SLAC-R-998
August 2012

Papers from U.S. Department of Energy Science Undergraduate Laboratory Internship Program (SULI 2011)

SLAC National Accelerator Laboratory, Stanford University, Menlo Park, CA 94025

Work supported by Department of Energy contract DE-AC02-76SF00515.

Table of Contents

SLAC-TN-12-017

INSPIRE and SPIRES Log File Analysis

Cole Adams
Science Undergraduate Laboratory Internship Program

Wheaton College

SLAC National Accelerator Laboratory

August 5, 2011

Prepared in partial fulfillment of the requirement of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Travis Brooks in the Scientific Computing division at SLAC National Accelerator Laboratory

Participant Signature: ___________________________________

Research Adviser Signature:______________________________

Work supported by US Department of Energy contract DE-AC02-76SF00515.

ABSTRACT

SPIRES, an aging high-energy physics publication data base, is in the process of being replaced by INSPIRE. In order to ease the transition from SPIRES to INSPIRE it is important to understand user behavior and the drivers for adoption. The goal of this project was to address some questions in regards to the presumed two-thirds of the users still using SPIRES. These questions are answered through analysis of the log files from both websites. A series of scripts were developed to collect and interpret the data contained in the log files. The common search patterns and usage comparisons are made between INSPIRE and SPIRES, and a method for detecting user frustration is presented. The analysis reveals a more even split than originally thought as well as the expected trend of user transition to INSPIRE

INTRODUCTION

SPIRES is the high-energy physics publication database that has been used for collaboration and communication between physicists since the late 1960s. However, forty years is a long time in terms of technology, and SPIRES is in the process of being replaced by a new, similar service called INSPIRE. INSPIRE builds off of the usefulness of SPIRES and hopes to provide the high-energy physics community with better tools for successful research. INSPIRE provides a much faster alternative to SPIRES with a greater depth of features. INSPIRE functions as the quality SPIRES content built on the modern Invenio digital document repository platform developed at CERN

While INSPIRE is an incredibly high-quality tool, it is not yet a finished product. Currently INSPIRE is in a public beta phase and it will become a full production service in the Fall of 2011. Any service can use improvement and the best way to ensure that this improvement is most beneficial is to get feedback from the users themselves. Having a user base that provides constant feedback about the website will give the developers a clear focus of what will help the users most.

At the moment, the user base is split between SPIRES and INSPIRE, and originally it was thought that roughly two-thirds of the users still were using SPIRES, based on a simple metric of total searches. There were two questions to be answered about these two groups:

1. What are the distinctions between the two groups?
2. Is the two-thirds division the correct one?

In order to answer these questions we analyzed the log files for user behavior to find a better usage metric and to describe the differences between user behavior in INSPIRE and SPIRES. The methods used in this project revealed the presence of a large number of robots that artificially increased the search counts on SPIRES, making the original usage metric a poor one. Using a new metric based on user session results in a division of almost 50-50. The original question about the distinction between these two groups is still important because the SPIRES users will need to switch to INSPIRE soon. There is an effort to try and get the remaining SPIRES to switch over to INSPIRE before SPIRES is shutdown so that these users can make the change without feeling forced, a situation that can create unnecessary work and pain for both users and developers alike.

METHODS

My part in the INSPIRE project is to build statistical tools that can assist developers in deciding how to encourage SPIRES users to make the switch to INSPIRE now, as well as determining functionality that INSPIRE lacks which can then be added. By looking at the usage patterns of both SPIRES and INSPIRE the developers can see what features users most want to retain from SPIRES and what features they use the most in INSPIRE. From that information the developers can guide the SPIRES users to those features in INSPIRE. It is also important to find what barriers prevent the switch to INSPIRE. There are a variety of possible reasons, such as a lack of knowledge about INSPIRE or possible missing features in INSPIRE. We can learn what the barriers might be through analysis of the usage patterns.

In order to learn about the usage patterns of SPIRES and INSPIRE, there are two ways to learn about the users. One way is to have the users continually submit feedback. This system is already in place and allows for the developers to communicate with their users. The second is to look through the log files and examine the search queries and site hits that the users have generated through using the website. This alternative method had only been implemented at a basic level, and it was the aim of this project to improve the depth and functionality of this technique.

Some of the data that exists in these log files that can be analyzed are session lengths, common search terms, search results, and geographical or institutional location of users. This information is useful to the developers in the transitional period between SPIRES and INSPIRE, and will continue to be important as they try to keep improving user experience on the website. It can directly help identify the barriers mentioned above; common search terms can show the differences between search patterns in SPIRES and INSPIRE, location can tell us if some regions are possibly unaware of the impending switch, and search results can let us know if people are finding what they are looking for, and the session information can tell us if people are either getting frustrated with INSPIRE or if they are pleased with the performance of the site. The data extracted from the logs can also be used to build more complex data such as reconstructing complete user sessions. The user session information can be used to track things such as user frustration, when a user is struggling to use a part of the website effectively. The administrative staff of INSPIRE could then see where this occurs and send a message to the user to assist them. Analysis of the logfiles may reveal additional uses of the logfile data.

There is a large amount of data associated with the log files resulting from over 15,000 searches performed by users every day. Since the log files contain every HTTP request to the website, there is a lot more data than just searches in the logs, and over time the data that needs to be processed can become fairly large. It would be extremely inefficient to process the data from scratch every time a different analysis needs to be done, so the data is cached at various stages to improve performance. This caching allows months worth of data to be processed in minutes.

The log file analysis was performed using Python, a free scripting language. The analysis contained in this paper was all done with a set of log files from May 1, 2011 through July 31, 2011; however, the analysis tools were designed with the intention of long term usage that will be able to repeat the analysis later. The most useful methods can potentially be added to some of the existing basic log file statistics that are already in place. This will make the analysis techniques more easily maintainable.

While analyzing some of the data early on it became apparent that there were a number of IP addresses from China that had an enormous number of searches. Upon further investigation, the searches they were performing appeared very scripted. These were considered to be robots and were not considered in further analysis.

RESULTS

Search Keyword Occurrence

The search keyword occurrence was extracted by examining each search query for any keywords or their aliases. For instance, 'au', and 'author' both map to the search keyword 'author'. The top ten search keywords for both INSPIRE and SPIRES are shown below in Table 1.

INSPIRE (June 2011)		SPIRES (June 2011)	
Search Keyword	Occurrences	Search Keyword	Occurrences
'author'	259243	'author'	278062
'field'	58225	'date'	251327
'title'	39090	'date-added'	237226
'date'	25316	'reportnumber'	89219
'reportnumber'	13824	'irn'	67713
'journal'	13512	'exactauthor'	65081
'exactauthor'	9801	'reference'	64565
'keyword'	4367	'title'	38420
'collaboration'	4300	'journal	25183
'reference'	4165	'keyword'	19672

Table 1: SPIRES Search Keyword Occurrences in INSPIRE and SPIRES

User Sessions

The user session data contains the searches that a user performed during a visit to either INSPIRE or SPIRES. The session data is constructed by looking at searches performed by a given IP (which is assumed to identify a unique user) and uses the time in between successive searches to decide whether or not the search reflects the start of a new session or a continuation of the current session. Both the time of the search and the search query are recorded for each search in a session. Figure 1 shows a histogram of the session searches for both INSPIRE and SPIRES. The data from both follows the same shape when the data is scaled to match the second points of both logs, with the exception of the first data point (which represents the single search session counts). For this reason, the multi-search session counts were considered distinctly from total session counts. A multi-search session represents a countable use of the website, so the multi-search session counts is taken as a superior metric to the total search counts.

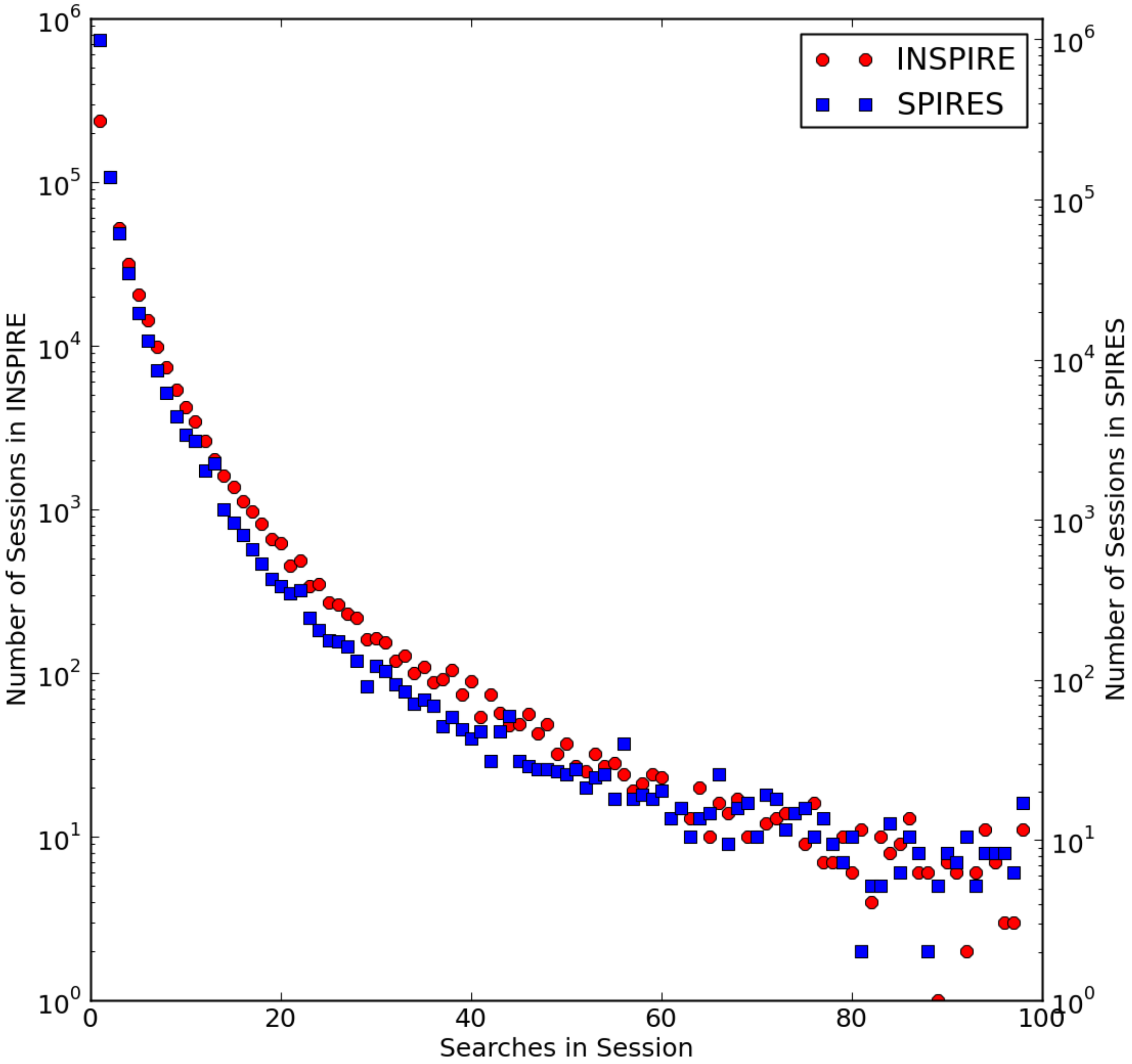


Figure 1. Histograms of the number of searches in a session.

Usage Statistics

Based on the session data shown in Figure 1, the two best measures of usage for INSPIRE and SPIRES are the number of unique users and the number of multiple search sessions. Figure 2 shows a plot of usage metrics over this time span.

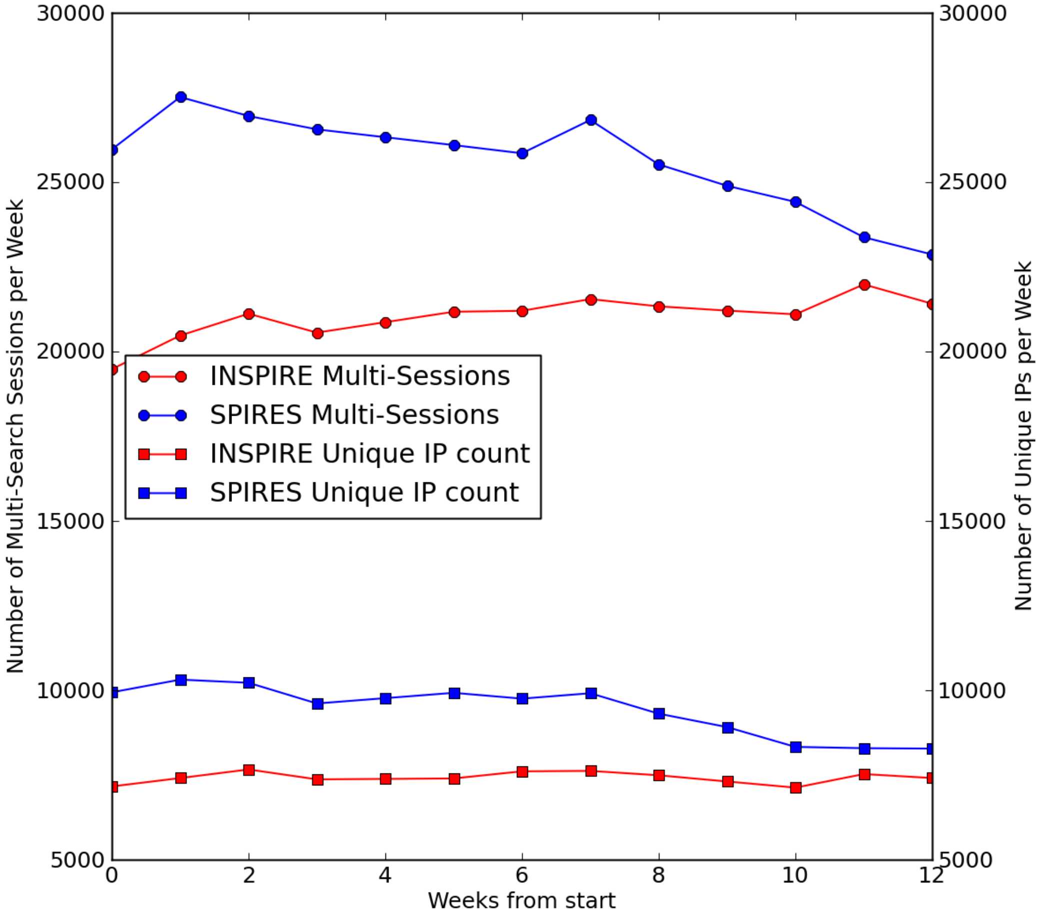


Figure 2. Plot of usage metrics (Multiple Session and Unique IP counts) over time

The usage can be broken down by country to show where the use of INSPIRE and SPIRES is most prominent as well as to give a good idea about where INSPIRE has been adopted in place of SPIRES. Figure 3 shows a bar plot of of the multiple search sessions in INSPIRE and SPIRES as well as the users who used both INSPIRE and SPIRES, broken down by their session count in each website. The segment labeled “Others” refers to any country that made up less than 5% of the whole in all of the categories.

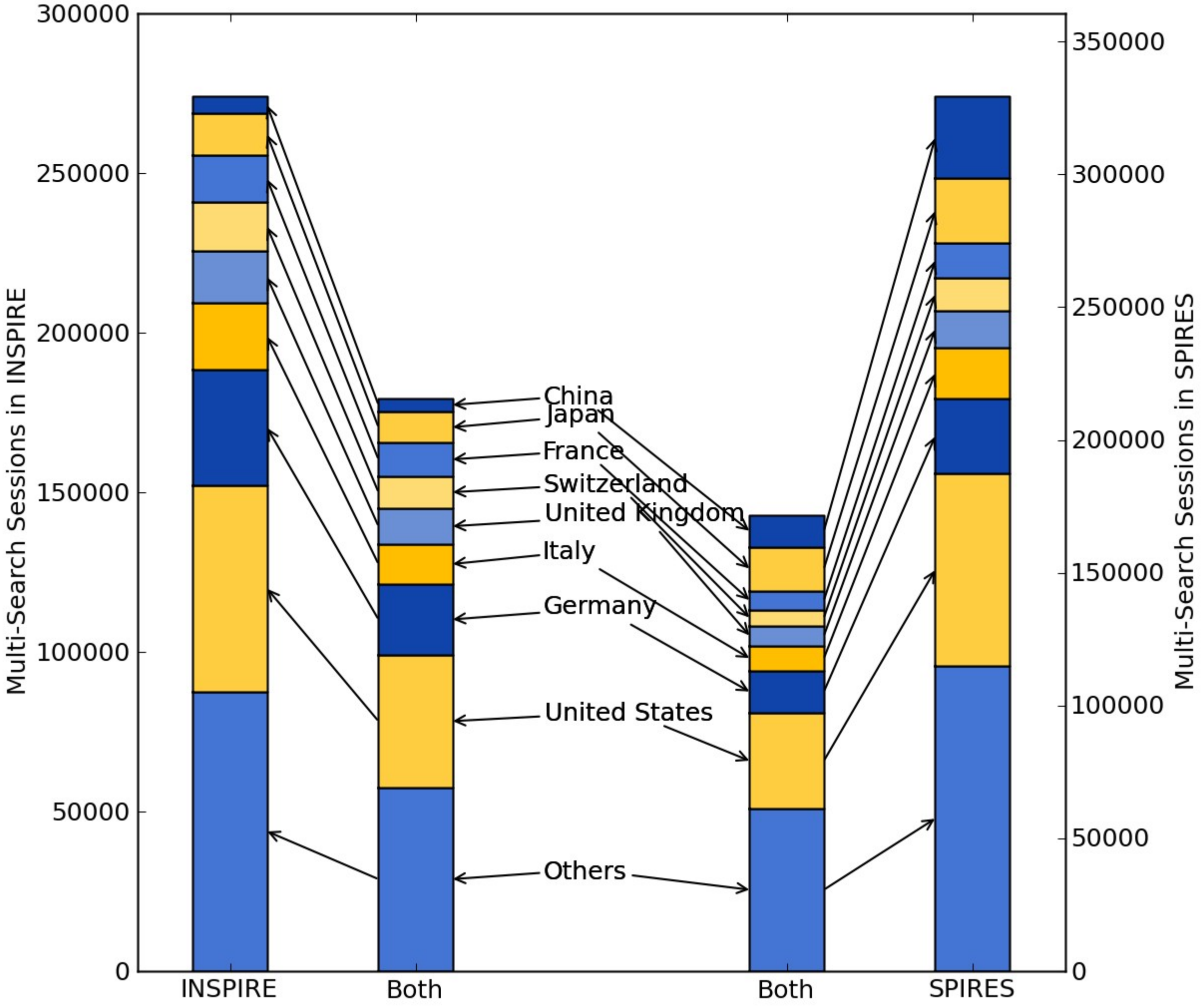


Figure 3. Bar plot of usage broken down into countries.

User Frustration Analysis

During a user's session it is possible that the user can experience a “frustration event”, where the user cannot seem to find the correct way to execute a given search query. Both SPIRES and INSPIRE use fielded metadata searches, which means that they have syntaxes that allow the users to specify the type of information they are searching on. This is useful in physics for searches such as “Higgs”, which is both an author and a frequent word in a title. When the search syntax doesn't work as expected it can lead to user frustration. This shows up in the logs as several consecutive searches by a user that are similar in structure and content. These frustrated sessions are recorded and divided into two categories based on the severity of the frustration events present in the session. Users in the extremely frustrated category have twice as many repeated similar searches as those in the frustrated category. Table 2 shows the frustration event counts for INSPIRE and SPIRES as well as the total number of sessions that are long enough to potentially contain a frustration event. The frustration events are usually centered on a search keyword, so the breakdown of the most common search keywords that appear in frustration events is shown in Table 3.

	INSPIRE		SPIRES	
	Frustrated Sessions	Total Sessions	Frustrated Sessions	Total Sessions
Frustrated	289	14613	197	13830
Extremely Frustrated	255	5967	220	5656

Table 2. “Frustration Event” counts for INSPIRE and SPIRES

INSPIRE		SPIRES	
'author'	37%	'reportnumber'	29%
'journal'	22%	'author'	26%
'field'	11%	'exactauthor'	15%
'title'	7%	'title'	9%
'reportnumber'	5%	'journal'	9%

Table 3. Common search keywords in “frustration events”

DISCUSSION AND CONCLUSION

The search keyword data shown in Table 1 can give a good estimate as to what INSPIRE is being used for. The high number of author searches demonstrates that the preferred search method is to search by author. It also gives an idea of what search keywords are most used in

SPIRES which can help to direct focus towards ensuring that the search engine in INSPIRE handles them properly. In this case, focus should be directed towards ensuring that the author search works properly.

The session data is highly interesting as a usage metric, particularly when the large variation in the SPIRES single search session is discounted(Figure 1). The two most general explanations are external links and automated systems. In the case of external links, a user is simply clicking a link that returns a SPIRES search page, and no actual search is performed user. In the logs, this would show up as a session with one search, and is not particularly interesting to look at. In the case of automated systems, there is perhaps some script that runs a search once every hour to refresh a list of articles that fits some search query. Both of these would artificially increase the number of single search sessions and would be sufficient reason to discount the point. The reason the same variation is not present in INSPIRE is perhaps due to how new it is relative to SPIRES. In future work it might be possible to identify robots and remove their searches entirely and correct the anomalous single search session count.

The usage data gives some useful results in terms of comparisons between INSPIRE and SPIRES. The usage time line in Figure 2 indicates that SPIRES usage is decreasing, while INSPIRE usage is increasing only slightly. There could be other trends that would occur in a system with constant usage that might mask an increasing trend in INSPIRE which would correlate to the decreasing trend of SPIRES. A seasonal effect that takes place from May 1 to July 31, such as summer vacations, could explain this.

The bar chart in Figure 3 helps break down the usage by the origin of the user. The plot shows that the United States makes up the plurality of both INSPIRE and SPIRES usage, but also that both services are clearly used internationally. The relative sizes of countries in

INSPIRE and SPIRES can identify where people have begun transitioning to INSPIRE from SPIRES. For instance, most of the European countries have a higher usage fraction in INSPIRE, which means means they have generally switched from SPIRES more than the United States, which has more of a half and half distribution. China has a large SPIRES usage metric, which could be due to the Chinese robots mentioned earlier that perhaps were not detected. In the future more of these can be identified and possibly removed. The “Both” columns can give an estimate of proportion is still in the transitional phase, as well as where the usage of those in this transitional phase is concentrated. This does not tell us about the users in the transitional stage, but it does ask questions about how the transitional process works. Do users gradually change, preferring INSPIRE for some tasks and SPIRES for others? Do they have a comparison phase, where they do similar searches on both to see the differences? These kind of questions could potentially be answered by examining in detail the sessions of the users that fall into the “Both” category.

The frustration event data in Table 2 indicates that INSPIRE has a higher occurrence of frustration events than SPIRES. This is because INSPIRE, as a newer service, is still being learned by its users. The data in Table 3 is more interesting, because it shows the kind of problems that cause the frustration events. The high occurrence of frustration events caused by author search keywords is due to misspellings in author names, something that is common in both SPIRES and INSPIRE. It can also show situations where a particular author is expected to appear in the search results but does not. The high occurrence of the journal search keyword in INSPIRE as opposed to SPIRES is due to a flaw in the code of search engine parsing that was present over the time period the data was taken from. This shows the usefulness of the frustration detection analysis in highlighting important issues for the developers. The frustration detection has already identified an important feature that is broken in INSPIRE, and will continue to prove a valuable tool. In the near future, the algorithm for frustration detection will be made a part of the INSPIRE web statistics module to continually watch for frustration events.

ACKNOWLEDGEMENTS

This work was supported by the Science Undergraduate Laboratory Internship program at SLAC National Accelerator Laboratory and the Department of Energy. I would like to thank my mentor, Travis Brooks, for his leadership in this project. I would also like to thank Joseph Blaylock, Valkyrie Savage, and Mike Sullivan for their assistance.

SLAC-TN-12-018

A Comparison of Image Quality Evaluation Techniques for Transmission X-Ray Microscopy

Peter J. Bolgert

Office of Science, Science Undergraduate Laboratory Internship Program

Marquette University

SLAC National Accelerator Laboratory
Menlo Park, California

August 17, 2011

Prepared in partial fulfillment of the requirement of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Yijin Liu at Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory.

Participant: ___________________________________
Signature

Research Advisor: ___________________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

Page

ABSTRACT

A Comparison of Image Quality Evaluation Techniques for Transmission X-Ray Microscopy. PETER J. BOLGERT (Marquette University, Milwaukee, WI, 53233), YIJIN LIU (Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025).

Beamline 6-2c at Stanford Synchrotron Radiation Lightsource (SSRL) is capable of Transmission X-ray Microscopy (TXM) at 30 nm resolution. Raw images from the microscope must undergo extensive image processing before publication. Since typical data sets normally contain thousands of images, it is necessary to automate the image processing workflow as much as possible, particularly for the aligning and averaging of similar images. Currently we align images using the "phase correlation" algorithm, which calculates the relative offset of two images by multiplying them in the frequency domain. For images containing high frequency noise, this algorithm will align noise with noise, resulting in a blurry average. To remedy this we multiply the images by a Gaussian function in the frequency domain, so that the algorithm ignores the high frequency noise while properly aligning the features of interest (FOI). The shape of the Gaussian is manually tuned by the user until the resulting average image is sharpest. To automatically optimize this process, it is necessary for the computer to evaluate the quality of the average image by quantifying its sharpness. In our research we explored two image sharpness metrics, the variance method and the frequency threshold method. The variance method uses the variance of the image as an indicator of sharpness while the frequency threshold method sums up the power in a specific frequency band. These metrics were tested on a variety of test images, containing both real and artificial noise. To apply these sharpness metrics, we designed and built a MATLAB graphical user interface (GUI) called "Blur Master." We found that it is possible for blurry images to have a large variance if they contain high amounts of noise. On the other hand, we found the frequency method to be quite reliable, although it is necessary to manually choose suitable limits for the frequency band. Further research must be performed to design an algorithm which automatically selects these parameters.

INTRODUCTION

Beamline 6-2c at the Stanford Synchrotron Radiation Lightsource (SSRL) is capable of Transmission X-ray Microscopy (TXM) at 30 nm spatial resolution [1]. A transmission X-ray microscope is conceptually identical to a familiar optical microscope, consisting of a light source, a condenser to focus light on the sample, a sample stage, and an objective to focus the transmitted light onto a digital sensor (Figure 1). In this context though, the word “light” refers to synchrotron X-ray radiation, ranging from about 5 to 14 keV per photon. Since X-ray wavelengths are shorter than those of visible light, the X-ray microscope is capable of much higher resolution.

Users from around the world have come to Beamline 6-2c to image a variety of biological and inorganic samples. While it depends on the specific TXM technique being used, the user will typically take thousands of images of a single sample. For example, the user might take images at 180 different angles, taking 20 identical pictures per angle, resulting in 3600 images. If the sample is larger than the field of view (FOV) of the microscope, it becomes necessary to construct a mosaic of sub-images at each angle. If it takes five exposures to cover the sample, the total increases to $5 * 3600 = 18{,}000$ images. Additionally, the user will periodically move the sample off to the side and take some reference images (perhaps several hundred). All of these images must be extensively processed before presentation-quality images and 3D models can be obtained. While the details of the image processing depend on the specific TXM technique used, a typical processing workflow is shown in Figure 2.

With such massive amounts of data, it is necessary to automate the image processing workflow as much as possible, especially for the alignment of similar images with each other. Image alignment is a recurring part of our workflow, as it appears in steps two (Aligning and

Averaging of Multiple Images), three (Mosaic Stitching), and five (Tomographic Reconstruction). As mentioned, the user will typically take 20 to 30 repeated images of the sample at each angle. In between every exposure, the sample stage will jitter slightly, causing these repeated exposures to be identical except for a small translation. Step two of our workflow consists of aligning these repeated images and then taking the average, which increases the signal-to-noise ratio (SNR), a step known as the "advanced average." Not only is it tedious to align images by eye, but a computer should be able to align images much more accurately than a human can. By automating tasks like this, users can focus more on their experiment and less on repetitive image processing.

One commonly used algorithm for aligning two images is known as phase correlation, which manipulates the two images in the frequency domain. First, we calculate the phase correlation function (*PC)* for images *a* and *b*, given by

$$PC = \frac{\mathcal{F}(a)\,\overline{\mathcal{F}(b)}}{|\mathcal{F}(a)||\mathcal{F}(b)|}, \tag{1}$$

where $\mathcal{F}(a)$ is the Fourier transform of image *a* and the overbar denotes complex conjugation [2]. For two images which are identical except for a pure translational shift, the phase correlation function will be given by

$$PC = e^{-2\pi i(\boldsymbol{u}\cdot\boldsymbol{f})}, \tag{2}$$

where $\boldsymbol{u}$ is a vector whose components are the horizontal and vertical shifts (in pixels) from image *a* to *b*, and ***f*** is the position vector in the *PC* matrix. Figure 3 shows the real part of the *PC* function for two images with $\boldsymbol{u} = (10\ pixels\ to\ the\ right, 10\ pixels\ down)$. This cosine wave points in the direction of $\boldsymbol{u}$ and the spacing is inversely proportional to the magnitude of $\boldsymbol{u}$. Since the Fourier transform of a complex exponential is a Dirac delta function [3], we find that $\mathcal{F}^{-1}(PC)$ contains a single spike whose coordinates correspond to the offset, $\boldsymbol{u}$. In other words,

$$offset = coordinates\ of\ \max\left(\mathcal{F}^{-1}(PC)\right). \tag{3}$$

By calculating the offset between the first image and every other image in the stack, we can accurately align all the images to the first image, resulting in a sharp average.

For images containing noise, the situation described above is not completely accurate. In particular, for two images containing high frequency noise, sometimes the algorithm will attempt to align noise with noise, yielding an unsatisfactory result. For example, consider a stack of 30 images of the same sample, which differ by a small translation. However, each image contains a sharp scratch in the exact same location (see Figure 4 for an example). In the case of this data set (which we call the "scratched data set"), there was a scratch on the digital sensor of the microscope. In the FFT, the scratch corresponds to high frequency power, and the phase correlation will align the scratches together, resulting in an offset of $\boldsymbol{u}$ = (0, 0). Since the alignment failed, the resulting average will be blurry (Fig. 4), tainting the rest of the workflow.

To remedy this problem, it is helpful to slightly blur the images during the phase correlation process. We start by multiplying the *PC* function by a two-dimensional Gaussian function of the form

$$f(x,y) = e^{-\frac{g}{2}(x^2+y^2)}, \tag{4}$$

where *g* is a blurring parameter. The greater *g* is, the narrower the Gaussian peak. Multiplying *PC* by a Gaussian masks the outer region of the *PC* function, which corresponds to high frequency noise in the original images. Once the parameter *g* passes a critical value, the maximum of $\mathcal{F}^{-1}(PC)$ shifts to the desired offset. Figure 5 shows the evolution of the *PC* function for two scratched images as *g* is increased. As *g* increases, the center spot begins to look more and more like the ideal sinusoid of Figure 3. Figure 5 also shows the evolution of $\mathcal{F}^{-1}(PC)$ with increasing *g*. You can see that there are two blobs competing for attention. The off-center blob corresponds to the features of interest (FOI). As g increases, both blobs fade, but the off-center blob fades more slowly. By g = 8, the maximum of $\mathcal{F}^{-1}(PC)$ occurs in the off-center blob, resulting in a correct alignment. Finally, Figure 5 also shows the evolution of the resulting averages when phase correlation is performed for the entire stack of the twenty "scratched" images. As you can see, when g crosses a critical value, the resulting average is as sharp as any of the starting images, but with much less noise (high SNR).

The process of blurring during phase correlation works well, but it requires manual user input for choosing g, slowing down the workflow. In order to make this process fully automatic we need an algorithm to evaluate image quality, particularly sharpness. If the averaged image is not sharp enough, the algorithm should adjust the blurring so that the alignment is optimized. All of this should occur without any input from the user. There are many examples of sharpness algorithms in the literature, many of which are surveyed in Ferzli and Karam [4]. In this paper, we will explore a variance-based algorithm and an FFT-based algorithm in an attempt to further automate our image processing workflow.

MATERIALS AND METHODS

All of the work for this project was performed using MATLAB R2010b, equipped with the Image Processing Toolbox. Every step in the TXM image processing workflow (Fig. 2) can be performed with TXM Wizard, a suite of Matlab software developed by Yijin Liu, Florian Meirer, and Phillip Williams. TXM Wizard can be downloaded for free at http://sourceforge.net/projects/txm-wizard. The program used to align and average images (step two in Fig. 2) is known as Advanced Average.

In order to test different sharpness algorithms, I used two test data sets. The first data set is the "scratched data set" described above. It consists of thirty 1024 x 1024 pixel TIFF images taken of an industrial catalyst. The second data set, consisting of fifty 1024 x 1024 pixel TIFF images, is known as the "root data set" since the images are of tiny plant roots (Figure 6). This data set has been doctored to contain both high and low frequency noise. The high frequency noise consists of small artificial scratches, which have the same position for each image. The low frequency noise consists of broad bright and dark patches which shift randomly from image to image. Figure 6 also shows the evolution of alignment for this data set. In the computer, these images are stored as 1024 x 1024 matrices with integer values ranging from 0 (black) to 255 (white).

Next I will describe the two sharpness algorithms which were used. The first is called the variance method. The variance of a data set (such as matrix of integers) is the square of the standard deviation and is equal to

$$var = \sum_i (x_i - \langle x \rangle)^2, \tag{5}$$

where x_i represents a data point and $\langle x \rangle$ is the mean of the data set. Now, it seems intuitively obvious that a sharp "advanced average" should have a larger variance than a misaligned, blurry "advanced average." In a sharp image, there will be very light pixels and very dark pixels, both deviating far from the mean. In a blurry image, the features are averaged out. The light pixels are not as light, and the dark are not as dark, resulting in a lower variance. This is demonstrated in Figure 7 for a random 20x20 matrix.

The second sharpness algorithm used is related to the Fast Fourier Transform (FFT) and is called the frequency threshold method. A 2D FFT is a pictorial depiction of the frequency distribution in a signal (Figure 8). Low frequencies appear in the center, and high frequencies away from the center. In the case of a digital image, we are concerned with the spatial frequency, that is, how quickly pixels values change as we move across the image. The highest frequency occurs when a black pixel (value 0) is adjacent to a white pixel (value 255), and this would show up at the fringes of the FFT. A quality image will contain sharp edges, with pixel values changing quickly from light to dark. This corresponds to high frequency power in the FFT. Thus we expect sharp images to contain more high frequency power than a blurry image.

To quantify the sharpness, we choose a suitable radius and then sum up all the absolute values of the FFT matrix which lie outside this value (technically, the term "power" refers to the absolute square of the FFT values, but we will use "power" to refer to the absolute value in this paper). However, we do not want to sum all the way to the outer boundary. While the features of interest in an image are sharp, it will generally take at least a few pixels to transition from light to dark. Any FFT power more than a certain distance from the origin corresponds to undesirable high frequency noise. Thus to quantify the sharpness of an image, it suffices to sum up the FFT power in between suitably sized concentric circles (Figure 8). The radii of the threshold circles (R_{in} and R_{out}) must be experimentally determined.

One way to determine suitable values of R_{in} and R_{out} is to simply compare FFT's of blurry and sharp images. Figure 8 shows FFT's of two "advanced averages" for the root set. In the blurry average (g = 0), you can see that while most power is concentrated in the center, there is power scattered all the way out to the edges, corresponding to the scratches. The sharp average (g = 5) has more power in the mid-range region of the FFT. Choosing $R_{in}= w/32$ and $R_{out}=w/16$ would be an appropriate choice for this data set, where w is the width of the cropped "advanced average."

Another way to choose R_{in} and R_{out} is to plot the power in the FFT's as a function of the distance from the center. Figure 9 shows such a plot for the root data "advanced averages" in Figure 8. For each pixel in the image, its distance from the origin was calculated and rounded to the nearest integer (say $r = r_0$). By adding together all elements at this rounded distance, we obtain $|FFT(r_0)|$. From Fig. 9 we can see that most of the power corresponding to the features of interest lies between $R_{in}= w/30$ and $R_{out} = w/10$, generally consistent with the values mentioned previously. In reality, choosing radii is somewhat arbitrary and it pays to look at the problem in more than one way.

In order to carry out my research I designed and built a Matlab graphical user interface (GUI) called Blur Master. A screenshot is shown in Figure 10. Blur Master is used to create "advanced averages" for a customizable range of *g* (blurring) values. After writing all the "advanced averages" to file, Blur Master automatically quantifies the sharpness of the averages using both the variance and frequency threshold methods. All images are cropped by 20% on each side before evaluating the sharpness. I also designed a sub-GUI, called Power Plotter to plot FFT power as a function of the distance from the center, which can be used to choose appropriate values for R_{in} and R_{out}. A screenshot of Power Plotter is shown in Figure 11.

RESULTS OF SHARPNESS TESTING

Table 1: Scratched Data Set
(R_{in}= $w/32$ and R_{out} = $w/16$ for freq. method)

g (blurring)	variance method	frequency method
0	270.7	1.012
1	270.7	1.015
2	270.9	1.019
3	270.9	1.021
4	261.0	1.028
5	263.3	1.190
6	282.0	1.580
7	286.5	1.667
8	313.9	1.987
9	313.9	1.987
10	313.9	1.985
11	314.0	1.984
12	314.0	1.989
13	314.0	1.989
14	314.0	1.988
15	314.0	1.988
16	314.0	1.988
17	314.1	1.990
18	314.1	1.991
19	314.1	1.990
20	314.1	1.990
30	314.1	1.991
40	314.1	1.990
50	314.0	1.986
60	314.0	1.985
70	313.9	1.983
80	313.8	1.977
90	313.8	1.975
99	313.7	1.966

Table 2: Root Data Set
(R_{in}= $w/32$ and R_{out} = $w/16$ for freq. method)

g (blurring)	variance method	frequency method
0	118.2	2.867
1	118.2	2.867
2	118.2	2.867
3	118.2	2.867
4	104.5	6.736
5	104.8	6.752
6	104.9	6.755
7	105.0	6.760
8	105.1	6.760
9	105.2	6.763
10	105.3	6.766
11	105.3	6.766
12	105.3	6.766
13	105.3	6.766
14	105.3	6.766
15	105.3	6.766
16	105.3	6.766
17	105.3	6.766
18	105.3	6.766
19	105.3	6.766
20	105.3	6.766
30	105.1	6.764
40	104.0	6.737
50	102.1	6.687
60	99.2	6.596
70	96.6	6.482
80	91.7	6.226
90	87.0	5.896
99	83.2	5.552

DISCUSSION

The tables above show the sharpness results for both the scratched and root data sets. For each value of *g*, we created an "advanced average" and quantified its sharpness. For the scratched data set, we see that both the variance and frequency threshold values start low, increase dramatically from $g = 5$ to 8, and then level off. These results are consistent with our visual sense of sharpness. The "advanced averages" start out blurry, start aligning between $g = 5$ to 8, and then look the same from that point on.

For the root data set, the results are mixed. While the frequency threshold algorithm performed well, the variance algorithm finds the *misaligned* averages to be the "sharpest"! In this case the variance contributed by the noise was greater than the variance contributed by the sharp features of interest. The variance method is not expected to perform consistently from one data set to another because it treats sharp features and high frequency noise on equal footing. Note that the same frequency parameters were the same for both data sets ($R_{in}= w/32$ and $R_{out} = w/16$). These values were found experimentally through trial and error. We conclude that these parameters should work for "typical" TXM images, although certainly more tests are needed.

Now that we have gained some insight into image quality evaluation, there are several directions for continued research. The literature is rich with various sharpness algorithms, including several which claim to be immune to noise [4], [5]. There also exist "no-reference" sharpness metrics which can align images without any prior knowledge of the images (i.e. it is not necessary to manually adjust parameters) [4], [5]. By implementing more sophisticated and robust algorithms, we can improve the ease and reliability of our GUI's.

In addition to adding more sophisticated algorithms, we would like to completely automate the advanced average process. In the ideal scenario, when images are loaded into the

Advanced Average GUI, the computer would automatically find the ideal *g* value, and then return only the sharp average. One way to achieve this would be to 1) calculate the "advanced average" at a default *g* (say *g* = 8), 2) evaluate the sharpness, 3) calculate the "advanced average" at adjacent *g*'s (*g* = 8 and *g* = 9), 4) evaluate these sharpnesses, and 5) continue outward until reaching a local maximum or a plateau. A plateau is reached when adjacent sharpnesses differ less than a specified amount (say 0.1%).

The main problem with this scheme is that is it is difficult to choose R_{in} and R_{out} ahead of time. Values that work for one data set may not work for another data set. It seems foolish to rely on "typical values" such as $R_{in} = w/32$ and $R_{out} = w/16$. We would like to choose these parameters automatically for each data set. We can potentially solve this problem by looking at the FFT vs. radius plot for one of the original images (see Figure 12). Since the features of interest are obviously sharp in the original image, this plot should display a local maximum at these frequencies. By evaluating the properties of this maxima (e.g. the full-width-at-half-max), suitable values of R_{in} and R_{out} could be chosen automatically.

While the "advanced average" blurring process is not yet automatic, the Blur Master GUI could still be useful for users in its current form. For example, if the user has multiple sets of similar data (e.g. 20 images for one angle, then 20 images for the next angle, etc.), he/she could run Blur Master on the first 20 images to find the best *g* value, and then apply the same blurring factor to all subsequent data sets.

In conclusion, two sharpness algorithms were used to evaluate the quality of alignment in TXM images. The frequency threshold method was found to be more reliable; however, it requires manual input from the user. Further research must be performed in order to automatically select these frequency parameters. In addition, “noise immune” metrics and “no-reference” metrics should be investigated.

ACKNOWLEDGMENTS

This work was supported by SLAC National Accelerator Center, the U.S. Department of Energy, and the Office of Science. I would like to give special thanks to Yijin Liu, who guided me, and to Allison Hume, who provided a stimulating working environment. Finally, I would like to thank anybody who took the time to read this research paper.

REFERENCES

[1] J.C. Andrews et al., “Nanoscale X-Ray Microscopic Imaging of Mammalian Mineralized Tissue,” *Microscopy and Microanalysis,* vol. 16, pp. 327 – 336, 2010.

[2] R. Szeliski, “Image Alignment and Stitching: A Tutorial,” *Foundations and Trends in Computer Graphics and Vision*, vol. 2, pp. 1 – 104, 2006.

[3] R.N. Bracewell, *The Fourier Transform and its Applications*, 3rd ed., New York: McGraw-Hill, 1999.

[4] R. Ferzli and L.J. Karam, “No-Reference Objective Wavelet Based Noise Immune Image Sharpness Metric,” *IEEE International Conference on Image Processing*, vol. 1, pp. 405 – 408, 2005.

[5] X. Zhu and P. Milanfar, “A No-Reference Sharpness Metric Sensitive to Blur and Noise,” *International Workshop on Quality of Multimedia Experience*, pp. 64 – 69, 2009.

FIGURES

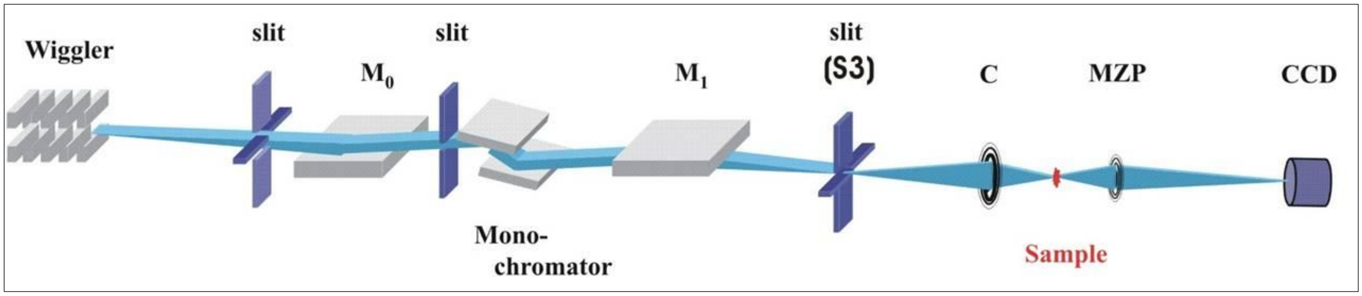


Figure 1: Schematic of the Transmission X-Ray Microscope at SSRL Beamline 6-2c. Synchrotron radiation, which contains a wide range of frequencies, is passed through the monochromator to select a single frequency. The condenser (C) focuses the X-rays on the sample, and the micro-zone plate (MZP) focuses the transmitted X-rays onto the Charge Coupled Device (CCD). The microscope also contains various slits and mirrors (M0, M1) to collimate and direct the beam.

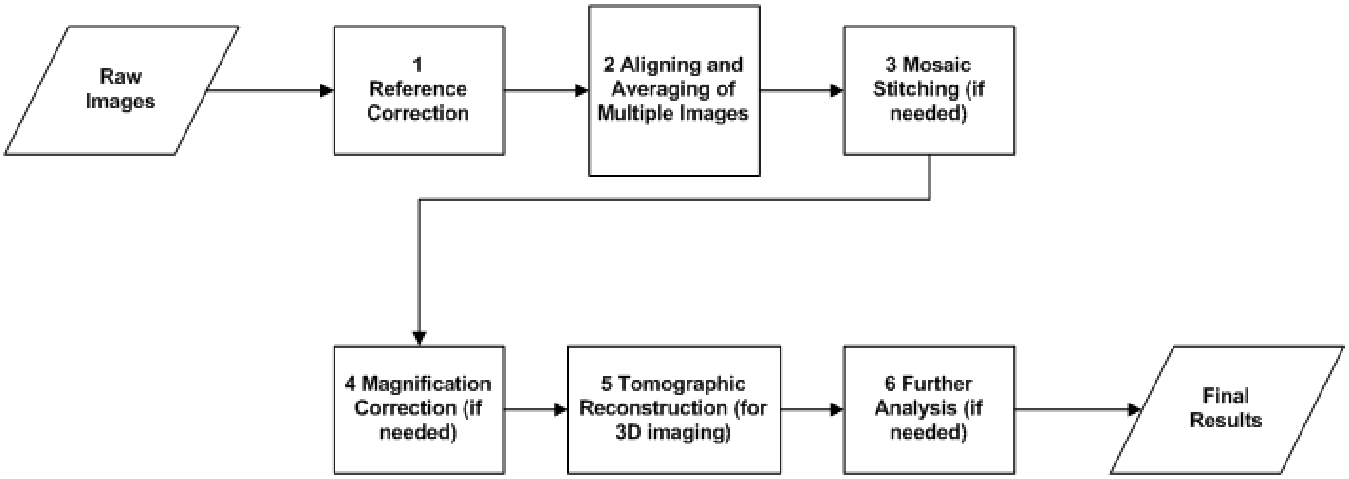

Raw Images
1 Reference Correction
2 Aligning and Averaging of Multiple Images
3 Mosaic Stitching (if needed)
4 Magnification Correction (if needed)
5 Tomographic Reconstruction (for 3D imaging)
6 Further Analysis (if needed)
Final Results

Figure 2: Image Processing Work Flow. Starting with raw images, the first step is to subtract references images, which are images taken with the sample moved outside the Field of View. Subtracting the references should eliminate the background and other types of noise (due to the digital sensor for instance). In step two, we align nearly identical images (see text for explanation) and then averaging them. Automating this process is the subject of the present paper. In step three, we stitch images together to create a larger mosaic. This is necessary when the sample is larger than the Field of View of the microscope. Step four, magnification correction, is needed when we image the sample at multiple X-ray energies. Magnification is a function of energy, and so we need to scale all images to a common size. Step five is used for tomographic techniques, in which we take images at multiple angles. The reconstruction algorithm converts these angled views to a series of cross-sectional slices, much like an MRI. Finally, further analysis may be necessary for more advanced techniques. In order to completely automate this workflow we need automatic image quality control, as discussed in the text.

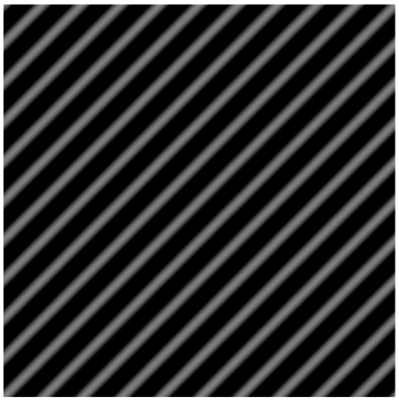


Figure 3: The real part of the phase correlation function for two images differing by a pure translational offset. The offset u is given by u = (10 pixels to the right, 10 pixels down). In the ideal situation, the real part of this function is a cosine pointing in the direction of the offset. The spacing, or wavelength, is inversely proportional to |u|.

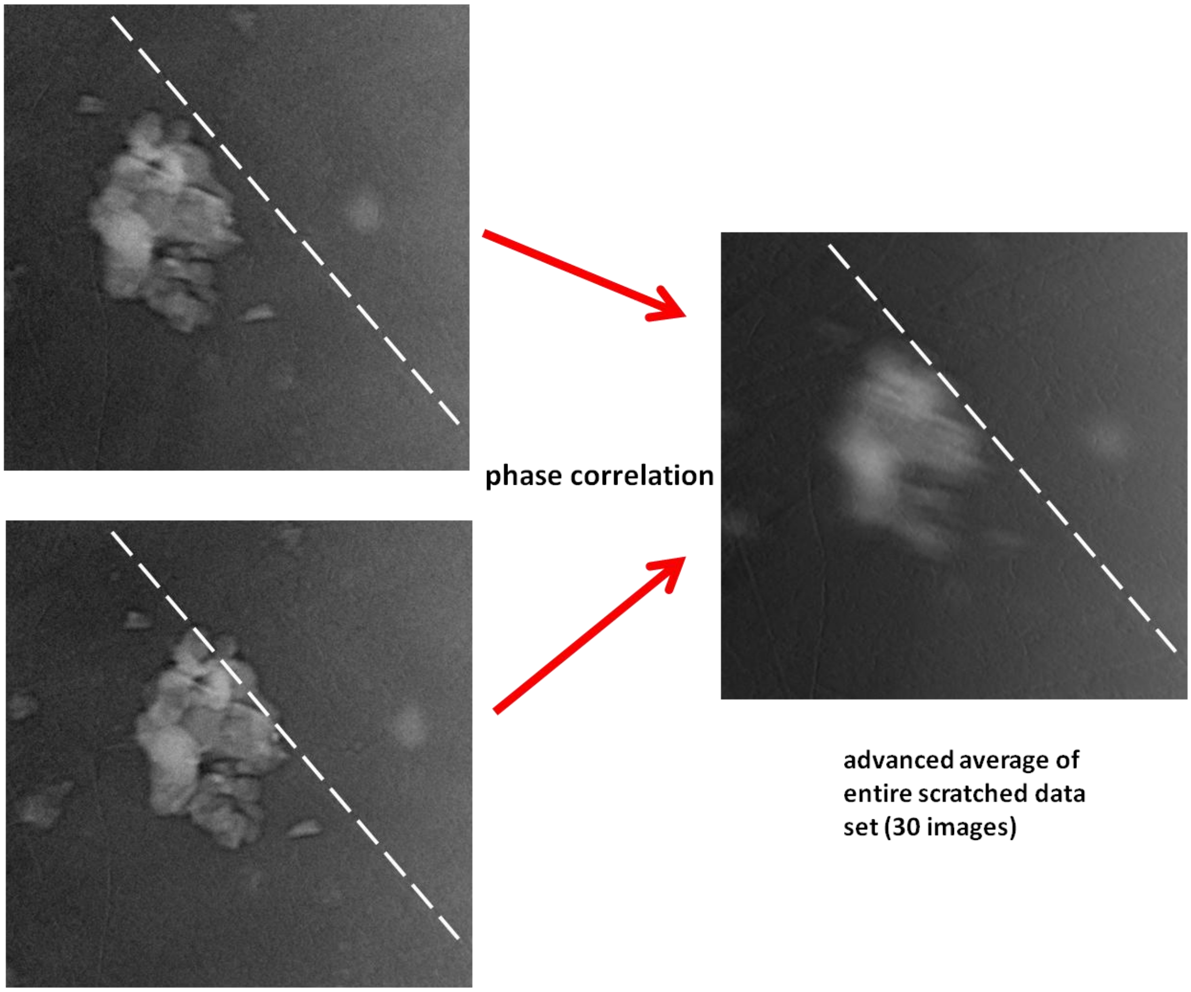


Figure 4: (LEFT) Two of the 30 images in the scratched data set. Both images have a faint scratch in the same location, while the catalyst is offset to the right in the second picture. Since the scratch is quite faint, its position is indicated by the dotted white line. (RIGHT) The "advanced average" of the entire image stack using phase correlation. The scratches all align, resulting in a sharp scratch and a blurry catalyst.

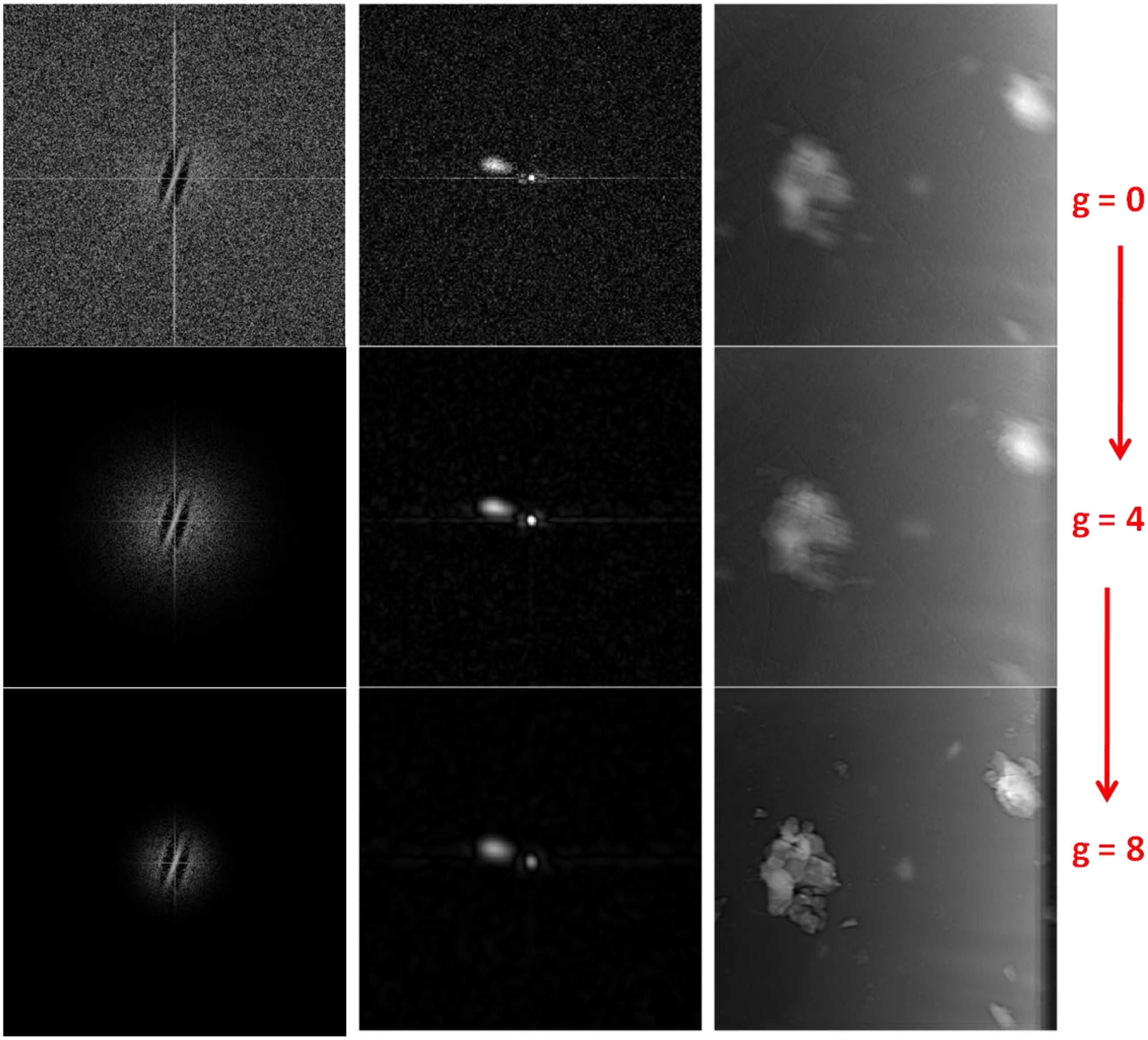


Figure 5: (LEFT) The evolution of the PC function (real part) between two scratched images as g increases. As g increases, the PC function begins to look like a cosine, reminiscent of Figure3. (MIDDLE) The evolution of the corresponding $\mathcal{F}^{-1}(PC)$ functions. By g = 8, the maximum is located in the off-center blob. (RIGHT) The evolution of the "advanced average" of the entire image stack with increasing g.

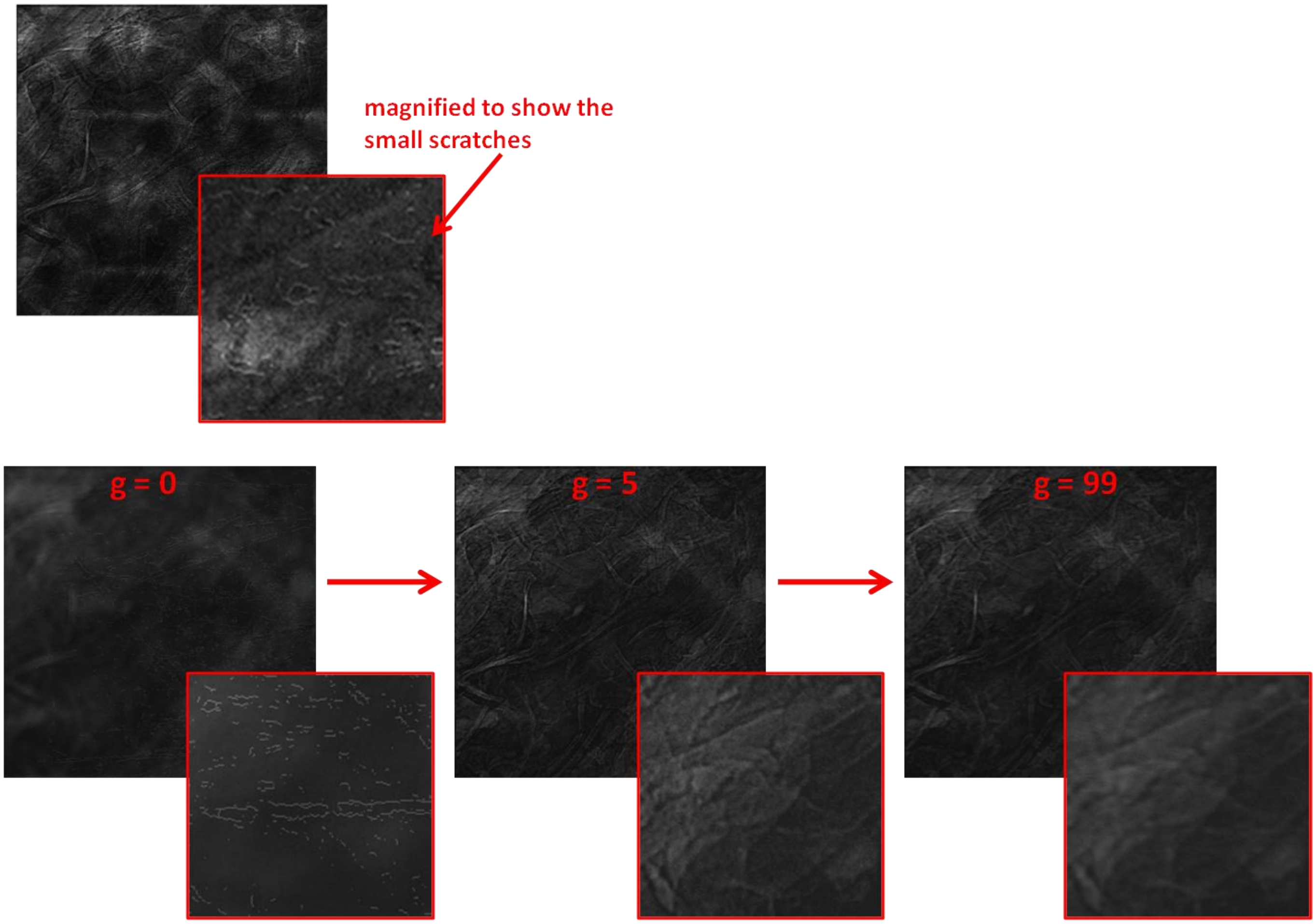


Figure 6: (TOP) One image from the root data set, with a blown up inset to show the high frequency artificial scratches. (BOTTOM) The evolution of alignment with increasing g for the root data set. For g = 0, the artificial scratches are aligned, resulting in blurry features. By g = 5, the features of interest are aligning properly. By g = 99, the "advanced average" is starting to get slightly blurred, which can be seen in the inset.

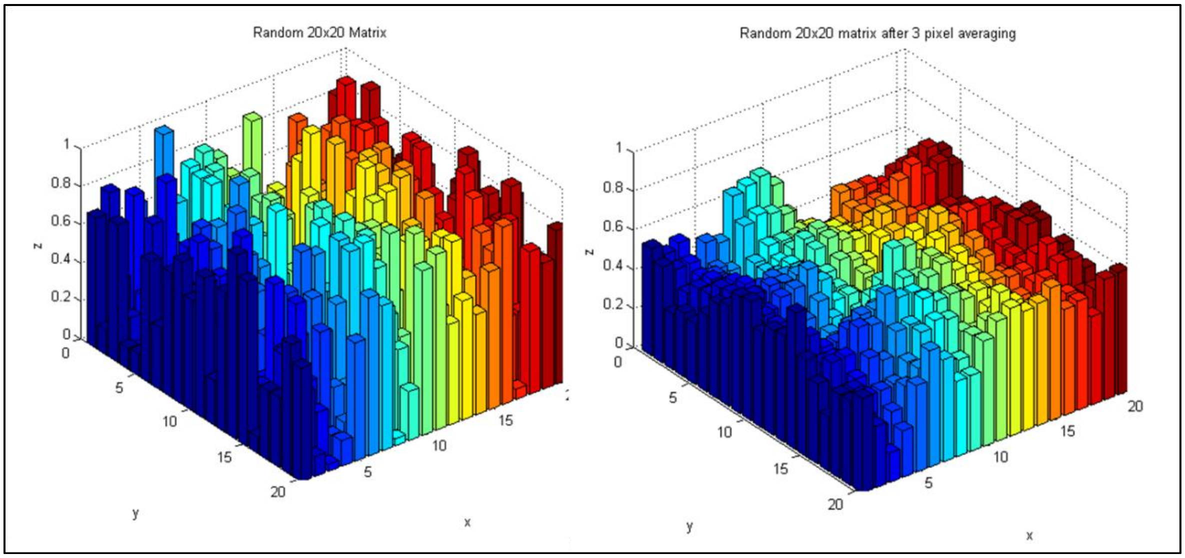


Figure 7: (LEFT) A random 20x20 matrix to represent a “sharp” image; (RIGHT) The same matrix after a 3x3 pixel averaging filter has been applied. This matrix represents a blurred image, and its variance has decreased.

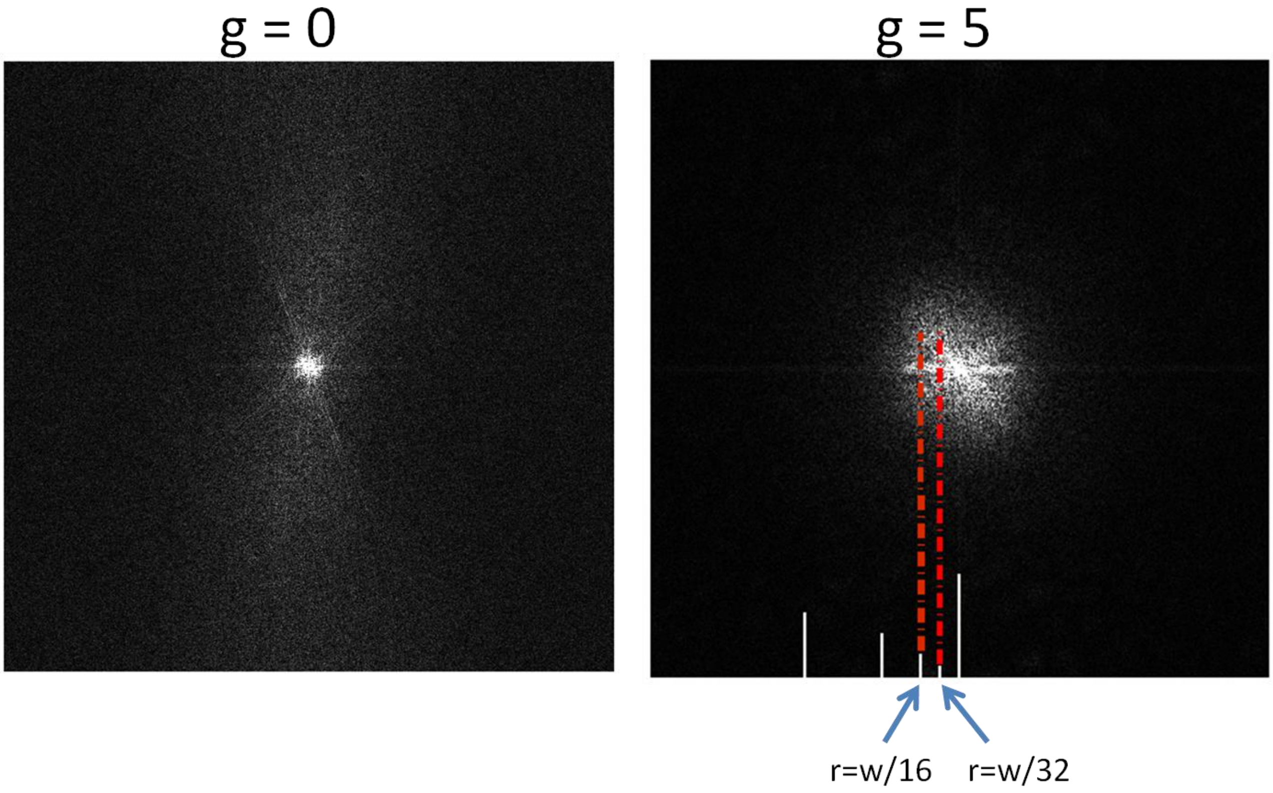


Figure 8: A comparison of the FFT’s of a blurry “advanced average” (g = 0) and a sharp average (g = 5). Most of the frequency power corresponding to the features of interest lies between r=w/32 and r=w/16, where w is the width of the cropped “advanced average.”

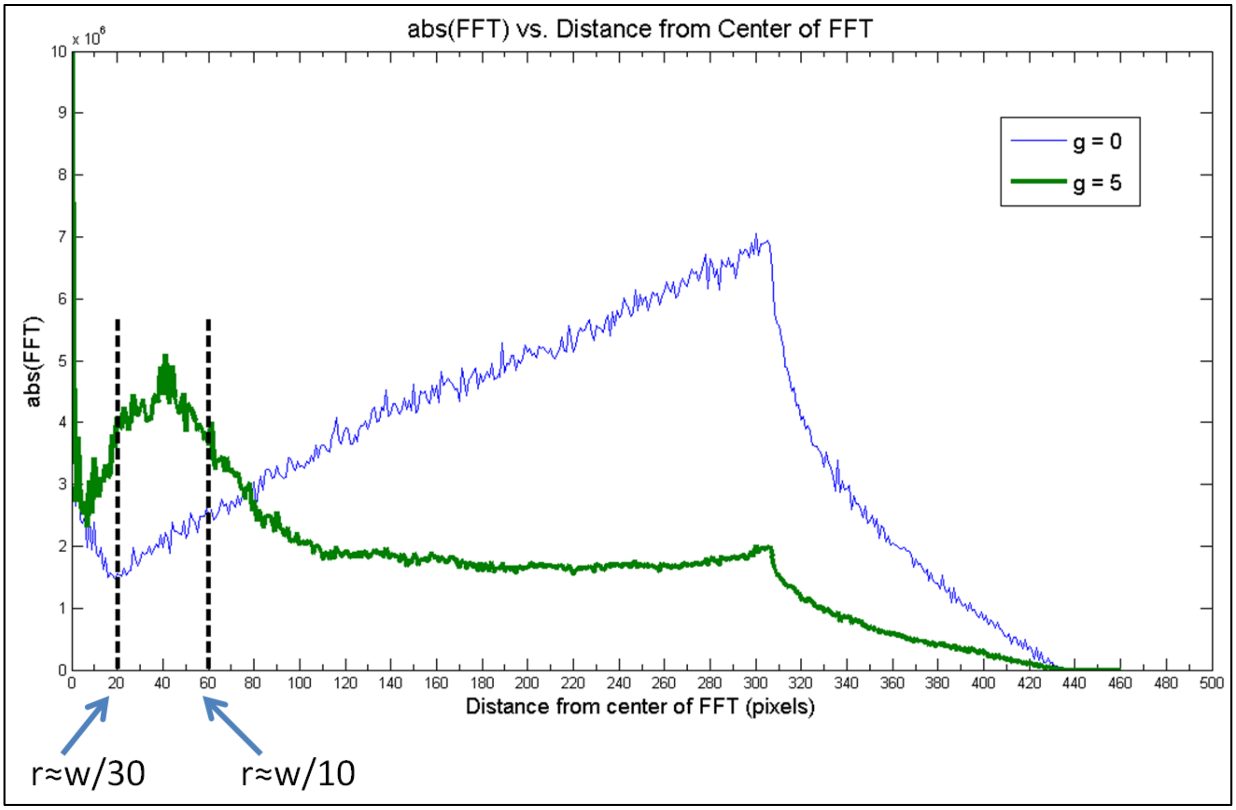
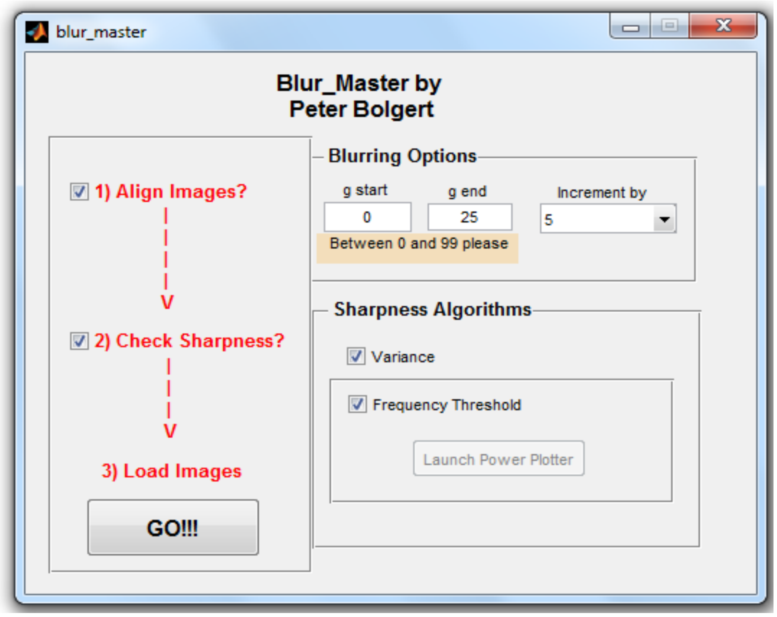


Figure 9: Plots of abs(FFT) vs. distance radius for "advanced averages," a blurry average (g = 0) and a sharp average (g = 5). These plots are for the root data set. In the sharp image (g = 5), there is a peak between r=w/30 and r=w/10, where w is the width of the cropped "advanced average." This peak corresponds to the sharp features of interest in the sharp average. The power drops off after about 300 pixels because we are starting to sum up the power in the corners of the FFT's, where the area is less.



Blur_Master by
Peter Bolgert

☑ 1) Align Images?

☑ 2) Check Sharpness?

3) Load Images

Blurring Options

g start: 0

g end: 25

Between 0 and 99 please

Increment by: 5

Sharpness Algorithms

☑ Variance

☑ Frequency Threshold

Figure 10: Screen shot of Blur Master. This GUI is used to produce "advanced averages" for a range of g (blurring) values, and also to quantify their sharpness using the two algorithms described in the text.

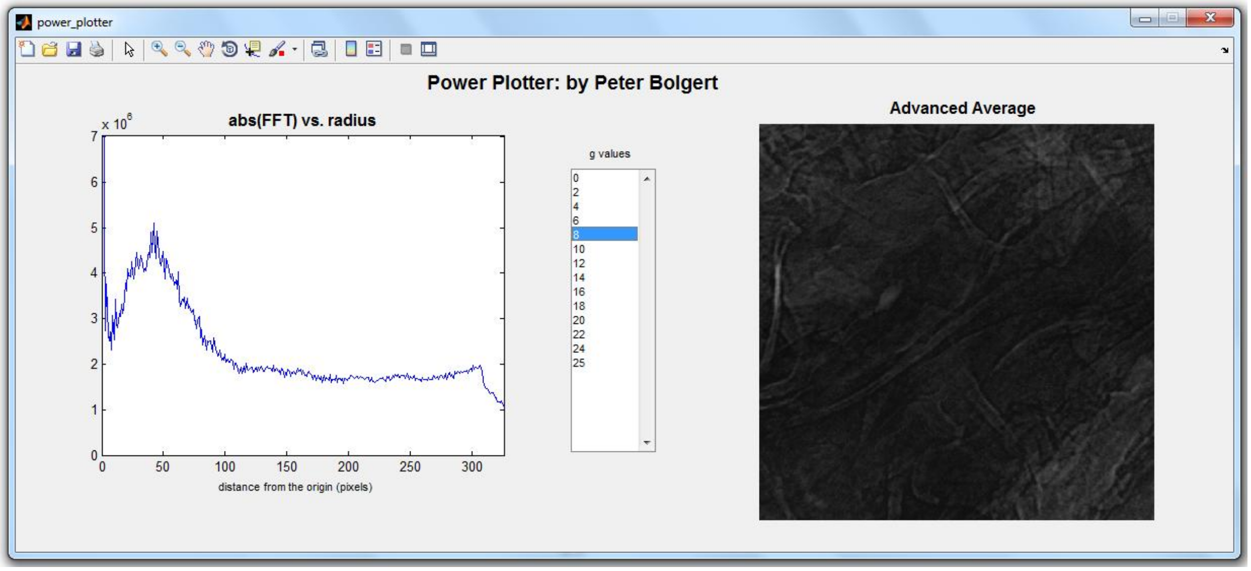


Figure 11: Screenshot of Power Plotter. This sub-GUI allows the user to simultaneously see each "advanced average" alongside a plot showing its FFT power vs. the distance from the FFT center, similar to Figure 9. This sub-GUI could be used to pick appropriate frequency bands to quantify the sharpness of subsequent data sets.

