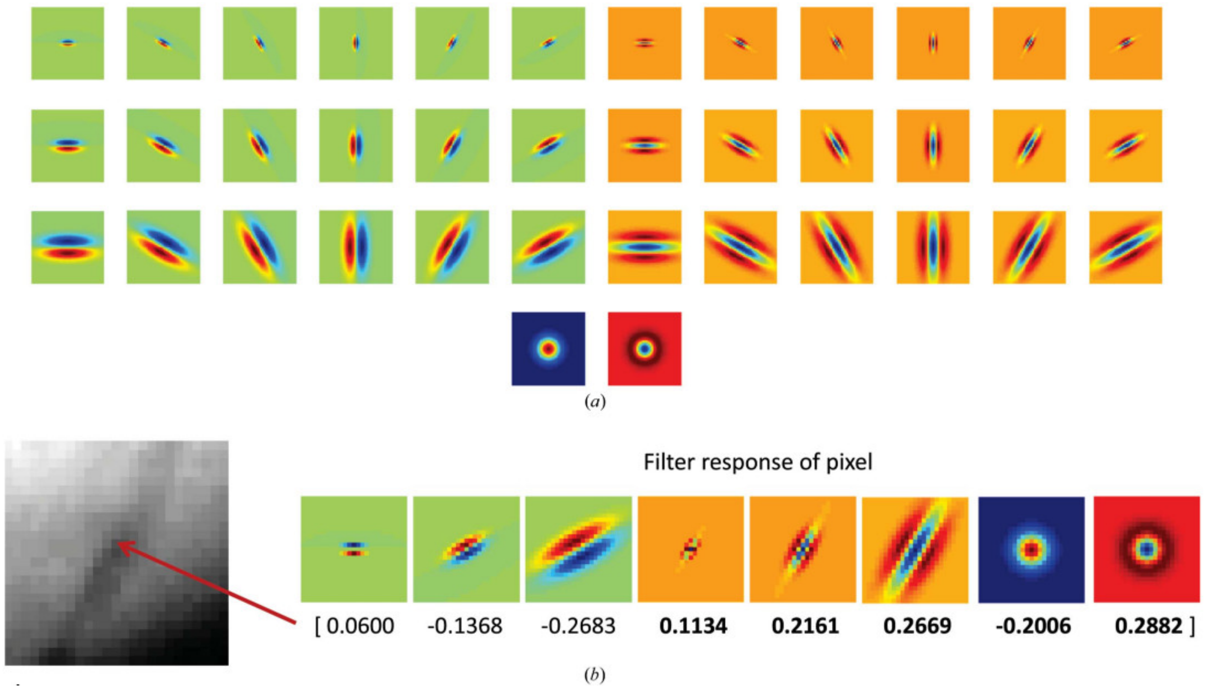
2.2.2 Image Filters

In their work, Malik, Belongie, *et al.,* 1999 applied a bank of 36 filters to the images in order to form the response vectors. Subsequently, Leung & Malik, 2001 extended the texton work from characterising simple textures to a method in which different lighting conditions were taken into account. In this study they described the filter bank as consisting of 48 filters which were made up of 36 elongated filters (6 orientations, 3 scales and 2 phases), 8 centre-surround derivative filters and 4 low-pass Gaussian filters. Later, Varma & Zisserman, 2005 introduced the MR8 (maximum response 8) filter bank which, in their analysis, was found to have a number of advantages over the set used by Leung & Malik, 2001. The MR8 filter bank contains 38 filters consisting of 6 edge and 6 bar filters at three scales each (36) plus a Gaussian and a Laplacian of Gaussian filter. This set was found to be superior in terms of enhanced feature detection and clustering. In addition, it is rotationally invariant and, since only the maximal response to each of the edge and bar filters is taken over each scale, the response vector for each pixel is 8 dimensional for the MR8 set, rather than 48 dimensional as it is after using the set proposed by Leung and Malik. Figure 2.2 (a) shows the MR8 filter bank and (b) shows the 8 dimensional vector response of a pixel.

2.2.3 Building a Texton Dictionary

To convert the filter response of each pixel into meaningful information, each response needs to be compared to a *texton dictionary* which has been built previously from example images of a particular texture class (Varma & Zisserman, 2005). This dictionary contains a set of texture prototypes for that class of texture, and each pixel in the image can be labelled with the texton from the dictionary that is the closest match in feature space. The image can then represented as a histogram of textons from the dictionary and can be classified by comparing this histogram to others from representative classes.

In order to make their dictionary, Ng, Dekker, Kroemer, *et al.,* 2014 started by using 100 images of crystal droplets that were selected to contain a wide range of precipitation types (precipitation patterns). Initial clustering of the filter response vectors led to a library of 1317 textons. Since many of these described the



(*a*)

(*b*)

Figure 2.2: Response of a pixel to the MR8 filter bank. (a) The MR8 filter bank proposed by Varma & Zisserman, 2005. (b) An example response of a pixel after taking the maximum response vector for each scale of the edge (green) and bar (orange) filters plus the Gaussian (blue) and Laplacian of Gaussian (red) filter. Taken from Ng, Dekker, Kroemer, *et al.,* 2014.

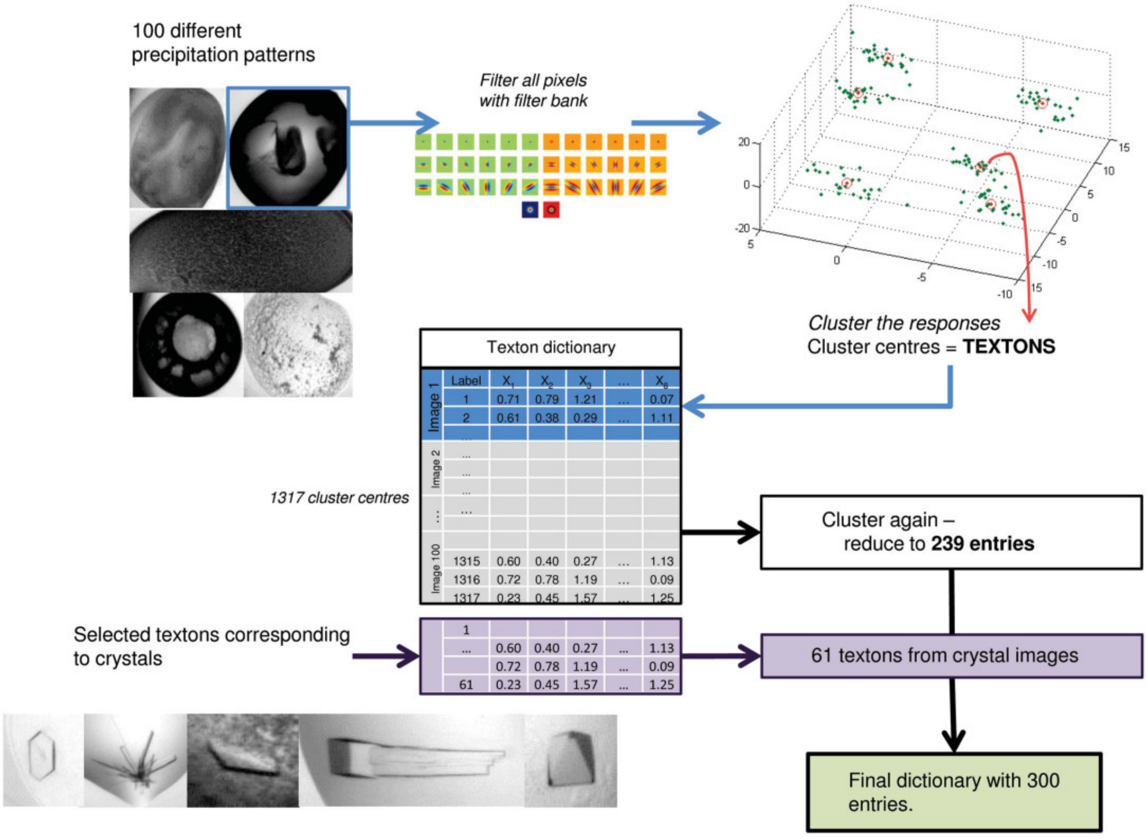
same features in different images, the textons were clustered again using an alternative method (Dirichlet Process-means, Kulis & Jordan, 2011) that removed these duplicates. This led to dictionary with 239 textons describing precipitation.

To represent crystals in the dictionary, 52 images that contained crystals were selected. These were filtered and then subjected to initial clustering in the same manner as for the precipitation images, however an extra selection step was added. In this step, visual selection was used to only take textons from the regions of the image that contained the crystals. Then the second clustering step was carried out. This resulted in 61 textons for crystals. The total number of textons in the final dictionary was therefore 300. A diagram showing this process can be seen in Figure 2.3.

2.2.4 Using the Textons

As stated previously, the texton dictionary is utilised to create a description of an image that can be compared to other images processed in the same way. Since each pixel is labelled with a texton, the image can be represented as a histogram of texton frequencies.

To aid the crystallographer, Ng, Dekker, Kroemer, *et al.,* 2014 utilised the texton method to rank crystallisation micrographs by the likely presence of crystals. This ranking allows the most ‘interesting’ crystallisation images to be viewed first, thereby increasing the chance that any crystals or crystalline precipitate will be noticed. In order to be able to rank the images, first they needed to train a classifier on a set of images that had been annotated by experts into different categories. They retrieved a set of 2501 ‘interesting’ images (those described as containing crystalline precipitates, micro-crystals or crystals) and a set of 3553 images that were ‘uninteresting’ (clear drops or precipitates) from the SGC database. These images were converted to texton distributions and used to train a two class random forest classifier (Breiman, 2001). Once trained, and given a set of textons, this classifier will then output a probability of the image being interesting or uninteresting, which can be used to directly rank them in an order to be displayed (Ng, Dekker, Kroemer, *et al.,* 2014). When tested, this method of ranking was found to lead to



Texton dictionary

	Label	X_1	X_2	X_3	...	X_8
Image 1	1	0.71	0.79	1.21	...	0.07
	2	0.61	0.38	0.29	...	1.11
	...					
Image 2	...					
	...					
	...					
...	...					
Image 100						
	1315	0.60	0.40	0.27	...	1.13
	1316	0.72	0.78	1.19	...	0.09
	1317	0.23	0.45	1.57	...	1.25

1					
...	0.60	0.40	0.27	...	1.13
	0.72	0.78	1.19	...	0.09
61	0.23	0.45	1.57	...	1.25

Figure 2.3: Procedure for generating the texton dictionary for crystallisation images. Taken from Ng, Dekker, Kroemer, *et al.*, 2014.

the crystals being found very high in the ranking order. On 196 test plates, a crystal was found in the first well displayed in 65% of plates and within the first 10 wells displayed in 94% of plates.

2.3 TeXRank

The TeXRank software application started as a viewer for images that had been ranked for the likely presence of crystals in the manner described in Section 2.2.4. As new algorithms and tools were written such as the ability to detect clear drops (described in Chapter 3) and calculate precipitation patterns, the software has increased in complexity (Ng, 2015). TeXRank is written in MATLAB and deployed as a standalone executable that runs on Microsoft Windows and requires MATLAB Complier Runtime (Mathworks) which is freely available. A schematic of the relationship between the software, the database and the image processing pipeline can be seen in Figure 2.4.

TeXRank is currently deployed at three sites, the SGC in Oxford, the Novartis Institutes for Biomedical Research, Basel, Switzerland and the Diamond Light Source synchrotron in Oxfordshire.

2.3.1 Image Processing Pipeline

After the robotic imaging system acquires the images, they are passed through a series of processing algorithms (Ng, 2015).

In summary, the images are converted to grayscale, contrast adjusted and passed through a classifier that determines if the droplets are faulty (such as empty wells or small droplets where either the protein or the cocktail solutions have not been added by the liquid handling robot). If the images pass this stage,

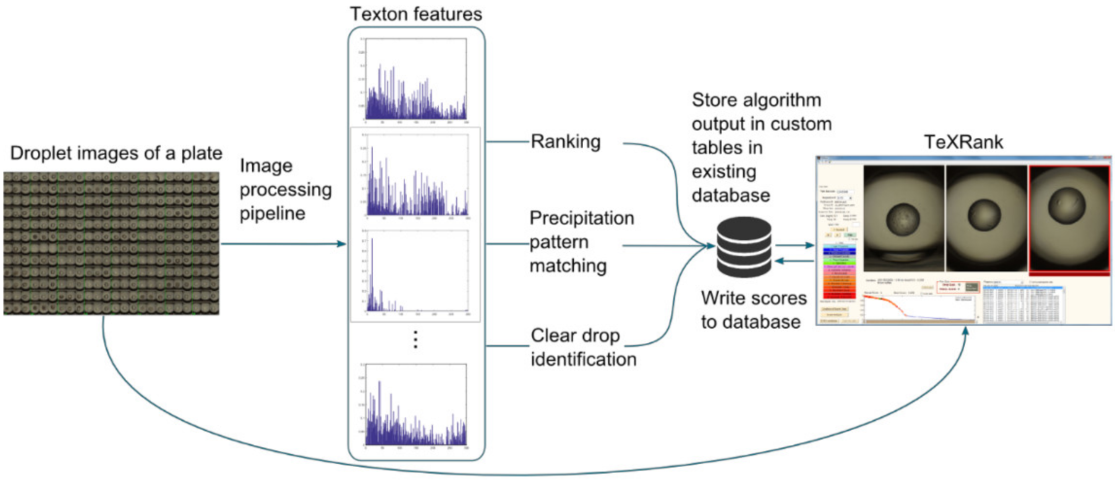

Droplet images of a plate
Image processing pipeline
Texton features
Ranking
Precipitation pattern matching
Clear drop identification
Store algorithm output in custom tables in existing database
Write scores to database
TeXRank

Figure 2.4: Schematic showing where TeXRank sits in relation to the image processing pipeline and database. The output of image analysis algorithms are stored in custom database tables, TeXRank displays information form these tables and also writes annotations given by the user to the database. Taken from Ng, 2015.

the droplets are segmented (i.e. the droplet section of the image is separated from the well background part of the image). Droplet segmentation is a whole area of study in itself and will not be discussed in detail here (for more information see Ng, Dekker, Kroemer, *et al.,* 2014). After segmentation, the droplet images are corrected for shadows around the edge using a selective Gamma correction method. Then, since the droplet boundary will dominate the filter responses when assigning textons, the droplet boundary is extended radially. This extended region is ignored when performing the texton analysis, but allows crystals at the edges of the drop to be detected rather than excluding these regions from analysis.

After these steps, the image is normalised and the texton distribution for the image is calculated from the dictionary, as described previously (Section 2.2.4).

2.3.2 Features

TeXRank was developed over the course of a 3 year PhD project by one person (Ng, 2015). There are various distinct parts to the software. These include processing scripts that are run hourly to check for any new images as well as user interfaces to the data generated by these algorithms. The data presentation interfaces will now be described briefly, these allow a user to view: (i) ranking results, (ii) analyses from precipitation pattern matching and (iii) results from algorithms that identify clear drops.

2.3.2.1 Ranking of Wells and Drops

Users are presented with an interface which links together the images in ranked order with various other tools such as the ability to view plate statistics, including a measure of drop quality and the amount of precipitation in the plate. In addition, a colour-coded view of image thumbnails for the whole plate links the images to the manually annotated scores given by crystallographers and a high resolution image can be viewed along with access to measuring tools (Ng, 2015).

2.3.2.2 Information from Precipitation Pattern Analyses

The precipitation patterns from experiments on a protein in a set of conditions can be compared to historical data in order to suggest new chemical conditions for follow up experiments. TeXRank has an interface that uses this precipitation information and allows the user to design a crystal plate of 96 follow up conditions. The data can be output in a format that can be used as input for a robotic liquid handling system that can prepare the new screen for the user.

2.3.2.3 Output from Clear Drop Identification

As will be described in Chapter 3, clear experimental drops can highlight conditions in which the protein is soluble. This information can then fed back into the protein purification protocol in order to increase the concentration of protein achievable which, is itself, often a limiting factor in obtaining useful crystals (Luft, Newman, *et al.,* 2014). TeXRank uses information from the clear drop analysis to allow the user to select a type of chemical component and view trends relating to the solubility of the protein across the subwells (which have differing mixing ratios of protein to cocktail solution), (Ng, 2015).

2.3.3 Additions to the Processing Pipeline and TeXRank

This project builds upon the current image processing pipeline and TeXRank software to add additional functionality. The first exercise will be to use information from the analysis of clear drops in order to mine the database for a set of conditions that can collected together and used as a tool to improve the solubility of proteins.

The image processing pipeline and TeXRank interface will then be adapted to be able to utilise the texton analysis techniques on images collected from experiments on proteins labelled with fluorescent dye. The hypothesis is that the combination of texture analysis and fluorescent imaging has the potential to be more powerful than either technique alone.

Chapter 3

Creating a Protein Solubility Tool

3.1 Background and Introduction

The first part of the project involves collecting a number of chemical cocktail conditions to be used as a tool for improving protein solubility. This chapter starts by discussing background information relating to this area. This will involve some chemical principles but they are important for understanding what is being attempted. The methods used and results returned will then be described, followed by an evaluation of how useful this approach was found to be and details of any problems that arose.

3.1.1 Solubility and Chemical Space

Solubility is intrinsically important to crystallography since, when a molecule crystallises, it must transition from a state in solution to a solid phase out of solution. There is a complex interplay of factors that determine whether the molecule comes out of solution as a precipitate or in the form of a crystal (Izaac, Schall, *et al.,* 2006). If a protein is not very soluble, it reduces the chances of being able to set up crystallisation experiments from which useable, large crystals can be grown. In fact, Collins, Tomanicek, *et al.,* 2004, found a significant correlation between maximising the solubility of a protein and obtaining crystals. Unfortunately, however, the solubility of any particular protein, which is made up from a specific sequence of amino acids, is very difficult to predict.

When purifying a protein, to be used in crystallography experiments, scientists will add a number of factors to water-based solutions in order to maximise the protein solubility and hence the yield of protein recovered from their starting material. The first component added is called a *buffer*. The molecules in the buffer determine the *p*H of the solution and keep it stable at that value. The second additive which is added to the purification solution is *salt*, usually common table salt (sodium chloride). The protein solubility can vary widely depending on the concentration of salt added and the *p*H. In addition to these additives, other chemicals or combinations of chemicals may be added which can affect solubility. Therefore, the experimenter is often faced with having to try many combinations and permutations of chemicals, concentrations and *p*H values in a trial and error manner.

As explained in Section 1.3, in each crystallisation experiment, some amount of the previously purified protein is mixed with a cocktail of other chemical components. In common with the additives described above, this cocktail will generally consist of a buffer at a particular *p*H, a salt (which may be sodium chloride, but is likely to be one of many other chemical salts) and another type of chemical termed a *precipitant* or *precipitating agent*. By definition, precipitants are agents that affect the solubility of the protein and these are often long-chain polymers. When a protein enters crystallisation trials, it is exposed to hundreds of different combinations of buffer, salt and precipitant in order to find some combination that will lead to crystals. Each combination of these components can be thought of as a point in *chemical space* and the protein is sampling a region of chemical space defined by the set of cocktail conditions that it has been mixed with.

Since chemical space is common ground for the scientists who want to crystallise their proteins and those who want to maximise the concentration of their proteins during purification, the crystallisation screen can give information that is useful for both parties. The crystallographer is looking for precipitation and crystallisation, whereas the purification scientist is looking for clear drops, i.e. combinations of chemical components which keep the protein in solution. The purification scientist can take these chemical components and feed them back into the purification process to maximise the concentration of protein that they can achieve in purification, which in turn, is correlated with being able to grow crystals.

3.1.2 Identifying Clear Drops

Collins, Stevens, *et al.,* 2005 used clear drops in crystallisation experiments to find additives to increase protein solubility. In their experiments, they inspected the crystallisation plates manually under the microscope in order to find clear drops. With the particular protein they were studying, there were 10 clear drops out of 192 experiments after one day. Out of those 10 drops, 5 contained a common buffer component and *p*H. By repurifying their protein in the presence of this buffer, they were able to double its concentration in solution when compared to their earlier attempts (from 8 mg ml^{-1} to 16 mg ml^{-1}). This higher concentration protein then entered crystallisation trials in the same 192 conditions (that had previously led to 182 precipitate wells and 10 clear wells). In this new experiment, crystals formed in 6 wells.

To extend this work, Ng, 2015 used the texton analysis method (described in Section 2.2.4) to identify clear drops. Whereas previously the texton distributions for images had been used to rank images in terms of 'interestingness', directly based upon the probability output by a random forest classifier, in the case of clear drops a random forest classifier was used to classify drops into distinct 'clear' and 'not clear' categories. The classifier was trained on a set of 968 clear drops and 1588 non clear drops. The fact that the image for each drop is linked to the corresponding chemical conditions in the database was then used to create the interface for TexRank (described in Section 2.3.2.3). In this interface, categories of chemical components (e.g. salt, buffer, polymer) can be selected and a mapping between the component and the clear drops in the set of experiments is displayed to the user. This output can then be used in order to make decisions regarding components to add to purification solutions.

Ng, 2015 went on to show how this solubility analysis can be useful in the context of improving the quality of crystals obtained for a particular protein. Previously, a concentration of only 3.8 mg ml^{-1} had been achieved for this protein and the quality of the resulting crystals meant that only a low resolution X-ray structure could solved. The results of the clear drop analysis were used to select new components for the protein solution. The buffer chemical, the *p*H and both the type and the concentration of salt were changed. In the new solution, the protein could be concentrated to more than 25 mg ml^{-1}. Crystallisation experiments were then set up in the same conditions as previously and this resulted in larger crystals from which a higher resolution structure could be solved.

3.1.3 How the Clear Drop Information is Stored at the SGC

In the context of this work there are two main SGC databases that are important:

Beehive - This is the main Oracle database used by all of the scientists at SGC in order to store details of proteins and experiments. It has a complex structure and contains information ranging from bioinformatic analyses of the proteins to results from biophysical experiments, links to electronic laboratory notebooks and details of any X-ray crystallography data collected. A custom GUI named Scarab has been written in-house that allows scientists to insert data into Beehive and run queries in a straightforward, consistent and safe manner.

Crystal Database - This Oracle database is independent of the Beehive database and runs on the machine connected to the crystal plate imaging robotics. The microscope images are stored on this machine and associated metadata are stored in this database along with schedules of when the plates should be imaged. In addition to Beehive, TeXRank writes to tables in this database, storing texton distributions, faulty droplet information and the clear drop scores as determined by output from the clear drop random forest classifier (described in Section 3.1.2). Also, when the texton distributions and clear drop scores are added to the Crystal database, a flag is added to say that the plate has been *autoscored*.

3.1.4 Analysing and Collecting Chemical Conditions

The first task in this section of the project involves retrieving chemical crystallisation conditions that are often linked to clear drops from the SGC database. This set of conditions will then form the basis of solubility tool made up of a set of chemical cocktails. The idea being that a purification scientist can mix their protein with this set of cocktails in order to find components to be added to their purification solutions and increase the concentration of protein achieved. This could therefore also increase the chance of crystallisation.

This part of the project was designed to be short (limited to two weeks) and relatively straightforward, in order to serve as an introduction to the structure of the database and the MATLAB language, before tackling the issue of texton analysis of fluorescence images from dye-labelled protein which will be described in Chapter 4.

All development was done in MATLAB version R2015b. Programs connected to the databases and performed SQL queries via the Oracle JDBC driver.

3.2 Initial Methods

3.2.1 Program plan

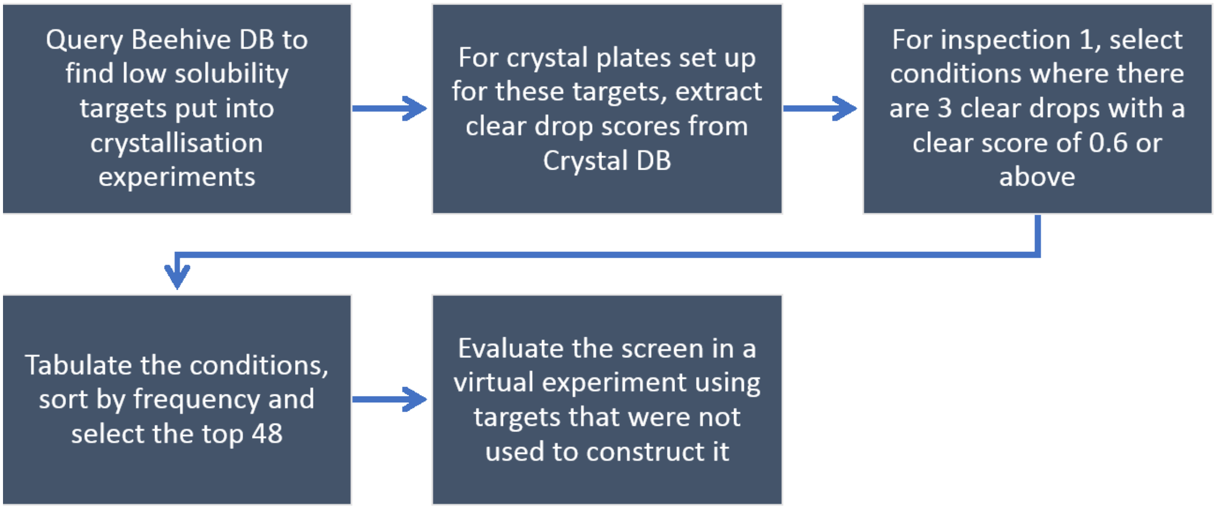

Query Beehive DB to find low solubility targets put into crystallisation experiments
For crystal plates set up for these targets, extract clear drop scores from Crystal DB
For inspection 1, select conditions where there are 3 clear drops with a clear score of 0.6 or above
Tabulate the conditions, sort by frequency and select the top 48
Evaluate the screen in a virtual experiment using targets that were not used to construct it

Figure 3.1: Flowchart for planned methods to collate and evaluate conditions for the solubility tool.

3.2.2 Beehive Database Query

In order to retrieve a list of low solubility proteins from the Beehive database, a query was set up in the Scarab interface. A simplified version of the SQL code for this query can be seen in Listing A.1 in Appendix A.

This query returns a list of proteins and all the crystal plate barcodes for each of these proteins where each crystal plate has been setup using one of 5 specified sparse-matrix screens[1]. The screen types searched for were (i) a screen originally developed based upon successful conditions from the Joint Center for Structural Genomics (JCSG), (Newman, Egan, *et al.*, 2005), (ii) the Hampton Crystal Screen (HCS), (iii) the Hampton Index screen (HIN) (see *Hampton Research Website* 2016) (iv) the Basic ChemSpace screen (BCS), (Chaikuad, Knapp, *et al.*, 2015) and (v) the Ligand Friendly Screen (LFS), (based on the

[1] A commercial collection of cocktails, based upon successful conditions found in research literature, that collectively cover a diverse region of chemical space. See Section 1.2.

PACT screen from Newman, Egan, *et al.,* 2005). The list is ordered by the maximum protein concentration used in the set of crystallisation experiments for that protein.

An initial maximum concentration of 7 mg ml^{-1} was used as cutoff to define a 'low solubility' protein. This resulted in a list of 54 proteins for which 313 crystal plates had been set up.

3.2.3 The Clear Drop Query Program

This program was written in order to determine the chemical cocktail conditions that are most likely to increase solubility amongst the group of proteins returned from the Beehive query.

- `Clear Drop Query` first reads in a list of crystal plate barcodes from a comma-separated text file that contains output from the Beehive database query.
- For each plate, a query is sent to the Crystal database, checking whether the plate images for a user-defined inspection number have been autoscored (see Section 3.1.3).
- If the plate has been analysed, then the clear drop scores are retrieved. Then a list of all the plate wells where the required number of wells that have a clear drop score above a user-defined threshold is collated.
- The list of clear well numbers is used, along with the plate barcode, in order to retrieve text strings containing details of the corresponding chemical conditions from the Beehive database.
- The conditions collected for each plate are then added to a frequency table, so that the 48 most commonly occurring cocktails can be selected to form the basis of the solubility screen.
- A frequency table of the parent screen names that the conditions come from.
- If a plate has not been autoscored, then the barcode is written to a text logfile.

It was decided to use images from the first inspection[2] as this gives a view of the immediate response of the protein to the chemical conditions. If the protein precipitates from solution immediately, it is unlikely to be a useful condition for purification. Wells where all three subwells had a clear drop score of greater than 0.6 were selected. Three drops were chosen to ensure that the protein was soluble in the presence of the chemical cocktail even in the drop where the protein solution and cocktail are mixed in a 2:1 ratio. This threshold for the clear drop score was chosen based on visual inspection of images such as those shown in Figure 3.2.

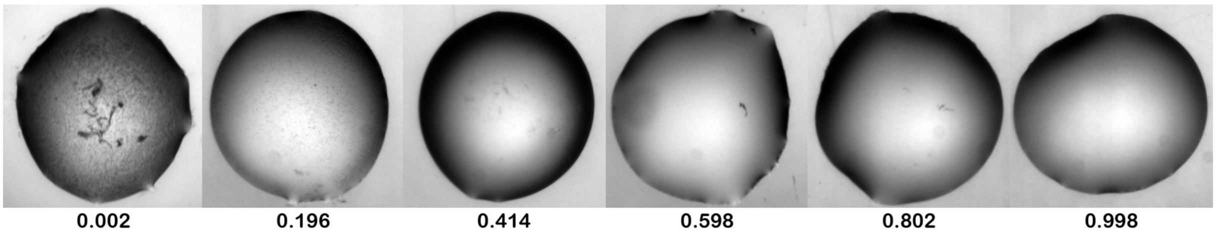


Figure 3.2: Images of drops from SGC, Oxford with their corresponding clear drop scores. These scores were calculated by the random forest classifier described previously (Section 3.1.2).

3.3 Initial Results

`Clear Drop Query` was run on the list of 313 crystal plate barcodes (selected previously, see Section 3.2.2). It was found that 196 plates (≈ 63%) had not been autoscored by TeXRank. This is likely to be due to: (i) crystal plate experiments set-up before TeXRank was in use and (ii) server or network outages

[2] The first set of images taken after the crystal plate has been set-up and stored in the robotic incubator.

when the connection was lost to the crystal database. Scoring by TeXRank is started by a script that is run hourly by the software utility `Cron`. When run, the script only checks for plates where new images have been recorded in the previous hour. In case of a long server or network outage, older image sets will have been ignored.

3.4 Method Refinement Round I

3.4.1 The Plate Catchup Program

Due to the initial results showing a large proportion of plates that had not been autoscored, another program, `Plate Catchup`, was written.

`Plate Catchup`:

- takes in the list of crystal plate barcodes that have not been autoscored along with the number of the inspection of interest (inspection 1 in this case). Since, in many cases the crystal drop images had been archived or stored in different locations to economise on disk space, the program then searches these folders for the correct set of images.
- If these image files are found, the program calls existing TeXRank functions, written by Jia-Tsing Ng, (Ng, Dekker, Kroemer, *et al.,* 2014) to analyse them and then to add the calculated texton distributions and clear drop information to the Crystal database.

3.4.2 Modification of the Beehive Database Query

To improve the quality of the data being analysed, it was decided to modify the original Beehive database query with a cutoff based upon the plate precipitation score. This is a score that is added to the Beehive database by TeXRank for each crystal plate analysed. It is calculated based upon the number of drops which are **not** clear as a proportion of the total number of drops (Equation 3.1).

$$\text{Plate Precipitation Score} = \frac{\sum(\text{Clear drop scores} < 0.5)}{\text{Number of subwells}} \times 10 \qquad (3.1)$$

It was decided to only return plates from the query if they had a precipitation score of greater than 5. This was done in order to exclude crystal plates that were set up with a protein concentration that was a long way below the protein solubility limit. In those cases, all of the plate drops would be clear, giving false positive results.

In addition a series of queries were performed in order to determine the proportions of plates that have been set up with each of the five sparse-matrix screens used in the original query. This was done in order to better understand the underlying data distribution that the results were being picked from.

3.5 Results From Refinement Round I

3.5.1 Use of Plate Catchup

Use of the `Plate Catchup` program reduced the number of plates found to not be autoscored from 196/313 (62.6%) to 95/313 (30.4%). It is thought that the images still missing have been transferred to tape backup.

3.5.2 Data Returned from the Modified Beehive Query

After modifying the Beehive query with the precipitation score cutoff, the results returned consisted of 27 proteins and 107 crystal plates.

3.5.3 Conditions Output by Clear Drop Query

Initially, the `Plate Catchup Program` was used to analyse any of plates which had yet to be autoscored and where images were available. After this, `Clear Drop Query` determined that 22 of the 107 plates returned from the Beehive query still had not been autoscored (20.6%).

A frequency histogram plot of the top 48 conditions output from the `Clear Drop Query` program for this set of plates can be seen in Figure A.1, Appendix A. These data show that the 48 conditions chosen by the program have a diverse effect on the population of proteins used to generate this data, with the 'top' condition seemingly able to solubilise 18 of the 27 proteins (67%), and the 'lowest' condition leading to three clear subwells for only 4 (14.8%) of the proteins. These results are likely to be misleading however, as closer inspection revealed that some of the proteins in the list had been tested multiple times with the same crystal screen. This could lead to the conditions that increase the solubility of these proteins being recorded multiple times.

A frequency histogram plot of the parent screens of all the conditions found by `Clear Drop Query` can be seen in Figure A.2, Appendix A. Here, the source of the conditions that led to clear drops can be analysed. There appears to be a large variation in the proportion of conditions that come from each parent screen type, with the Hampton Crystal Screen (HCS) responsible for 268 instances of 3 clear subwells whereas the Basic ChemSpace screen (BCS) only appears to have led to one instance of 3 clear subwells for the combination of proteins and plates tested. Although this may be due to an intrinsic property, where particular screens contain conditions that are more suited to increasing protein solubility, this may also be due to a bias in the underlying data. In addition, since some of the proteins had been tested multiple times with the same screen, the data would be skewed even further towards those particular screens used for these repeat experiments.

3.5.4 Breakdown of Plate Screen Types Setup at SGC

Figure 3.3 shows graphs representing the output from querying the Beehive database in order to determine the proportions of different screen types used at SGC Oxford. This gives an insight into the make-up of the data that is searched by the `Clear Drop Query` program. During the time that the SGC has been operating, the screen versions have been updated with minor changes which mean that there are now subpopulations of screen versions amongst the main five screen types.

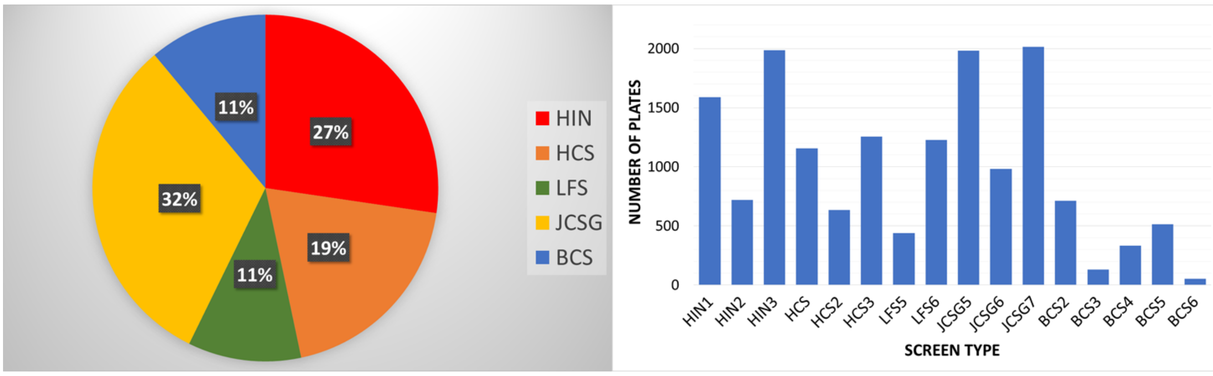


Figure 3.3: Graphs showing a breakdown of the proportions of the main screen types for crystal plates set up at SGC Oxford. Left: Pie chart showing the proportions of the 5 main screen types used. Right: Bar chart showing detail of the numbers of plates set up for each of the screen versions that make up the main screen types.

3.6 Method Refinement Round II

In light of the results collected up to this point, and in order to be able to evaluate the use of the solubility tool, it was decided to standardise the dataset further. There is a need for the group of chemical conditions from which the data for any test set is drawn to be exactly the same as the group used to generate the set of solubility screen conditions. Results output by the Clear Drop Query program (Section 3.5.3) appeared to show possible problems with the underlying data, as shown by the large variation in the parent screen data (Figure A.2, Appendix A) and closer inspection appeared to show repeat experiments in the dataset which could cause bias.

The results of a Beehive database (Section 3.5.4) showed that the screen types were made up of a number of screen versions which vary in a minor way (adjustments to one or two chemical conditions, data not shown) from the other screens of the same type. Even if the components in two corresponding wells remained the same between screen versions, the format of the description string stored in the database often varied, with components being listed in differing order or using alternative spellings. This made it difficult to match condition strings when aggregating and tabulating data.

3.6.1 Further Modification of the Beehive Database Query

It was decided to limit the Beehive database query to only return those plates set up using one of the three most frequently used screen types, HIN, HCS and JCSG. Together these three screen types made up 78% of the plates setup with coarse screens (as can be seen in Figure 3.3).

In addition, the cutoff for the maximum crystal plate protein concentration was increased from 7 to 11 mg ml^{-1}. This was done in order to increase the number of proteins returned from the Beehive database query so that there would be a sufficient number of samples for evaluation involving an independent test set.

3.6.2 The Select Targets Program

The `Select Targets` program was written in order to further standardise the list of proteins used for generating and evaluating the solubility tool.

- Given an inspection number and the output of the modified Beehive query, the program reduces the list of proteins to only those for which autoscored data exists for the three screens, HIN, HCS and JCSG, and for that particular inspection.
- Repeated plate set-ups are removed (if the protein has been setup with the same screen multiple times). This reduces the list of crystal plates to just three for each protein, one JCSG, one HCS and one HIN plate.
- The screen versions are then standardised to HCS3, HIN3 and JCSG7 so that the combined set of chemical components can be defined with consistent description strings.
- In addition, the program outputs a list of crystal plate barcodes which were found to have not been autoscored. This list can then be used as input for the `Plate Catchup` program (described previously, Section 3.4.1) to ensure that data is available for as many plates as possible.

3.6.3 The Collect Screen Conditions Program

This program was written to enable construction of the solubility tool with evaluation of how effective it is.

- `Collect Screen Conditions` first takes in the data structure output by `Select Targets`. This data contains a list of proteins with corresponding crystal plate barcodes.

- The program then partitions the list into a **test set** of 25% of proteins selected at random and a **data set** containing the other 75% of proteins.
- `Collect Screen Conditions` then calls the `Clear Drop Query` program on the test set and the data set respectively in order to collect a table of conditions resulting in clear drops for each set of proteins.
- The solubility screen is formed from the 48 most frequently occurring conditions for the data set.
- In addition, a randomised screen of 48 conditions is also generated from a list of all the 265 unique chemical cocktails found in the HIN3, HCS3 and JCSG7 screens. This is created to aid in evaluating the effectiveness of the screen.
- For each protein in the 25% test set, the program then determines how many of the chemical conditions present in the solubility screen would have led to 3 clear subwell drops (the set intersection between the solubility screen and the cocktails that led to clear drops for the test proteins).
  - This procedure is then repeated, but using the solubility screen created from randomised conditions.
- The list of proteins in the test set and the number of conditions leading to soluble protein (3 clear drops) for the solubility screen versus the randomised screen is then displayed.

3.7 Results From Refinement Round II

3.7.1 Output from the Beehive Database Query and Select Targets Program

The newly modified Beehive database query returned names and crystal plate barcodes for 114 proteins. This data was fed into the `Select Targets` program, which resulted in a list of 41 proteins for which autoscored data was available (inspection 1).

3.7.2 Output from the Collect Screen Conditions Program

The 48 conditions that make the solubility tool, as selected by the `Collect Screen Conditions` program can be seen in Table 3.1. As is the case for the results returned from `Clear Drop Query` in Refinement Round I (A.1, Appendix A), there appears to be a large range in the number of proteins that are solubilised by the 48 conditions which make up the solubility tool. The ‘top’ condition (2M ammonium sulfate, 0.1M acetate, *p*H 4.5) leads to 3 clear subwell drops for 16 out of the 31 proteins (51.6%) used to generate the tool. For the bottom four conditions in the table, only 5 of the proteins are solubilised (16.1%). The decision to choose 48 conditions was arbitrary and could be varied depending on a chosen frequency cut-off if a different sized dataset is available in the future. Unlike the results from Refinement Round I, however, the data from which these results have been generated is standardised and it is therefore much more likely to be robust.

Details of the 31 human proteins (75% of the original 41 proteins) that made up the data set, from which this solubility tool was generated, are listed in Table A.1, Appendix A. It can be seen from this table that the average of the maximum crystal plate concentrations for this group of proteins is 8.8 mg ml^{-1} which is higher than than the initial cutoff of 7 mg ml^{-1} that was set for the Beehive database query to define the notion of a ‘low solubility’ protein. Potential issues, such as this, will be discussed further in the evaluation section (Section 3.8).

A list of the 48 chemical conditions that were selected at random and used for evaluation can be seen in Table A.2, Appendix A.

Table 3.2 shows the results of the ‘virtual experiment’ carried out by the Collect Screen Conditions program where each protein in the independent test set is tested against the solubility screen and then

Chemical Cocktail Condition	**Frequency**
2M ammonium sulfate, 0.1M acetate, *p*H 4.5	16
30% PEG400, 0.2M magnesium chloride, 0.1M HEPES, *p*H 7.5	12
0.49M sodium phosphate monobasic, 0.91M potassium phosphate dibasic	12
20% PEG6000, 0.1M citrate, *p*H 5.0	12
1.1M ammonium tartrate	12
30% PEG400 0.1M cadmium chloride 0.1M acetate, *p*H 4.5	11
20% PEG3350 0.2M ammonium formate	11
30% jeffamine M-600 0.05M cesium chloride 0.1M MES, *p*H 6.5	10
2.4M sodium malonate	10
40% PEG300 0.1M citrate, *p*H 4.2	10
30% PEG4000 0.2M ammonium acetate 0.1M citrate, *p*H 5.5	9
25% tert-butanol 0.1M tris, *p*H 8.5	9
0.02M magnesium chloride 22% polyacrylic acid 5100 0.1M HEPES, *p*H 7.5	9
0.4M sodium/potassium tartrate	9
0.1M ammonium acetate 17%(w/v) PEG10000 0.1M bis-tris, *p*H 5.5	9
3M sodium chloride 0.1M bis-tris, *p*H 6.5	9
30% 2-propanol 0.2M ammonium acetate 0.1M tris, *p*H 8.5	8
0.4M sodium phosphate monobasic 0.4M potassium phosphate monobasic, 0.1M HEPES, *p*H 7.5	8
1.4M sodium citrate tribasic, 0.1M HEPES, *p*H 7.5	8
20% PEG6000, 0.1M bicine, *p*H 9.0	8
0.2M ammonium sulfate, 30% PEG8000	8
0.2M sodium chloride, 25% PEG3350, 0.1M tris, *p*H 8.5	8
0.2M ammonium sulfate, 30% PEG2000MME, 0.1M acetate, *p*H 4.5	7
40% PEG300, 0.2M calcium acetate, 0.1M cacodylate, *p*H 6.5	7
10% PEG6000, 0.1M bicine, *p*H 9.0	7
16% PEG8000, 20% glycerol, 0.16M calcium acetate, 0.1M cacodylate, *p*H 6.5	7
0.8M succinic acid	7
12% PEG3350, 0.005M $CoCl_2$, 0.005M $CdCl_2$, 0.005M $NiCl_2$, 0.005M $MgCl_2$, 0.1M HEPES, *p*H 7.5	7
10% PEG8000, 8% ethylene glycol, 0.1M HEPES, *p*H 7.5	7
0.2M L-Proline, 10% PEG3350, 0.1M HEPES, *p*H 7.5	7
1M sodium acetate, 0.04M cadmium sulfate, 0.1M HEPES, *p*H 7.5	7
10% PEG6000, 2M sodium chloride	6
0.5M sodium chloride, 0.01M magnesium chloride, 0.01M cetrimonium bromide	6
2M sodium chloride, 0.1M acetate, *p*H 4.5	6
1.26M ammonium sulfate, 0.2M lithium sulfate, 0.1M tris, *p*H 8.5	6
1.6M ammonium sulfate, 0.1M sodium chloride, 0.1M HEPES, *p*H 7.5	6
3M sodium chloride, 0.1M HEPES, *p*H 7.5	6
25% PEG3350, 0.2M lithium sulfate, 0.1M bis-tris, *p*H 5.5	6
0.8M sodium phosphate monobasic, 0.8M potassium phosphate dibasic, 0.1M HEPES, *p*H 7.5	6
20% PEG3350, 0.2M sodium malonate	6
2M ammonium sulfate	6
1.6M ammonium sulfate, 10%(v/v) dioxane, 0.1M MES, *p*H 6.5	6
0.2M lithium sulfate, 25% PEG3350, 0.1M bis-tris, *p*H 5.5	6
20% PEG3000, 0.2M zinc acetate, 0.1M HEPES, *p*H 7.5	6
0.4M ammonium phosphate monobasic	5
30% MPD, 0.2M magnesium acetate, 0.1M cacodylate, *p*H 6.5	5
0.8M sodium/potassium tartrate, 0.1M HEPES, *p*H 7.5	5
12% PEG20000, 0.1M MES, *p*H 6.5	5

Table 3.1: The 48 chemical cocktail conditions that make up the solubility tool. These conditions came from the HIN3, HCS3 and JCSG7 screens. They are ordered by the number of times that a condition led to three clear subwells for a protein in the data set (Frequency).

also the randomised screen for a comparison of effectiveness. These results form the basis for evaluation of the utility of this tool and will be discussed in the next section.