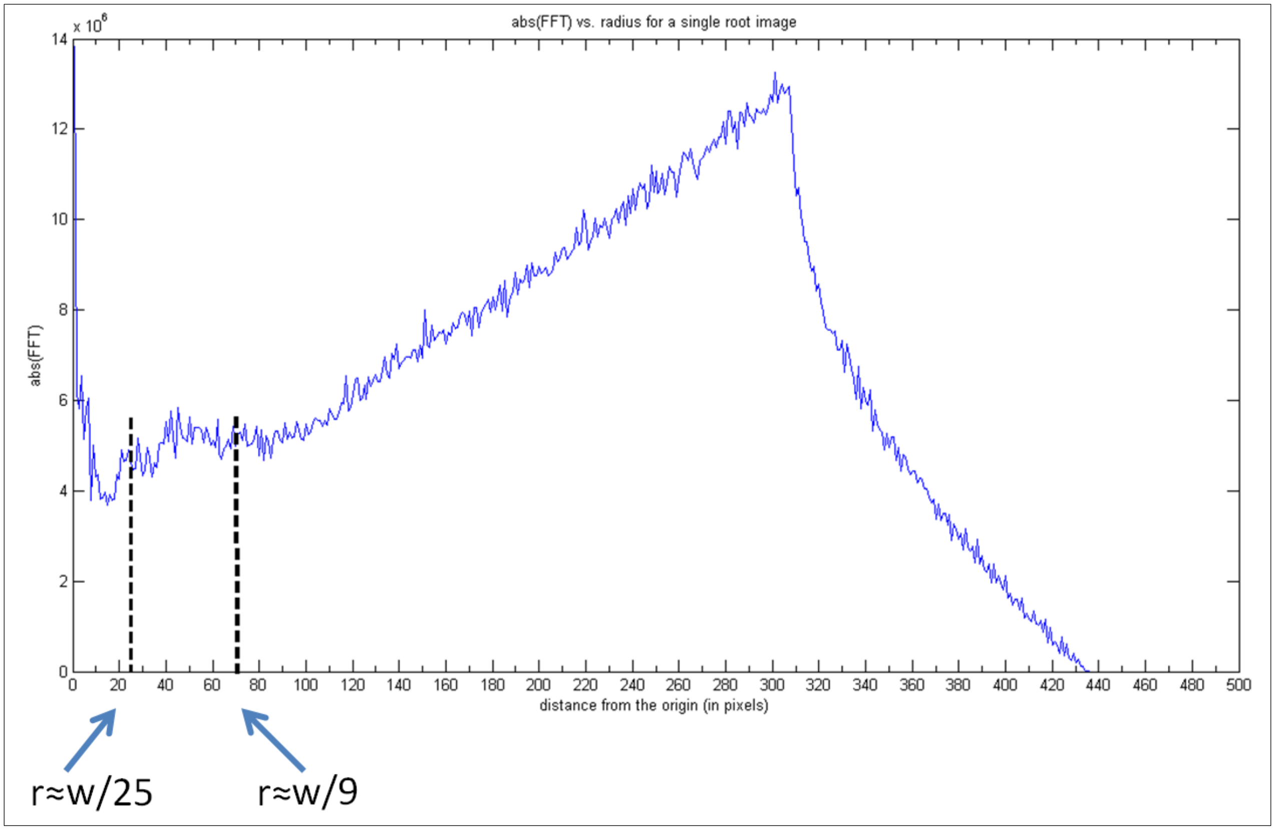


Figure 12: Plot of abs(FFT) vs. radius for a single root image. While this image contains high frequency noise, you can see that there is a small local max between r=w/25 and r=w/9, where w is the width of the cropped image. This local maximum corresponds to the sharp features of interest in the original image. We expect any sufficiently sharp "advanced average" to show a similar hump.

SLAC-TN-12-019

Investigating the Magnetorotational Instability with Dedalus, an Open-Source Hydrodynamics Code

Keaton J. Burns

Office of Science, Science Undergraduate Laboratory Internship (SULI)
University of California Berkeley
SLAC National Accelerator Laboratory
Menlo Park, CA 94025

August 17, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Jeffrey S. Oishi with the Kavli Institute for Particle Astrophysics and Cosmology.

Participant: ________________________________

Signature

Research Advisor: ________________________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Investigating the Magnetorotational Instability with Dedalus, an Open-Source Hydrodynamics Code. KEATON J. BURNS (University of California Berkeley, Berkeley, CA 94720) JEFFREY S. OISHI (Kavli Institute for Particle Astrophysics and Cosmology, Menlo Park, CA 94025)

The magnetorotational instability is a fluid instability that causes the onset of turbulence in discs with poloidal magnetic fields. It is believed to be an important mechanism in the physics of accretion discs, namely in its ability to transport angular momentum outward. A similar instability arising in systems with a helical magnetic field may be easier to produce in laboratory experiments using liquid sodium, but the applicability of this phenomenon to astrophysical discs is unclear. To explore and compare the properties of these standard and helical magnetorotational instabilities (MRI and HRMI, respectively), magnetohydrodynamic (MHD) capabilities were added to Dedalus, an open-source hydrodynamics simulator. Dedalus is a Python-based pseudospectral code that uses external libraries and parallelization with the goal of achieving speeds competitive with codes implemented in lower-level languages. This paper will outline the MHD equations as implemented in Dedalus, the steps taken to improve the performance of the code, and the status of MRI investigations using Dedalus.

INTRODUCTION

Astrophysical accretion discs are discs of material orbiting around a star or black hole. For decades, an outstanding theoretical problem in the physics of accretion discs was the identification of an efficient mechanism for transporting angular momentum outward, since viscous transport was unable to account for the observed rates of infalling material [1]. In 1991, Balbus and Hawley discovered a linear instability that occurs in discs with a poloidal magnetic field [2]. This magnetorotational instability (MRI) is generally accepted as the mechanism responsible for angular momentum transport and turbulence in accretion discs.

Laboratory experiments seeking to produce and study the MRI face a key difficulty: the toroidal field perturbations required by the MRI must be produced by induction if an entirely poloidal external magnetic field is applied. However, liquid metal flows with sufficiently high magnetic Reynolds numbers (Rm) to produce these induction effects tend to have extremely high hydrodynamic Reynolds numbers (Re) ($\frac{Rm}{Re} \approx 10^{-5}$ for these materials), and hence tend to be turbulent even without contributions from the MRI [3]. One solution to this issue is to apply a helical magnetic field in the laboratory, exciting a helical magnetorotational instability (HMRI). Simulations of the MRI and HMRI into their non-linear phases may provide information pertaining to the applicability of experimental HRMI results to the processes occurring in astrophysical discs.

METHODOLOGY: DEDALUS DEVELOPMENT

Dedalus is an open-source hydrodynamics code, written in Python 2.7. It is a pseudospectral code, meaning it uses fast Fourier transforms (FFTs) to compute spatial derivatives. It was designed to be flexible and easy to use, with the FFTs handled by external libraries and/or parallelization to abate the performance penalties of using a high-level language. The code makes extensive use of object-oriented programming, facilitating the modular implementation of different domain representations and physics. Dedalus is currently hosted on a public Bitbucket repository[1].

MHD Equations

Incompressible MHD is primarily governed by four equations[2] [3]: the Navier-Stokes equation with the Lorentz force and viscosity, the induction equation with magnetic diffusivity, the mass continuity equation, and Gauss's law for magnetism:

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla p}{\rho_0} + \frac{\mathbf{F_L}}{\rho_0} + \nu \nabla^2 \mathbf{u}, \tag{1}$$

$$\partial_t \mathbf{B} = \nabla \times (\mathbf{u} \times \mathbf{B}) + \eta \nabla^2 \mathbf{B}, \tag{2}$$

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0 \xrightarrow{\text{incomp.}} \nabla \cdot \mathbf{u} = 0, \tag{3}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{4}$$

Expanding the Lorentz force as

$$\mathbf{F_L} = \frac{(\nabla \times \mathbf{B}) \times \mathbf{B}}{4\pi} = \frac{\mathbf{B} \cdot \nabla \mathbf{B}}{4\pi} - \frac{\nabla B^2}{8\pi}, \tag{5}$$

Eq. (1) becomes

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla P_{tot}}{\rho_0} + \frac{\mathbf{B} \cdot \nabla \mathbf{B}}{4\pi\rho_0} + \nu \nabla^2 \mathbf{u}, \tag{6}$$

$$P_{tot} = p + \frac{B^2}{8\pi}. \tag{7}$$

[1] http://bitbucket.org/jsoishi/dedalus

[2] All presented here in Gaussian units, and with ∂_t indicating $\frac{\partial}{\partial t}$.

[3] With $\mathbf{u}$ the velocity field, $\mathbf{B}$ the magnetic field, p the pressure, ρ the density, ν the kinematic viscosity, and η the magnetic diffusivity.

Using the identity $\nabla \times (\mathbf{A} \times \mathbf{B}) = \mathbf{A}(\nabla \cdot \mathbf{B}) - \mathbf{B}(\nabla \cdot \mathbf{A}) + (\mathbf{B} \cdot \nabla)\mathbf{A} - (\mathbf{A} \cdot \nabla)\mathbf{B}$ along with the constraining equations (Eqs. (3) and (4)), Eq. (2) becomes

$$\partial_t \mathbf{B} = \mathbf{B} \cdot \nabla \mathbf{u} - \mathbf{u} \cdot \nabla \mathbf{B} + \eta \nabla^2 \mathbf{B}. \tag{8}$$

Spectral Implementation

In a periodic domain, any sufficiently smooth field variable can be represented by its discrete Fourier decomposition on the grid. That is, for some function f,

$$f(\mathbf{x}, t) = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} \hat{f}(\mathbf{k}, t) e^{i x_i k_i}, \tag{9}$$

$$\hat{f}(\mathbf{k}, t) = \frac{1}{\sqrt{N}} \sum_{\mathbf{x}} f(\mathbf{x}, t) e^{-i x_i k_i}. \tag{10}$$

Pseudospectral codes, such as Dedalus, use FFTs to evolve partial differential equations in Fourier space, where they become systems of ordinary differential equations. This is because the spatial derivatives in the governing equations become inexpensive multiplications under the Fourier transform: $\nabla \xrightarrow{\text{FT}} i\mathbf{k}$. The spectral implementation follows Maron and Goldreich [4].

Taking the Fourier transforms of Eqs. (6), (8), (3), and (4) yields

$$\partial_t \hat{\mathbf{u}} = -\widehat{\mathbf{u} \cdot \nabla \mathbf{u}} - \frac{i\mathbf{k}\hat{P_{tot}}}{\rho_0} + \frac{\widehat{\mathbf{B} \cdot \nabla \mathbf{B}}}{4\pi\rho_0} - \nu k^2 \hat{\mathbf{u}}, \tag{11}$$

$$\partial_t \hat{\mathbf{B}} = \widehat{\mathbf{B} \cdot \nabla \mathbf{u}} - \widehat{\mathbf{u} \cdot \nabla \mathbf{B}} - \eta k^2 \mathbf{B}, \tag{12}$$

$$i\mathbf{k} \cdot \hat{\mathbf{u}} = 0, \tag{13}$$

$$i\mathbf{k} \cdot \hat{\mathbf{B}} = 0, \tag{14}$$

the Fourier space evolution and constraining equation pairs, respectively. Taking the scalar product of $i\mathbf{k}$ with Eq. (11) leads to an expression for the total pressure:

$$\frac{\hat{P_{tot}}}{\rho_0} = \frac{i\mathbf{k} \cdot \widehat{\mathbf{u} \cdot \nabla \mathbf{u}}}{k^2} - \frac{i\mathbf{k} \cdot \widehat{\mathbf{B} \cdot \nabla \mathbf{B}}}{4\pi\rho_0 k^2}. \tag{15}$$

The nonlinear terms ($\widehat{\mathbf{u} \cdot \nabla \mathbf{u}}$, $\widehat{\mathbf{u} \cdot \nabla \mathbf{B}}$, $\widehat{\mathbf{B} \cdot \nabla \mathbf{u}}$, and $\widehat{\mathbf{B} \cdot \nabla \mathbf{B}}$) in Eqs. (11), (12), and (15) are computed in real space. To eliminate aliasing effects, the 2/3 rule is utilized, zeroing any mode with a k component greater than or equal to 2/3 of the Nyquist wavenumber in that direction. This zeroing is done prior to every reverse Fourier transform, after every forward Fourier transform, and at each temporal evolution.

Temporal Integration

The Fourier space ODEs, along with the initial conditions specified at the start of the simulation, form an initial value problem that is integrated using explicit Runge-Kutta methods, specifically the second-order midpoint method. For simulations with viscosity and/or magnetic diffusivity, an integrating factor is used to evaluate the normally linear steps used to construct the Runge-Kutta stages. Consider Eq. (11) for a specified mode $\mathbf{k}$, with the non-viscous terms considered to be constant during an integration step:

$$\partial_t \hat{\mathbf{u}}(t) + \nu k^2 \hat{\mathbf{u}}(t) = \mathbf{RHS}. \tag{16}$$

This is an equation of the form $y'(x) + P(x)y = Q(x)$, which has the exact solution

$$y(x) = \frac{\int Q(x)M(x)dx}{M(x)}, \tag{17}$$

where $M(x) = e^{\int P(x)dx}$ is called the integrating factor. Hence the solution of Eq. (16) at time $t + dt$ is found to be

$$\hat{\mathbf{u}}(t + dt) = \Big[\hat{\mathbf{u}}(t) + \frac{\mathbf{RHS}}{\nu k^2}(e^{\nu k^2 dt} - 1)\Big]e^{-\nu k^2 dt}. \tag{18}$$

Shearing Box

To study the effects of the MRI, local simulations are performed in which the computational domain represents a small part of an astrophysical disc. The domain is taken to be a co-rotating box, whose left edge is a distance r_0 from the axis of rotation, and whose length in each dimension is much less than this fiducial radius. In the co-rotating frame, the unit vector $\mathbf{e_x}$ is in the outward radial direction, and the unit vector $\mathbf{e_z}$ is along the axis of rotation. The box is rotating with an angular velocity $\mathbf{\Omega_0} = \Omega_0\mathbf{e_z} = \Omega(r_0)\mathbf{e_z}$.

The radial dependence of angular velocity in a Keplerian disc, $\Omega(r) = \sqrt{GM}r^{-3/2}$, gives rise to a linear shear flow in this domain: with the domain moving at the angular velocity of the left (inner) edge, the fluid in the box will shear in the x direction with a velocity of $-\frac{3}{2}\Omega_0 x\mathbf{e_y}$, as shown in Fig. 1. This shear motivates the construction of a domain representation and a corresponding physics implementation to handle MHD in a box with an arbitrary local linear shear.

Consider an arbitrary power-law shearing profile, $\Omega(r) = Cr^S$ (which arises from an attractive force of magnitude $\rho_0 C^2 r^{2S+1}$ and gives rise to a linear background shear in the local frame with velocity $S\Omega_0 x\mathbf{e_y}$. Hence $C = \sqrt{GM}$ and $S = -3/2$ for Keplerian rotation). In this case, the centrifugal ($-\mathbf{\Omega_0} \times (\mathbf{\Omega_0} \times \mathbf{r}) = {\Omega_0}^2 r\mathbf{e_x}$) and attractive ($-C^2 r^{2S+1}\mathbf{e_x}$) accelerations partially cancel (via approximations utilizing $x, y, z \ll r_0$):

$$\mathbf{a_{rad}} = ({\Omega_0}^2 - C^2 r^{2S})r\mathbf{e_x} \approx -2S{\Omega_0}^2 x\mathbf{e_x}. \tag{19}$$

With this radial acceleration and the Coriolis acceleration ($-2\mathbf{\Omega_0}\times\mathbf{v}$), Eq. (1) for the velocity field in the rotating frame, $\mathbf{v}$, becomes

$$\partial_t \mathbf{v} + \mathbf{v}\cdot\nabla\mathbf{v} = -\frac{\nabla p}{\rho_0} + \frac{\mathbf{F_L}}{\rho_0} + \nu\nabla^2\mathbf{v} - 2S{\Omega_0}^2 x\mathbf{e_x} - 2\mathbf{\Omega_0}\times\mathbf{v}. \tag{20}$$

Decomposing $\mathbf{v}$ into the background shear flow and velocity perturbations ($\mathbf{v} = S\Omega_0 x\mathbf{e_y} + \mathbf{u}$), Eq. (20) becomes

$$\partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} = -\frac{\nabla p}{\rho_0} + \frac{\mathbf{F_L}}{\rho_0} + \nu\nabla^2\mathbf{u} + 2\Omega_0 u_y\mathbf{e_x} - (2+S)\Omega_0 u_x\mathbf{e_y} - S\Omega_0 x\partial_y\mathbf{u}. \tag{21}$$

To account for the background shear in the evolution of the perturbations, the remesh-free approach of Brucker et al [5] was implemented. Due to the shear in the local frame, the forward and reverse Fourier transforms must be modified to maintain periodicity along the shearing direction:

$$\hat{f}(\mathbf{k},t) = \sum_{\mathbf{x}} f(\mathbf{x},t)e^{-i(x_i k_i - S\Omega_0 x t k_y)}, \tag{22}$$

$$f(\mathbf{x},t) = \sum_{\mathbf{k}} \hat{f}(\mathbf{k},t)e^{i(x_i k_i - S\Omega_0 x t k_y)}. \tag{23}$$

Hence, the fixed-grid wavevector $\mathbf{K}$ becomes a function of the Lagrangian (shearing) wavevector $\mathbf{k}$ and time, $\mathbf{K}(\mathbf{k},t) = (k_x - S\Omega_0 t k_y, k_y, k_z)$, and when transforming to Fourier space, the derivative operators become

$$\partial_x \to i(k_x - S\Omega_0 t k_y) = iK_x, \tag{24}$$

$$\partial_y \to ik_y = iK_y, \tag{25}$$

$$\partial_z \to ik_z = iK_z, \tag{26}$$

$$\partial_t \to \partial_t - iS\Omega_0 x k_y. \tag{27}$$

The shearing box implementations of Eqs. (11) and (15) are then

$$\partial_t \hat{\mathbf{u}} = -\widehat{\mathbf{u}\cdot\nabla\mathbf{u}} - \frac{i\mathbf{K}\hat{P_{tot}}}{\rho_0} + \frac{\widehat{\mathbf{B}\cdot\nabla\mathbf{B}}}{4\pi\rho_0} - \nu K^2\hat{\mathbf{u}} + 2\Omega_0\hat{u_y}\mathbf{e}_{\hat{\mathbf{x}}} - (2+S)\Omega_0\hat{u_x}\mathbf{e}_{\hat{\mathbf{y}}}, \tag{28}$$

$$\frac{\hat{P_{tot}}}{\rho_0} = \frac{i\mathbf{K}\cdot\hat{\mathbf{M}}}{K^2} - \frac{i\mathbf{K}\cdot\hat{\mathbf{L}}}{4\pi\rho_0 K^2} - \frac{2i\Omega_0\hat{u_y}K_x}{K^2} + \frac{(1+S)2i\Omega_0\hat{u_x}K_y}{K^2}. \tag{29}$$

Eq. (8) becomes, in velocity perturbations,

$$\partial_t\mathbf{B} = \mathbf{B}\cdot\nabla\mathbf{u} - \mathbf{u}\cdot\nabla\mathbf{B} + \eta\nabla^2\mathbf{B} + S\Omega_0 B_x\mathbf{e_y} - S\Omega_0 x\partial_y\mathbf{B}, \tag{30}$$

and the shearing box implementation of Eq. (12) becomes

$$\partial_t\hat{\mathbf{B}} = \widehat{\mathbf{B}\cdot\nabla\mathbf{u}} - \widehat{\mathbf{u}\cdot\nabla\mathbf{B}} - \eta K^2\mathbf{B} + S\Omega_0\hat{B}_x\mathbf{e}_{\hat{\mathbf{y}}}. \tag{31}$$

FFTs and Parallelization

Although Dedalus is written in Python, the FFTs dominate the computational cost of a simulation. Optimizing the FFTs largely negates the performance penalties of using a high-level language, while maintaining the ease of use and speed of development of Python.

Outsourcing the FFTs from Python to FFTW, a C-based library that optimizes FFT routines based on local hardware, results in a substantial speed improvement over Python's (i.e. NumPy's) built-in FFT algorithms. Much greater gains can be made on a graphics processing unit (GPU) using Nvidia's CUDA architecture to compute the FFTs and other calculations.

Finally, MPI-based parallelization has been implemented, allowing a single simulation to simultaneously run as N separate tasks. To achieve this, the computational domain is evenly divided among the N tasks along the k_z direction in Fourier space. The reverse Fourier transform is then accomplished in a series of steps, as depicted in Fig. 2. First, each task performs a 1D IFFT in the k_x direction on its dataset, and the shearing phase shift is applied as in Eq. (23), if necessary. Second, an MPI All-To-All call is issued, in which each task evenly divides its data N times in the x direction, and sends the Nth slab to the Nth task. Each task then stacks the N slabs it has received, and performs a 2D IFFT in the k_y and k_z directions. The resulting datasets have gone through a full 3D IFFT, and the data is evenly divided among the tasks along the x direction in real space. The parallelized forward transform is the reverse of this process.

RESULTS

Code Verification and Testing

Multiple tests were performed to verify the MHD implementation in Dedalus. First, the motion of a shear-Alfvén wave, an eigenmode of the linearized MHD equations which travels along magnetic field lines, was tested. A constant background magnetic field $\mathbf{B_0}$ was setup across the domain, along with the velocity and magnetic field perturbations corresponding to a shear-Alfvén eigenmode. Shear-Alfvén waves were found to propagate with the correct phase velocities based on their direction of travel and the magnetic field strength: $v_A \propto B_0 \cos\theta$, where v_A is the speed of a shear-Alfvén wave and θ is the angle between the magnetic field and the wavevector.

This test was repeated with nonzero viscosity and magnetic diffusion. Fig. 3 shows the resulting amplitude and phase velocity plots for several seeded modes, which behave as expected. The velocity and magnetic field amplitudes decay exponentially, with rates proportional to wavevector magnitude, as expected from Eq. (18), and the phase velocities are constant in time and of the expected magnitude (here $\mathbf{B_0}$ is directed along $(1, 0, 0)$).

The advection of a low amplitude magnetic field was also tested. A small ($|\mathbf{B}| \approx 10^{-3}$), random magnetic field was imposed on a constant fluid velocity of order unity. The magnetic field forced velocity perturbations of the same order, and both fields were advected with the background fluid velocity. This behavior was expected, since the non-linear terms are negligible with such small field perturbations.

The shearing box implementation was tested using the swinging wave test of Lithwick [6]. The wavefronts of the vorticity waves are seen to turn with the local domain's shear, as expected. Snapshots of the wave at several times are shown in Fig. 4.

Science Results

At the time of this report, the parallelization techniques described above are being finalized and tested. While they are critical for high-resolution simulations quantitatively studying the MRI, linear and early non-linear behavior can be investigated at lower resolution. These MRI simulations begin with a vertical magnetic field and small ($\approx 10^{-6}$) random velocity perturbations. Channel modes, precursors to turbulence in the MRI, are seen to form and grow exponentially. The mode formation is seen in Fig. 5, which depicts $\mathbf{u_x}$ in one xz plane at several times. The modes lie primarily in xy planes and are nearly axisymmetric, as expected. Higher resolution simulations following these modes further into the non-linear domain will be performed utilizing the parallel architecture of Dedalus.

DISCUSSION AND CONCLUSIONS

Dedalus is among the first entirely open-source spectral MHD codes. The Python development environment facilitates ease of use and code development, and the project's emphasis on object-oriented techniques have helped make Dedalus a very modular code which can be easily adapted to study a variety of problems. The parallelized FFT algorithms, the use of external libraries, and the incorporation of GPU-based calculations using CUDA all contribute substantially to the performance of the code. With shearing-box simulations underway, the MHD capabilities of Dedalus may help identify points of comparison and departure between current lab-based HMRI experiments and the standard MRI operating in astrophysical accretion discs.

ACKNOWLEDGMENTS

Dedalus is currently being developed by Jeffrey S. Oishi, Oliver Hahn, Keaton J. Burns, and Greg Peairs. This project was made possibly by the U.S. Department of Energy's SULI program, and the KIPAC group at SLAC National Accelerator Laboratory and Stanford University.

REFERENCES

[1] G. Lesur and P.-Y. Longaretti, "Impact of dimensionless numbers on the efficiency of magnetorotational instability induced turbulent transport," *Monthly Notices of the Royal Astronomical Society*, vol. 378, no. 4, pp. 1471–1480, Jul. 2007. [Online]. Available: http://doi.wiley.com/10.1111/j.1365-2966.2007.11888.x

[2] S. A. Balbus and J. F. Hawley, "A powerful local shear instability in weakly magnetized disks. I - Linear analysis. II - Nonlinear evolution," *The Astrophysical Journal*, vol. 376, pp. 214–233, 1991. [Online]. Available: http://adsabs.harvard.edu/abs/1991ApJ...376..214B

[3] O. N. Kirillov and F. Stefani, "ON THE RELATION OF STANDARD AND HELICAL MAGNETOROTATIONAL INSTABILITY," *The Astrophysical Journal*, vol. 712, no. 1, pp. 52–68, Mar. 2010. [Online]. Available: http://arxiv.org/abs/0911.0067

[4] J. Maron and P. Goldreich, "Simulations of Incompressible Magnetohydrodynamic Turbulence," *The Astrophysical Journal*, vol. 554, no. 2, pp. 1175–1196, Jun. 2001.

[Online]. Available: http://arxiv.org/abs/astro-ph/0012491http://adsabs.harvard.edu/abs/2001ApJ...554.1175M

[5] K. A. Brucker, J. C. Isaza, T. Vaithianathan, and L. R. Collins, “Efficient algorithm for simulating homogeneous turbulent shear flow without remeshing,” *Journal of Computational Physics*, vol. 225, no. 1, pp. 20–32, Jul. 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0021999106004827

[6] Y. Lithwick, “Nonlinear evolution of hydrodynamical shear flows in two dimensions,” *The Astrophysical Journal*, vol. 670, pp. 789–804, 2007.

FIGURES

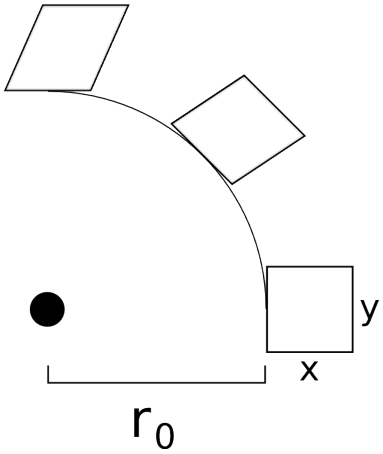


Figure 1: The motion of fluid tracers initially aligned with the edges of the co-rotating local domain, as viewed from an inertial frame. Lower angular velocity at larger radii results in a local linear shear across the x direction of the local domain. The size of the box relative to r_0 is exaggerated for clarity.

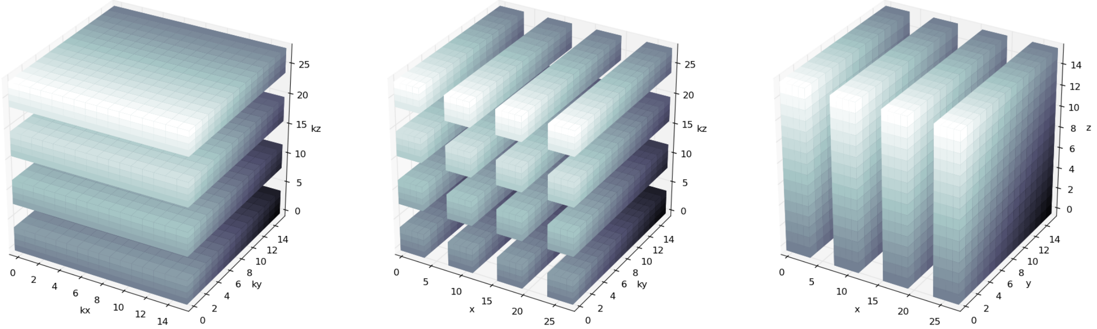


Figure 2: Data distribution using MPI parallelization: task cuts along k_z in Fourier space, data passing during MPI All-To-All call, and task cuts along x in real space.

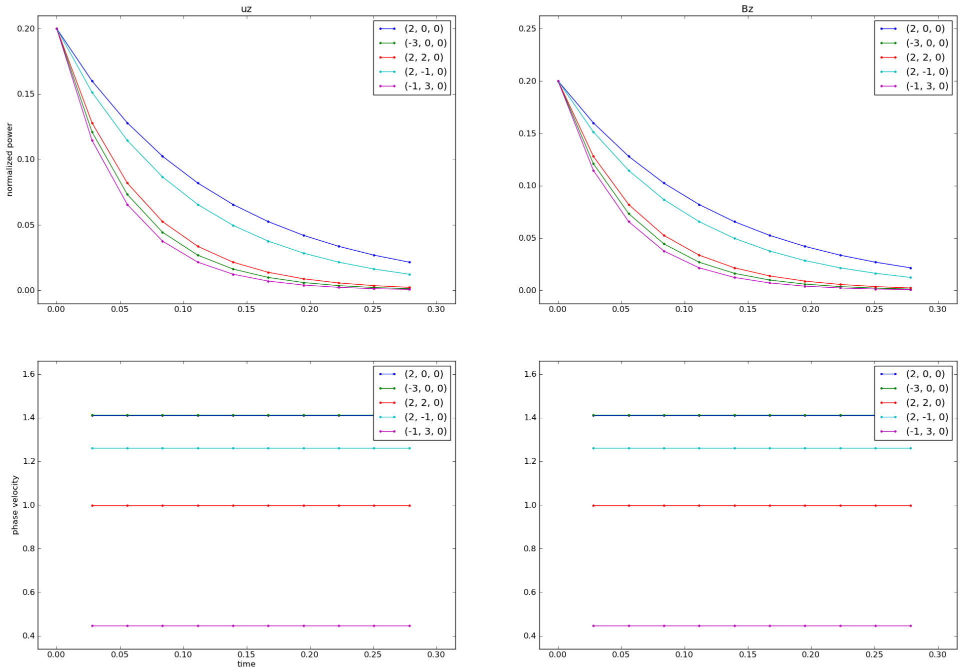


Figure 3: Amplitude and phase velocity evolution of several shear-Alfvén waves, with viscosity and magnetic diffusivity.

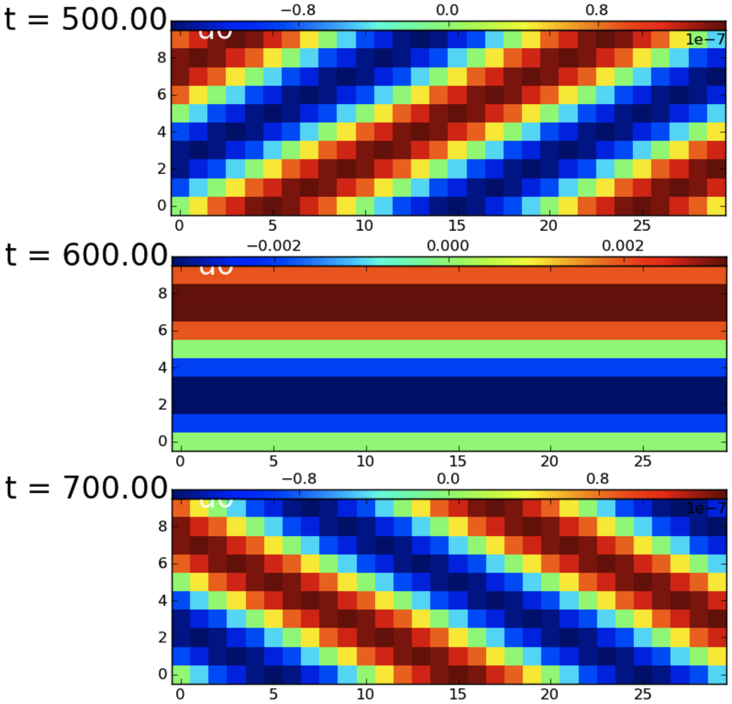


Figure 4: Swinging wave demonstrating local linear shear ($\mathbf{u_x}$ depicted).

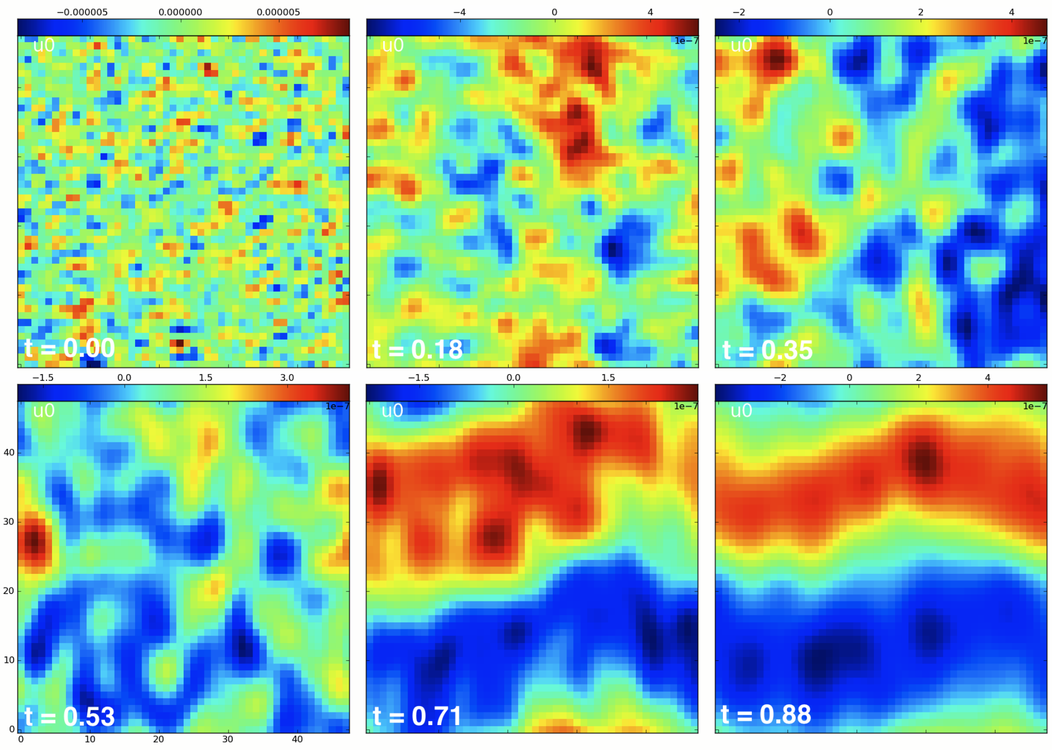


Figure 5: Formation of channel modes ($\mathbf{u_x}$ depicted in an xz plane).

SLAC-TN-12-020

Simulations and Analysis of an Infrared Prism Spectrometer for Ultra-short Bunch Length Diagnostics at the Linac Coherent Light Source

Julie Cass

Office of Science, Science Undergraduate Laboratory Internship Program

University of Notre Dame

SLAC National Accelerator Laboratory
Menlo Park, California

August 19, 2011

Prepared in partial fulfillment of the requirement of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Josef Frisch in the ICD Accelerator Physics & Engineering Department at SLAC National Accelerator Laboratory.

Participant: ___________________________________
Signature

Research Advisor: ___________________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

ABSTRACT

Simulations and Analysis of an Infrared Prism Spectrometer for Ultra-short Bunch Length Diagnostics at the Linac Coherent Light Source.

Julie Cass (University of Notre Dame, Notre Dame, IN 46556), Josef Frisch (ICD Accelerator Physics & Engineering, SLAC National Accelerator Laboratory, Stanford, CA 94025).

We require a means of determining the minimum achievable bunch length at the Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory. Current methods of measurement in the time domain do not have the resolution to measure ultra-short bunch lengths below 10 fs. An alternative method has been designed, using a single-shot near- to mid-infrared prism spectrometer for bunch length diagnosis in the frequency domain. This design requires the precision alignment of sensitive optical equipment and complete understanding of the spectrum reaching the spectrometer's detector surface. This paper provides information on preliminary simulations designed to model beam behavior in the optical set-up and at the detector surface for comparison to measured values as an indication of the current achievable optical alignment and the study of optimal detector orientation. The results of these simulations indicate the necessity for a method of greater precision in optical alignment and further study of achievable spectrometer range.

INTRODUCTION

The Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory is a groundbreaking free electron laser (FEL), using magnetic fields to oscillate high-energy electrons and generate ultrafast X-ray pulses [1]. This FEL is capable of producing pulses with ultra-short bunch lengths, enabling scientists to image molecules and atoms in motion, pushing the limits of technology in pursuit of a deeper understanding of fundamental chemical, biological, and physical properties of materials [2]. The interpretation and understanding of the data collected in these experiments relies on a complete understanding of the parameters used in driving the FEL. While the minimum achievable bunch length at LCLS is known to have a value below 10 fs, the precise value remains unknown. Measurement of this parameter requires a mechanism capable of probing pulses shorter than 10 fs, a scale too small to be resolved by existing time domain detectors. Scientists at the LCLS have developed an alternative design implementing a broadband infrared prism spectrometer to probe for this parameter in the frequency domain.

This new approach is centered on the concept that information about the bunch length can be extracted from the spectrum of optical transition radiation generated by the electron beam. This radiation can be characterized as a relativistic effect, resulting in the emission of electromagnetic radiation as a charge particle approaches a conducting foil. The implementation of an inverse Fourier transform of the transition radiation spectrum allows for the inversion of spectral information for analysis in the time domain. This provides an absolute means for calculation of the FEL bunch length from the information contained in the transition radiation spectrum. The wavelength range of interest for the transition radiation from the LCLS beam is within the near- to mid-infrared range.

Transition radiation from FEL electrons is expanded and collimated using precision optics [3] and steered to the spectrometer prism. The index of refraction of the prism material is wavelength dependent, creating a dispersion relation between the wavelengths comprising the radiation beam and the angle of refraction as the light emerges from the prism. The dispersed light is focused onto the detecting surface of a pyroelectric detector array [4]. The resulting spectrum will then be used in the inverse Fourier analysis to determine the FEL bunch length. It is crucial for the success of this approach that the spectrometer is designed with precisely aligned optics for propagation of transition radiation from the foil to the detector. For this purpose, a method for analyzing the behavior of the beam in the optical system of the spectrometer has been determined. We require a calculation of both how the beam would behave for an ideal system and for a system containing aberrations with the introduction of alignment errors.

METHODS AND MATERIALS

Spectrometer Design

The infrared spectrometer design is comprised of an electronic system to drive and collect data from the pyroelectric line detector and an optical set-up employing lenses and mirrors to expand, collimate and steer the beam towards this detector. A diagram of the optical system as it would be implemented in the LCLS is provided in Figure 1.

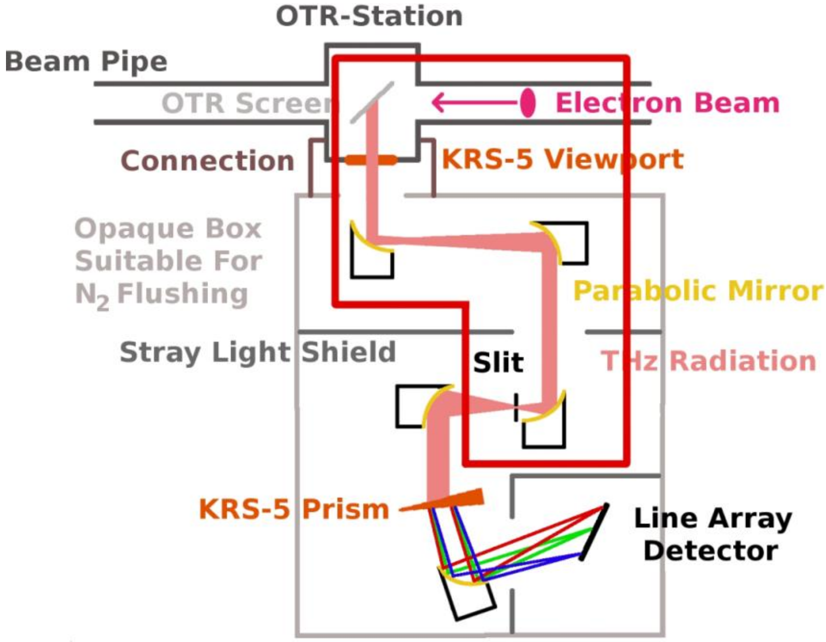


Figure 1. Preliminary optical layout for the prism spectrometer

The mirrors used in this optical design are gold-coated, 90-degree, off-axis parabolic mirrors (OAPs). Preliminary calculations for this project showed that the optical transition radiation form the LCLS is likely to fall in the infrared wavelength range. These mirrors were chosen for their reflectivity in the wavelength range of interest. They are particularly difficult to align, due to the challenge in locating the off-axis central point. A Helium-Neon (HeNe) laser emitting light in the visible range at 632.816 nm was initially used for positioning of the system elements and a Thallium Bromo-Idodide (KRS-5) crystal was selected to act as the spectrometer prism, dispersing the transition radiation over the detector. This material was chosen for its dispersive and transmission properties in the range of wavelengths of interest.

MATLAB Simulation for Idealized Modeling

As proper alignment is essential to our project, we required a calculation of the tolerances for beam behavior while navigating the optical system. A mathematical interpretation of the optical system was programmed in MATLAB [6] to graphically simulate the size of the beam as it is passed through various optical elements using “ray transfer matrix analysis” [7]. The programming software MATLAB was selected for this simulation due to its capabilities in matrix-based calculation. In the ray transfer matrix approach to optical analysis, light rays are represented as vectors in terms of height *r*, and angle of inclination θ relative to the optical axis. Each optical element or length of free space propagated by the beam is represented by 2 x 2 matrix; when the initial light ray is multiplied by the ray transfer matrix, the result is a new vector, containing information about the outgoing beam height and angle of inclination. Examples of ray transfer matrices for light refraction through a lens with focal length *f* and the propagation of a ray through free space of length *d* are provided in Equations 1 and 2 respectively.

$$\begin{pmatrix} r_f \\ \theta_f \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} r_i \\ \theta_i \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} r_f \\ \theta_f \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_i \\ \theta_i \end{pmatrix} \tag{2}$$

If we represent each successive optical element or length of free space (at 100 μm intervals) in the system as ray transfer matrix M_i, we can represent the compounded effects of the entire optical system with a single ray transfer matrix M:

$$M = \prod_{i=1}^{n} M_i \tag{3}$$

where n is the total number of optical elements and 100 μm intervals of free space in the system. While basic ray transfer matrix analysis is designed for modeling ideal behavior of light rays with negligible waist size, our system required ray transfer analysis of Gaussian beam propagation, as the HeNe laser used to align the system produces a Gaussian beam. In order to incorporate the effects of diffraction into our design, we generated a vector to represent the initial Gaussian beam in terms of a complex beam parameter *q*, calculated by the following equation:

$$\frac{1}{q} = \frac{1}{R} - \frac{i\lambda}{\pi w^2} \qquad (4)$$

where *q* is the complex beam parameter, *R* is the radius of curvature of the beam, λ is the wavelength of the beam and w is the beam waist size or beam diameter [8]. With this notation, the equation modeling the transformation of the initial beam to a final beam through ray transfer analysis takes on the modified form:

$$\begin{pmatrix} q_f \\ 1 \end{pmatrix} = k \cdot M \cdot \begin{pmatrix} q_i \\ 1 \end{pmatrix} \qquad (5)$$

where the beam is now represented by a vector containing the complex beam parameter and *k* is a constant chosen to maintain the normalization of the bottom component of the beam vector. Carrying out this multiplication and solving for the inverse *q* value leaves us with:

$$\frac{1}{q_f} = \frac{M(2,1) + {}^{M(2,2)}\!/\!_{q_i}}{M(1,1) + {}^{M(1,2)}\!/\!_{q_i}} \qquad (6)$$

giving us a way to find the new waist size from the definition of the beam parameter *q* and the elements contained in the ray transfer matrices and initial beam parameter. From this information, the MATLAB code calculates the beam waist at each element and free space interval and outputs a graph of these waist sizes.

This simulation was designed with matrices representing ideal versions of each optical element; off-axis parabolic mirrors and lenses are treated according to the thin lens approximation. The errors in alignment are not accounted for; the program offers an outline for behavior in a system with infinite precision as a guideline for proper alignment and a means of comparison between ideal beam behavior and the behavior, or beam sizes, seen in the actual set-up.

ZEMAX Simulation

Disagreements between the measured beam sizes and those calculated in the MATLAB simulation were anticipated due to limitations in the precision aligning of optics. We wished to determine the spectrometer tolerances achievable considering these limitations. For a more accurate representation of the design and calculation of this resolution, the optical simulation software ZEMAX was used [9]. This software provides the ability to trace rays of light as they are reflected, refracted, and dispersed in an optical system.

The spot size was analyzed at two crucial points in the optical system: the focal points of the first and third OAPs, with the latter corresponding to the proposed location of the line array detector. As these are the two tightest focuses in the optical design, they are excellent test points for the design tolerances in angular misalignments. The beam sizes measured and simulated in this project are the $1/e^2$ beam diameters. For analysis of the beam size at the focal point of the first OAP, a ZEMAX simulation of a single 90-degree OAP with a 50.8 mm focal length and mirror diameter perpendicular to the optical axis of 50.8 mm was generated. The 3D plot of this mirror produced by ZEMAX is provided in Figure 2.

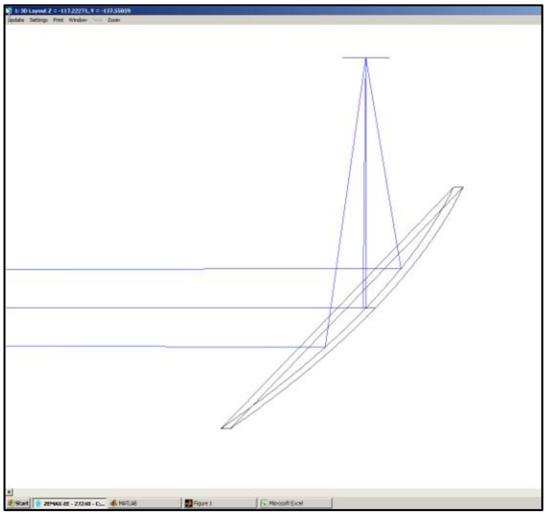


Figure 2. ZEMAX ray tracing of reflection off of a 90-degree off-axis parabolic mirror. Light rays are incident from the left side of the image and are focused by the OAP.

As the HeNe laser used has a beam output diameter 2w_0 of 2 mm and is expanded in a pair of lenses, or a telescope, to approximately 17 times its initial size, the entrance beam size for the simulation was set to 33 mm. The beam size at the OAP focus was first calculated for the case of a perfectly aligned mirror. Mirror tilts on the order of a tenth of a degree about the horizontal, vertical and optical axes were added in keeping with the estimated tolerances for physical alignment of the optics. These angles were adjusted until a beam size similar to the beam size measured on the optical table was reached. The tilts needed to cause these beam sizes were recorded for comparison with estimated order of misalignment error.

The design for testing beam sizes due to aberrations at the final detector location was more complex. While the functionality of the spectrometer depends on the dispersion of wavelengths through the prism, this dispersion also introduces chromatic aberrations in the focal plane of the spectrum. Previous simulations of the beam path indicate that, due to these chromatic aberrations, the wavelength-dependent focal plane is not perpendicular to the optical axis; a 44 degree tilt of the detector plane was required to minimize these effects [10]. We wished to create a new simulation of this portion of the spectrometer design to account for differences in mirror diameters and beam sizes, as well as refocusing the system to the best focus for all wavelengths of dispersed light. From this model we could then test the angular tolerances and determine the new optimal detector tilt.

The simulation was first designed to focus several wavelengths of light at the detector to simulate actual conditions for application in LCLS. While chromatic aberrations make perfect foci impossible to produce, we could optimize the optical system in ZEMAX to achieve minimum beam diameters and sharpest possible foci. A simple example of the process is shown for a basic plano-concave lens in Figure 3a-b. When the ZEMAX function 3D-plot is implemented for a simple lens design (Figure 3a) we can see the unclear focus produced as different rays of light focus at different points.

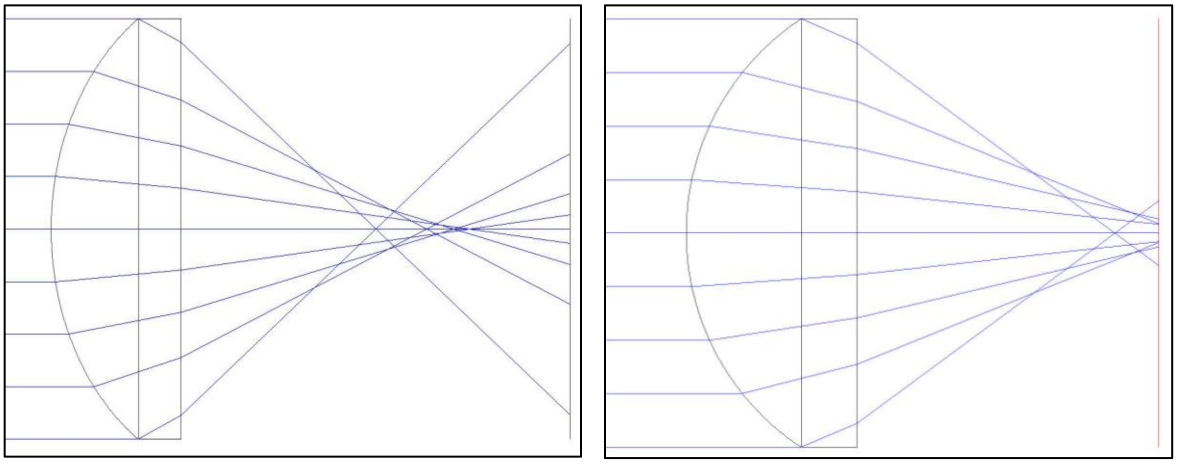


Figure 3a and 3b. ZEMAX tracing of light rays intiated at the far left of the image, are refracted through plano-concave lens with spherical aberrations onto an image plane (line at far right), shown before and after focusing optimization

Owing to the chromatic aberrations, the "focal length" specified for the lens is no longer the point of smallest beam radius. Instead, we defined a new adjusted focal length as the point where the tightest beam waist is located. Through use of the Merit Function Editor in ZEMAX, system parameters specified by the user in the Lens Data Editor as variables may be adjusted, allowing for system optimization. The Geometric Encircle Energy (GENC) merit function was used for the focusing of elements in this simulation. GENC places the image plane closest to the point at which a user-specified fraction of the beam energy falls within a user-specified beam radius, constraining beam size at the image plane. Using the Optimization feature of ZEMAX allows the software to adjust any “variable” parameters in the lens data editor, such as distance from the object to the image plane, in order to meet the target encircled energy values as closely as possible. Figure 3b displays the resulting 3D plot when the GENC merit function was used to constrain the greatest possible amount of beam energy to the smallest possible radius at the image plane. This technique for image focus, once understood for simple cases as the lens shown above, was then applied to our optical design for precision focusing of OAPs about the prism and detector for a measurement of the new detector tilt needed and angular tolerances in beam alignment. The GENC target fraction of beam energy set in order to simulate the $1/e^2$ beam diameter for comparison with the measured beam sizes.

With a design completed for the spectrometer dispersion of several wavelengths of light, the model was adjusted to simulate only the beam path of the HeNe laser and the system was re-optimized for this single wavelength. A ZEMAX plot of this model is given in Figure 4.

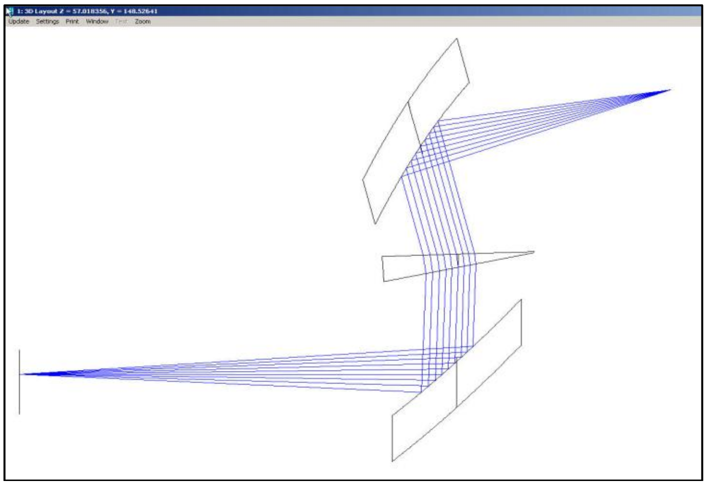


Figure 4. ZEMAX plot of OAP focusing of HeNe light onto the detector plane (line on the far left) after traveling through the final two OAPs and prism

To determine the angular adjustments needed to cause a beam size similar to the measured beam size, coordinate breaks with angular tilts were added to the design. Angular tilts about the optical, vertical, and horizontal axes were added and the system was optimized for each new design angle. The tilts were adjusted until the simulated beam size was similar to the measured beam size and the resulting tilts were compared to the expected magnitude of tilt misalignments. We could then consider whether the aberrations in measured beam sizes were in fact due to misalignments and whether they were close enough to the values predicted by MATLAB and ZEMAX to be negligible in spectrometer functionality.

RESULTS

Figure 5 below displays a plot of the beam waist results produced through the implementation of the MATLAB optical simulation code.

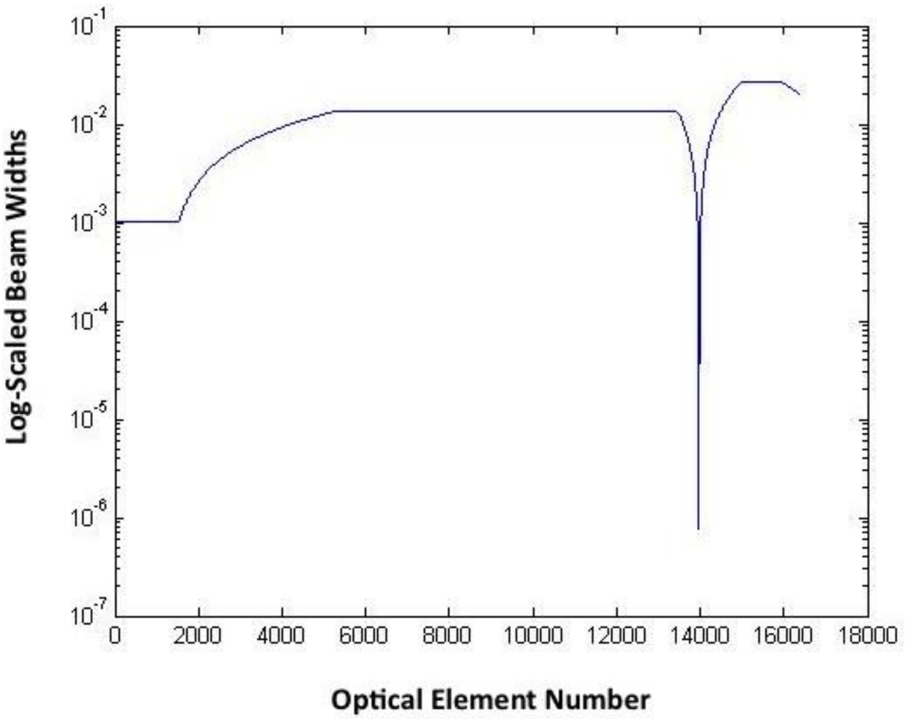


Figure 5. Log-scaled plot of beam size at various points in the spectrometer optics predicted by MATLAB simulation

The Optical Element Numbers given in the x-axis represent which step through the optical system each element falls at. The step size for this system was 0.1 mm. Each Optical Element is either a distance step of this size or an optical element itself.

At the first focal point of interest, following light reflection from the first parabolic mirror, we measured a beam diameter of approximately 172 μm. Our simulations in MATLAB predict a beam diameter of approximately 65μm. To determine whether the magnification of the measured beam size by three times the predicted size could be the result of misalignments, we compared these numbers to those generated by the ZEMAX simulation. The Spot Diagram display in ZEMAX providing the beam diameter at this focal point through general ray tracing predicted a beam size of 0 μm without any mirror tilts or aberrations. Tilts of approximately 0.4 degrees both along the optical axis and perpendicular to the optical table were needed in order to produce a beam with diameter equal to the measured beam diameter.

At the focal point of the third parabolic mirror, with no detector tilt implemented, we measured the beam diameter of approximately 280 μm. MATLAB simulations predict a beam size of approximately 37.5 μm at this location. It was determined that the disagreements between these two values were also likely the result of limits in precision alignment of the system, though these effects seemed much greater at the detector location than at the focal point of the first parabolic mirror. This hypothesis was tested with the study of beam sizes predicted by the ZEMAX model. Without aberrations added, this simulation predicted a beam diameter of 14 μm. Tilts of approximately 1 degree both about the vertical, horizontal, and optical axes were needed to cause beam sizes comparable to the measured beam size. A table containing all relevant measured and simulated beam sizes is provided in Table 1.

	Measured	**MATLAB**	**ZEMAX**	**ZEMAX Tilts Required**
50.8 mm OAP Focal Point	**172** μm	**65** μm	**0** μm	**0.4°**
Detector Surface	**280** μm	**37.5** μm	**14** μm	**1°**

Table 1. Table of measured and simulated beam diameters (μm) at two points of interest in the optical system, and simulation tilts needed to produce the measured beam sizes

With a new simulation designed to match the design parameters of our spectrometer, multiple wavelengths of light were added to the simulation and the detector tilt and distance were optimized for focusing of all wavelengths. A new detector tilt of 45 degrees was determined to minimize chromatic aberrations at the detector plane. While this tilt provided a decent focus for all wavelengths of light, this design created a dispersion of approximately 16 mm. The detector, however, has a length of 14 mm and not all wavelengths of light transmitted by the KRS-5 prism would fit on the detector for analysis. Ray tracing images of this portion of the system and light focusing at the detector are given in Figure 6. A spot diagram of all wavelengths falling on the detector surface is given in Figure 7 and a table listing the wavelengths plotted in these simulations is given in Table 2.

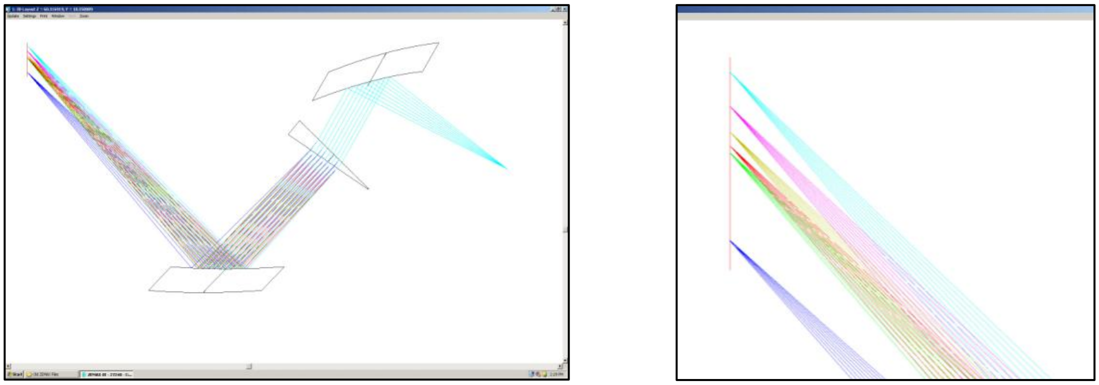


Figure 6. Ray tracing for focusing of several wavelengths onto the detector plane. Left: The simulated beam initiates at the far right, is reflected by the 105.6 mm OAP, disperses through the prism, reflects off of the 177.8 mm OAP and falls on the detector surface. Right: Close-up of all wavelengths focusing onto the detector plane.

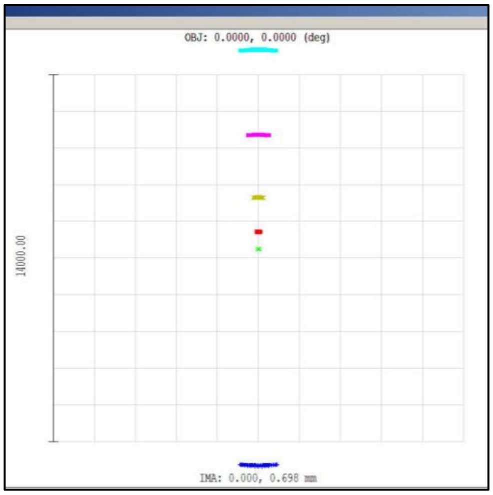


Figure 7. Spot diagram for the detector surface. The vertical axis represents the length of the detector array. Each wavelength is confined to a thin strip along the vertical or detector axis

Wavelengths (μm)	
Blue	**0.632**
Green	**3.39**
Red	**10.6**
Gold	**20.0**
Magenta	**30.0**
Turquoise	**39.0**

Table 2. List of wavelengths (μm) plotted in the ZEMAX simulation shown in Figure 7.

DISCUSSION AND CONCLUSION

Comparisons of the measured beam size and MATLAB simulated beam sizes at both the focal point of the 50.8mm OAP and the detector plane yielded measured results larger than those simulated by MATLAB. This result was consistent with our hypothesis that discrepancies in these beam sizes could be the result of minor errors in alignment. When the system was modeled in ZEMAX and tilts were added to simulate mirror misalignment, we found that the tilts necessary to match the measured beam sizes were less than or equal to one degree. We believe that precision in mirror alignment is limited to tolerances greater than one degree. However, the aberrations found in our system are great enough, that we believe a superior method for precision alignment is needed for the spectrometer to function properly. We will require further research and testing to develop a method for fine-tuning this alignment for a closer match between the predicted and measured beam sizes.

Spectrometer design considerations originally predicted the transmission range of KRS-5 prism to be the limiting factor in the range of wavelengths detected in the spectrometer. Through our ZEMAX simulations, we found that no detector tilt can be implemented in our current design to allow all wavelengths transmitted by this prism. This result implies that either our spectrometer detection range will be bounded for larger wavelengths at 38 μm rather than 40 μm or the design will need to be changed. We believe that this decrease in detector range is small enough that the current spectrometer design will be sufficient. The spectrometer contains aberrations large enough that a new method for alignment is needed but has a detection range sufficient to produce the spectrum needed for FEL bunch length diagnosis.

ACKNOWLEDGEMENTS

This research was carried out at the SLAC National Accelerator Laboratory between May 23 and August 19, 2011 under the direction of Josef Frisch. The author would like to thank the United States Department of Energy and the SLAC National Accelerator Laboratory for supporting this research through the creation, organization, and funding of the Summer Undergraduate Laboratory Internship (SULI) program. She would also like to thank her mentor Joe Frisch and research partners Gilles Dongmo-Momo and Kiel Williams. Thanks also go to Alan Fisher for his guidance in optical simulation design.

REFERENCES

[1] Y. Ding *et al.*, "Measurements and Simulations of Ultralow Emittance and Ultrashort Electron Beams in the Linas Coherent Light Source", PRL 102, 254801, 2009

[2] Linear Coherent Light Source http://lcls.slac.stanford.edu/

[3] K. Williams, "Optical Design of a Broadband Infrared Spectrometer for Bunch Length Measurement at the Linac Coherent Light Source," SLAC National Accelerator Laboratory SULI Program 2011, Palo Alto, CA

[4] Pyreos Ltd, http://www.pyreos.com/

[5] G. Dongmo-Momo, "Infrared Spectroscope for Electron Bunch-length Measurement: Heat Sensor Parameters Analysis," SLAC National Accelerator Laboratory SULI Program, Palo Alto, CA 2011

[6] MATLAB 7.0.4, The MathWorks Inc.

[7] Bahaa E. A. Saleh and Malvin Carl Teich (1991). *Fundamentals of Photonics*. New York: John Wiley & Sons. Section 1.4, pp. 26-36

[8] Gaussian Beams http://www.rp-photonics.com/

[9] Radiant ZEMAX LLC, http://www.zemax.com

[10] C. Behrens et al., "Design of a Single-Shot Prism Spectrometer in the Near- and Mid-Infrared Wavelength Range for Ultra-Short Bunch Length Diagnostics", DIPAC'11, Hamburg, Germany, 201

SLAC-TN-12-022

Infrared Spectroscope for Electron Bunch-length Measurement:

Heat Sensor Parameters Analysis.

Gilles Dongmo - Momo

Office of Science, Science Undergraduate Laboratory Internship (SULI) Program

Towson University
Baltimore, Maryland

SLAC National Laboratory
Menlo Park, California

08/20/2011

Prepared in partial fulfillment of the requirements of the Office of Science, U.S. Department of Energy's SULI under the direction of Dr. Josef Frisch at the Linac Coherent Light Source (LCLS).

Participant: ______________________________

Signature

Research Advisor: ______________________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Table of Contents

ABSTRACT:

Infrared Spectroscope for Electron Bunch-length Measurement: Heat Sensor Parameters Analysis. Gilles Dongmo - Momo (Towson University, Baltimore, MD 21252) Josef Frisch (ICD Accelerator Physics & Engineering department, SLAC National Laboratory, Menlo Park CA 94025)

The Linac Coherent Light Source (LCLS) is used for many experiments. Taking advantage of the free electron laser (FEL) process, scientists of various fields perform experiments of all kind. Some for example study protein folding; other experiments are more interested in the way electrons interact with the molecules before they are destroyed. These experiments among many others have very little information about the electrons' x-ray produced by the FEL, except that the FEL is using bunches less than 10 femtoseconds long. To be able to interpret the data collected from those experiments, more accurate information is needed about the electron's bunch-length. Existing bunch length measurement techniques are not suitable for the measurement of such small time scales. Hence the need to design a device that will provide more precise information about the electron bunch length. This paper investigates the use of a pyreoelectric heat sensor that has a sensitivity of about 1.34 micro amps per watt for the single cell detector. Such sensitivity, added to the fact that the detector is an array sensor, makes the detector studied the primary candidate to be integrated to an infrared spectrometer designed to better measure the LCLS' electron bunch length.

INTRODUCTION:

The X-rays produced by the LCLS have the same profile in the time domain as the electrons producing them. More information about the electrons can be translated to more information about the x-rays. Many techniques were developed to characterize electrons. To give accurate information about the LCLS' electron bunch, those techniques were able to approximate the electron bunch length to be shorter than 10 femto-seconds. [1] But the need to get better image resolution out of the LCLS's Free Electron Laser becomes a necessity added to the fact that current experimental efforts require a more precise measurement for LCLS electron pulse duration, new bunch-length measurement techniques are being developed[2]. An infrared spectrometer was designed as a means of indirectly measuring the pulse duration. [3]. Due to the difficulty of directly measuring the electron bunch length, measurement efforts must focus on indirectly extracting the pulse duration from other properties of the beam.

For this characterization purpose, an innovative spectrometer designed to measure the optical transition radiation spectrum of the electrons as they pass through a foil can provide such an indirect measurement. Getting the wavelength distribution of this radiation and then calculating its inverse Fourier transform will provide a space-time profile of the electron bunch. The optical transition radiation is mostly in the infrared range which makes it hard to work with, especially in the optics alignment and setup. In the final setup to be used at LCLS, the radiation produced from the electrons will go through a set of optical elements [3] including gold-coated mirrors, irises, gold-coated off-axis parabolic mirrors, and a prism to stop at a detector. [4]

METHODS:

The project aim is to use infrared spectroscopy to get information on the electrons from the optical transition radiation (OTR) that are produced when electrons approach and pass through a thin metal foil. [5] The ultimate spectrometer design is expected to direct the OTR beam to a prism that will disperse infrared light from about 0.6 to 40 micro meters. This covers the infrared range most relevant to the OTR emitted by the LCLS pulses. The primary setup was done on an optical bench using lasers of different wavelengths to simulate the OTR. The IR sensor is a heat detector that will be connected to a computer.

Most available spectrometers are designed to detect a particular wavelength of light or a very narrow range of wavelengths. The Spectrometer that is being designed is intended to be sensitive to a wide range of infrared light for wavelengths varying from 0.6 nanometers to 40 nanometers. The infrared spectrometer includes an optical set up that ends with an infrared detector. For this setup, a piezoelectric heat sensor was utilized. The PY-LA-S-128 from PYREOS is a line sensor array made of 128 two and a half-millimeter long sensor elements, each lined up over 14.3 millimeters. Each element of the detector constitutes a pixel and is made of a thin film of pyroelectric PZT (Lead zirconate titanate) material. PZT is a material that develops a voltage difference when experiencing a temperature, mostly infrared light change; these characteristics make this innovative sensor the perfect candidate for the project. [6] Most detectors of this type are made of a single cell element. The detector used in this experiment is sensitive a wide band in the IR spectrum. [7]

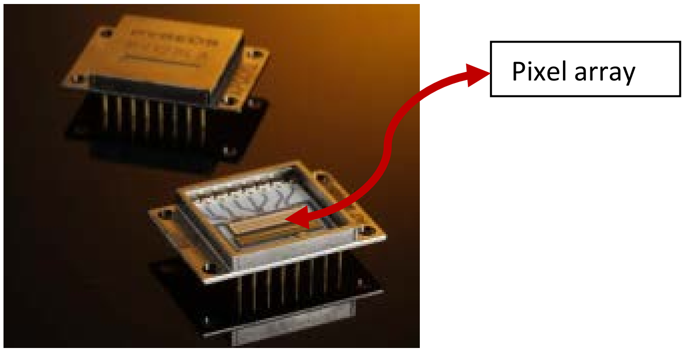


Figure 1: Line sensor.

The detector is mounted on a board on which all the electronics necessary to the data acquisition are setup. The detector was purchased with the board and the necessary software allowing us to upload collected data during the experiment. [8]

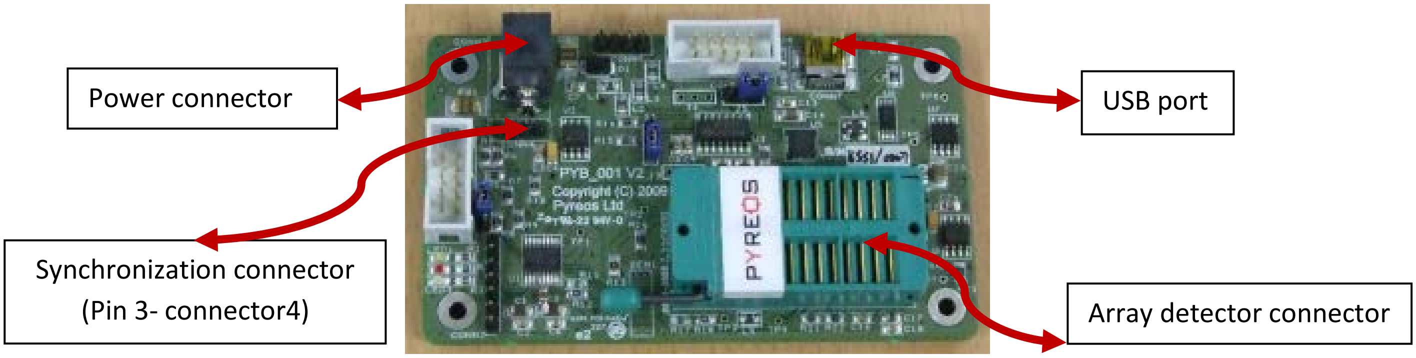


Figure 2: Sensor PCD board.

In an attempt to improve the understanding of the data collected by the detector, some preliminary data were collected only from the room temperature variations. Several modifications were then made to reduce temperature fluctuations due to room lighting. A black box was made to shield as much of the room's light as possible. On the inside of the box, an LED was inserted so most of the change in light recorded by the sensor came from the LED flashes.

The sensor board is equipped with a set of pins each outputting specific data. Pin 3 of connector 4 shown on Figure 2 out puts the pulse necessary to drive the LED. This pulse is synchronized with the sampling frequency of the detector which can be digitally modified through the software. In the initial setup, an LED was connected and driven by the board. Sets of data were recorded for different sampling frequencies values. The raw data were collected and saved in an EXCEL file. Each column represented a pixel, and each alternating row was the voltage reading of the detector during a period of the data collection time. The data reading happened during two periods; when the LED was on and when it was off

To calculate the actual change in brightness of the LED, an algorithm was created in EXCEL to take the difference between consecutive rows of the EXCEL sheet with each row representing the LED on and LED off recorded values at each pixel. The result of this calculation yielded an array of signal intensities. Later, the average signal for each pixel was taken so a graph of signal intensity vs. pixel could be plotted.

To compare how much noise the room light and other electronics was creating relative to the signal, the average root mean square (RMS) voltage of the signal was calculated from the most lit pixels. From the array of data calculated earlier, the standard deviation of the signal was calculated. The average RMS voltage and the actual signal values were calculated for each frequency and plotted to compare the sensitivity of the detector across a range of frequencies. This RMS signal was then compared to the average signal for each pixel on the same plot.

After interpreting the data collected from the initial setup, other experiments were performed. They consisted in covering parts of the detector with regular tape and observing the detectors behavior in different conditions. Figure 3 represents the latest setup used.

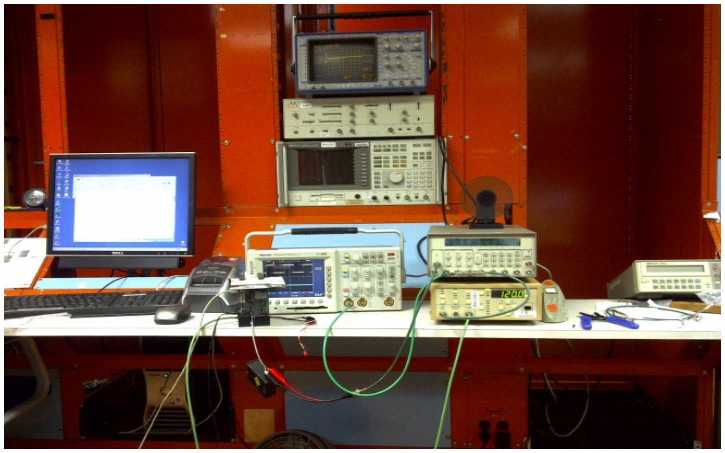


Figure 3: electronics setup to test the detector.