SGC ID	**Max. Conc. (mg ml^{-1})**	**Soluble Conditions(Tool)**	**Soluble Conditions(Random)**
SRPK1A	10.6	0	0
BRD8A	5.68	8	5
KCTD8A	7.0	19	10
LIMK2A	10.0	7	3
KLHL1A	8.6	0	0
PRKCL2A	6.0	0	1
CHD1LA	8.0	5	0
UGT1A1A	10.0	1	0
STK39Z	10.0	0	0
KCTD1A	10.0	0	0
Average	**8.6**	**4.0**	**1.9**

Table 3.2: The results of the 'virtual experiment' carried out by the Collect Screen Conditions program. The 10 proteins in the test set (independent from the data used to generate the solubility tool) are listed. For each protein the number of well conditions that led to 3 clear subwells when tested with either the solubility tool or a random set of 48 conditions is listed. Max. Conc. refers to the highest concentration of the protein used to set up crystal plates.

3.8 Evaluation

Since the time spent on this initial part of the project was limited to two weeks (self-imposed), there was not time to carry out an analysis by testing the solubility screen with a protein in the laboratory. It would then have been possible to determine whether the maximum protein concentration achievable could be increased using this tool in a real-world setting. To gain a definitive answer, the tool would have to be tested on several proteins and statistical tests used to determine if any result were significant.

Instead, a 'virtual' experiment was included in the Collect Screen Conditions program (results in Section 3.7.2). This was done by using a test set of proteins, independent of the set used to create the solubility screen tool. The intersection between this condition set, formed from those that resulted in 3 clear subwell droplets for the test proteins, and the 48 condition set that forms the solubility screen tool was calculated. For comparison the intersection with a randomised set of 48 chemical conditions was also calculated. The output of this analysis can be seen in Table 3.2.

From this table it can seen that the average maximum crystal plate protein concentration (8.6 mg ml^{-1}, standard deviation 1.8 mg ml^{-1}) is similar to that observed for the group of proteins selected for the data set (Table A.1, Appendix A, average 8.8 mg ml^{-1}, standard deviation 2.0 mg ml^{-1}), this implies that neither the test set or data set of proteins are biased towards either more or less soluble proteins.

When the test set of 10 proteins was 'exposed' to the conditions that make up the solubility tool (listed in Table 3.1), an average of 4.0 conditions per protein would have led to three clear subwell drops whereas when the same set of proteins were exposed to 48 conditions selected at random, there were an average of 1.9 clear conditions per protein. This implies that the solubility tool (and the methodology used to create it) performs better than random. To test whether this difference is significant, a Wilcoxon signed-ranks test was carried out. This test is appropriate since it is non-parametric (the data tested here can not be assumed to be normally or *t*-distributed) and the variables are paired and randomly selected. In addition, since the assumption is that the solubility tool should be better than random conditions, a one-tailed version of the test is appropriate (Sokal & Rohlf, 2012). The probability of the null hypothesis returned from this test is 0.0469 which means that the two results are significantly different (at the 5% level).

As these data are formed from real-world experiments, even if the experiments were not carried out initially for this purpose, this adds weight to the notion that this methodology for creating a solubility tool is likely to be of use in the laboratory. This is especially true since the difference between the number of conditions leading to clear wells as a result of the solubility screen is significantly greater than when tested against a random set.

The definitive answer still be gained, of course, is whether this tool can be used to find conditions that can be taken back into the purification of a protein in order to increase solubility and improve the chances of growing a high quality crystal. In addition, there are potential improvements that can be made to this method, particularly with respect to the quality of data being mined and standardising this data to be sure of any conclusions. These will now be discussed briefly along with a description of other challenges faced in this part of the project.

3.8.1 Insights and Challenges

In addition to producing a solubility tool that will be of use and creating a set of programs to generate and evaluate the tool, this part of the project was designed to serve as an introduction to (i) MATLAB (a language I had never used) (ii) the structure of the data stored at SGC, Oxford and (iii) the codebase of TeXRank before modifying image processing algorithms in the next section of the project.

The most time-consuming task was to begin to understand the codebase of TeXRank. This incorporates three years of work by one individual and although, for the most part, the code is well commented, no version control was used in development and the only documentation available is a PhD thesis (Ng, 2015). As a result, gaining a clear picture of what the various collective parts of the program do, working out where the data is stored and retrieved from and shedding light on the reasoning behind how the algorithms are structured was not trivial.

The next biggest challenge was to be able to use real world data from the laboratory and to reduce it down to a set that was consistent and from which robust, statistical conclusions could be drawn. Even structured data stored in databases can be messy, particularly with regards to strings of chemical components that can entered in varying orders and with alternative spellings, as was the case here. In addition, the fact that there were gaps in the historical data with regards to both the crystal plates that had been automatically analysed by TexRank and the archived images that were available for processing was unforeseen. However, this collection of data is still an amazing resource and these problems will only become an issue if a particular specific task, such as the one embarked upon here, reveals them.

Finally, the process of writing programs, analysing the output, refining the methods and repeating was a useful challenge to undertake. This allowed flexibility in the approach used and enabled adaptation to unforeseen problems. In particular, the methods for standardising the sets of data from which the screen conditions were chosen and then being able to perform a statistically robust evaluation would not have been possible without repeated re-evaluation as this part of the project progressed.

Chapter 4

Detecting Fluorescent Protein Crystals

4.1 Background and Introduction

Taking images of crystallisation droplets under white light has a number of drawbacks, for example, when a crystal is obscured by a large amount of precipitate it may not be seen at all. Also, occasionally some of the chemical components in the crystallisation cocktail themselves form crystals (salt crystals) which can be mistaken for a protein crystal, giving a false positive.

One promising method to overcome these issues is to label a small percentage of the protein molecules in the sample with a fluorescent dye before setting up the crystallisation experiments (Forsythe, Achari, *et al.,* 2006; Pusey, Paley, *et al.,* 2007). Now, when images are recorded under the correct wavelength of light, any protein crystals should 'light up' and be obvious to see.

In this part of the project, the texton analysis methodology described in Chapter 2 will be adapted to accept and analyse fluorescence images of labelled protein. This will be a novel combination of techniques and has the potential to be more powerful than either method used separately.

This chapter will first describe the literature on fluorescent labelling and previous work on image analysis of micrographs from these experiments. Much of the work on this technique has been carried out by Marc L. Pusey of iXpressGenes Inc, Huntsville, Alabama and Dr. Pusey has kindly agreed to share images for use in this project. Then the methods used to adapt the pipeline for image processing, create a new texton dictionary and train a classifier will be described along with description and evaluation of the results.

4.1.1 Trace Fluorescent Labelling

Fluorescent substances absorb light at one wavelength and then emit light at a different, longer wavelength. For example, the compound carboxyrhodamine absorbs light at a wavelength of 525 nm and emits light at 555 nm, both in the green range. The technique of trace fluorescent labelling (TFL) involves covalently linking a dye such as carboxyrhodamine to a low percentage ($\leq 0.2\%$) of protein molecules in a sample. The crystallisation experiments can the be set up as normal but for imaging, the droplet is lit with a green LED corresponding to the excitation wavelength (Pusey, Barcena, *et al.,* 2015). If the protein crystallises, the dye, now concentrated in the crystal, fluoresces brightly. Therefore, labelling protein in this way allows easy detection of crystals even when buried in precipitate (Forsythe, Achari, *et al.,* 2006). An example of micrographs of TFL protein crystals under white and green light can be seen in Figure 4.1.

4.1.2 Fluorescence Image Analysis

An advantage of using TFL with a dye such as carboxyrhodamine, that fluoresces in the visible spectrum, is that any crystals or crystalline behaviour can be seen by eye. However, automated image analysis still has the same potential advantages in the case of experiments containing trace-labelled protein as it does for unlabelled experiments. There are two main studies on the automated analysis of images from TFL experiments and these will now be looked at in turn (Sigdel, Pusey, *et al.,* 2013; Sigdel, Pusey, *et al.,* 2015).

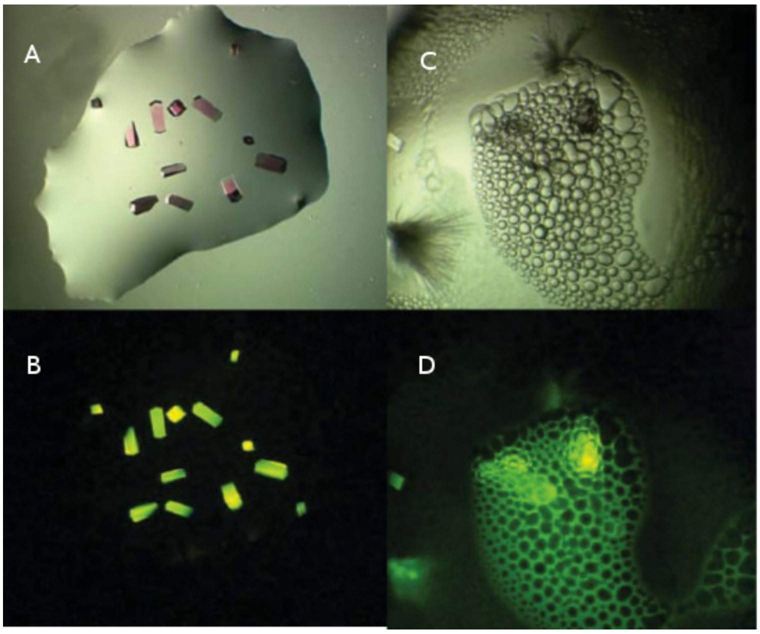


Figure 4.1: Micrographs, taken under white and green light, of droplets containing trace fluorescent labelled protein. Images A and B, protein with 0.5% of molecules labelled with carboxyrhodamine shown in white light and green light. Images C and D, A crystal drop in which some foam has accumulated shown in white light and then in green light where it can seen that crystals are hidden behind the foam. Adapted from Forsythe, Achari, *et al.,* 2006.

4.1.2.1 Study 1. Sigdel, Pusey, *et al.,* 2013

Sigdel, Pusey, *et al.,* 2013, described an image analysis system that categorised micrographs of TFL protein experiments into (i) non-crystals, (ii) likely leads and (iii) crystals. In this work, they extracted 45 features for each image and classified them based on a multilayer perceptron (MLP) neural network. The first stages in the image processing pipeline involved reducing the size of the image and applying an image noise filter. After this three copies of the image were made and converted into binary (black and white) images using different techniques.

Image 1 - converted to greyscale before Otsu's dynamic thresholding algorithm is applied (Otsu, 1979).

Image 2 - thresholded using a 90th percentile green intensity algorithm since, in experiments labelled with carboxrhodamine, regions of the image corresponding to crystals have a higher green intensity than red or blue.

Image 3 - thresholded using a maximum green intensity algorithm, similar to that used for image 2. However, this generally resulted in a smaller foreground (white) portion of the image than for the 90th percentile green intensity algorithm.

The three binary images were then used as masks on the colour image in order to define the foreground and the background regions. For each masked image, six features relating to intensity (e.g. average image intensity and standard deviation in intensity for foreground and background regions) were extracted. A technique called connected component labelling was then used to extract blobs from each of the binary images. Nine features relating to these blobs were then calculated (e.g. blob number, size and measure of symmetry).

For classification, the authors compared using a single MLP classifier and a max-class ensemble. The max-class ensemble was constructed in order to reduce the chances of missing a crystal and consisted of one MLP classifier for each thresholding technique and one additional MLP used for all of the features combined. The class chosen for the image was then taken as the maximum of all four classifiers.

As the authors admit, there is no perfect classification system. When testing on a sample set of 2250 images, using the MLP classifier alone, they achieved an accuracy of 90% with 97% of non-crystals detected and 2% of crystals missed. The max-class ensemble of classifiers achieved an accuracy of 88%, the number of false negatives was lower than for a single MLP but the number of false positives was higher. The ensemble was 99% accurate when classifying non-crystals and 1.2% of crystals were missed.

Overall, the positive aspects of using this technique were that (i) it was fast (3 s per image to extract features and classify), (ii) that when using the ensemble of MLP classifiers only 1.2% of crystals were missed and (iii) that the classification of non-crystal images was very high. However, the system was found to not distinguish well between the categories of 'likely lead' and crystal, with the precision of the 'likely leads' category being 70% and the precision for the crystals category being 69% when using the max-class ensemble. Although it is suggested that perhaps better lighting and fewer out of focus images would improve this issue, in the context of the results of this study, it means that the images would need to be manually reviewed by a crystallographer to separate them into these two categories.

4.1.2.2 Study 2. Sigdel, Pusey, *et al.,* 2015

In the next study, Sigdel, Pusey, *et al.,* 2015 changed their focus to analysing the temporal change in sequences of crystallisation images of protein trace labelled with carboxyrhodamine. Since a crystal droplet is imaged at multiple time intervals, each of these images can be compared and changes, such as growth of crystals or increase in crystal number, detected. Their methodology for this particular study consisted of three stages, (i) identifying crystals, (ii) comparing a sequence of images of the droplets identified in the first stage to find changes and (iii) trying to predict further crystal growth. Since their analysis only concentrated on experiments where crystals were found, all other sequences of images were ignored. The first stage will now be described, but since the second and third stages are not relevant for this project, they will be overlooked.

Crystal identification was carried out using a combination of Otsu's dynamic thresholding (Otsu, 1979) and Canny edge detection (Canny, 1986). After thresholding to generate a binary image, very large (greater than 2.5% of the image) and very small (less than 0.01% of the image) foreground regions were removed. If there were still foreground regions in the image, then Canny edge detection was applied. Any closed edges were then detected and defined as likely to be crystals. Example images showing this process can be seen in Figure 4.2.

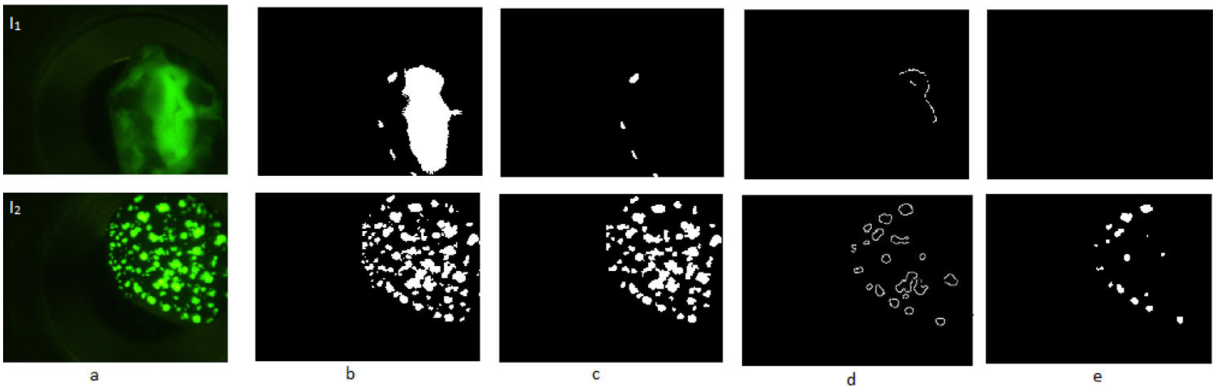


Figure 4.2: Crystal detection in micrographs of carboxyrhodamine trace fluorescent labelled protein. I_1 (a) image unlikely to contain crystals, I_2 (a) image containing crystals, (b) Otsu thresholded images, (c) after small and large regions removed, (d) after Canny edge detection, (e) after detection of closed regions (likely crystals). Taken from Sigdel, Pusey, *et al.,* 2015.

This crystal detection method was tested on three crystal plates (864 experiments), the total number of drops that contained crystals was 46. The algorithm successfully identified 41 of these experiments (≈ 89%), in addition there were 11 false positives (experiments flagged as containing crystals which contained none). The authors concluded that this method is very good at rejecting experiments which do not contain crystals, which is what they wanted to achieve before taking the images through more analysis stages.

The image thresholding and edge detection analysis in this study is straightforward compared to the extraction of 45 image features and neural network classification described in the authors' earlier work (Section 4.1.2.1). In this (2015) study, the authors were only interested in detecting the presence of crystals, whereas their earlier work attempted to classify the image into three distinct categories. The sample image set tested for the 2013 study was also much larger (2250 images) than for the 2015 study (864 images) and it is difficult to make direct comparisons in terms of accuracy when both were attempting to achieve different goals. However, both methodologies were around 99% accurate in classifying images as non-crystals and whereas only 1.2% of crystal-containing images were missed by the ensemble of MLP classifiers, $\approx$ 11% of crystal-containing experiments were missed when using thresholding and edge detection alone.

4.1.3 Combining Florescence and Texture Analysis

Trace fluorescent labelling is a useful technique for visualising regions of high protein concentration in experimental droplets and, as described previously, texton image analysis (Chapter 2) is a useful technique for analysing and describing precipitation patterns and crystallinity. Combining the two methodologies may not be straightforward due to the need to accommodate differences in the lighting conditions and image features (such as background) between image types. Using both methods may have advantages over texton analysis in terms of detecting crystals hidden in precipitate and reducing false positives from salt crystals. This in turn, could lead to more reliable ranking of images than using the current technique on standard white light images. Adapting the TeXRank algorithms to perform texton analysis on images of fluorescent crystals will now be investigated, with a comparison to the techniques described in the two studies that were reviewed in Section 4.1.2.

4.2 Methods to Adapt the Image Processing Pipeline

As described previously (Section 2.3.1), once droplet images are recorded, they are passed through a number of processing algorithms in order to produce the texton distribution for the image. This is termed the image processing pipeline.

In this project, the images used were kindly provided by a collaborator, Dr. Marc L. Pusey of iXpress Genes Ltd., Alabama, US. These images were provided in folders, separated according to crystal plates. Standard white light images for each subwell and corresponding green light (fluorescence) images were provided in subfolders (288 images for each plate, resulting from 96 wells × 3 subwells). Each image was provided as an RGB JPEG of size 1920 × 2560 pixels. In all image processing methods used here, including texton analysis, the images are scaled by a factor of 0.25 to 480 × 640 pixels as this was shown previously to not affect the outcome of the analysis but increases the speed of computation (Ng, Dekker, Kroemer, *et al.,* 2014).

The crystal plates, from which the collaborator's images had been obtained, were of a different specification to ones used at the SGC. In addition, since these experiments had been set up using manual liquid dispensing, the volume of the crystal droplets was much greater than those produced by the liquid dispensing robot at the SGC. These two factors meant that the droplets in the crystal plates looked markedly different to those in the images for which the TeXRank image processing pipeline was designed. Figure 4.3 shows a comparison of the image types.

As a result of this, a new method needed to be developed for masking the white light images (the previous method used for segmenting the drop from the image background is described in Section 2.3.1). In addition, various changes also needed to be made to the pipeline in order to analyse the images taken under green light.

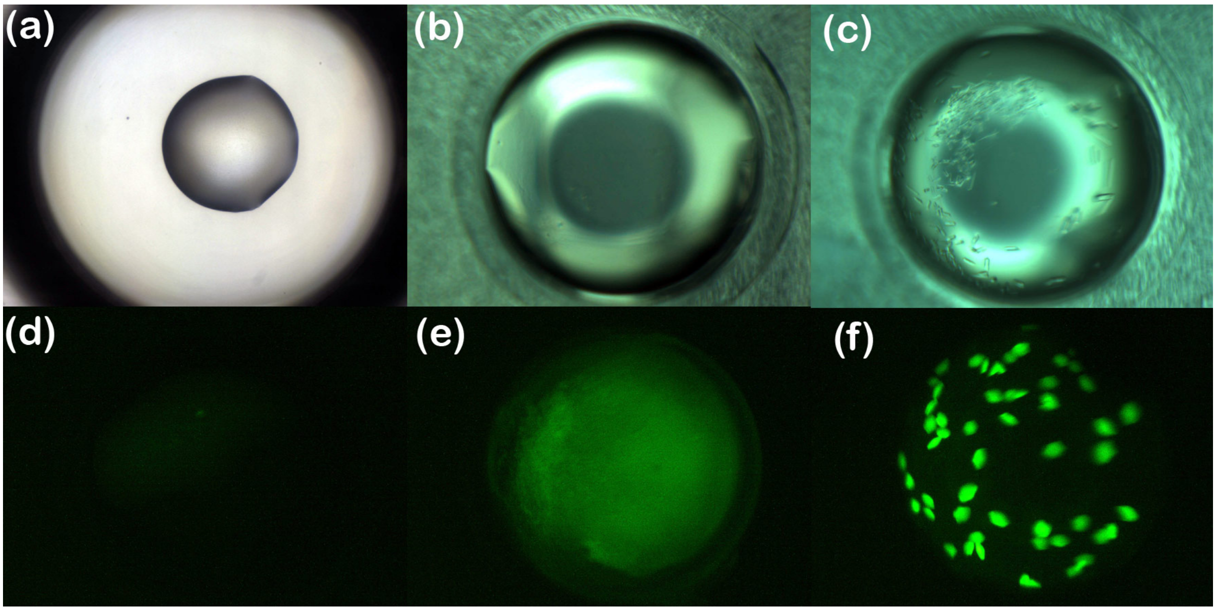


Figure 4.3: Example white light and green light (fluorescence) images. (a), A crystal drop micrograph from SGC, Oxford taken under white light. (b), A crystal drop micrograph taken under white light and provided by our collaborator (no crystals). (c), Another example of a white light image as in (b), but with crystals present. (d), A green light image provided by our collaborator (No precipitate or crystals present). (e), A green light image as in (d), but with precipitate present. (f), A green light image with crystals present.

4.2.1 Masking of White Light Images

Two methods were attempted to separate the droplet area of the white light image from the surrounding pixels. These were (i) aligning to a template based upon minimisation of a least-squares difference in pixel values and (ii) detecting the droplet region using a circular Hough transform.

4.2.1.1 Aligning Images to an Averaged Image Template

Before separating the droplet area from the rest of the subwell in an image, the original TeXRank image processing pipeline first aligns the whole subwell to a grayscale template, created by averaging a number of subwell images (Ng, 2015). This alignment is done by using a translation function in combination with another function to minimise the least-squares difference between pixel values from the translated image and the template. After this, TeXRank uses another algorithm to separate the droplet from the rest of the subwell (this stage is not needed for the new image processing pipeline, since the drop fills the whole subwell area).

In order to use this technique for the new sets of images, so that the area surrounding the drop can be masked out, five of the white light images provided by our collaborator were averaged together. Then the translation and minimisation functions from TeXRank were used to try and align images to this template. This was done with the aim of being able to mask a fixed area of the image once it had been aligned.

This technique was found not to work successfully (data not shown). This may have been due to a number of reasons, for example, large differences in initial alignment of the images, issues with image focus and variability in lighting conditions.

4.2.1.2 The Generate Mask from Image Program

Since the edge of the subwell can be seen to form a dark circle surrounding the droplet, determining the position of this circle is one strategy for starting to mask the droplet area.

A program called `Generate Mask from Image` was written to perform masking based on this method.

- This program takes in an image that has been converted to grayscale. This image is then contrast-adjusted and a MATLAB function, `imfindcircles`, is called to perform a circular Hough transform and detect dark circles within a specified range of radii. The `imfindcircles` function returns the x, y coordinates of the circle centre along with the radius.
- This circle information is then used by `Generate Mask from Image` to create a logical mask of the same pixel dimensions as the image, with the area inside the circle set to 'True' and the area outside of the circle set to 'False'.

The mask returned from the program can then be applied to the image. An overview of the masking process can be seen in Figure 4.4.

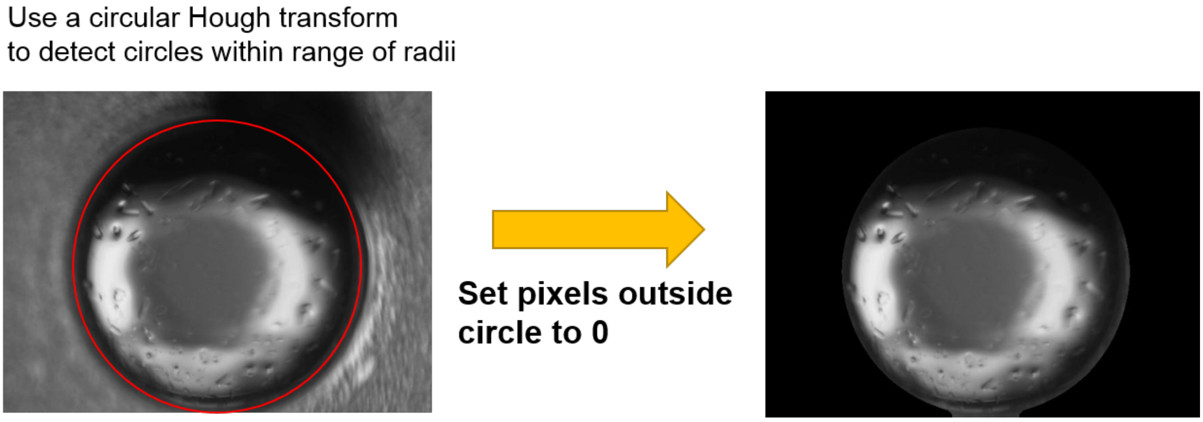

Use a circular Hough transform
to detect circles within range of radii
Set pixels outside
circle to 0

Figure 4.4: Overview of white light image masking process performed by the Generate Mask from Image program.

4.2.2 Masking of Green Light Images

The first step in all attempts to prepare the green light images for analysis was to extract the green channel from the RGB images. This produced a grayscale image with the only the information of relevance for analysing fluorescence.

Initial attempts at masking the images collected under green light involved simply using the mask calculated by `Generate Mask from Image` when given the corresponding white light image as input. The regions of the image that corresponded to the 'False' values of this mask were given a pixel intensity of 0. However, it quickly became apparent that the positions of the subwells did not correspond between the two image types as they had been collected independently from one another.

4.2.2.1 The Mask Fluo Image Program

The second approach involved writing another program, `Mask Fluo Image`.

- First, a pre-made logical mask is loaded. This mask was created by averaging 5 white light images and using the resulting image as input to the `Generate Mask from Image` program (thereby producing a circular mask with an approximation of the radius of the droplets).
- In order to increase the speed of the maximisation function, the grayscale image (green channel of the RGB) and mask are downsized to 120 × 160 pixels before being passed to an image translation function. This function translates the image in order to maximise the summed intensity of the pixels in the masked image and returns the translation vector. Downsizing the image and mask in this manner was found to reduce the image masking time from 10.0 s for a 480 × 640 pixel image to 1.1 s. A constrained minimisation function is used in this process, again to reduce the search time, especially in the cases where the images are black (no precipitation or crystallisation observed and therefore no fluorescent signal, e.g. Figure 4.4, (d)).

- The translation vector is multiplied by the inverse of the image scaling factor (used to downsize the image in the previous step), before the pre-calculated mask is applied to the original size grayscale image.

An overview of the steps performed by the `Mask Fluo Image` program can be seen in Figure 4.5.

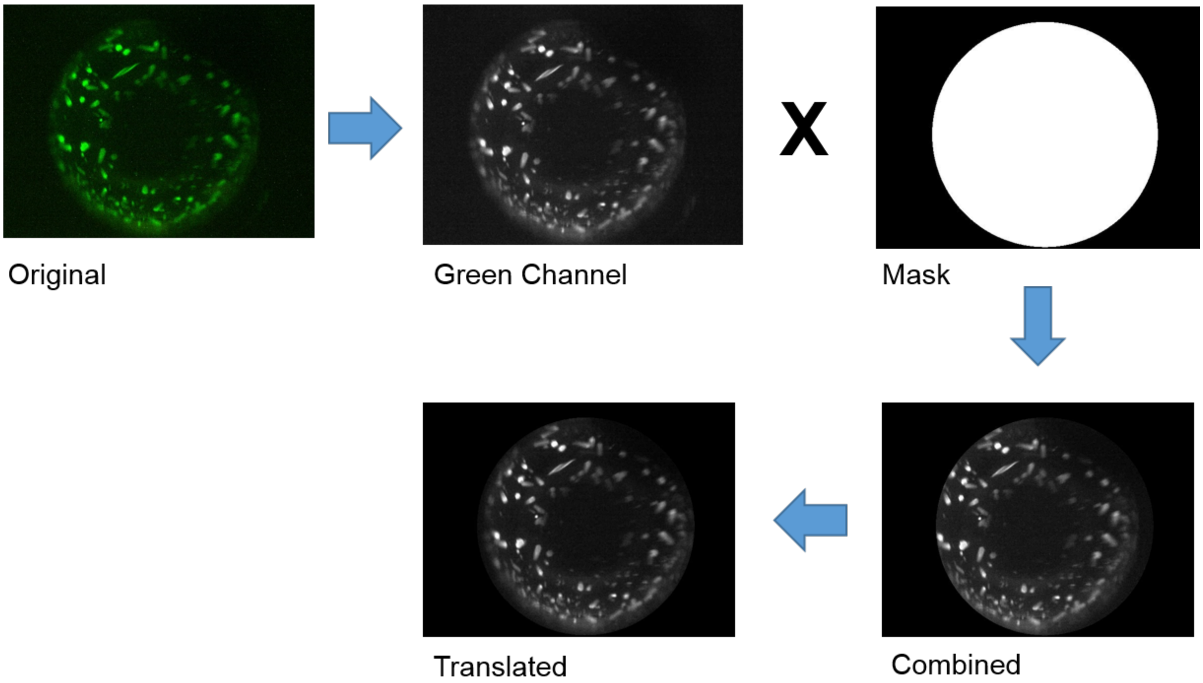


Figure 4.5: Overview of masking process performed by the Mask Fluo Image program.

4.2.3 Testing the Images with Texton Analysis

The masked white light and green light images were input into the texton analysis function of TeXRank. The output was visually assessed by viewing images where each pixel had been labelled with the closest texton from the 300 member dictionary (described previously, Section 2.2.3) and the corresponding texton histograms. The texton analysis functions as well as those used to label images with textons and output the texton histograms, used in section, were all written by Dr. Jia-Tsing Ng (Ng, 2015).

4.2.3.1 Output from Texton Analysis of White Light Images

Figure 4.6 shows examples of two white light images and the outputs from processing them with the texton analysis algorithms. It can be seen from this figure that effects of the uneven droplet lighting are being picked up by the texton analysis. For example, the droplet image in Figure 4.6, (a) is poorly lit on the north side (when viewed as a compass) and is saturated on the west side. These lighting effects can be seen in the texton-labelled image, (b). In addition, it can be seen that Figure 4.6, (d) does not show a clear drop, but one that contains textured precipitate. However, very little of this precipitate texture is being detected in the texton-labelled image, (e), due to uneven lighting.

The two texton frequency histograms (Figure 4.6, (c) and (d)) are dominated by textons corresponding to the lower range of texton numbers (textons 0 to 75, displayed as blue in images (b) and (e)). Since the textons in the dictionary are arranged by magnitude, those in the lower range (0-75) can be thought of as the ‘clear drop’ textons, those in the middle of the range (76-239) as the ‘precipitate textons’ and those at the top of the range as the ‘crystal textons’ (240-300). Therefore, it would be expected that the histogram in Figure 4.6, (c) should contain a higher proportion of crystal textons (in the 240-300 range) than seen

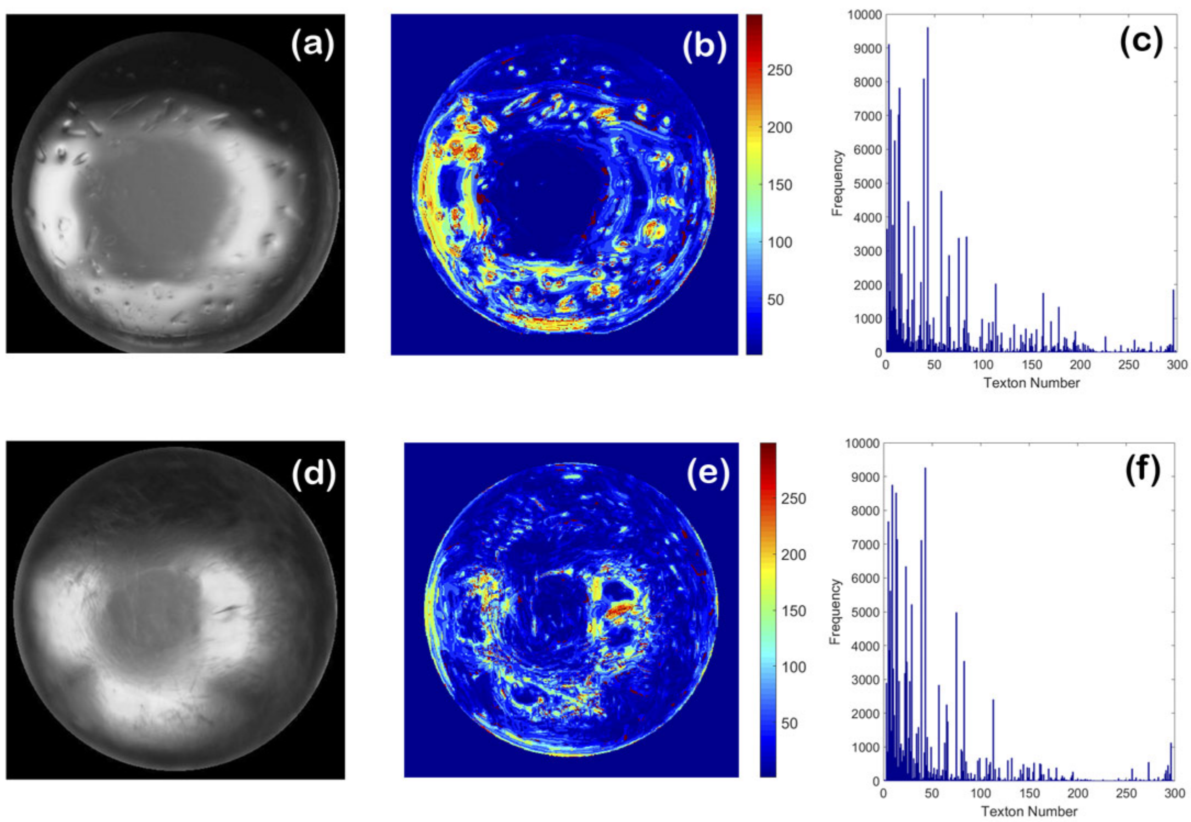


Figure 4.6: Examples of white light images, their texton labels and texton frequency histograms. Image (a), A droplet containing crystals. Image (b) The texton-labelled image corresponding to image (a). Graph (c), The texton frequency histogram corresponding to image (b). Image (d), A droplet containing precipitate. Image (e), The texton-labelled image corresponding to image (d). Graph (f), The texton frequency histogram corresponding to image (e).

here and the histogram in Figure 4.6, (f) should contain higher proportion of precipitate textons (in the 76-239 range) than seen in this case.

In addition, it should be noted that there are some artefacts around the edge of the image mask (Figure 4.6, (a) and (d)) that are detected by the texton analysis (images (b) and (e))

4.2.3.2 Output from Texton Analysis of Green Light Images

Figure 4.7 shows examples of two green light images and the outputs from processing them with the texton analysis algorithms. The crystal containing regions of the images (Figure 4.7(a) and (b)) are clear to see when looking at the corresponding texton-labelled images (b) and (e). There does however, appear to be an issue with image noise being recorded as texture (in the 75-240 range of the textons), as can be observed in the texton frequency histograms (Figure 4.7(c) and (f)). This noise is especially apparent in dark regions of the images.

4.2.4 Further Adjustments to the Processing Pipeline

As a result of the observed output from the texton analysis algorithms on the masked white light images (see Section 4.2.3.1) and green light images (see Section 4.2.3.2), various changes needed to made to the pipeline in order to ensure that reliable results could be obtained for both sets of images.

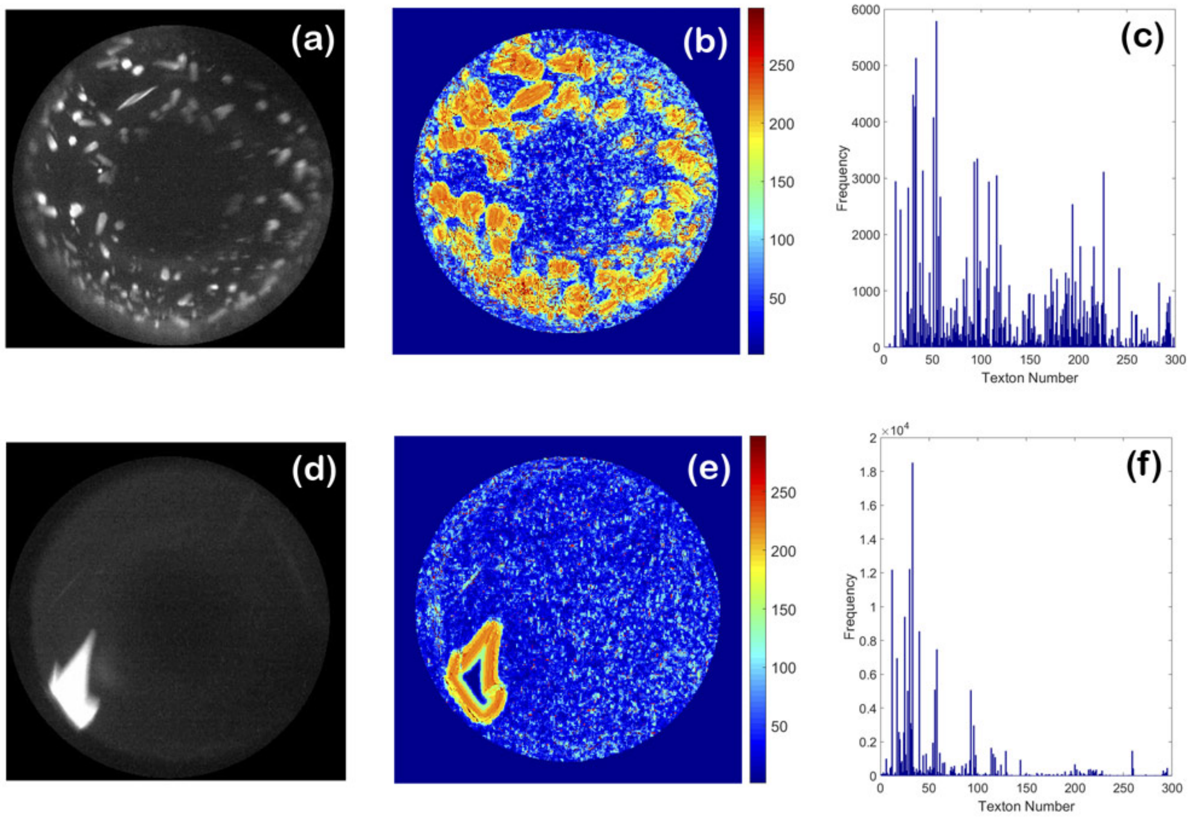


Figure 4.7: Examples of green light images, their texton labels and texton frequency histograms. Image (a), A droplet containing small crystals. Image (b) The texton-labelled image corresponding to image (a). Graph (c), The texton frequency histogram corresponding to image (b). Image (d), A droplet containing one large crystal. Image (e), The texton-labelled image corresponding to image (d). Graph (f), The texton frequency histogram corresponding to image (e).

4.2.4.1 Background Lighting Subtraction for White Light Images

To reduce issues caused by uneven lighting of the white light images (seen in Figure 4.6, Section 4.2.3.1), a technique was used that involved creating a ‘background’ image and subtracting the pixel intensities of this image (scaled by a factor of $\frac{2}{3}$) from the original. The background image was generated by performing a morphological opening operation (an image erosion followed by an image dilation) using a disk-shaped structuring element of size 17 pixels for the 480 × 640 grayscale images (Fisher, Perkins, *et al.,* 2004). The size of the structuring element was chosen based upon experiments on a number of images to ensure that the background image created did not contain the crystal detailing.

A diagram showing an overview of the background subtraction process can be seen in Figure 4.8. The scaling of the background image pixel values before subtraction by $\frac{2}{3}$ was decided on as a good compromise as subtracting the full value was too severe as a correction. The effects of this correction in addition to the mask modification described in Section 4.2.4.2 can be seen in Figure 4.9.

It is apparent from viewing the texton labelled images before (Figure 4.9, (a) and (d)) and after (Figure 4.9, (b) and (e)) the background correction is applied that the level of image detail detected by the textons is enhanced and the effect of the uneven lighting is decreased. This is reflected in the texton histograms plotted in panels (c) and (f) of the same figure. Here it can be seen that in comparison to the equivalent graphs, (Figure 4.6, (c) and (d)), the distributions from the corrected images are less dominated by particular low index number (clear drop) textons. As a result, those textons in the precipitate (76-239) or crystal (240-300) range are more pronounced in the histogram.

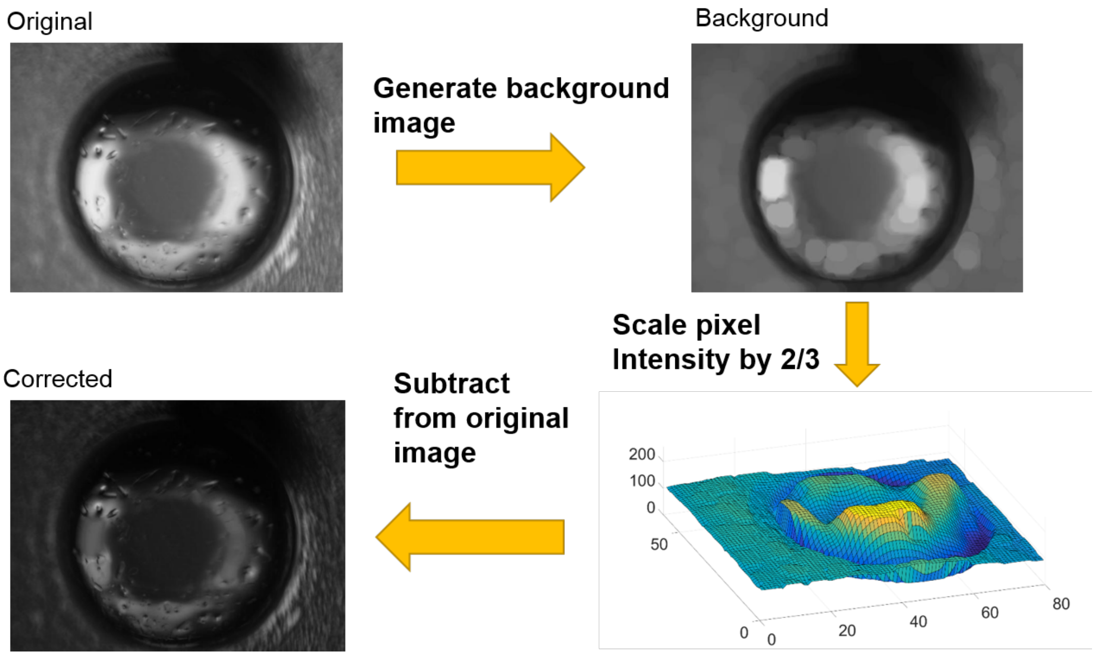

Original
Generate background image
Background
Scale pixel Intensity by 2/3
Corrected
Subtract from original image
200
100
0
50
0
0
20
40
60
80

Figure 4.8: Overview of the lighting correction process. A 'background' image is created by using a morphological opening operation with a disk-shaped structuring element. The pixel values of the background image are scaled before being subtracted from the original image.