Later in the experiment, a single cell pyroelectric detector (Figure 4) was calibrated for potential use in the alignment of the invisible lasers that will be used in the final design of the spectrometer. The single cell detector was connected to a lock-in amplifier so the current output can be read at a specific frequency. The single cell detector was placed on the optical table after the chopping wheel, a mirror and two lenses used to focus the beam [2] Later, the single cell sensor was replaced by a diode detector that gave information about the power of the beam.

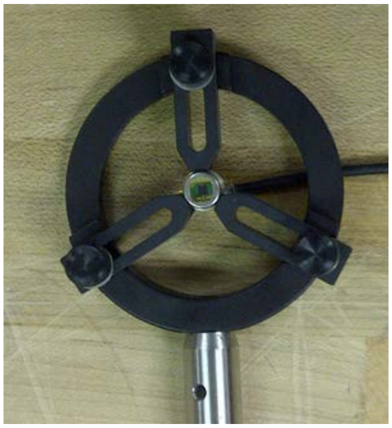


Figure 4: Pyroelectric single cell detector mounted on a holder.

RESULTS:

To calibrate the detector, its sensitivity had to be probed by a strong light source, the LED light source was run using the pulse generated from the third pin of the fourth connector of the detector. (Figure 2) It was convenient to use a regular LED light source for this purpose. When using the detector, several anomalies in the data were noticed, indicating some deficiencies in the data collection process. These deficiencies were not easily assigned to the hardware or the software of the detector. To determine the origin of the problems, several tests were performed. These tests yielding data sets that were also challenging to interpret. Figure 5 shows one signal recorded during these tests.

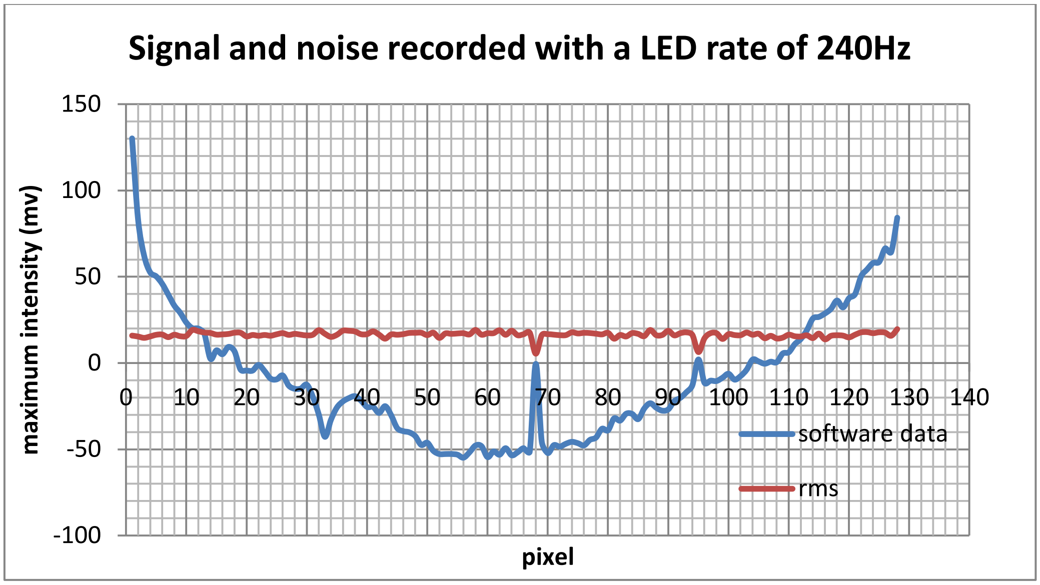


Figure 5: Data collected when running the LED at a frequency of 240Hz, the LED is placed at the center of the array detector, around pixel 60.

The data shown in Figure 5 represents the output power the sensor recorded when an LED was placed at the center of the array, directly above it. The data yielded an inverted bell-curve. This contrasted with the expectation of a Gaussian distribution. The most intensely-illuminated pixels yielded negative intensities, while the least intensely-illuminated pixels yielded positive ones. Also, the whole spectrum represented had an average intensity of -9.22 mv which is too small to represent the average intensity of the LED. These effects suggest that the detector may be AC-Coupled rather than DC-Coupled as described in the company-provided user manual.

These observations were made for each available frequency up to 240 Hz. The data were recorded for frequencies at 10Hz intervals.

The manufacturer of the detector was initially contacted in an effort to determine how the detector used raw data. The received information did not fully characterize the observed data, some of which was in direct contradiction to the supplied user manual. The EXCEL algorithm described in the previous section was developed to get a better understanding of the software functioning. Data as the ones produced by the software from the raw data collected from the board were calculated. Then, the noise and the signal were compared for the whole range of frequencies of interest. From that data, the observation that the noise was very small compared to the signal was made. When averaging the signal and noise of the 30 most lit pixels and averaging those for each frequency, we determined that the signal was 43.3 times higher than the noise. Figure 6 shows the plot of the signal compared to the noise. One can notice that the noise is very small relative to the signal.

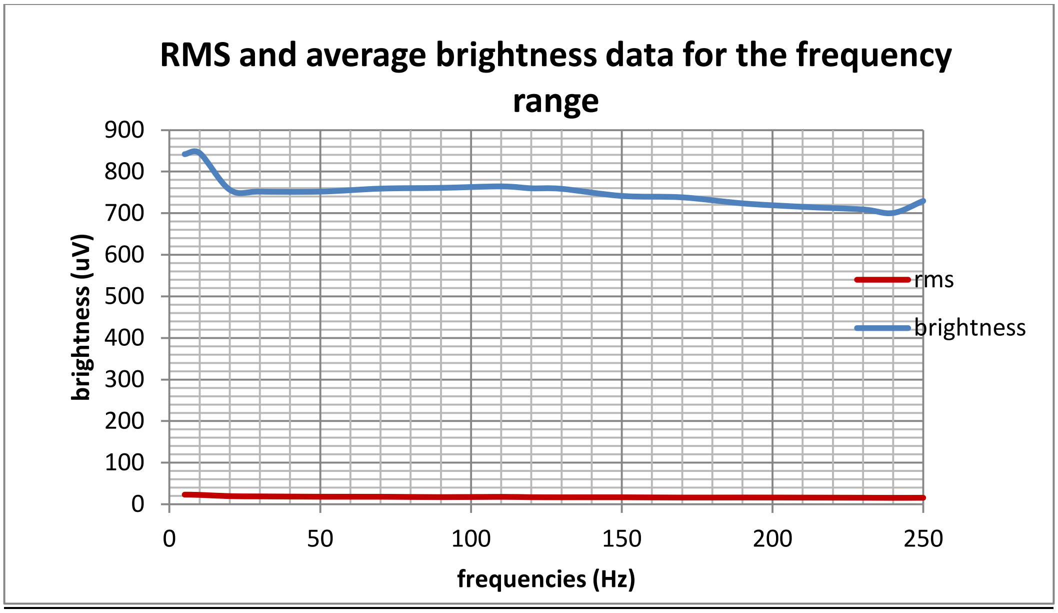


Figure 6: Signal compare to the noise for frequencies from 10Hz to 240Hz.

The software calculations appeared reasonable since the results could be reproduced using a program written in the lab. The only concern now was to know what type of information were the raw data yielding. During this part of the experiment, a dark tape was placed at the center of the array detector. The plot in Figure 7 shows that the detector presented characteristics similar to the initial data collected. These results suggested that the program may have inverted the raw data it received. However, this was inconsistent with the observed behavior of dead pixels on the detector across different tested frequencies.

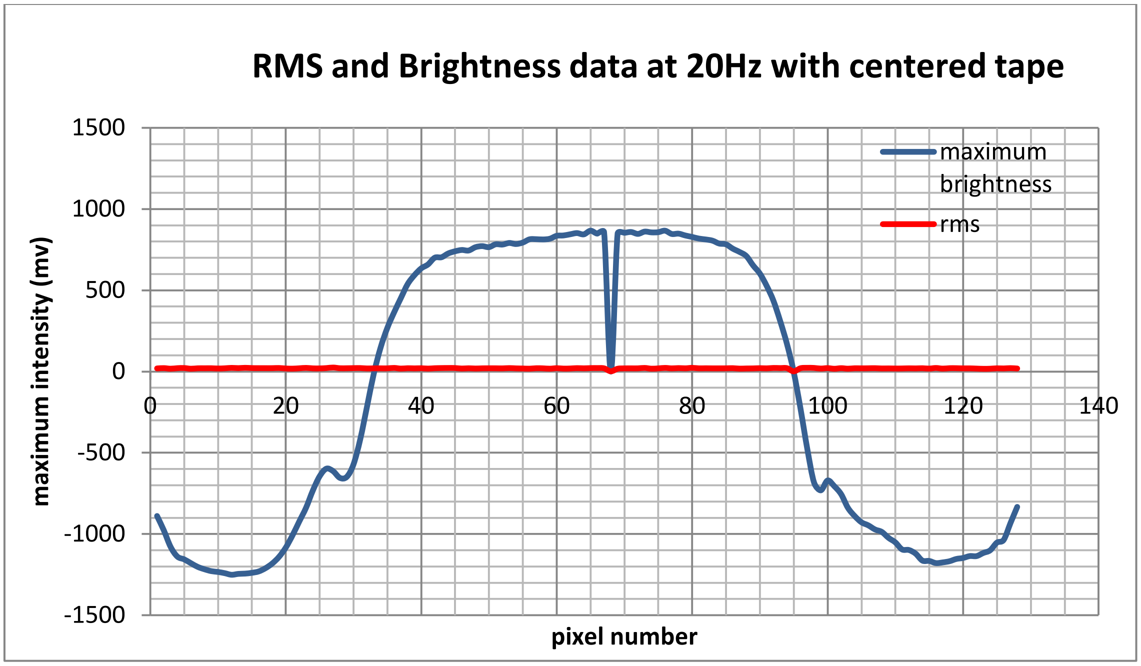


Figure 7: Data collected from the PYREOS sensor with a centered tape, the graph suggests that the software averages the overall signal to zero.

Another issue with the data was observed when the experiment was run at a sampling rate of 220Hz. (Figure 8) instead of a Gaussian curve, the expected inverse bell shaped curve was observed with the minimum brightness at the center of the detector, where the tape was placed. This completely contradicted the according to which the data were just inversed.

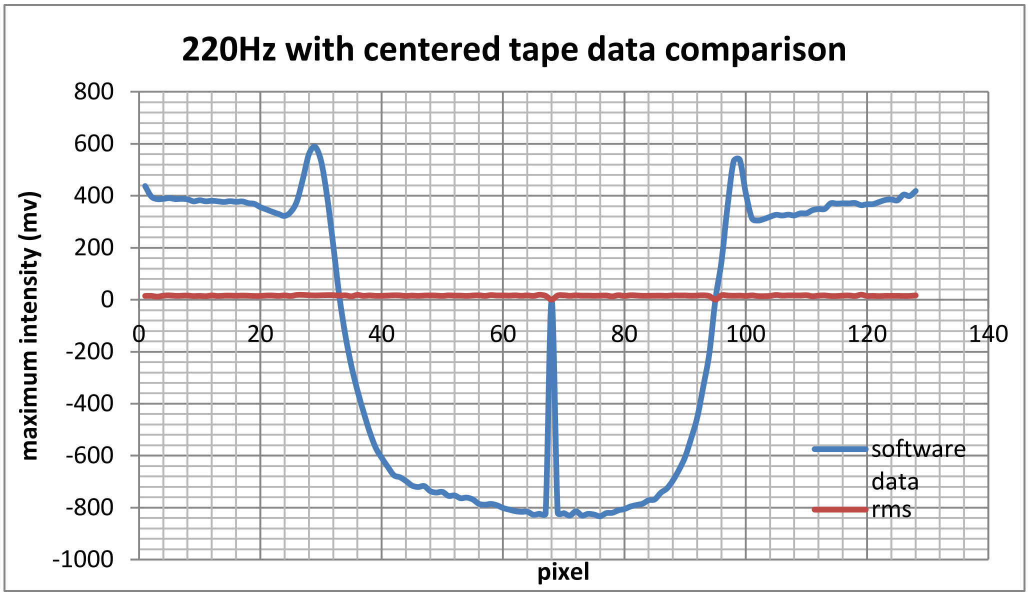


Figure 8: Data collected from the PYREOS sensor at 220 Hz sampling frequency with a centered tape, the graph suggests that the software averages the overall signal to zero.

In Figure 8, the tape is centered so the observed plot is the shape we expected while the previous data presented the opposite characteristics even though the parameters were the same throughout the experiment. This observation further suggested AC-Coupling in the detector. In this configuration, the slightest change in lightning can result in the change in the whole spectrum. It was suggested that an additional experiment be performed to determine the parameters needed to optimize the detector system. But changing the software parameters and observing the behavior of the detector in real time.

The detector allows for sample rate and data collection interval to both be adjusted. The times that are modifiable are: the ‘VVR closed time’ (VVR = Voltage VGainStage Reset) and the ‘VDR closed time’ (VDR = Voltage DGainStage Reset) those two times determine the duration of the detector not recording data: Dead time. Modifying the dead time affected the integration time which is the time during which the detector collects data.

According to the detector's manufacturer, the integration time is related to the other parameters as follow:

$$integration\ time\ (us) = \frac{1}{sample\ frequency}\ (us) - \ VDR\ time(us) - 3(us) \qquad (1)$$

Data were collected for a range of sampling frequencies and dead time. Figure 11 shows the table of the data collected, the integration time predicted by the user's manual.

Sampling frequencies (Hz)	VDR time (us)	integration time (us)
240	375	3788.666667
240	400	3763.666667
240	646	3517.666667
230	800	3544.826087
220	1000	3542.454545
210	1200	3558.904762
201	1400	3572.124378
197	1500	3573.142132
129	4000	3748.937984

Figure 8: Calculated Integration time for several frequencies to be compared With the delay time measured.

This calculated integration time seemed very high compared to what was observed simultaneously on an oscilloscope (Figure 9). The realization that the time during which the software takes data does not change while the dead time and "the integrating time" decrease was quickly made.

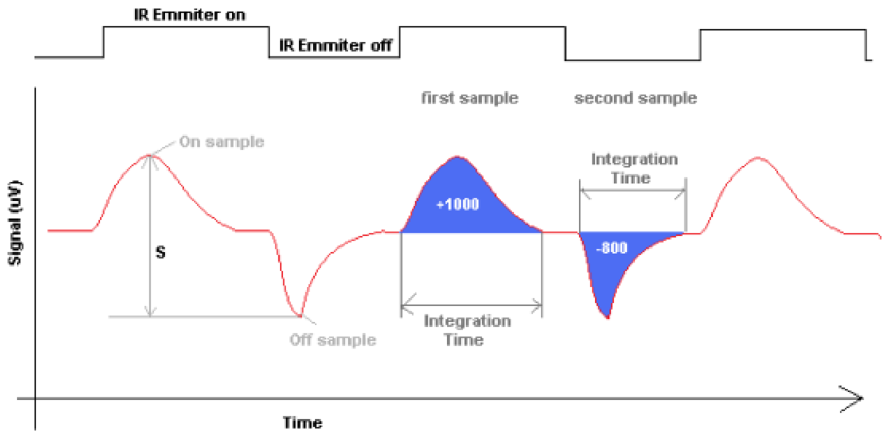


Figure 9: Generated signal and integration time. On this graph from the detector's user's manual, the integration time covers most of the signal period both the on and off phase.

While performing this experiment, it appeared that the integration time was different from the expected time. Using the previously described setup, modifications to the phase difference between the LED pulse and the sampling rate were made. A very small change in this delay time induced using the pulse generator completely nullified the signal when the calculated integration time suggested otherwise. Figure 10 shows how delaying the pulse may cancel the signal, represented by the blue areas under the red curve. This plot shows that the sampling frequency will have to be shifted by a considerable amount for the signal to disappear but continuing the delay will make the signal reappear with a potential change in sign.

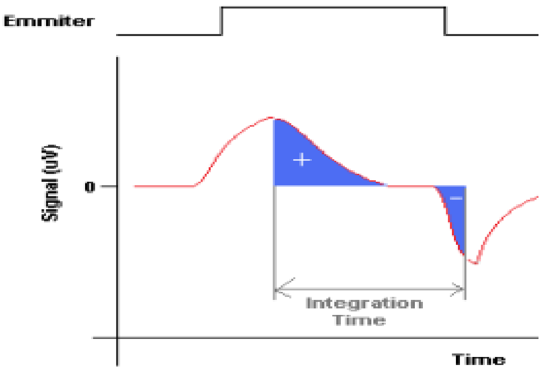


Figure 10: a delay time will result in the cancelation of the signal.

For the same parameters and dead times as the ones described above, the pulse was delayed in time, and the amount of time it took the detector to lose the signal was recorded. The delay times for several VDR values were recorded in Figure 11 for several frequencies.

Sampling frequencies (Hz)	VDR time (us)	delay (us)
240	375	0
240	400	74
240	646	258
230	800	350
220	1000	494
210	1200	634
201	1400	1270
197	1500	over
129	4000	over

Figure 11: the delay time measured for several frequencies to be compared to the calculated integration time.

Figure 12 shows a plot of both the integration time calculated and the delay time needed for the signal to disappear on the oscilloscope. One can see that the delay time is much too short compared to the integration time to cause the signal to disappear.

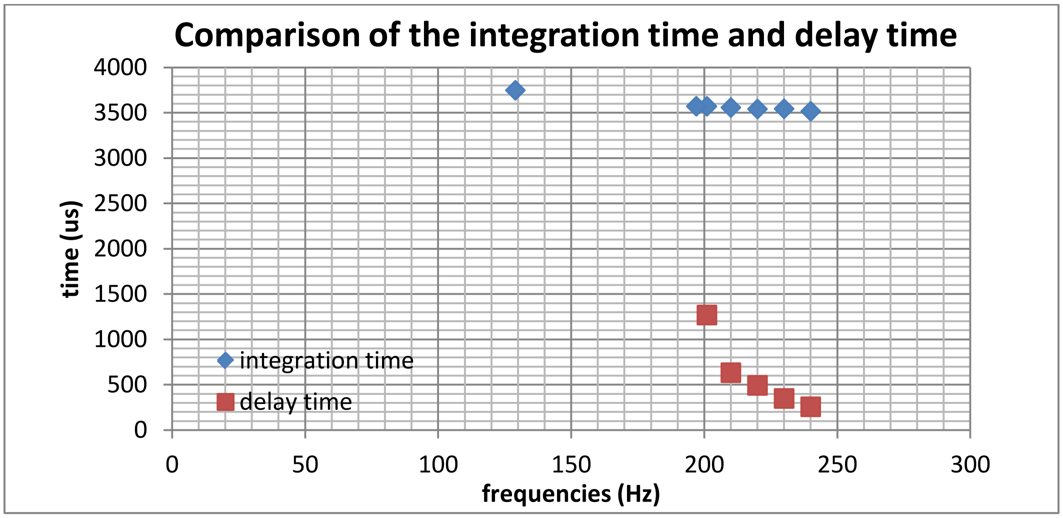


Figure 12: Comparison of the delay time and integration time. One would expect the delay time to be much larger for the signal to disappear.

Simultaneously, the sensitivity of a single cell, one-pixel pyroelectric detector was measured. A chopper wheel was driven at a frequency of 100 Hz immediately after the origin of a test laser. The power sensor, placed after the chopping wheel, a mirror, and two lenses to focus the beam, recorded a power output by the laser of 0.531 mW. The same set up was used to determine the current output by the pyroelectric detector; the locking amplifier displayed a current reading of 0.71 nA. From this we can infer that the single cell detector has sensitivity of about 1.34uA/w. at 100Hz for the Helium- Neon laser of wave length 633nm.

Discussion and Conclusion:

The fact that the data is flipped for some frequencies is believed to result from the fact that the program used by the software does not distinguish between subtracting the minimum intensity by the maximum and vice-versa. So one can simply multiply the whole set of data by a -1 and will only have to deal with the AC-Coupling.

The performed test indicated that the detector used provided us with hard to analyze data. We were able to confirm that the issue was not a software deficiency since the program used to get the reading from the raw data was written and gave results identical to the original detector software. This observation, coupled with variation in the data, suggests a hardware deficiency. The detector is thought to be an AC coupled device since the program appears to average the data it receives to 0 when the parameters are changed. Taken together, this evidence suggests a problematic user manual description of the detector integration time. Further experiment with the timing will allow the determination of the integration time for the whole range of frequencies studied.

Even though a full understanding of the data collected was not achieved, an understanding of its functions and data treatment methods was engaged. With no other candidate for its replacement, understanding the detector's operating scheme is a big step toward the completion of the Infrared spectrometer.

The integration time determination is crucial for the data acquisition phase of the project, and will be a major focus of future work with the detector. The calibration of the sensors depends greatly on the amount of signal collected by the detector, and the right timing must be determined to optimize the reliability of the data.

Acknowledgement:

I would like to thank Josef Frisch for being an outstanding mentor; Alan Fisher for his tremendous help and support throughout the project; Mark Petree and Georges Burgess for their assistance in the laboratory work; Stephen Rock for directing the program, all SULI students for the social environment with which they surrounded program. I would also like to thank the department of energy for funding the Science undergraduate Laboratory Internship. Finally, thanks to the SLAC national laboratory and Stanford University for hosting the program.

References:

[1]Josef Frisch, Tonee Smith, et al, *Accelerator physics and engineering update*, (Jan 6, 2011) Department updates

[2] C. Behrens, A. F. *Design of a single-shet prismspectroeter in the near infrared and mid infrared wavelength range for ultra-short Bunch Length diagnostics***.** DIPAC'11, Hamburg, Germany, 201

[3] Williams, K.(2011). *Optical Design of a Broadband Infrared Spectrometer for Bunch Length Measurement at the Linac Coherent Light Source*

[4] Cass, J. (2011). *Simulations and Analysis of an Infrared Prism Spectrometer for Ultra-short Bunch Length Diagnostics at the Linac Coherent Light Source.*

[5]Alan Fisher, *From LCLS to LHC: Light from Electrons, Protons, even Lead ions, (jan 6, 2011)* LHC Synchrotron Light monitor, THz studies, Alan Fisher.

[6] Lead Zirconate Titanate http://en.wikipedia.org/wiki/PZT

[7]Tim Chamberlain, (. (2009, december). *Thin Film Pyroelectric.* Retrieved july 2011, from Pyreos.com: http://www.pyreos.com/images/docs/PYDK_LAS_Demo_Kit_User_Guide__JWrelease_Jan2010.pdf

[8]Tim Chamberlain, (n.d.). *www.pyreos.com*. Retrieved july 1, 2011, from Pyreos Ltd: http://pyreos.com/products/line-sensors.html

SLAC-TN-12-021

Design & Fabrication of a High-Voltage Photovoltaic Cell

Jennifer Felder

Office of Science, Science Undergraduate Laboratory Internship (SULI)
North Carolina State University
SLAC National Accelerator Laboratory
Menlo Park, CA

12 August 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Chris Kenney at the Particle Physics and Astrophysics (PPA) Division of SLAC National Accelerator Laboratory.

Participant: ______________________________

Signature

Research Advisor: ______________________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Design & Fabrication of a High-Voltage Photovoltaic Cell. JENNIFER FELDER (North Carolina State University, Raleigh, NC 27615) CHRIS KENNEY (Particle Physics and Astrophysics (PPA) Division of SLAC National Accelerator Laboratory, Menlo Park, CA 94025)

Silicon photovoltaic (PV) cells are alternative energy sources that are important in sustainable power generation. Currently, applications of PV cells are limited by the low output voltage and somewhat low efficiency of such devices. In light of this fact, this project investigates the possibility of fabricating high-voltage PV cells on float-zone silicon wafers having output voltages ranging from 50 V to 2000 V. Three designs with different geometries of diffusion layers were simulated and compared in terms of metal coverage, recombination, built-in potential, and conduction current density. One design was then chosen and optimized to be implemented in the final device design. The results of the simulation serve as a feasibility test for the design concept and provide supportive evidence of the effectiveness of silicon PV cells as high-voltage power supplies.

INTRODUCTION

Background Information

Solar photovoltaic (PV) energy is produced by converting light energy to electrical energy. The energy of light, composed of photons, is determined by the color and therefore frequency of that light. As explained by the photoelectric effect, when the energy of that light is high enough, it excites the electrons on the surface of a material, causing them to move to a higher energy level. While in normal materials these excited electrons return to their original state relatively quickly, in photovoltaic devices these excited electrons are driven to create a current; this is known as the photovoltaic effect. The electrons do not return to their lower energy state because of an asymmetry that is inherent to this type of materials, and a potential difference is created. PV cells are developed to take advantage of this effect, and thus have a unique functionality: they conduct as diodes in a dark environment and in light they gain a charge, producing a photovoltage. [1]

One distinguishing feature of the PV device designed in this project is that it is fabricated on float-zone silicon wafers. There are two processes typically used to fabricate silicon wafers, namely the float-zone and Czochralski methods. Both processes extract a single crystal as the final product; however, the Czochralski does so using a melt from which the crystal is extracted directly, while the float-zone technique passes a liquid phase through a polycrystalline rod. Using the float-zone method provides several advantages, in particular the fact that the wafers have fewer impurities, which improves the performance of the final device. However, there also exists a tradeoff in using this method: the wafers are much more expensive than those prepared using other methods. [1]

This PV device also has a comparative advantage in terms of efficiency, as a result of the application for which it was developed. The device was designed to act as a power source for the Enriched Xenon Observatory (EXO) experiment. The thickness of the material, surrounding temperature, and wavelength of the incident light all contribute to the efficiency of a PV device. Because the environment of the experiment is known and relatively constant, the surrounding temperature and wavelength of the light to be converted to electrical energy is known. The PV cell can therefore be designed to absorb light near the peak in its absorption spectrum, increasing the efficiency of the device.

Prior Work: Silicon Solar Cells in the Past Decade

In the past decade in particular, efforts have been made to significantly increase the efficiency, and therefore utility, of PV devices. A twenty-four percent efficient silicon solar cell was achieved by altering the design of such cells in several ways. First, the silicon cell was covered in a layer of SiO_2; this reduced the recombination losses at the surface of the cell covered in oxide. Second, the surface was textured with an inverted pyramid pattern, reducing external light reflection of the top cell surface and increasing the internal light reflection of the rear cell surface. This increased the ability of the cell to absorb infrared light. Finally, the addition of a double layer antireflection coat composed of a MgF_2/ZnS compound decreased the reflective losses of the cell. [2]

Another promising advance in this area of research was made by using an alternate texturing pattern on the cell surface. Instead of the previously described inverted pyramid scheme, cells have also been fabricated with a honeycomb pattern, which consists of wells in the shape of hemispheres aligned adjacently on the cell surface. This, when combined with the previously described technique of covering the cell in oxide, increased the efficiency of monocrystalline cells from 24.0% to 24.4% and of multicrystalline cells from 18.6% to 19.8%. The symmetric pattern, while reducing the angle and height of surface features, allowed for more internal reflection; more light was absorbed by the cell, increasing its efficiency at wavelengths that are more difficult for PV devices to absorb. [3]

Motivation & Purpose

While solar cells have been fabricated with success for some time, their applications are still limited. These limiting factors include the low output voltage and somewhat low efficiency typical PV devices yield. This project aims to explore producing a single PV device having an output voltage up to 2000 V. Such a high-voltage DC power supply could substantially improve the utility of renewable energy sources such as silicon solar cells.

This high-voltage PV device was designed to be a power source for the EXO experiment. The goal of this experiment is to use the 136 isotope of Xenon to observe neutrinoless double beta decay. While it has not yet been observed, being able to observe this process would lead to two major discoveries: discerning whether the neutrino is its own antiparticle and measuring the half-life of the neutrinoless double-beta decay process, which would help in estimating the mass of the neutrino. The experiment also has importance in improving the current limit on the neutrino mass estimate. [4]

METHODS

Design Process

In beginning the mask design process, a list of features necessary for the PV cell was first developed. These features included both fabrication test features and device features. The purpose of the fabrication tests included in the design was to verify the functionality of the following aspects of the device: the conductivity of the metal, the conductivity of the N and P diffusion layers, the contact between both the metal and N diffusion layer and the metal and P diffusion layer, diode functionality, trench hopping and the field oxide etch and glass etch patterns. Each of these aspects required a separate fabrication feature, as this allows the various properties of the device to be tested independently of one another. In this way, if the final PV cell does not work properly after fabrication, it will be relatively easier to determine the failing feature and to diagnose and fix the problem.

Once the fabrication test features were designed, the design of the actual PV device was developed. High voltage PV cells were designed having dimensions of both 50 μm and 100 μm per side. Three main cell designs were developed for the mask, each having a different geometry of the diffusion layers; these designs for the 100 μm square cell are included in Figure 1. In this figure, the bright green layer is the field oxide (trench) etch pattern, the orange is the diffusion-to-metal contact, the light green is the N-type diffusion, and the pink is the P-type diffusion layer. Different devices were designed to produce various output voltages ranging from 50 V to 2000 V. To determine how many cells were needed in each module for a desired output voltage, several parameters of the silicon PV cells had to be calculated.

The intrinsic doping concentration, n_i, of the substrate, silicon, is known to be approximately 10^{10} cm^{-3}, which was included as the theoretical value in Table 3. Next, the built in potential, ϕ_j, of the individual PV cell was calculated using Equations 1 and 2, in which N_A and N_D are the active acceptor and donor dopant concentrations respectively, and V_T is the thermal voltage of a material at a specific temperature, in this case of silicon at 300 K.

$$\phi_j = \frac{kT}{q} \ln\Big(\frac{N_A N_D}{n_i^2}\Big) \tag{1}$$

$$V_T = \frac{kT}{q} = 0.0259V \tag{2}$$

Finally, the depletion width, ω_{do}, of the cell was calculated using Equation 3 and Equation 4; here ϵ_s represents the permittivity of a material, which in this case is silicon, and q is the charge of an electron.

$$\omega_{do} = \sqrt{\frac{2\epsilon_s}{q}\Big(\frac{1}{N_A} + \frac{1}{N_D}\Big)\phi_j} \tag{3}$$

$$\epsilon_s = \epsilon_o \epsilon_r \tag{4}$$

Table 1 shows the actual variable values used to calculate the parameters of the PV device. The resulting calculated parameters of the PV device, defined in the formulae above, were then calculated to produce theoretical results for those cell parameters.

The calculated built-in voltage of the PV cell was used to calculate how many cells connected in series were needed to produce various output voltages. Each module has a square geometry; therefore, the next largest perfect square to the number of cells needed was chosen to determine the size of each module. The number of cells in each PV cell is summarized in Table 2.

Different layouts of the cells were also considered, including linear, circular, and hexagonal designs. Several features were added to the PV device, including the ability to bypass a particular cell, connections for external switches, and side-entrance photo switches. Bypass diodes were necessary because of the series connections between individual cells: if one failed, the entire device may not function. Therefore, adding bypass diodes prevents the entire device from failing in the event of an individual cell failing. [1]

Simulation Process

The PV device was designed in software first to test the various parameters of the designs being considered. Three basic designs for the individual PV cells were first simulated; each had a different geometry of the diffusion layer, matching those in the mask design, with all other parameters being equal. These simulated cell designs are shown in Figure 2. In this figure, dark green layer is the support wafer, light green is the Silicon-On-Insulator (SOI) layer, red is the N-type diffusion, and blue is the P-type diffusion. The main purpose of the simulation process was to assess the performance of each design with respect to a few parameters. The first parameter that was examined was the Shockley-Read Hall (SRH) recombination. As recombination limits the current generation capability of the cell, having low recombination is desirable. Another factor of each cell design that was modeled was the built-in potential. The doping geometry that maximizes the built-in potential is more advantageous, as it would require fewer individual cells to produce a given voltage, reducing the size of the overall device. Finally, the conduction current of each design was also studied, especially in the region between the N-type and P-type diffusion areas. Using these parameters, the three main designs were compared, to determine which geometry of the dopants provided optimal performance.

After examining the simulated performance, the design shown in Figure 2(c) was chosen to be implemented in the actual PV device, and so some parameters were then adjusted to optimize the performance of that particular design, leading to a fourth cell design. All of the wafers on which the PV devices were to be fabricated had N-type substrates. For this reason, any N-type doped layers would simply function as ohmic contacts. Therefore, the area of P-type dopant was maximized. The minimum area of the N-type dopant and of the distance between the two dopant types was mainly determined by the limitation in precision of the alignment equipment used to align the mask to the wafer.

Fabrication Process

This high-voltage PV cell is being fabricated at the Stanford Nanofabrication Facility. The first step was to remove impurities from and oxidize the wafers on which the PV cells were to be fabricated. To do so, the wafers went through the following wet-etch procedure.

- 20 minutes in 9:1 $H_2SO_4 : H_2O_2$ mixture; rinsed with deionized water; dried with a spin dryer

- 10 minutes in a 4:1 $H_2SO_4 : H_2O_2$ bath to remove organic particles; deionized water rinse
- 30 seconds in a 50:1 HF mixture for oxide etching; rinsed with deionized water
- 10 minutes in 5:1:1 H_2O:H_2O_2:HCl bath to remove metal ions; rinsed with deionized water; cycle through spin dryer
- Oxidation process: WET 850 Tylan process [5]; variable time set to 20 minutes

With the oxidation process complete, there are still several steps that need to be completed to finish fabricating the device. Thus far, only the first step of the fabrication process has been completed; device fabrication and testing will be continued in future work.

RESULTS

Three parameters of the individual PV cell were calculated, as previously described; these calculated values comprise the theoretical results of those cell parameters. These same properties were measured using the simulation software, to assess the validity of the calculated values and provide a more extensive theoretical basis on which to judge the final device. These results are summarized in Table 3. As shown in the table, the calculated and simulated values agree fairly closely, within a margin of error. This adds credibility both to the calculations and simulation results. As measuring ω_{do} directly in the simulation was only possible by estimating a value off of a plot, Equation 3 was used with the value of ϕ_j measured from the simulation to get a more precise value for comparison.

In addition to the three basic parameters of each cell that were calculated, other factors of a PV cell are critical to the device performance, yet cannot be easily or directly calculated by hand. For this type of properties, simulation is imperative to comparing different designs. One such factor is the SRH recombination of the simulated cells. An ideal PV cell would have minimal recombination, which limits the output current of the cell. The SRH recombination plots for the various designs are included in Figure 3.

Another property of PV cells that was simulated was the conduction current density. A functional cell needs conduction between contacts on the N and P diffusion layers; therefore the conduction current density is expected to be higher between contacts and minimal elsewhere. It is also desirable to have higher conduction current on the surface of the cell, as it is that portion of the device that is designed to be conductive. The plots of the conduction current density for each design are shown in Figure 5.

To simulate the operation of the cell as a photovoltaic device, several parameters had to be specified. To effectively simulate the environment in which this device could potentially operate, the optical generation of the device was set to 10^{18} $cm^{-3}s^{-1}$. The diffusion implants in the actual PV cell are approximately 2 μm deep; assuming a Gaussian distribution with a peak value of 10^{19} cm^{-3} of the diffusion layer, the standard deviation was adjusted to achieve this in the simulation. Finally, to bias the cell to 0 V, a resistive contact with a value of 10^5 Ω was connected to the diffusion contacts.

DISCUSSION

Some properties of the individual cells are easier to compare in the mask designs, while others are more evident in simulation. When designing the mask for the PV cells, enough metal has to be placed on the cell to provide adequate contact to the diffusion layers, to create a strong electrical contact. However, the area of the cell covered by metal should be minimal, as metal prevents light from being absorbed by the cell. Comparing the layout designs in Figure 1, it is clear that the design in Figure 1(c) has the best performance in minimizing the area of metal coverage.

Several trends emerge from examining the simulation results of the four designs. In all of the designs the SRH recombination is higher in the area between the two diffusions. As recombination limits the output of a PV cell, the area with high recombination should be minimized. The optimized design has the least area between diffusions, and therefore the most desirable recombination profile, as shown in Figure 3(c). In contrast, the design in Figure 3(a) shows the greatest amount of recombination in its profile.

The designs were also compared in terms of built-in potential. Comparing the built-in potential simulation plots for each device in Figure 4, the potential does not vary more than 0.01 V between designs. This is an important factor when comparing cells, as the built-in potential determines how many cells are needed to produce a given output voltage, which determines the size of the overall device. However, in this case, the built-in potential for all designs compared is the nearly same, as expected because the built-in potential is only dependent upon the doping concentrations and intrinsic concentration of the substrate, which is constant for all devices. This is therefore not a determining factor in choosing a cell design.

Finally, the conduction current density was compared among designs. A higher conduction current is expected between diffusion contacts: the current generated in the cell must flow between the contacts to be an effective power source. As the area of the designs being compared is constant, the design with the best conduction current density profile can be assumed to have the most ideal conduction current profile as well. The design in Figure 5(c) has the most favorable profile; in contrast, Figure 5(a) has the least ideal conduction current properties.

CONCLUSION

Based on the comparisons drawn among the different cell designs, a design was chosen to be implemented in the PV device. Because it had minimal metal coverage, an acceptable conduction current density profile, and the lowest recombination rate among the designs, the design shown in Figure 2(d) was chosen to be implemented in the mask design.

There are several important implications that arise from this project. The design and particularly the simulation portion of this project serve as a feasibility test for the concept of high-voltage PV cells. While the final fabrication of the high-voltage PV device has not yet been completed, the preliminary results of the simulations support the theory of such a device having practical functionality. Once the device is fabricated and tested, further conclusions as to the possibility of implementing PV devices in a wider range of applications will be assessed.

There are many applications of this technology and future work to be done on this specific project. The first step in continuing this research is to finish fabricating the PV device. After fabrication, the device will be tested to assess its functionality. Its effectiveness at producing the desired output voltage will be determined, to conclude the feasibility of such a device to be used as a high-voltage DC power supply. In addition to having an impact on scientific experiments such as EXO and other work at SLAC, society could also benefit if such devices were implemented on a large scale.

ACKNOWLEDGMENTS

This work was funded by the United States Department of Energy and the SLAC National Accelerator Laboratory Science Undergraduate Laboratory Internship (SULI) program. In addition, I would like to thank Chris Kenney, Astrid Tomada, Jasmine Hasi, and Julie Segal for the opportunity to conduct this research, for training me in the nanofabrication facility,

and for helping in many ways throughout the project. Finally, I would like to acknowledge the contribution of Sentaurus Synopsis simulation software to this project.

REFERENCES

[1] J. Nelson, *The Physics of Solar Cells.* London: Imperial College Press, 2003. 7 July 2011.

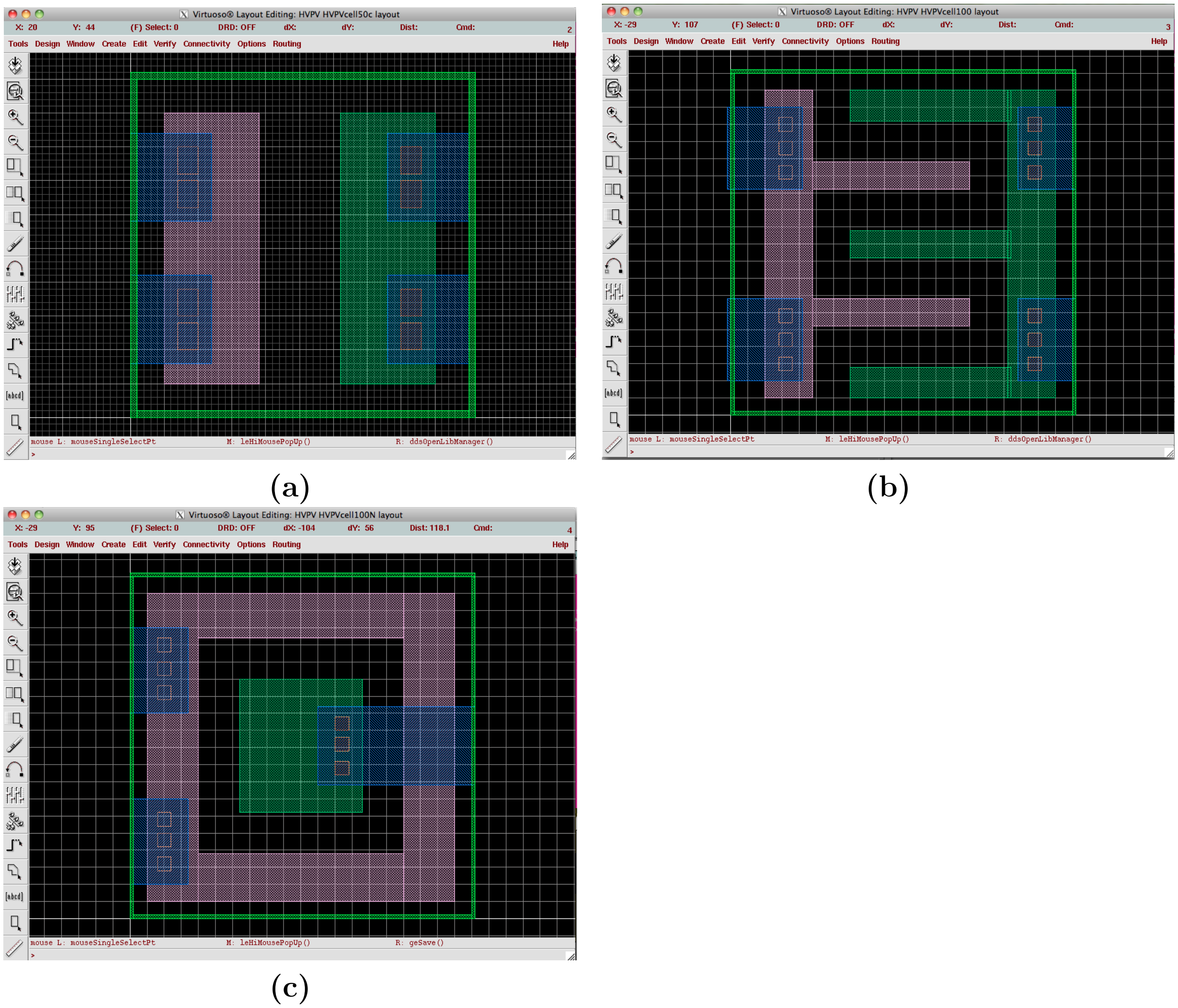
[2] J. Zhao, A. Wang, P. Altermatt, and M. A. Green, "Twenty-four percent efficient silicon solar cells with double layer reflection coatings and reduced resistance loss," *Applied Physics Letters*, vol. 66, pp. 3636-3638, 26 June 1995.

[3] J. Zhao, A. Wang, M. A. Green, and F. Ferrazza, "19.8% efficient "honeycomb" textured multicrystalline and 24.4% monocrystalline silicon colar cells," *Applied Physics Letters*, vol. 73, num. 14, pp. 1991-1993, 5 October 1998.

[4] Enriched Xenon Observatory, "EXO: About the Experiment," 20 August 2010, http://www-project.slac.stanford.edu/exo/ .

[5] Stanford Nanofabrication Facility, "WET850 Tylan Recipe," http://snf.stanford.edu/Equipment/tylanrecipes/wet/WET850 .

FIGURES

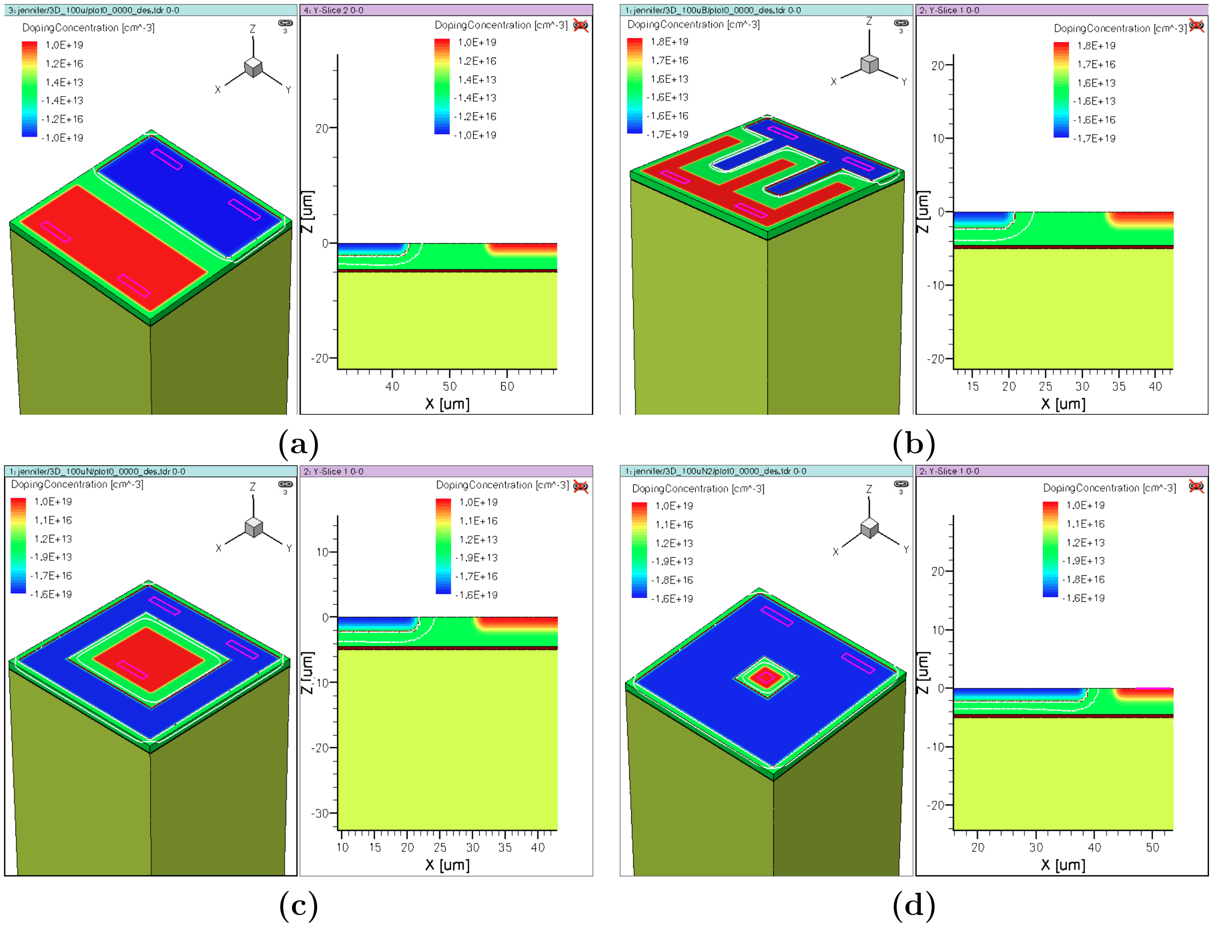

Virtuoso® Layout Editing: HVPV HVPVcell50c layout
X: 20 Y: 44 (F) Select: 0 DRD: OFF dX: dY: Dist: Cmd: 2
Tools Design Window Create Edit Verify Connectivity Options Routing Help
mouse L: mouseSingleSelectPt M: leHiMousePopUp() R: ddsOpenLibManager()
>
Virtuoso® Layout Editing: HVPV HVPVcell100 layout
X: -29 Y: 107 (F) Select: 0 DRD: OFF dX: dY: Dist: Cmd: 3
Tools Design Window Create Edit Verify Connectivity Options Routing Help
mouse L: mouseSingleSelectPt M: leHiMousePopUp() R: ddsOpenLibManager()
>
Virtuoso® Layout Editing: HVPV HVPVcell100N layout
X: -29 Y: 95 (F) Select: 0 DRD: OFF dX: -104 dY: 56 Dist: 118.1 Cmd: 4
Tools Design Window Create Edit Verify Connectivity Options Routing Help
mouse L: mouseSingleSelectPt M: leHiMousePopUp() R: geSave()
>

(a)

(b)

(c)

Figure 1: Mask Design of Individual PV Cells for (a) Design 1, (b) Design 2, and (c) Design 3



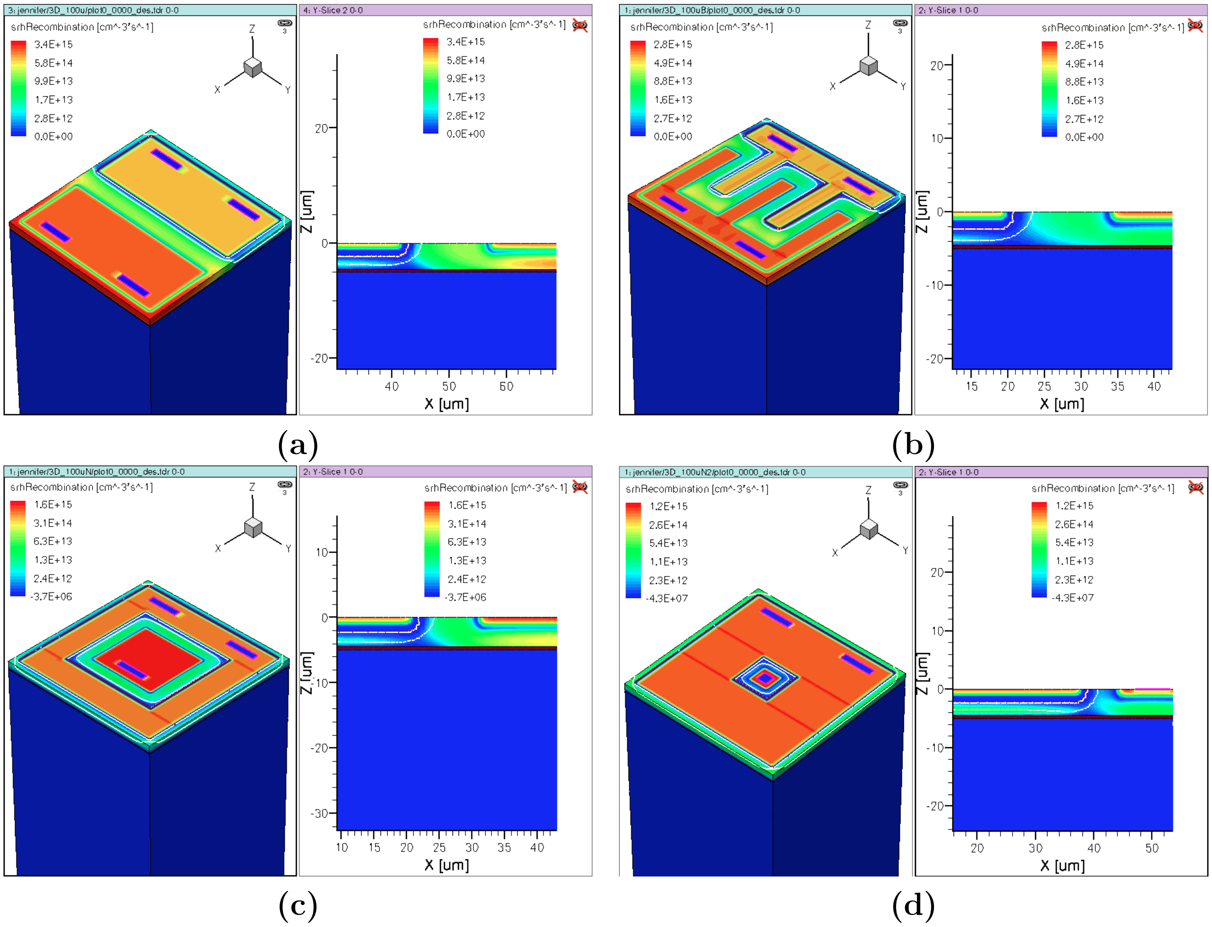
(a)

(b)

(c)

(d)

Figure 2: Doping Profile Simulation Plots for (a) Design 1, (b) Design 2, and (c) Design 3



(a)

(b)

(c)

(d)

Figure 3: SRH Recombination Simulation Plots for (a) Design 1, (b) Design 2, and (c) Design 3

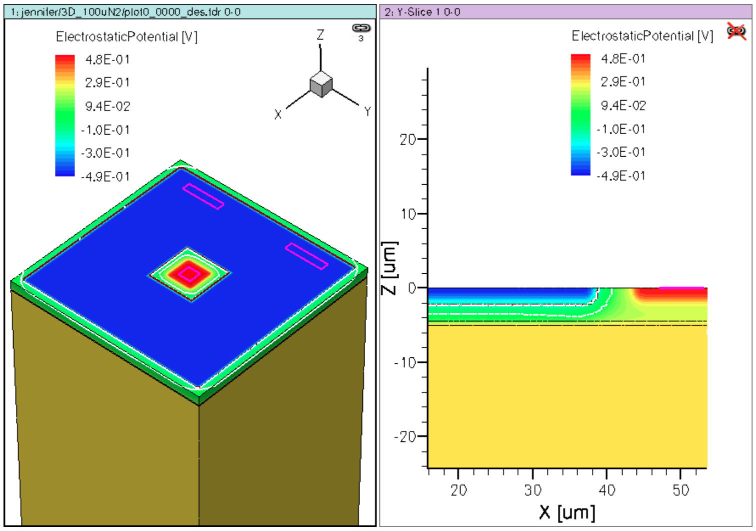
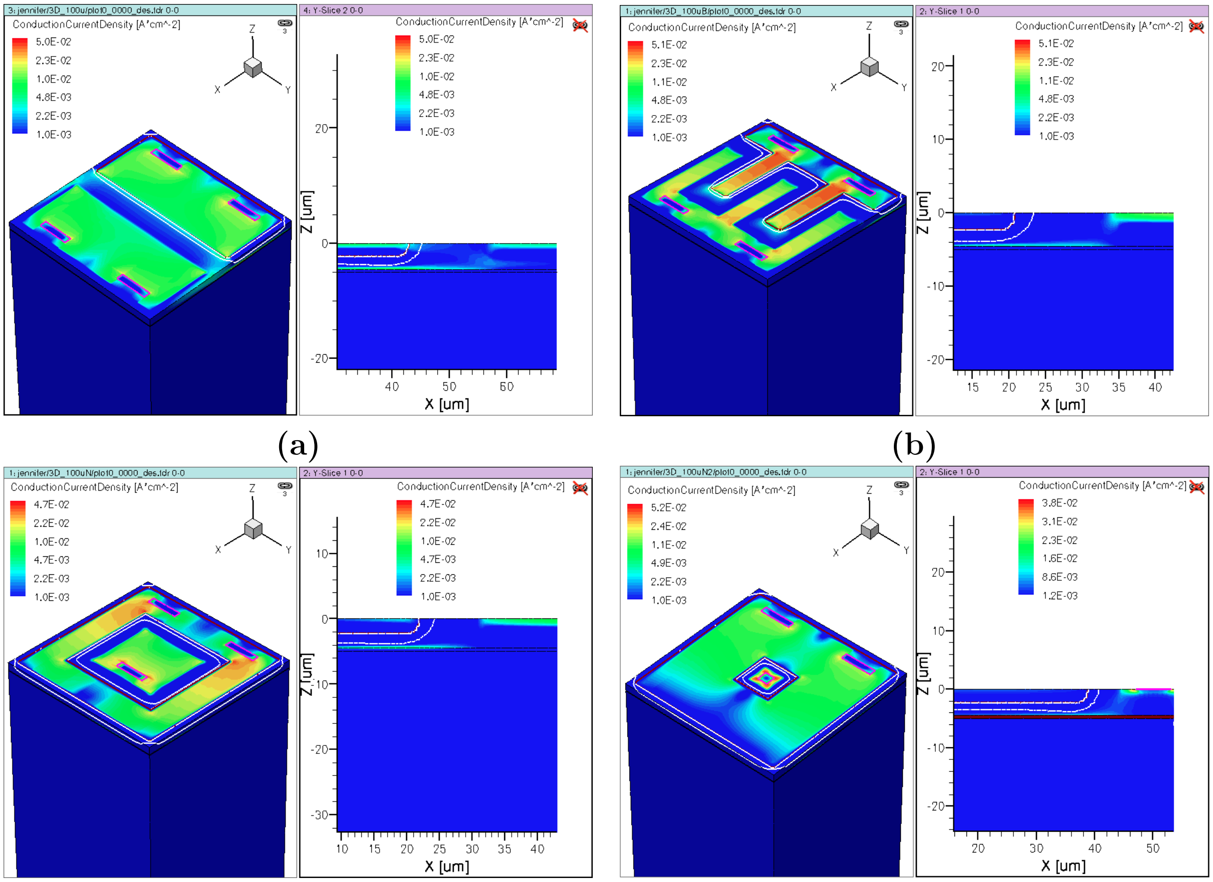


Figure 4: Electrostatic Potential Plots of Design 4, Optimized Design 3



(a) **(b)**

(c) **(d)**

Figure 5: Conduction Current Density Simulation Plots for (a) Design 1, (b) Design 2, and (c) Design 3

TABLES

Parameter	Value	Units
N_A	$2 \cdot 10^{15}$ (10^{19})	cm^{-2} (cm^{-3})
N_D	$2 \cdot 10^{15}$ (10^{19})	cm^{-2} (cm^{-3})
μ_n	1300	cm^2/V·cm
μ_p	440	cm^2/V·cm
ρ	10	kΩ · cm
ϵ_r	11.7	
ϵ_o	$8.85 \cdot 10^{-14}$	$\frac{F}{cm}$
q	$1.609 \cdot 10^{-19}$	C

Table 1: Values of Various Parameters Used in the Calculations In Equations 1-6

Output Voltage (V)	Cells Needed	Sq. Matrix	Actual Number Cells
50	52	8 x 8	64
100	104	11 x 11	121
200	207	15 x 15	225
500	516	23 x 23	529
1000	1,031	33 x 33	1,089
2000	2,062	46 x 46	2,116

Table 2: Summary of Cells Needed in PV Cell Modules to Produce Given Output Voltages

Parameter	Theoretical Value	Simulated Value	Units
n_i	$1.01 \cdot 10^{10}$	$1.5 \cdot 10^{10}$	cm^{-3}
ω_{do}	0.0166	0.01580	μm
ϕ_j	1.07	0.970	V

Table 3: Values of Calculated & Simulated Parameters of PV Cell

SLAC-TN-12-023

Dark Photon Search at *BABAR*

Ross N. Greenwood

Office of Science, Science Undergraduate Laboratory Internship (SULI)
Massachusetts Institute of Technology
Stanford National Accelerator Laboratory
Stanford, CA

August 12, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Arafat Gabareen Mokhtar at the Stanford National Accelerator Laboratory.

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Dark Photon Search at *BABAR*. ROSS N. GREENWOOD (Massachusetts Institute of Technology, Cambridge, MA 02139) ARAFAT GABAREEN MOKHTAR (Stanford National Accelerator Laboratory, Stanford, CA 94025)

Presented is the current progress of a search for the signature of a dark photon or new particle using the BaBar data set. We search for the processes $e^+e^- \to \gamma_{ISR}A', A' \to e^+e^-$ and $e^+e^- \to \gamma_{ISR}\gamma, \gamma \to A', A' \to e^+e^-$, where γ_{ISR} is an initial state radiated photon of energy $E_\gamma >= 1$ GeV. Twenty-five sets of Monte Carlo, simulating e^+e^- collisions at an energy of 10.58 GeV, were produced with different values of the A' mass ranging from 100 MeV to 9.5 GeV. The mass resolution is calculated based on Monte Carlo simulations. We implement ROOT's Toolkit for Multivariate Analysis (TMVA), a machine learning tool that allows us to evaluate the signal character of events based on many of discriminating variables. TMVA training is conducted with samples of Monte Carlo as signal and a small portion of Run 6 as background. The multivariate analysis produces additional cuts to separate signal and background. The signal efficiency and sensitivity are calculated. The analysis will move forward to fit the background and scan the residuals for the narrow resonance peak of a new particle.

INTRODUCTION

Based on observational evidence from astronomical surveys of the internal motions of distant galaxy clusters, it is believed that electromagnetically interacting matter contributes to only about 4% of the mass-energy of the universe. Current theory suggests that the remainder is made up of two additional components in a ratio of approximately three to one, respectively: dark energy, which is responsible for accelerating the metric expansion of space, and dark matter - nonbaryonic matter particles with nonzero mass (inferred from its gravitational effects on nearby visible matter) that may experience the weak or strong nuclear forces, or yet undiscovered interactions.

One candidate constituent of dark matter is a theoretical dark photon (symbol: A'). The dark photon is a massive, weakly interacting boson whose existence is postulated in models of supersymmetry [2]. It may serve as the force-carrying gauge boson for particle interactions unique to the dark sector, or even provide a means for interaction between dark matter and Standard Model particles. Theory suggests the possibility of dark photon production in electron-positron annihilation events. By the process $e^+e^- \to A' \to f^+f^-$, the electron-positron pair may annihilate directly to a dark photon, which would decay to a fermion pair. Alternatively, the e^+e^- pair annihilates to a virtual Standard Model photon, which produces a dark photon via an intermediate virtual fermion loop [2].

The BaBar experiment (1999 - 2008) recorded asymmetric collisions of electrons and positrons supplied by the linear accelerator at the SLAC National Accelerator Laboratory. BaBar focused on the study of CP violation by production of B mesons, hadrons containing a quark and an anti-quark of which one has bottom quark flavor (b or $\bar{b}$). B mesons can be produced during the decay of the $\Upsilon(4S)$ ($b\bar{b}$) meson. During Runs 1-6, collisions took place at center-of-mass energies corresponding to the masses of $\Upsilon(4S)$. The collision energy was tuned to the masses of the $\Upsilon(2S)$ and $\Upsilon(3S)$.

We searched the *BABAR* data set for evidence of a dark photon or a new particle, probing a mass range from 100 MeV to 9.5 GeV. Events featuring e^+e^- annihilation preceded by emission of an energetic photon ($E_\gamma >= 1\,\mathrm{GeV}$) from either particle were selected. By requiring initial state radiation (ISR), we access a range of values for the relativistic mass m_{ee} of the e^+e^- system at the collision event. The main background process (those with the same final state as the sample) are radiative Bhabha scattering ($e^+e^- \rightarrow \gamma_{ISR}e^+e^-$) and two-photon production ($e^+e^- \rightarrow \gamma_{ISR}e^+e^-\gamma\gamma, \gamma\gamma \rightarrow e^+e^-$) where one of the e^+e^- pairs goes undetected.

If a dark photon exists and we implement the appropriate selection criteria, the particle will manifest in the data as a narrow resonance peak in the candidate event count plotted against the e^+e^- invariant mass. In this analysis, the significance is optimized as

$$\frac{S}{\sqrt{B}+\frac{3}{2}}$$

, where S is the number of signal events and B is the number of background events. We begin with a blind analysis using a small portion of the available data, composed of events with a dielectron mass corresponding to the $\Upsilon(4S)$ resonance ($10.58\,\mathrm{MeV}/c^2$). After perfecting our optimization techniques, we will unblind the analysis to include all compatible data from Runs 1-7.

MATERIALS AND METHODS

Twenty-five Monte Carlo simulation production (SP) modes with discrete dark photon generated mass values ranging from 100 MeV to 9.5 GeV were produced. Bhabha scattering and two-photon production were also simulated. Two terms are used to describe the mass with respect to Monte Carlo simulations: the *generated mass* is a parameter submitted to the event generator as the true mass of the simulated dark photon with a resonance of zero width; the *reconstructed mass* is the invariant mass of the e^+e^- system calculated from the kinematics of the generated event and simulated to account for detector resolution.

We began with a data set containing approximately 2.3% of Run 6, which contains events with pre-ISR m_{ee} tuned to the $\Upsilon(4S)$ mass of 10.58 GeV. We require at least two e tracks and one photon with energy greater than 1 GeV in the final state. Up to ten additional photons are allowed with energies less than 100 MeV. If multiple candidate pairs of e tracks are available, the pair with the best vertex fit (best χ^2) is selected, such that there is only one candidate e^+e^- pair per event. Our analysis is focused on dark photon events with an e^+e^- final state, though $\mu^+\mu^-$, $\pi^+\pi^-$, and $e^\pm\mu^\mp$ final states are also considered for the purpose of understanding the background.

The analysis thus far can be summarized in the following steps:

1. Resolution $\sigma(m)$

   Quantify the spread of the reconstructed dark photon mass produced by a Monte Carlo simulation.

2. Multivariate Analysis (MVA)

   ROOT's Toolkit for Multivariate Analysis (TMVA) package is a machine learning tool that allows us to evaluate the most likely character of events (signal vs. background) based on many discriminating variables [1]. Train a TMVA program to recognize signal events by comparing sample signal events generated in Monte Carlo simulations to the sample background, a small portion of Run 6.

3. Signal efficiency $\epsilon(m)$

   Determine the rate of successful reconstruction of events generated in Monte Carlo.

4. Sensitivity

   Predict the potential discovery.

Resolution

The reconstructed mass histogram for each Monte Carlo mode was fit to the sum of two Crystal Ball functions. The widths and means of the two functions are set equal, but the exponential tails are on opposite sides and otherwise free to vary. The histograms and fits are found in Figure 1. The width is considered as the resolution at the mean value of the mass. The twenty-five resolution values were plotted with the generated mass of each Monte Carlo mode and fit to a third order polynomial.

Multivariate Analysis

We used TMVA to characterize events based on nine variables:

[Unless otherwise stated, kinematic calculations were performed in the center-of-momentum frame of the e^+e^- *system after initial state radiation.]*

- $EXUnc$: Unconstrained total energy of $\gamma_{ISR}e^+e^-$ calculated in the laboratory frame
- PTX : Transverse momentum of $\gamma_{ISR}e^+e^-$
- $XvtxProb$: Vertex probability
- $e2cms$: Energy of the 2nd highest energy photon
- apl : Positive cosine angle between γ_{ISR} and the plane defined by the lepton tracks
- $mmiss2$: Squared missing mass
- $costhmiss$: Cosine polar angle of the missing momentum vector
- $thglfCM$: Angle between γ_{ISR} and the fastest e track
- $thfCM$: Polar angle of the e^- track

The TMVA classification methods employed in this analysis are Boosted Decision Tree (BDT), Likelihood, and Fisher. An assessment of each method's performance based on signal efficiency and background reduction, sampled in Figure 3, identified BDT as the most successful classifier. We ran a separate TMVA analysis for each Monte Carlo mode, and interpolated the results to produce a continous function of the BDT output parameter mva_{BDT} for the full range of m_{ee}.

To determine the most appropriate cut on mva_{BDT}, we used the well understood J/ψ resonance peak as a reference. We varied the mimimum mva_{BDT} value in a cut of the Run 6 background and fit the region of the J/ψ to the stum of a Gaussian and a linear function using the RooFit package in ROOT. We then calculated the signal-to-background ratio of the J/ψ peak. The data and fits are shown in Figure 4 and the results summarized in Figure 5. We found the cut $mva_{BDT} > 0.2$ to produce the best signal-to-background ratio while maintaining an acceptable level of signal retention.

Signal Efficiency

The efficiency was calculated for each Monte Carlo simulation as the ratio of the number of reconstructed events to the number of generated events. The results are plotted in Figure 7.

$$\epsilon(m_{ee}) = \frac{N_{rec}}{N_{gen}}. \tag{1}$$

Sensitivity

The sensitivity was calculated as the expected error on the branching ratio σ_{BR}, the fraction of the total e^+e^- annihilation cross section that includes dark photon production, corresponding to a 5σ discovery.

$$\sigma_{BR} = \frac{5\sqrt{\alpha \cdot N_{bkg}(m)}}{\epsilon(m) \cdot N_{ee}} \tag{2}$$

α is the ratio of the number of events in Runs 1-7 to the number of events in Run 6. N_{bkg} is the number of background events with reconstructed masses close to m_{ee}, within $1.5\sigma(m_{ee})$, where $\sigma(m_{ee})$ is the resolution. N_{ee} is the total number of electron-positron collision events in Runs 1-7.

DISCUSSION AND OUTLOOK

In summary, we have made significant progress in a search for a dark photon signature using the *BABAR* data set, calculating the simulated resolution, establishing cuts based on multivariate analysis, determining our signal efficiency, and predicting the sensitivity. However, the analysis is still underway, and has yet to produce conclusive results. The next step is to fit the reconstructed m_{ee} distribution. We will first fit the background, including recognized features (i.e. J/ψ) and assuming that no other signal is present. We will then iterate through the range of possible reconstructed A' masses, attempting to fit to a double sided Crystal Ball function at each mass value on top of the background. We aim to produce a measure of the likelihood that a discovery resonance peak sits atop the background at that mass value. Finally, we will unblind the analysis to include all available *BABAR* data from Runs 1-7.

ACKNOWLEDGMENTS

I would like to thank my project mentor Arafat Gabareen Mokhtar for his genuine effort and guidance during my two months at SLAC. I would also like to thank the SULI staff for their hard work in making the program an unforgettable success: Stephen Rock, Maria Mastrokyriakos, and Anita Piercey. Finally, my fellow SULI students have greatly enriched the other aspects of my summer life, and I greatly value their friendship.

REFERENCES

[1] M. Backes, et al. *Toolkit for Multivariate Analysis in ROOT, User's Guide*. European Organization for Nuclear Research (CERN). (2009).

[2] R. Essig, et al. "Probing dark forces and hidden light sectors at low-energy e^+e^- colliders." SLAC National Accelerator Laboratory. (2009).

FIGURES

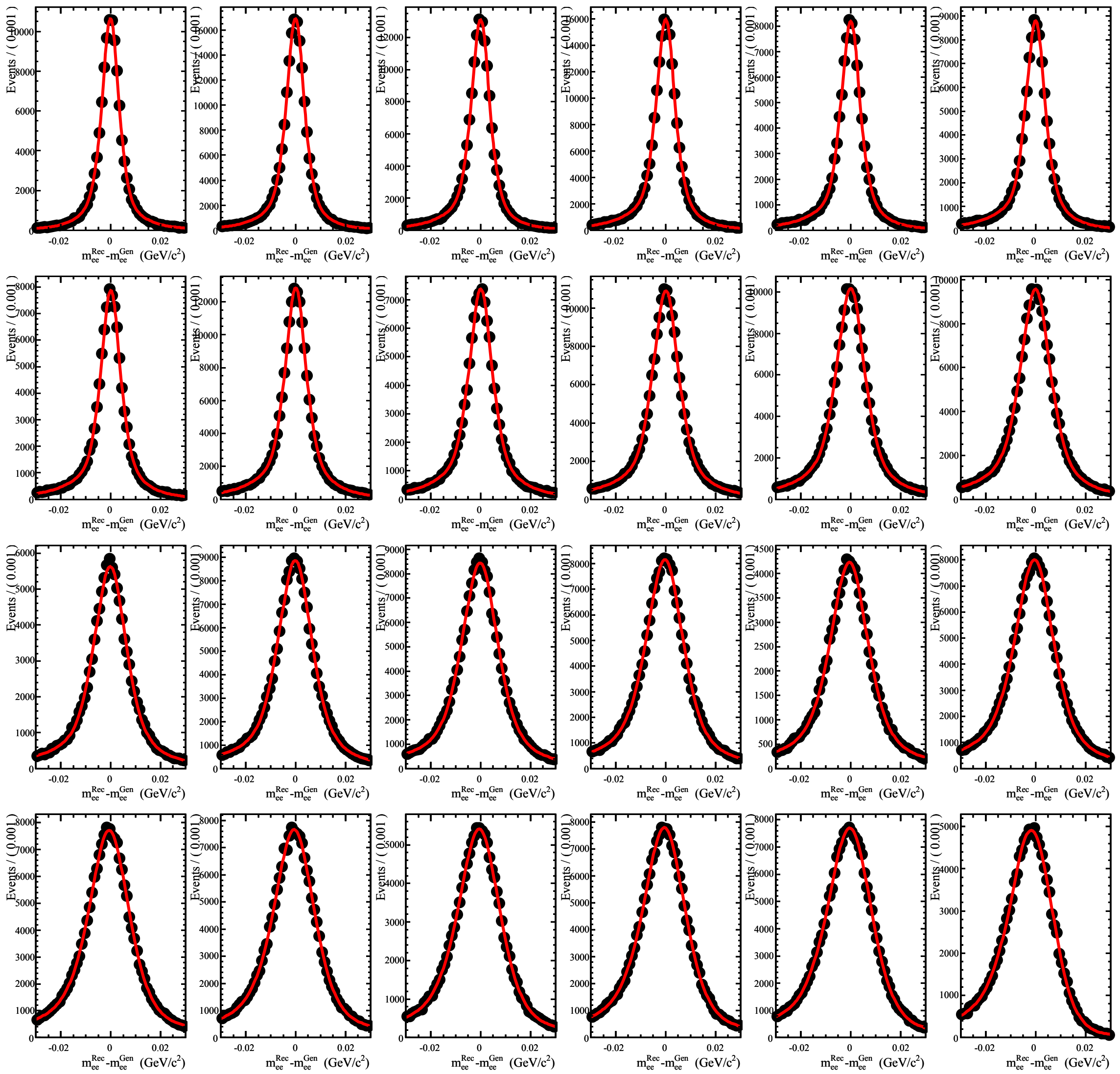


Figure 1: The reconstructed mass histograms and double Crystal Ball fits for 24 Monte Carlo modes.