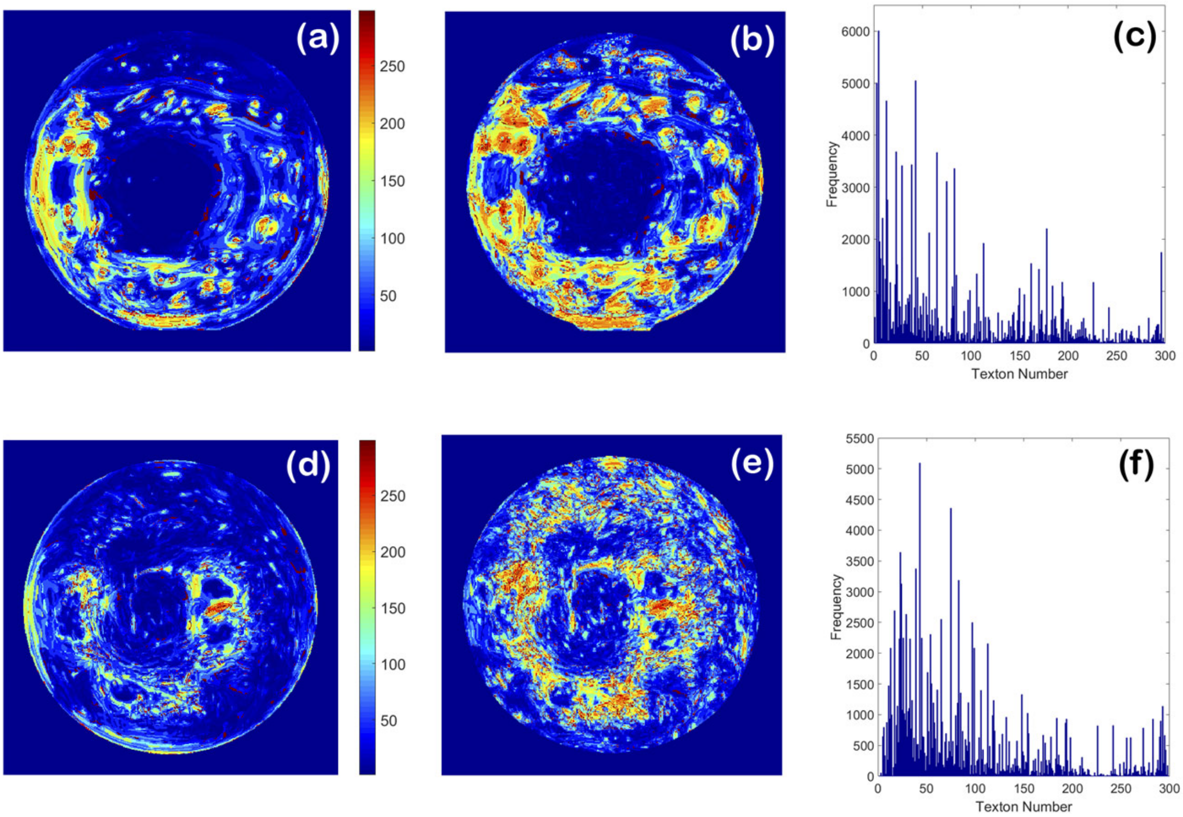


Figure 4.9: The effect of image corrections on the texton analysis of white light images. Image (a), The texton-labelled image corresponding to Figure 4.6, image (a) before corrections were applied. Image (b), The texton-labelled image corresponding to (a) with corrections applied. Graph (c), The texton frequency histogram corresponding to image (b). Image (d), the texton-labelled image corresponding to Figure 4.6, image (d) before corrections were applied. Image (e), The texton-labelled image corresponding to image (d) with corrections applied. Graph (f), The texton frequency histogram corresponding to image (e).

4.2.4.2 Modification of the Generate Mask from Image Program

In addition to the background lighting correction, the artefacts around the edge of the image mask (seen in Figure 4.6, Section 4.2.3.1) were corrected by applying an image erosion using a disk-shaped structuring element (size 11 pixels) on the 480 × 640 mask created by the `Generate Mask from Image` program. This had the effect of reducing the size of the circle in the image mask, thereby reducing the artefacts around the edge of the mask, as can be seen in the corrected images in Figure 4.9.

4.2.4.3 Changes to the processing of Green Light Images

The issues with image noise in the green light images (seen in Figure 4.7) were reduced in two ways.

1. By reducing the 'salt noise'[1] in the images. This noise was especially apparent in dark regions of the images. The images of droplets that contain neither precipitates or crystals are entirely black and so this noise was more apparent in these cases (data not shown). This was achieved by applying a morphological opening operation to the image using a disk-shaped structuring element of size 3 pixels (Fisher, Perkins, *et al.,* 2004). This noise reduction was applied to the the full size (1920 × 2560 pixel) images before they were scaled down in size for analysis. This ensured that important texture detail was not lost.
2. The selective Gamma correction step (described briefly in Section 2.3.1 and in more detail by Ng, Dekker, Kroemer, *et al.,* 2014) was removed. This process is designed to selectively boost shadow regions in the white light images. This correction is not appropriate for the green light images however, and when applied, it was found to increase image noise in dark regions.

Figure 4.10 shows the results of alterations on the output from texton analysis. It can be seen from texton-labelled images (b) and (e) that the level of noise in the dark areas of the images are greatly reduced. The detection of the crystal regions in the images does not appear to be affected, however.

When comparing the texton histograms before (Figure 4.7, (c) and (f)) and after the alterations were made, it can be seen that the textons around number 100 (in the precipitate region of the distribution) contribute proportionally less to the histograms when the noise is reduced. However, the low number textons now dominate more. This is desirable as these images only contained crystals and not precipitate so it could be said that reduction of signal from precipitate textons is therefore an improvement.

4.3 Building a New Texton Dictionary

Following on from studying the output of the texton analysis algorithms on the white light and green light images provided by our collaborator (Section 4.2.3) and the resulting alterations made to the processing pipeline for both types of image (Section 4.2.4), the next stage was to decide whether it is was sufficient to use the existing texton dictionary, created previously on white light images from the SGC, on the green light images. As the images of fluorescent drops appear starkly different to those collected under white light, and to ensure that the textons in the dictionary gave good coverage of the texture prototypes found in the precipitation and crystal regions found in these images, it was decided to create a new dictionary of textons specifically for the green light images. The original program for building the dictionary, written by Jia-Tsing Ng (Ng, 2015), was modified to include the changes to green light image processing described in Section 4.2.4.3. The modified program is named `Append Flu Textons To Dictionary`.

4.3.1 Image Selection

As described in Section 2.2.3, the first step in building a texton dictionary involves selecting a number of images showing various textures of interest. All of the images were viewed and a group of 158 green

[1] White pixels, sparsely distributed across the image.

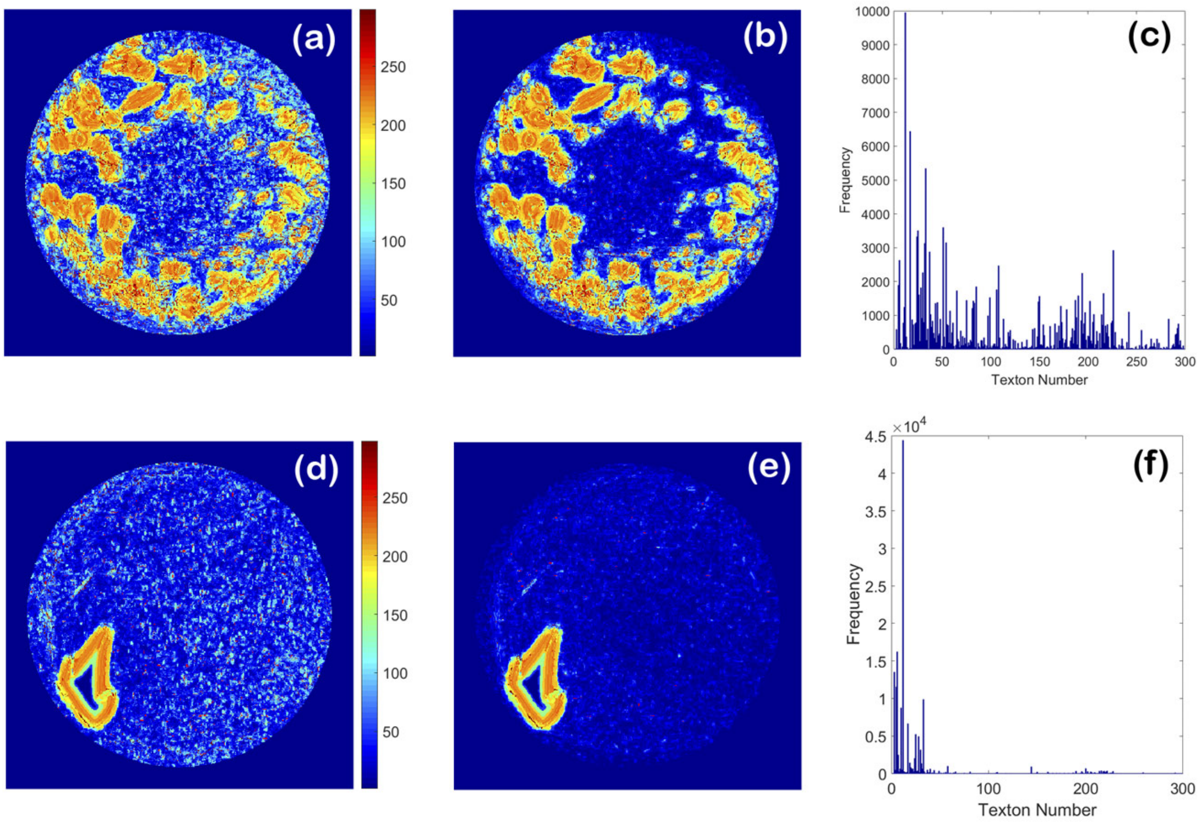


Figure 4.10: The effect of alterations to the processing pipeline on the texton analysis of green light images. Image (a), The texton-labelled image corresponding to Figure 4.7, image (a) before alterations were made. Image (b), The texton-labelled image corresponding to (a) after alterations applied. Graph (c), The texton frequency histogram corresponding to image (b). Image (d), the texton-labelled image corresponding to Figure 4.7, image (d) before alterations were made. Image (e), The texton-labelled image corresponding to image (d) after alterations applied. Graph (f), The texton frequency histogram corresponding to image (e).

light images containing precipitation were placed into one group and 57 images containing crystals were placed into another group.

4.3.2 Selecting Precipitation Textons

The 158 precipitation images were cropped so that just the region of interest in the image remained. All of these cropped images can be seen in Figure B.1, Appendix B.

Each image is filtered with the MR8 filter bank (Varma & Zisserman, 2005) and the resulting 38 dimensional vector of filter responses is reduced down to an 8 dimensional vector (since only the maximum response for the edge and bar filters at each scale are kept, as described in Section 2.2.2). This is exactly the same method used for the standard analysis technique used to convert an image into a texton histogram.

In this case, however, since a dictionary is being built, rather than being used to label the image pixels, the image response vectors that have been calculated are then clustered in order to create the image texture prototypes to be added to the dictionary. The clustering method used is a Gaussian Mixture Model technique that uses variational Bayesian model selection rather than the maximum likelihood method (Corduneanu & Bishop, 2001). In short, this version of model selection is advantageous, as prior knowledge of the number of clusters is not needed in order to reach a good solution. Instead, the model starts with a large number of clusters which can then be trimmed down to the final group based upon minimising a parameter related to mixing of components (in the Gaussian Mixture Model).

The clusters calculated in this way for each image are added to a starting texton dictionary. The generation and clustering of the textons from the set of precipitation images took around 22 h on all 4 cores of an Intel i7 processor on a Windows machine with 8 GB of RAM. This initial dictionary contained 3120 textons.

As many of the textons in this starting dictionary are likely to be the similar to one another, the whole dictionary is clustered again to reduce them down to a smaller set of distinct prototypes. This time, a different clustering method, Dirichlet Process (DP) means is used (Kulis & Jordan, 2011). This technique allows points far from a cluster to form their own cluster but with a penalty imposed, so that cluster centres that are over a certain threshold distance away can be kept as separate textons, but duplicate textons are removed (Ng, 2015). After this step, the dictionary contained 138 textons.

4.3.3 Selecting Crystal Textons

The 57 crystal-containing images were also cropped to leave only the area of interest and thereby save processing time. These images can be seen in Figure B.2, Appendix B.

The calculation of the image response vectors and the first stage of clustering was carried out in the same way as for the precipitation images (Section 4.3.2). However, after the textons for each of the images had been clustered, the ones to be kept were selected manually from a texton-labelled image. This way, those textons relating to the crystals in the images could be chosen over those generated from other regions. Figure 4.11 shows an example of the images used for selection of textons in this manner.

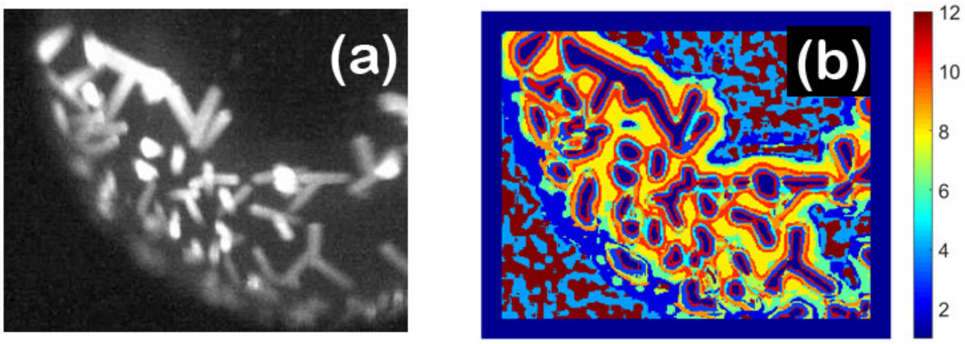


Figure 4.11: Selecting crystal textons from labelled images, Image (a), The crystal region of interest. Image (b), An image labelled with the colours referring to clustered filter response vectors (textons). In this case, 12 textons have been found from image (a). The indices of the textons describing the crystals are entered into a command line interface and added to the dictionary. The other textons are discarded.

This method lead to 268 crystal textons being found. After clustering by DP means, this set was reduced to 73.

4.3.4 The Final Fluorescent Protein Texton Dictionary

The 138 precipitate and 73 crystal textons were combined and then clustered once more by Dirichlet Process means. This led to a final fluorescent protein texton dictionary containing 188 members. As with the white light texton dictionary, these textons are ordered by magnitude.

4.3.4.1 Comparing the White Light and Fluorescent Dictionaries

The most obvious difference between the original dictionary, created on white light images and the new fluorescence dictionary, created using green light images, is the size. The original dictionary contains 300 textons versus the 188 member fluorescence dictionary. The dictionaries were both created using the same parameters for clustering. It could be speculated that the fluorescence dictionary is smaller since in the green light images, unlike the white light images, the prescence of a protein crystal is clearly defined by an increase in intensity of the green channel in the images, which often saturates, thereby loosing all texture detail. In the case of white light images, the translucent crystals and precipitates have a higher detectable content of texture information.

4.3.5 Overview of the Two Image Pipelines

A figure showing the white light and green light image pipelines side by side can be seen in Figure B.3, Appendix B.

4.4 Data Preparation for Machine Learning

Before training classifiers for the white light and green light images, the data for training and evaluation needed to be generated.

First, the set of images from our collaborator were reduced down into those just those with corresponding pairs of white and green light images. This reduced set consisted of 20 crystal plates with pairs of subwell images. There were therefore 5760 (288 × 20) green light images and 5760 white light images in the data set.

4.4.1 The Analyse Image Folders Programs

In order to calculate texton distributions for the images in the dataset, two programs were written, `Analyse Flu Image Folders` and the, very similar, program `Analyse White Image Folders`.

These programs:

- Take an input path to a folder and then for each image in the subfolders within, mask the image using the appropriate program (described in Section 4.2) and save the masked image.
- Calculate the texton distribution for the image using the appropriate dictionary (described in Section 4.3).
- Save the texton distribution along with the filepaths to both the original image and the masked image in a data structure.

After running these two programs on the images in the data set, the final, single, data structure contained 5760 rows, each corresponding to one subwell in the 20 crystal plates. Each row contained filepaths and texton distributions for the white light image and the green light image of the same droplet, thereby keeping the these data linked together.

4.4.2 Manually Scoring Images

To be able to train a classification algorithm and to evaluate classification performance, a ‘ground truth’ is needed. To train the classifier for white light images, currently in use at SGC, Oxford, this ground truth came from the scoring data for droplets which is recorded in the Crystal database. This data is generated when crystallographers look through images of their experiments and give them scores from a predefined scale (described previously in Section 1.3). This is, or course, a laborious process and is the reason why ranking the most interesting drops so that they can be viewed first is so useful. The images for this study, however, did not have scores attached. Therefore one of the stages in preparing data for machine learning and evaluation of effectiveness was to score these images manually.

4.4.2.1 Scoring GUI

To ease the process of scoring the set of 11520 images, a graphical user interface (GUI) was written with the aid of the GUIDE (GUI development environment) module in MATLAB. A screenshot of this GUI can be seen in Figure 4.12.

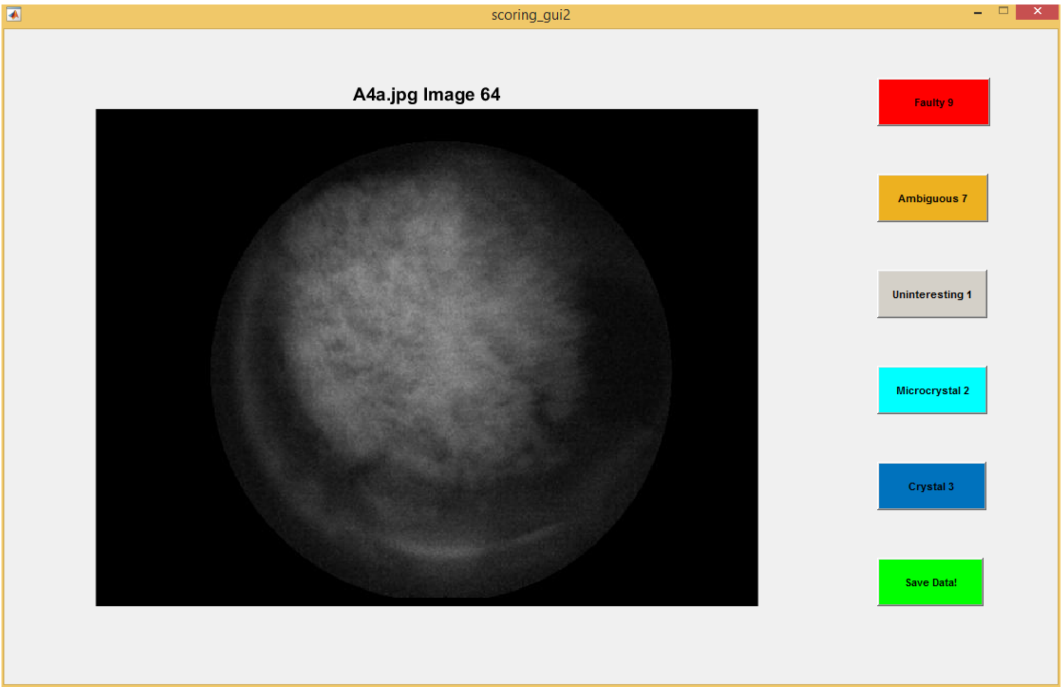


Figure 4.12: The image scoring GUI. Each masked image is viewed in turn and a score recorded to a data structure by either clicking a button or pressing a key.

The GUI reads in the data structure that was created by the `Analyse Image Folder` programs and which contains the filepaths to the masked images. Each image is shown in turn and a score added to the data structure by either clicking on one of the buttons or pressing a key. The scores given to the images and their corresponding categories are displayed in Table 4.1. The white light and green light images were scored independently of each other.

Score	**Category**
1	Uninteresting (empty or precipitate)
2	Microcrystal
3	Crystal
7	Ambiguous (small droplet or poor quality image)
9	Faulty (image masking fault)

Table 4.1: Scores given to the images and corresponding categories.

After scoring the images had been completed, the data structure contained all of the information needed to train machine learning algorithms and, in addition, contained data on the reliability of the masking programs and the proportion of ambiguous images.

4.4.3 The Evaluate Scoring Program

This program was written to analyse the scoring data and to exclude ambiguous results before training classifiers. If there was a fault with one image in a green light/white light pair, then both were removed. In addition, to ensure consistency of the data to be used for training, the program uses an exclusive-NOR (XNOR) operation in order to ensure that only pairs of images that are both scored as 'interesting' (either crystal or microcrystal) or both scored as 'uninteresting' are included in the final set of training and evaluation data in order to remove ambiguities. For example, if there is a salt crystal in the drop, the white light image would be marked as 'interesting' but the green light image would be 'uninteresting'. Therefore, hopefully, this process excluded some false positives as well as false negatives from the training data.