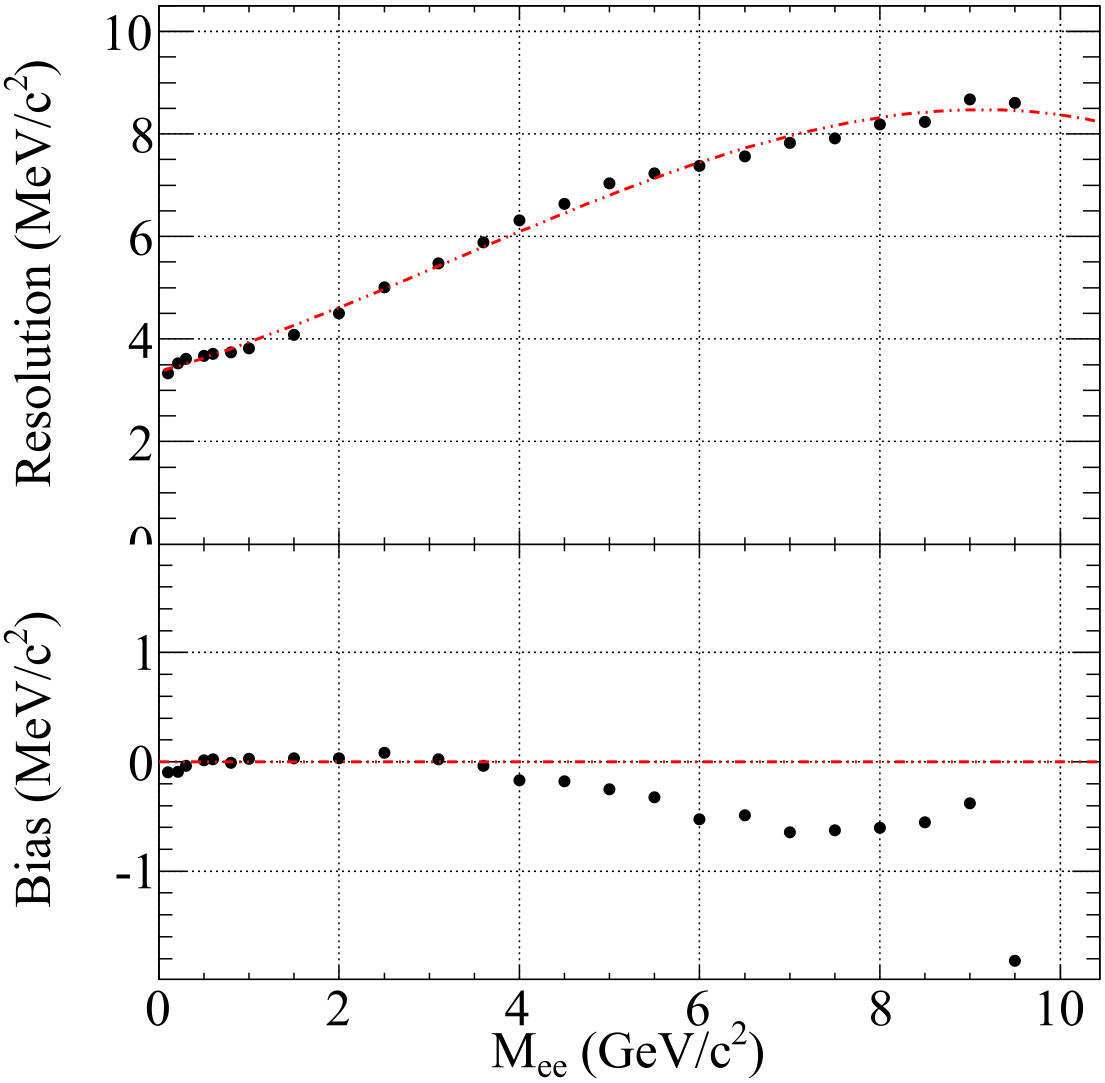


Figure 2: The resolution (top) and bias (displacement of the mean of the fit function) (bottom), plotted against m_{ee}.

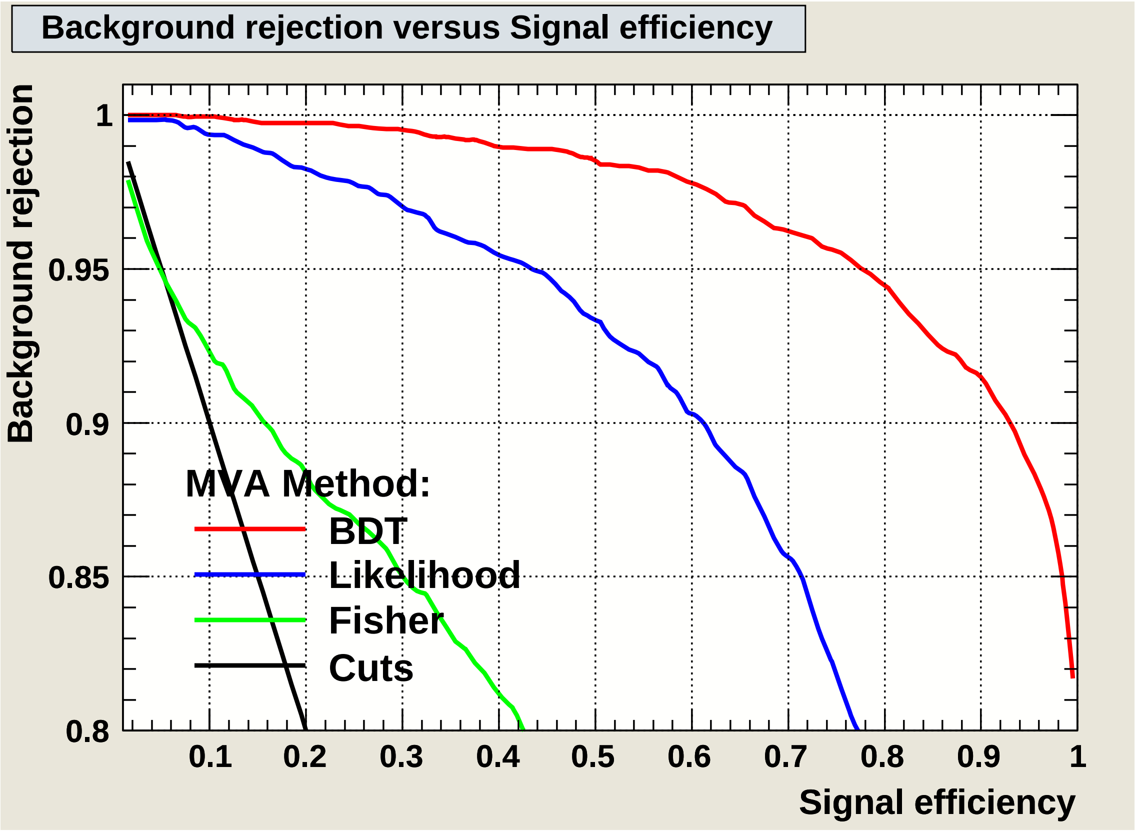


Figure 3: Background rejection plotted against signal efficiency for various MVA-based cuts.

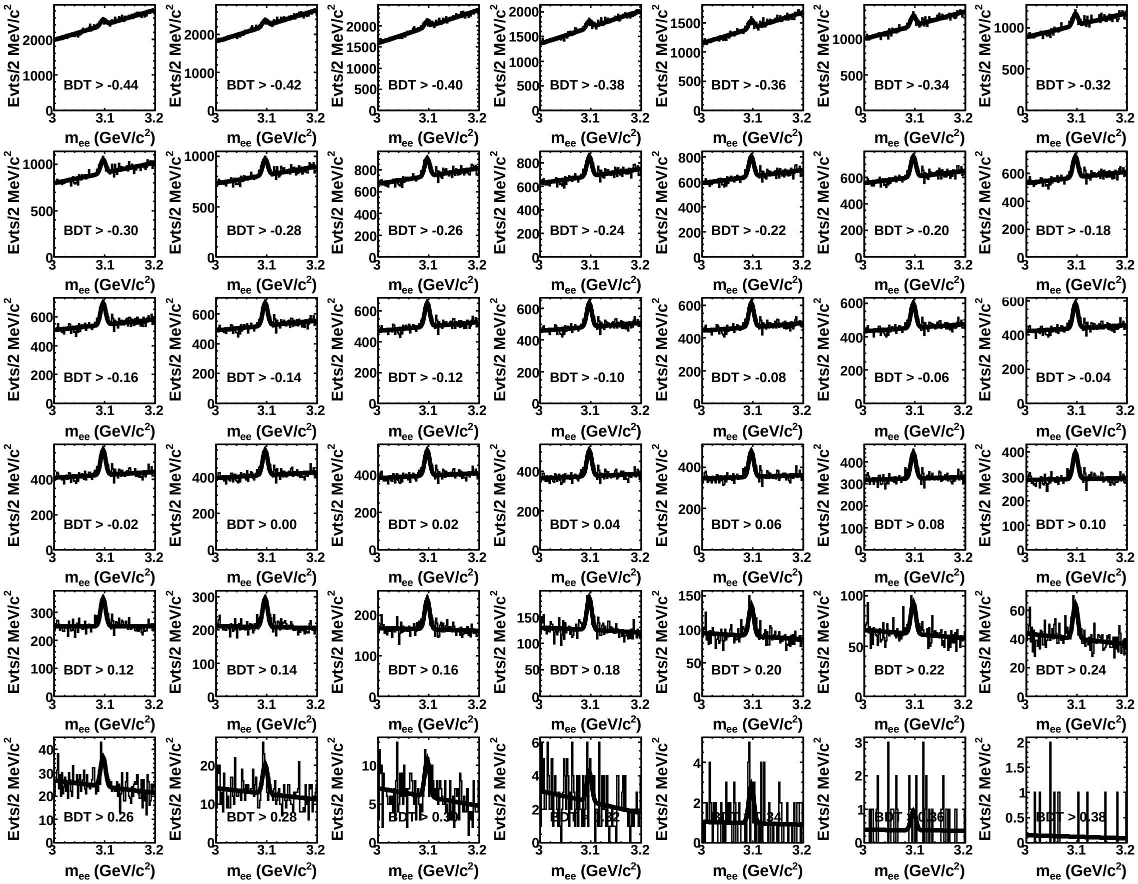


Figure 4: Fits to the J/ψ resonance for different values of minimum mva_{BDT}

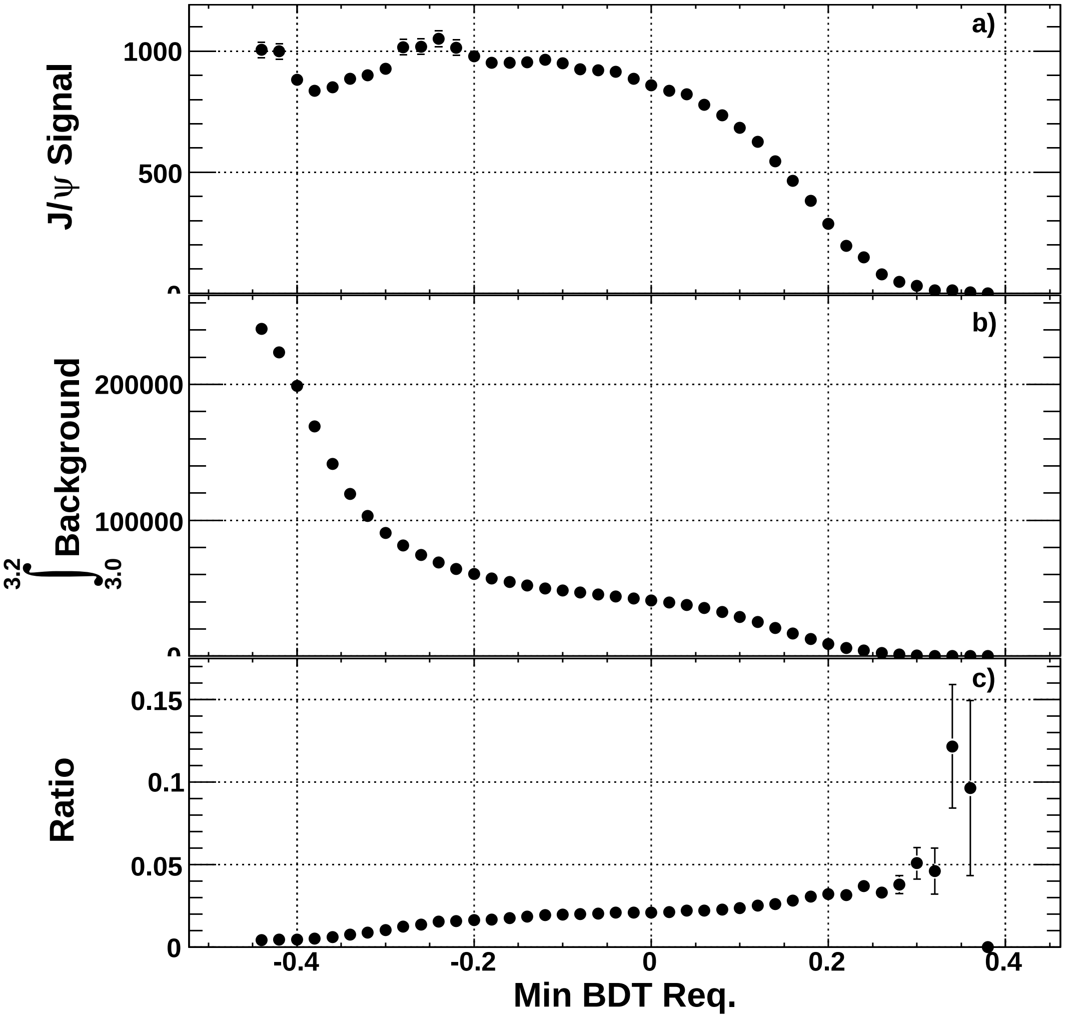


Figure 5: The normalizations of the J/ψ resonance (top) and background (middle), and the signal-to-background ratio (bootm) plotted against minimum required value of mva_{BDT}. At 0.2 on the x-axis, corresponding to a cut of $mva_{BDT} < 0.2$, there is a local maximum in the signal efficiency.

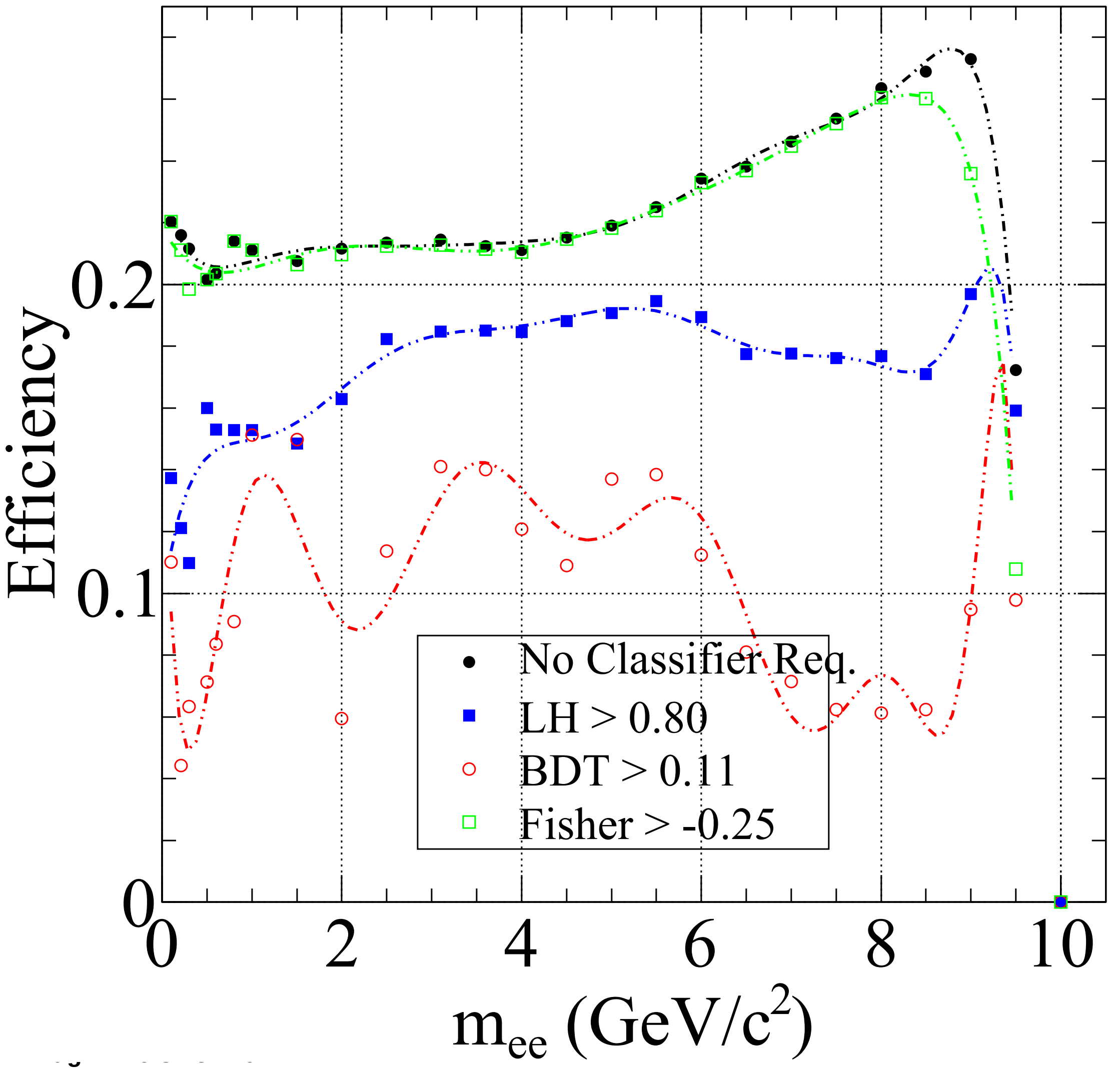


Figure 6: The signal efficiency $\epsilon(m_{ee})$ calculated from 25 Monte Carlo modes with fits to high order polynomials. Also plotted are the efficiencies calculated after applying the cuts from several different MVA methods.

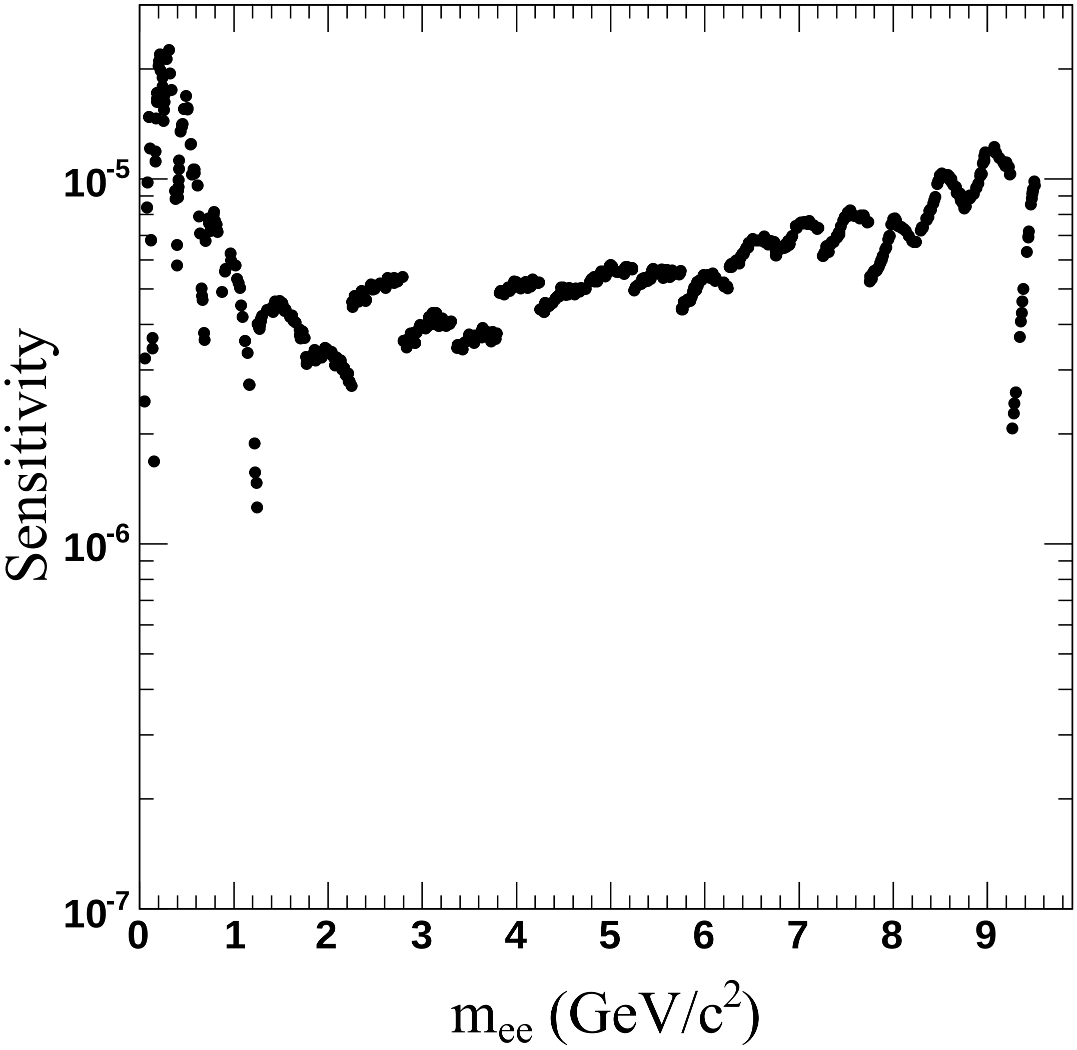


Figure 7: Sensitivity plotted against m_{ee}. Discontinuities are caused by the evaluation of the sensitivity based on parameters specific to each separate Monte Carlo mode. The sensitivity calculation at m_{ee} uses parameters from the Monte Carlo with generated mass closest to m_{ee}.

SLAC-TN-12-024

Using Dynamic Quantum Clustering to Analyze Hierarchically Heterogeneous Samples on the Nanoscale

Allison Hume

Office of Science, Science Undergraduate Laboratory Internship (SULI)
Princeton University
SLAC National Accelerator Laboratory
Menlo Park, California

August 19, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Marvin Weinstein at the SLAC National Accelerator Laboratory.

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Table of Contents

Abstract

Using Dynamic Quantum Clustering to Analyze Hierarchically Heterogeneous Samples on the Nanoscale. ALLISON HUME (Princeton University, Princeton, NJ 08544) MARVIN WEINSTEIN (SLAC National Accelerator Laboratory, Menlo Park, CA 90425)

Dynamic Quantum Clustering (DQC) is an unsupervised, high visual data mining technique. DQC was tested as an analysis method for X-ray Absorption Near Edge Structure (XANES) data from the Transmission X-ray Microscopy (TXM) group. The TXM group images hierarchically heterogeneous materials with nanoscale resolution and large field of view. XANES data consists of energy spectra for each pixel of an image. It was determined that DQC successfully identifies structure in data of this type without prior knowledge of the components in the sample. Clusters and sub-clusters clearly reflected features of the spectra that identified chemical component, chemical environment, and density in the image. DQC can also be used in conjunction with the established data analysis technique, which does require knowledge of components present.

Introduction

The ability to study hierarchically heterogeneous materials on the nanoscale can lend insight to problems ranging from the chemistry of batteries to techniques of ancient artists. A hierarchically heterogeneous sample is a sample that contains multiple chemical components that appear homogenous on the nanoscale, but heterogeneous on a micron or millimeter scale. The Transmission X-ray Microscopy (TXM) group at beam line 6-2c in the Stanford Synchrotron Radiation Light-source (SSRL) has the ability to image such samples with a 30 nanometer resolution and field of view large enough to see the hierarchical structure.

The data collected for a sample is a set of images taken over a range of energies from which the TXM group can determine a full X-ray Absorption Near Edge Structure (XANES) spectrum for each pixel. XANES goes beyond identifying the elements present in a sample, determined by the edge-jump in the spectrum, and provides information about the surrounding chemistry, such as the oxidation state of the element. This information is present in the fine structure in the energy range higher than the edge-jump. [1] The TXM group, however, is not able to analyze this data without knowing what chemical components are present in the sample. The current data analysis procedure compares the absorption spectrum for each pixel to spectra for compounds known to be in the sample using R^2 fitting. A colored map of the sample is created by assigning the known spectra to colors such as red and blue and coloring each pixel according to how similar it is to the "red" or "blue" spectra. Any pixels with extremely poor R^2 values are examined to find unknown components. This method is effective for samples in which the composition is known and the only information needed is the distribution, but a new method is necessary for samples of unknown composition.

Dynamic Quantum Clustering (DQC) is a method of data mining that does not require any previous information about the composition of the sample. It performs unsupervised clustering to identify energy spectra that have similar characteristics. To test the applicability of DQC to TXM-XANES data, a sample was used that had been analyzed by the original method as well. The sample came from a piece of ancient Roman pottery made from hematite (iron(III) oxide) and hercynite (iron(II) aluminate). The images are taken as a cross-section of the interface between the two oxides with hematite on the surface and hercynite underneath. This data set consists of a series of 146 images, each with close to one million pixels, which determines the spectra for each pixel. The spectra are noisy and the combination of this noise and the large data size creates a significant data mining challenge to find groups of pixels with similar spectra.

Methods

DQC is an algorithm that maps a problem of unsupervised clustering to a problem in quantum mechanics and provides a visual animation of the cluster formation. An outline of the algorithm appears below and a full description of the theory and methods of computation can be found in *Dynamic quantum clustering: a method for visual exploration of structures in data* by Marvin Weinstein and David Horn. [2]

In DQC each data point is described as a Gaussian wave function in an n-dimensional parameter space. The full data set is then associated with the sum of these individual Gaussians:

$$\Psi(\vec{x}) = \sum_i e^{\frac{-(\vec{x}-\vec{x_i})^2}{2\sigma^2}} = \sum_i \psi_i(\vec{x_i}) \tag{1}$$

Choosing a larger value for σ causes the individual wave functions to have more overlap while a smaller value decreases overlap. This in turn affects the level of definition of the extrema of the full wave function.

The wave function is then taken to be the ground state of the Hamiltonian:

$$(-\frac{\sigma^2}{2}\mathbf{\nabla}^2 \ + V(\vec{x}))\Psi(\vec{x}) = E\Psi(\vec{x}), E = 0 \tag{2}$$

The full wave function therefore determines some potential $V(\vec{x})$ that will have minima determined by the maxima of the wave function, which correspond to the most most dense regions of the data in parameter space:

$$V(\vec{x}) = \frac{\sigma^2}{2\Psi(\vec{x})}\mathbf{\nabla}^2\Psi(\vec{x}) \tag{3}$$

The potential, is less sensitive to the value of σ than the full wave function so the choice of value for that variable is less important than it is for other data mining methods that use a sum of Gaussians.

The individual data points are them evolved in time according to the time-dependent Schrödringer equation:

$$i\frac{\partial\psi_i(\vec{x_i},t)}{\partial t} = (-\frac{\mathbf{\nabla}^2}{2m} + V(\vec{x}))\psi_i(\vec{x_i},t) \tag{4}$$

The value chosen for the mass of the "particle" has an effect on the amount of tunneling that occurs as the data evolve, as well as the speed at which the evolution occurs.

The data is a matrix: each row corresponds to the 146 point energy spectrum of a single pixel. After removal of the pixels for which there was no data, slightly over half a million pixels were left. Singular value decomposition (SVD) was then performed. SVD is a method of factorizing a matrix that allows the original matrix to be approximated as a sum of a series of terms that show increasing detail. In this case, since each row of the matrix is a spectrum, recreating the matrix from a subset of the 146 singular components reduced the noise in the spectra. The more singular components used to recreate the original matrix, the closer the spectra will be to their original shape. The first five components were used for the rest of the initial analysis. Five features were enough to retain the important features in the shape of the spectra, while reducing the noise of the spectra greatly. This can be seen in Figure 1 which shows an example spectrum taken from the original matrix in green and the same spectrum reconstructed from the first give components in black.

DQC is able to operate in high dimension; however, the speed of computation scales linearly with the number of dimensions. So reduction from 146 to five dimensions in the data matrix increased the speed of computation. The most important step taken to reduce the time of computation, however, was the selection of template states. Template states are essentially an approximate basis: a subset of data points, linear combinations of which can reproduce all of the other states up to a given accuracy. By choosing about 1500 states that spanned the full data and only evolving these states, the computation became possible on a desktop workstation. A description of the DQC Maple library, including further detail on the process of selecting template states can be found on Weinstein's website. [3] In this analysis, all of the states were used to construct the potential, the template states were evolved through 75 time-steps using the potential, and the remaining states were expressed as approximate linear combinations of the template states to fully evolve the data.

After the 75 frame animation of the evolution was completed, the result was examined, and clusters and sub-clusters were visually identified. In this analysis, a cluster is considered as a group of data points which is separate from all others in any dimension and a sub-cluster is any visible structure within a cluster such as connected strands. The data were then colored based on the visible structure, the original image was reconstructed from those colors, and the clusters were used to understand the structure and distribution of the data. A selection of reconstructed spectra, as well as the average spectra, was examined from each cluster and sub-cluster to determine the aspects of the spectra that affected the clustering of the data. Four clusters and the sub-structure of two of the clusters were examined in depth. The colors used are shades of red, shades of blue, and shades of green. The three shades of red (dark red, red, and pink) and the shades of blue (blue and cyan) each correspond to the sub-structure of a cluster while the shades of green (green and dark green) correspond to two clusters.

Results

As DQC had never been used for this type of data, the first important result was that the data clustered. Figure 2 shows the final frame of the evolution and illustrates how both the clustering and sub-clustering appear. It is important to note that the size of the cluster in the final frame does not necessarily correspond to the number of points it contains. In theory, each isolated point in the final frame must be considered as an individual cluster. Watching the animation is useful to estimate the size of clusters. Figure 3 shows the final frame again but illustrates how the clusters and sub-clusters were colored. In practice, most of the isolated points were grouped with other clusters.

The picture that was recreated based on the clustering of the data appears in Figure 4(a) colored according to how each pixel was colored in the final frame of the animation. It is important to note that no information about the location of the pixels in the image was used during the clustering process. Figure 4(b) shows the picture that was created from the original data analysis method for comparison. Information about pixel location *was* used by the TXM group to create this picture. Again, the different shades of color in the DQC picture come from the structure of the data. The different shades of red and green in the original, however, were added by examining the amplitude of the curves after the fitting was performed and adding white to the lower amplitude curves to represent lower intensity pixels. The blue piece of the picture was identified in the original analysis through the poor fit those pixels had with both the hematite and hercynite spectra. In the DQC analysis those pixels clustered to a point within four frames and remained separated for the rest of the animation.

Examining the spectra in each cluster provided information about what aspects of the curves were important in the clustering process. Several groups of clusters and sub-clusters were seen to have spectra of similar shape, but different intensities. Figure 5 shows the averages of these groups, as well as a selection of curves from each group to illustrate how the intensity of spectra comes out in the clustering.

Most of the large clusters corresponded to groups of spectra with different shapes. Figure 6 shows 100 random spectra from each of the three clusters whose spectra had the most distinct shapes. Upon examining the spectra in the green and red clusters, it was confirmed that the green spectra correspond to the spectra of hematite, and the spectra in the main red band correspond to the spectra of hercynite. The locations of hematite and hercynite in the image (locations of green and red pixels in Figure 4(a)) matched what was expected based on the original analysis. The information about the location of hematite and hercynite, although confirmed by the other analysis method, were therefore determined purely from similarities in spectra in the DQC analysis.

It is clear (both in Figure 6 and Figure 5) that the green spectra are tightly grouped while the red and blue spectra contain a much larger spread of curve shapes. This information is reflected in the sub-clustering of the three clusters. The green cluster shows no sub-structure while the red cluster shows three large strands, among other features, and the blue shows a tight node and two diffuse strands. The sub-structure of the red and blue clusters were examined in depth. One strand of the red cluster was determined to correspond to a different chemical environment, as seen in

Figure 7. Figure 7(a) shows only the red cluster in a two dimensional projection of the last frame of the animation with one strand colored yellow. Figure 7(b) shows 20 random spectra from both the red strand and the yellow strand. It shows that the yellow strand corresponds to an apparently different chemical environment from the hercynite from the rest of the red. This is visible in the image in Figure 4(a) as the yellow pixels mostly appear on the interface between the hematite and the hercynite. The blue cluster has a band that resembles the spectrum of mostly pure iron but also contains curves that resemble hercynite. Although the blue cluster appeared as a point on the animation by the fourth or fifth frame, the sub-structure of this cluster is visible with the other clusters removed. Figure 8(a) contains a view of the sub-structure of this cluster from the fifth frame of the animation. Figure 8(b) contains sample spectra from the different components of the structure and Figure 8(c) contains the average spectra from the two components. Both of these plots show that the curves more closely resembling hercynite or pure iron separate in the sub-clustering. It is interesting to note in the image in Figure 4(a) that the light blue or pure iron appears as a coherent band.

Discussion and Conclusion

The DQC algorithm successfully identified structure in the XANES data. The animation of the evolution shows detailed clustering and sub-clustering which correspond to physically important information in the curves. Upon examination of the spectra present in the various clusters it appears that the most important information in forming the clusters comes from the shape of the pre-edge and edge jump. Thus, different components are the first groups to cluster out. Clusters containing spectra for hematite, hercynite, and relatively pure iron contained pixels from areas where those components were expected to appear based on the original analysis, which confirms that DQC is able to correctly identify different chemical components in the data. Some sub-structure of clusters identified the chemical environment of a compound that is present. This was seen in the mapping of apparent differences in the hercynite spectra to the visible strands of the red cluster.

The other information that is important in the cluster formation is the intensity of the spectra.

This was seen in how well some separate clusters and large sub-features corresponded to the intensity information that had been added in the original picture. Although learning that clustering occurs partially based on intensity was informative for understanding how DQC responds to XANES data, in the future clustering will most likely be performed on normalized data. This is because intensity is not physically interesting, and it can be added to the image afterwards, in the same way it was added to the TXM image. Clustering on normalized data will most likely simplify the structure that needs to be examined which will make it easier to find physically interesting information.

The blue cluster (pixels containing mostly iron) represented an example of a "strong" cluster, or a cluster that forms almost immediately and does not join other clusters later. It was important to see that DQC was sensitive to this cluster because iron and hercynite have similar spectra and the pixels in the iron cluster represented less than 1% of the data set. This was a confirmation of how well DQC works with this type of data. The characteristic differences in the spectra is barely visible to the eye when examining the spectra reconstructed from five SVD components but the algorithm is sensitive enough to small changes in the spectra that this tiny subset of the data clustered immediately and remained separate from the data.

It is also important to note that the two data analysis methods can be used well in tandem - DQC provides information about what parts of the image might be interesting to study in depth but, by using the traditional method to create a second picture, the user can also identify areas of the image that should be examined further in the DQC animation. Any area of the image identified by the user can be examined in the animation and so the user can request specific information such as how a smaller group of pixels clusters. The differences between the two pictures are also important and will generally reflect areas that should be examined in depth. This provides another motive for using the two data analysis techniques together. The traditional method is currently faster but DQC is capable of providing more detail and work is continuing to speed up the algorithm. DQC holds great potential for use in the analysis of XANES data in the TXM group.

Acknowledgements

I would like to thank my mentor Marvin Weinstein for all of his help and support as well as Apurva Mehta, Florian Meirer, and Yijin Liu for their time and assistance. I would also like to thank the Department of Energy for funding this program and for everyone at the SLAC National Laboratory for being welcoming and helpful. Thank you to Stephen Rock, Maria Mastrokyriakos and Anita Piercy for all of the work they put into making the summer go smoothly.

REFERENCES

[1] F. Meirer, Y. Liu, A. Mehta. Mineralogy and morphology at nanoscale in hierarchically heterogeneous materials. June 24, 2011.

[2] M. Weinstein, D. Horn. Dynamic quantum clustering: a method for visual exploration of structures in data. SLAC-PUB-13759.

[3] M. Weinstein. DQCOverview1. `http://www.slac.stanford.edu/~niv/index_files/DQCOverview1.html`

FIGURES

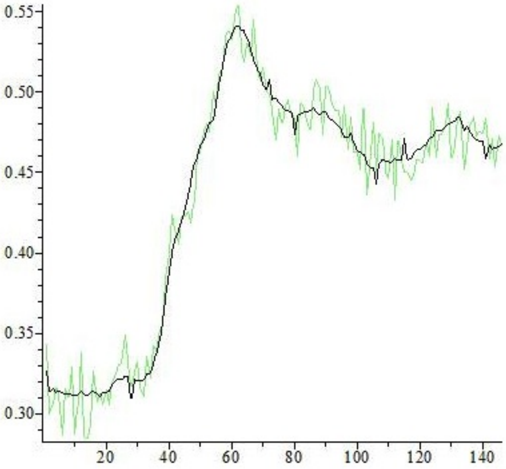
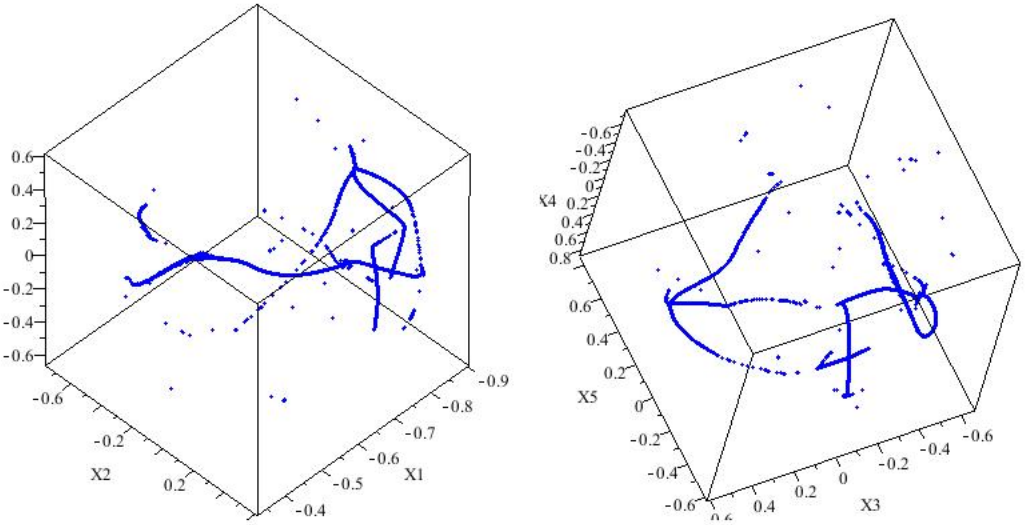


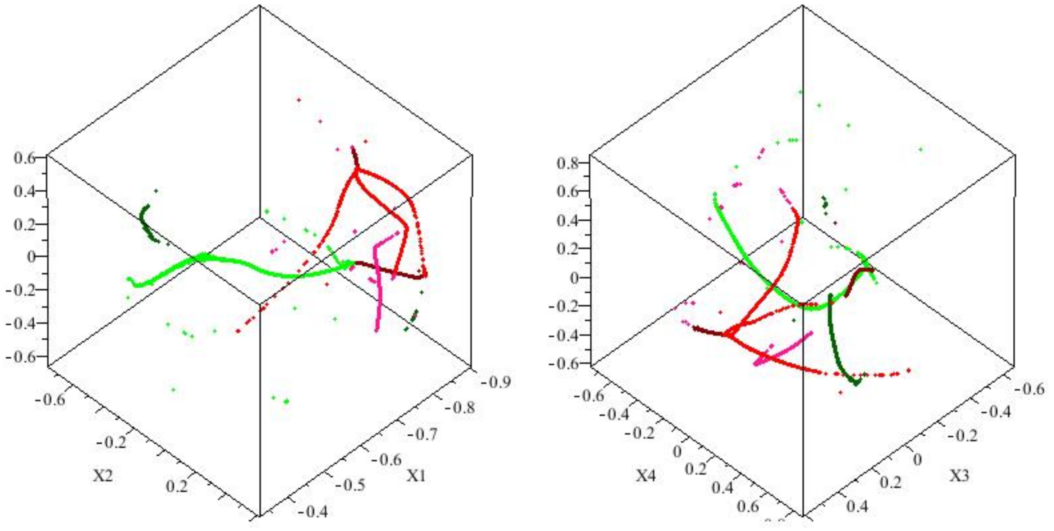
Figure 1: In this illustration of SVD reconstruction the black curve is a sample spectrum taken from the data and the green curve is the same spectrum reconstructed from the first five singular components.



(a) Dimensions 1, 2, 3

(b) Dimensions 3, 4, 5

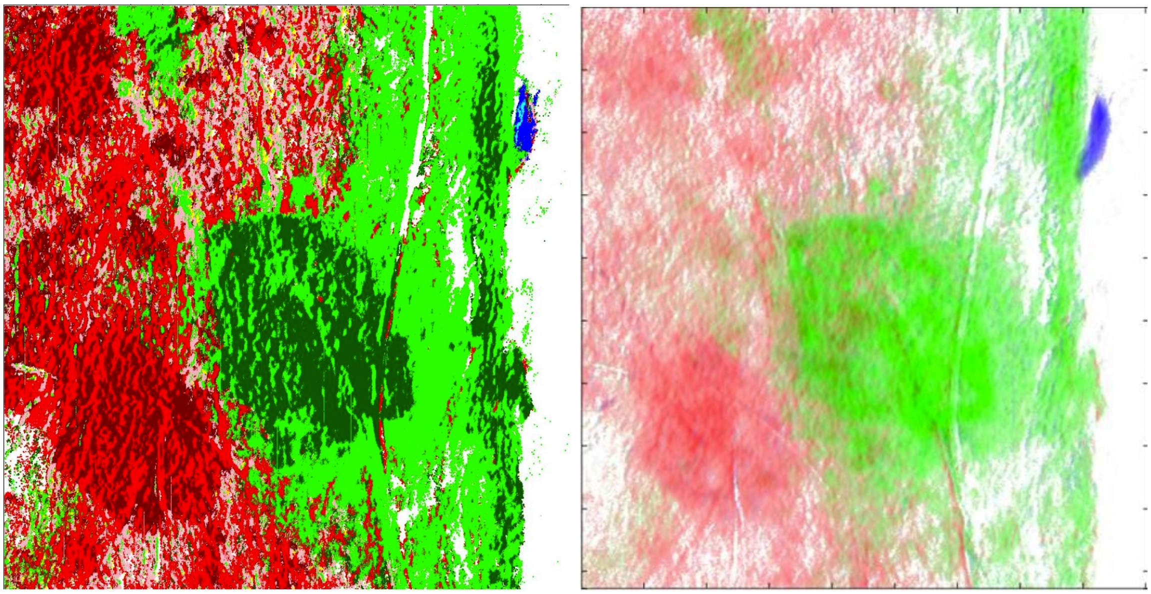
Figure 2: Two views of the last frame of the DQC animation that show all five dimensions of the data and illustrate how the data separated into clusters, as well as how many of the clusters contain visible sub-structure.



(a) Dimensions 1, 2, 3

(b) Dimensions 3, 4, 5

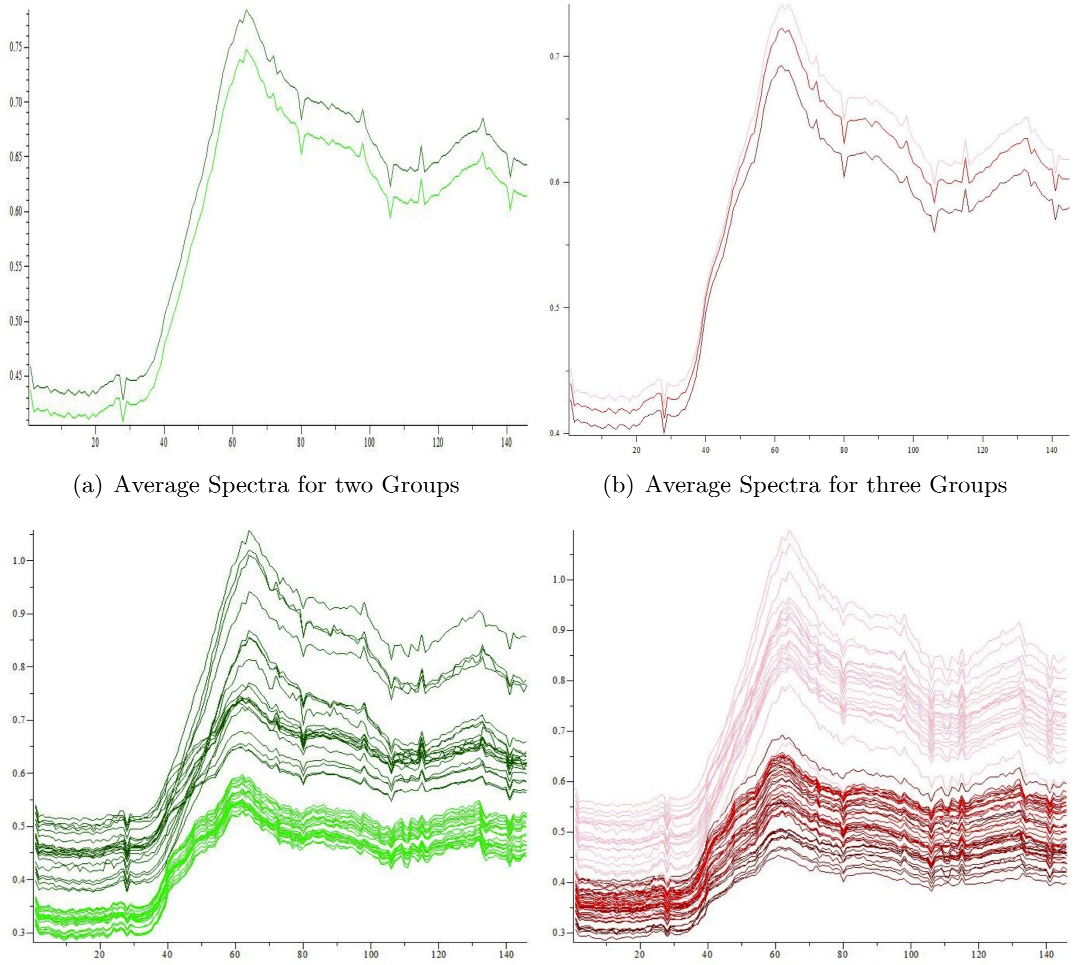
Figure 3: Two views of the last frame of the DQC animation that show all five dimensions of the data and illustrate how clusters and sub-clusters were colored.



(a) Image Created by DQC Analysis

(b) Image Created by Fitting Analysis

Figure 4: Two images of the data with the left image colored based on the unsupervised clustering of the data using DQC analysis and the right image colored based on R^2 fitting to previously known components.



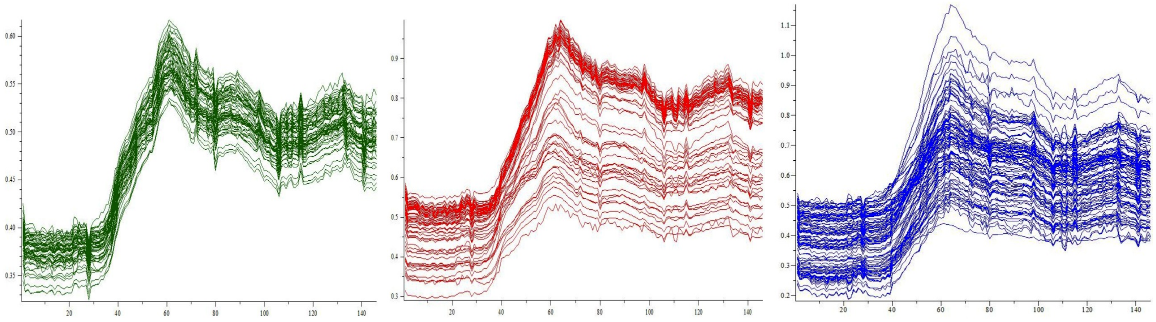
(a) Average Spectra for two Groups

(b) Average Spectra for three Groups

(c) 100 Random Spectra for two Groups

(d) 100 Random Spectra for three Groups

Figure 5: The amplitude differences seen between spectra of several clusters and sub-clusters that have spectra of similar shape. There were two major intensities seen in the shape in green and at least three seen in the shape in red. The different intensities are illustrated using the average reconstructed spectra for a group and 100 random reconstructed spectra for each group.

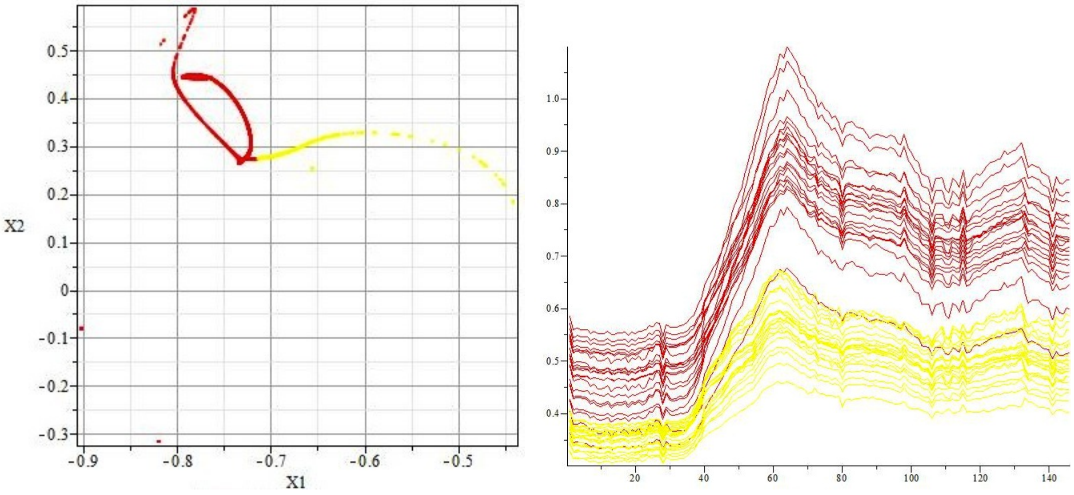


(a) Hematite

(b) Hercynite

(c) Mostly Iron

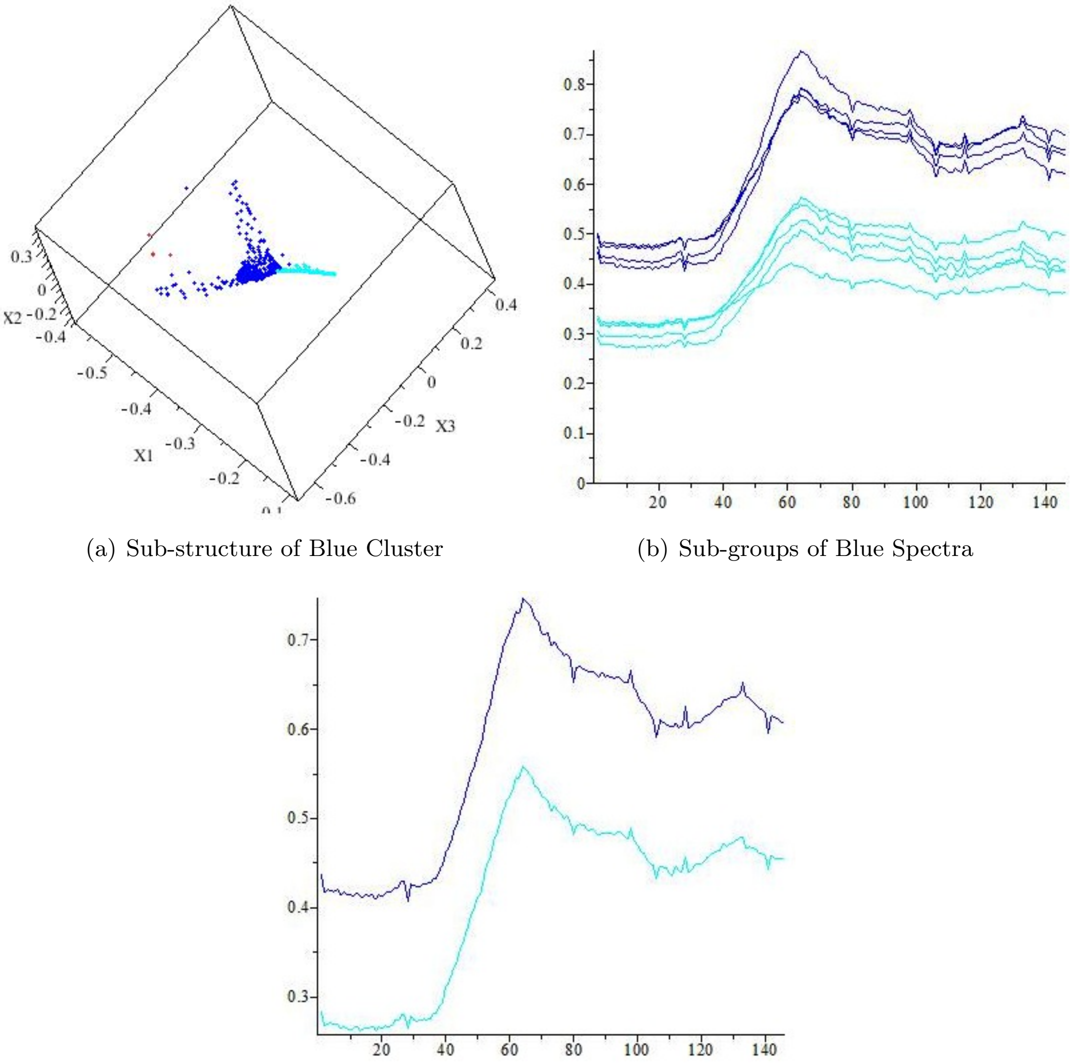
Figure 6: 100 random reconstructed spectra from three clusters that illustrate the three main spectra shapes, and therefore the three main chemical components seen in the data.



(a) Sub-structure of Red Cluster

(b) Sub-groups of Red Spectra

Figure 7: The sub-structure of the red cluster is visible in the last frame of the animation. This sub-structure reflects differences in the shape of the spectra, corresponding to different phases of hercynite.



(a) Sub-structure of Blue Cluster

(b) Sub-groups of Blue Spectra

(c) Averages of Sub-groups of Blue Spectra

Figure 8: The sub-structure of the blue cluster is visible in the fifth frame of the animation when the rest of the data is removed. This sub-structure reflects differences in the shape of the spectra, corresponding to how much iron is present. Curves more closely resembling hercynite or pure iron separate in the sub-clustering. The lighter blue curves show the flattening of the spectra after the edge-jump that is characteristic of iron.

SLAC-TN-12-025

Representing Range Compensators with Computational Geometry in TOPAS

Forrest N. Iandola

University of Illinois at Urbana-Champaign

and SLAC National Accelerator Laboratory, Menlo Park, CA 94025

August 19, 2011

Prepared in partial fulfillment of the requirements of the Department of Energy's Science Undergraduate Laboratory Internship under the direction of **Joseph Perl**.

Participant: ______________________

Signature

Research Advisor: ______________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

In a proton therapy beamline, the range compensator modulates the beam energy, which subsequently controls the depth at which protons deposit energy. In this paper, we introduce two computational representations of range compensator. One of our compensator representations, which we refer to as a *subtraction solid*-based range compensator, precisely represents the compensator. Our other representation, the *3D hexagon*-based range compensator, closely approximates the compensator geometry. We have implemented both of these compensator models in a proton therapy Monte Carlo simulation called TOPAS (Tool for Particle Simulation). In the future, we will present a detailed study of the accuracy and runtime performance trade-offs between our two range compensator representations.

INTRODUCTION

Proton therapy is one of the best methods available for treating cancer [1]. The number of proton facilities is steadily increasing, and proton therapy is entering the mainstream of radiation therapy. TOPAS (Tool for Particle Simulation) is a four year NIH-funded project of SLAC, University of California, San Francisco (UCSF) and Massachusetts General Hospital to make Monte Carlo particle transport simulation faster and easier to use for proton therapy [2] [3] [4]. TOPAS enables researchers and clinicians to refine proton therapy delivery system designs and individual patient treatment plans to improve proton therapy's therapeutic ratio (improved tumor control, reduced side effects). TOPAS offers a collection of ready-made geometries for modeling proton therapy beamlines, and TOPAS uses Geant4 for fundamental physics processes.

In most proton therapy facilities, an extensive beamline lies between the proton source and the patient. This beamline includes components such as collimators, magnets, a range modulator, scatterers, and a range compensator. The range compensator (or simply *compensator*) sits close to the patient. The fundamental purpose of a range compensator is to modify the proton beam energy so that the protons deliver energy to the correct location in a patient. This study seeks to represent compensators with computational geometry. While a wide variety of computational geometry methods could be applied to compensator representation, we focus on fast and accurate methods. This paper presents two compensator representations, which the authors have recently implemented in TOPAS.

BACKGROUND

The Range Compensator

In proton treatment facilities, medical staff typically prescribe a radiation dose (in Greys) to a patient's tumor. To ensure that the correct dose is delivered to the correct location, the protons must have a specific energy. The necessary energy may vary within the treatment region. A *range compensator* allows a beamline to deliver protons with the correct energy (and corresponding depth of deposition).

In many proton therapy facilities, a non-Monte Carlo treatment planning software such as XiO [5] designs a unique range compensator for each patient. In some facilities, medical physicists use a Monte Carlo simulation such as TOPAS to validate the compensator design. Therapy facilities typically construct compensators by using a milling machine to drill a number of holes out of a cylinder of Lucite or other clear plastic. Each drill hole may have a unique depth, such as in Figure 1. The thickness of the Lucite proportionally reduces the protons' energy. Lower-energy protons deposit most of their energy at a shorter depth in the patient than higher-energy protons. In short, the range compensator modulates the beam energy, which subsequently controls the depth at which protons deposit energy.

Subtraction Solids in Geant4

Geant4 supports boolean solid combinatorial geometry [6]. Subtraction solids are a type of Geant4 boolean solid. Subtraction solids enable a small solid to be cut out, or *subtracted*, from a larger solid [7]. The pseudo code in Figure 2 serves as a simple example of subtraction solid geometry. As shown in Figure 3, Geant4 also allows nested subtraction of solids. In addition, subtracted solids are allowed to overlap. Therefore, for the example in Figure 3, Geant4's ability to track particles is not impaired if the subtracted cylinders, `smallCylinder1` and `smallCylinder2`, overlap. In contrast, geometrical components that are not comprised of boolean solids are not allowed to overlap in Geant4.

RANGE COMPENSATOR MODEL WITH SUBTRACTION SOLIDS

Given the compensator milling machine instructions in .Decimal format [8], we subtract each drill hole from the solid Lucite cylinder. We subtract the cylinders in the same order that the milling machine would drill the cylinders. Thus, we produce an exact model of the real compensator. The pseudo code in Figure 3 constructs a compensator comprised of a bigCylinder with n drill holes subtracted. Figure 4 offers a Geant4 visualization [9] of a subtraction solid compensator with twelve holes subtracted.

Our (unpublished) preliminary results show that, especially for compensators with fifty or more drill holes, tracking particles through a subtraction solid-based compensator is quite slow. To remedy this, we present a hexagon-based compensator approximation in the next section.

“3D HEXAGON”-BASED RANGE COMPENSATOR APPROXIMATION

In this section, we present an approximate compensator representation. We approximate the drill holes with a collection of “3D hexagons,” as shown in Figure 5. Each 3D hexagon has the depth of the drill hole that it represents.

We choose 3D hexagons, because they can be clustered without overlap or gaps. (Recall that, while boolean solids allow overlap, other forms of Geant4 geometry do not allow overlap.) The lack of overlap among the hexagons allows us to insert a volume of air in the shape of each 3D hexagon into the Lucite cylinder. However, unlike the precise compensator model produced with subtraction solids, the 3D hexagon model does not correctly represent the drill hole depth in regions where drill holes overlap.

Our preliminary results show that particle tracking through a hexagon-based compensator is as up to twenty times faster than through a subtraction solid-based compensator. We will present the finalized version of these results in a future study.

CONCLUSIONS AND FUTURE WORK

In this paper, we introduced two range compensator representations, which we have implemented in TOPAS. In the near future, we will present a detailed study of the accuracy and runtime performance trade-offs between our two compensator representations.

ACKNOWLEDGMENTS

The authors thank Jan Schuemann, Harald Paganetti, Jungwook Shin, and Bruce Faddegon for their helpful suggestions. Thanks to the U.S. Department of Energy and the National Institutes of Health for their generous support.

REFERENCES

[1] H. Suit, S. Goldberg, A. Niemierko, A. Trofimov, J. Adams, H. Paganetti, G. Chen, T. Bortfeld, S. Rosenthal, J. Loeffler, and T. Delaney, "Proton beams to replace photon beams in radical dose treatments." *Acta Oncologica*, vol. 42, no. 8, pp. 800–808, 2003.

[2] J. Perl, J. Shin, J. Schuemann, B. Faddegon, and H. Paganetti, "Topas: A fast and easy to use tool for particle simulation," in *Joint American Association of Physicists in Medicine (AAPM) and Canadian Organization of Medical Physicists (COMP) Meeting*, August 2011.

[3] J. Perl, J. Schuemann, J. Shin, B. Faddegon, and H. Paganetti, “Topas: A fast and easy to use monte carlo system for proton therapy,” in *Presented at the 50th Meeting of the Particle Therapy Co-Operative Group (PTCOG)*, May 2011.

[4] T. Akagi, T. Aso, B. Faddegon, A. Kimura, N. Matsufuji, T. Nishi, C. Omachi, H. Paganetti, J. Perl, T. Sasaki, D. Sawkey, J. Scheumann, J. Shin, T. Toshito, T. Yamashita, and Y. Yoshida, “The ptsim and topas projects, bringing geant4 to the particle therapy clinic,” in *Joint International Conference on Supercomputing in Nuclear Applications and Monte Carlo 2010 (SNA + MC 2010)*, October 2010.

[5] M. Lee, “Xio treatment planning system from elekta cms software optimizes clinical treatment at rinecker proton therapy center,” Elekta, Inc. Press Release, October 2009, http://www.elekta.com/healthcare_international_press_release_20070632.php.

[6] G. Cosmo, “Modeling detector geometries in geant4,” in *IEEE Nuclear Science Symposium (IEEE-NSS)*, vol. 1, October 2003, pp. 479–481.

[7] ——, “The geant4 geometry modeler,” in *IEEE Nuclear Science Symposium (IEEE-NSS)*, vol. 4, October 2004, pp. 2196–2198.

[8] B. E. Willis, “Commissioning of adac pinnacle for imrt utilizing solid compensators,” Master’s thesis, Oregon State University, 2008.

[9] J. Allison, M. Asai, G. Barrand, M. Dönszelmann, K. Minamimoto, J. Perl, S. Tanaka, E. Tcherniaev, and J. Tinslay, “The geant4 visualisation system,” *Computer Physics Communications*, vol. 178, no. 5, pp. 331–365, 2008.

FIGURES

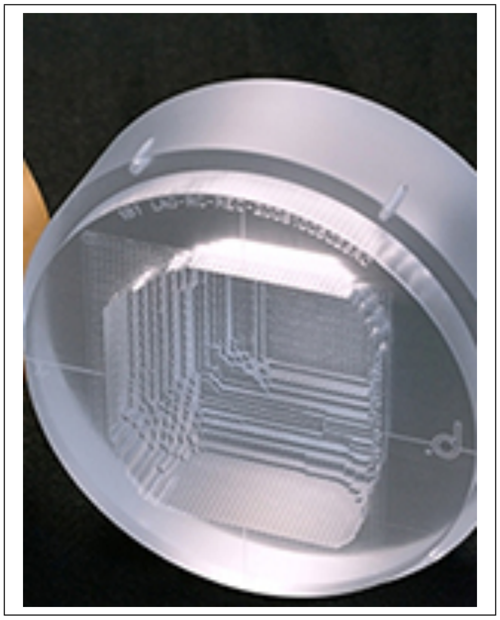


Figure 1: Example range compensator

```
newSolid = bigCylinder - smallCylinder
```

Figure 2: Pseudo-code for simple subtraction solid example

```
newSolid1 = bigCylinder - smallCylinder1
newSolid2 = newSolid1 - smallCylinder2
...
Compensator = newSolid(n-1) - smallCylinder(n)
```

Figure 3: Pseudo-code for repeated subtraction solid example

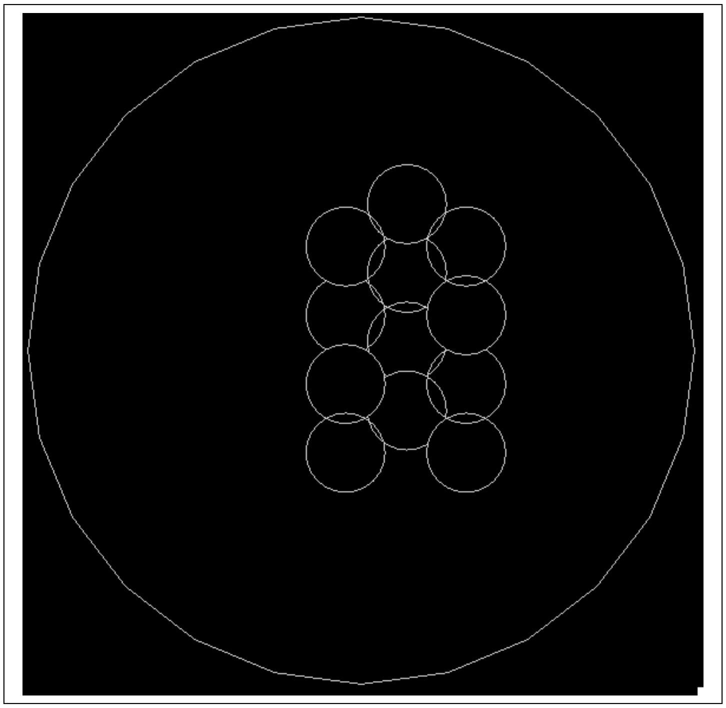


Figure 4: Compensator modeled with subtraction solids

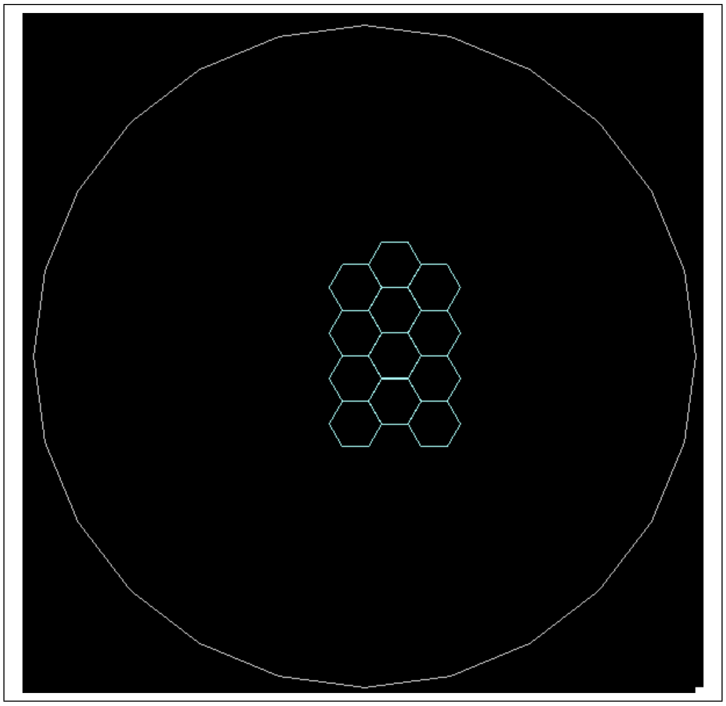


Figure 5: Compensator modeled with 3D hexagons

SLAC-TN-12-026

Channeling through Bent Crystals

Stephanie Mack

Office of Science, Science Undergraduate Laboratory Internship (SULI)

University of Ottawa

SLAC National Accelerator Laboratory

Menlo Park, CA

August 19, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of H.-Ulrich (Uli) Wienands at the Stanford University, SLAC National Accelerator Laboratory.

Participant: ______________________

Signature

Research Advisor: ______________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Channeling through Bent Crystals. STEPHANIE MACK (University of Ottawa, Ottawa, ON K1N 6N5), H.-ULRICH (ULI) WIENANDS (Stanford University, SLAC National Accelerator Laboratory, Menlo Park, CA 94025).

Bent crystals have demonstrated potential for use in beam collimation. A process called channeling is when accelerated particle beams are trapped by the nuclear potentials in the atomic planes within a crystal lattice. If the crystal is bent then the particles can follow the bending angle of the crystal. There are several different effects that are observed when particles travel through a bent crystal including dechanneling, volume capture, volume reflection and channeling. With a crystal placed at the edge of a particle beam, part of the fringe of the beam can be deflected away towards a detector or beam dump, thus helping collimate the beam. There is currently FORTRAN code by Igor Yazynin that has been used to model the passage of particles through a bent crystal. Using this code, the effects mentioned were explored for beam energy that would be seen at the Facility for Advanced Accelerator Experimental Tests (FACET) at a range of crystal orientations with respect to the incoming beam. After propagating 5 meters in vacuum space past the crystal the channeled particles were observed to separate from most of the beam with some noise due to dechanneled particles. Progressively smaller bending radii, with corresponding shorter crystal lengths, were compared and it was seen that multiple scattering decreases with the length of the crystal therefore allowing for cleaner detection of the channeled particles. The input beam was then modified and only a portion of the beam sent through the crystal. With the majority of the beam not affected by the crystal, most particles were not deflected and after propagation the channeled particles were seen to be deflected approximately 5mm. After a portion of the beam travels through the crystal, the entire beam was then sent through a quadrupole magnet, which increased the separation of the channeled particles

from the remainder of the beam to a distance of around 20mm. A different code, which was developed at SLAC, was used to create an angular profile plot which was compared to what was produced by Yazynin's code for a beam with no multiple scattering. The results were comparable, with volume reflection and channeling effects observed and the range of crystal orientations at which volume reflection is seen was about 1 mrad in both simulations.

INTRODUCTION

Crystals contain atoms arranged in a lattice which forms planes. When accelerated particles enter a crystal parallel to a plane they can be trapped by the nuclear potentials, a process called channeling [1]. If the crystal is bent elastically then particles may follow the curvature of the planes. This concept can then be applied to beam collimation: if a crystal is placed to intercept the halo particles of a beam these can be deflected away from the beam to a detector or beam dump. This may improve the collimation of the beam which is important for preventing damage to equipment and improving the noise to signal ratio of experiments using the beam. Previous channeling experiments have examined this effect at circular accelerators including the RD22 experiment at the H8 beamline at the European Organization for Nuclear Research (CERN) [2] and the Tevatron at Fermilab [3].

An experiment to be proposed for the Facility for Advanced Accelerator Experimental Tests (FACET) at SLAC National Accelerator Laboratory (SLAC) will study the effectiveness of a bent crystal for collimation of the 23 GeV positron/electron beam using planar channeling with a silicon (110) crystal. The CRY_AP code written in Fortran by Igor Yazynin and modified by Robert J. Noble [4] provides the baseline for the study of the bent crystal. The effects of volume reflection, volume capture, channeling and dechanneling were investigated with a beam of positively charged particles passing through the crystal. Volume capture happens when a particle is not initially channeled, but as it passes through successive planes it eventually makes a critical angle with a plane and becomes channeled. A dechanneled particle is one that is initially channeled between the planes but then exits. These processes are largely caused by multiple scattering of the particle. Volume reflection occurs when the particles hit the planes of the crystal at a shallow angle. They are thus reflected off the plane and do not get trapped between the planes and are only slightly deviated from a straight path.

The crystal parameters were optimized for FACET beam energy. A more realistic input beam was then prepared and the crystal placed to intercept the edge of the beam. The effectiveness of channeling after passing through the crystal and propagating a certain distance to a detector was investigated. Beyond the crystal a quadrupole magnet was then added to the code to improve separation of channeled particles from the remainder of the beam to improve the experimental setup. Another code written at SLAC [5], which employs a different method of computation for determining the particles' positions and angles, was then compared to the results of Yazynin's code. This work focuses on channeling of positrons as the process is less efficient with electrons due to their being attracted by the nuclei.

MATERIALS AND METHODS

The main tool used for analysis of positively charged particles passing through a bent crystal was the FORTRAN code CRY_AP written by Igor Yazynin and modified by Robert J. Noble [4]. The code uses a truncated Monte Carlo method for determining the position and angle of the particle as it exits the crystal. It includes the volume reflection, volume capture, dechanneling and channeling effects. Nuclear and multiple scattering are also taken into account.

This code was modified to make a more realistic comparison to experimental conditions in FACET. The divergence of the beam was added in to the initialization of the position of the particles. A drift space was added so that after passing through the crystal, the particles propagated five meters and their angle and position were calculated at that point to determine what a detector would measure. Furthermore, instead of initializing the particles with a random number generator, a datafile was constructed with an initial position and angle assigned to each particle with a phase space ellipse similar to what would be expected in the accelerator. This input beam was then implemented into the code and positioned so only the edge of the beam was intercepted by the crystal. These particles were also propagated in vacuum space.

In FACET it would be possible to propagate the particles further from the crystal in order to allow more separation between the main beam and the channeled portion, but they would pass through a quadrupole magnet. Therefore the code was adjusted to take into account the manipulation of the beam due to the magnet. The transform matrix was obtained from a magnet-lattice model of FACET.

The relevant FACET parameters were calculated for positrons and are included in Table 1. As well, the critical angle for channeling in silicon (110) crystals was calculated using the formula

$$\Theta_c = \sqrt{\frac{2 * U_{max}}{pv}}(1 - \frac{R_T}{R_m})$$

where U$_{max}$ is the maximum potential energy, p is the momentum, v is the velocity, R$_T$ is the Tsyganov radius and R$_m$ is the bending radius. The critical angle was found to be 42.6 μrad.

The second code performs detailed tracking to decide the particle's angle and position. It uses many steps through the crystal and averages the atomic potential along the planes according to the Lindhard Potential [1].This code does not calculate multiple scattering effects on the particles.

RESULTS

The angular profile plot for the crystal was used primarily for the analysis. The deflection of the particles as a function of the crystal orientation, the angle the crystal planes make with the incoming beam, is plotted. For a 10mm long crystal with a 10m bending radius with multiple scattering effects turned off in the code, a basic view of the amorphous region, volume reflection, and channeled particles is outlined in green in Fig. 1. When multiple scattering is included, the beam width increases significantly as seen by the red data points, and the volume capture and dechanneled particles are clearly observed.

Upon implementing the propagation of the particles to a detector, a histogram was created in order to see the distribution of the particles. For the amorphous region, in Fig. 2 the histogram shows a peak of particles centered around the middle, with no displacement, with a certain width of the distribution of the beam due to scattering. Fig. 3 illustrates a slice of the crystal orientation which includes volume reflection and volume capture, seen as two clearly defined peaks of intensity. In the channeling orientation of the crystal, the channeled particles have the highest intensity relative to the rest of the beam (Fig. 4).

In order to optimize the crystal parameters, the same analysis was done for three different bending radii and crystal lengths but the same overall bending angle. In Fig. 5 is shown the angular plots overlaid for a 10m/10mm, 5m/5mm, 2.5m/2.5mm bending radius/crystal length. These particles were also propagated 5m after passing through the crystal and their distribution is plotted in Fig. 6.

A phase space plot of the new input beam is shown in Fig. 7. This beam was offset so that only a portion would pass through the crystal, which was placed at the zero position, and the rest of the beam would not be affected by it. The portion traveling through the crystal is highlighted in green. Fig. 8 shows the same phase space plot of the particles at the exit face of the crystal, for a crystal of a 5m bending radius and a length of 5mm. The distribution of particles is plotted in Fig. 9 upon propagating 5m. Then in place of just propagating 5m in vacuum space, after the crystal the entire beam was sent through a quadrupole magnet (analogous to the what is in place in the beamline at FACET) and the distribution of the particles after the magnet is plotted in Fig. 10.

Using the code developed at SLAC [5], the result of the angular profile was compared to what was observed with Yazynin's code. The angular profile plot in Fig. 11 demonstrates that there are strong similarities; it should be noted that the vertical line parallel to the main beam in the figure is an artifact of the code dependent on the number of steps taken.