4.4.3.1 Output from Evaluate Scoring

`Evaluate Scoring` identified that the masking fault rate for white light images (circular Hough transform) was 3.2% whereas for the green light images (maximised sum of image intensity) it was 1.0%.

Excluding the image pairs where one or more was faulty or ambiguous led to the removal of 3302 image pairs (57.3%). In addition to over-saturated and out-of-focus images, this proportion was so high due to the large number of semi-full subwells which were found when viewing the white light images. Since masking was based upon simply finding the circular subwell in the image, this set of images were marked as ambiguous, as the droplet edge and empty regions of the subwell were likely to lead to spurious textures being detected. Table 4.2 shows the numbers of images assigned to each category once those which were faulty or ambiguous were removed.

	Faults/Ambiguous removed		**After XNOR**	
	White Light	**Green Light**	**White Light**	**Green Light**
Uninteresting	2192	2218	2148	2148
Microcrystal	102	113	90	60
Crystal	164	127	106	136
(Interesting)	266	240	196	196

Table 4.2: Breakdown of categories assigned to images after faulty and ambiguous pairs had been removed. The 'interesting' category is the sum of 'microcrystal' and 'crystal'. 'After XNOR' refers to the images remaining after an exclusive-NOR operation was applied in order to remove pairs where one image was assigned as 'interesting' and the other 'uninteresting'.

After the XNOR operation, an additional 114 pairs of images had been removed, leaving 2344 pairs of images in the final data set.

4.5 Training the Classifiers

The aim of this exercise is to train a classifier for a set of green light images and a separate classifier for images taken of the same droplets, but under white light. Training a classifier requires a set of predictors (in this case the frequency of each of the 300 white light textons or the 188 fluorescent textons) and a response ('interesting' or 'uninteresting'). Now that the data set had been reduced to pairs of scored, unambiguous images, each with an associated texton distribution, all the information required was now in place.

4.5.1 Partitioning the Data

Ultimately, aside from comparing the cross validation statistics generated during training, the performance of the two classifiers needed to be evaluated in terms of ranking sets of images. With this in mind, and given the numbers of 'interesting' and 'uninteresting' images remaining after removing ambiguities in the data set, it was decided to split the data into a number groups of image pairs that represented 'mini-crystal plates'. Each of these groups is similar to a hypothetical crystal plate in that it has number of 'interesting' and a number of 'uninteresting' images and it is the job of the classifier to rank the interesting plates to the top of the 'pile' to prevent the scientist from having to look through all of them in order to find a crystal.

4.5.1.1 The Data Partition program

This program divides the data (2148 'uninteresting' and 196 'interesting' image pairs) into 48 sets/'stacks' of 48 image pairs. Each stack contains 44 randomly selected 'uninteresting' and 4 randomly selected 'interesting' image pairs. These image pairs are shuffled into random positions in each stack.

A visualisation of the data output by this program can be seen in Figure B.4, Appendix B. Arranging the data like this meant that the data set consisted of 192 'interesting' and 2112 'uninteresting' image pairs.

4.5.1.2 The Training Set

The training set was formed by first taking the first 24 columns of the data set, resulting in a set that contained 96 'interesting' and 1056 'uninteresting' image pairs. The group of 'uninteresting' pairs was then reduced to 96 by random selection.

Before training, the white light and fluorescent texton distributions for the 'interesting' and 'uninteresting' images needed to extracted and normalised. Then an appropriate score of 1 for 'interesting' and 0 for 'uninteresting' appended to these normalised features. A program called `Extract and Normalise` was written to do this. Normalisation was carried out so that the texton distributions had a sum of 1.

4.5.2 Training the Random Forests

Training random forests was carried out in the Classification Learner module of MATLAB. An ensemble of 250 learners was trained on the white light data and a separate ensemble of the same size was trained on the fluorescence data. Classification statistics were calculated based upon 5-fold cross-validation. These statistics along with an evaluation of the ranking performance of both classifiers will be described in the next section (Section 4.6)

4.6 Evaluation

In this section the utility of texton analysis of green light images from crystallisation experiments containing trace-fluorescent labelled (TFL) protein will be compared to using the same texture analysis technique on standard white light images. The relative performance of the classifiers for the white light and green light images will be evaluated in terms of cross-validated statistics and ranking of the stacks of images in the test set of data (the 24 columns of data that were not chosen to form the training set in Section 4.5.1.2).

Ranking performance will be compared to the that of the classifier currently in use SGC, Oxford (Ng, Dekker, Kroemer, *et al.,* 2014) and, in addition, classification performance will be evaluated in comparison to the studies on detecting TFL protein crystals that have already been published in the literature (Sigdel, Pusey, *et al.,* 2013; Sigdel, Pusey, *et al.,* 2015, described earlier in this chapter, Sections 4.1.2.1 and 4.1.2.2).

4.6.1 Descriptors of Classification Model Performance

The 5-fold cross-validation gave an accuracy of 84.9% for the random forest trained on the white light images with the 300 texton features and accuracy of 92.7% for the random forest trained with the 188 fluorescence textons on the green light images. Confusion matrices and receiver operating characteristic (ROC) curves for the two classifiers can be seen in Figure 4.13. The area under the ROC curve was 0.928 for the white light image classifier and 0.965 for the green light image classifier.

As it has been shown that the area under the ROC curve is equivalent to the Wilcoxon test of ranks (Hanley & McNeil, 1982), this statistic alone can be used to describe the ranking performance of the the two classifiers, since this area is also the probability that any randomly chosen 'interesting' image

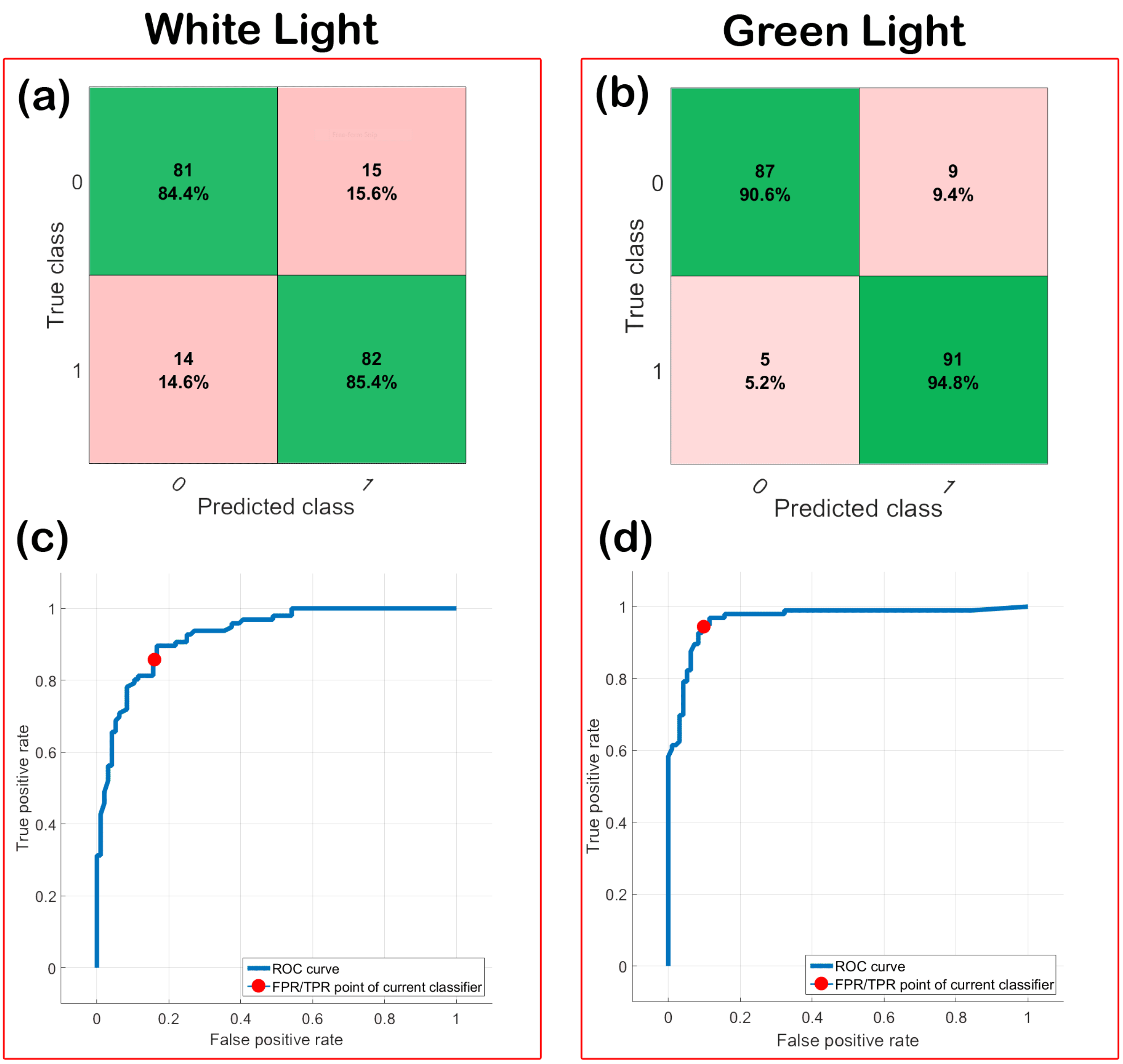


Figure 4.13: Statistics from 5-fold cross-validation of random forest classifiers. (a), Confusion matrix for the classifier trained on white light images. (b), Confusion matrix for the classifier trained on green light images. (c) Receiver operating characteristic (ROC) curve for the white light image classifier, area under curve = 0.928. (d) ROC curve for the green light image classifier, area under curve = 0.965. Red dots on curves indicate the False Positive Rate/True Positive Rate points of the classifiers.

will be ranked higher than any randomly chosen ‘uninteresting’ image. Therefore it can stated that the classifier for the green light images is superior to the classifier for the white light images in terms of ranking performance as well as accuracy (when using the cross-validated statistics generated from the training set). However, it is still interesting to see how the classifiers perform in terms of ranking on a test set of data.

4.6.2 Comparison of White Light and Green Light Image Ranking

Previously, the 48 stacks/columns of 48 image pairs were split in order to produce a training set and a test set of data. Now, the performance of the two random forest classifiers will be evaluated in terms of ranking either the white light image or the green light image in all of these pairs. Figure B.5, Appendix B, shows the test set of data with categories shown for the white light image in each pair. The dividing line between an image categorised as a crystal and one categorised as a microcrystal is subjective, and as such, the differences in these categories between the white and green light images should not be given much weight. The distinction is made in case of any obvious difference in ranking of these two types which could give clues as to the reasons why some ‘interesting’ images maybe given a higher rank (appear lower in the viewing order) than others.

In order to rank the data, a program, `Shuffle Rank`, was written. This program takes in the texton distributions of one of the image types from this data structure and ranks each stack according to the posterior probabilities output by the appropriate classifier. A figure showing the results of ranking the white light images and green light images in this way is shown in Figure 4.14.

It can be seen from Figure 4.14 that the ranking of the ‘interesting’ green light images is better than that for the white light images. The two ‘interesting’ white light images that were ranked most poorly were in ranks 42 and 33. The two ‘interesting’ green light images that were ranked most poorly, however, were in ranks 26 and 19. For the white light images, 16.7% of the ’interesting’ images were given a rank of grater than 10 whereas the equivalent figure for the green light images was 4.2%. For the two ’interesting’ green light images that were ranked most poorly, it can be seen that there is not much contrast between the crystal and the heavy precipitate in the background, however, the crystals do appear to be detected by the texton analysis (seen in the corresponding texton analysis images). For the ’interesting’ white light image that was ranked most poorly (column 3, row 46 in the middle panel of Figure 4.14) the cluster of crystals on the north side of the droplet are barely registered by the texton analysis (as can be seen in the corresponding texton-labelled image). The small crystal (on the east side half-way between the centre and edge of the drop) that can be seen in second most poorly ranked image(column 11, row 33 in the same panel) does get picked up by the analysis, however.

4.6.3 Comparisons to the Literature

In this section, the performance of the green light classifier will be compared to studies in the literature which either used texton analysis for ranking crystal images (Ng, Dekker, Kroemer, *et al.,* 2014) or focussed on automatically detecting crystals of Trace Fluorescent Labelled (TFL) protein (Sigdel, Pusey, *et al.,* 2013; Sigdel, Pusey, *et al.,* 2015, described in Section 4.1.2).

4.6.3.1 Comparison to Ng, Dekker, Kroemer, *et al.,* 2014

The random forest classifier, trained by Ng, 2015, that is currently used at SGC, Oxford was trained on 2501 ’interesting’ images and 3553 ‘uninteresting’ images and using an ensemble of 500 learners. The area under the ROC (AUROC) curve in this case was 0.942. Therefore, given the fact, mentioned earlier (Section 4.6.1), that the AUROC curve can be used to describe the ranking performance of a classifier, this implies that the 250 member random forest classifier trained on green light images in this study is superior (AUROC curve = 0.965). The accuracy of the classifier created by Ng, Dekker, Kroemer, *et al.,* 2014 is also lower at 89.3% compared to the 92.7% for the classifier from this study, although too much should not be read into accuracy scores as the classifier here is solely being used for ranking, so no classification threshold has been chosen.

In order to compare like with like, the fluorescent texton training data prepared previously (Section 4.5.1.2) was used to train a random forest with 500 learners instead of the 250 member forest evaluated earlier in this study. The new classifier has an AUROC curve of 0.972 which increases the difference between the fluorescence classifier and the one trained on white light images to 3%. This is important to note, since the white light images from the SGC that were used to train this classifier are very different to the to those used in this particular study (as shown in Figure 4.3, Section 4.2). In particular, (although the SGC images also suffer from focus and exposure problems or faulty droplets on occasion), the lighting across the droplets is much more consistent for the droplet images from the SGC. Therefore the fact that the ranking of green light images is still better (from cross validation on the training sets), validates how useful the TFL protein technique is when compared to standard imaging. A large scale ranking study needs to be performed for a fair comparison however.

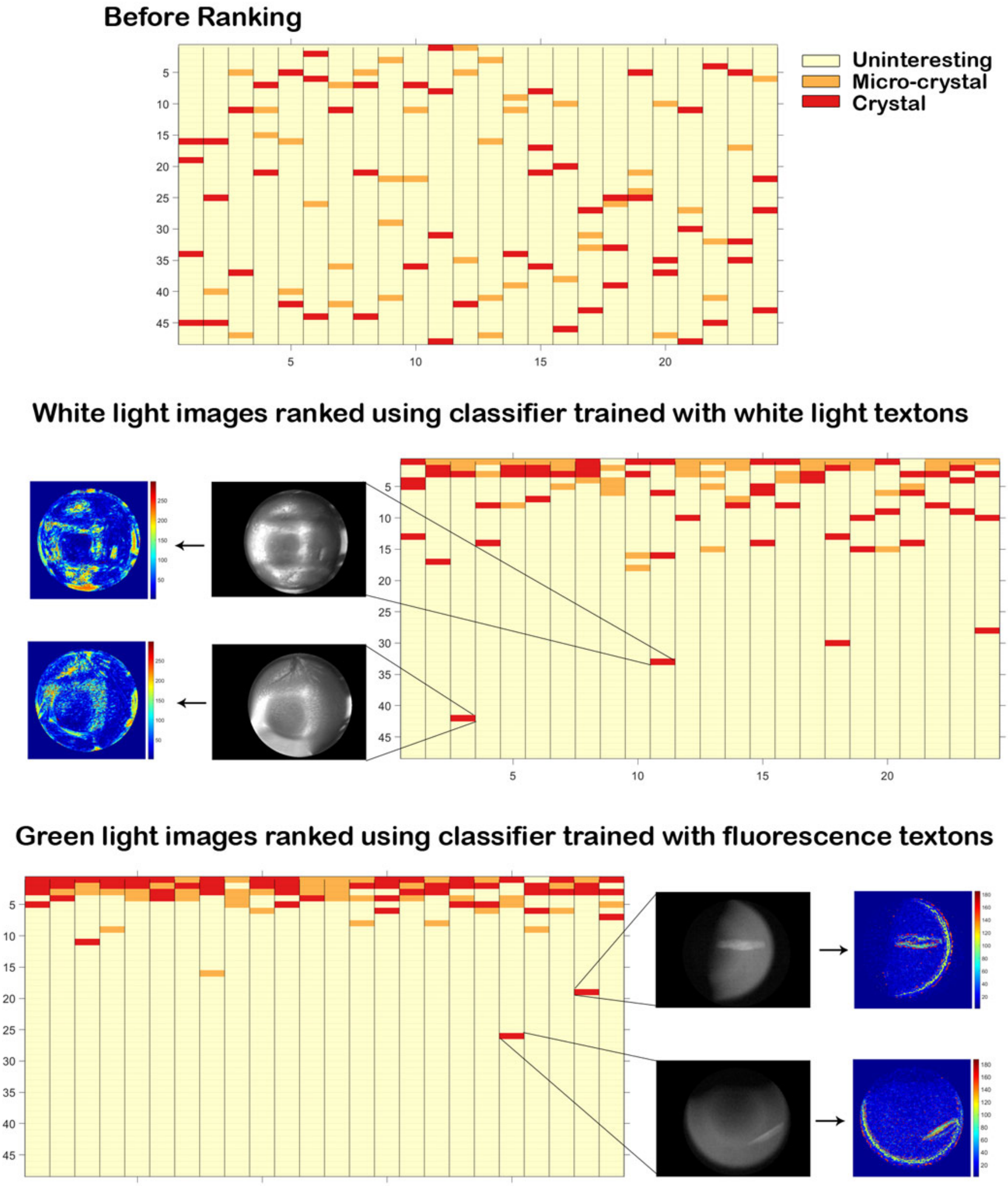


Figure 4.14: A comparison of the ranking of the test set by the two random forest classifiers. Top panel, The test set of 24 stacks/columns of 48 image pairs before ranking. Middle panel, The stacks after the white light images were ranked by the white light classifier. Bottom, The stacks after the green light images were ranked by the fluorescence classifier. Images and corresponding texton-labelled images are shown for the two that were ranked most poorly.

4.6.3.2 Comparison to Sigdel, Pusey, *et al.*, 2013

In this study (originally described in Section 4.1.2.1), the authors express the performance of their multilayer perceptron (MLP) neural network in terms of accuracy and false negatives when tested on 2250 images containing TFL protein. Since the random forest classifier in this current project is used for ranking rather than classification, the cross-validated scores from training of the classifier will be used for comparison (Figure 4.13, Section 4.6.1).

In terms of number of crystals missed by the classifier (false negative rate), the MLP neural network and the max-class ensemble of classifiers described by Sigdel, Pusey, *et al.,* 2013 have a much lower (2% and 1.2% respectively) false negative rate than our classifier (5.2%). The accuracy of their neural network and ensemble were also higher in terms of detection of non-crystals (true negative rate) with rates of 97% and 99% respectively, compared to 90.6% for the green light classifier trained in this study. The overall accuracy of our random forest classifier is higher (92.7%) than the 90% accuracy achieved by their MLP neural network. However, Sigdel, Pusey, *et al.,* 2013 were classifying their images into three classes, non-crystals, crystals and likely-leads so this needs to be taken into account when comparing the data.

4.6.3.3 Comparison to Sigdel, Pusey, *et al.,* 2015

The second study on images of TFL protein (originally described in Section 4.1.2.2) concentrated on crystal detection rather than classification into three categories. Therefore it is much more appropriate to compare the data from our green light classifier with the results from this particular paper. The thresholding and edge detection technique in this paper had a true positive rate of ≈ 89% whereas our green light classifier has a true positive rate of 94.8%. In addition, their technique had false positive rate of 23.9% compared to 9.4% for our classifier. From this comparison, texton analysis in conjunction with a random forest classifier seems to be a more accurate system than the methods described in this paper.

4.6.3.4 A Final Word on the Comparisons

Ideally, to perform a robust comparison between texton analysis of experiments containing TFL protein and the literature, the data set analysed and the end-goal would be the same across the comparisons. Therefore it would be better to evaluate the classification of a large number of images using our classifier for comparison with the literature, however the end-goal for our classifier is to rank images rather than classify. Because of this it is difficult to make a direct and meaningful comparison unless it is with a study where ranking is also the aim.

4.6.4 Insights and Challenges

This part of the project has given me a fascinating insight into image processing, texture analysis and machine learning. In a similar fashion to the first part of the project, however, it has been a challenging undertaking. Even though the first part of the project also served as an introduction to MATLAB, the structure of the data and the code base, the knowledge required for this, the second part of the project was almost completely separate. Once more, understanding how all the existing pieces of code fitted together was a time consuming task but a useful exercise.

There was a large amount of image processing background to be leant, (for example morphological opening with structured elements). Again, dealing with real world data meant that imperfections such as noise or out-of-focus images had to be dealt with on occasion. The large number of files to be processed and scored, along with computationally expensive tasks such as clustering vector responses meant that a lot of patience and organisation was required before this part of the project was at a stage to yield reliable results.

The machine learning aspect to this section of the project involved complex concepts but I am glad to have got the opportunity to have hands on experience using real data. In this case, doing is the certainly the best way of learning.

Once again, as in the rest of the project, explaining a complicated topic, with a large body of background information, in a way that is clear but not verbose has been a great challenge but a satisfying one to attempt.

Chapter 5

Summary and Conclusion

This document and each experimental section within includes a review of the relevant literature and background information required to understand this study. The topic is large and contains many biochemical concepts. However, attempts have been made to translate these ideas from the complex information in the literature into details that are directly relevant to the work to be undertaken. This review of the literature found that:

- **Scientists go to great lengths to try and obtain a crystal of a protein**, setting up thousands of microscopic experiments and thereby generating thousands of images that have to be viewed (Chapter 1).
- **The use of textons is a promising technique in texture analysis of crystallisation micrographs.** Much work has previously been done on detecting the 'best case' scenario for experiments, (i.e. crystals), but analysis of texture allows precipitation outcomes and clear drops to be included in addition to crystallisation (Chapter 2).
- **Improving protein solubility can also improve the chances of growing a crystal.** The automated evaluation of clear drops can be used to determine conditions to be fed back into experiments to achieve this increased solubility. These facts add weight to the notion that identifying clear drops in historical data, collating this information into a set of conditions and using this set as a solubility tool could be useful to increase crystallisation success rates.
- **Labelling protein molecules with a fluorescent dye can be useful to aid the detection of crystals.** To date, only two studies have been published on computational analysis of images taken of crystallisation experiments that utilise this labelling technique. The first study extracted image features and used machine learning to classify them, only 1.2% of crystals were missed, however 31% of them were placed into the wrong classification group. In the second study, image thresholding and edge detection were used to detect crystals, however $\approx$ 11% of crystal-containing experiments were missed altogether. Therefore there is room for improvement in analysis of these images (Chapter 4).