DISCUSSION AND CONCLUSIONS

The angular profile plots match the overall form expected from previous research [2] with the volume reflection, capture, dechanneling and channeling effects taken into account in the code. Comparing the angular profiles in Fig. 1, the effect of multiple scattering is easily seen. The beam width increases significantly from a narrow beam without the multiple scattering. As well, the volume capture effect and dechanneling are dependent on scattering. The scattered particles far from the main part of the beam (i.e. when the absolute value of the deflection is greater than approximately 0.5mrad) are largely due to nuclear scattering and large angle multiple scattering.

Using this angular profile to determine the orientation of the crystal to observe different channeling effects, the distribution of particles upon propagating 5m past the crystal was investigated using a crystal with a bending radius of 10m and a length of 10mm. In the amorphous region, where the beam interacts minimally, the distribution in Fig. 2 shows a single peak with a certain width which can be attributed to the scattering effects. Upon changing the orientation of the crystal to observe volume capture and volume reflection in Fig. 3, the distribution changes significantly. The wider peak on the left representing the volume reflected portion of the beam illustrates that it would be difficult to use a detector to observe this effect since the displacement of these particles from zero is minimal compared to the width of the peak. There is some noise between the two peaks which reduces the clarity of the peaks. The volume captured particles have a higher intensity with a narrower displacement distribution than the volume reflected particles.

At channeling orientation of the crystal, channeling of the particles is easily detected in Fig. 4. There is a small peak in the distribution where the beam is minimally deflected. The intensity of the channeled particles is observed around a deflection of 5mm. However, there is significant noise between these two peaks due to the dechanneled particles. In order to minimize this effect and optimize channeling, different bending radii and crystal lengths were compared. As observed in Fig. 5 the beam width (the amount of multiple scattering) decreases with crystal length. As well, the density of dechanneled particles can be seen to be less with the 2.5m bending radius than the 5m or 10m bending radius since the particles are less dense in that area on the plot. When using Fig. 5 and Fig. 6 to compare the different bending radii, one can observe the decreased multiple scattering with a decrease in the bending radius. The intensity of channeled particles almost doubles as the bending radius goes down by a factor of four and the angle of deflection for these particles remains constant. The noise due to dechanneled particles between the two peaks does decrease with the shorter bending radius. This would improve the ability to cleanly detect the channeled particles in an experimental setup. However, mechanical constraints will hinder the feasibility of attaining a crystal with a very small bending radius. Therefore, the remaining simulations used a 5mm long crystal with a 5m bending radius.

The code originally initializes the particles with a fixed position and angle then adds a randomized element using a random number generator. To better simulate the beam that would be in FACET, a data file with particles assigned positions and angles that make a phase space ellipse as seen in Fig. 7 was created. Instead of passing the entire beam through the crystal, since the crystal would be used for collimation, the beam was offset from the crystal so only a portion of the beam actually passes through as illustrated with the green portion of the beam. At the exit face of the crystal the main part of the beam remains unchanged but the portion going through the crystal is subject to dechanneling, channeling, volume capture/reflection as is illustrated in Fig. 8. One can see that of the part of the beam going through the crystal, the particles are mostly either not deflected, volume reflected or channeled with some dechanneling occurring. Since the majority of the beam is not passed through the crystal there is a high intensity of particles that are not deflected in Fig. 9. The channeled particles represent less than 5% of the beam but are still detectable at a deflection of approximately 5mm.

Beyond where the crystal would be placed in FACET is a quadrupole magnet. This would affect the beam and so using the input beam in Fig. 7, again with the portion highlighted in green actually going to the crystal, the entire beam was then passed through the magnet and the particle distribution plotted in Fig. 10. Similar to Fig. 9, the majority of the beam is unaffected causing a large peak close to zero. The channeled particles are deflected approximately 20mm with less than 5% channeled.

A comparison was made between the angular profile of a beam with no multiple scattering, with the entire beam going through the crystal, and the deflection as a function of crystal orientation between the code developed at SLAC (Fig. 11) and Yazynin's code. The volume reflection and channeling can be easily observed. The range of orientations of the crystal for which volume reflection occurs is approximately 1mrad, comparable to the range seen in Fig. 1. Further modifications need to be made to the code, including removing the artifact vertical band parallel to the beam and possibly including multiple scattering effects.

Bent crystals have applications for beam collimation. A particle beam can be passed through a crystal and depending on the orientation of the crystal, various effects can be observed. A shorter bending radius along with a shorter length of crystal improved the detection of channeled particles after they have propagated 5m past the crystal. Propagating the particles through the quadrupole magnet increases the distance the channeled particles are deflected. The placement of a detector could also be investigated further to improve the experimental setup. As well, the code could be modified to include negatively charged particles. It appears that a bent crystal could be used to channel a portion of the positron beam in FACET.

ACKNOWLEDGMENTS

This research was made possible by the United States Department of Energy Summer Undergraduate Laboratory Internship (SULI). I would like to thank SLAC National Accelerator Laboratory and Stanford University for hosting me this summer. Lastly I would like to thank Uli Wienands for providing me with this interesting project and mentoring me throughout my time at SLAC.

REFERENCES

[1] S. Møller "High-energy channeling – applications in beam bending and extraction". Nucl. Instrum. Meth. A, vol. 361, no. 3, 403-420, July 1995

[2] W. Scandale, "High-Efficiency Volume Reflection of an Ultrarelativistic Proton Beam with a Bent Silicon Crystal," Phys. Rev.Lett., vol. 98, no. 15, 154801, Apr. 2007

[3] R.A. Carrigan, "Beam extraction studies at 900 GeV using a channeling crystal". Phys. Rev. ST Accel. Beams, vol. 5, no. 4, 043501, Apr. 2002

[4] R.J. Noble, "Proton-Nucleus Scattering Approximation and Implications for LHC Crystal Collimation," Nucl. Instrum. Meth. A, vol. 623, no. 3, 872-882, Nov. 2010

[5] U. Wienands, SLAC National Accelerator Laboratory, private communication

FIGURES

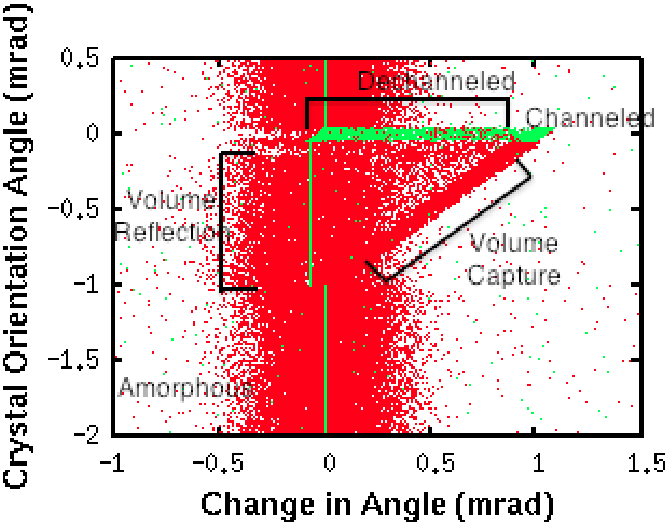


Figure 1: Angular profile of the crystal representing the angular deflection of the particles as a function of the crystal orientation. The red points represent when multiple scattering is included and the green points are when it is not included in the calculations. The crystal in 10mm in length and has a 10m bending radius. The labels indicate the effects that occur at various crystal orientations. The crystal is aligned with the beam at an orientation angle of 0.

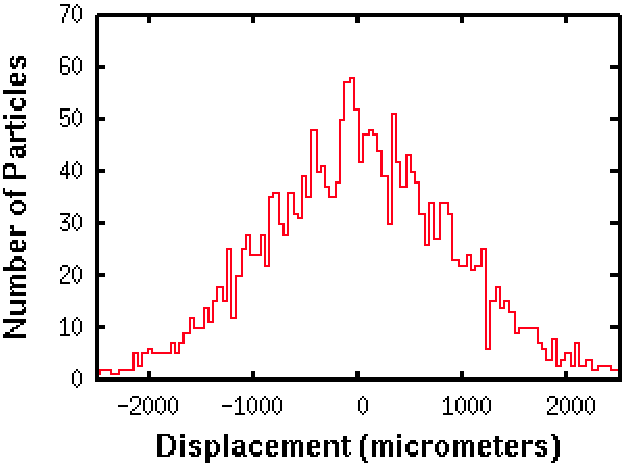


Figure 2: Histogram of the displacement of particles after propagating five meters beyond the crystal in vacuum space. The orientation angle of the crystal is from 0.2mrad to 0.3mrad representing the amorphous region.

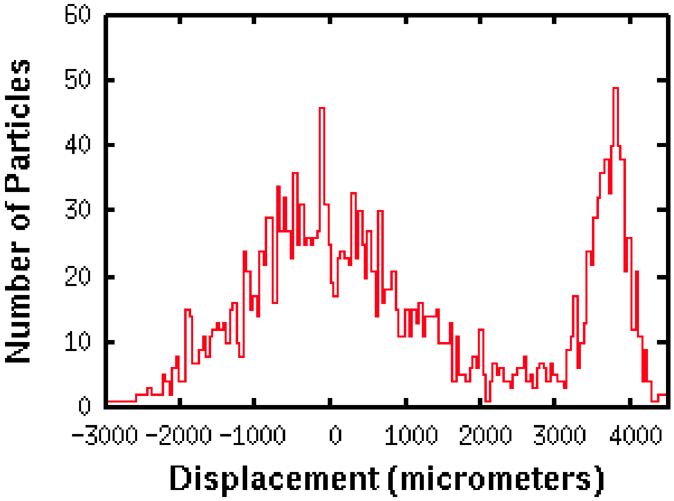


Figure 3: Histogram of the displacement of particles after propagating five meters beyond the crystal in vacuum space. The orientation angle of the crystal is from -0.3mrad to -0.2mrad representing a region with volume reflection (peak on the left) and volume capture (peak on the right).

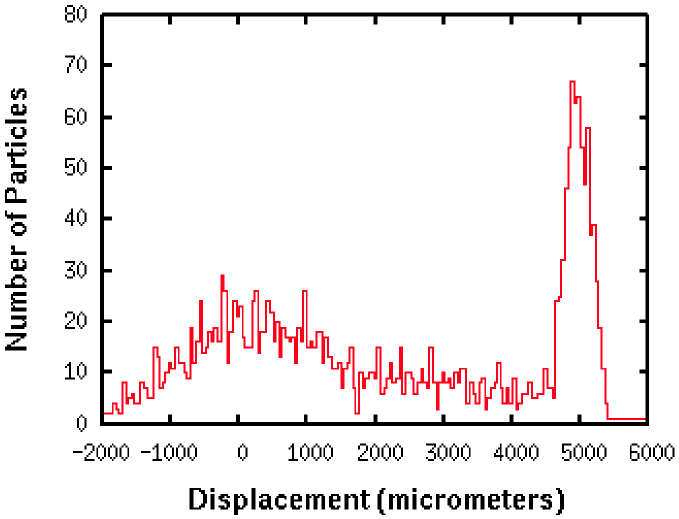


Figure 4: Histogram of the displacement of particles after propagating five meters beyond the crystal in vacuum space. The orientation angle of the crystal is from -0.051mrad to 0.049mrad. The peak on the right is channeled particles and the unaffected particles are centered around 0.

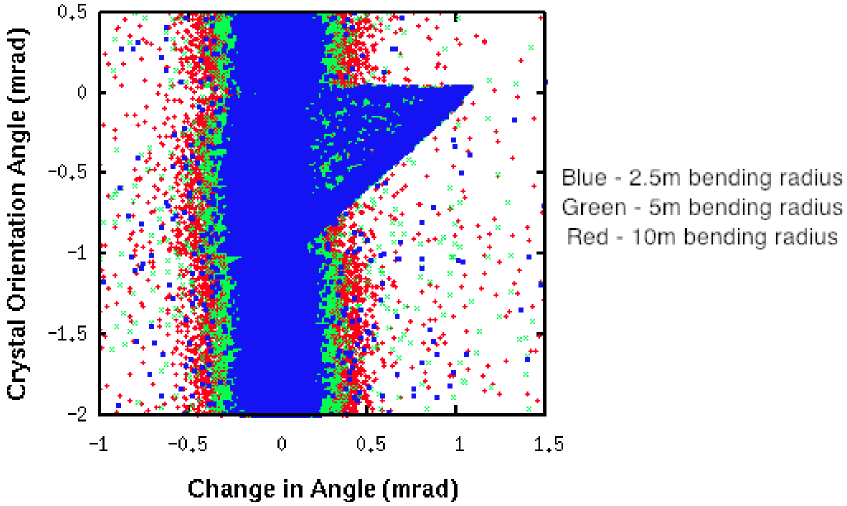


Figure 5: Angular profile with 10m/10mm, 5m/5mm, 2.5m/2.5mm bending radius/crystal length. Demonstrates the decrease in multiple scattering as the bending radius decreases.

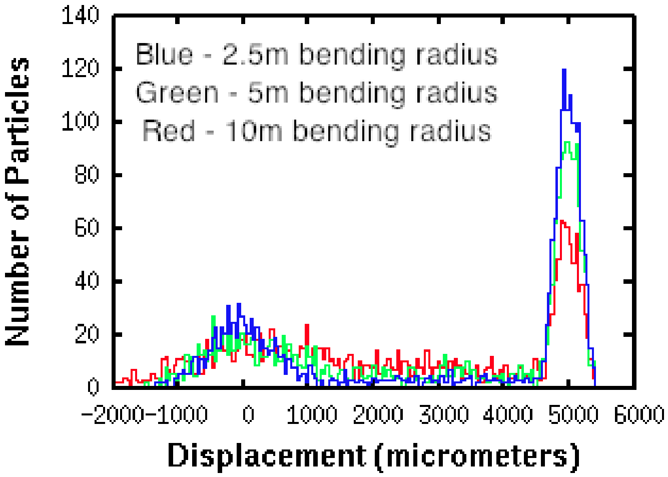


Figure 6: Histogram of the displacement of particles after propagating five meters beyond the crystal in vacuum space. The orientation angle of the crystal is from -0.051mrad to 0.049mrad corresponding to the channeling region. Shown for three different bending radii and crystal lengths.

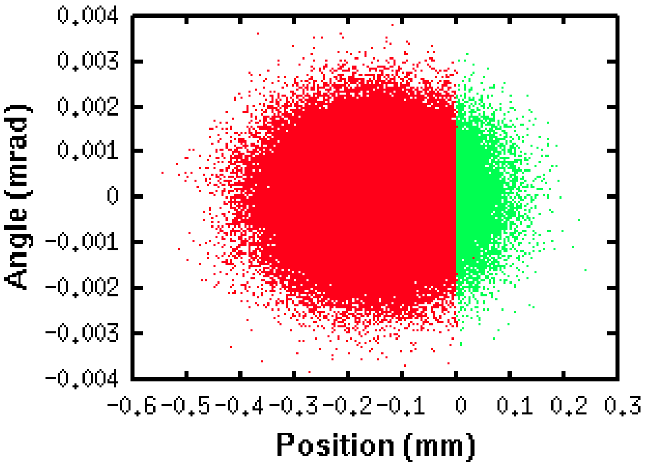


Figure 7: Phase space of angle vs position for the input positron beam. The crystal is placed at position=0mm. The beam is offset and only the portion highlighted in green passes through the crystal. The remainder of the beam propagates unaffected.

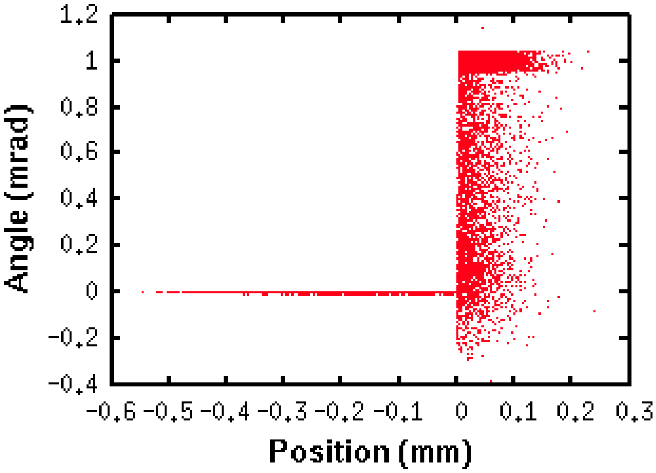


Figure 8: Phase space of angle vs position of positron beam at exit face of the crystal. Note the scale difference from Figure 7.

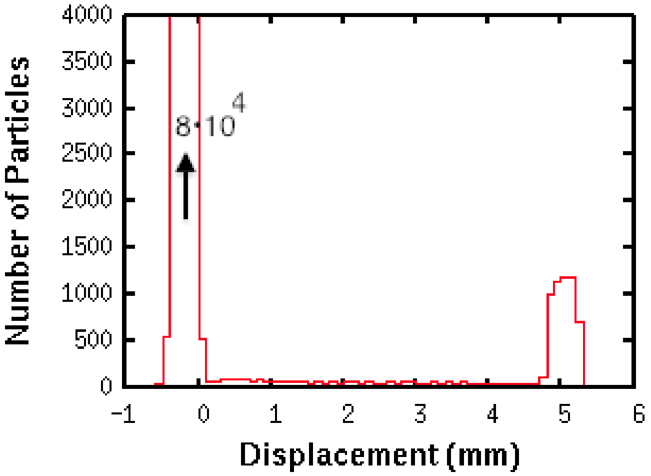


Figure 9: Histogram of the displacement of particles after propagating five meters beyond the crystal in vacuum space with only a portion of the beam passing through the crystal as illustrated in Fig. 7. The majority of the beam is not displaced, the left peak is cut off due to the scale of the graph but reaches its maximum at approximately 80 000 particles. The peak on the right represents the channeled particles.

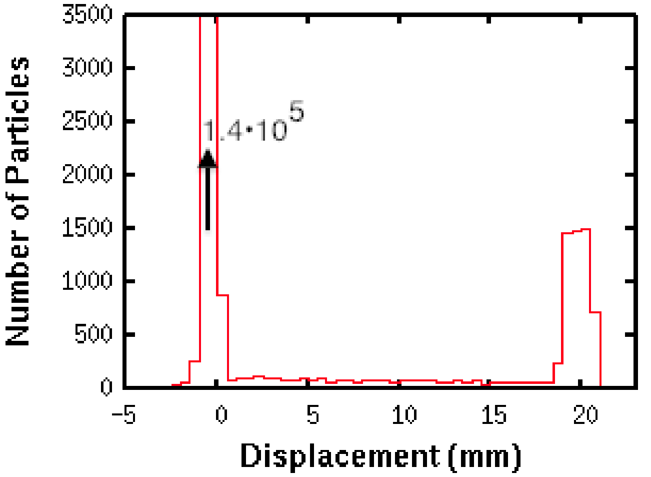


Figure 10: Histogram of the displacement of particles after propagating through a quadrupole magnet after a small portion of the beam passes through the crystal as illustrated in Fig. 7. Note the scale. The peak on the right represents the particles that were channeled and the remainder of the beam remains unchanged as represented in the peak on the left.

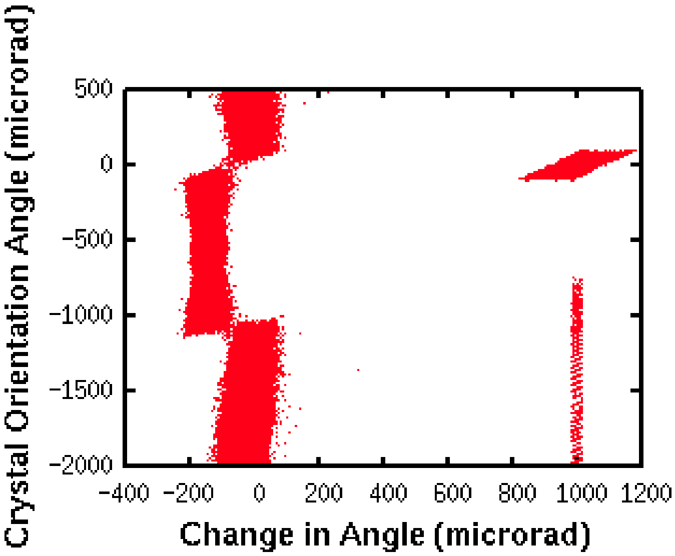


Figure 11: Angular profile of the crystal representing the angular deflection of the particles as a function of the crystal orientation. The crystal is aligned with the beam at an orientation angle of 0. This plot is the result of the code developed at SLAC. The vertical line parallel to the beam at the change in angle = 1000microrad is an artifact of the code dependent on the number of steps. This plot was made using 64000 steps.

TABLES

Table 1: FACET Parameters

Energy	23 GeV
Emittance	3 μm·rad
Lorentz Factor	45 010
Beam Width	81.6 μm
Divergence	0.82 μrad

SLAC-TN-12-010

Magnet Lattice Design for the Transmission of Power Using Particle Beams

Daniel Marley

Office of Science, Science Undergraduate Laboratory Internship (SULI)
North Carolina State University
SLAC National Accelerator Laboratory
Menlo Park, CA

12 August 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of James Welch at the SLAC Accelerator Directorate.

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Magnet Lattice Design for the Transmission of Power Using Particle Beams. DANIEL MARLEY (North Carolina State University, Raleigh, NC 27695) JAMES WELCH (SLAC Accelerator Directorate, Menlo Park, CA 94025)

As the amount of electricity generated by renewable energy sources continues to increase, the current method of power transmission will not serve as an adequate method for transmitting power over very long distances. A new method for transmitting power is proposed using particle beams in a storage ring. Particle beams offer an incredibly energy efficient alternative to transmission lines in transmitting power over very long distances. A thorough investigation of the magnet lattice design for this storage ring is presented. The design demonstrates the ability to design a ring with stable orbits over a 381.733 km circumference. Double bend achromats and FODO cells are implemented to achieve appropriate β functions and dispersion functions for 9-11 GeV electron beams.

INTRODUCTION

Although the manner for transmitting power in the United States is well established, there still exist some shortcomings of using transmission lines to efficiently transmit power over very long distances. This paper evaluates the current state of the electrical power system (the grid) and explores the idea of using particle beams for power transmission with a discussion of the magnet lattice that has been designed for such an endeavor.

Background Information

Electrical power is transmitted from power plants to consumers through various changes in voltage over multiple power lines. Transmission lines transfer most of this power over very long distances, at a few hundred kilovolts, from the power plants to substations. From the substations, lower-voltage (a few kV) lines carry power to primary consumers and then pass through another step-down transformer (120 and 240V) before entering housing developments. The transmission line is a critical piece of the electrical grid in that it first receives the raw power from the generators and is responsible for distributing the power at a high efficiency to substations. The voltages at which these lines operate include 138, 230, 345, 500, and 765 kV. These transmission lines currently retain an average of about 93% efficiency and increasing the voltage is one way of improving the efficiency by minimizing resistive losses [1]. However, they still grow approximately linearly with distance and eventually become an unfavorable form of power transmission [2]. This property of transmission lines hinders the ability to share excess power with distant areas experiencing a power deficit. Transmission lines are also becoming less capable of providing enough power to serve the increased demand from consumers [3].

Motivation & Purpose

In recent years there has been an ongoing shift from fossil fuels to renewable energy sources. The two main types of renewable energy sources of interest are nuclear and solar power because they provide the most promise for widespread distribution. Nuclear power is already a popular energy source with 32 states in the continental United States (US) hosting at least one nuclear reactor [4]. While most of these reactors are located in the eastern half of the US, there is a surge of solar power production in the western half of the US, with California producing the most of any state [5]. Nuclear reactors, however, are being relocated to different areas of the country that contain a less dense population. As the efficiency and power production of these forms of renewable energy sources continue to grow, and investment in transmission lines continues to decline, there is an ever-increasing need to transmit the power over very long distances. Solar power is generated largely in the desert of the southwestern US where the population density is very low. Because nuclear and solar power plants are being placed in locations were very few people live, there is a need to transmit the power they produce over very long distances. Transmitting this power over very long distances (from the power plants where no one lives to areas of high population density) motivates us to propose a new method for transporting power across the country, and possibly between continents. Fig. 1 depicts a schematic of this process and shows possible locations for generators and loads.

Particle accelerators are routinely used in high-energy physics experiments around the world. The particle beams that make these accelerators useful are tremendously energetic (up to 7 TeV at the Large Hadron Collider) and typically run from a few MeV to a few GeV, or five orders of magnitude greater than that of current transmission lines. Along with the tremendous voltages at which particle accelerators operate, they are also incredibly energy efficient. The primary source of energy loss is synchrotron radiation, which occurs when the particles undergo centripetal acceleration in a storage ring, described by (for electrons):

$$P_{lost} = 88.5\frac{E^4 I}{2\pi\rho}\Delta\theta \tag{1}$$

with the energy, E, measured in GeV, the current, I, in amperes, and the radius of curvature, ρ, in meters. Incorporating particle beams into the grid will allow for cost-effective power distribution over incredibly long distances (e.g., around the United States or from North to South America). To adequately determine the feasibility and viability of using particle beams for power transmission, multiple factors need to be researched: the current condition of the grid, the energy efficiency of particle beams, the cost for materials and operation, the status of other proposed technologies compared to that of particle beams, and the design of the arrangements of magnets, or lattice, to focus and steer the beams. The work presented explores the magnet lattice design for the future electron storage ring for particle transmission.

METHODS

To design the magnet lattice for an electron storage ring, the transfer matrix method was applied to generate ring designs and analyze the electron stability. The following section describes the mathematical formalism of the transfer matrix method applied to design the magnet lattice and the software used to calculate beam stability and the lattice design. The lattice design was created for an example ring that nearly encloses the United States, joining the Eastern, Western, and Texas interconnects. The size of this ring represents the magnitude and scope of this project to efficiently transmit power over very long distances.

Transfer Matrix Method

In an electron storage ring, electrons remain in orbit due to magnetic forces from dipole, quadrupole and even sextupole magnets. These magnets steer and focus the electrons around the ring with an approximately Gaussian distribution about the ideal orbit. To quantify the behavior and properties of the electrons, a unique coordinate system has been defined with respect to the ideal orbit: $\hat{s}$ along the beam line, $\hat{x}$ in the radial direction, and $\hat{y}$ in the transverse direction, as noted in (Fig. 2) [6].

The lattice, or series of magnets and drift spaces (areas free of magnetic forces) that define an accelerator, design in this ring currently contains only quadrupoles and dipoles, although sextupoles can be added later if deemed necessary. The effect that these magnets have on electrons can be calculated using K, the focusing function:

$$K_x \equiv \frac{1}{\rho^2} - \frac{ec}{E}B_1(s) \qquad \& \qquad K_y \equiv \frac{ec}{E}B_1(s) \tag{2}$$

with $B_1(s) = \partial B_y/\partial x$, e the electron charge, c the speed of light, and E the energy of the particle of interest [7]. The focal length for the quadrupole magnets is $f = 1/(Kl)$, with l the length of the quadrupole. The focus/defocus character of quadrupole is a result of $\nabla \times \mathbf{B} = 0$ so $dB_x/dy = -dB_y/dx$. That is a positive gradient in one direction implies a negative gradient in the other. In the limit of small oscillations, the linear equations of motion for an electron in an accelerator can be formed and then solved to reveal the position and phase functions based on the focusing function and structure of the lattice. The linear equation of motion is $x'' + K_x(s)x = 0$ (can substitute x with y and the derivatives are taken with respect to s) and the corresponding position solutions are [7]

$$x(s) = \begin{cases} a\cos\left(\sqrt{K}s + b\right) & K > 0 \\ as + b & K = 0 \\ a\cosh\left(\sqrt{|K|}s + b\right) & K < 0 \end{cases} \tag{3}$$

The solutions produced will combine into a pseudo-periodic function that does not repeat itself exactly, but does follow an approximate sine/cosine shape.

For a given set of initial conditions, $x(s_0)$ and $x'(s_0)$, a matrix equation can be written that describes the current position, $x(s)$, and phase, $x'(s)$, functions due to the beam elements (quadrupole, dipole, or drift space)

$$\mathbf{x}(s) \equiv \begin{pmatrix} x(s) \\ x'(s) \end{pmatrix} \qquad \mathbf{x}(s) = \overleftrightarrow{\mathbf{M}}(s, s_0)\mathbf{x}(s_0) \tag{4}$$

The form of this equation provides the necessary information for determining the transfer matrix, $\overleftrightarrow{\mathbf{M}}$ for each beam element.

$$K > 0: \quad \overleftrightarrow{\mathbf{M}}(s, s_0) = \begin{pmatrix} \cos(\sqrt{K}l) & \frac{1}{\sqrt{K}}\sin(\sqrt{K}l) \\ -\sqrt{K}\sin(\sqrt{K}l) & \cos(\sqrt{K}l) \end{pmatrix} \tag{5}$$

$$K = 0: \quad \overleftrightarrow{\mathbf{M}}(s, s_0) = \begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix} \tag{6}$$

$$K < 0: \quad \overleftrightarrow{\mathbf{M}}(s, s_0) = \begin{pmatrix} \cosh(\sqrt{|K|}l) & \frac{1}{\sqrt{|K|}}\sinh(\sqrt{|K|}l) \\ \sqrt{|K|}\sinh(\sqrt{|K|}l) & \cosh(\sqrt{|K|}l) \end{pmatrix} \tag{7}$$

For the sector dipole magnet of length l, $K > 0$ $(= 1/\rho^2)$ and the matrix simplifies to

$$\overleftrightarrow{\mathbf{M}}(s, s_0) = \begin{pmatrix} \cos(l/\rho) & \rho\sin(l/\rho) \\ -\dfrac{1}{\rho}\sin(l/\rho) & \cos(l/\rho) \end{pmatrix}. \tag{8}$$

For n beam elements, Eq. (4) can be rewritten as

$$\mathbf{x}(s) = \Big[\overleftrightarrow{\mathbf{M}}_1(s_1, s_0) \cdot \overleftrightarrow{\mathbf{M}}_2(s_2, s_1) \cdot \overleftrightarrow{\mathbf{M}}_3(s_3, s_2) \cdots \overleftrightarrow{\mathbf{M}}_n(s, s_{n-1})\Big]\mathbf{x}(s_0). \tag{9}$$

This matrix product results in a unit matrix [7]. Not only does this help calculate final position and phase functions, but also the beam stability is contained within this matrix. If the Trace of the final matrix is less than or equal to 2 ($|\mathrm{Trace}(\overleftrightarrow{\mathbf{M}})| \leq 2$) there exists a stable orbit. This result is derived from the requirement that the eigenvalues of the matrix satisfy the equations: $\lambda_1 = 1/\lambda_2$, $\lambda_1 + \lambda_2 = \mathrm{Trace}(\overleftrightarrow{\mathbf{M}})$, and $\lambda^2 - \mathrm{Trace}(\overleftrightarrow{\mathbf{M}})\lambda + 1 = 0$ for either eigenvalue [9]. This has to be calculated for both x and y to determine stability in both directions.

Simulation Process & Beam Parameters

With this theoretical formalism in place, *Mathematica* 8 was used to compute the final matrix produced from the beam elements. Because the trace of a matrix is easily determined, the stability of the beams can be calculated in a quick and efficient manner. This allows for the calculation and frequent manipulation of beam parameters to fully optimize the design and explore many options in very little time. To calculate the final matrix, the parameters of the beam need to be defined. These parameters are collected in (Table 1).

After determining the beam stability in *Mathematica*, the Methodical Accelerator Design (MAD) software (v 8.52), developed by CERN, was used to input the lattice design and determine the β functions in x and y, along with the dispersion relation and an output of the coordinates of each beam element. The β functions are very important in accelerator physics and are derived from the general solution to the linear equation of motion $x'' + K_x(s)x = 0$:

$$x(s) = \sqrt{\varepsilon\beta}\cos(\phi - \theta) \tag{10}$$

where ε is the emittance of the beam (area in phase space) and β describe the lateral focusing properties of the lattice. From this equation one can see the maximum amplitude of the beam is $\sqrt{\varepsilon\beta}$ ($\equiv \sigma$, the standard deviation). For this lattice, the amplitude should remain below 1 mm to ensure the beam does not become too wide and begin to be lost to the enclosure and other nonlinear effects. MAD can determine more complex features of the beam (nonlinear terms, beam position monitors, horizontal and vertical kickers, RF cavities, etc.) than *Mathematica*, and allows the user to input constraints on the system. MAD is also capable of determining stability, but the location of the instability is unknown and difficult to find. Being able to input constraints allows the user to have some freedom in writing a lattice because MAD can calculate the optimal values. However, in some instances the determined values are not ideal, and it is best to re-compute them using *Mathematica*.

It was pre-determined that there would be two main lattice elements for the beam: Double bend achromat (DBA) and a FODO cell. The best mechanism for controlling the dispersion of the beam was to implement a DBA. Two dipole magnets with a focusing quadrupole magnet, in $\hat{x}$, placed between them comprise a DBA. The usefulness of a DBA is that it prevents electrons of different energies that travel through the dipole magnet from taking a different path by eliminating dispersion outside of the DBA. As electrons come through the first dipole, they are refocused by the quadrupole to pass through the second dipole and emerge back in the electron bunch as if there had been no bend at all (Fig. 3). Outside of a DBA it is optimal to include either a doublet or a triplet of quadrupoles to ensure the beam is not lost and that it remains stable while entering and exiting a DBA [9]. Multiple drift cells are needed to cover the large distances planned by this ring. A cell that consists of only drift sections and quadrupoles is known as a FODO cell (a segment of a storage ring that is symmetric about its center), outlined in Fig. 4. These two design elements were implemented into *Mathematica* and MAD to yield a stable orbit of electrons around a storage ring.

RESULTS

After applying the transfer-matrix formalism in *Mathematica* and then inputing the results into MAD, three separate cell designs were combined to create a single lattice design. Two stable cells were created that incorporated a smaller bend radius ($\equiv$ Cell 1) and a larger bend radius ($\equiv$ Cell 2) through two different DBAs, and a third FODO cell ($\equiv$ Cell 3) was created. The plots of the β functions can be seen in Fig. 5,6 & 7.

Fig. 5 (Cell 1) represents the DBA with a smaller radius, Fig. 6 (Cell 2) the DBA with a bigger radius, and Fig. 7 (Cell 3) an example drift section. Table 2 displays the maximum β functions in x and y, the total length, and maximum dispersion for each Cell. The β functions, dispersion, and total length vary between the different cells because they have different quadrupoles, drift spaces, and dipoles. Both Cell 1 and 2 have multiple quadrupoles implemented into the design. They have different dipoles form one another to achieve different bend angles as well. Because Cell 3 is just a FODO cell, the quadrupoles all consist of the same strength and the same separation between them. From the values used to create these cells, it is possible to see the wide variety in values necessary to gain stability in the electron orbit simply based on the type of DBA needed.

After the files were created and the stability was verified in *Mathematica*, MAD was used to combine these cells into one final ring design (Fig. 8). The ability to vary quantities and set constraints in MAD made it possible to match β functions across the various cells.

Constraints were also applied to ensure the dispersion remained zero outside of the DBAs and to attempt to keep the β functions below 2000 m. A stable beam was created that is 2355.38 m long, a maximum dispersion of 0.099596 m and a maximum β_x and β_y of 2188.89 m and 2254.96 m, respectively. These values are not below the desired maximum, but exceed that value by less than 15% each.

DISCUSSION AND CONCLUSIONS

From the results it was possible to find an electron storage ring that would yield stable electron motion and cover approximately 381.733 km. This length is short of the desired 10,000 km, but with further work on the MAD program 10,000 km should be easily achievable by implementing more FODO cells and multiple DBAs. The power of MAD to iterate through the beam elements and vary them such that the constraints could be met allowed for the design to be created.

The desired parameters for this lattice made it particularly difficult in finding a large enough ring that produced a stable orbit. As the lattice became longer and longer, the β functions became larger and larger. The size of the FODO cells were restricted by the size of the β functions. The DBAs also became affected by the β function size as well. Meeting the dispersion requirement was completed by varying the strength of the quadrupole in the DBA; the β function would grow too large if the quadrupole was too far or too close to the surrounding dipoles. For future designs, allowing the β functions to increase may make it possible to more quickly find a stable orbit for the ring. Also, as a result of the given parameters, the electron beam will contain greater than 300 MJ worth of energy. This is a tremendous amount of energy traversing the beam and how this energy should be dissipated needs to be addressed in future beam designs. Regardless of the difficulties in meeting the pre-determined parameters, MAD made it much easier than expected to find a stable orbit because of its ability to vary parameters and constrain variables.

This lattice design is not the only possible design. If more constraints are revealed, or if some can be removed, the values that were formed for the quadrupoles, dipoles, and drift sections may be changed to yield different lattice designs. Some of these possible constraints may include the strength of permanent magnets that can be manufactured, the width of the pipe and tunnel, and the cost of tunneling, among others. From here there are multiple steps that should be taken to turn this example lattice into the final design including nonlinear terms, RF cavities, vertical bends to match the curvature of the earth, and creating more complex DBA cells. It is also important that the terrain be more thoroughly analyzed and a ring be designed that can more efficiently navigate the landscape.

With this design, a first step in demonstrating the possibility of this project has been taken and it proves the feasibility in at least achieving stability in a ring of this magnitude. Transmitting power over very large distances has gained plausibility from this design and encourages the exploration of other factors in developing this technology, such as the cost of construction, manufacturing, cost of operation, and economic advantage over current transmission lines and other competing technologies.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy Office of Science through the Summer Undergraduate Laboratory Internship (SULI) Program. I am very appreciative of my mentor Jim Welch for not only the opportunity to work on this interesting and very challenging project, but for also helping me continuously throughout the summer and better exposing me to accelerator physics. I would like to thank the director, Steve Rock, for providing a positive learning environment, and for exposing me to many new areas of physics.

REFERENCES

[1] U.S. Energy Information Administration. (2009, November 19). Data on electricity transmission and distribution losses. Retrieved from http://www.eia.gov/tools/faqs/faq.cfm?id=105&t=3

[2] L. Paris, G. Zini, M. Valtorta, G. Manzoni, and N. De Franco (1984). Proceedings from International Conference on Large High Voltage Electric Systems, 1984 Session: *Present Limits of Very Long Distance Transmission Systems.* Paris, France: Global Energy Network Institute.

[3] Schewe, P.F. (2007). *The Grid: A Journey Through the Heart of Our Electrical World.* Washington, DC: Joseph Henry Press

[4] U.S. Nuclear Regulatory Commission. (2011, May 19). Operating Nuclear Power Reactors. Retrieved from http://www.nrc.gov/info-finder/reactor/

[5] Gelman, R., Hummon, M., McLaren, J., and Doris, E. (2010, October). *2009 US State Clean Energy Data Book.* U.S. Department of Energy: Washington, DC.

[6] Sands, M. (1970). The Physics of Electron Storage Rings: an introduction. Prepared for the US Atomic Energy Commission under contract no. AT(04-3)-515.

[7] Lee, S.Y. (1999). *Accelerator Physics.* River Edge, NJ: World Scientific Publishing Co.

FIGURES

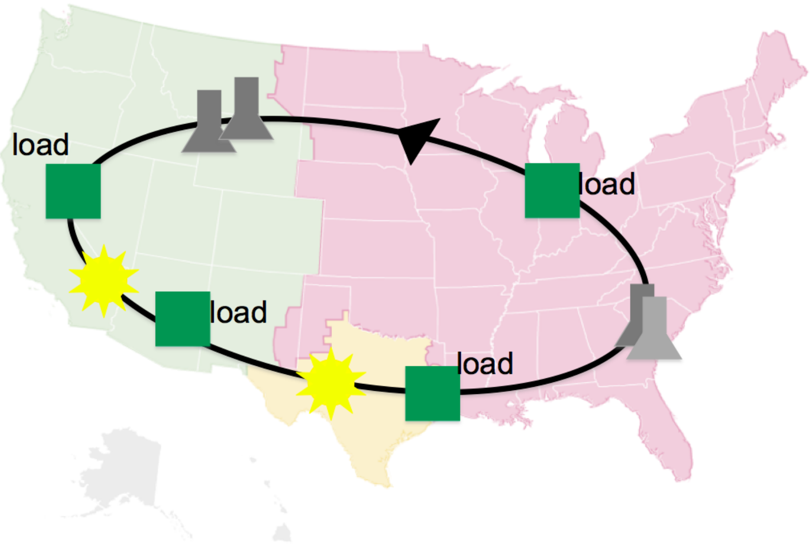


Figure 1: Outline of final storage ring with possible generator and load locations. Overlaid on a map of the US with the 3 interconnects colored: East (Pink), West (Green), Texas and (Orange).

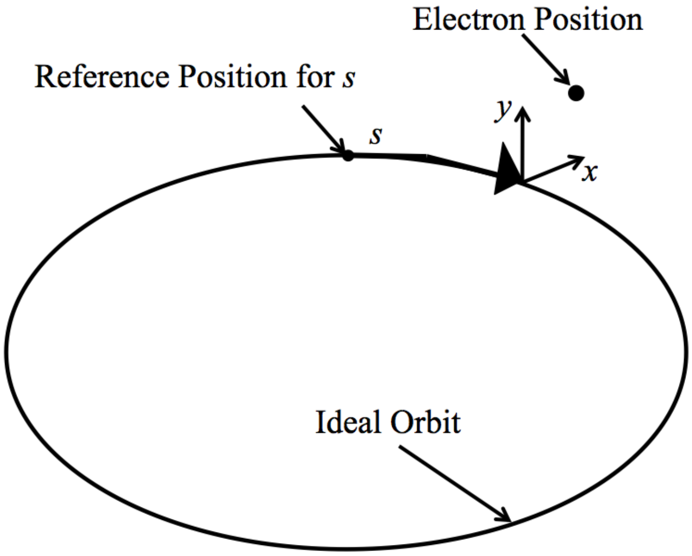


Figure 2: Schematic depicting the coordinate system for electrons in a storage ring.

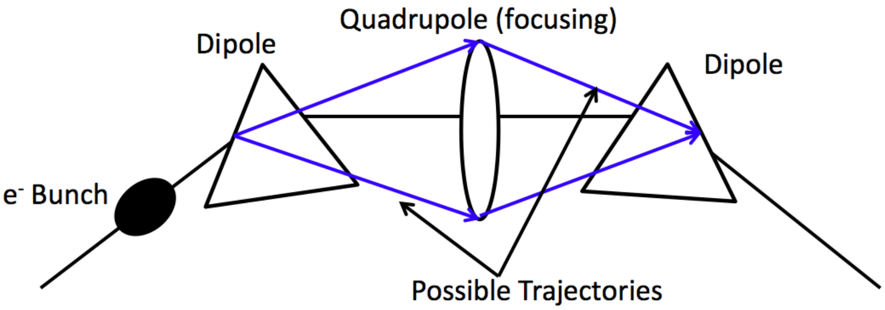


Figure 3: Schematic of a double bend achromat (DBA).

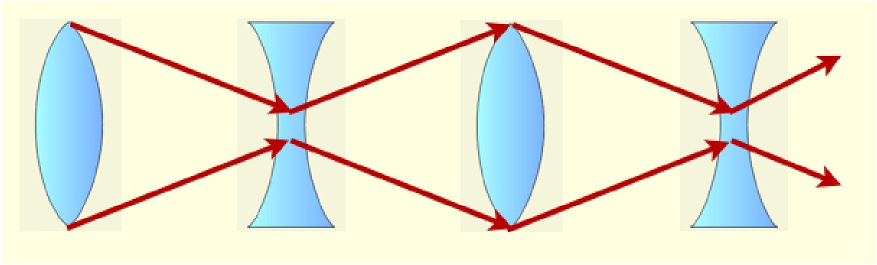


Figure 4: Diagram of a FODO lattice using optical lenses to represent quadrupoles. The spaces between each lens represent the drift spaces in a lattice.

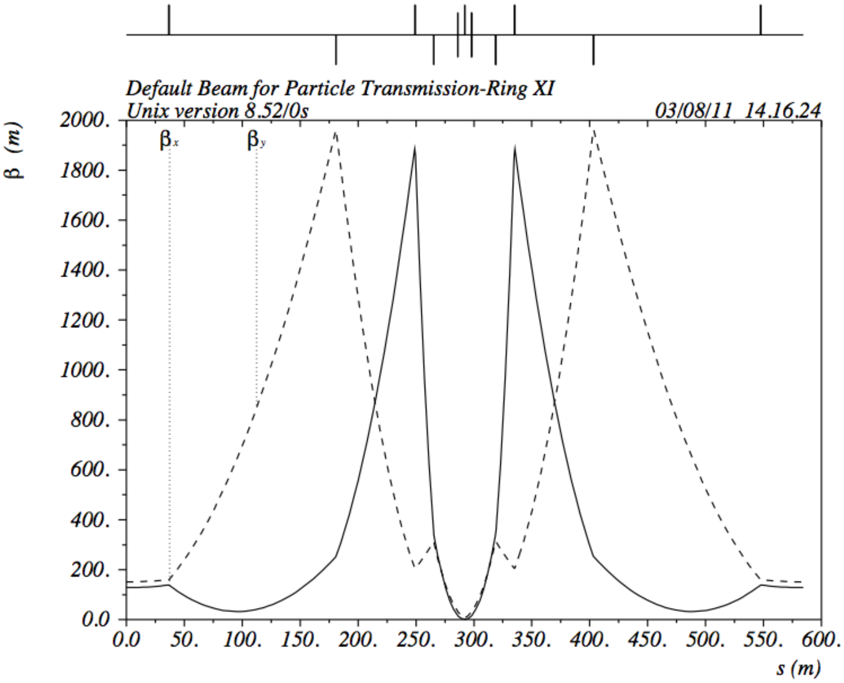


Figure 5: β function for stronger DBA, smaller radius.

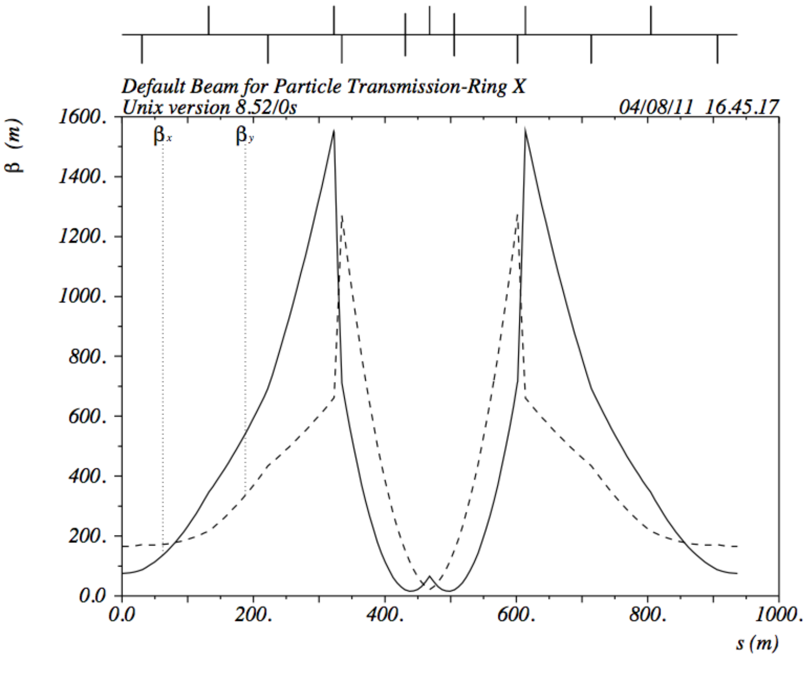


Figure 6: β function for weaker DBA, larger radius.

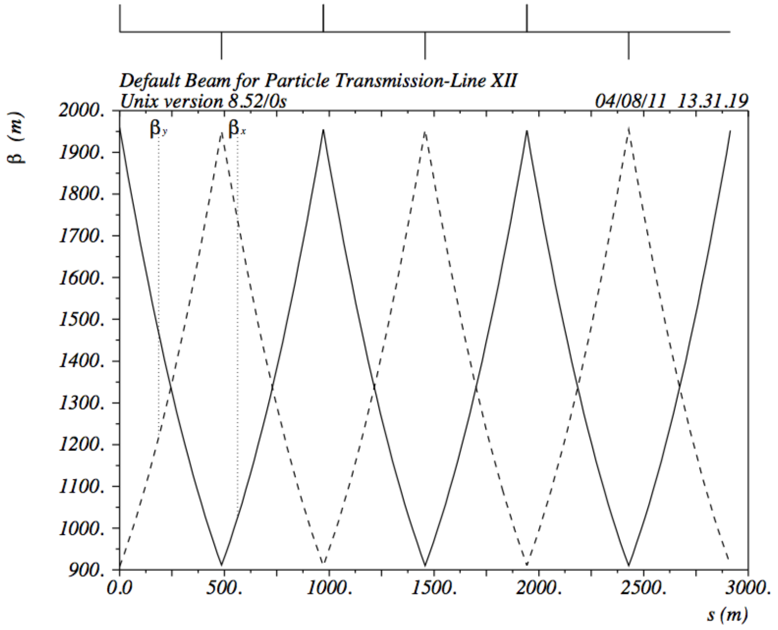


Figure 7: β function for FODO cell.

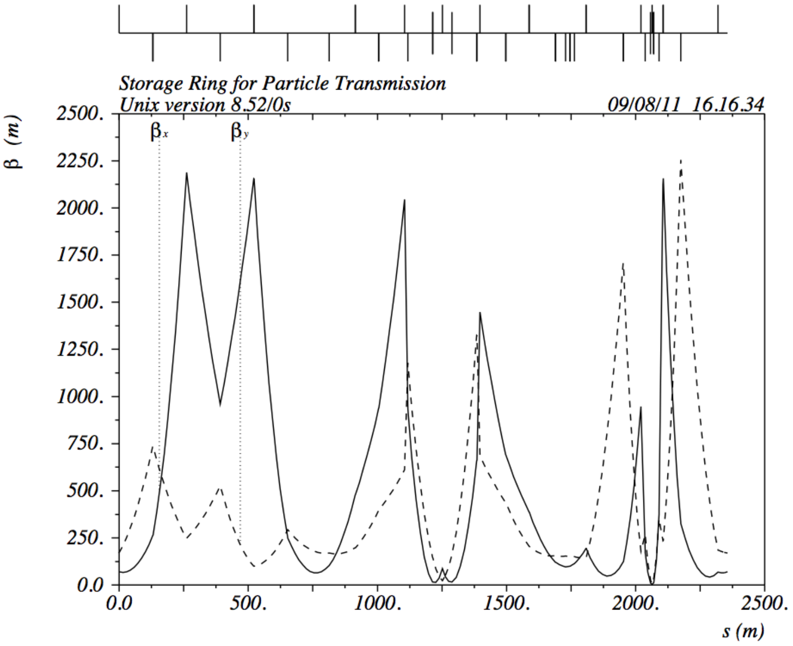


Figure 8: β function for combined cell.

TABLES

Parameter	**Value**	**Parameter**	**Value**
Circumference	10,000 km	Dispersion (max)	0.1 m
Beam Energy	10 GeV	Dipole Radius (min)	100 m
Beam Current	1 A	Dipole Field	0.1 T
Emittance (ε)	5×10^{-10} m	Quadrupole Gradient (max)	10 T/m
$\beta_{x,y}$ max	2000m	σ	$\sim$1 mm

Table 1: Parameters for the storage ring.

Cell	β_x max. (m)	β_y max. (m)	Dispersion max. (m)	Total Length (m)
1	1884.85	1960.75	0.099424	584.24
2	1551.37	1271.54	0.099595	936.50
3	1954.72	1954.72	0.00	2914.0

Table 2: Maximum β and dispersion values and the total length for the Cells in the magnet lattice.

SLAC-TN-12-009

Bunch Profiling Using a Rotating Mask

Mitchell Miller

SLAC National Accelerator Laboratory

Summer Undergraduate Laboratory Internship (SULI) Program

08/19/2011

Illinois Institute of Technology Physics 2012

Mentor: Alan Fisher

Work supported by US Department of Energy contract DE-AC02-76SF00515.

SLAC National Accelerator Laboratory, Menlo Park, CA 94025

Abstract

The current method for measuring profiles of proton bunches in accelerators is severely lacking.   One must dedicate a great deal of time and expensive equipment to achieve meaningful results. A new method to complete this task uses a rotating mask with slots of three different orientations to collect this data. By scanning over the beam in three different directions, a complete profile for each bunch is built in just seconds, compared to the hours necessary for the previous method. This design was successfully tested using synchrotron radiation emitted by SPEAR3. The profile of the beam was measured in each of the three desired directions. Due to scheduled beam maintenance, only one set of data was completed and more are necessary to solve any remaining issues. The data collected was processed and all of the RMS sizes along the major and minor axes, as well as the tilt of the beam ellipse were measured.

INTRODUCTION

The Large Hadron Collider (LHC) is the premier high energy physics experiment, using some of the most advanced technology on the planet. This colossal machine, 27 kilometers around, now collides 3.5 TeV protons (7 TeV center of mass) and will collide at twice this energy in 2014, the highest collision energy ever achieved. At the opposite end of the scale spectrum, the proton beam is only a few microns across at an interaction point (IP, where the beams collide in the four detectors around the ring), requiring extremely precise control and measurements. It is, therefore, very useful to directly measure the beam cross-section. One would imagine that this task could prove very difficult, but with the clever application of a few pieces of basic hardware, one can easily achieve this goal at a low cost.

Outside the detectors, the transverse size of the beams is much larger. At the injection energy, 450 GeV, the beam size is approximately 1 mm RMS (root mean square, the standard deviation) and at full energy this size drops to 0.3 mm. As the protons fly around the enormous ring, they, like all charged particles, emit synchrotron radiation. When low energy charged particles are bent by an outside influence, such as a magnetic field, they emit radiation with an intensity that varies with the cosine of the angle to the direction of travel. However, when the particle is accelerated to near the speed of light, this wide spread is focused into a narrow, forward beam. The spectrum is also shifted to higher energies. For electrons the midpoint of the emission spectrum for a bending magnet is called the "critical frequency" follows:

$$\omega_c = \frac{3}{2}\gamma^3 \frac{c}{\rho} \tag{1}$$

Where ω_c is the critical frequency, γ is the relativistic factor, c is the speed of light, and ρ is the radius of the bending magnet. The x-ray beam is commonly used for experiments in a wide range of disciplines, from condensed matter to biological studies, but synchrotron radiation, particularly the visible emission, is often used to understand the profile of the beam of particles that created it. Optics can image the synchrotron light onto an ordinary CCD to provide video of the transverse profile of the beam, but this requires electron beams or very high energy proton beams, and these still often need image intensifiers. Measurement of these profiles usually requires a camera with a gate faster than the spacing between bunches. The camera is placed in the beam of synchrotron light, at an image point of the particle beam, and records the profile generated by each proton bunch. These bunches traveling around the ring are separated only by a few nanoseconds, requiring an incredibly fast shutter. In this very brief period of time, very little light is gathered by the camera. To compensate for this, many exposures are taken over many separate passes of the same bunch. This means that several seconds are necessary to obtain meaningful statistics. Since there are hundreds or even thousands of bunches in most accelerator rings, measuring a complete profile of every bunch can take several hours. The LHC is a particularly difficult case, with 2808 bunches spread over 89 microseconds with 25 nanoseconds between each bunch. With a typical integration time of two seconds per bunch, a full measurement of every bunch would take nearly two hours. This is obviously not ideal, since the bunch size can change drastically in that time.

One alternative to this method significantly cuts down on the time necessary to take data, but only provides a mathematical description of the bunch cross-sections, not a tangible image. By passing a rotating mask consisting of three slots of different orientation, one horizontal, vertical, and one at a 45° angle, in front of the beam, one can build an intensity profile in the three different dimensions. From the light that passes through the slot, one can find the Gaussian shape of the bunch that created it. With this information in three dimensions, the shape and tilt of the overall Gaussian ellipse can be calculated.

Each bunch returns to the optical system every 89 microseconds. During this time, the wheel has only moved a small amount. This provides a very large number of data points to reconstruct the shape of the bunch. Using simple computer software, one can index each point to a specific bunch and then reconstruct all of the bunches separately from a single set of three slots. This entire process can be conducted in less than a second, improving performance time by orders of magnitude.

MATERIALS AND METHODS

The main materials necessary for this project are a photomultiplier tube, rotating mask, and a driving motor. To test this method, visible light from the SPEAR3 storage ring at SSRL was used in place of LHC synchrotron radiation. More details regarding these components are found below.

PHOTOMULTIPLIER TUBE

This project requires a small, fast photomultiplier tube to collect intensity data from bunches. One other necessary feature of this device is resistance to radiation damage. The entire apparatus being developed in this project will eventually sit in the tunnel at the LHC. This is a very extreme environment, with very high levels of ambient radiation from the accelerator. The current choice for PMT is the Hamamatsu R7400U. The "U" specifies a fused-silica window rather than glass to avoid radiation damage, since glass is easily darkened by the ambient radiation that is present in the LHC tunnel. This PMT is small, less than 20 mm. in any dimension, and also has a very fast response time, just below 2 ns[1]. Since the radiation being measured is entirely in the visible spectrum, no additional equipment, such as scintillators, is necessary.

ROTATING MASK

Another very important component of this project is the rotating mask. An earlier version of this component was designed and tested by another student working under Dr. Fisher in a previous SULI summer program[1]. The wheel has a radius of 100 mm with slots centered at 80 mm. There are four sets of three slots, each cut so that they will be at precisely the right orientation as they pass over the image of the beam, which intercepts the wheel at a 45° angle from its X or Y axis. The slots are laser cut to be approximately 15 microns in a very thin foil which is then sandwiched between two plates of aluminum to protect it. The design schematics for this component are shown in Fig. 1.

OSCILLOSCOPE

To record the data taken from a scan of the mask, an oscilloscope with a very long memory and high sample rate was used. The Tektronix DPO4104 was chosen on account of its wide range of features. This oscilloscope is able to scan five gigasamples per second and record ten million points in its memory. The high resolution, resulting from the sampling rate, allows for signals from individual bunches to be resolved it high detail. Since the signal resolution is so high, however, a large amount of memory is needed to record all the data from one slot.

MOTOR

The final mechanical component of this project is the motor that turns the mask. The motor chosen is the Maxon Sensor DC Motor 118748. This motor is then attached to a gear-head with a 53:1 reduction factor. This provides the motor with a huge range of speeds, allowing for a great deal of versatility in the number of data points taken. Additionally, a resolver head is attached to the rear of the motor which outputs angular position in the form of a cosine and sine signal. These two signals can be converted to an angular position using simple software to track which slot is providing a signal at a given time. An example of this output is shown in Fig. 2.

SSRL

Due to the high levels of ambient radiation in the LHC tunnel which this apparatus will eventually be used in, it is convenient to test the setup in a safer and more accessible environment. The Stanford Synchrotron Radiation Lightsource (SSRL) provides an excellent opportunity to test the behavior and diagnostic abilities of the rotating wheel system. At SSRL, electrons orbit a storage ring at 3 GeV, creating extremely intense beams of radiation. This radiation ranges from infrared to hard x-rays. For this experiment, the range of radiation was restricted to visible light, since that is the only type available for detection at the LHC. Although the energy range is vastly different, most other beam properties are similar between the LHC and SSRL. Both have bunches of particles that emit radiation, which is the central focus of this project. However, one major drawback is that SSRL has no signal digitizer that can record data from each bunch as it passes in real time. Therefore, at SSRL, the data must be recorded and processed at a later time. This results in a very large data file that is difficult to work with.

One should note that there are significant differences between the light measured at SSRL compared to the ultimate goal of diagnostics at LHC. The intensity and energy of the radiation produced at SSRL is orders of magnitude higher than that of the LHC. Since this light was so intense, several filters were necessary to prevent the PMT used in this project from being overwhelmed. The saturation of the PMT caused a major distortion in the output signal. This was caused by insufficient time between an successive signals for the capacitors within the

PMT to recharge. To remedy this issue, the rotational speed of the wheel was increased, causing the PMT to be exposed to light for a shorter time, the driving voltage of the PMT was decreased, reducing the output gain, and a filter was placed in the beam path, decreasing the amount of light being measured. All of these changes, although resolving the PMT discharge issue, significantly reduced the signal intensity. Therefore, an amplifier was used to magnify the output signal. Additionally, the time necessary for a bunch to travel around the storage ring at SSRL is approximately 100 times faster than at the LHC. This makes the signal analysis much more difficult because the peaks generated by each bunch are compressed very closely together.

METHODS

The test of this apparatus was conducted on the diagnostic beam line at SSRL. This is a beam line that is limited to only visible radiation. A large optical table was shared with two other experiments to provide a stable, organized test area. A light tight box was constructed around the mask and PMT housing to reduce ambient light. This setup is shown in Fig. 3. The light coming from the beam line was first focused with a large lens to a point approximately two meters away. This lens results in a demagnification of the image by a factor of 8. Along this axis, several mirrors were inserted at different points so each experiment could acquire the light as well as a neutral density filter to reduce the amount of light reaching the PMT, preventing overload. The light for this project was then passed through another lens to magnify the image by a factor of 3.25, resulting in a net magnification of 0.4063.

In order to obtain a precise focus and magnification measurement, an electronic camera was placed at the exact point where the mask would be located. A glass slide with a USAF resolution test pattern, consisting of a series of lines with different sizes, was placed in the path of the beam. Using software written in MatLab by Jeff Corbett, both the beam and this image were accurately focused. The test pattern also provided a very precise measurement of the magnification power of the optical set up.

After the focus and magnification were successfully measured, a light-tight enclosure was built to ensure that only photons from the synchrotron source were being recorded. The PMT was isolated in a specially designed box and mounted to an optical breadboard. The motor and wheel were then carefully aligned so that the PMT would focus at 45° from the horizontal axis of the wheel. This ensured that the angles of the slots were exactly 0°, 90° and 45° when they passed in front of the PMT. Another layer of light reduction was then put in place using black cardboard and tape to further seal the PMT from ambient light. In addition to all of these measures, the room in which the experiment was conducted was also dark during data collection.

Finally, to collect data, a high voltage power supply applied -600 volts to the PMT. A DS 345 signal generator provided a 10 kHz sine wave signal to the motor resolver, allowing for position measurements. Four signals were then recorded using the oscilloscope: the resolver drive signal, the cosine output, the sine output, and the PMT signal. A slight modification to the wheel was made in the form of dark tape placed over all but one slot. The record stored by the oscilloscope was then limited to the time interval around the passage of just one slot. To resolve the fast PMT signals, the data was digitized at two gigsamples per second. A full wheel rotation would create a record of billions of points. By examining one slot at a time, the resolution was maintained while being limited by the oscilloscope's ten megasample memory. This insured that position of the slot was known very accurately, since the data was not being processed in real time. Under normal working conditions, the sampling rate at the LHC will be such that only the relevant data is accessed. However, at SSRL, this is not the case, so a large amount of extraneous data is recorded, making precise position measurements difficult.

The final step of this project was the development of a Python script to process these signals. The latest version of the script reads all of the data, and then finds the average angular velocity of the wheel using the cosine, sine, and resolver input signals. Since the cosine and sine signals are simply amplitude modulations of the drive signal, finding the peak of the drive signal and checking the value of the cosine and sine signals will give an accurate angular position.

The script then scans the PMT data for the peak generated by the timing bunch, which is a large, isolated bunch of electrons used for diagnostic purposes. By isolating each of the timing bunch peaks, a Gaussian profile is built that corresponds to that bunch only, allowing for the imaging of single bunches. This can be repeated with any of the bunches, but the timing bunch is the simplest for testing purposes. The profile from each scan were then fit to find the size of the beam in real space.

RESULTS AND DISCUSSION

This project has successfully shown that this technique is a viable method for measurement of individual bunch profiles. A scan in each of the three axes was successfully obtained. From that data, the Gaussian profiles were fit and their sizes were obtained. Figure 4, 5, and 6 each display the scans after being converted into real space with position on the x-axis and intensity on the y-axis. The Y profile scan is particularly interesting due to the non-gaussian shape. This abnormality is most likely the result of improper focusing of the beam or aberrations caused by the optics used to focus the beam.

After fitting a Gaussian peak to each of these profiles, several beam parameters were extracted for each including height and σ. These values for each axis are shown in Table 1. The two sets of values for σ result from two different calculation methods. The first, labeled as 'automated', uses the motor resolver head signal to calculate angular velocity. Due to the sampling method used in this project discussed above, this has a large amount of error. One can see in Fig. 2 that the exact angular position is not well defined, so this measurement, especially when automated with computer software, is not particularly accurate. The second value uses a hard coded value for angular velocity. This value was measured using an oscilloscope to find the time necessary for a complete revolution of the wheel.

One should note the relatively large discrepancy between the expected values and those which were measured. This error results from diffraction effects of the beam as it passes through the numerous optical devices before being imaged as well as the incredibly high amount of precision necessary to properly focus the beam. A lens out of place by less than a millimeter can change the image size by an enormous magnitude. This is a result, however, of the current setup and not a defect with the apparatus.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Alan Fisher, for allowing me to work on this project. It was a very valuable experience. I would also like to thank Jeff Corbett for assisting me in optical alignment when Alan was not available. Finally, I would like to thank the US Department of Energy for providing funding for this project.

REFERENCES

[1] Lee, Christopher J. *Bunch by Bunch Profiling with a Rotating X-Ray Mask*. Science Undergraduate Laboratory Internship (SULI), 22 Aug. 2007. Web. 15 Aug. 2011.

APPENDIX A: Figures and Tables

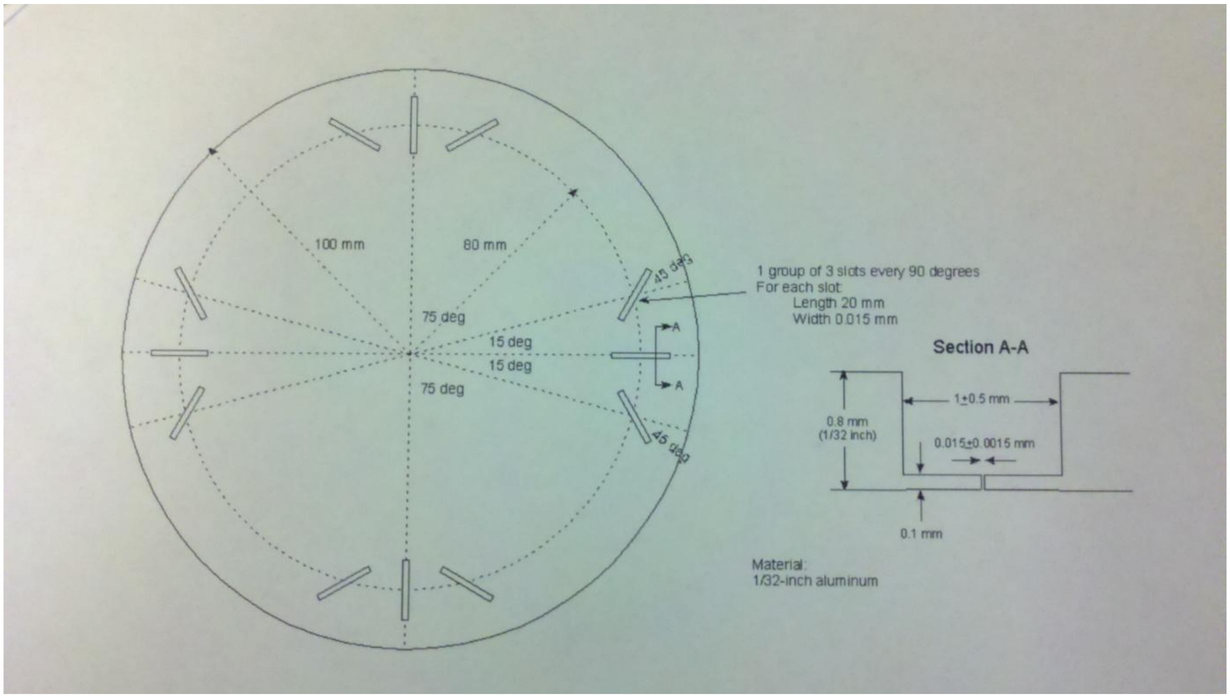


Figure 1: Design schematics for the rotating light mask. One can see four sets of slots at different angles, generating each of the three profiles.

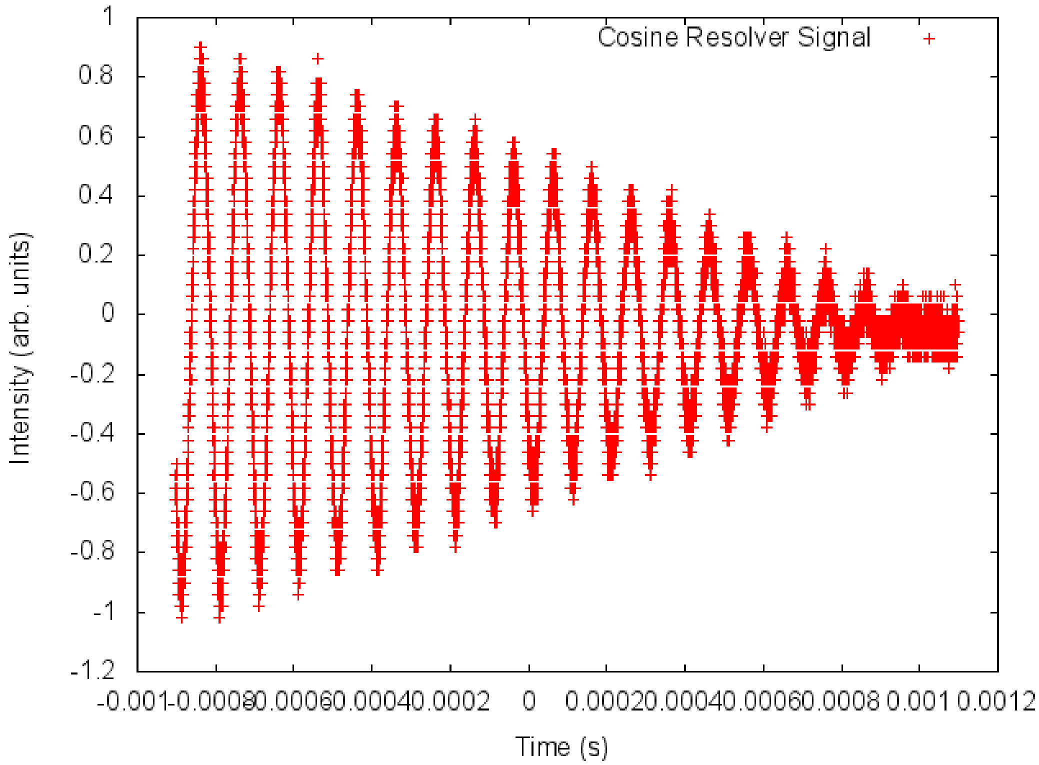


Figure 2: The cosine output from motor resolver for the X Profile scan. Note the envelope in which the 10 kHz resolver sits within.

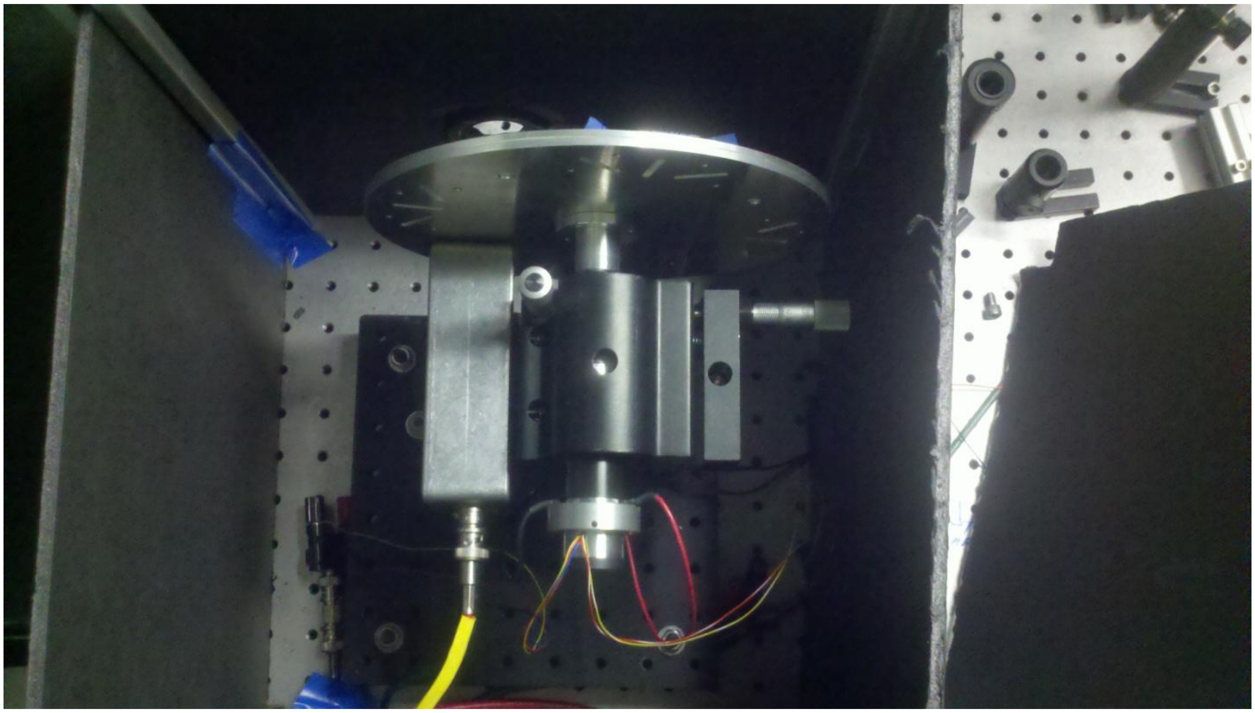


Figure 3: Light tight box setup. One can see the PMT box (large yellow cable) to the left of the motor and resolver outputs (thin, multi-colored wires) in the lower section of the image.