Whilst executing the project, two main areas have been investigated, namely (i) using results from the analysis of crystallisation droplet images in order to create a tool to improve the solubility of proteins (Chapter 3) and (ii) the use of texton analysis and machine learning to assess experiments containing protein labelled with fluorescent dye (Chapter 4).

An evaluation has been carried out for both of these parts. For the solubility tool, the evaluation took the form of a 'virtual' experiment where a set of independent data from laboratory crystallisation experiments was tested against the tool *in silico*. **This tool was compared to a set of random conditions and was found to be better with statistical significance ($p < 0.05$).**

For the work relating to detecting fluorescent crystals, matched pairs of droplet images, collected under different lighting conditions were used to train two random forest classifiers. The techniques of using texton analysis alone or in conjunction with fluorescence could then be directly compared, both in terms of descriptors of classification model performance and in terms of ranking performance on a test set of image pairs. This led to the conclusion that **the combination of using Trace Fluorescent Labelled (TFL) protein and texton analysis allows for more effective ranking of 'interesting' experimental outcomes than texton analysis of standard white light images alone.**

Also, this combination of technologies was compared to studies in the literature on automated detection of TFL protein crystals. The texton method in combination with labelling was found to be more accurate, in terms of ranking and based upon areas under the ROC curve than using texton analysis on white light images from the SGC (Ng, Dekker, Kroemer, *et al.,* 2014). It was found not to as effective at classification as feature extraction from green light images in combination with a multilayer perceptron neural network (Sigdel, Pusey, *et al.,* 2013) (although it is difficult to directly compare the data) and more accurate than using a combination of thresholding and edge detection for finding flourescent protein crystals Sigdel, Pusey, *et al.,* 2015.

In conclusion, this project has produced an evaluated tool, and the methodology to recreate it, that may allow scientists to increase the solubility of the protein that they are working on, thereby increasing their chances of obtaining a crystal of high quality. In addition, the changes made to the image processing pipeline of TeXRank and the creation of a new texton dictionary, as well as a random forest classifier, should aid the experimenter in identifying a crystal and reduce the number of images that need to be viewed in doing so. Ultimately, once in possession of such a crystal, they can then obtain an accurate atomic structure of the protein in order to answer questions related to disease processes and the function of the molecule.

Chapter 6

Future Work

Here are some suggestions for future work, some of which is likely to happen and some that it would have been interesting to carry out had more time been available.

6.1 The Solubility Tool

- **Standardising the screen conditions data and retrieving missing images from tape archives.** Having consistent chemical component description strings in the database would allow and a larger set of conditions and screens to be included in the creation of a solubility tool. In turn, this is likely to improve the effectiveness of the tool. Having access to more images would increase the amount of data available and increase the number of proteins that could be included in the creation and evaluation of the tool.
- **Monte-Carlo cross-validation of the virtual experiment.** This idea is more speculative, but it would have been interesting to repeat the creation and validation of the solubility screen tool a large number of times and then generate statistics based upon the averaged results.

6.2 Combining Texton Analysis and Fluorescent Protein Labelling

- **Analysis of the minimum texton dictionary size required for effective ranking/classification.** This could be perhaps be performed using principle component analysis (PCA).
- **Creating an ensemble from the white light and fluorescence classifiers.** Could the ranking performance be improved further than using either classifier alone?
- **Training the fluorescence classifier with a larger set of images.** It should be possible to improve performance as only 96 images in each class were used to train the classifier in this study. This is likely to happen, since SGC, Oxford has just invested in new imaging equipment that can record images of fluorescent labelled protein.
- **Preparation of a manuscript for publication based on this work.** This is also likely to happen since this is a novel contribution to the field and my external supervisor is keen to share this information with the scientific community.

Bibliography

1. Bern, M, Goldberg, D, Stevens, RC & Kuhn, P. Automatic classification of protein crystallization images using a curve-tracking algorithm. *Journal of Applied Crystallography* **37,** 279–287 (April 2004).
2. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (October 2001).
3. Canny, J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8,** 679–698 (November 1986).
4. Chaikuad, A, Knapp, S & von Delft, F. Defined PEG smears as an alternative approach to enhance the search for crystallization conditions and crystal-quality improvement in reduced screens. *Acta Crystallographica Section D: Biological Crystallography* **71,** 1627–1639 (August 2015).
5. Collins, BK, Stevens, RC & Page, R. Crystallization Optimum Solubility Screening: using crystallization results to identify the optimal buffer for protein crystal formation. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **61,** 1035–1038 (Pt 12 November 2005).
6. Collins, BK, Tomanicek, SJ, Lyamicheva, N, Kaiser, MW & Mueser, TC. A preliminary solubility screen used to improve crystallization trials: crystallization and preliminary X-ray structure determination of Aeropyrum pernix flap endonuclease-1. *Acta Crystallographica Section D: Biological Crystallography* **60,** 1674–1678 (September 2004).
7. Corduneanu, A & Bishop, CM. Variational Bayesian model selection for mixture distributions. *Artificial intelligence and Statistics* **2001,** 27–34 (2001).
8. Cross, GR & Jain, AK. Markov Random Field Texture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5,** 25–39 (January 1983).
9. Cumbaa, CA & Jurisica, I. Protein crystallization analysis on the World Community Grid. *Journal of Structural and Functional Genomics* **11,** 61–69 (March 2010).
10. Fisher, R, Perkins, S, Walker, A & Wolfart, E. *Morphology - Opening.* Image Processing Learning Resources. <http://homepages.inf.ed.ac.uk/rbf/HIPR2/open.htm> (2004).
11. Forsythe, E, Achari, A & Pusey, ML. Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography* **62,** 339–346 (March 2006).
12. *Hampton Research Website* <https://hamptonresearch.com/product_detail.aspx?cid=10&sid=182&pid=568> (2016).
13. Hanley, JA & McNeil, BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (April 1982).
14. Haralick, RM. Statistical and structural approaches to texture. *Proceedings of the IEEE* **67,** 786–804 (May 1979).
15. Izaac, A, Schall, CA & Mueser, TC. Assessment of a preliminary solubility screen to improve crystallization trials: uncoupling crystal condition searches. *Acta Crystallographica Section D: Biological Crystallography* **62,** 833–842 (July 2006).
16. Jancarik, J & Kim, SH. Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography* **24,** 409–411 (August 1991).
17. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **290,** 91–97 (March 1981).
18. Kulis, B & Jordan, MI. Revisiting k-means: New Algorithms via Bayesian Nonparametrics. *arXiv:1111.0352 [cs, stat].* arXiv: 1111.0352. <http://arxiv.org/abs/1111.0352> (visited on May 6, 2016) (November 2011).
19. Leung, T & Malik, J. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision* **43,** 29–44 (June 2001).
20. Liu, R, Freund, Y & Spraggon, G. Image-based crystal detection: a machine-learning approach. *Acta Crystallographica Section D: Biological Crystallography* **64,** 1187–1195 (Pt 12 November 2008).

21. Luft, JR, Newman, J & Snell, EH. Crystallization screening: the influence of history on current practice. *Acta Crystallographica. Section F, Structural Biology Communications* **70,** 835–853 (Pt 7 June 2014).
22. Malik, J, Belongie, S, Shi, J & Leung, T. *Textons, contours and regions: cue integration in image segmentation* in *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999* The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999. **2** (1999), 918–925 vol.2.
23. McPherson, A & Gavira, JA. Introduction to protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications* **70,** 2–20 (January 2014).
24. Newman, J *et al.* Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallographica Section D: Biological Crystallography* **61,** 1426–1431 (October 2005).
25. Ng, JT, Dekker, C, Reardon, P & von Delft, F. Lessons from ten years of crystallization experiments at the SGC. *Acta Crystallographica Section D: Structural Biology* **72,** 224–235 (February 2016).
26. Ng, JT. *Precipitation Pattern as a Proxy for Protein Crystallization* PhD thesis (University of Oxford, September 2015).
27. Ng, JT, Dekker, C, Kroemer, M, Osborne, M & von Delft, F. Using textons to rank crystallization droplets by the likely presence of crystals. *Acta Crystallographica Section D: Biological Crystallography* **70,** 2702–2718 (Pt 10 September 2014).
28. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9,** 62–66 (January 1979).
29. Pusey, ML, Paley, MS, Turner, MB & Rogers, RD. Protein Crystallization Using Room Temperature Ionic Liquids. *Crystal Growth & Design* **7,** 787–793 (April 2007).
30. Pusey, ML *et al.* Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications* **71,** 806–814 (July 2015).
31. *RCSB Protein Data Bank Website* <http://www.rcsb.org/> (2016).
32. Russo Krauss, I, Merlino, A, Vergara, A & Sica, F. An Overview of Biological Macromolecule Crystallization. *International Journal of Molecular Sciences* **14,** 11643–11691 (May 2013).
33. *SGC | Structural Genomics Consortium Website* <http://www.thesgc.org/> (2016).
34. Sigdel, M, Pusey, ML & Aygun, RS. CrystPro: Spatiotemporal Analysis of Protein Crystallization Images. *Crystal Growth & Design* **15,** 5254–5262 (2015).
35. Sigdel, M, Pusey, ML & Aygun, RS. Real-Time Protein Crystallization Image Acquisition and Classification System. *Crystal Growth & Design* **13,** 2728–2736 (July 2013).
36. Sokal, RR & Rohlf, FJ. *Biometry : the principles and practice of statistics in biological research* 4th (W.H. Freeman, New York, 2012).
37. Spraggon, G, Lesley, SA, Kreusch, A & Priestle, JP. Computational analysis of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography* **58,** 1915–1923 (November 2002).
38. Varma, M & Zisserman, A. A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision* **62,** 61–81 (April 2005).
39. Ward, KB, Perozzo, MA & Zuk, WM. Automatic preparation of protein crystals using laboratory robotics and automated visual inspection. *Journal of Crystal Growth* **90,** 325–339 (July 1988).
40. Watts, D, Cowtan, K & Wilson, J. Automated classification of crystallization experiments using wavelets and statistical texture characterization techniques. *Journal of Applied Crystallography* **41,** 8–17 (February 2008).
41. Wilson, J. Towards the automated evaluation of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography* **58,** 1907–1914 (November 2002).
42. Xie, X & Mirmehdi, M. in *Handbook of Texture Analysis* 375–406 (Imperial College Press, 2008).
43. Zuk, WM & Ward, KB. Methods of analysis of protein crystal images. *Journal of Crystal Growth* **110,** 148–155 (March 1991).

Appendix A

Solubility Tool Data

A.1 Initial Beehive Query

```
SELECT T1."PROTEIN_ID" "Protein", CONCAT(DISTINCT T4."Barcode1") "Xtal Plate
    Barcode", MAX(T2."CONCENTRATION") "Xtal Plate Protein Concentration"
FROM "SGC"."PROTEIN" T1
JOIN "SGC"."XTAL_PLATES" T2 ON T2."SGCPROTEIN_PKEY" = T1."PKEY"
JOIN "SGC"."XTAL_SCREENID" T3 ON T3."SGCXTALPLATEID_PKEY" = T2."PKEY"
JOIN (SELECT T4."BARCODE" "Barcode1", T4."BARCODE" "Barcode2" FROM
    "SGC"."XTAL_PLATES" T4) T5 ON T5."Barcode2" = T2."BARCODE" WHERE T3."SCREENNAME"
    LIKE 'JCSG' OR LIKE 'HCS' OR LIKE 'HIN' OR LIKE 'BCS' OR LIKE 'LFS')
GROUP BY T1."PROTEIN_ID"
ORDER BY MAX(T2."CONCENTRATION")
```

Listing A.1: Simplified Beehive database query. The actual query that was run involved many more table joins and the exclusion of protein classes outside of the scope of this study (such as integral membrane proteins and protein-protein complexes).

A.2 Solubility Screen Conditions from Refinement Round I

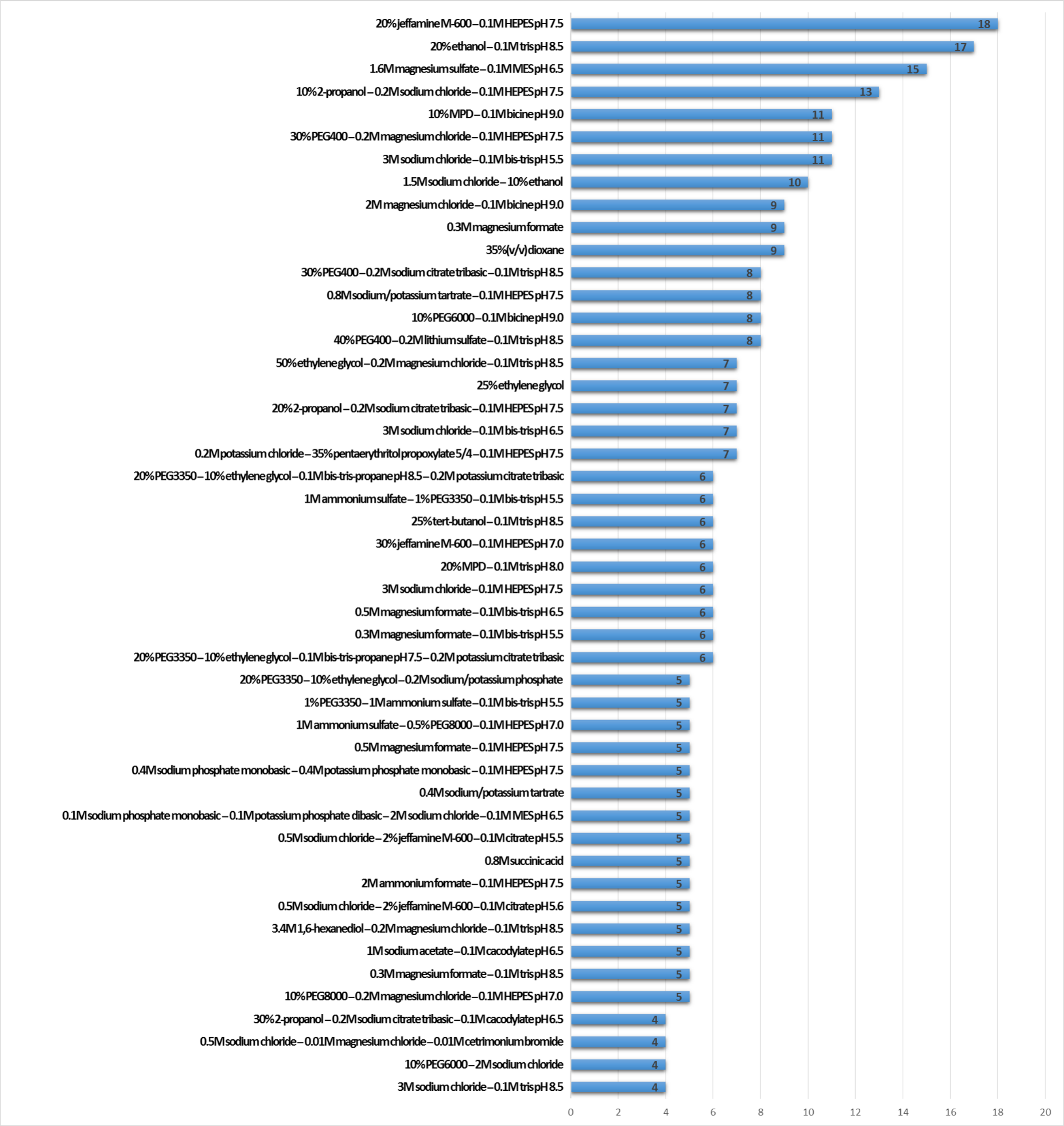


Figure A.1: Frequency histogram of the top 48 conditions found by the Clear Drop Query program. Refinement round I. Only plates with a precipitation score of > 5 were included. Three clear subwell drops, each with a clear drop score of > 0.6 were needed for a well condition to be recorded.

A.3 Parent Screens for Solubility Conditions from Refinement Round I

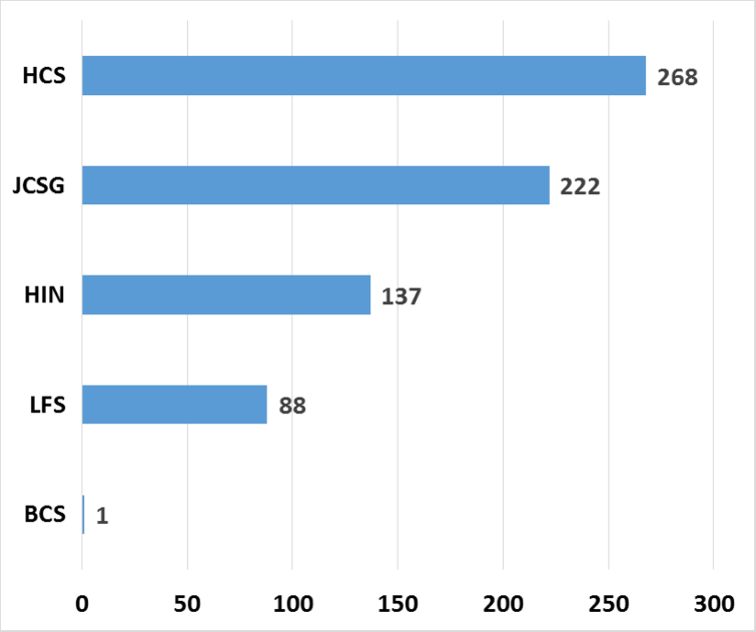


Figure A.2: Frequency histogram of the parent screens for all conditions found by the Clear Drop Query program. Refinement round I. Only plates with a precipitation score of > 5 were included. Three clear subwell drops, each with a clear drop score of > 0.6 were needed for a well condition to be recorded. This data shows parent screen types for all conditions recorded, not just the top 48.

A.4 Proteins that Formed the Data Set for Generating the Solubility Tool

SGC ID	**Protein Description**	**Species Name**	**Max. Conc.(mg ml^{-1})**
AIPL1A	aryl-hydrocarbon-interacting protein-like 1 isoform 1	*Homo sapiens*	10.6
ALPK1MMA	alpha-kinase 1	*Mus musculus*	9.5
ASB11A	ankyrin repeat and SOCS box-containing protein 11	*Homo sapiens*	11.0
CYFIP2A	cytoplasmic FMR1-interacting protein 2	*Homo sapiens*	3.3
DCLRE1AA	DNA cross-link repair 1A	*Homo sapiens*	10.0
DCLRE1CA	DNA cross-link repair 1C [isoform a]	*Homo sapiens*	10.0
DOPVA	pup deamidase/depupylase	*Mycobacterium tuberculosis*	10.3
EGLN3A	EGL nine homolog 3	*Homo sapiens*	10.0
FAM83FA	family with sequence similarity 83, member F	*Homo sapiens*	10.0
FAM83GA	family with sequence similarity 83, member G	*Homo sapiens*	10.1
FERA	fer (fps/fes related) tyrosine kinase	*Homo sapiens*	6.5
FKBP10A	FK506 binding protein 10, 65 kDa	*Homo sapiens*	7.4
ITKA	IL2-inducible T-cell kinase	*Homo sapiens*	10.0
JIKA	JIK: STE20-like kinase	*Homo sapiens*	3.5
KBTBD8A	kelch repeat and BTB domain-containing protein 8	*Homo sapiens*	10.0
KCTD18A	BTB/POZ domain-containing protein KCTD18	*Homo sapiens*	10.0
KIAA1361A	serine/threonine protein kinase TAO1 homolog	*Homo sapiens*	6.5
MAGEA3A	melanoma antigen family A, 3	*Homo sapiens*	10.0
MAPK3A	mitogen-activated protein kinase 3	*Homo sapiens*	10.7
MAPK6A	mitogen-activated protein kinase 6	*Homo sapiens*	10.1
MRE11AA	meiotic recombination 11 homolog A isoform 1	*Homo sapiens*	7.8
PB1A	polybromo 1 [isoform a]	*Homo sapiens*	6.5
PCTK2A	PCTAIRE protein kinase 2	*Homo sapiens*	7.9
PIP3EA	phosphoinositide-binding protein PIP3-E	*Homo sapiens*	9.5
POP2RSZ	type III effector protein popp2	*Ralstonia solanacearum*	10.1
RIPK4A	ankyrin repeat domain 3	*Homo sapiens*	8.0
SND1A	staphylococcal nuclease domain containing 1	*Homo sapiens*	10.9
SP110C	SP110 nuclear body protein [isoform c]	*Homo sapiens*	7.21
TRIB2A	tribbles homolog 2	*Homo sapiens*	6.9
TRIM28A	tripartite motif-containing 28	*Homo sapiens*	8.0
ZFYVE16A	zinc finger FYVE domain-containing protein 16	*Homo sapiens*	10.0
		Average	**8.8**

Table A.1: Details of the 31 proteins that formed the data set used to generate the solubility screen tool. Max. Conc. refers to the highest concentration of the protein used to set up crystal plates. (Refinement round II, Section 3.7.2)

A.5 Evaluation: Random Screen Conditions

Chemical Cocktail Condition
0.1M sodium chloride, 20% PEG500MME, 0.1M bicine *p*H 9.0
20% MPD, 0.1M tris *p*H 8.0
20% PEG8000, 0.2M magnesium chloride, 0.1M tris *p*H 8.5
1.6M sodium citrate tribasic
30% PEG4000, 0.2M magnesium chloride, 0.1M tris *p*H 8.5
40% MPD, 0.1M tris *p*H 8.0
0.1M ammonium acetate, 17%(w/v) PEG10000, 0.1M bis-tris *p*H 5.5
10% PEG6000, 0.1M bicine *p*H 9.0
3.4M 1,6-hexanediol, 0.2M magnesium chloride, 0.1M tris *p*H 8.5
0.2M ammonium citrate, 20% PEG3350
0.2M ammonium sulfate, 25% PEG4000, 0.1M acetate *p*H 4.5
16% PEG8000, 20% glycerol, 0.16M calcium acetate, 0.1M cacodylate *p*H 6.5
20% PEG6000, 0.1M citrate *p*H 5.0
2M ammonium sulfate, 0.1M acetate *p*H 4.5
10% PEG6000, 2M sodium chloride
0.2M ammonium sulfate, 30% PEG5000MME, 0.1M MES *p*H 6.5
30% MPD, 0.2M ammonium acetate, 0.1M citrate *p*H 5.5
30% PEG400, 0.2M magnesium chloride, 0.1M HEPES *p*H 7.5
20% PEG3000, 0.2M zinc acetate, 0.1M HEPES *p*H 7.5
0.1M magnesium formate, 15% PEG3350
12% PEG3350, 0.005M $CoCl_2$, 0.005M $CdCl_2$, 0.005M $NiCl_2$, 0.005M $MgCl_2$, 0.1M HEPES *p*H 7.5
28% PEG400, 0.2M calcium chloride, 0.1M HEPES *p*H 7.5
3M sodium chloride, 0.1M acetate *p*H 4.5
20% ethanol, 0.1M tris *p*H 8.5
2M ammonium sulfate, 0.1M bis-tris *p*H 6.5
0.2M sodium chloride, 25% PEG3350, 0.1M bis-tris *p*H 6.5
20% PEG3000, 0.1M citrate *p*H 5.5
40% PEG300, 0.1M citrate *p*H 4.2
50% PEG200, 0.2M sodium chloride, 0.1M sodium/potassium phosphate *p*H 7.5
20% PEG3350, 0.2M magnesium formate
0.2M ammonium acetate, 25% PEG3350, 0.1M bis-tris *p*H 6.5
30% MPD, 0.2M sodium chloride, 0.1M acetate *p*H 4.5
5% PEG1000, 40% ethanol, 0.1M citrate *p*H 4.2
0.01M cobalt chloride, 20% polyvinylpyrrolidone, 0.1M tris *p*H 8.5
20% jeffamine M-600, 0.1M HEPES *p*H 7.5
25% PEG1000, 20% glycerol
20% PEG3350, 0.1M succinic acid
30% PEG2000MME, 0.15M potassium bromide
25% PEG3350, 0.1M HEPES *p*H 7.5
0.2M magnesium chloride, 25% PEG3350, 0.1M bis-tris *p*H 5.5
2.1M DL- malic acid
45% MPD, 0.2M ammonium acetate, 0.1M HEPES *p*H 7.5
0.2M ammonium acetate, 45% MPD, 0.1M bis-tris *p*H 6.5
0.05M ammonium sulfate, 30% pentaerythritol ethoxylate 15/4, 0.1M bis-tris *p*H 6.5
25% PEG3350, 0.1M tris *p*H 8.5
10% 2-propanol, 0.2M zinc acetate, 0.1M cacodylate *p*H 6.5
5% MPD, 10% PEG6000, 0.1M HEPES *p*H 7.5
20% PEG6000, 0.1M bicine *p*H 9.0

Table A.2: A list of the 48 chemical cocktail conditions that were selected at random from the set of 265 unique conditions found in the HIN3, HCS3 and JCSG7 sparse-matrix screens. These were used for evaluation of the test set of proteins (Refinement round II, Section 3.7.2)

Appendix B

Fluorescent Crystal Detection

B.1 Building the Texton Dictionary

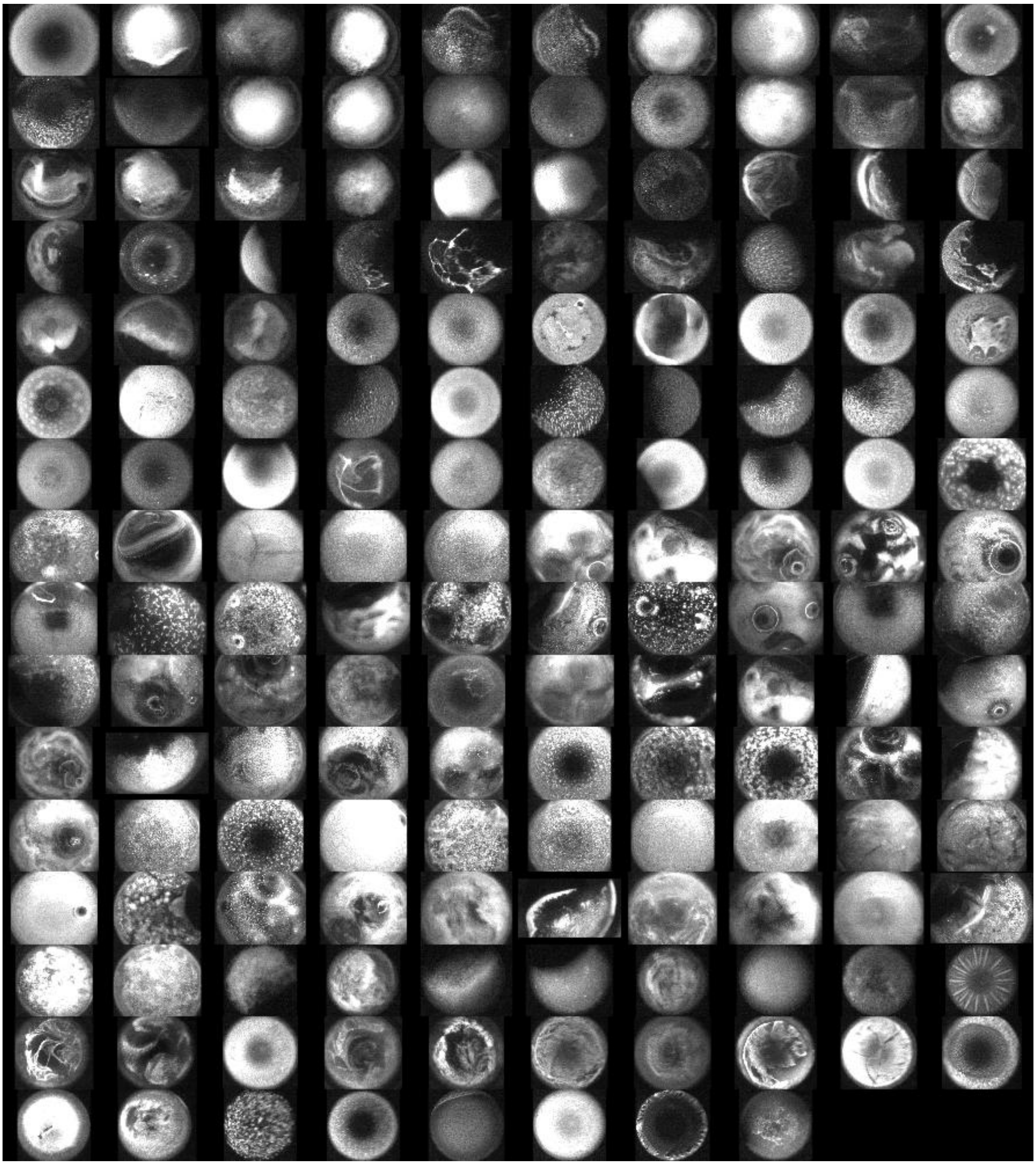


Figure B.1: The 158 cropped precipitation images, collected under green light that were used to build the dictionary.

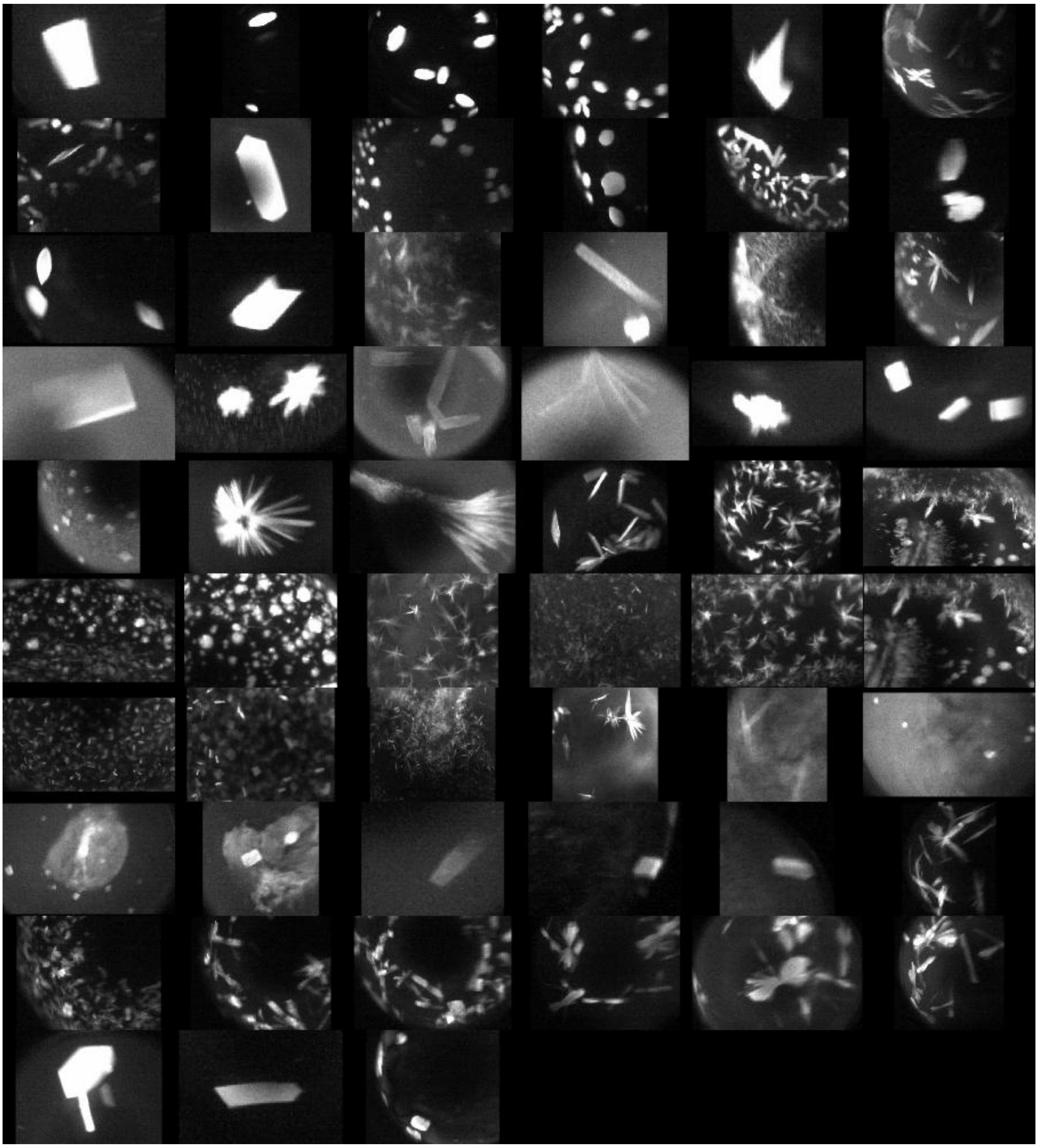


Figure B.2: The 57 cropped crystal images, collected under green light that were used to build the dictionary.

B.2 Overview of the Imaging Pipeline for Both Image Types

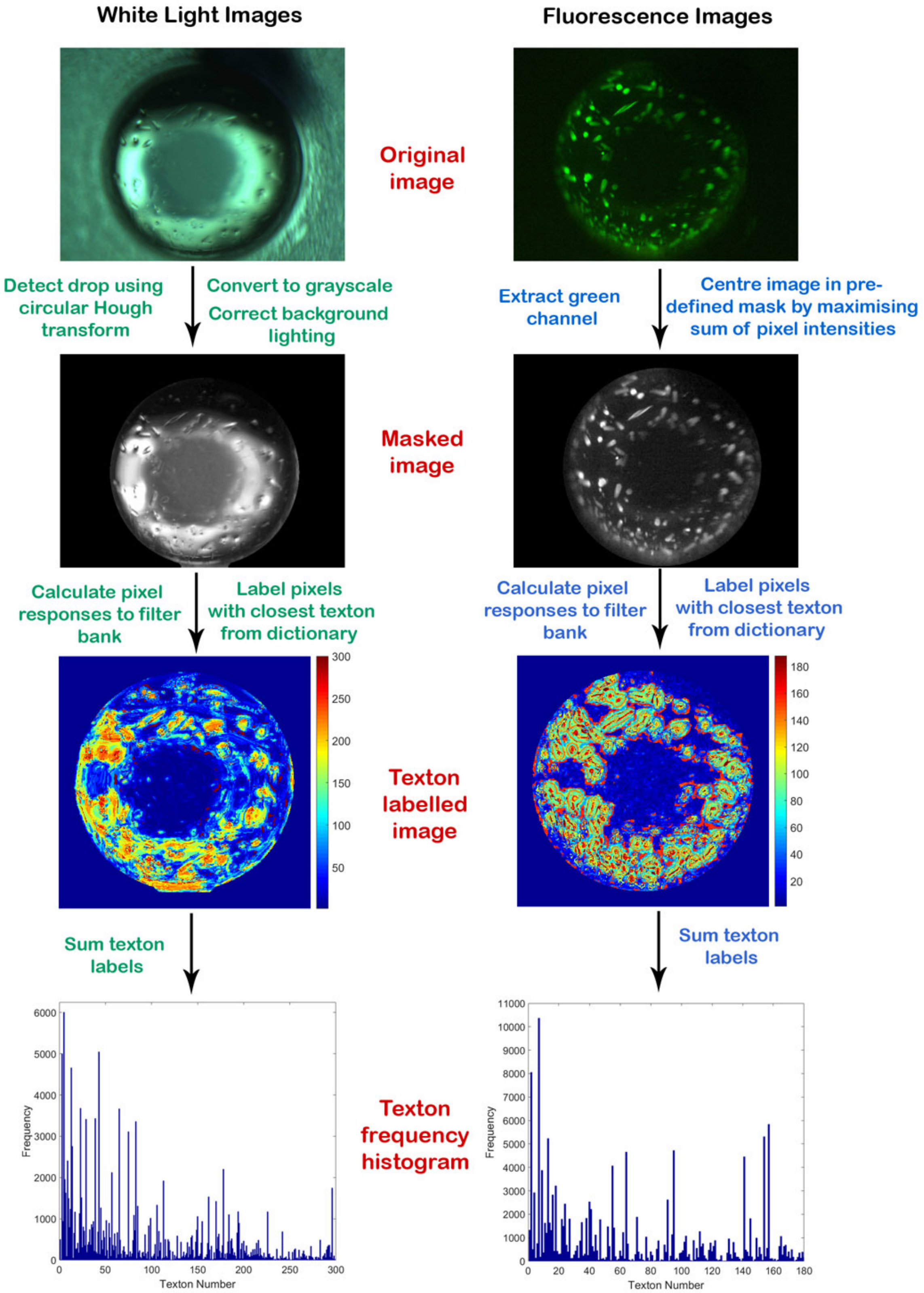

White Light Images
Fluorescence Images
Original image
Detect drop using circular Hough transform
Convert to grayscale
Correct background lighting
Extract green channel
Centre image in pre-defined mask by maximising sum of pixel intensities
Masked image
Calculate pixel responses to filter bank
Label pixels with closest texton from dictionary
Calculate pixel responses to filter bank
Label pixels with closest texton from dictionary
300
250
200
150
100
50
180
160
140
120
100
80
60
40
20
Texton labelled image
Sum texton labels
Sum texton labels
Texton frequency histogram
Frequency
Texton Number
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180

Figure B.3: An overview of the image processing pipeline for both image types.

B.3 Partitioning the Data

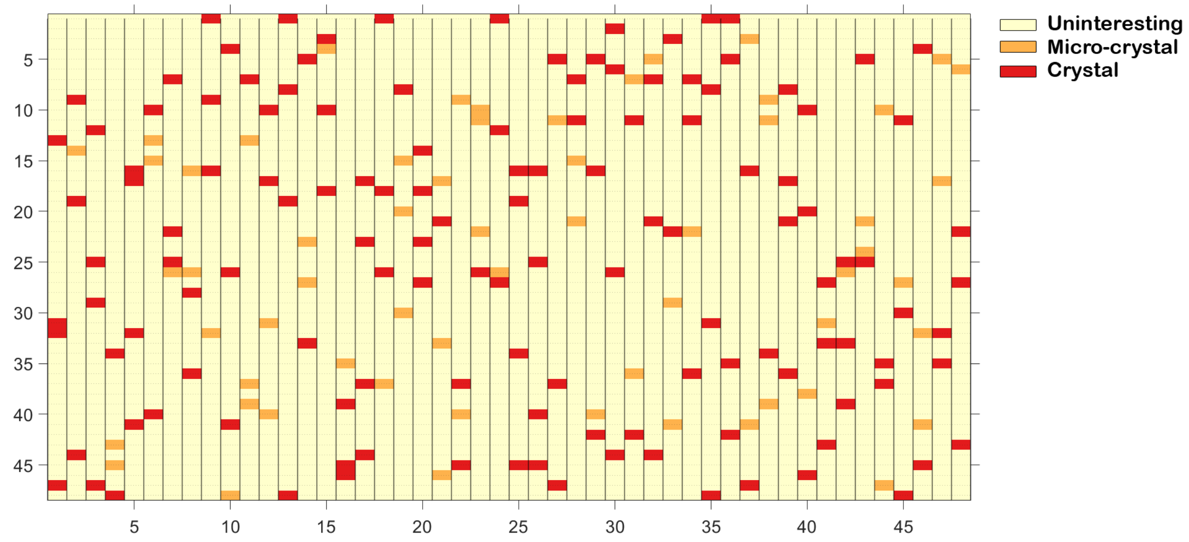


Figure B.4: The data set of images divided in to 48 randomised stacks/columns of 48 images. Each column represents one 'mini-plate'.

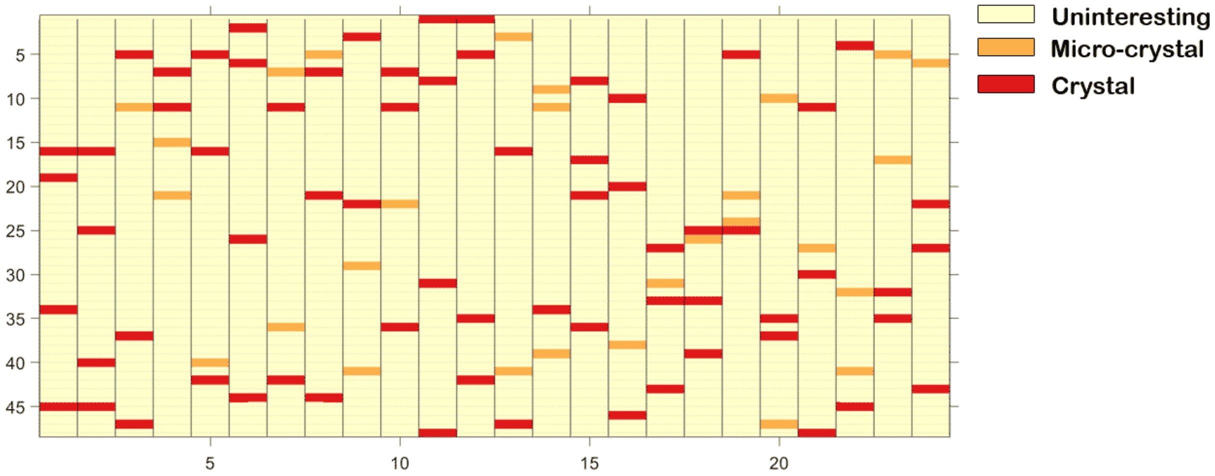


Figure B.5: The test set of image pairs before ranking with white light image categories shown.

B.4 Comparison to the Literature

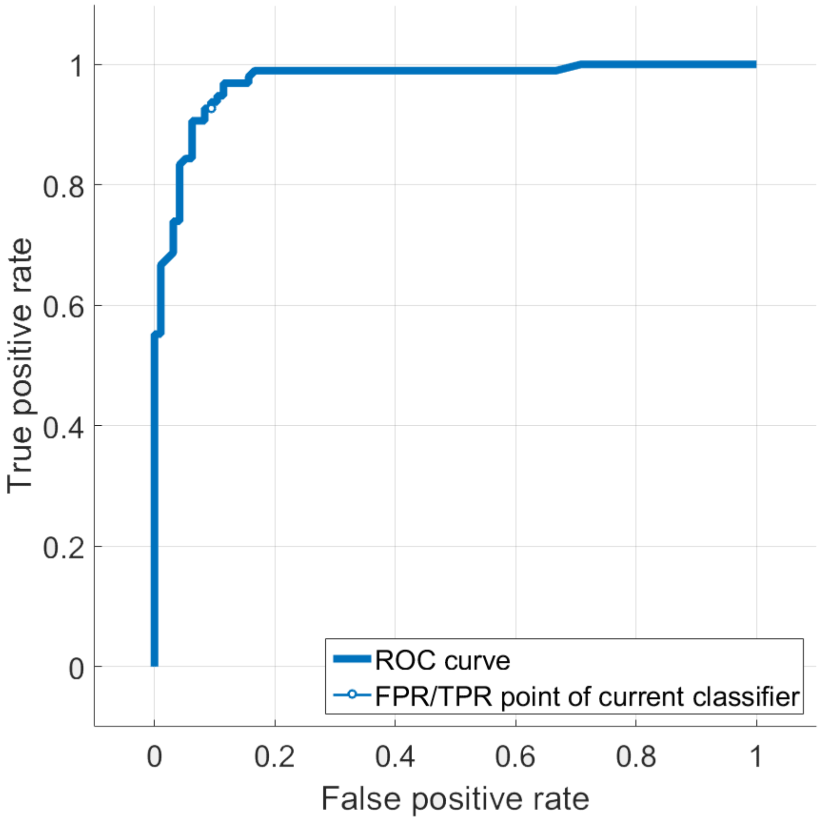


Figure B.6: Reciever operating charcteristic curve for a random forest with 500 learners and trained on the texton distributions from green light images. Area under curve = 0.972.