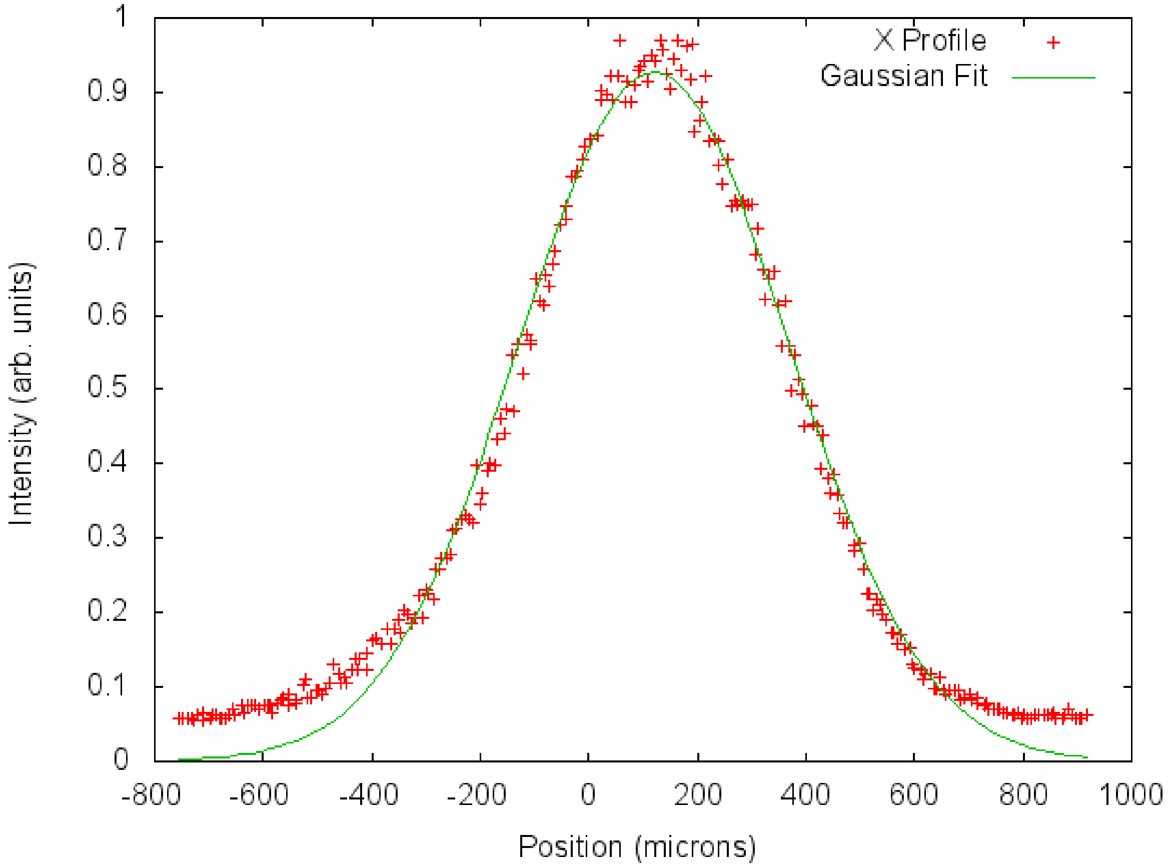


Figure 4: The X profile scan taken on the SSRL Diagnostic Beamline showing the Gaussian fit.

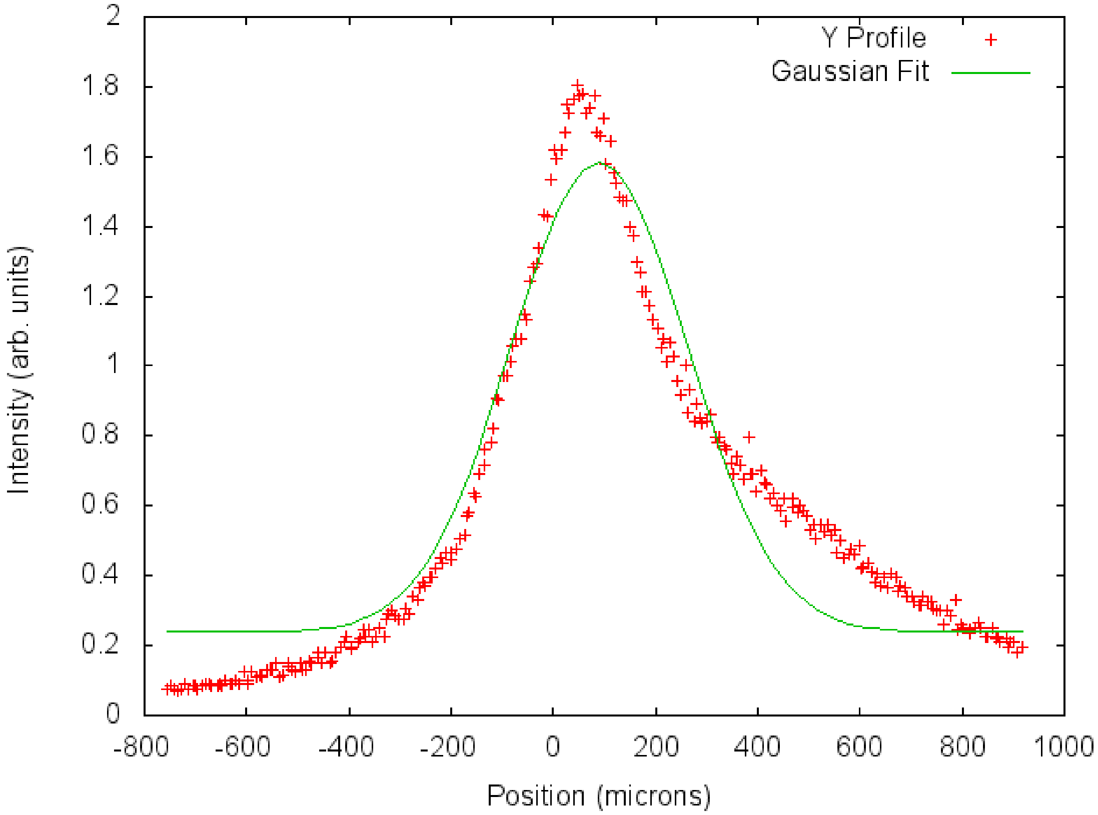


Figure 5: The Y profile scan taken on the SSRL Diagnostic Beamline showing the Gaussian fit.

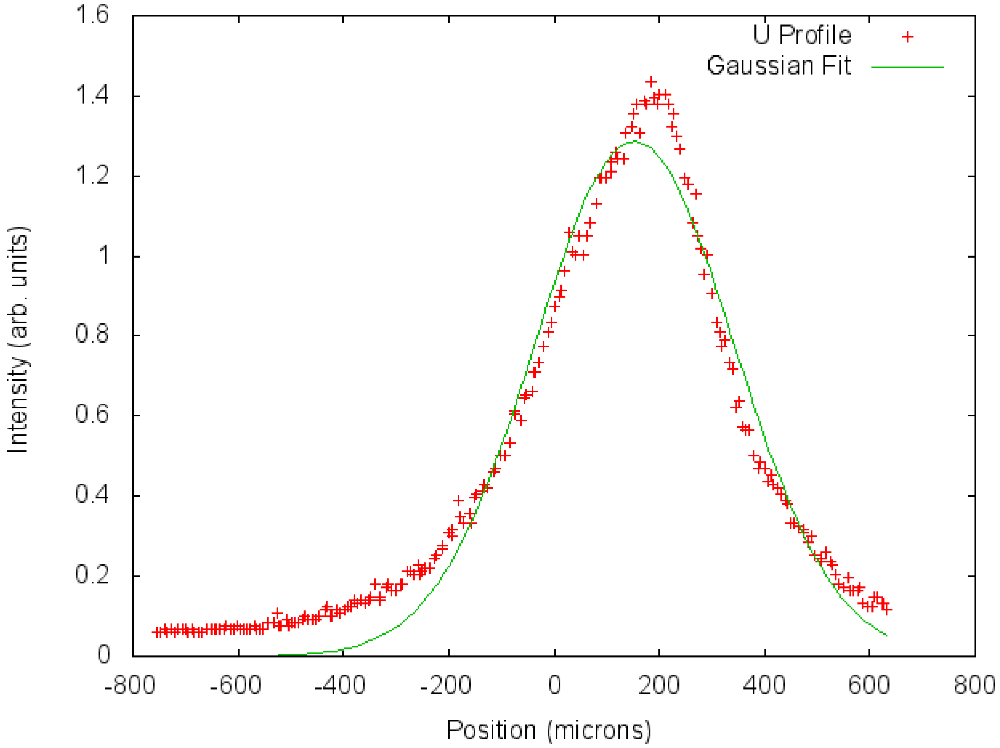


Figure 6: The U profile scan taken on the SSRL Diagnostic Beamline showing the Gaussian fit.

Table 1: Profile σ values

Profile	Expected σ	Measured σ (automated)	Measured σ (measured ω)
X	60.94	321.98	234.76
Y	20.31	282.22	169.98
U	Unknown	262.86	189.13

APPENDIX B: Code

```
# Read data from scope containing 5 signals
# Converts time to distance using angle resolver
#
# Data is read using numpy.genfromtxt which requires a file open in
# 'binary read' mode.  The cosine signal is then scanned for maxima
# to build a carrier frequency profile.  This is used to calculate
# an average angular frequency for use in time to position conversion
# Next, the time and intensity data are read and maxima are found.
# These points correspond to the 'mondo' bunches that were observed
# during data collection.  The time scale is then converted to
# position using the previously calculated angular frequency.  This
# data is all saved to a text file for further use.
#
# Mitch Miller
# 7/22/11

import sys
import numpy as np
import math as m

##pathList = ["E:\TestProfile.csv"]
##pathList =
[r"E:\XProfile_600V_1ND_721.csv",r"E:\YProfile_600V_1ND_721.csv",r"E:\UProfile_600V_1ND_721.csv"]
pathList = ["/media/2423-A37F/XProfile_600V_1ND_721.csv","/media/2423-
A37F/YProfile_600V_1ND_721.csv","/media/2423-A37F/UProfile_600V_1ND_721.csv"]

for path in pathList:

    ## Define Variables
    numberOfPeaks = 256  ## 256 for full scan
    inFile = open(path,'rb')
    outPath = path.replace('.csv','.peaks.csv')
    anglePath = path.replace('.csv','.anglepeaks.csv')
    outFile = open(outPath,'wb')
    angleOutFile = open(anglePath,'wb')
    max = 0
    sinMax = 0
    cosMax = 0
    counter = 1
    peakCounter = 0
    omegaCounter = 0
    omegaPeaks = 20  ## 20 for full scan
    gearRatio = (64. / 3375.)
    maxCounter = 0
    pointCounter = 0
    maxArray = np.array([[0,0,0,0,0,0]])
    angleArray = np.array([[0,0,0]])
    radius = 80000.  ## In microns
    omega = 0
    avgOmega = 0
    ## Read data from file
    sys.stdout.write('Begin Reading Data \n')
    sys.stdout.write('This will take up to 30 minutes \n')
    ##data = np.genfromtxt(inFile, dtype=None, delimiter=",")#, skip_header=16,
names=['time','cos','sin','res','inten'])  ##WINDOWS
    data = np.genfromtxt(inFile, dtype=None, delimiter=",", skiprows=16, names=['time','cos','sin','res','inten'])  ##
LINUX

    numberOfPoints = np.size(data,axis=0)




    ## Find peaks for Cos and Sin
    ## NOTE: Only cos peaks are measured directly,
    ## sin peaks may have some error.  Use cos for
    ## calculations if possible.
    sys.stdout.write('Begin Angle Peak Search\n')
    pointsPerSection = numberOfPoints/omegaPeaks
    while omegaCounter < omegaPeaks:

        counter = omegaCounter * pointsPerSection
        cosMax = 0
        sinMax = 0

        while counter < ((omegaCounter + 1) * pointsPerSection):

            if data[counter][1] > cosMax:

                cosMax = data[counter][1]
                sinMax = data[counter][2]
                time = data[counter][0]

            counter  = counter + 1

        dummyArray = np.array([[time,(cosMax),sinMax]])
        angleArray = np.concatenate((angleArray,dummyArray))
        omegaCounter = omegaCounter + 1



    ## Calculate average omega for time to distance conversion
    sys.stdout.write('Begin Omega Calculation\n')
    counter = 1
    angleArray = np.delete(angleArray, [0], axis = 0)
    # Find the maximum of the array and normalize to ensure [-1 < data < 1]
    absoluteValueArray = np.absolute(angleArray)
    arrayMax = absoluteValueArray.max(axis=0)
    if arrayMax[1] < 1:
        arrayMax[1] = 1
    if arrayMax[2] < 1:
        arrayMax[2] = 1

    while counter < omegaPeaks:

        deltaT = angleArray[counter][0] - angleArray[counter-1][0]
        deltaTheta = m.tan((angleArray[counter][2] / arrayMax[2])/(angleArray[counter][1] / arrayMax[1])) -
((angleArray[counter-1][2] / arrayMax[2])/(angleArray[counter-1][1] / arrayMax[1]))        ##deltaTheta =
m.acos(angleArray[counter][1]) - m.acos(angleArray[counter-1][1])
        if deltaTheta != 0 and m.sqrt(deltaTheta**2) < 0.25:
            omega = deltaTheta / deltaT
            avgOmega = avgOmega + omega
            pointCounter = pointCounter + 1
        counter = counter + 1

    avgOmega = avgOmega / pointCounter
    avgOmega = avgOmega * gearRatio
    avgOmega = avgOmega * avgOmega
    avgOmega = m.sqrt(avgOmega)
    sys.stdout.write('Avg Omega = ')
    sys.stdout.write(str(avgOmega))
    sys.stdout.write('\n')


    ## Save angle peaks
    np.savetxt(angleOutFile, angleArray, delimiter = ',')



    ## Scan each section for largest value and save inensity
    ## and position.
    sys.stdout.write('Begin Main Function\n')
    counter = 0
    pointsPerSection = numberOfPoints/numberOfPeaks
    while peakCounter < numberOfPeaks:

        counter = peakCounter * pointsPerSection
        max = 0

        while counter < ((peakCounter + 1) * pointsPerSection):

            ## Record point if it is larger than the previous
            if data[counter][4] > max:

                max = data[counter][4]
                time = data[counter][0]
                maxCounter = counter

            counter = counter + 1

        ## Convert time to position and record all relevant data in maxArray
        ## [Time, Position, Cos, Sin, Drive, Intensity]
        position = radius * avgOmega * time
        dummyArray =
np.array([[time,position,data[maxCounter][1],data[maxCounter][2],data[maxCounter][3],max]])
        maxArray = np.concatenate((maxArray,dummyArray))

        ## Print current peak number and omega for debug
        ##sys.stdout.write('Counter = ')
        ##sys.stdout.write(str(peakCounter))
        ##sys.stdout.write('\n')

        peakCounter = peakCounter + 1




    ## Delete first row [0,0,0,0,0,0] from array
    maxArray = np.delete(maxArray, [0], axis = 0)




    ## Save array of maximums to a text file
    ##output = ''.join('Average Omega = ',str(avgOmega),'\n')
    ##outFile.write(output)
    sys.stdout.write('Saving Data\n')
    np.savetxt(outFile, maxArray, delimiter = ',')




    sys.stdout.write('File Complete\n')
    inFile.close()
    outFile.close()
```

SLAC-TN-12-011

Measuring Cluster Relaxedness

Blythe Moreland

University of Michigan

Office of Science, Science Undergraduate Laboratory Internship Program (SULI)

SLAC National Accelerator Laboratory

Menlo Park, California

August 18, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy’s Science Undergraduate Laboratory Internship, under the direction of Risa Wechsler at the Kavli Institute for Particle Astrophysics and Cosmology at SLAC.

Participant: Blythe Moreland ___________________________________

Research Advisor: Risa Wechsler ___________________________________

Work supported in part by US Department of Energy contract DE-AC02-76SF00515.

Contents

1 Introduction

When is a dark matter halo "relaxed"? In our efforts to understand the structure of the universe, dark matter simulations have provided essential grounds for theoretical predictions. These simulations provide a wealth of ways of parameterizing and measuring the features of astronomical objects. It is these measurements on which we base comparisons of our world and our attempts to re-create it. One of the essential questions dark matter simulations help address is how dark matter halos evolve. How does one characterize different states of that evolution? The focus of this project is identifying cluster relaxedness and how it relates to the internal structure of the halo.

A dark matter simulation consists of an N-body simulation which takes an initial set of positions and velocities of the dark matter particles and evolves them under the influence of gravity [6]. Though scientists have so far not been able to detect dark matter particles, the information from these simulations is still valuable especially given the relationship between dark matter halos and galaxy clusters. Galaxies sit within dark matter halos and recent evidence points to filaments of dark matter forming the framework on which galaxy clusters grow [7]. A dark matter halo is a collapsed group of gravitationally bound dark matter particles. Subsets of bound particles form subhalos or substructures. The dark matter simulation is carried out over time - with decreasing redshift (z) or increasing scale factor ($a = \frac{1}{1+z}$). (Thus, $z = 0$ or $a = 1.0$ is present-day.) The merger history of a halo can be represented pictorally by a *merger tree.* A major merger event occurs when a structure joins the main halo with the mass ratio between it and the main halo being above a certain threshold. These events mark important points in the halo's evolution. And it is at these events that one hopes, and perhaps is more likely, to relate measures of relaxedness to this mass accretion.

Cluster relaxedness is not a well-defined concept. Rather a set of qualities are defined that one expects a "relaxed" cluster to have. One expects a relaxed halo to have a roughly isotropic density distribution. Most of the particles should be part of the main halo rather than bound in substructures. Taking into account kinematic information of the halo, one does not expect a proportion of the particles' energy in kinetic energy that goes far beyond virial equilibrium.

With our measures of cluster relaxedness, we want to investigate its relationship to these major merger events. We'll first look at how measurements of different aspects of relaxedness relate to each other and to other aspects of a halo's internal structure. We'll then look at how these measurements behave in response to major mergers.

2 Methods

2.1 The simulation

Our dark matter halos are from the Rhapsody (Re-simulated Halo Sample for Statistical Observable-mass Distribution Study) catalog [9]. The sample includes 110 halos with virial mass $\log_{10} M_{vir} = 14.8 \pm 0.05\,\mathrm{M}_{\odot}/h$ and this puts the halos in the regime of "massive" halos. The parent simulation is called Carmen of the LasDamas suite of simulations that originally ran with a 1000 Mpc/h size box and 1120^3 collisionless particles and a softening scale of 25 kpc/h. Wu's process of curating these massive halos was to select halos in a high but narrow mass bin, zoom-in on them and re-simulate them at higher resolution. Such a bin omits enough particles to be able to run for high resolution but admits a statisticially appreciable number of halos. The resulting smoothing scale is 6.7 kpc/h. This resolution gives us significant substructures and a rich, unique probe into halo evolution. The data cited in the ensuing sections is the output of the halo-finder and merger tree at 100 "snapshots" of the simulation. These snapshots are spaced equally in $\log a$, thus the time in gigayears (Gyr) between each snapshot is not constant.

2.2 The measurements

The relaxedness measurements (which will be shortened to RMs) considered here are taken from the literature and in particular from Neto's paper on CDM halo structure [4]. They are described below:

(i) *Concentration*:

Navarro, Frenk & White [3] proposed a simple model to approximate the density profile of a dark matter with two free parameters:

$$\frac{\rho(r)}{\rho_{crit}} = \frac{\delta_c}{(r/r_s)(1+r/r_s)^2} \tag{1}$$

ρ_{crit} is the critical density of the universe - the density required for a closed universe. δ_c is called the characteristic density and r_s, the scale radius. This profile has done well at parameterizing dark matter halos close to virial equilibrium. The NFW profile indicates two major regimes where different power laws dominate. This r_s tells us something about the shape of the density profile as it corresponds to the transition between the two power laws. The *concentration* measurement is defined as $c_{vir} = r_{vir}/r_s$ where r_{vir} is the virial radius of the halo. This value relates to how much of the halo is in this denser region.

(ii) *NFW fit error* :

Since the NFW profile best describes large, relaxed halos. The χ^2 value [4] used to evaluate the goodness of fit when fitting the profile may be an indicator of relaxedness:

$$\chi^2 = \frac{1}{N_{bins}-1} \sum_{i=1}^{N_{bins}} \frac{1}{w_i} [\log_{10} \rho_i - \log_{10} \rho_{NFW}(\delta_c; r_s)]^2 \tag{2}$$

Where w_i is the number of particles in bin i.

For (i) and (ii) the fitting process plays a role in determining the values. The IDL function MPFIT.pro was used to minimize the χ^2 value. The halo is binned initially in 32 bins spaced equally in $\log_{10}(r/r_{vir})$, with the innermost bin beginning when $\log_1 0(r/r_{vir}) = -2.5$

and the outermost bin ending at r_{vir}. Then bin edges were removed until the bin was at least $.02Mpc$ in width, which is around 3× the softening length. The density is spherically averaged by counting the particles within each bin and dividing by the volume of the bin. In the calculation of χ^2, the bin is weighted by the number of particles it contains. When the fitting was particularly difficult, MPFIT would sometimes return a value of 0.0 or 1.0 for 2 at low snapshot number. This seemed to be a feature of how the value was stored within the program so we ignore these values in our analysis. Lastly, the minimum number of particles to consider an NFW fit was 50 and this was the threshold used for all other RMs as well.

(iii) *Substructure mass fraction* [4]:

f_{sub} is the total mass inside substructures whose centers are within the virial radius of the center of the main halo divided by the mass of the main halo. This value should reasonably be between 0 and 1 but may rise above 1 if there is complex subhalo structure contained within the halo as well as a large subhalo whose center lies at the outskirts of r_{vir}.

(iv) *Center of mass displacement* [4]:

Take the difference between the average position of particles in the halo (the center of mass) with the position of the particle at the bottom of the potential well. A low s value describes a radially symmetric mass distribution.

$$s = \frac{|r_c - r_{cm}|}{r_{vir}} \tag{3}$$

(v) *Virial fraction* [4]:

Compute $2T/|U|$ where T is the total kinetic energy of the halo particles within r_{vir} (with respect to the halo center) and U the gravitational potential. This value addresses the dynamic state of the halo and how close it is to dynamic equilibrium.

3 Results and discussion

This section is organized as follows: Section 3.1 deals with measurements taken to establish a connection between properties of the halo at low redshift and characterizations of the full trajectory of the halo's growth. Section 3.2 focuses on behavior of RMs around major mergers. Note also that while the Rhapsody catalog officially contains 110 halos, 106 were put through the following analyses.

Definition of Major Merger: In this project, a major merger is said to occur at timestep t_i when $M_{i-1}/M_i < 0.70$ where M_i is the mass within the virial radius of the center of the halo at time i. This gives a distribution of most-recent MM as seen in Figure 3:

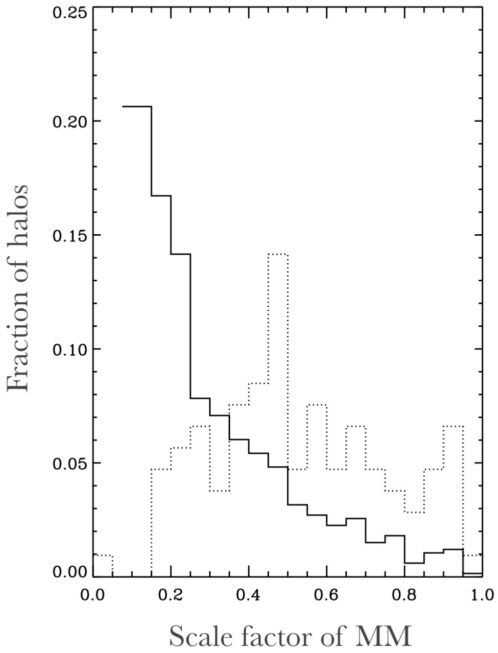


Figure 1: Fraction of all major mergers occuring within a scale factor bin (solid). The fraction of halos' most recent MM occurring within the bin is overplotted in the dotted line.

3.1 Relaxedness measures and halo properties

Figure 3.1 gives an array of plots with the RMs plotted against one another with information taken at $a = 1.0$ and Table 3 gives Spearman ranked-correlations. The clearest relations appear between f_{sub} and s, f_{sub} and c, and f_{sub} and s. The more mass is in substructures, which orbit the center, the less likely it is for the distribution to be isotropic. Thus in cases of higher f_{sub} and s values, the mass has a wider spatial distribution, skewing the spherically averaged concentration calculation higher.

Concentration has been shown to correlate with formation time [8]. Defining when a halo has "formed" can be a bit arbitrary, though it is clear some sort of time marking when halo gained some proportion of its mass would differentiate halos. While the final mass of all 106 halos lie within a narrow mass range, their rates of accumulating mass can be quite diverse. $a_{1/2}$ is defined as $\frac{1}{1+z_{1/2}}$ where $z_{1/2}$ is the redshift at which the halo first exceeds half of its final mass ($.5M_0$). Figure 3.1 shows the distribution of $a_{1/2}$ values for the halos.

Another way of characterizing the formation history is by fitting the full trajectory of the mass accretion history (MAH). In [8], this can be fitted by a simple power law with one parameter α and this model is adapted here (S is an arbitrary threshold):

$$\frac{M(a)}{M_0} = e^{-\alpha z} \tag{4}$$

Where M_0 is the mass of the halo at $z = 0$. And the characteristic epoch of formation, denoted a_c, can be derived from:

$$a_c = \frac{a_0 \alpha}{S} = \frac{\alpha}{2} \tag{5}$$

Since a_0, our intial scale factor, is 1, and the same value of S as [8] is taken. While $a_{1/2}$ has less scatter in its calculation, it also does not incorporate as much information of the

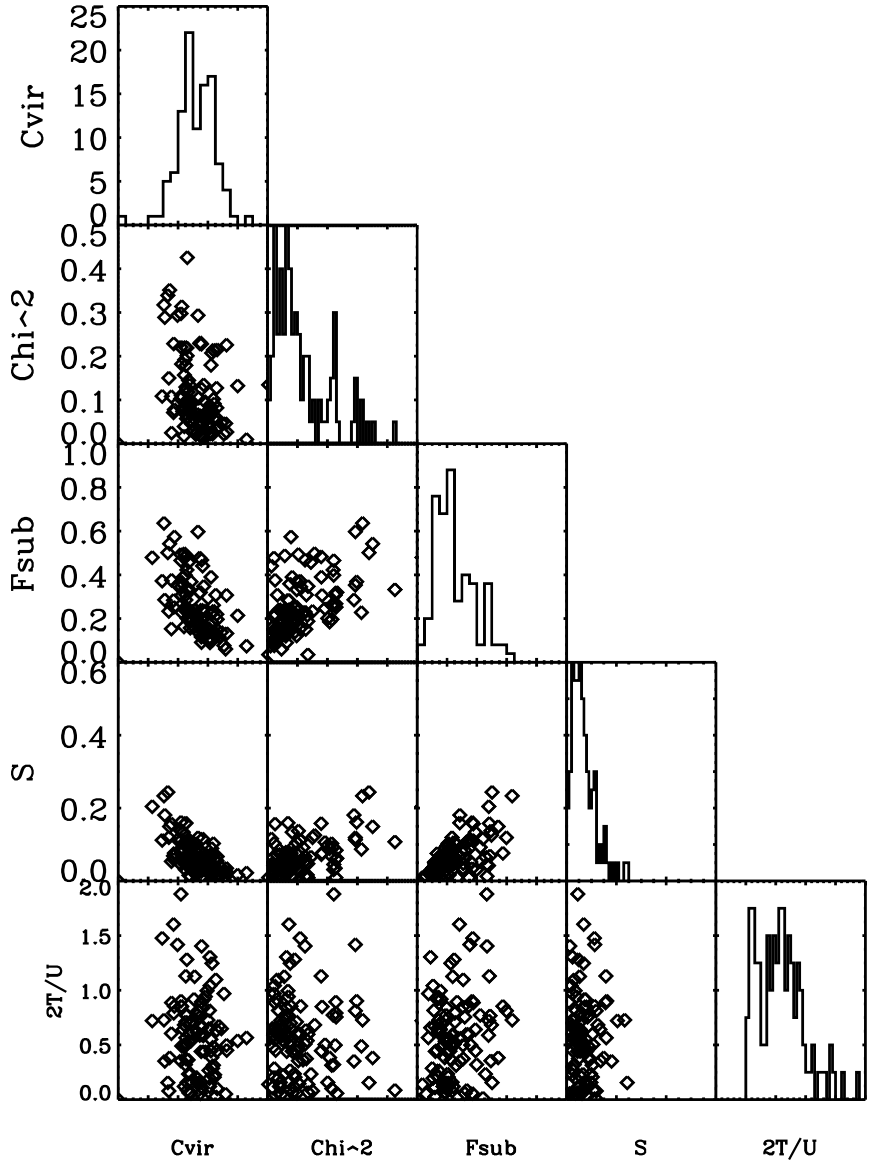


Figure 2: Off-diagonal plots are scatter plots of the relaxedness measurements at $a = 1.0$. On the diagonal are histograms of the values.

MAH as a_c does. These characterizations of the MAH are plotted against the RMs taken at $a = 1.0$ in Figure 3.1 with correlation statistics in Table 7. It happens that the correlation between a_c and last MM time is higher than that for $a_{1/2}$ but for both measurements we get strong anti-correlations with concentration. This agrees with the strong correlation between

Figure 3: Table of correlation and significance values for the relationships examined in 3.1. The values are given as they appear left to right and top to bottom in 3.1

Measure vs. Measure	**R - correlation**	***p*-value**
C_{vir} vs. χ^2	-0.367836	9.70663e-05
C_{vir} vs. f_{sub}	-0.587884	2.78713e-11
χ^2 vs. f_{sub}	0.579312	6.26867e-11
c_{vir} vs. s	-0.579909	5.92881e-11
χ^2 vs. s	0.447961	1.31010e-06
f_{sub} vs. s	0.607516	3.97083e-12
c_{vir} vs. virial ratio	-0.117528	0.227964
χ^2 vs. virial ratio	-0.000186086	1.00000
f_{sub} vs. virial ratio	0.189561	0.0505157
s vs. virial ratio	0.0116872	0.904910

Figure 4: Table of correlation and significance values of relaxedness values and concentration with the time of the last major merger.

Measure vs. Last MM	**R - correlation**	***p*-value**
C_{vir}	-0.435212	2.80388e-06
f_{sub}	0.437618	2.43454e-06
s	0.278723	0.00364790
χ^2	0.291659	0.00230319
virial ratio	0.0778430	0.425470

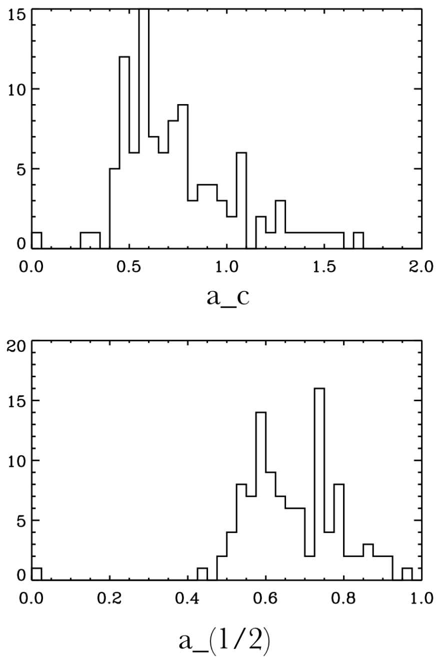


Figure 5: At left, a histogram of the a_c (top) and $a_{1/2}$ (bottom) values calculated for the 106 halos considered in the sample. At right, the $a_{1/2}$ distribution is reproduced but with smaller bin size to show detail

$z_{1/2}$ and c_{200} in [2]. Though in that paper, the correlation between $z_{1/2}$ and c_{200} is stronger than that between c_{200} and f_{sub}. But they find both correlations to be strong. In general, among the common measurements, all of our $|R_s|$ values are higher than those found in [2] but the relative values between relations are indeed similar.

Among relaxedness measures, χ^2 has the weakest correlation. Furthermore, one can still see a strong distinction in the MAHs when halos are taken from the ends of the distribution of the RMs at $a = 1.0$. For the measurements of c_{vir}, f_{sub}, s, χ^2, and $2T/U$, the halos with the top and bottom 20% values (21 halos) are taken and at each timestep the median value of their mass history is plotted. The results are seen in Figure 8. How well the measure does in distinguishing the two regimes could be taken as the maximum separation between the two averages for each measurement. Incidentally, the measurement with the largest separation is concentration. At this point it is noted that the virial ratio correlates poorly to the other RMs and as seen in Figure 8 does little to separate out halos' histories. This may be due to the omission of a surface pressure term [5] but the measurement will be dropped from most of following discussion.

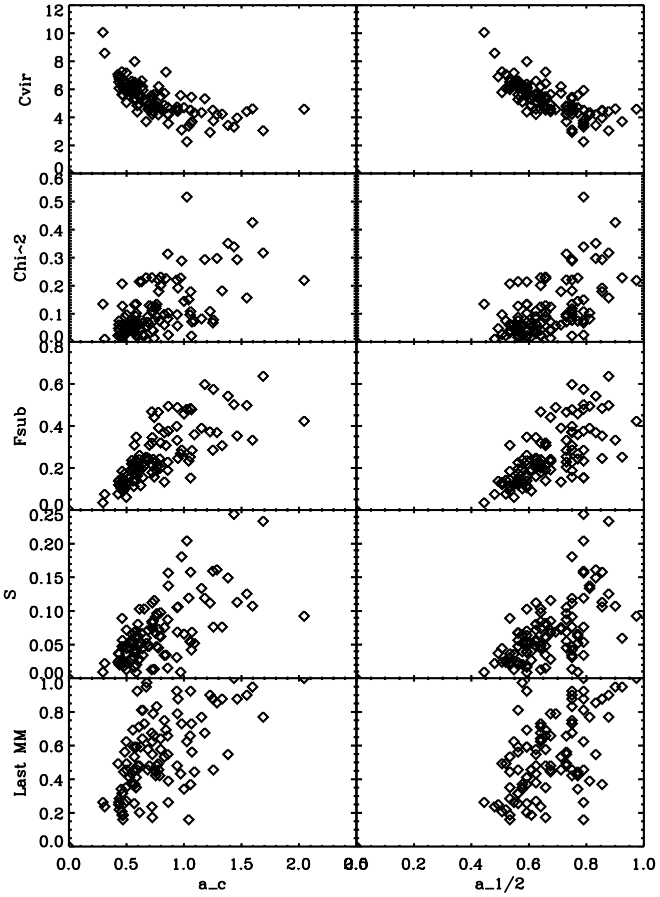


Figure 6: Different definitions of formation time are plotted against RMs at $a = 1.0$ and concentration. See Table 7 for correlation values

3.2 Relaxedness and major mergers

How recent states of relaxed measures correlate with the full trajectory of the MAH is affected by how such measures behave around major merger events. A major merger implies an unrelaxedness but the response seen in the RMs may not be seen directly, nor solely, at the reported time of the major merger since a merger may be ongoing for several snapshots. There is also a question of where do we want to look for MMs in most of our analysis. At

Figure 7: Table of correlation and significance values for the relationships examined in 3.1. The values are given as they appear left to right and top to bottom in 3.1

Formation time vs. Measure	**R - correlation**	p**-value**
a_c vs. C_{vir}	-0.746015	2.98065e-20
$a_{1/2}$ vs. C_{vir}	-0.743025	5.05537e-20
a_c vs. χ^2	0.583755	4.13028e-11
$a_{1/2}$ vs. χ^2	0.541274	1.74920e-09
a_c vs. f_{sub}	0.810633	3.74928e-26
$a_{1/2}$ vs. f_{sub}	0.737966	1.21578e-19
a_c vs. s	0.651351	3.04152e-14
$a_{1/2}$ vs. s	0.621948	8.68429e-13
a_c vs. MM time	0.592732	1.74371e-11
$a_{1/2}$ vs. MM time	0.509901	2.0259e-08

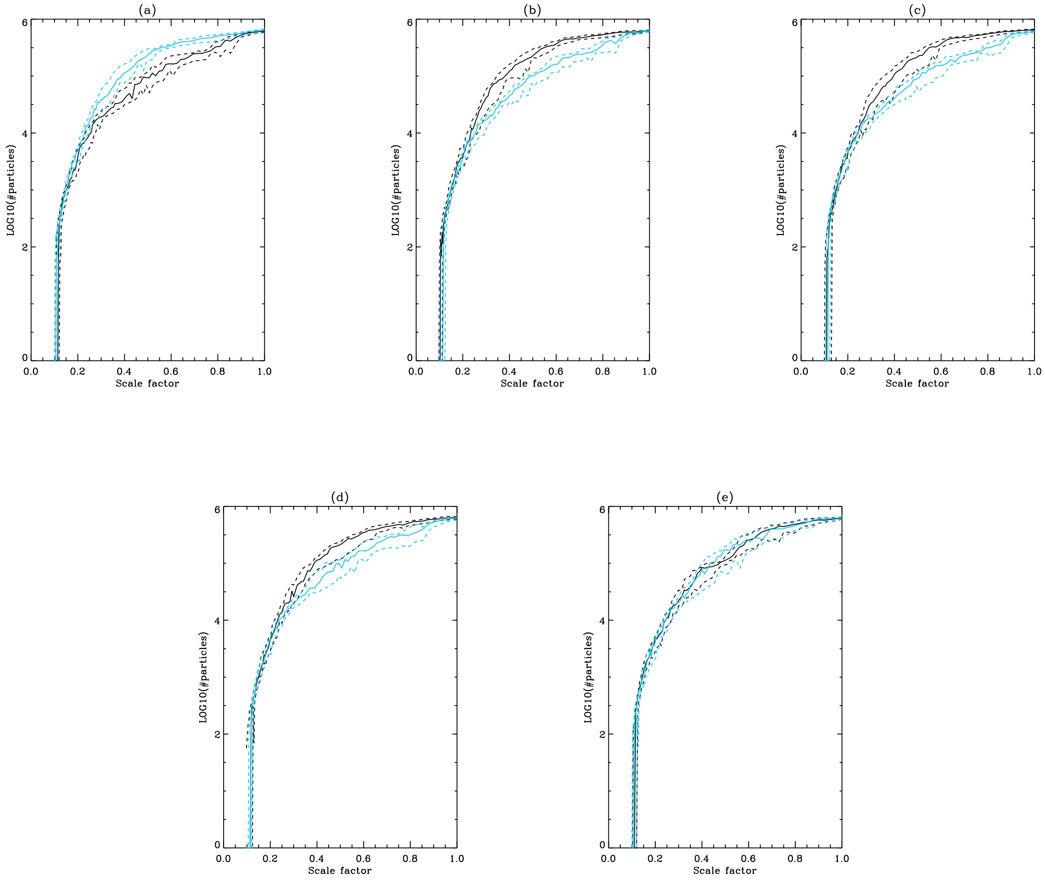


Figure 8: Solid lines show the averaged MAH for top 21 halos (black) and bottom 21 halos (blue) for 5 measures. Dashed lines show 25% and 75% values. (a) C_{vir}, (b) f_{sub}, (c) s, (d) χ^2, (e) virial ratio.

early times, the mass may jump considerly with the addition of only 1000 particles. In Figure 3.2 is plotted the distribution of relaxed measures across all recorded snapshots on the left, and at all major mergers on the right with values of the 183 major mergers occurring after $a = .4$ overplotted. Looking at the MM at all times introduces too much noise from the measurements of relatively few particles whereas the distribution for $a > .4$ shows a larger in mean value.

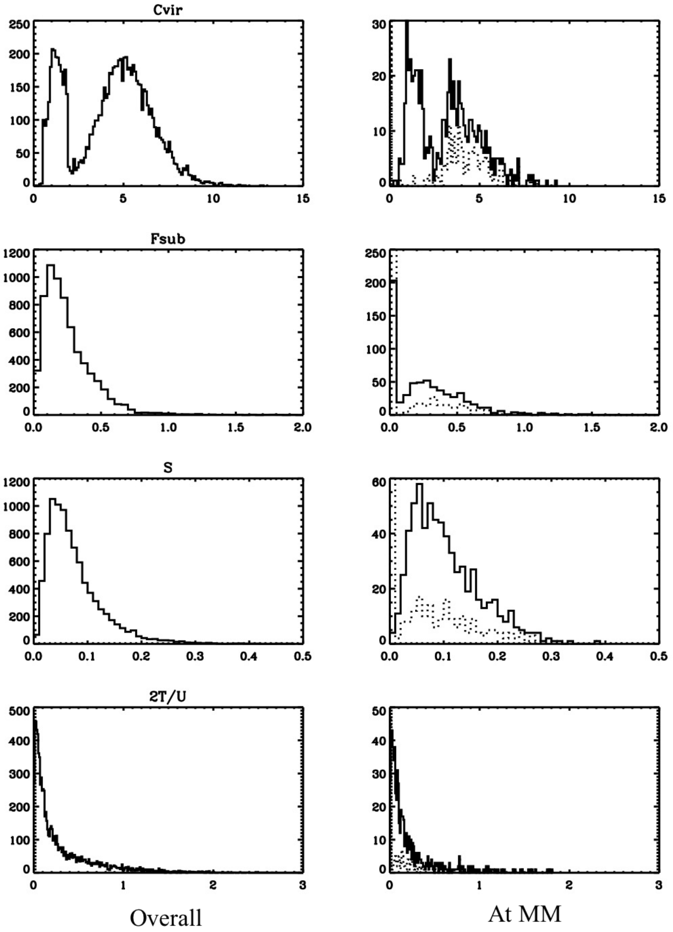


Figure 9: At left are the distributions of the RMs across all recorded snapshots and at left the distributions of values that occur at MM. In particular, the values for MMs occurring after $a = .4$ are shown in the dotted line.

Very simply, one can look at the initial response of the RMs to the major merger. Figure 10 shows the values of f_{sub}, and s normalized to the value taken on at the MM and for halos whose last MM occurred after $a = .4$ for 3 snapshots before and 4 snapshots after the major merger (this resulted in 68 halos). f_{sub} and s where chosen since they have the strongest correlations to formation time and unlike concentration have little systematic error in their measurements. The lines are colored to reflect the jump in mass from the previous snapshot to this one, thus bluer lines indicate smaller jumps in mass and green/yellow lines more so. Since the lines are converge to 1 at time 0.0, the radial distribution of colors on both sides of the vertical through that point give some indication of how f_{sub} immediately changes around the MM for different mass jumps. Particularly for f_{sub} the green lines are concentrated within 26^o of the vertical, and the most dramatic drops from before the MM to the MM are by lower mass jumps, for which the merger may have already mostly happened. Little besides an intuititon for the varied responses in the structure to the major merger can be deduced from Figure 3.2, so Figure 3.2 gives boxplot information for bins .5 Gyr in width. f_{sub} values depart from the MM value in a much narrower range than does s though there is a 60% spread in values between the upper and lower quartiles. Given that s often exhibits a periodic nature when looked at across snapshots - presumably as structures cross and re-cross the halo - this seems reasonable. In this mass range it is noted that younger halos are less able to disrupt their substructures [1] so the extended presence of mass in substructures is also unsurprising. Furthermore, our definition of MM time helps enable this since it is set when these particles first enter the virial radius.

It is logical to then ask how long a halo stays “unrelaxed”. Again f_{sub} and s are considered. To determine when a halo first becomes unrelaxed after a MM, the thresholds set by [4] were first used. These were $f_{sub} < .1$ and $s < 0.07$. For s, this threshold still allowed us to see a significant number of halos to settle. For f_{sub}, only 3 out of 68 settled within the time from their MM to $a = 1.0$. It took those three halos, 11,18, and 17 snapshots to get below .1. The average f_{sub} value at the time of major merger was .509 among these 68 halos. So to look at more points, a threshold of .20 was selected. In this case, only 8 were not able to relax and these halos were given a “time to relax” of ∞. Interestingly, there seemed no evident

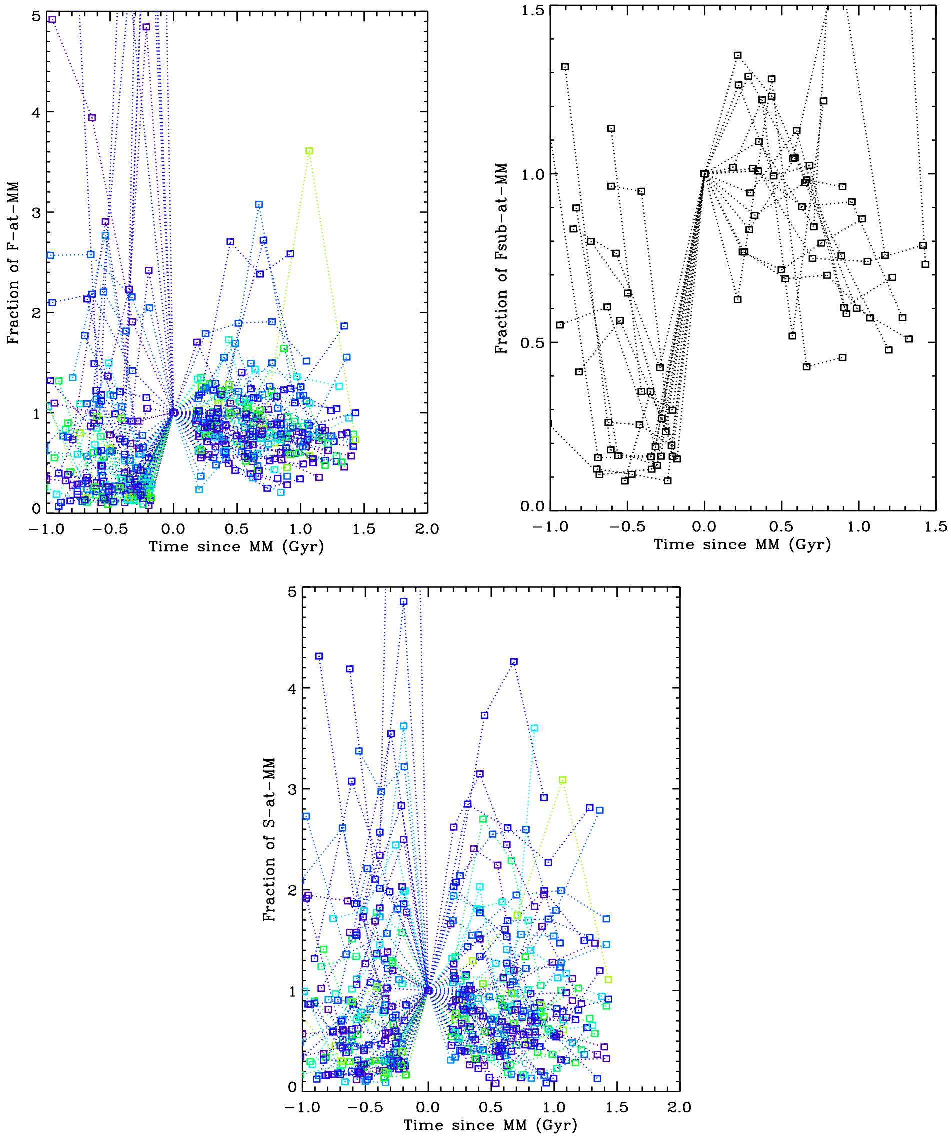


Figure 10: A stacked view of all the MMs occurring after $a = 0.04$ and within 4 snapshots of $a = 1.0$ where the RM value normalized to $RM(a_{mm})$ is plotted against the time in Gyr from the major merger. A negative value implies such a time before the MM. The value at left is for f_{sub} and on the bottom$s$. The top right isolates the halos with $\log_{10}(M_f/M_i) > .5$. The line color attaches lower values of the mass jump ($\log_{10}(M_f/M_i)$) to the bluer end of the spectrum and higher values to the redder.

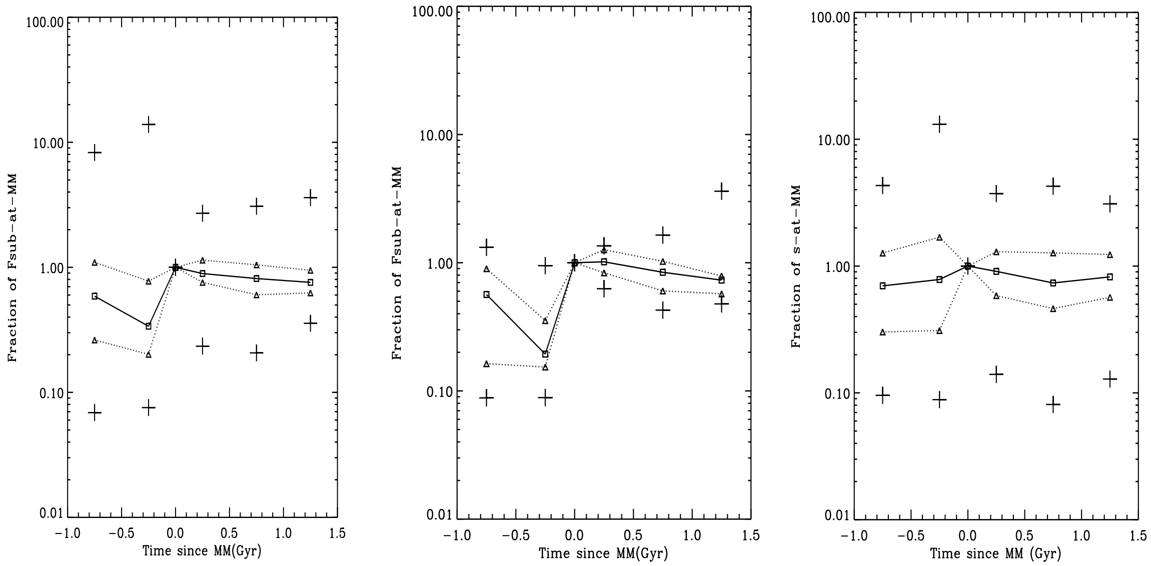


Figure 11: These plots describe the information in Figure 10. Time is binned by 0.50 Gyr. The solid line connects the medians, the dotted lines the upper and lower quartiles and the plus signs the maximums and minimums.

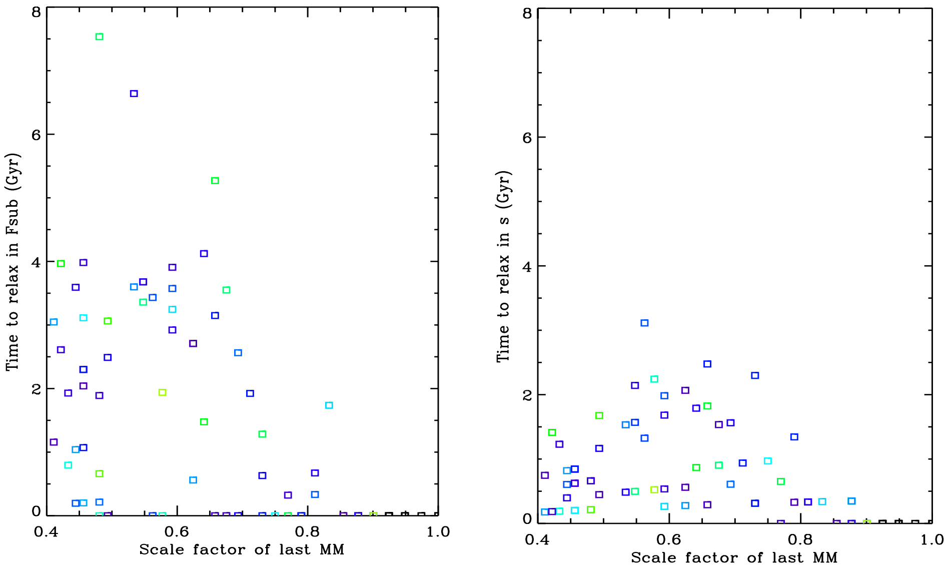


Figure 12: The scale factor of a halo's last MM and the period afterwards that it remained in an "unrelaxed" according to different RM measures. The colors correspond to the mass gained at that snapshot, lower - bluer, higher-redder.

correlation between the period of unrelaxedness and time of last MM (and again, we are considering only the most recent MM for a given halo) and to the mass ratio occurring at the MM (see Figure 12. Possible effects may come from the f_{sub} curves being un-smoothed.

Halos that we would not expect to relax as quickly maybe only dip below the threshold for a single snapshot. Something may emerge if the halos were separated by formation time or similarly by the fraction of their final mass gained at this step. This may also be a sampling issue since this analysis considered MMs in a wide time range and yet there stills seems to be evidence that by $a = 1.0$, traces ostensibly left by MMs provided enough information with which correlate formation time with RMs.

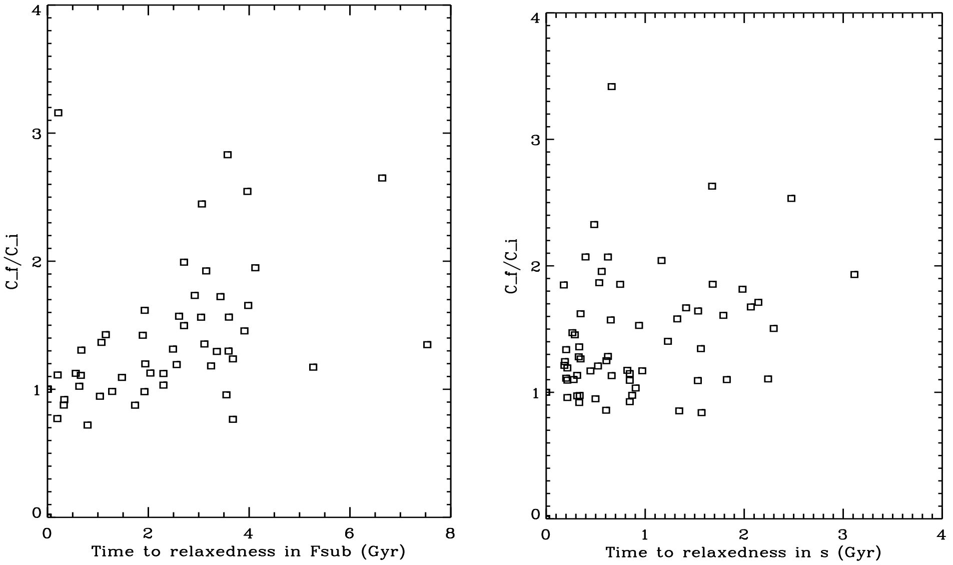


Figure 13: Relating the period of time it look for the RM to go below the relaxedness thresholds - chosen here as 0.20 for f_{sub} and 0.07 for s - and the change concentration underwent during this time period. The change we observe is the ratio c_f/c_i where c_f is the concentration when the RM becomes relaxed and c_i the the concentration when the MM occurred.

An interesting relationship we do observe was with respect to how this period of unrelaxedness related to how the concentration changed over that time period. In Figure 13 we see the ratio of final to initial concentrations taken with respect to this period of unrelaxedness after a MM. The f_{sub} plot at left (R_s value of 0.687901) suggests that the shorter this period of unrelaxedness, the smaller gain in concentration. Yet Figure 12 would motivate that this happens in a way that is not completely dependent on the amount of mass gained at the time of the MM. The absence of correlation with respect to the time s is unrelaxed may be because of the smaller timescales in comparison to f_{sub}. Though even when we looked at the ratio for the f_{sub} time period with the time spent unrelaxed in s, there was little correlation.

4 Conclusion and future work

With sample of 106 massive halos we were able to examine relationships among relaxedness measures and see how much relaxedness measurements relate to the evolution of the halo (by way of chosen formation times). These correlations in the end, however, are directly influenced by how these relaxedness measures respond to the most evident drivers of unrelaxedness - major mergers. When we focused on f_{sub} and s we saw what we expected in terms of the halo remaining unrelaxed for longer periods of time with respect to f_{sub}. What will certainly be interesting to look further into is how the initial concentration of the halo, the mass jump at the major merger, and scale factor of the major merger perhaps influence together the time the halo remains unrelaxed. Future work may involve looking at other major mergers besides the most recent ones and incorporate other systems of analysis aside from correlation. A technique of time sequence analysis may be particularly exciting.

5 Acknowledgements

This research was made possible by the United States Department of Energy Summer Undergraduate Laboratory Internship (SULI). I would like to change SLAC National Accelerator Laboratory and Stanford University for hosting me this summer. I would also like to thank my project advisor Risa Wechsler and her graduate student Hao-Yi Wu for their time, guidance, and support in this project.

References

[1] Gao, L. *et al*, 2011, The statistics of the subhalo abundance of dark matter haloes, *Monthly Notices of the Royal Astronomical Society*, vol. 410, Issue 4, 2309-2314p.

[2] Jeeson-Daniel, A., Dalla Vecchia, C., Haas, Marcel R., and Schaye, J., 2011, The Correlation Structure of Dark Matter Halo Properties, *Monthly Notices of the Royal Astronomical Society*

[3] Navarro, J. F., Frenk, C. S., and White, S. D. M., 1996, The Structure of Cold Dark Matter Halos, *Astrophysical Journal*, v.462, 563p.

[4] Neto, A. F., Gao, L., *et al.*, 2007, The Statistics of Λ CDM halo concentrations, *Monthly Notices of the Royal Astronomical Society*, vol. 381, Issue 4, 1450-1462p.

[5] Shaw, L., Weller, J., and Ostriker, J. and Bode, P., 2006, Statistics of Physical Properties of Dark Matter Clusters, *The Astrophysical Journal*, vol. 646, 815-833p.

[6] Steves, Bonnie A. and Maciejewski, A.J., 2001, *The restless universe: applications of gravitational n-body dynamics to planetary, stellar and galactic systems*, J W Arrowsmith Ltd, Bristol, 133p.

[7] Scoville, N., Aussel H. *et al.*, 2006, Large Structure and Galaxy Evolution in COSMOS at $z < 1.1$, *Bulletin of the American Astronomical Society*, vol. 38, 1211p.

[8] Wechsler, R., Bullock, J., Primack, J., Kravtsov, A., and Dekel, A., 2002, Concentrations of Dark Halos from their Assembly Histories,*The Astrophysical Journal*, vol. 568, 52-70p.

[9] Wu, H. *et al*, In preparation

SLAC-TN-12-012

An Improved Technique for Increasing the Accuracy of Photometrically Determined Redshifts for “Blended” Galaxies

Ashley Marie Parker

Marietta College, Marietta, Ohio

Office of Science, Science Undergraduate Laboratory Internship Program

SLAC National Accelerator Laboratory

Menlo Park, California

August 13, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy’s Science Undergraduate Laboratory Internship, under the direction of Dr. Deborah Bard, Kavli Institute for Particle Astrophysics and Cosmology at SLAC.

Participant: ________________________________

Research Advisor: ________________________________

Work supported in part by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

Abstract

An improved technique for increasing the accuracy of photometrically determined redshifts for “blended” galaxies. ASHLEY M. PARKER (Marietta College, Marietta, OH 45750) DEBORAH J. BARD (Kavli Institute for Particle and Astrophysics, Menlo Park, CA 94025)

The redshift of a galaxy can be determined by one of two methods; photometric or spectroscopic. Photometric is a term for any redshift determination made using the magnitudes of light in different filters. Spectroscopic redshifts are determined by measuring the absorption spectra of the object then determining the difference in wavelength between the “standard” absorption lines and the measured ones, making it the most accurate of the two methods.

The data for this research was collected from SDSS DR8 and then separated into blended and non-blended galaxy sets; the definition of “blended” is discussed in the Introduction section. The current SDSS photometric redshift determination method does not discriminate between blended and non-blended data when it determines the photometric redshift of a given galaxy. The focus of this research was to utilize machine learning techniques to determine if a considerably more accurate photometric redshift determination method could be found, for the case of the blended and non-blended data being treated separately. The results show a reduction of 0.00496 in the RMS error of photometric redshift determinations for blended galaxies and a more significant reduction of 0.00827 for non-blended galaxies, illustrated in Table 2.

Introduction

The goal of this project is to utilize Sloan Digital Sky Survey's (SDSS) data along with machine learning techniques to ultimately increase the reliability of photometric redshift analysis for blended and non-blended galaxies. Currently in SDSS database, self adjusting algorithms are used to determine photometric redshifts for all objects as a set, although none of these algorithms are optimized for galaxies [1]. This summer research aims determine if more accurate photometric redshift measurements will result from looking at only galaxy data and separating blended from non-blended.

A "blended" object is defined by the SDSS database as a light source (e.g. galaxy, star, etc.) for which intensity analysis shows multiple intensity peaks in the single light source, meaning there are multiple objects present [1]. Figure 1 shows a simple illustration of the deblending process, taken from a paper on the SDSS deblending algorithm [2]. Within the database the "frames pipeline" analyzes the data to determine if a light-emitting object is blended. If so, a de-blending algorithm is used to separate the multiple objects into "child" objects whose spectra add to become the "parent" image [1]. After de-blending these "child" light sources are all treated as separate objects, and are categorized as blended data. Objects that are flagged as blended are given a unique parentID number greater than zero, otherwise parentID is set to zero for the non-blended [1].

This parentID numbering is used in SDSS's newest data release, DR8, which covers approximately one third of the sky and includes all spectroscopic measurements that will be taken with this imaging camera [1]. This research project will use data acquired from SDSS which includes approximately 900,000 galaxies for which both photometric and spectroscopic data has been recorded. The reason both photometric and spectroscopic methods will be analyzed is that an accurate redshift measurement, assuming that the spectroscopic is the “true” redshift, must exist to test results from the new machine learning techniques.

A photometric redshift is measured using photometry, a method of looking at the light from an object through 5 standard filters (u, g, r, i, z) and using the overall magnitudes per filter to determine the redshift. The average wavelengths for the SDSS filters are shown in Table 1 [1]. Photometry is much less time consuming than the alternate method of spectroscopic redshift determination. In order to spectroscopically measure redshift there must be significantly more light collected for the object, so the full spectrum can be seen rather than just the intensities per filter. This makes spectroscopic far more accurate, however due to the large amount of telescope time it requires, the photometric is the most commonly used method. Telescope time is even more precious when it is used for a large scale sky survey, this is the reasoning behind efforts to increase the accuracy of the photometric method.

This is, to my knowledge, an original research project which will yield a photometric redshift determination method, specifically for blended or non-blended galaxies, which could be implemented in the next generation of sky survey databases, namely LSST, Large Synaptic Survey Telescope. This increase in accuracy of photometric redshift measurements will impact many scientific measurements, which rely on redshift, such as the study of large scale structure of the universe and gravitational lensing.

The goal was to find a method which yielded results for blended galaxies which were more accurate than SDSS's photometric redshift values. The goal for the non-blended galaxies was to determine a method equivalent to SDSS's method, although we did not expect to be capable of doing a significantly more accurate determination on this data set.

In the future I will investigate the scientific application of the, more accurate, photometric redshifts. Ultimately this work is hoped to be useful for the future LSST database which will not contain spectroscopic redshift measurements for all objects, and will therefore need to make use of the photometric method.

Methods

The initial step of the project was to learn SQL, Structured Query Language, which is used to write queries that acquire data from SDSS. The multitude of data available on SDSS's database makes it ideal for “training” a machine learning program such that the example data should show nearly every variation in galaxy type. The data used for the “training” consisted of approximately 100,000 non-blended galaxies and 800,000 that were blended. The difference in sample size was unintentional; it is a product of the fact that most of SDSS’s galaxy data is blended.

This project made use of CasJobs DR8, a program which utilizes SQL queries to acquire large amounts of data, from the most recent data release. The queries allowed for request of specific useful quantities such as: spectroscopic redshift measurement, two separate photometric redshift measurements using “random forest” and “robust fit” methods, magnitudes in the bands u, g, r, i, z, parentID and uncertainty measurements for all relevant quantities. An example of one SQL query for non-blended galaxy data, which was used for this research, is shown in Figure 3.

Data was requested for all objects which are of the type “galaxy” and downloaded for use in the ROOT data analysis framework. Data was requested for both blended and non-blended objects separately, so that a determination could be made if the photometric redshift measurements for blended galaxies are less accurate than measurements of non-blended objects. There was an investigation into which of the predetermined photometric redshift determination methods is most accurate, which was determined to be the “robust fit” technique.

For this research a machine learning technique, specifically TMVA, was used in order to find a more accurate method for determination of photometric redshift for blended galaxies. The Toolkit for Multivariate Analysis (TMVA) is a 'ROOT-integrated environment' which allows multivariate regression techniques to be used to analyze large data sets [3]. A TMVA regression script template was edited to include all necessary information about the studied galaxies. The goal of the script is to start with only the galaxy magnitudes in the 5 different bandwidth filters (u, g, r, i, z) and from that determine a redshift which is in good agreement with the spectroscopic value, which is considered the “correct” redshift. Within the script various methods attempt to determine redshifts, afterwards a given method's results were compared with the spectroscopic redshift determined by SDSS to see which method gave the best approximation.

The TMVA regression script was “trained” separately for blended and non-blended galaxies. All available TMVA methods were tested on the data, which include: PDERS, PDERSkNN, KNN, LD, FDA_GA, FDA_MC, FDA_MT, FDA_GAMT, MLP, SVM, BDT, and BDTG. All method titles are acronyms which describe the underlying mathematics of the method, there are far too many methods to describe them all in sufficient detail in this paper, for specific information regarding the individual methods see reference [3], the TMVA users guide.

The data analysis began by determining the accuracy of the current SDSS photometric method by comparing it to the spectroscopic data, which was needed to show the improvement of the machine learning methods used. In this study the accuracy of a given method was determined by an RMS error, given by:

$$\sigma_z \quad \equiv \sqrt{\overline{(\delta z)^2} - (\overline{\delta z})^2} \qquad \text{Equation 1.}$$

Where $\delta z = z_{method} - z_{spectroscopic}$, with z being the determined redshift and the overhead bars in the equation representing when the average is taken [4].

Results

On the far right of Figures 4 and 5 the graph shows the difference in redshifts as determined by spectroscopic and SDSS photometric methods, for non-blended and blended data respectively. From these figures, 4 and 5, it is clear that for both data sets there are large discrepancies between the photometric and spectroscopic redshift determinations. It is also clear from the shapes of the curves that the blended and non-blended data are distinctly different in their shape and reaction to the SDSS photometric method, thus showing the need for this research. The RMS errors for non-blended and blended data for the SDSS photometric method are shown in Table 2, 0.0724 and 0.04587, respectively.

In order to determine the most accurate of the TMVA methods, many histograms were produced showing the photometrically determined redshifts to be compared with the spectroscopic. These are depicted in Figures 7 and 8, which show the redshifts for the 7 most accurate TMVA methods and the SDSS photometric method. The best method should have a 2 peak profile similar to the graphs on the far left of figures 4 and 5, which show the spectroscopic or “true” redshift. Notice that the values end promptly at zero and do not go negative. Since the current “standard” cosmological theory infers that the universe is expanding, all galaxies should be moving away from the Milky-Way, making all redshift values of galaxies positive, therefore a negative value is considered non-physical.

Figures 9 and 10 show the “best” (lowest RMS error) TMVA method redshift values as a function of the spectroscopic redshift and compare those to the SDSS photometric data as a function of spectroscopic redshift, for blended and non-blended data respectively. When comparing figures 9 and 10, it becomes obvious that the non-blended data seems very tight where the blended data shows a large spread, which can partially be contributed to the blended data set being much larger but is also believed to be attributed to shortcomings of the SDSS deblending algorithm.

Discussion and Conclusion

Table 2 shows that for blended galaxy data the KNN method gave the most accurate results which are slightly more accurate than the SDSS determination. Table 2 also shows that for non-blended galaxies the MLP method yields the most accurate results for redshift, significantly more accurate than the SDSS determination. This was not an expected result; this research began with the hypothesis that a better photometric method could be determined for blended galaxy data and a similar, but not significantly more accurate, method would be found for non-blended data. This hypothesis was exactly the opposite of what was determined by the research: a significantly more accurate photometric redshift determination method was found for non-blended galaxy data while only a slight improvement was made on the blended galaxy data.

It is clear, from figures 7 and 8, that the BDT, LD and FDA_GAMT methods, first row right, second row left and third row right, respectively in both figures 7 and 8, are not the best methods due to the fact that they clearly include many non-physical redshift values. The BDTG method, shown on first row right in both figures 7 and 8, is also clearly not the best method due to the large amount of noise in the data which leads to the strange shape of the curve. The PDERSkNN method, shown in figures 7 and 8, third row left, yielded an unusual pile-up of redshift determinations at 0 which is not representative of the actual data. The data from these figures was input into Equation 1, along with the spectroscopic data from figures 4 and 5, in order to determine the RMS errors for all methods, which are displayed in Table 2.

In conclusion, a significant improvement over the SDSS photometric redshift determination method was made for non-blended galaxies. This will have far reaching effects on how photometric redshifts are determined for future large-scale sky-surveys, such as LSST. This work should be taken into account when the de-blending algorithms for the LSST database are being developed since the blended data clearly behaves differently from the non-blended due to the effects of the de-blending process. These improved photometric redshift determination methods should also be applied to existing data so that a more accurate representation of the universe can be seen.

Acknowledgments

I would like to express my gratitude to my mentor, Dr. Deborah Bard, for her assistance with this project. Dr. Bard was invaluable in my research effort; providing her guidance as well as many useful references along with assistance in troubleshooting and de-bugging my many ROOT macros. I would also like to thank Dr. Steven Rock, director of the SULI program at SLAC, for his efforts in making this year's SULI program a success. This research was funded by the U.S. Department of Energy, and the Kavli Institute for Particle Astrophysics and Cosmology.

References

[1] http://sdss.org, July 2011.

[2] http://www.astro.princeton.edu/~rhl/photomisc/deblender.ps.gz, August 2011.

[3] “TMVA Users Guide”, http://tmva.sourceforge.net/, August 2011.

[4] http://arxiv.org/PS_cache/astro-ph/pdf/0605/0605303v2.pdf, August 2011.

Figures

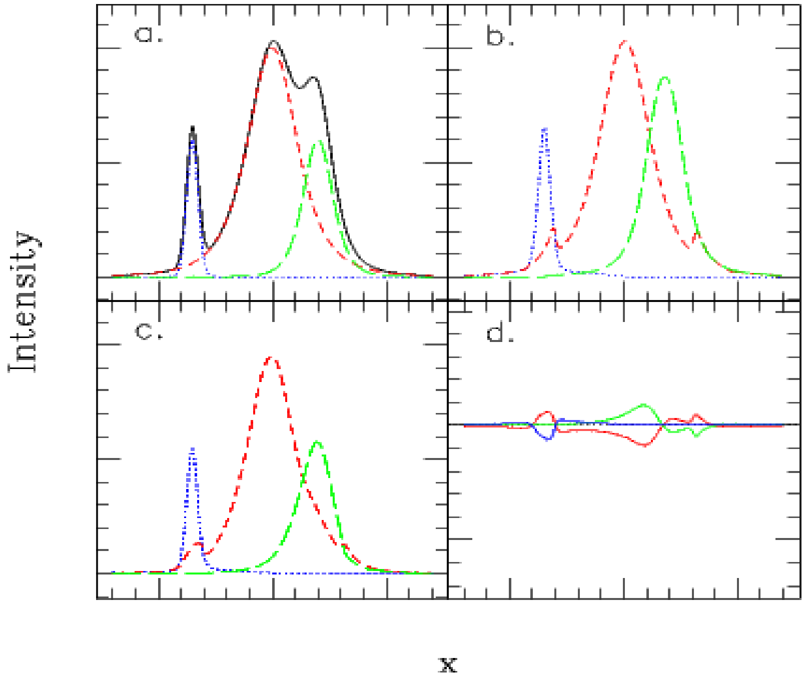
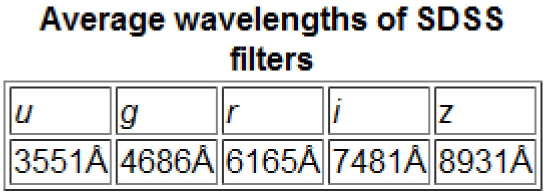


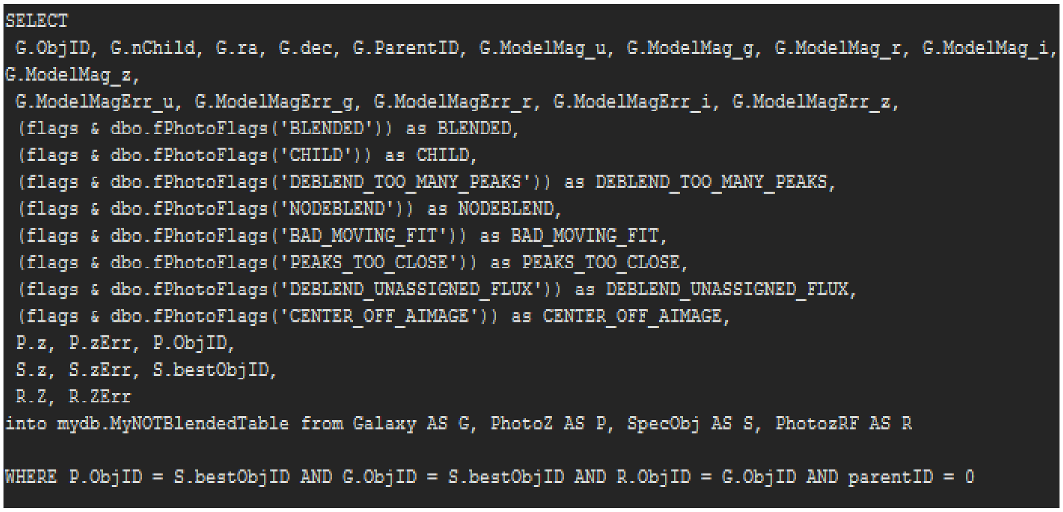
Figure 1: a) here is a “blended” parent object, solid black line, consisting of 3 child objects shown as colored dotted lines. b and c) shows the corresponding deblended children d) shows the difference between the sum of the children and the original parent, illustrating the imperfection of the method.



Average wavelengths of SDSS filters

u	*g*	*r*	*i*	z
3551Å	4686Å	6165Å	7481Å	8931Å

Table 1: The average wavelength, in angstrom, of the 5 model magnitude filters for SDSS DR8 data $_1$.



```
SELECT
 G.ObjID, G.nChild, G.ra, G.dec, G.ParentID, G.ModelMag_u, G.ModelMag_g, G.ModelMag_r, G.ModelMag_i,
G.ModelMag_z,
 G.ModelMagErr_u, G.ModelMagErr_g, G.ModelMagErr_r, G.ModelMagErr_i, G.ModelMagErr_z,
 (flags & dbo.fPhotoFlags('BLENDED')) as BLENDED,
 (flags & dbo.fPhotoFlags('CHILD')) as CHILD,
 (flags & dbo.fPhotoFlags('DEBLEND_TOO_MANY_PEAKS')) as DEBLEND_TOO_MANY_PEAKS,
 (flags & dbo.fPhotoFlags('NODEBLEND')) as NODEBLEND,
 (flags & dbo.fPhotoFlags('BAD_MOVING_FIT')) as BAD_MOVING_FIT,
 (flags & dbo.fPhotoFlags('PEAKS_TOO_CLOSE')) as PEAKS_TOO_CLOSE,
 (flags & dbo.fPhotoFlags('DEBLEND_UNASSIGNED_FLUX')) as DEBLEND_UNASSIGNED_FLUX,
 (flags & dbo.fPhotoFlags('CENTER_OFF_AIMAGE')) as CENTER_OFF_AIMAGE,
 P.z, P.zErr, P.ObjID,
 S.z, S.zErr, S.bestObjID,
 R.Z, R.ZErr
into mydb.MyNOTBlendedTable from Galaxy AS G, PhotoZ AS P, SpecObj AS S, PhotozRF AS R

WHERE P.ObjID = S.bestObjID AND G.ObjID = S.bestObjID AND R.ObjID = G.ObjID AND parentID = 0
```

Figure 3: This SQL query is requesting data on non-blended galaxies where both spectroscopic and photometric data is available.

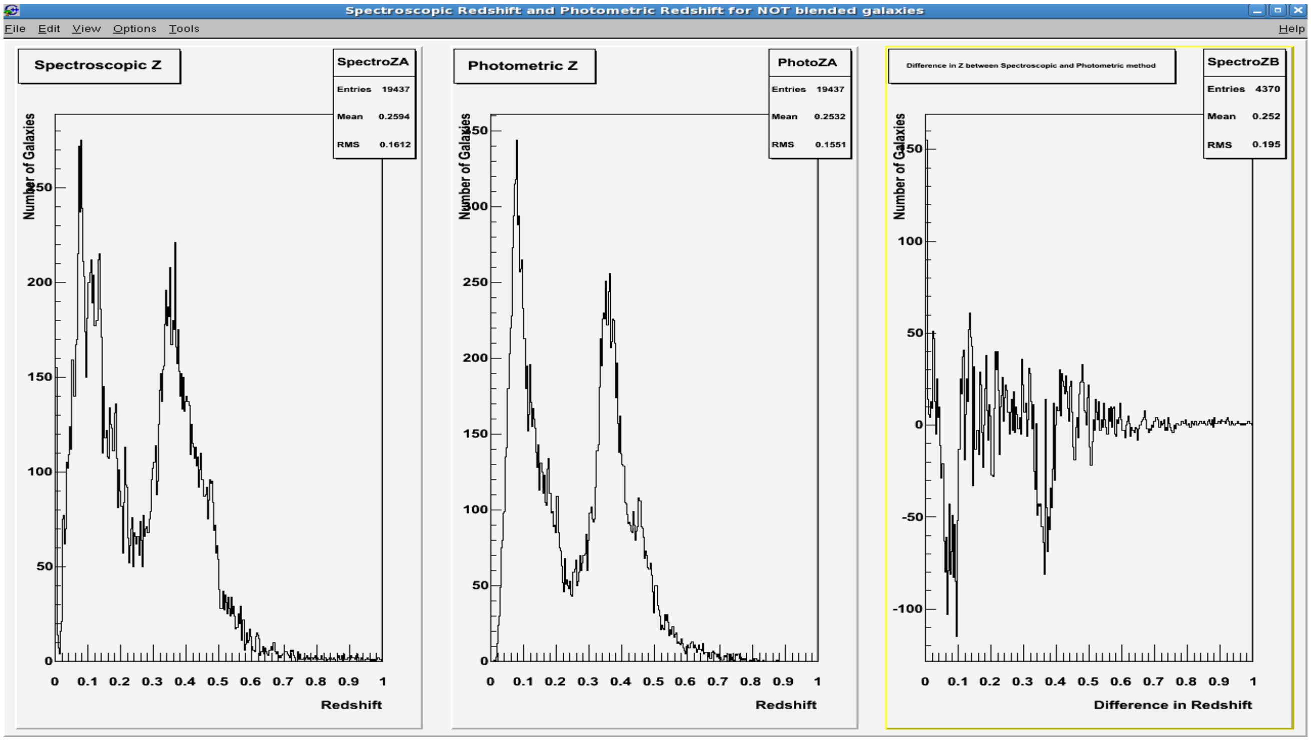


Figure 4: Shows non-blended data: on the far left is the spectroscopic redshift determination, middle shows the SDSS photometric method and on far right is the difference between the two.

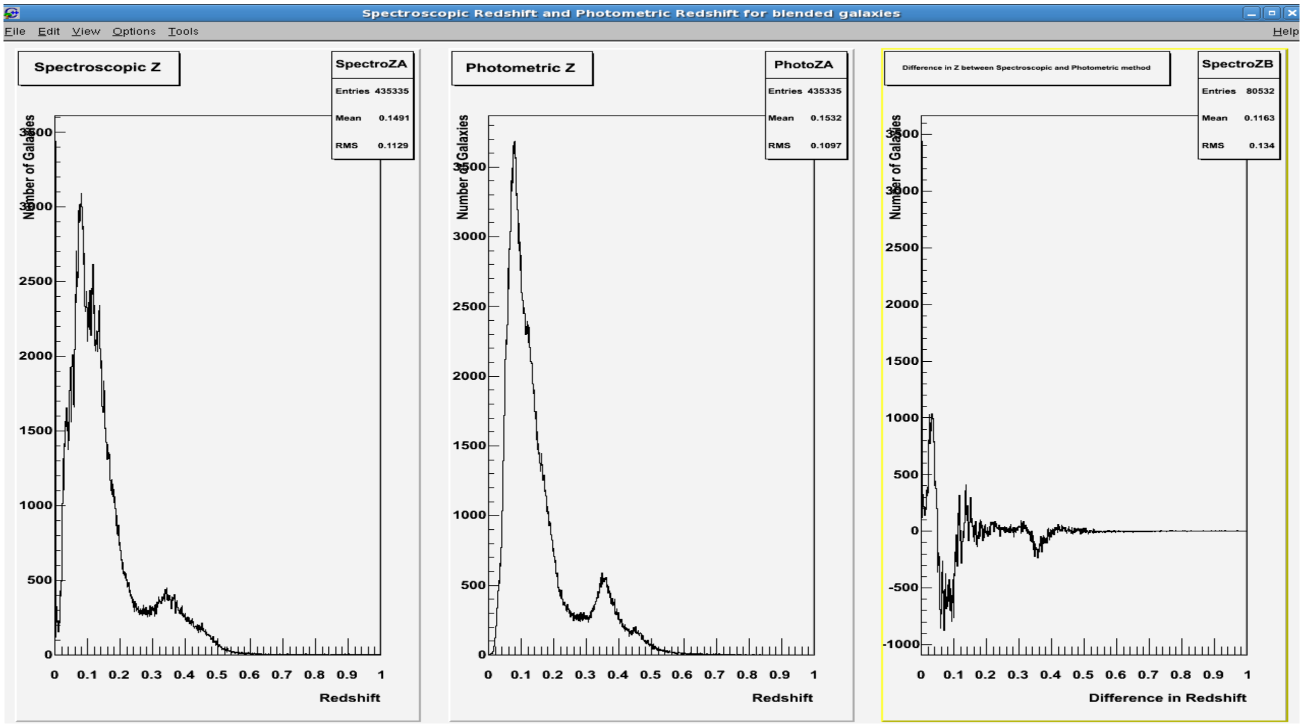
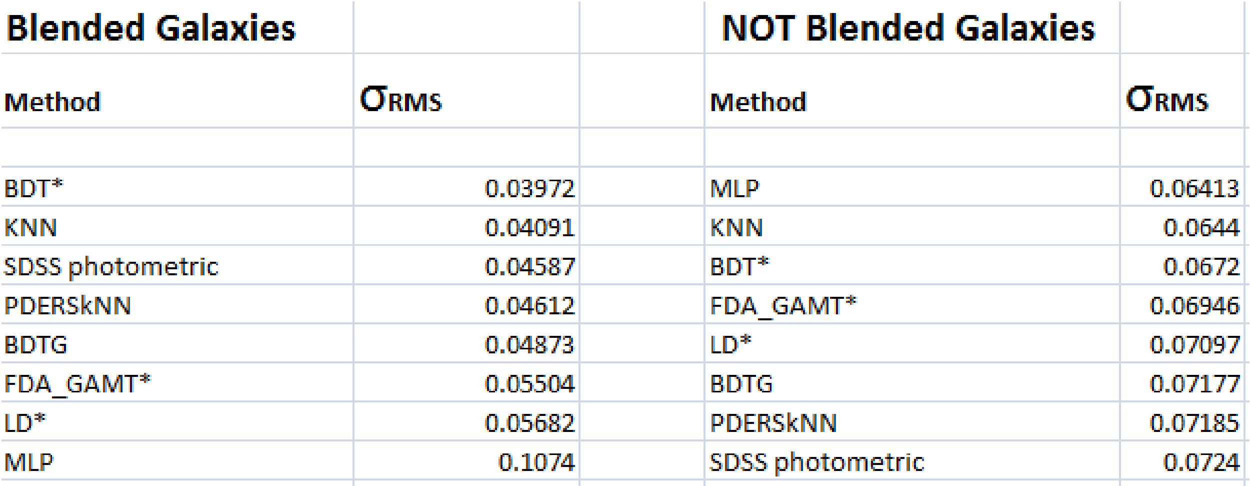


Figure 5: Shows blended data: on the far left is the spectroscopic redshift determination, middle shows the SDSS photometric method and on far right is the difference between the two.



Blended Galaxies			NOT Blended Galaxies	
Method	σ_{RMS}		Method	σ_{RMS}
BDT*	0.03972		MLP	0.06413
KNN	0.04091		KNN	0.0644
SDSS photometric	0.04587		BDT*	0.0672
PDERSkNN	0.04612		FDA_GAMT*	0.06946
BDTG	0.04873		LD*	0.07097
FDA_GAMT*	0.05504		BDTG	0.07177
LD*	0.05682		PDERSkNN	0.07185
MLP	0.1074		SDSS photometric	0.0724

Table 2: Above are the RMS errors, in accending order, computed using Equation 1, of the various TMVA machine learning techniques as compared to the SDSS photometric method. The * next to some methods is to indicate the fact that although the RMS may be low, because it was calculated on the redshift interval of 0 to 1, it is not the best method due to the large number of non-physical redshift determinations.

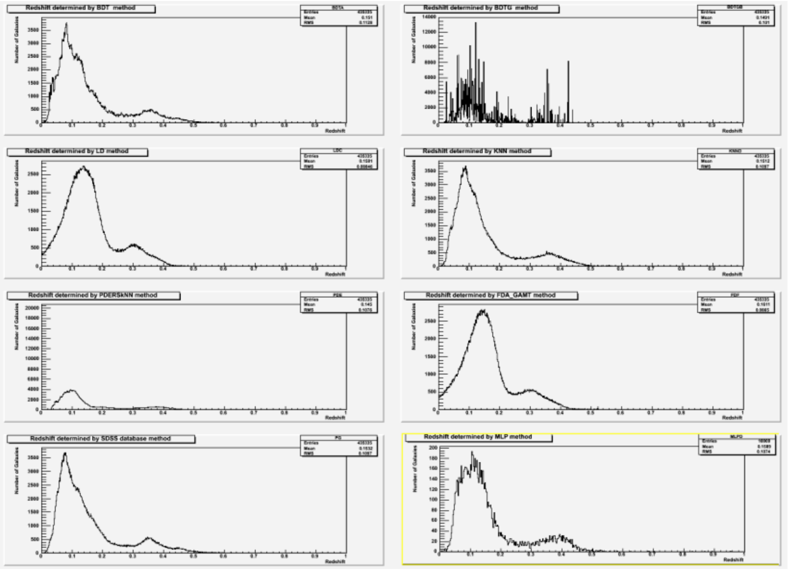


Figure 7: The above figure shows the redshift results of various methods for blended data. Method titles from left to right beginning with top; BDT, BDTG, LD, KNN, PDERSkNN, FDA_GAMT, SDSS photometric, MLP.

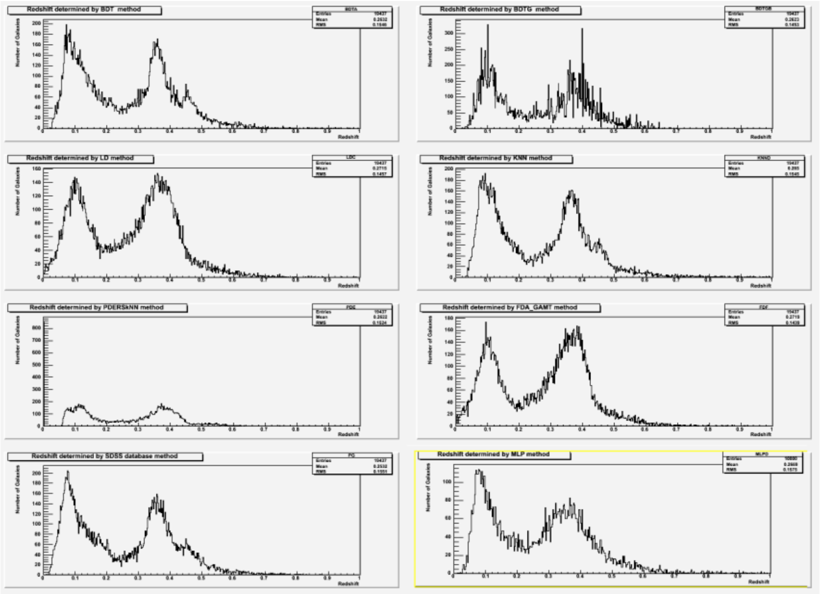


Figure 8: The above figure shows the redshift results of various methods for non-blended data. The method titles from left to right beginning with top; BDT, BDTG, LD, KNN, PDERSkNN, FDA_GAMT, SDSS photometric, MLP.

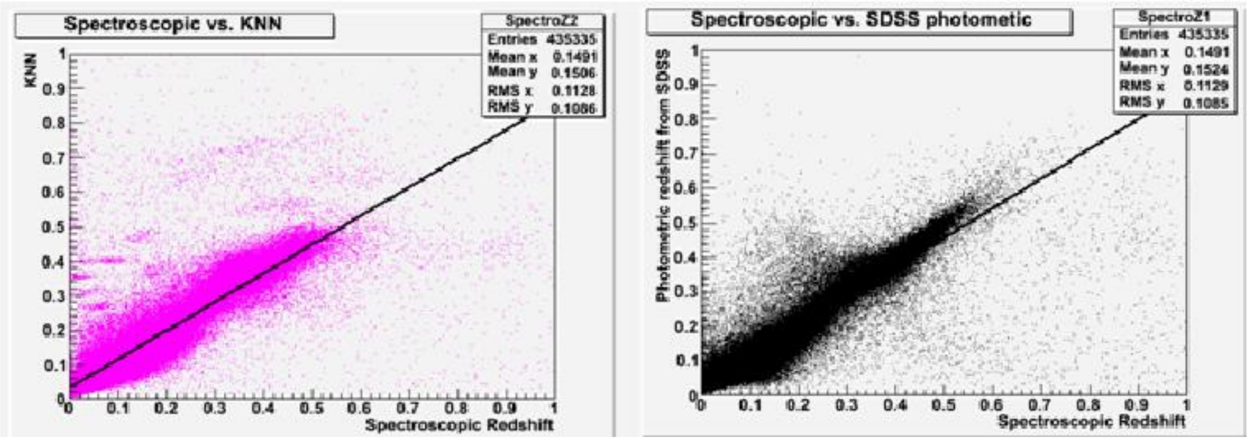


Figure 9: (Left) The “best” redshift determining TMVA method for blended data, KNN, as a function of the spectroscopic redshift.(Right) The SDSS photometric data as a function of spectroscopic redshift.

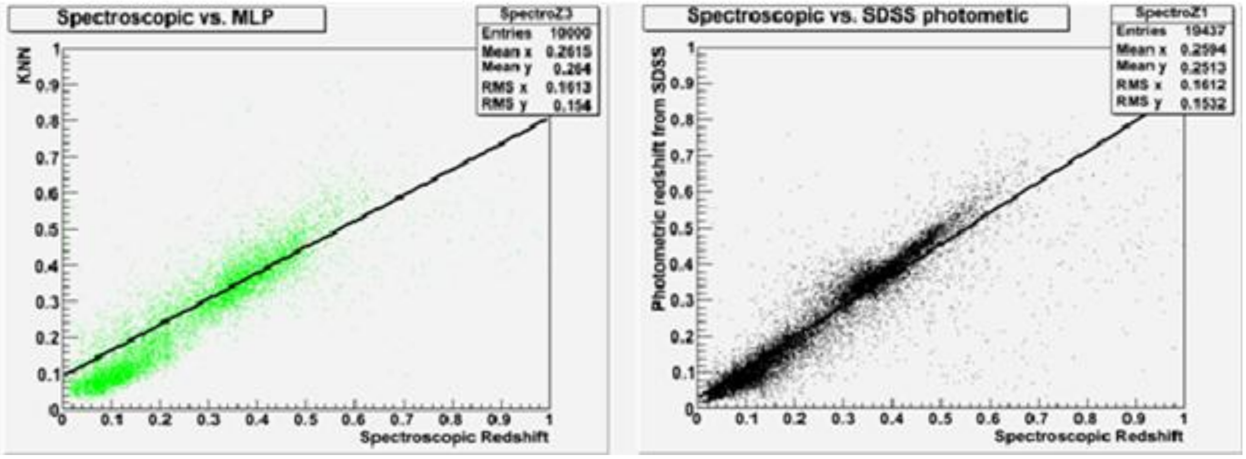


Figure 10: (Left) The “best” redshift determining TMVA method for non-blended data, MLP, as a function of the spectroscopic redshift.(Right) The SDSS photometric data as a function of spectroscopic redshift.

SLAC-TN-12-013

Characterizing the Nanoscale Layers of Tomorrow's Electronics

An Application of Fourier Analysis

Christopher Bishop Payne
Princeton University

United States Department of Energy
Office of Science, Science Undergraduate Laboratory Internship Program
SLAC National Accelerator Laboratory
Menlo Park, California

August 20th, 2011

Prepared in partial fulfillment of the requirement of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Apurva Mehta of the SSRL division of SLAC National Accelerator Laboratory.

Participant: ______________________________
Signature

Research Advisors: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Abstract

Characterizing the Nanoscale Layers of Tomorrow's Electronics: An Application of Fourier Analysis. CHRISTOPHER PAYNE (Princeton University, Princeton, NJ 08544) APURVA MEHTA (Stanford Synchrotron Radiation Lightsource (SSRL) Division of SLAC National Accelerator Laboratory, Menlo Park, CA 94025) MATTHEW BIBEE (Stanford University, Stanford, California, 94305)

Thin film applications are of great interest to the semiconductor industry due to the important role they play in cutting edge technology such as thin film solar cells. X-Ray Reflectivity (XRR) characterizes thin films in a non-destructive and efficient manner yet complications exist in extracting these characteristics from raw XRR data. This study developed and tested two different algorithms to extract quantity of layers and thickness information on the nanometer scale from XRR data. It was concluded that an algorithm involving a local averaging technique revealed this information clearly in Fourier space.

I. Introduction

The next generation of electronic devices will power an increasingly technology dependent world, requiring industry to be able to manufacture semiconductor chips that are both smaller and composed of new materials. The semiconductor industry has rapidly grown over the past three decades as they have pushed the limits of fabrication technology into the nanometer resolution range. This exponential decrease in device size was hypothesized by Gordon Moore in 1965[i] and has led to increasingly cheaper devices with more processing power per area than their more expensive predecessors. Additionally, nearly all semiconductors are now made of silicon, however the use of new semiconductor materials could yield electronics that are more efficient and can operate under greater extremes.

To unlock these favorable attributes of these next generation semiconductors, we need to be able to characterize and analyze materials with a resolution of a few nanometers. Many different techniques have been developed to perform this analysis, yet the one we will focus on is called X-Ray Reflectivity (XRR). This technique quantifies parameters of materials by revealing the number of layers of which the material is composed of and each layer's corresponding thickness, density, and roughness value.[ii] These parameters are discussed in more detail in the theory section, but for now, it is sufficient to understand that XRR allows the user to know the physical structure of a material on the nanometer scale. It should also be noted that unlike other techniques, XRR does not destroy the sample nor does it require specially prepared samples such as Transmission Electron Microscopy[iii] demands.

The specific focus of this project was to characterize the layers present on silicon carbide (SiC) wafers that GE is in the process of developing for future use in a wide range of applications that include wind turbines and hybrid-electric vehicles. [iv] It is important to keep in mind though that the XRR analysis techniques discussed in this paper can be applied to a vast range of cutting edge semiconductor applications that involve nanometer scale layers such as thin film solar cells and power transistors. Having the capability to characterize these nanometer layers in the non-destructive manner XRR allows, is a crucial step on the path to developing the electronics of the future that the world urgently seeks.

II. Materials and Methods

i. Reflectivity Theory

When a multilayer system is illuminated by an x-ray beam of wavelength (λ), the reflected beam profile measured as a function of the incidence angle, theta (Θ), contains information about the layer stratigraphy, density and roughness. Figure 1 shows a simple case of one thin layer on an infinitely thick substrate. The reflected beam in this case is a coherent summation of the beam reflected from the top surface and the beam reflected from the buried interface. The intensity of the reflected beam is then a function of the path length difference between the beams reflected from these two interfaces. This path length difference is shown in red in Figure 1.

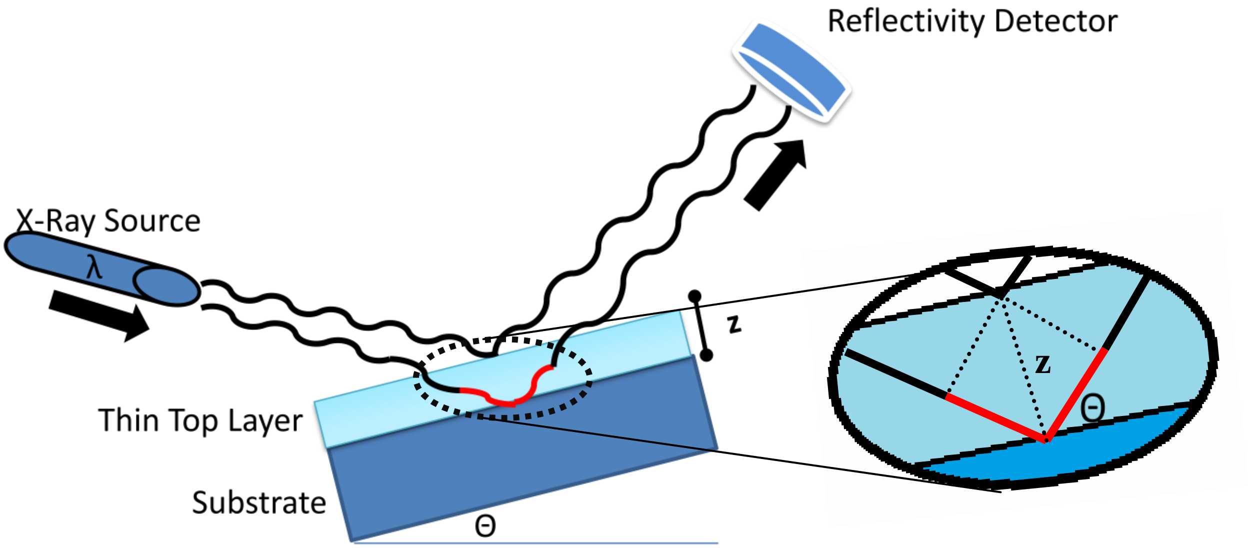


Figure 1 **XRR Experimental Setup**

Note the path length difference highlighted in red. This path length difference is a function of Θ and Z as shown in Equation 1.

The path length difference that modifies the phase of the reflected wave can be approximated as

$$Extra\ Distance = 2z\sin(\theta) \quad (1)$$

This extra distance corresponds to a phase difference at the point that the reflected beams interfere at the detector, dependent on the λ of the incident beam:

$$Phase\ Difference\ At\ Detector = \frac{2z\sin(\theta)}{\lambda} \quad (2)$$

The beams reflected from the two interfaces interfere constructively if this phase difference is a multiple of 2π and destructively if the phase difference is an odd multiple of π. This simple illustration, therefore, suggests that the intensity of the reflected beam is a function of the incidence angle (Θ) and is modulated to contain information about the layer thickness.

Under a more rigorous derivation, but still under the kinematic approximation, the reflected beam intensity is:

$$Intensity(S) \approx \frac{1}{S^4}\left|FT\left(\frac{d\rho(z)}{dz}\right)\right|^2 \quad (3)^{v}$$

with $\rho(z)$ being the electronic density a distance z (see Figure 1) into the sample and S defined as:

$$S = \frac{2\sin(\theta)}{\lambda} \quad (4)^{vi}$$

Equation 3 can be rewritten in the following manner:

$$Intensity(S) \approx \frac{1}{S^4} FT\left(\frac{d\rho(z)}{dz} * \frac{d\rho(z)}{dz}\right) \quad (5)$$

As the reflected beam intensity is a function of the electron density gradient into the sample, it contains information about not only the thickness of the individual layers, but also the interfacial roughness and density difference between adjacent layers[vii]. Figure 2 represents an ideal situation in which $\rho(z)$of a particular layer is uniform and $\rho(z)$ instantaneously changes at a layer interface.

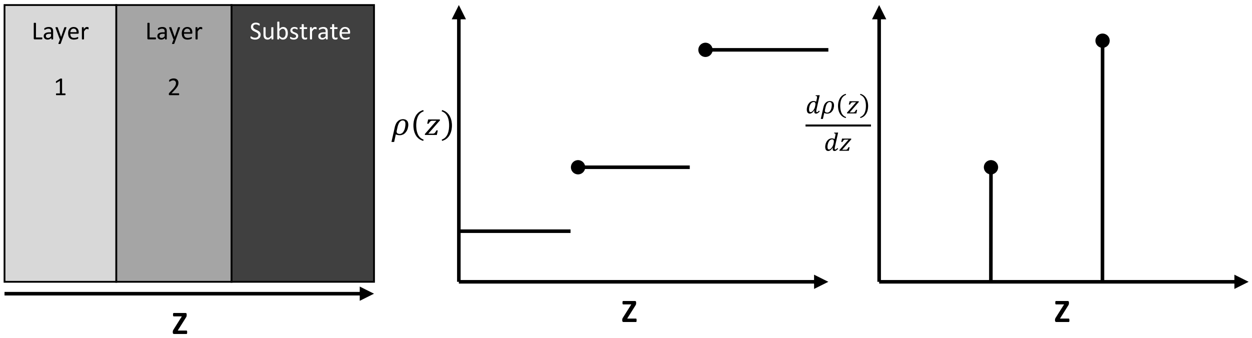


Figure 2 Relating density gradients to the sample's layers.

As Figure 2 depicts, the non-zero values of dρ(z)/dz correspond to layer boundaries expressed in terms of z, thus we must arrange Equation 5 to get dρ(z)/dz in terms of intensity. The 1/S^4 term in Equation 5 is a fall-off term which attenuates the intensity value over even a small S domain and is valid if the interfaces have no roughness. Experimentally we found this fall-off term ranges from 1/S^4 to approximately 1/S^6 for real multilayer samples, thus we developed an algorithm to calculate this fall-off term on a case by case basis so that it could be removed from the data. The algorithms that were tested are referred to as dN and 2N in the Results and Discussion section. With this fall-off term removed by using this algorithm, Equation 5 now has the form:

$$Intensity(S) - \text{ Falloff Term} \approx FT(\frac{d\rho(z)}{dz} * \frac{d\rho(z)}{dz}) \quad (6)$$

The final step is to take the inverse Fourier Transform which will yield the self-convolution of dρ(z)/dz:

$$FT^{-1}(Intensity(S) - \text{ Falloff Term}) \approx \frac{d\rho(z)}{dz} * \frac{d\rho(z)}{dz} \quad (7)$$

Shortly, using some simulations, we will explore how the layer characteristics are encoded in Equation 7.

The main challenge in converting the original equation, Equation 3, into Equation 7, is to accurately calculate the fall-off term so that it can be removed. The following sections describe two algorithms for calculating the fall-off term. Subsequently, we will test the range of validity of Equation 7, and explore how the various characteristics of the layers and layer stratigraphy are encoded in the formulation suggested by Equation 7.

ii. Fall-Off and Conversion Algorithm Development Outline

In order to find a solution to the problem outlined previously, I used MATLAB to create two algorithms that applied two different methods to remove the fall-off term from the intensity data. Independent of which method was used, the result was then transformed into Fourier space where sample characteristics such as number of layers present and layer thicknesses were compared with the information available in Fourier space. As will be discussed later, additional characteristics such as roughness or density of the layers are thought to be contained within the Fourier space, but were not fully explored in my study.

To optimize the fall-off subtraction algorithm and to test the range of validity of Equation 7, I used a MATLAB program called Multig, developed by Anneli Munkholm and Sean M. Brennan. I inputted sample characteristics such as the number of layers present and their corresponding thicknesses into the program and it outputted a simulated intensity versus Θ curve. I then applied my algorithm to the simulated data to test whether the algorithm could return the sample characteristics I had simulated. I used this simple process to develop and tune my algorithm until the algorithm yielded the correct characteristics.

Below, the entire process of simulating raw intensity versus Θ data and then transforming it into the accessible form of Equation 7 is outlined in five steps. Note that the two different fall-off removal methods, dN and 2N, are introduced in Step 3, Fall-off Removal.

Step 1, Simulate intensity versus Ө data using Multig

In this example, a 2 nm layer of aluminum (Al) between a 20 nm layer of silicon oxide (SiO_2) is simulated on a SiC substrate (Figure 4). The Multig program then outputs a simulated intensity versus Ө profile (Figure 3).

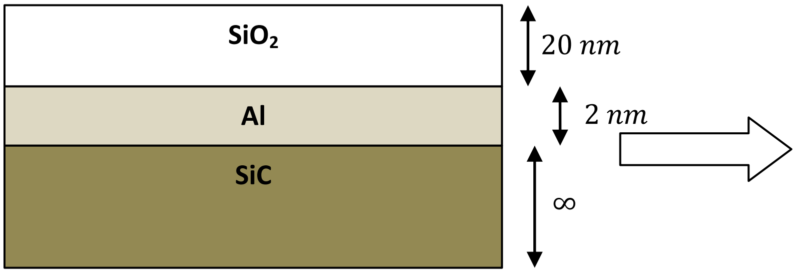


Figure 4 **Simulated sample inputted into Multig, no roughness in either layer.**

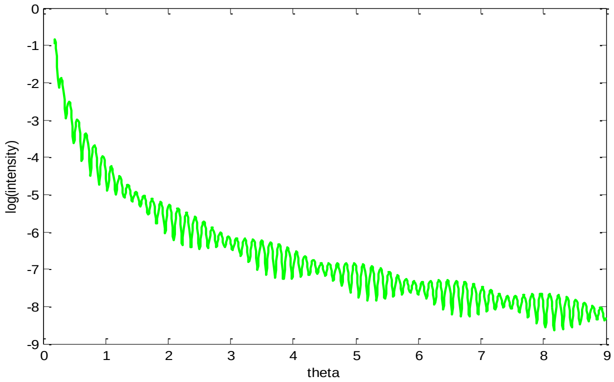


Figure 3 **Resulting log(Intensity) versus Ө simulated curve at 12keV**

Step 2, Restrict the Ө Domain

X-rays do not penetrate the sample until Ө exceeds a critical angle. The kinematic approximation is known to be inaccurate near the critical angle, therefore, data points below a certain Ө threshold are removed from the intensity versus Ө data set to be operated on. Additionally, the reflectivity curve for higher Ө values becomes progressively noisy, accordingly, data points above a certain threshold are removed as well. Techniques for selecting favorable threshold values, referred to as *theta_low_clip* and *theta_high_clip*, respectively, are discussed later. Additionally, the $\log_{10}$ of the intensity data is taken to make it easier to operate on the data over approximately eight orders of magnitude.

Step 3, Fall-off Removal

Apply the dN Method

The main idea behind this method is to take an averaged derivative of the intensity versus Ө signal as described by Poust and Goorsky in *Enhanced Fourier Transforms for X-Ray*

Scattering Applications, which will minimize the original fall-off component present in the resulting dataset[viii]. The discrete version of this, as we are dealing with a finite dataset is described as:

$$I_j^{dN}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{I(\theta_{j+i})-I(\theta_{j-i})}{\theta_{j+i}-\theta_{j-i}} \quad (8)$$ [ix]

Expressed in words, the algorithm takes the ***j***$^{\boldsymbol{th}}$ intensity & Θ data point and replaces it with the 'local' average derivative of the intensity. The 'local' area is defined by the dataset ranging from θ_{j-N} through θ_{j+N} and is centered on the ***j***$^{\boldsymbol{th}}$data point. By applying this process, the fall-off term is removed leaving a dataset in the form of Equation 6. As recommended by Poust and Goorsky, N is selected so that it is high enough to average out noise fluctuations but kept substantially less than the period of any possible thickness signals. As a result of taking the derivative, the dataset now trails the original dataset by a quarter of a period of the high frequency oscillations present in the original dataset.

Or, alternatively apply the 2N Method

The central idea behind this method is to construct a smooth and monotonic curve that removes the fall-off term from the original dataset while keeping all of its original frequency content. This is done by taking the local average of the intensity versus Θ data and recording it as the fall-off value at that point. As alluded to in *Enhanced Fourier Transforms for X-Ray Scattering Application*[x]:

$$I_j^{2N\_Falloff}(\theta) = \frac{1}{2N+1}\sum_{i=j-N}^{j+N} I(\theta_i) \quad (9)$$

The method is called '2N' as the local average includes the ***j***$^{\boldsymbol{th}}$ point plus the N data points between θ_{j-N} and θ_j plus the *N* points between $\theta_j$and θ_{j+N}. Effective selection of N will be discussed in detail later.

After quantifying the fall-off value at each point, we now subtract it from the corresponding original intensity values to remove the fall-off.

$$I_j^{2N}(\theta) = I_j^{Original}(\theta) - I_j^{2N\_Falloff}(\theta) \quad (10)$$

The last step in the 2N method is to take the anti-log of all the intensity values that have had the fall-off removed from them, $I_j^{2N}(\theta)$.

Regardless of whether the dN or 2N method is used, the fall-off has now been subtracted from the data, resulting in a dataset in the form of Equation 6, shown graphically in Figure 5.

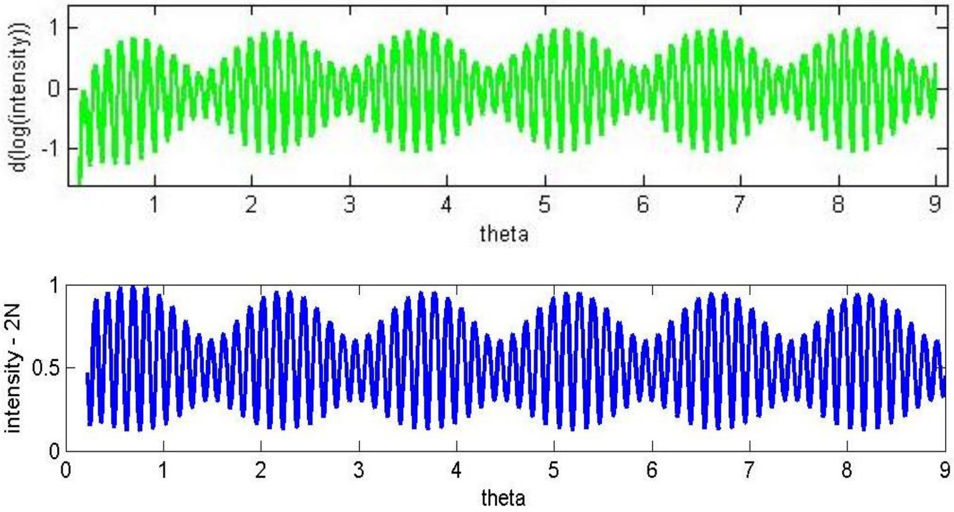


Figure 5 Graphical view of dataset after fall-off has been removed using dN (Top green graph) and 2N(Bottom blue graph)

Note, the dN result includes negative values, as it is the derivative of the log of the intensity. On the other hand, the 2N result is entirely positive because it is the antilog of the log of intensity minus the fall-off term:

$2N\ Result(Blue\ Graph) = 10^{I^{2N}}$ with I^{2N} defined in Equation 10

Step 4, Convert Ɵ into S

The Ɵ values are converted to S according to Equation 11:

$$S = \frac{4\pi \sin(\theta - \theta_{critical})}{\lambda} \quad (11)$$

Note that this has an extra coefficient of 2π compared to Equation 4 and that $\theta_{critical}$ is subtracted off because there is no interference for Θ below $\theta_{critical}$ as there is no x-ray penetration of the sample. This equation favorably converts the Θ axis [degrees] into S [m^{-1}]. For the more advanced reader, this transforms Θ into a value related to momentum space.

Step 5, Fourier Space Analysis

Lastly, we take the inverse Fourier transform of the extracted oscillation shown in Figure 5 in order to quantitatively know the number of layers and their corresponding thicknesses. In MATLAB, this is done using the built in fft() command and then discarding the results that correspond to negative frequencies – a mathematical byproduct of any application of fft(). The final result, expressed graphically, is the approximation of the convolution of the density gradient of the sample with itself as outlined in Equation 7. Figure 6 is the result from the simulation we started with:

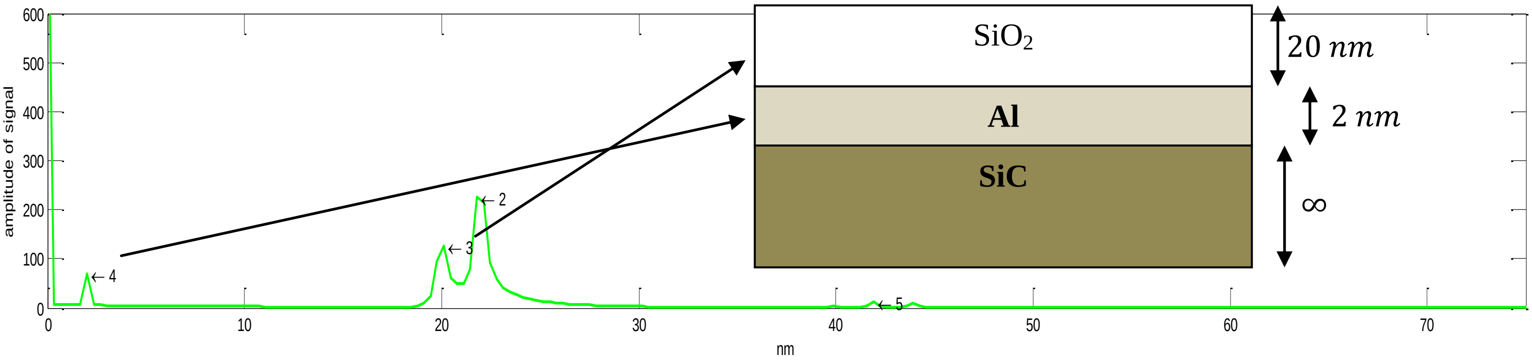


Figure 6 **Fourier transform result**

Note peak 4 corresponding to the 2nm layer and peak 3 corresponding to the 20 nm layer. Peak 2 is simply the sum of these two layers, equaling about 22nm. Also, we see the usefulness of transforming theta[degrees] into S[m^{-1}], which results in an axis expressed in units of meters after the Fourier transform.

III. Results

I will now discuss the results from testing the above algorithm against Multig simulations that included no roughness and roughness. This testing was designed to determine which of the dN or 2N method for Step 3, Fall-off Removal yielded structures closer to ones used for the simulation. Also, the parameters *theta_low_clip*, *theta_high_clip*, and *N* were varied to see how to select them to achieve a consistently accurate algorithm.

i. dN Results With No Roughness Introduced

Using a simulation with no roughness present, such as the simulation of the 2nm of Al and 20 nm of SiO_2 discussed earlier, dN successfully extracted the number of layers present in the model and their corresponding thickness. It should be noted that although the oscillations resulting from applying the dN method in Step 3, Fall-off Removal are $\frac{\pi}{4}$ out of phase with the original high frequency oscillations, it does not alter the frequency of the oscillations and thus does not alter the result of Step 5, Fourier Space Analysis.

One evident shortcoming of the dN method is the existence of a low frequency residue evident in Figure 7:

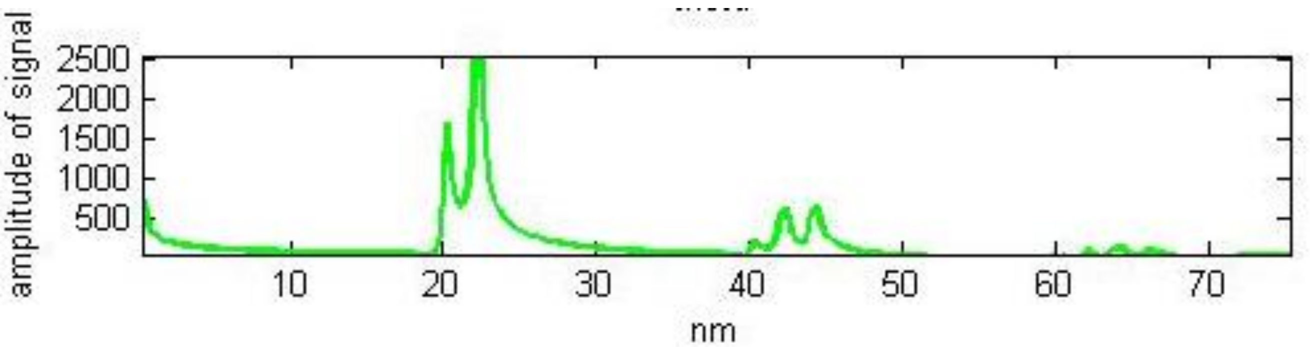


Figure 7 Fourier transform result using the dN method on simulated data equivalent to that in Figure 4. Test run at 12keV, with a *theta_low_clip* of .2 and an *N* value of 3. The original data spans 9 degrees of ϴ.

This gradual low frequency curve spans from approximately 0 to 10 nm and 'drowns out' a peak that should be visible at 2nm in Figure 7. The existence of this 2nm layer is then only revealed through the difference in peaks at about 22 and 20 nm.

Another observation of the dN results in Fourier space is the relatively strong second harmonics signals present around approximately 45 nm in Figure 7. The presence of pronounced harmonics such as these indicates the dataset that is being Fourier transformed is not very sinusoidal in nature.

Lastly, the characteristics of a multilayer stack that are most readily evident from this method are the number and thickness of layers present. This capability has only been tested on one and two layer samples that are situated on a substrate.

ii. dN Results on simulated roughness

Under actual experimental conditions, roughness is present on the layers of any real sample being tested since no semiconductor sample can be perfectly smooth. To simulate this, the same simulation above is analyzed, yet with 1.5 Å of roughness added to the SiO_2 layer. Figure 8 show how the roughness affects the intensity versus Θ signal and the corresponding application of the dN method, while Figure 9 specifically shows the Fourier transform result.

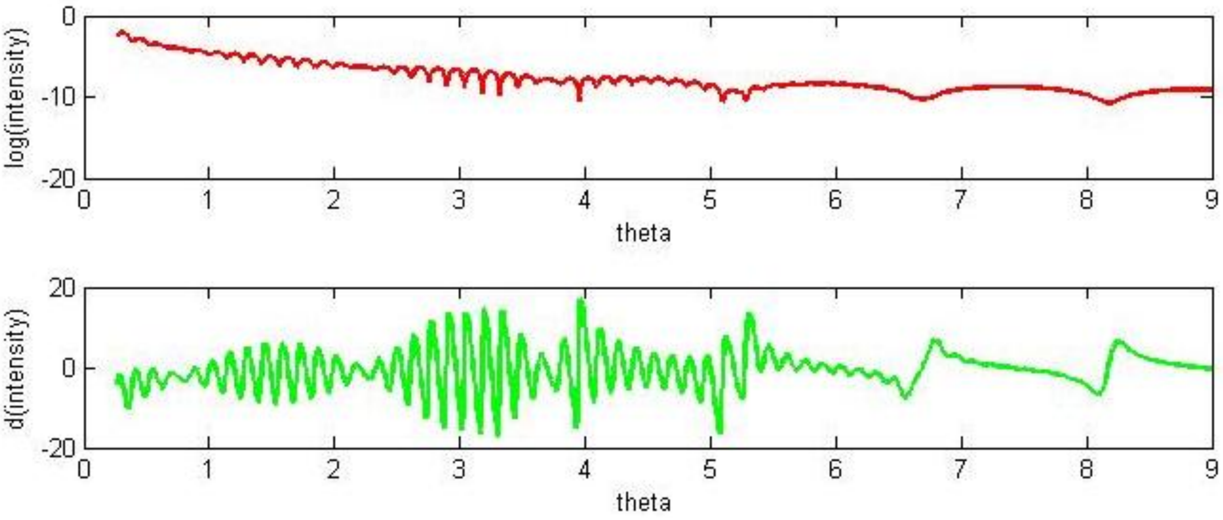


Figure 8 **The top graph is the intensity versus theta dataset for a simulated sample defined in Figure 4 – with 1.5 Å of roughness added to the SiO_2 layer. The second graph is the result of applying the dN method in Step 3, Fall-off Removal.**

Note how the high frequency component is greatly attenuated, especially after 7 degrees of theta. In turn, this affects the derivative in this region too, introducing a pulse like signal.

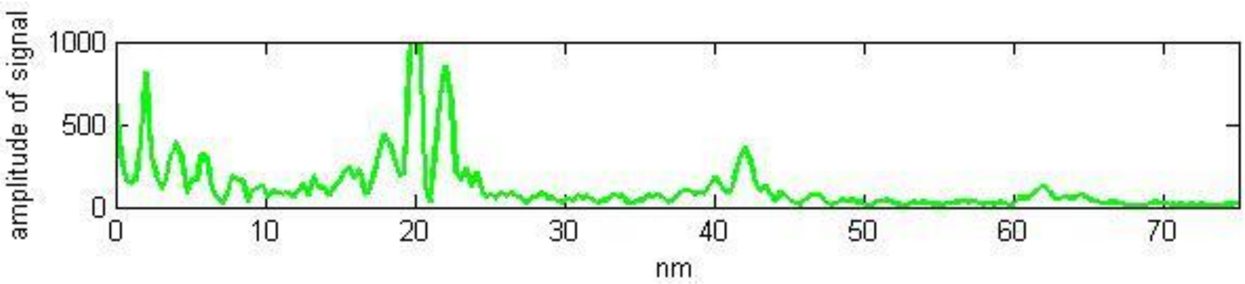


Figure 9 **The <u>Step 5, Fourier Space Analysis</u> result of the data shown in Figure 8.**

Note the many different peaks that have been introduced by the addition of roughness, as compared to Figure 7 when no roughness was present

As can be seen in Figure 8, the roughness greatly attenuates the high frequency oscillation for larger ϴ. Although the Fourier transform result in Figure 9 is very noisy, a sizable peak is now centered over approximately 2nm. Additionally, the peak with the largest amplitude is now over 20nm, rather than the largest amplitude peak representing the summation of both layers (22nm). The Fourier transform also contains many large “ringing” oscillations, which do not correspond to any real layer spacing.

iii. 2N Results With No Roughness Introduced

Using a simulation with no roughness, such as the 2nm of Al and 20 nm of SiO_2 discussed earlier, the 2N method extracted the number of layers present in the model and their corresponding thickness. This success is illustrated in the Figure 10 Fourier transform:

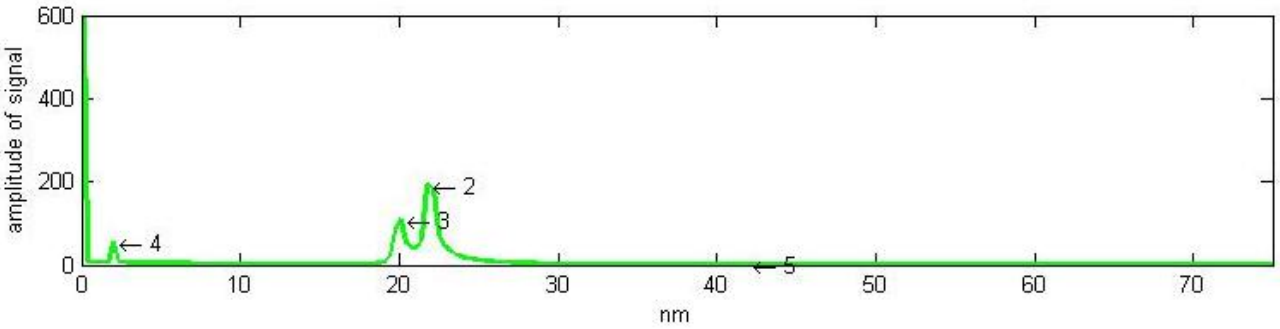


Figure 10 **The Fourier transform result from <u>Step 5, Fourier Space Analysis</u> using the 2N method on the dataset defined in Figure 4 with *N* set to 18 and *theta_low_clip* equal to 0.18˚.**

Note the existence of peak 4, indicating the 2nm layer. This peak is not visible in Figure 7, when the dN method is used.

Observing Figure 10, we see three distinct peaks. The first, peak four, corresponds to a thickness of 2.01 nm while the third corresponds to 20.1 nm. Peak two is approximately the sum of these two layers. We notice that the second harmonics are extremely small as indicated by peak 5, suggesting that dN method enhances the higher harmonic contents. One anomaly is noted for a peak that appears to correspond to 0 nm. The reason for its existence and its physical meaning is not understood. We have only used the Fourier method so far to extract information about the number and thicknesses of layers. It is likely that the height, width and the shape of the peaks in the Fourier space contain information about roughness and density of the layers, but we haven't explored those characteristics yet.

iv. 2N Results on simulated roughness

Figure 11 shows the result of running the 2N algorithm on the same roughness simulation described previously in Figure 8 and 9 in the dN section

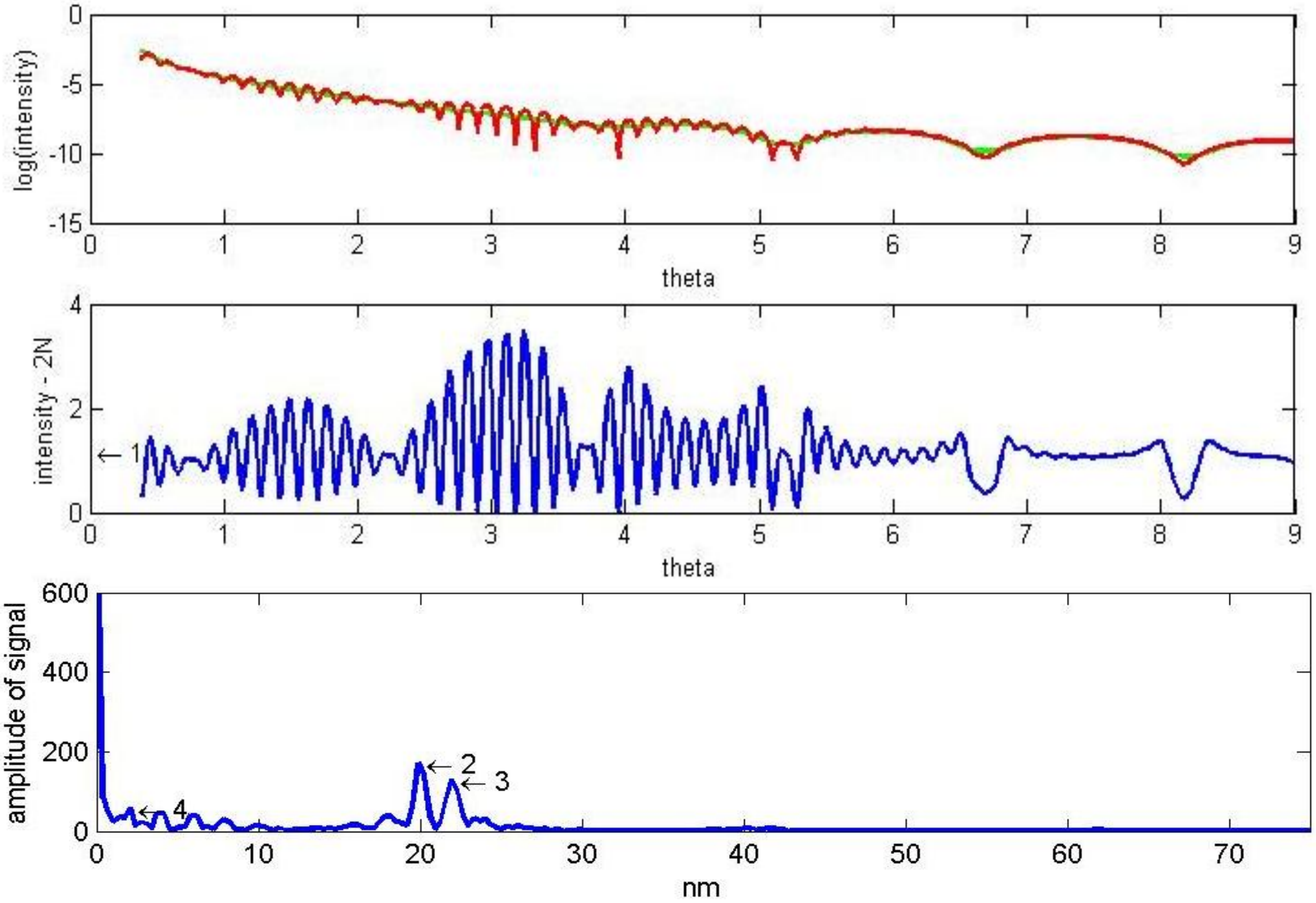


Figure 11 The top graph is the intensity versus theta dataset for a simulated sample defined in Figure 4 – with 1.5 Å of roughness added to the SiO_2 layer. The second graph is the result of applying the 2N method in <u>Step 3, Fall-off Removal</u>. The bottom graph is the Fourier transform result from <u>Step 5, Fourier Space Analysis</u> using the 2N method. *N* is set to 18 and *theta_low_clip* equal to 0.3°.

Note the introduction of noise from the roughness, especially in the low frequency regions.

Just as with the dN simulation, there is the introduction of additional oscillations that do not correspond to any real layer thickness, especially in the low frequency region, though they are not as pronounced as in the case of dN method. These roughness-driven additional oscillations are rounded peaks, while the thickness of the 2 nm layer is denoted by a pointy peak of greater amplitude, as shown in peak 4 in Figure 11. Also, the peak with the largest amplitude (peak 2) now corresponds to 20nm, rather than the largest amplitude peak representing the summation of both layers (peak 4). The same effect occurred using the dN method.

v. Results of altering the theta_low_clip and theta_high_clip bounds on both algorithms

The most sensitive input parameter for both algorithms appears to be the lower Ө bound used in Step 2, Restrict the Ө Domain, a parameter defined as *theta_low_clip*. To highlight this fact, Table 1 illustrates the effect of altering the *theta_low_clip* used in applying the 2N method. For additional context, the location of the *theta_low_clip* is plotted in Figure 12.

·

theta_low_clip Used [˚]	2 nm Layer Result [nm]	20 nm Layer Result [nm]	Total Deviation [%]
🔴 0	1.97	20.03	1.78
🟢 0.22	2.02	19.86	1.78
🟠 0.33	2.38	20.46	23.58
🔵 0.38	2.05	19.89	3.72
⚫ 0.45	2.42	20.40	25.13
🟡 0.75	2.50	20.42	27.50

Table 1 Results of changing the lower bounds of theta, *theta_low_clip* , when the 2N method is applied to the simulation detailed in Figure 4 but with 1.5 Å of roughness added to the SiO_2 layer. *N* is held constant at 20 throughout. The colored dots correspond to the points on the Step 3, Fall-off Removal graph in Figure 12 below.

Total deviation is the percent that the 2nm result differs from the actual value plus the percent that the 20nm result differs from the actual value.

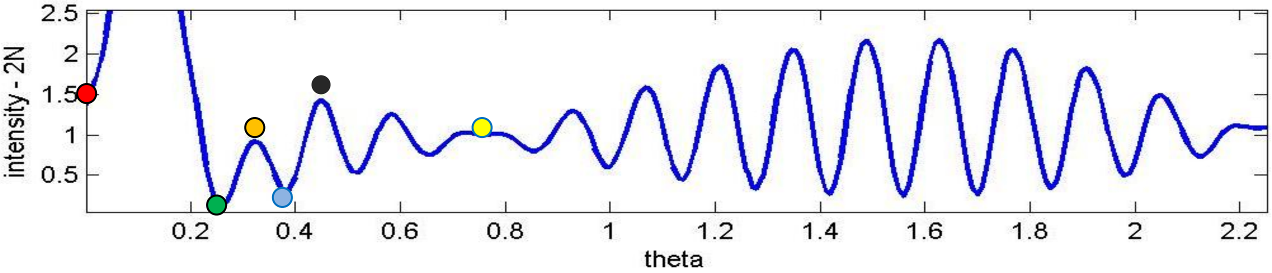
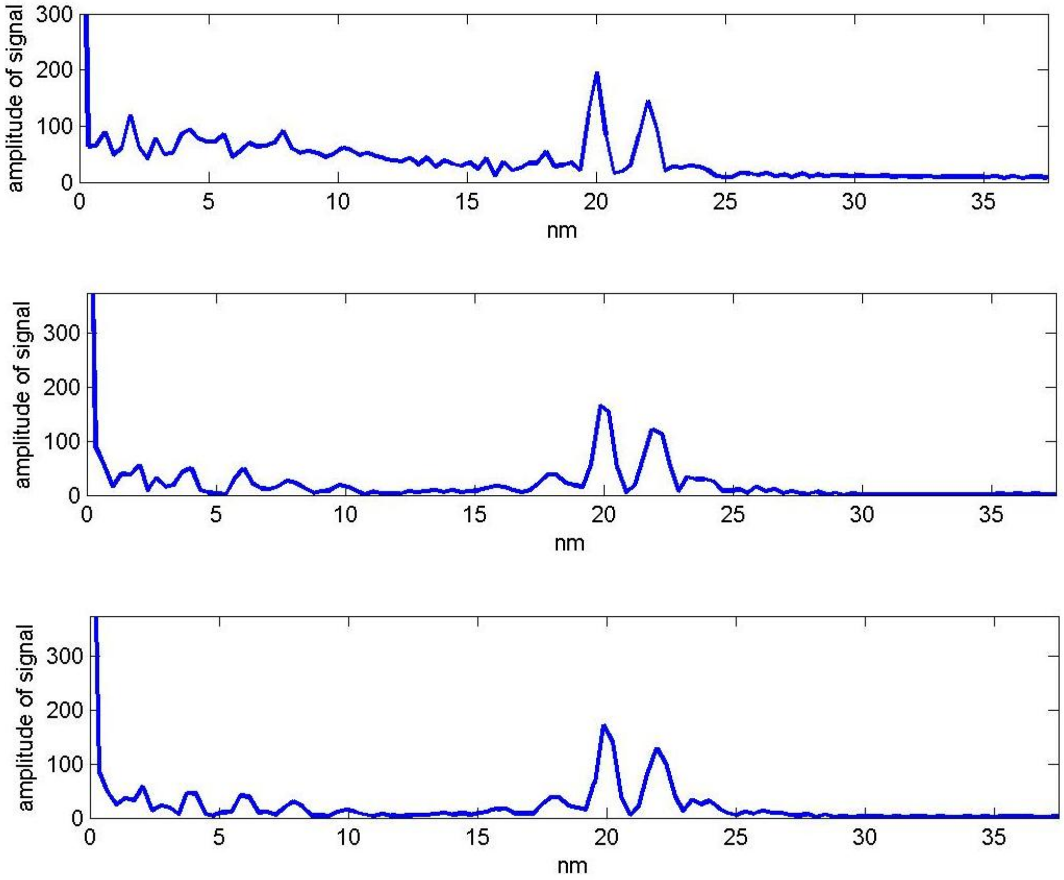


Figure 12 A graph of the first part of the extracted oscillations resulting from the Step 3, Fall-off Removal with colored dots corresponding to Table 1.

Note the large spike below theta equal to 0.2 and how the oscillations then settle into a regular pattern after a theta equal to about 0.8.

To further analyze these findings, Figure 13 displays the Fourier transform of the three *theta_low_clips* that have the smallest deviations.



Theta_low_clip = 0

Total Deviation = 1.78%

Theta_low_clip = 0.22

Total Deviation = 1.78%

Theta_low_clip = 0.38

Total Deviation = 3.72%

Figure 13 Fourier transform graphs, the results of Step 5, Fourier Space Analysis on the dataset described in Table 1. All elements are held constant between the graphs, except for *theta_low_clip* values chosen because they deviated the least in Table 1. Note that although the *theta_low_clip* value of 0.38 thicknesses deviates the most, the false peaks in the transform are rounded, leaving the true peak at approximately 2nm pointed. The peaks in the other transforms are prominently pointed in nature, which could lead to multiple false readings of small layers

Looking at the results of Table 1, Figure 12, and Figure 13 it is observed that choosing the *theta_low_clip* to be the second minima in the high frequency oscillation, 0.38 ° in the example case, yields the best combination of layer thickness and true layer detection accuracy. Choosing alternative *theta_low_clip*s that are more accurate in terms of layer thickness appear to falsely detect other thin layers.

Additional tests found that using a *theta_high_clip* always lowered the accuracy of the algorithm when used against various simulations. Tests were not tried on actual data.

v. Results of altering N value used for the 2N method

It was determined that the most sensible way to choose an appropriate N value was to observe the visual output of Step 3, Fall-off Removal, both the fall-off line created by the 2N method and the corresponding graphic with this fall-off line removed. As noted in Figure 14 and

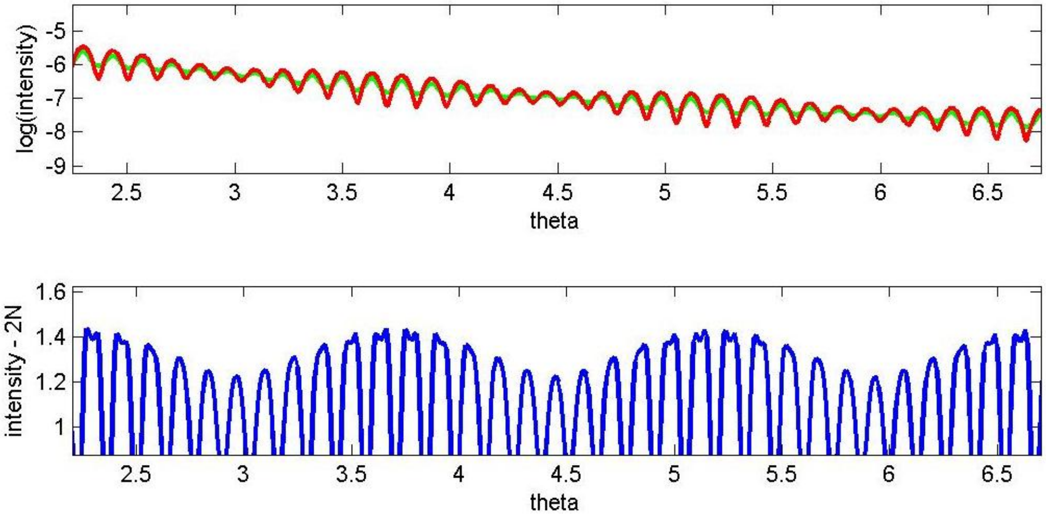


Figure 14 An *N* value of 4 is used on the dataset defined in Figure 4. The green line in the top picture indicates the fall-off the 2N method is calculating.

In this case, *N* is too small, thus the green fall-off calculation too closely matches the original function resulting in the ‘clipping’ seen in the lower blue oscillation graph.

Figure 15, an appropriate N value allows the full oscillation to be extracted, rather than an oscillation with its peak clipped – such as Figure 14.

It is concluded that an *N* value equivalent to half the points in one period of the high frequency oscillation accurately maps the fall-off function present in the original data. Mathematically, this is intuitive because the 2N method will span a full period of an oscillation when it takes a local average.

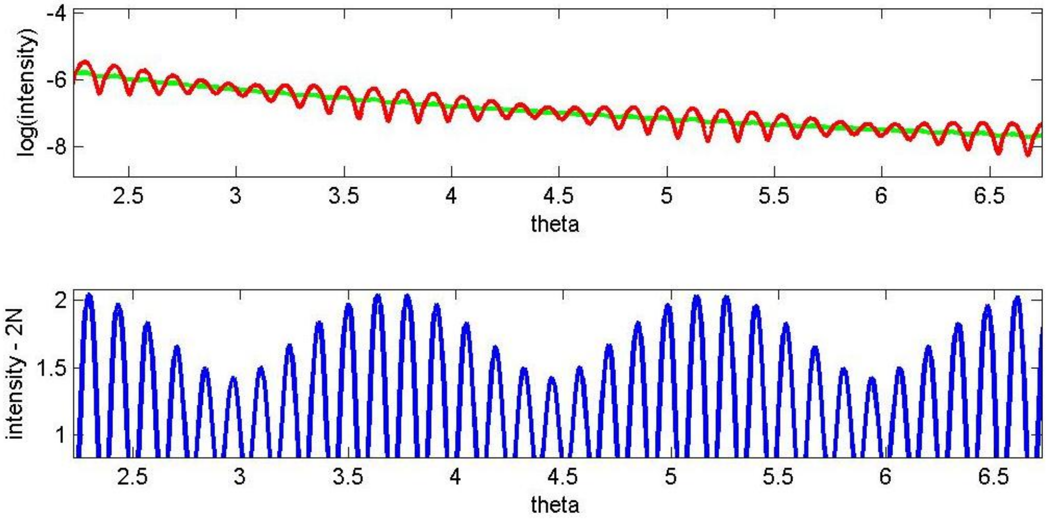


Figure 15 **An *N* value of 6 is used on the dataset defined in Figure 4. The green line in the top picture indicates the fall-off the 2N method is calculating.**

In this case, the green line accurately captures the fall-off with N chosen to be half the number of points in a period of the high frequency oscillation.

IV. Discussion & Conclusion

Separating the reflected beam intensity into a smooth, monotonic falloff term and an oscillatory terms, and then Fourier transforming the oscillatory terms is a quick and powerful way of extracting key characteristics of the layers from a sample containing multiple layers.

Based on the simulation test results described in the Results section, it appears that 2N fall-off subtraction distorts the Fourier space data less than the dN method. Therefore, it is recommended that the 2N algorithm be employed to assist in extracting information from the

XRR data, particularly the number of layers present and their thicknesses. The key distinction that sets the 2N algorithm apart from the dN algorithm is its low frequency sensitivity, a characteristic that is critical to the mission of detecting thin layers on samples. The low frequency artifact that exists in the dN method, as manifested in Figure 7, is simply unacceptable.

As noted in the Results section, it is recommended a *theta_low_clip* corresponding to the second minima of the high frequency oscillation be used with the 2N method. This recommendation is done with caution, due to the sensitivity of this parameter and it's recommended that this parameter's affect on the output of the algorithm be studied additionally. No *theta_high_clip* should be used, however, it may be useful to remove a high theta region of experimental data that traditionally contains a high noise to signal ratio – this needs to be further explored. Experience has determined that using an *N* value equal to at least half the number of points in the high frequency period yields the cleanest extraction of the fall-off function.

It is concluded that using the 2N fall-off removal method combined with the Fourier analysis is a powerful technique that can reveal the presence of nanoscale layers and their corresponding thicknesses. This study did not focus on the intensity and the shape of the peaks in the Fourier space, but there is some preliminary evidence that they may hold information about layer densities and order. It is recommended that this be explored in more detail.

Even if no connection is established, the current capabilities of this technique are very useful in analyzing new semiconductor structures that often contain multiple layers of varying thicknesses. Just knowing the number of layers present and their thicknesses eliminates many important unknown variables, making modeling of additional characteristics such as layer order and density less labor intensive. This partial information is often sufficient enough to guide the generation of appropriate stratigraphic models which can then be refined against the raw XRR data using conventional goodness-of-fit metrics, ultimately resulting in a reliable characterization of complex multilayer samples.

V. Acknowledgments

This work was supported by the SULI Program, U.S. Department of Energy, Office of Science. I would like to thank my mentor Apurva Mehta and Matthew Bibee, a graduate student at Stanford and a former SULI intern, for their assistance in addition to all the administrators of the SULI program for this wonderful and enriching experience. Finally, I would like to thank Yan Gao, of GE, for providing interesting samples for testing the effectiveness of the analysis approach taken here.

VI. References

[i] Moore, Gordon E. (1965). "Cramming more components onto integrated circuits" . Electronics Magazine. pp. 4.

[ii] Toney, Michael and Brennan, Sean. “Measurements of carbon thin films using x-ray reflectivity”. Journal of Applied Physics 66 (4). 15 August 1989.

[iii] Discussion with Apurva Mehta. August 4th, 2011.

[iv] Fronheiser, Jody and Tilak, Vinayak. “Motivation for SiC Device Development and Relevance to NIST/GE Contract on SiC/SiO2 Interface Measurement Technologies”. March 27th, 2011.

[v] O. Durand. “Characterization of multilayered materials for optoelectronic components by high-resolution X-ray diffractometry and reflectometry: contribution of numerical treatments”. Thin Solid Films Volume 450, Issue 1, Proceedings of Symposium M on Optical and X-Ray Metrology for Advanced Device Materials Characterization, of the E-MRS 2003 Spring Conference. 22 February 2004. Pg. 51-59

[vi] O. Durand.

[vii] Toney, Michael and Brennan, Sean

[viii] Poust, Benjamin and Goorsky, Mark. “Enhanced Fourier Transforms for X-Ray Scattering Applications”. Fourier Transforms – Approach to Scientific Principles. InTech. 2011

[ix] Poust, Goorsky

[x] Poust, Goorsky

SLAC-TN-12-014

Research and Design of a Sample Heater for Beam Line 6-2c Transmission X-ray Microscope

Veronica Policht

Office of Science, Science Undergraduate Laboratory Internship (SULI)
Loyola University, Chicago
SLAC National Accelerator Laboratory
Stanford, CA

August 12, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Johanna Nelson at the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory.

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Research and Design of a Sample Heater for Beam Line 6-2c Transmission X-ray Microscope.

VERONICA POLICHT (Loyola University, Chicago, Chicago, IL 60660) JOHANNA NELSON (Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Stanford, CA 94025)

There exists a need for environmental control of samples to be imaged by the Transmission X-Ray Microscope (TXM) at the SSRLs Beam Line 6-2c. In order to observe heat-driven chemical or morphological changes that normally occur *in situ*, microscopes require an additional component that effectively heats a given sample without heating any of the microscope elements. The confinement of the heat and other concerns about the heaters integrity limit which type of heater is appropriate for the TXM. The bulk of this research project entails researching different heating methods used previously in microscopes, but also in other industrial applications, with the goal of determining the best-fitting method, and finally in designing a preliminary sample heater.

INTRODUCTION

Facilities

SLAC National Accelerator Laboratory is one of 13 facilities in the United States equipped with a permanent source of synchrotron x-rays. Synchrotron radiation is produced at SLAC in the Stanford Synchrotron Radiation Lightsource (SSRL) by an electron beam subject to magnetic manipulation along a curved path. The bending of the electron beam's path results in a change in the beam's momentum and the emission of synchrotron x-rays which can be used to probe structures on the molecular level. The SSRL is equipped with facilities for several types of characterization techniques including x-ray diffraction, x-ray absorption spectroscopy, and macromolecular crystallography. Another type of technique which utilizes synchrotron x-rays is Transmission X-ray Microscopy (TXM). Transmission X-ray Microscopy is used to produce images of the sample in a way similar to Transmission Electron Microscopy (TEM), but with several marked advantages.

Beam Line 6-2c at the SSRL consists of a TXM, which utilized intensified synchrotron x-rays from an electron beam fed through a 54 pole wiggler. A beam of synchrotron x-rays builds as the electron beam travels sinusoidally through the wiggler's magnets of alternating polarity. Transmission X-ray Microscopy is a full field microscope, similar to a visible light microscope or TEM, that functions by training a focused, monochromatic beam of x-rays onto a sample where the rays are either absorbed or transmitted. Data is collected from the transmitted rays by a detector placed downstream of an objective lens. The TXM at Beam Line 6-2c can be characterized by the high image resolution of 30 nm, large depth of field (20 - 60 μm), and a variety of imaging techniques. Unlike TEM, the high penetration depth of the synchrotron x-rays allows thick samples to remain largely intact and make it possible to image a complete structure. The x-rays range in energy from 5 - 14 keV and by changing the energy of the x-ray beam incident on the sample, information can be gathered about the chemical structure of a sample. An advantage of using TXM for imaging is the ability to easily construct 3D tomography of a sample by collecting 2D images at many angles. The TXM uses hard x-rays to image a wide variety of samples including geological samples, lithium-ion batteries, and biological samples [1].

There is very limited available space around the sample stage of the TXM on Beam Line 6-2c. Between the pinhole, where the x-ray beam exits, and the zone plates downstream of the sample is roughly 1 inch of space for the sample and any additional components, as depicted in Fig. 1.

Introduction of Problem

To gain a better understanding of a sample's morphology and behavior, some samples require *in situ* conditions be maintained during imaging. Certain molecular and chemical reactions only occur at pressures and temperatures outside of those present in the open air of the TXM hutch and therefore need to be placed in an artificial reaction environment with an additional microscope component. The focus of this project has been to develop a heating device to be used at Beam Line 6-2c for the samples previously alluded to as well as other potential sample types.

The important characteristics of a heating device are as follows: it delivers well controlled, confined and measurable heat within a given temperature range and is capable of maintaining a given temperature for several hours as data is gathered. The heater will ideally perform between 50-500 °C. Because the Beam Line equipment is also used by other users, the design of this heater must be easily installed and removed. Additionally, the heater device must physically be able to fit in the sample space of the TXM. A final consideration is the adaptability of the heater for different sample types; the two most general sample formats used in the TXM are either flat samples or samples contained within a capillary. An example of a heaters applicability is the real-time imaging of thermally driven reactions or the observation of a samples behavior and character across a temperature spectrum. For example, a recently proposed project for use of SSRL facilities would require their samples be kept at a temperature between 65 and 80 °C. The proposers wish to conduct TXM and X-ray Absorption Near Edge Structure (XANES) imaging in order to study the nucleation and growth of a ZnO crystal film [2]. XANES is used to better understand chemical composition by examining the absorption of x-rays below, throught, and aboce the x-ray energy necessary to excite a core electron of a specific element.

There are many uses for heating devices in lightsource science and the ability to heat a sample adds another dimension to our understanding of elemental behavior over a spectrum of temperatures. Previously, the use of sample heaters in visible light and electron beam microscopes has successfully imaged the performance of materials at elevated temperatures. Therefore, by examining these heating devices, I plan to gather information for the design of a similar heating device to be used for x-ray light based equipment. The heating methods under consideration in this paper are three types of microelectromechanical resistive heating and two types of infrared heating: silicon wafer-based resistive, Kapton-based resistive, ceramic-based resistive, laser-based and infrared-based.

Introduction to Sample Heaters

A heater system is composed of the heating apparatus and a measurement apparatus. The two general methods of heating we have considered for use in our sample heater are infrared radiation (IR) and resistive heating methods. Infrared heat sources can be either coherent like lasers or incoherent like heat lamps. Infrared radiation trained onto the sample face is partially reflected and partially absorbed, resulting in sample heating. Heat lamps are radiative heaters, which consist of a heating element such as a metal filament that upon application of a large current radiates infrared light. Resistive heaters function by running a current through a metal wire with some internal resistance such that internal collisions of the moving electrons generate heat. This heat dissipates over the surface of the metal electrode and through the air. These general methods for heating function based upon Joule heating that running a current through a resistor will result in heat: $Q \propto I^2R$, where Q is heat, I is current, and R is resistance. Resistance is a characteristic quality of a material and depends upon cross sectional area. Note that heat is only proportional to the right side of the equation.

There are several methods by which the temperature of a heater can be determined. Those considered in this project are direct measurement with a thermocouple, calibration of the heater before use, and calculating the local electrical resistance. All three of these methods are based upon electronic principles and take material properties into account. Calculating the temperature of a resistive heater requires information about the metal's resistance and the current as Q is proportional to power, P: $Q \propto P$. A thermocouple functions by measuring the current gradient across a junction of two different types of metal. By knowing the resistance of the respective metals you can calculate the temperature at the junction with the information about the current. There are few methods of heater calibration including using a heater on a test sample which has very well understood transitional properties. By comparing the well known materials transition states, such as melting, with the associated voltage and current applied to the heater you can calibrate the heater's resulting temperature.

For the heating methods which require radiation to travel through the air, the Planck radiation function effectively describes emitted power per unit area of the emitting element:

$$I = \frac{2hc^2}{(e^{\frac{hc}{\lambda kT}} - 1)\lambda^5}, \tag{1}$$

where I is the energy per unit time per unit surface area, λ is the wavelength of the emitted light, T is the temperature of a black body, h is the Planck constant, c is the speed of light, and k is the Boltzmann constant.

METHODS

My portion of this project consisted of investigating and working on a preliminary design for a heating device that would ideally adhere to all the previously stated parameters of removability and the constraints of its heating profile. After researching different methods of heating, I constructed a decision matrix. From this decision matrix I have deduced the best heat source options for Beam Line 6-2c. At this point I designed a preliminary heating apparatus, but due to time constraints did not see through to the beginning stages of fabrication.

The method by which research was conducted began with searching for any academic papers that used a given heating method. The papers that contained pertinent information were collected and used for reference material for constructing the decision matrix. Papers were either found by conducting Internet searches using Google Scholar, ArXiv, or the Stanford academic journal search engine [3, 4, 5]. In some instances a certain paper was found in another paper. Where there were no academic papers available relating to a certain heating method, information was gathered on the heating device from textbooks, manufacturer websites, and speaking with manufacturers' engineers. Textbook resources were discovered through Google Books and the manufacturer websites were found through Google search [6].

Based upon the gathered information, I then constructed a decision matrix; the decision matrix compares and ranks each heating method based upon how well it meets the previously determined desirable characteristics. Each of these characteristics has been given a weight 1 through 3, where a weight of 3 denotes more importance than a weight of 1. Finally, the ranks given to the heaters are multiplied by the characteristic's weight, all of which are then summed. The heating method that has the lowest weighted rank is determined to be the best fit for our needs [Tables 3 & 4].

RESULTS

Infrared Heaters

Infrared Heat Lamp Heaters

Infrared heat lamp heaters are on the whole very effective at heating large areas with high intensity. The heat emitted by these lamps depends upon their color temperature, the temperature at which a black body would radiate with the same color. There exists commercially available infrared heating filaments, which operate within the desired temperature range decided upon for our sample heater. The main challenge of designing an infrared heat lamp heater for use in the TXM is managing to confine the heat to the sample face. As Fig. 1 shows, the components of the TXM near the sample stage are very close to where the desired heat would be. In order to preserve the integrity of these TXM elements we must ensure the elimination of all stray heat from the sample. To combat this problem we could use a small IR heat filament and create an apparatus such that the beam of infrared radiation would be focused and manipulated onto the sample. A preliminary design for such a heating apparatus is pictured in Fig. 2, and features the use of a reflecting mirror to direct the beam and a Cassegrain reflector to focus the beam to a small spot size. The use of a Cassegrain reflector is based upon the design for a Fourier Transform Infrared Spectrometer [7]. Another way to confine the IR heat would be implementing radiation-containment chamber made of ceramic.

Overall using an infrared heat lamp would be relatively costly in terms of purchasing components and the labor that would be required to build a viable heater. An infrared heat lamp would be difficult to use especially when the sample stage must be moved during imaging. A typical heat lamp alone costs around $200. A recommendation I would suggest for making such a system easier to use would be placing the reflecting mirror and Cassegrain reflector on a mechanical stage, which could be controlled by computer. In this way, the two moveable components of the heater could be configured to move in sync with the sample stage. A final aspect to consider is the trade-off between filament size and the emissivity of the filament. Emissivity a is dimensionless quantity that relates a material's radiation of energy to the energy emitted by a black body at the same temperature. Emissivity depends on several factors including the thickness of a material; thinner materials will have a reduced emissivity.

Laser Heaters

Lasers are reliable, coherent heat sources; due to their coherence they can be manipulated in terms of intensity, beam size, polarization, and direction fairly easily. The use of an infrared laser to heat samples has been previously used in an x-ray setting in laser-heated diamond anvil cells [8, 9]. These cells simultaneously pressurize and heat a sample and are able to reach temperatures above 1200 °C. In these instances, a double-sided laser technique is implemented using Nd:YLF lasers. The use of two different lasers allows for minimal axial temperature gradient on the sample face. This is done by choosing to have the lasers function in complementary modes, one in Gaussian and the other doughnut-shaped modes. As a result, the axial temperature gradient is approximately 30 °C, whereas a similar single-laser system has an axial temperature gradient of hundreds of degrees. Another positive aspect of double-sided heating is its level of precision; this system implements several optical devices to obtain a heating area of 10 - 25 μm [9].

The drawback of having the double-laser system is that the apparatus, pictured in Fig. 3, requires a lot of space and the entire set-up would be very expensive. A laser similar to those recommended, Diode Pumped Infrared 1030 nm CrystaLaser, is $7,600, including its power supply [10]. A third reason this method is a poor fit is that the fabrication is too complicated for the scope of this project. This is a very precise and controlled heating method though, so I would recommend it in the case where someone could be dedicated to the design and fabrication of this specific apparatus.

Resistive Heaters

Resistive heaters include a broad range of heaters which most generally are composed of a base substrate with a metal electrode either embedded in or deposited on top of the base. Some resistive heaters have additional layers of insulation and electrodes placed on top of the two previously mentioned most basic layers. The heat character, dimensions, and physical character depend mostly upon the materials used for the base substrate and the metal electrode. For this reason, I have addressed three common types of resistive heaters; silicon-based, Kapton-based, and ceramic-based; separately. Figure 4 shows 3 examples of each of the three previously mentioned kinds of resistive heaters. Between the three types of resistive heaters the main areas of disagreement are the 1) operational temperature range, 2) cost, 3) manufacturing process difficulty, and 4) sample applicability.

Micro-Hotplate/Microelectromechanical System (MEMS) Heaters

Micro-hotplates and MEMS are fabricated using a silicon wafer as their base substrate forming a frame and acting as a heat sink. Upon a silicon frame there is a membrane, which forms a window for radiation to penetrate. These heaters have been used as sample heaters in TEM systems because of their relative strength and thermal properties. Silicon has a high heat capacity and is strong at thicknesses above 50 nm, but is brittle at thicknesses less than 50 nm. Because of silicon's high heat capacity there is a large heat-drop off between the metal electrodes and the heater stand. The fabrication method of these silicon-based heaters has been performed and re-worked by many different research groups and as such these heaters have been developed to function in a broad temperature range. MEMS and micro-hotplates are able to operate at a wide range of temperatures with a highest mentioned operating temperature of 800 °C and the average operating temperature of 500 °C [11, 12, 13, 14, 15, 16].

The silicon-based heaters perform well and offer a homogeneous heat distribution on the sample. Beckel, *et al.* specifically mentioned that the use of an integrated heater to anneal a crystallized film sample caused it to crack, but also that the annealing of said film took 15% of the time it took using other techniques [14]. Another paper tracked the drift of nanoparticles in their FOV and concluded that the heat from a spiraled Pt heater was well distributed. The findings are made based upon how consistent the drift rate of said particles was as it moved along a given path, giving a good measure of local heat gradient [17]. The use of resistive heating could effectively heat the sample at a steady, tunable, and readable temperature. Several previously used systems that require the use of a micro-hotplate have used either a heating element with multiple connections such that it can simultaneously be controlled and read, or a system with dedicated components for heating and for temperature measurement.

Several of the papers that I have read have stated in various levels of detail the exact fabrication methods which indicate that this would be a difficult apparatus to pursue without outside help. These apparatuses are produced using several nanofabrication and chemical processes and would only be able to be performed in the cleanroom environment of an outside nanofabrication lab. Generally, fabrication employs several machining and microfabrication techniques. A few examples of the type of processes are lift-off technique for the deposition of the Pt heating element and low-pressure chemical vapor deposition for the placement of a SiN membrane layer on top of the heating element [11, 12, 13, 14, 15, 16]. Additionally, E.A. Olson, *et al.* details specific instructions on how to best choose your conducting material [16].

There are also many manufacturers and distributors of silicon-based resistive heaters. Many of the companies that produce these also offer customization services to ensure that the heater meets the customers specifications. The trade off between fabricating the apparatus ourselves and buying one from an outside source is cost. One silicon-based micro-heater (10 mm x 5 mm x 0.2 mm, gold electrode, 350 nm thick) is $82.00 without any customization; requesting custom dimensions increases the price [18]. Though one heater is not on it's own expensive, most companies have minimum order sizes and the outcome is a large, expensive order of heaters. Another drawback of using the silicon-based heaters is that it, because silicon is not very flexible, it would only apply homogeneous heat to flat samples. A final thing to consider is that, as already mentioned, the devices were designed for and used for TEM and SEM measurements. The use of the TEM and SEM places the micro-hotplate in a vacuum and as such we cannot assume that our observed heat drop-off will not be the same as what had been observed in their papers. That said, the spacing between the electrode patterns would be larger than the x-ray beam diameter and so, expecting that the heat will be homogeneous in that region is not a terrible assumption.

Kapton Heaters

Kapton heaters use Kapton film as their base substrate. Kapton film is characterized by its excellent thermal and physical properties, which make it a good component for a prospective sample heater. Kapton film's limits have been well tested and so many of its thermal properties are known. Kapton has a thermal conductivity on the order of $10^{-1}\ \frac{W}{mK}$, which correlates to high insulative properties. For comparison, Table 1 lists several materials and their thermal conductivities [19].

There are several companies which specialize in the production of Kapton heaters. The majority of these companies advertise customizability in terms of dimensions of the heater and of the metal foil materials used for the metal electrode. The dimensions of the heater, such as thickness and distance between the electrodes, depend upon the metal material chosen.

The problem met with using a Kapton-based heater is the maximum operational temperature. Table 2 shows three Kapton-heater companies and the associated maximum temperatures attainable. These listed temperatures are well below what we had hoped for in a maximum heating temperature (500 °C). Kapton film itself maintains its structural integrity up to temperatures as high as 400 °C [19]. This would indicate that the temperature limitations stem from the metal used for the electrode or the adhesive used in fabrication.

Overall I think that Kaptons physical properties would make it a good material for a heater. Kaptons flexibilty means it could homogeneously heat both planar surfaces and surfaces which are gently curved. However, as far as prefabricated heaters go those available will not operate in the temperature range we require.

Ceramic Heaters

Ceramic materials are used as heater substrates because of their favorable physical and electrical properties including good strength, high temperature stability, low dielectric constant, and high electrical insulation. Ceramic infrared heaters are often used in industrial manufacturing as well as domestic and pet care applications. More so than silicon- and Kapton-based heaters, ceramic-based heaters are limited in terms of the dimensions at which they can be produced to function. Ceramic heaters can be manufactured in many shapes including squares, circles, and cylinders with the metal wires being embedded on the face for the square and circle and within the inner portion for the cylinder. This offers some flexibility in applicable samples as you could potentially use a square or circle for flat samples and a cylindrical for a capillary-bound sample. Unfortunately, the heat is not easily confined in the case of the square or circular heater with there being too much risk of the TXM elements surrounding the sample being affected by stray heat. This leaves the cylindrical heater, which is an excellent option for samples contained within a capillary.

There are several cylindrical ceramic based heaters available commercially and offer some level of customizability. A ceramic heater generally consists of a ceramic shell with a metal coil embedded in the wall of the shell. The wires are seldom left uncovered as they reach very high temperatures; clamshell furnace elements reaching temperatures greater than 1100 °C [20]. The metal wires are usually either covered completely with a thin layer of ceramic or are partially covered by ceramic bars.

Another option is fabricating a ceramic-based heater by threading a hollow cylinder of Macor machinable ceramic with a metal wire. Macor retains structural integrity up to 1000 °C and has similar chemical and thermal properties to ceramics used in the fabrication of commercial furnaces. A 1/2” diameter 3 long cylinder of Macor costs $32.00 [21]. Fabrication of this type of heater would be relatively easy with the most difficult portion being threading the wire in the Macor. A preliminary design of a ceramic-based heater produced in-house at SLAC is pictured in Fig. 7.

CONCLUSIONS

Based upon the decision matrix in Tables 3 and 4, the best option for our sample heater is the Kapton-based resistive heater. While the Kapton heaters are not able to function through the entirety of our desired temperature range, the benefits in terms of size, cost, and ease of production make Kapton a good option for the current needs of Beam Line 6-2c. Because the Kapton heater only is applicable to flat or gently-curved surfaces and because it can only operate up to 200 °C, we have chosen a second heating type to cover a wider sample range. The second heater type is the cylindrical ceramic-based heater. Ceramic heaters are relatively easy to produce, have several methods of measuring the temperature, and operate very well within the desired temperature range. The combination of these two heating methods covers most sample types for imaging in the TXM and covers the characteristics we desired in a sample heater. Figures 5 and 6 show SketchUp images of the proposed Kapton-based sample heater to be used on flat and gently-curved sample types [22]. The heater is based upon a Kapton heater distributed by OMEGA Industries [23].

ACKNOWLEDGMENTS

I would like to thank the Department of Energy and the faculty at SLAC who have made this internship possible for myself and many other undergraduate students. I would like to thank Johanna Nelson for mentoring me throughout this project and for ensuring that I had a meaningful experience at SLAC. Finally, I'd like to thank the other SULI student interns who helped make this summer wonderful.

REFERENCES

[1] J. Andrews, S. Brennan, C. Patty, K. Luening, P. Pianetta, E. Almeida, M. C. H. van der Meulen, M. Feser, J. Gelb, J. Rudati, A. Tkachuk, and W. B. Yun, "A high resolution, hard x-ray bio-imaging facility at ssrl," *Synchrotron Radiation News*, vol. 21, no. 3, p. 26, 2008.

[2] A. Cruickshank, "In-situ txm of the nucleation and growth of electrodeposited zno nanocrystals," June 2011, proposed Usage.

[3] Google Scholar Database. [Online]. Available: http://scholar.google.com/

[4] ArXiv: Open access Scholarly Database. [Online]. Available: http://arxiv.org/

[5] Stanford Research E-Library. [Online]. Available: http://www-group.slac.stanford.edu/library/

[6] Google Books Internet Database. [Online]. Available: http://books.google.com/

[7] S. Slobodan and Y. Ozaki, Eds., *FT-IR Imaging Hardware.* John Wiley and Sons, 2011, sections 3.1-3.2.

[8] Y. Meng, G. Shen, and H. K. Mao, "Double-sided laser heating system at hpcat for in situ x-ray diffraction at high pressures and high temperatures," *Journal of Physics: Condensed Matter*, vol. 18, no. 25, p. S1097, 2006. [Online]. Available: http://stacks.iop.org/0953-8984/18/i=25/a=S17

[9] G. Shen, M. L. Rivers, Y. Wang, and S. R. Sutton, "Laser heated diamond cell system at the advanced photon source for in situ x-ray measurements at high pressure and temperature," *AIP*, vol. 72, no. 2, pp. 1273–1282, 2001.

[10] T. Zaki, August 2011, person correspondance with CrystaLaser technical engineer. [Online]. Available: www.crystalaser.com

[11] D. Briand, A. K. B. van der Schoot, U. Wiemar, N. Barsan, W. Gpel, and N. de Rooij, "Design and fabrication of high-temperature micro-hotplates for drop-coated gas sensors," *Sensors Actuators B*, vol. 68, pp. 223–233, 2000. [Online]. Available: www.elseview.com/locate/snb

[12] E. de Smit, I. S. J. F. Creemer, G. H. Hoveling, M. K. Gilles, T. Tyliszczak, P. J. Kooyman, H. W. Zandbergen, C. Morin, B. M. Weckhuuysen, and F. M. F. de Groot, "Nanoscale chemical imaging of a working catalyst by scanning transmission x-ray microscopy," *Nature*, vol. 456, pp. 222–226, 2008.

[13] M. Z. nd E.A. Olson, R. Twesten, J. Wen, L. Allen, I. Robertson, and I. Petrov, "In situ transmission electron microscopy studies enabled by microelectromechanical system technology," *J. Mater. Res.*, vol. 20, no. 7, pp. 1802–1807, 2005.

[14] D. Beckel, D. Birand, A. Bieberle-Hütter, J. Courbat, N. F. de Rooij, and L. J. Gauckler, "Electrical conductivitiy of lscf thin films for sofc on micro hotplates," work performed by members of ETH Zürich and the University of Neuchatel.

[15] M. Jeule and L. Gauckler, "Miniaturised arrays of tin oxide gas sensors on single microhotplate substrates fabricated by micromolding in capillaries," *Sensors and Actuators B*, vol. 93, pp. 100–106, 2003. [Online]. Available: www.elseview.com/locate/snb

[16] E. A. Olsen, M. Y. Efremov, M. Zhang, Z. Zhang, and L. H. Allen, "The deisgn and operation of a mems differential scanning nanoncalorimeter for high-speed heat capacity measurements of ultrathin films," *Journal of Microelectromechanical Systems*, vol. 12, no. 3, pp. 355–364, June 2003.

[17] J. F. Creemer, S. Helveg, G. H. Hoveling, S. Ullmann, A. M. Molenbroek, P. M. Sarro, and H. W. Zandbergen, "Atomic-scale electron microscopy at ambient pressure," *Ultramicroscopy*, vol. 108, pp. 993–998, April 2008.

[18] M. Cox and P. Anastasi, July 2011, person correspondance with Silson Ltd regarding the purchasing of a micro-hotplate. [Online]. Available: http://www.silson.com/

[19] *Summary of Properties for Kapton Polyimide Films*, DuPont. [Online]. Available: http://www2.dupont.com/Kapton/en_US/assets/downloads/pdf/summaryofprop.pdf

[20] The Mellen Company; produce ceramic cylindrical furnaces and heaters. [Online]. Available: http://www.mellencompany.com/

[21] Astro Met, Inc.; distributor of Macor glass ceramic produced by Corning Incorporated. [Online]. Available: http://www.astromet.com/macor.htm

[22] Free 3D Modeling Software by Google. [Online]. Available: http://sketchup.google.com/

[23] OMEGA; produce Kapton Flexible Heaters. [Online]. Available: http://www.omega.com/

[24] G. J. Nelson, W. M. Harris, J. J. R. Izzo, K. N. Grew, W. K. S. Chiu, Y. S. Chu, J. Yi, J. C. Andrews, Y. Liu, and P. Pianetta, "Three-dimensional mapping of nickel oxidation states using full field x-ray absorption near edge structure nanotomography," *AIP*, vol. 98, no. 17, p. 173109, 2011.

[25] Newport Corporation; produce IR Heating Elements. [Online]. Available: http://newport.com/

FIGURES

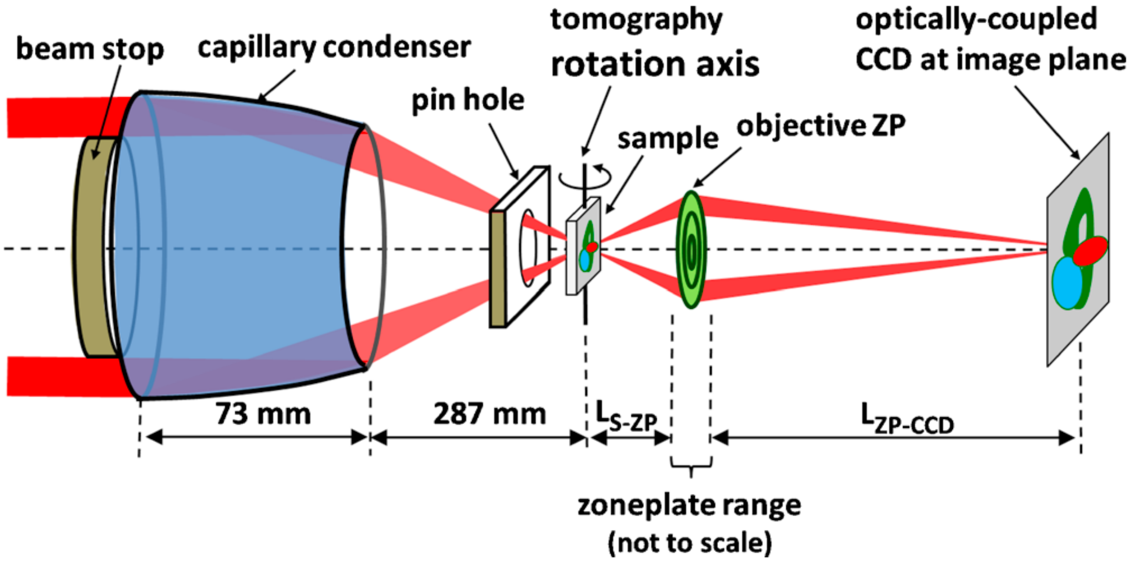


Figure 1: TXM schematic of the optics involved in full field nanotomography. The x-rays are focused by the capillary condenser and unfocused rays are obstructed by the pin hole and beam stop. The x-rays are focused onto the sample and the transmitted rays are transformed by the objective zone plate and are finally detected by an optically-coupled CCD. The sample is able to rotate about an axis in cases where tomography is desired. [24]

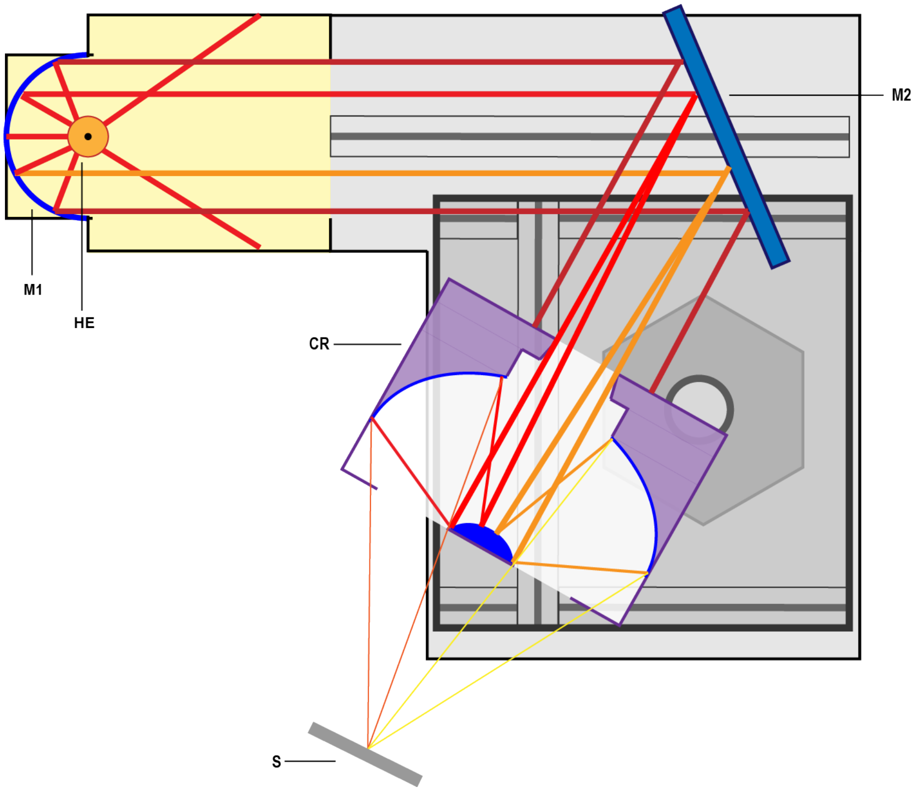


Figure 2: Schematic depicting possible IR Heat Lamp Apparatus. The Heat Emitter (HE) is backed by a primary reflecting mirror (M1) which directs the radiation towards a second reflecting mirror (M2) which is also acts as a directing mirror. The radiation is then reflected into a Cassegrain Reflector (CR) where it is focused onto the sample (S) face. Both M2 and CR are placed on mechanical tracks such that they are able to move as adjustments are needed.

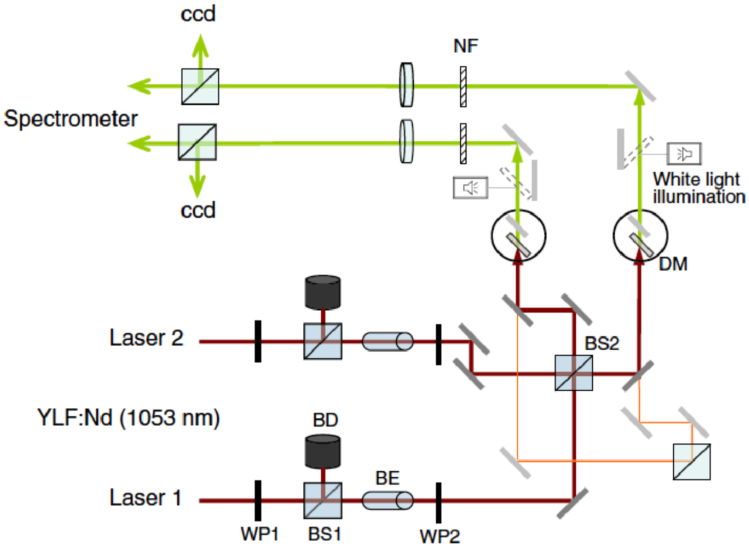


Figure 3: Schematic of the optics involved in manipulating and focusing the laser beams. In this diagram WP represent wave plate, BS represent beam splitter, BD represents beam dump, BE represents beam expander, DM represents dichroic mirror, ccd represents a CCD camera. WP1, BS1, and BD regulate the power of the laser beam. The beam is then combines and redistributed by WP2 and BS2. [8].

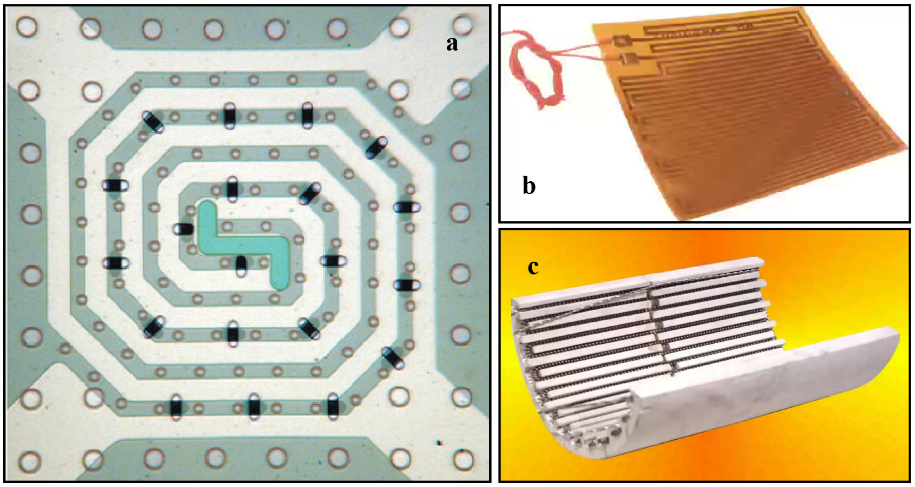


Figure 4: Depicted are various resistive heaters. 4a A silicon-based MEMS with spiraled Pt electrode pattern. [17] 4b A Kapton-based flexible heater with metal-foil meandering electrode pattern and two leads. [23]. 4c A ceramic-based heater where the ceramic half-cylinder is embedded with metal wire, which is partially obscured with additional ceramic bars to keep wire from contacting sample. [20].

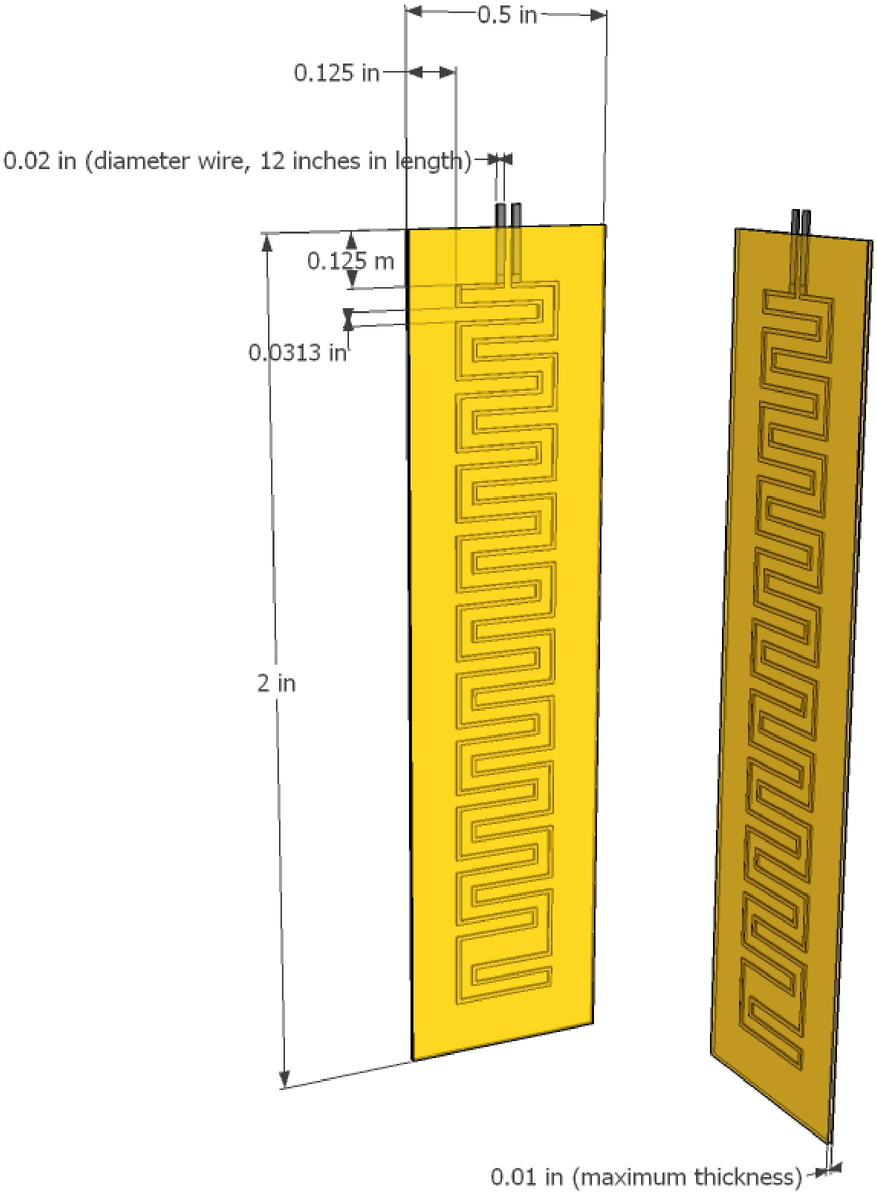

0.5 in
0.125 in
0.02 in (diameter wire, 12 inches in length)
0.125 m
0.0313 in
2 in
0.01 in (maximum thickness)

Figure 5: SketchUp image of the proposed Kapton Flexible Heater with two 12” FEP insulated 24 gauge wire leads. OMEGA Flexible Heater, 0.5” x 2” ; spacing between etched foil electrodes 1/32”; from heater edge to electrode edge 1/8”; maximum heater thickness 0.01” [23].

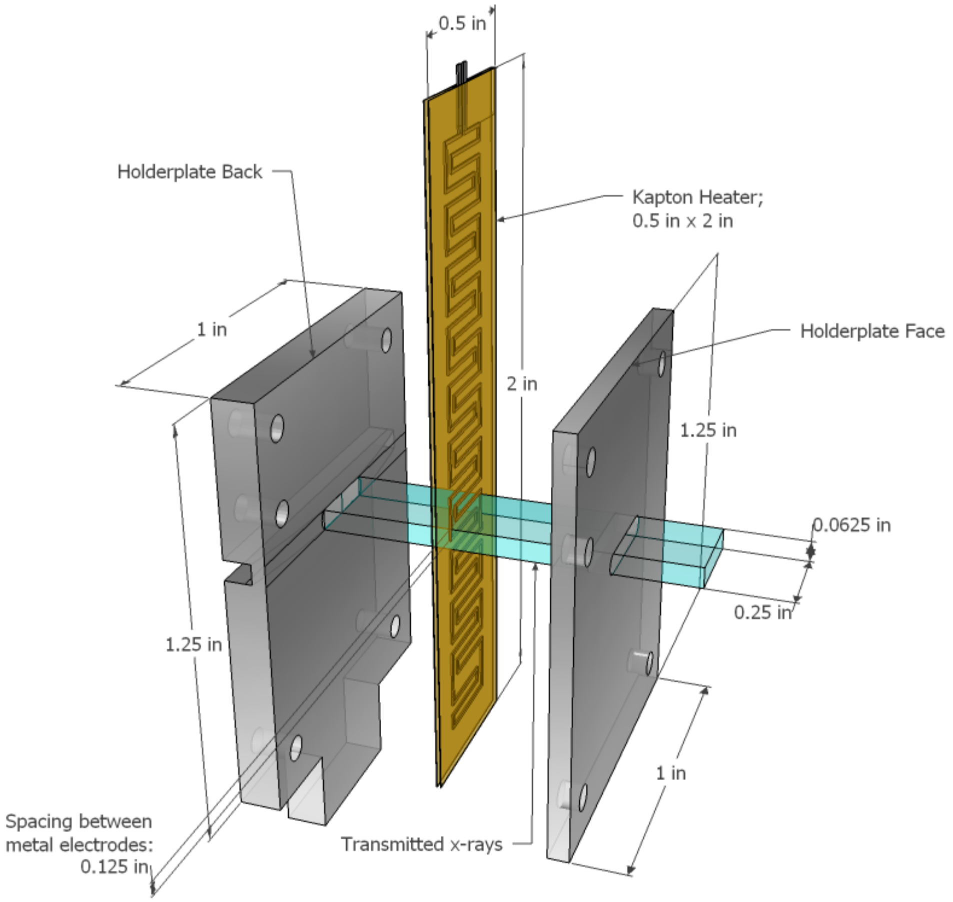


Figure 6: SketchUp image of the expanded sample holder. The proposed Kapton heater from Figure 5 is positioned between the front and back sample holderplates. Holderplates are made of aluminum and were designed by Johanna Nelson. A projection of the spatial range available for x-rays to be transmitted shown in blue; note that the actual beam is on the order of 15-30 μm. The openings in the holderplates are 1/4” x 1/16”.

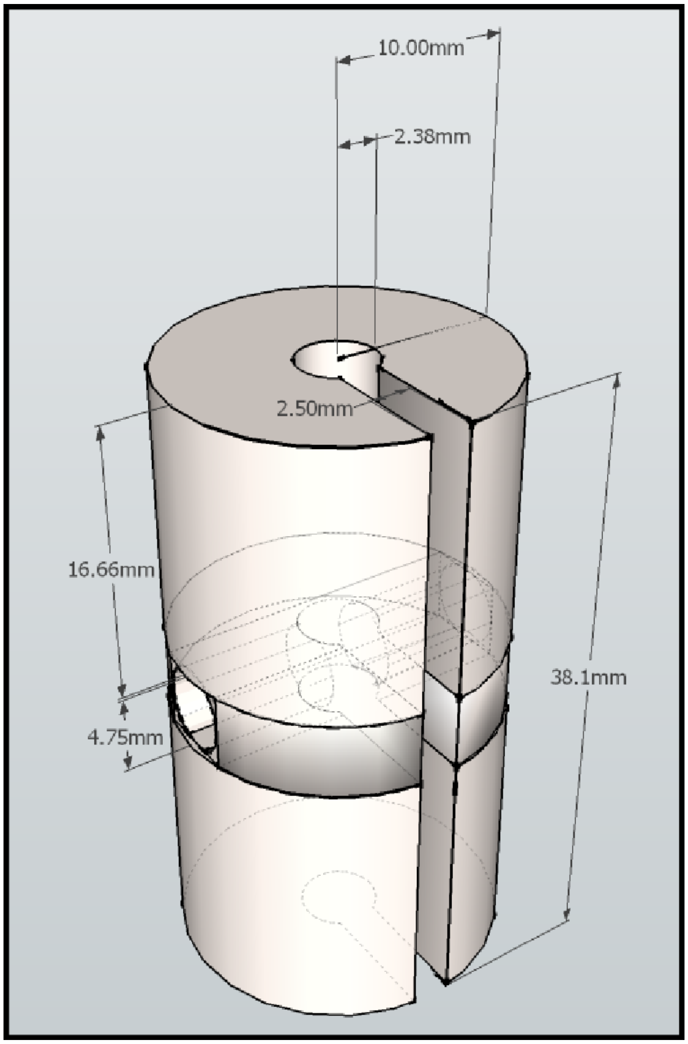


Figure 7: SketchUp image of a 3D model of a ceramic-based resistive heater that would be fabricated in-house using machinable Macor ceramic. The ideal design for such a ceramic heater would have the wire partially embedded within the ceramic. This design would feature a solid ceramic body with a vertical slit, which enables the user to slide the heater around a capillary. The ceramic cylinder would feature a circular, horizontal gap to allow x-ray transmittance through the sample. This design could be mounted externally to the table adjacent to the sample stage, allowing the sample to rotate for tomography while keeping the heater still. [21]

TABLES

Material	Air	Wood	Kapton	Water	Glass	Macor Ceramic	Aluminum	Copper
k [W/mK]	0.025	0.04	0.37	0.6	1.1	1.46	237	401

Table 1: Materials and their thermal conductivities [W/(mK)]. Kapton polyimide films thermal conductivity than glass and ceramic but is larger than air and wood; its thermal conductivity is one thousandth of copper.

Company	Operating Temp. Extrema of the Kapton Heating (°C)	Heater Thickness (nm)	Spacing Between Electrodes (mm)	Minimum Dimensions (mm x mm	Cost per Heating Unit[†]
Durex Industries	Low -195, **High: 200**	0.1524-0.3048; **Standard: 0.178**		25.4 x 25.4 /	
OMEGA	-200, **200**	**0.203**	0.79	6.35 x 6.35 /	$31
Birk Design and Manufacturing	N/A, **120/260**[‡]	0.07-0.254		12.7 x 12.7 /	

Table 2: Companies which produce customizable Kapton-based resistive heaters and their respective characteristics and dimensions. [†]5 W/in^2, 0.5" x 2"; [‡]Acrylic bonded Kapton has max. capability of 120

	Infrared Heat Lamp	**Infrared Laser**	**Silson Silicon-based MEMS**	**OMEGA Kapton®-based Resistive**	**Mellen Cylindrical Ceramic-based Furnace**
(a) Controlled	Power Supply	Power Supply	Power Supply	Power Supply	Power Supply
(b) Confined	Must use optics to confine heat small spot size	Use optics to manipulate laser beam; already small spot size	Heater region is 0.5 mm x 0.5 mm; high drop off from electrode pattern to outer silicon	Heat is confined to electrode region; heating area begins 3.17 mm (1/8") from outer edge of heater	Confined within the cylinder; holes necessary for x-ray transmission through sample may be concern for stray heat
(c) Measurable	Can calculate temperature incident on sample face and use a thermocouple on IR source	Calculate temperature at the sample interface using the Planck radiation function (1); use an imaging spectrograph	Can calculate the local resistance; calibrate the heater; use a thermocouple on the heater face	Can calculate the local resistance; calibrate the heater; use a thermocouple on the heater face	Can calibrate; use a thermocouple
(d) Temperature Range (50-500 °C)	Some IR heaters can heat samples to 700 °	Can heat above 1200 °C	Below 400 °C; currently use a Au electrode, in the process of R&D for a Pt electrode which would operate at higher temperatures	Below 200 °C	Below 1100 °
(e) Removable	Would be installed off-stage or on the side of the TXM	Installed off TXM; would be difficult to remove and install. If possible, install above TXM (out of the way).	Placed with sample on electrode side; removable	Placed on sample, side doesn't matter; removeable	Placed around sample; depending upon the model would be clamped around or slid over the sample
(f) Size Range	Size of IR heaters varies greatly; estimate size would be a 10" cube	Requires a rather large optical stage as well as several satellite data-gathering components; very large and bulky	Minimum dimensions are 10 mm x 5 mm x 0.5 mm (height x width x thickness)	Minimum size 12.7 mm x 12.7 mm (0.5" x 0.5")	Minimum diameter 19.05 mm (0.75"), any length
(g) Possible Sample Types	All; difficulty for those to be rotated for tomography	All; difficulty for those to be rotated for tomography	Flat samples only	Flat or gently curved samples only	Cylindrical samples only; those in capillary tubes
(h) Cost	> $239.00 [a]	> $15,000 [b]	$81.40 [c]	$31.00 [d]	N/A
(i) Level of In House Manufacturing Difficulty	High; optics and fabrication of Cassegrain reflector	High; optics and mounting large components	High; nanofabrication required	Medium; require nanofabrication but less difficult than silicon-based	Medium; require machining Macor ceramic and threading wire through

Table 3: Summary of each heating elements applicability to the determined criteria for our desired heater. The criteria are labeled a through i. [a]IR Emitter, Model 6363 [25]; [b]IR Laser (1054 nm), model CL1030-200 [10]; [c]1" x 1" Micro-hotplate, Au electrode pattern [18]; [d]0.5" x 2" KHLV-0502, 5W/in^2 [23]

	a	b	c	d	e	f	g	h	i	
Weight	1	2	2	3	1	3	1	3	3	
Rank (1-best, 5-worst)										**Weighted Rank**
Heat Lamp	1	5	3	5	4	4	3	3	4	72
Laser	1	1	1	1	5	5	1	5	5	59
Silicon	1	2	4	2	1	1	5	2	3	43
Kapton	1	3	5	2	1	2	2	1	2	**41**
Ceramic	1	4	1	4	3	3	5		1	**43**

Table 4: Each of the criteria a through i were weighted, as shown. Then the heating methods were ranked for each of the criteria, 1 being the best and 5 the worst. The total indicates the sum of the heaters rank for a respective criteria multiplied by that criterias weight. The best scores are the lowest.

SLAC-TN-12-015

Extracting $b\bar{b}$ Higgs decay signals using multivariate techniques

W Clarke Smith

Office of Science, Science Undergraduate Laboratory Internship (SULI)

George Washington University

SLAC National Accelerator Laboratory

Menlo Park, CA

August 17, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Timothy Barklow with the ATLAS group at SLAC National Accelerator Laboratory.

Participant: ______________________

Signature

Research Advisor: ______________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Extracting $b\bar{b}$ Higgs decay signals using multivariate techniques. W CLARKE SMITH (George Washington University, Washington, DC 20052) TIMOTHY BARKLOW (ATLAS group at SLAC National Accelerator Laboratory, Menlo Park, CA 94025)

For low-mass Higgs boson production at ATLAS at $\sqrt{s} = 7$ TeV, the hard subprocess $gg \to h^0 \to b\bar{b}$ dominates but is in turn drowned out by background. We seek to exploit the intrinsic few-MeV mass width of the Higgs boson to observe it above the background in $b\bar{b}$-dijet mass plots. The mass resolution of existing mass-reconstruction algorithms is insufficient for this purpose due to jet combinatorics, that is, the algorithms cannot identify every jet that results from $b\bar{b}$ Higgs decay. We combine these algorithms using the neural net (NN) and boosted regression tree (BDT) multivariate methods in attempt to improve the mass resolution. Events involving $gg \to h^0 \to b\bar{b}$ are generated using Monte Carlo methods with Pythia and then the Toolkit for Multivariate Analysis (TMVA) is used to train and test NNs and BDTs. For a 120 GeV Standard Model Higgs boson, the m_{h^0}-reconstruction width is reduced from 8.6 to 6.5 GeV. Most importantly, however, the methods used here allow for more advanced m_{h^0}-reconstructions to be created in the future using multivariate methods.

INTRODUCTION

When formulating the Standard Model of particle physics (SM), theorists initially had trouble explaining the origin of mass for bosons and fermions. In general, they reasoned that a symmetric system could not evolve into an asymmetric state. As a consequence, though, neither $W^{\pm}$ nor Z^0 bosons would be massive. Since $W^{\pm}$ and Z^0 are massive, they hypothesized the existence of a so-called *Higgs mechanism* that allows for spontaneous symmetry breaking, thereby giving such force-carrying bosons mass.

A Toroidal LHC ApparatuS (ATLAS), one of the six experiments at the Large Hadron Collider (LHC), is currently attempting to confirm the empirical existence of the Higgs mechanism by observing its scalar remnant: h^0, the *Higgs boson*[1]. Teams within ATLAS are attempting to do this by detecting the particles that characteristically result when these bosons decay. Higgs bosons can be created in proton-proton collisions, or 'events,' at ~ 7 TeV in the center-of-mass frame. The most promising Higgs decay modes to detect are $WW^{(*)}$[2], $ZZ^{(*)}$, $\tau^+\tau^-$, and $\gamma\gamma$ [1]. However, for the low Higgs masses ($115 \lessapprox m_h \lessapprox 140$ GeV) that this project is interested in, the Higgs boson decays into $b\bar{b}$ approximately 66% of the time[3] via the hard subprocess $gg \to h \to b\bar{b}$. We call *signal events* those that include this process (Figure 1) and *background events* those that include a different $b\bar{b}$-producing hard subprocess, such as $gg \to b\bar{b}$ (Figure 2). Unfortunately, almost every $b\bar{b}$-dijet detected results from a background event.

In order to successfully detect the Higgs boson, we seek to exploit its intrinsic narrow mass width of several MeV [2]. Theoretically, this will manifest itself as a sharp, narrow spike at m_h in the $b\bar{b}$-dijet mass ($m_{b\bar{b}}$) plot, allowing for the Higgs boson to be easily observed. However, reconstructing $m_{b\bar{b}}$ from event information proves difficult due to the so-called 'jet

[1] There may be multiple Higgs bosons; we reduce to the singular case. Also, hereafter we denote the Higgs boson h instead of the more accurate h^0 (the symbol for a scalar, light Higgs boson).

[2] The superscript $(*)$ denotes a virtual particle.

[3] Calculated for a 120 GeV SM Higgs boson using Pythia Monte Carlo generated events.

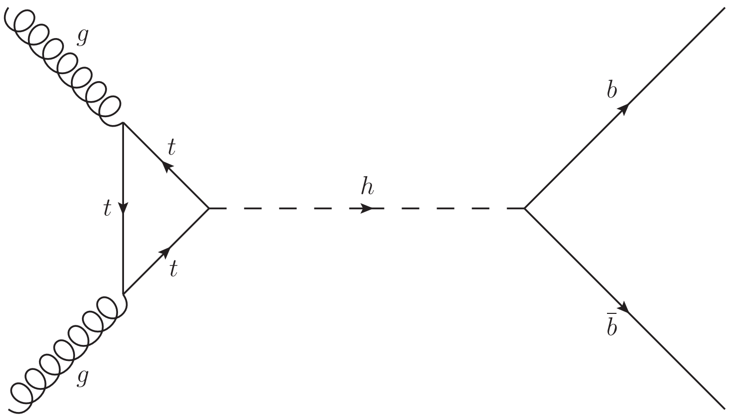


Figure 1: Feynman diagram of $gg \to h \to b\bar{b}$ (Higgs boson production from gluon fusion via a virtual t-quark loop followed by $b\bar{b}$ Higgs decay), the hard subprocess in signal events.

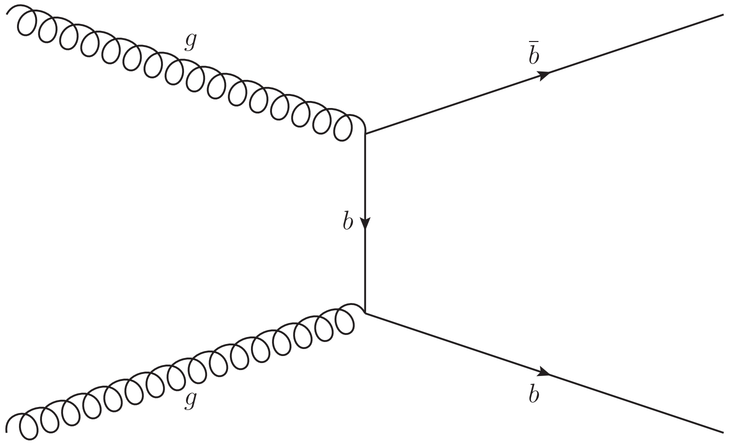


Figure 2: Feynman diagram of $gg \to b\bar{b}$ (production of $b\bar{b}$ from t-channel gluon fusion via the exchange of a virtual b-quark), an example of a hard subprocess in background events.

combinatorics problem,' that is, our inability to identify jets accurately. Although the $b\bar{b}$-dijets can be efficiently distinguished (b-tagged) [3], the b- and $\bar{b}$-quarks constantly radiate and reabsorb gluons before hadronizing and some gluons will be radiated but not reabsorbed. Those gluons are called *final state radiation* (FSR) and they are precisely those whose jets are difficult to identify. To solve this problem, there have been many *mass-reconstruction* *algorithms* developed to isolate such FSR-jets, but their inability leads to a reconstructed $m_{b\bar{b}}$ plot with a dull, essentially nonexistent bump at m_h. Realizing that the plot of background events vs. $b\bar{b}$-dijet masses experiences gradual decay over tens of GeV and so should the analogous reconstructed background $m_{b\bar{b}}$ plots, we restrict ourselves to the case of signal events, where $m_{b\bar{b}} = m_h$. In other words, reconstructing $m_{b\bar{b}}$ from signal events alone is sufficient to extract the m_h spike from the background $m_{b\bar{b}}$ plot. Hence, the goal of this project is to specifically construct an improved m_h-reconstruction algorithm[4].

Motivated by the quantity and unknown degree of interdependency of the observables measured by ATLAS, we use several types of *multivariate analysis*[5] (MVA) to search for patterns in Monte Carlo (MC) generated *pp* collisions in attempt to reconstruct $m_{b\bar{b}}$ more effectively. One MVA method that we use is the *neural network* (NN), a composition of many linear and nonlinear functions which maps sets of input values into models of the target values [4]. For our purposes, the target is $m_{h,true}$.

This approach is not altogether new; others have used various MVA techniques to enhance Higgs decay signals in many of the modes. For example, Hakl et al. [5] used NNs[6] and Takahashi [1] used another type of MVA, boosted decision trees, to improve the signal-to-background ratio for $b\bar{b}$ and $WW^{(*)}$ Higgs decays, respectively. In fact, boosted decision trees have shown impressive results in comparison to NNs — in a *b*-tagging performance evaluation, boosted decision trees demonstrated an efficiency improvement of 35% [6]. Instead of using a classification NN or boosted decision trees to distinguish signal events from background events like Hakl et al. and Takahashi, we use a variety of regression NNs and boosted regression trees to reconstruct m_h.

[4] Note that "mass-reconstruction algorithm" is taken to mean "$m_{b\bar{b}}$-reconstruction algorithm."

[5] Multivariate analysis is merely drawing inferences from data sets with multiple statistical variables.

[6] Hakl et al. used NNs with switching units optimized via a genetic algorithm.

MATERIALS AND METHODS

All regression MVA methods search for patterns in statistical variables in an attempt to better model the target, in this case, the true Higgs mass. If the method is not supplied any information aside from the variables, it will be entirely unable to model the target and it will not be able to produce an output; it needs to be "taught." We systematically feed the method variables as well as targets. It then develops a mechanism for predicting the target (computing outputs). This is called *training*.

Event generation with Pythia

For the data from which to select the variables and targets, we generate signal pp collisions using MC methods with Pythia. The Pythia program is a powerful particle physics event generator that includes a collection of methods for modeling the evolution of few-body hard processes to complex multihadronic final states. In Pythia event generation, the user is allowed to select the physical models used, the hard subprocesses involved, and, if applicable, the Higgs mass [7]. We use SM physics, the process $gg \to h \to b\bar{b}$, and m_h = 90, 100, 110, 120, 130, 140, and 150 GeV to generate 10^5 (signal) events at each m_h. Pythia's main weakness is that it does not include next-to-leading-order perturbative calculations and thus is insensitive to events with a low total *transverse momenta* p_T, that is, the momentum perpendicular to the beam axis. The high total p_T of $gg \to h \to b\bar{b}$ makes Pythia ideal for our low mass Higgs decays [7]. We do not use data taken by ATLAS. The reasons for this are twofold: we avoid complications caused by experimental uncertainties and we have absolute control over the type of event that is generated. For instance, we can choose whether to generate a $b\bar{b}$ or $ZZ^{(*)}$ Higgs decay, an event with m_h = 90 GeV or m_h = 150 GeV, or an event with one FSR-jet or five. Moreover, the only detector simulation is an angular cut is applied to remove all hadrons that escape the detector geometry; even neutrinos are generated.

Event processing with ROOT

For each hadron detected by ATLAS, the azimuthal angle ϕ, the *pseudorapidity* $\eta = -\ln\tan\left(\frac{\theta}{2}\right)$, and p_T are measured, providing us with its geometric and energetic image[7]. Additionally, we form the observable $\sum_i p_{T,i}$ for particles i in a given event to gauge the total transverse momentum.

Using a given mass-reconstruction algorithm, first the detected hadrons are partitioned into their respective jets. This partitioning is somewhat trivial — there is not much variation between algorithms. For each jet, η, ϕ, and p_T are computed from the η_i, ϕ_i, and $p_{T,i}$ measured for each hadron i within that jet. Second, the $b\bar{b}$-dijets and FSR-jets all resulting from $b\bar{b}$ production are isolated. This is done with cuts in the $\eta\phi$ plane, where the metric $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ can be constructed, providing a notion of distance between components of the event. Each mass-reconstruction algorithm uses a value of ΔR to classify all jets within a certain vicinity of the tagged $b\bar{b}$-dijets as FSR-jets from the b- and $\bar{b}$-quarks. From this jet selection, $m_{b\bar{b}}$, $p_{T,b\bar{b}}$, and $\eta_{b\bar{b}}$ are calculated. Finally, the algorithms compute $\Delta R_{b\bar{b}}$, the "distance" between b- and $\bar{b}$-jets, from the algorithmically determined $b\bar{b}$-dijet information.

It then seems reasonable that some algorithms would be better for certain types of events than others, depending primarily on their method of selecting those jets that result from $b\bar{b}$ production. For example, some mass-reconstruction algorithms may be very good at determining $m_{b\bar{b}}$ in events with a large fraction of jets being emitted due to *initial state radiation*[8] (ISR), while others may be better suited for events with little ISR activity.

We use ROOT, the leading data analysis framework for particle physics, to process the raw output of the Pythia signal event generation. For every Pythia MC generated event,

[7] Here, θ is the angle with respect to the beam axis.

[8] Initial state radiation is those gluons that radiate but do not get reabsorbed prior to the pp collision.

m_h, $p_{T,h}$, and η_h are reconstructed using 25 of the most effective mass-reconstruction algorithms. Furthermore, these 25 algorithms are divided into five families. The algorithm families have slightly different methods for partitioning hadrons into jets while algorithms within a family have distinct ΔR values for identifying those FSR-jets that result from $b\bar{b}$ production. So, for the same events, η_b, $\eta_{\bar{b}}$, $p_{T,b}$, $p_{T,\bar{b}}$, and $\Delta R_{b\bar{b}}$ are reconstructed using the five algorithm families. Note that no ϕ information is reconstructed; this is due to the fact that the pp collisions are azimuthally symmetric. Together, these collections of outputted event information are the input values for the MVA methods. We call these input types *variables*; they are listed in Table 1. Obviously, not all 101 variables will be used for every MVA technique.

Input Variables	*Quantity per Event*	*Description*
$\sum_i p_{T,i}$	1	Total event transverse momentum
η_b, $\eta_{\bar{b}}$	5 each	Reconstructed b, $\bar{b}$-jet pseudorapidities
$p_{T,b}$, $p_{T,\bar{b}}$	5 each	Reconstructed b, $\bar{b}$-jet transverse momenta
η_h	25	Reconstructed Higgs pseudorapidity
$p_{T,h}$	25	Reconstructed Higgs transverse momentum
$m_{h,reconstruction}$	25	Reconstructed Higgs mass
$\Delta R_{b\bar{b}}$	5	Reconstructed "distance" between b, $\bar{b}$-jets

Table 1: A comprehensive list of MVA input variables.

Finally, in ROOT we declare one further characteristic of the generated events: the true Higgs mass as specified in Pythia. This $m_{h,true}$ is not a variable but the target of the MVA regression; it is only used in training and for performance evaluation.

MVA training and testing

At this point, we have $\sim 7 \times 10^5$ Pythia generated signal events each with one of seven Higgs masses as well as values for 101 different variables. The goal is then to construct mappings from the 101 variables onto the set of seven Higgs masses by invoking multiple MVA methods, which have been shown to be superior to uni- and bivariate techniques for most purposes [1]. While each mass-reconstruction algorithm is insufficient on its own, feeding the outputs of 25 different algorithms into an advanced MVA technique vastly improves the m_h regression. Due to the ease of implementation and the customizability of its MVA methods, we use the Toolkit for Multivariate Analysis (TMVA) to construct such mappings. TMVA is included in the ROOT package and offers thirteen MVA techniques for *classification* (directly separating signal from background events), eight of which can also be used for *regression* (creating a regression function that models the target). We restrict ourselves to the two most promising of the eight methods suitable for our regression task: neural nets and boosted regression trees.

Previous NN analyses of this kind used the ROOT class `TMultiLayerPerceptron` which allows for any number of intermediate functions as well as one of six different learning methods to be used [2]. The advantages of using TMVA are the facility of using boosted regression trees, an increased number of adjustable method-specific parameters, and most importantly, the ability to repeat analyses quickly using different parameters, variables, MVA methods, and data sets [8].

In a feedforward NN[9], there is an input layer, a hidden layer(s), and an output layer (Figure 3). The nodes are called *inputs* in the input layer, *neurons* in the hidden layers, and *outputs* in the output layer. Each variable corresponds to an input and the targets correspond to the outputs. Each neuron in the first hidden layer maps a linear combination of the inputs through a specified nonlinear function σ. Neurons in subsequent hidden layers repeat this process using neurons in the previous hidden layer. Additionally, there can be "empty" neurons and inputs called *biases* which are merely weights independent of prior layers added to increase the ability of the NN.

[9] Other types of NNs (those with feedback loops, etc.) are neglected because they are overkill for our relatively standard regression task.

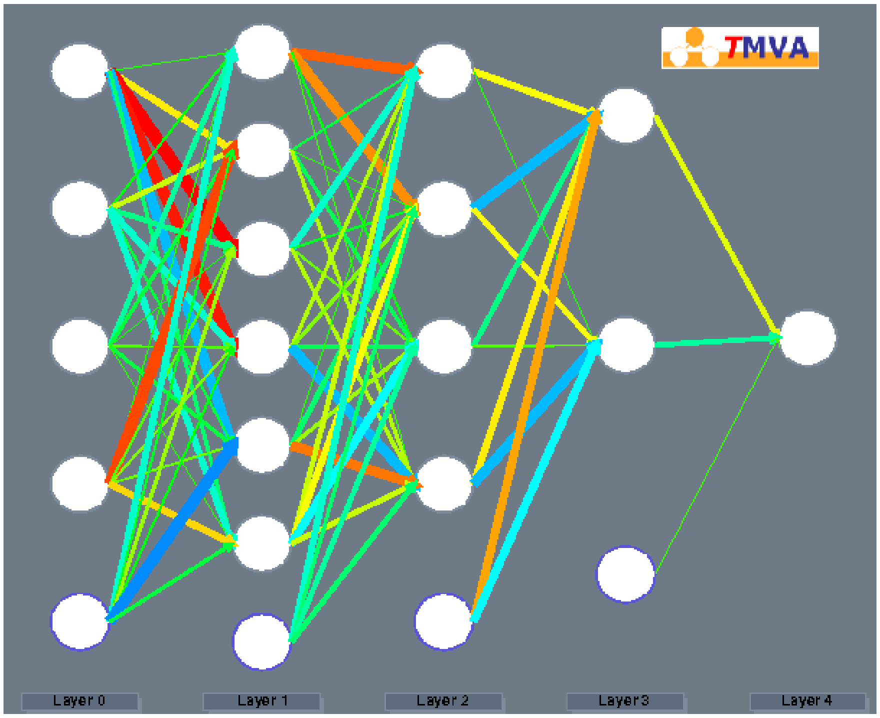


Figure 3: Feedforward NN with three hidden layers with 6, 4, and 2 neurons, respectively, as well as biases.

It may seem that having more neurons and hidden layers allows for a more complicated and therefore capable NN to be formed. This is true up to a point — while additional neurons and hidden layers enable the neural net to find more subtle relationships, it eventually reaches a peak beyond which the neural net is being *overtrained*. In that region, neural nets find patterns that are not necessarily representative of legitimate relationships, but are merely due to the finite size of the event sample. Moreover, there is not good reason to use excess hidden layers or neurons, for the Weierstrass approximation theorem dictates that any continuous function can be approximated by a sufficient number of neurons in one hidden layer [9].

MVA for particle physics has been dominated by the use of NNs over the past decade. In past several years, though, boosted decision and regression trees have emerged as a powerful tool for MVA. First of all, a *regression tree* is an MVA method grown by splitting the training sample of events into two parts for each value of each variable, depending on that value. Then, the variable and value that produces the optimal split ("best" division of the training sample into high and low target value $m_{h,true}$ parts) is chosen as the first branching of the regression tree. This process is repeated until the user-specified maximum number of leaves is reached. *Boosted regression trees* (BDTs[10]), then, are merely an advanced form of regression trees. Boosting is achieved by allowing the tree to predict the target $m_{h,true}$ for each of the training events; those events which have their targets incorrectly predicted are given additional weight depending on how poor the prediction was. The regression tree is then regrown; this boosting process goes through a user-specified number of iterations, eventually arriving at a final boosted regression tree (Figure 4). Typically a larger number of maximum leaves and boosting iterations allows for a more robust network, but BDTs are very susceptible to overtraining so these parameters need to be chosen carefully. In attempt to minimize the effects of overtraining, once the BDT is grown and boosted, it is *pruned* by removing statistically insignificant leaves.

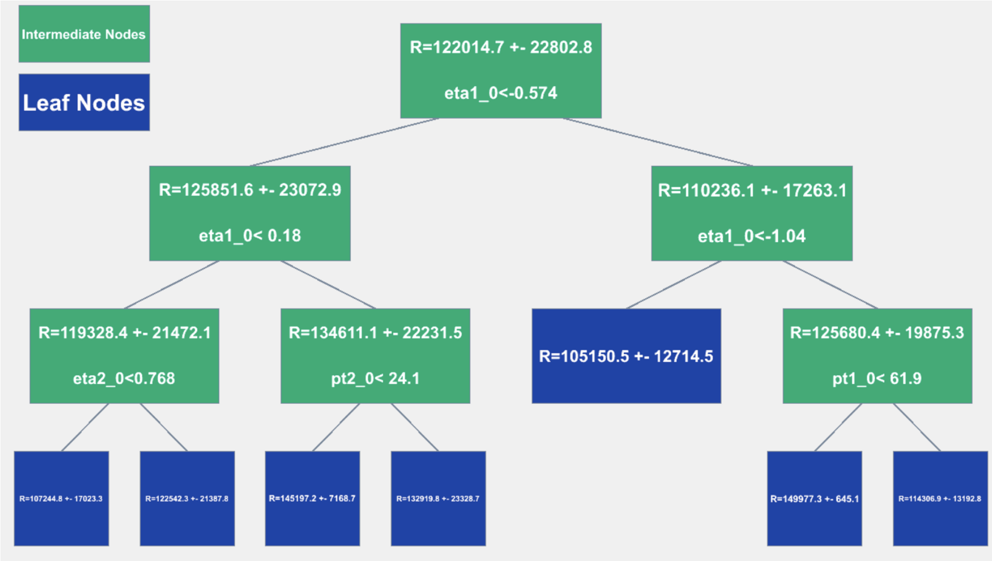


Figure 4: The 99th iteration of a boosted regression tree used for m_h regression.

[10] The acronym "BDT" is adopted for both boosted regression trees as well as boosted decision trees out of convention.

As mentioned before, event processing yields 101 variables with which to train the MVA methods. In order to increase efficiency and reduce computing time, redundant and ineffective variables are removed. Redundant variables are those which are very highly correlated with other variables used, meaning the underlying event information that they hold is very similar. Ineffective variables are those which do not show a significant correlation with the target. Clearly, redundant variables are useless while ineffective variables slow down and hinder the training process.

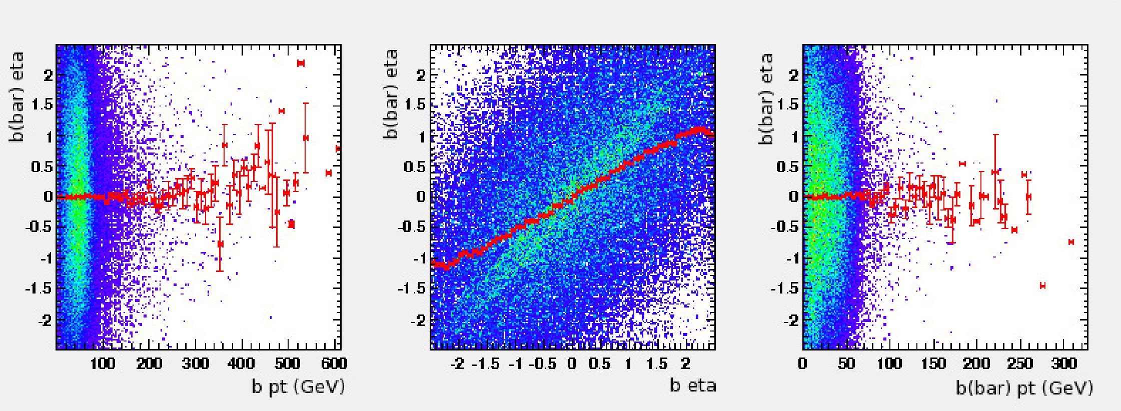
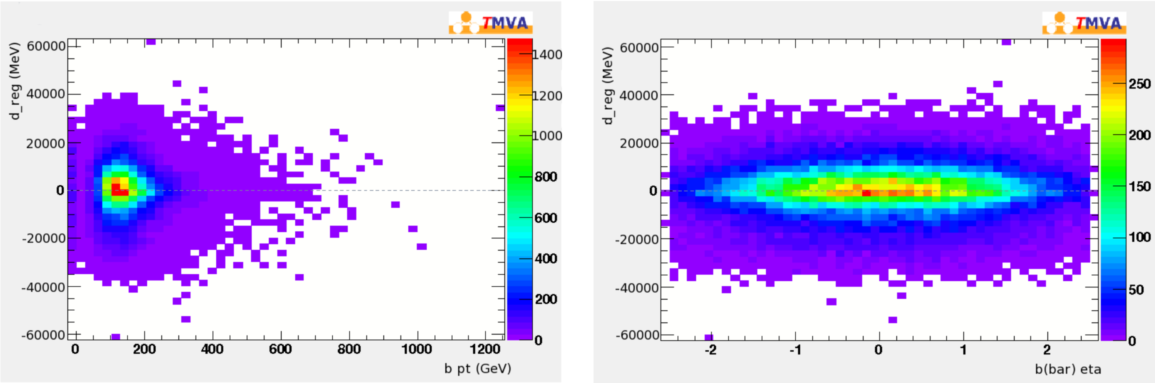


Figure 5: Plots of $\eta_{\bar{b}}$ vs. $p_{T,b}$, $\eta_{\bar{b}}$ vs. η_b, and $\eta_{\bar{b}}$ vs. $p_{T,\bar{b}}$ showing redundancy in the center plot.

A simple correlation-to-target ranking performed by TMVA allows us to reduce the number of variables from 101 to 40 relatively easily. The 25 reconstructed Higgs masses are the most highly correlated with $m_{h,true}$ and so they are not removed. Then, the correlations are similar from variable to variable, so $\sum_i p_{T,i}$ as well as two of each $\Delta R_{b\bar{b}}$, η_b, $\eta_{\bar{b}}$, $p_{T,b}$, $p_{T,\bar{b}}$, η_h, and $p_{T,h}$ are used for the initial training. Then, scatter plots of the variables vs. one another (Figure 5) as well as two-dimensional histograms of regression output deviation $d_{reg} := m_{h,reco} - m_{h,true}$ vs. the variables (Figure 6) are used to determine redundancy and effectiveness, respectively. Clearly the efficacy of a variable depends on the MVA method for which it is training. As such, we construct two sets of variables, a NN set and a BDT set, which are slightly different but contain ∼30 variables.



(a) The high efficacy of $p_{T,b}$.

(b) The low efficacy of $\eta_{\bar{b}}$.

Figure 6: Regression output deviation d_{reg} vs. variable plots used to determine variable contribution to the MVA method.

Once the variables have been selected to maximize efficiency, the various parameters of the NN and the BDT need to be optimized. There are three types of NN implemented in TMVA: the Clermont-Ferrand NN, the ROOT NN (`TMultiLayerPerceptron`), and the multilayer perceptron (MLP) NN. We use the MLP NN due to its improved speed and flexibility over the other two types of neural net [8]. Five of the most important parameters for the MLP are `NCycles`, `HiddenLayers`, `NeuronType`, `NeuronInputType`, and `TrainingMethod`. `NCycles` is merely the number of training cycles the NN undergoes and `HiddenLayers` specifies the number of hidden layers and how many neurons are in each of them. In general, we want to have as many cycles, hidden layers, and neurons as possible in order to have a robust and sensitive network while not overtraining it. The `NeuronType` is the NN's nonlinear function σ. The options include the sigmoid function $1/(1+e^{-rv})$, the hyperbolic tangent function $(e^v - e^{-v})/(e^v + e^{-v})$, and the radial function $e^{-v^2/2}$. We have found the best results using the hyperbolic tangent function[11]. The option `NeuronInputType` specifies how the linear combinations are taken within each neuron. They can be sums, sums of squares,

[11] "Best" is ambiguous; here, we refer to the MLP with the lowest d_{reg} root mean square (RMS).

or sums of absolutes; we use regular sums. Finally, the `TrainingMethod` can be either the back-propagation (BP) method or the Broyden-Fletcher-Goldfarb-Shannon (BFGS) method. BFGS requires less iterations to receive the same result as BP, but it costs a large amount of computing power when the NN is complicated [8]. As such, we use BP with complex hidden structures and BFGS with simple hidden structures.

For the BDT, some of the adjusted parameters are `NTrees`, `BoostType`, `SeparationType`, `nEventsMin`, and `PruneMethod`. `NTrees` is merely the number of times the regression tree is boosted and regrown while `nEventsMin` is the minimum number of events in any given leaf node. Increasing the former and decreasing the latter heightens the BDT's sensitivity and capability so long as overtraining is avoided. For regression, the `BoostType` options are narrowed to adaptive boost (AdaBoost.R2) and gradient boost (GradientBoost). The difference between the two algorithms is the function they use to determine how poor the BDT's estimation of the target value is. Because neither AdaBoost.R2 nor GradientBoost is particularly advantageous, we use two BDT implementations: one for each type of boosting. Another important parameter is `SeparationType`, which determines how the training sample is split at each node. There are five settings available, but we find the default Gini-Index to produce the best results. Finally, for regression tasks, the `PruneMethod` can either be cost complexity or nothing. We use the cost complexity method to recover from overtraining.

With selected variables and optimized parameters for MLP, BDT (AdaBoost.R2), and BDT (GradientBoost) MVA methods, we feed the $\sim 7 \times 10^5$ Pythia MC generated events into TMVA. Half of these events are used for training and half for testing.

RESULTS

For each MVA method, we build a complicated nonlinear function from the set of selected variables into the set of possible m_h values through training. Then, that function's performance is evaluated through *testing*, in which the remaining half of the events have their variables mapped to reconstructed Higgs masses ($m_{h,reco}$). From TMVA we ultimately receive $\sim 3 \times 10^5$ pairs of $m_{h,reco}$ and $m_{h,true}$ data. These results can be depicted in three ways.

First, the total MVA method RMS can be calculated by RMS $= \sqrt{\langle {d_{reg}}^2 \rangle}$. This RMS can then be compared for multiple methods to determine which method has the overall least difference between reconstructed and true Higgs mass. The RMS values for the MLP, the BDT with AdaBoost.R2, and the BDT with GradientBoost methods are presented in Table 2 along with some extra information. Note that half of the values in Table 2 are "truncated," meaning they were calculated from only those events whose d_{reg} values lie within two standard deviations of the mean, i.e., in the middle 95%. On the other hand, the truncated RMS roughly represents the difference between $m_{h,reco}$ peak values and $m_{h,true}$ values. Finally, while $\langle d_{reg} \rangle$ is sensitive to sign, comparing the truncated and non-truncated entries for a given method provides shape and location information about the d_{reg} plot. For instance, the magnitude and sign of the difference indicates the degree and sign of the skewness of d_{reg}. Additionally, the size of $\langle d_{reg} \rangle_T$ reflects how centered about $m_{h,true}$ the $m_{h,reco}$ plot is.

MVA Method	*RMS* (MeV)	$\langle d_{reg} \rangle$ (MeV)	RMS_T (MeV)	$\langle d_{reg} \rangle_T$ (MeV)
MLP	1.25×10^4	47	9.73×10^3	1.21×10^3
BDT with AdaBoost.R2	1.38×10^4	631	1.25×10^4	1.29×10^3
BDT with GradientBoost	1.24×10^4	-1.49×10^3	8.99×10^3	-281

Table 2: A performance summary for all MVA methods utilized.

Second, we can plot d_{reg} on its own and against $m_{h,true}$ (Figure 7). These plots tell us whether the MVA method is under- or over-estimating the Higgs mass and for which $m_{h,true}$. Third and finally, for any one of 90, 100, 110, 120, 130, 140, or 150 GeV, we can plot both the reconstructed Higgs mass (Figure 8).

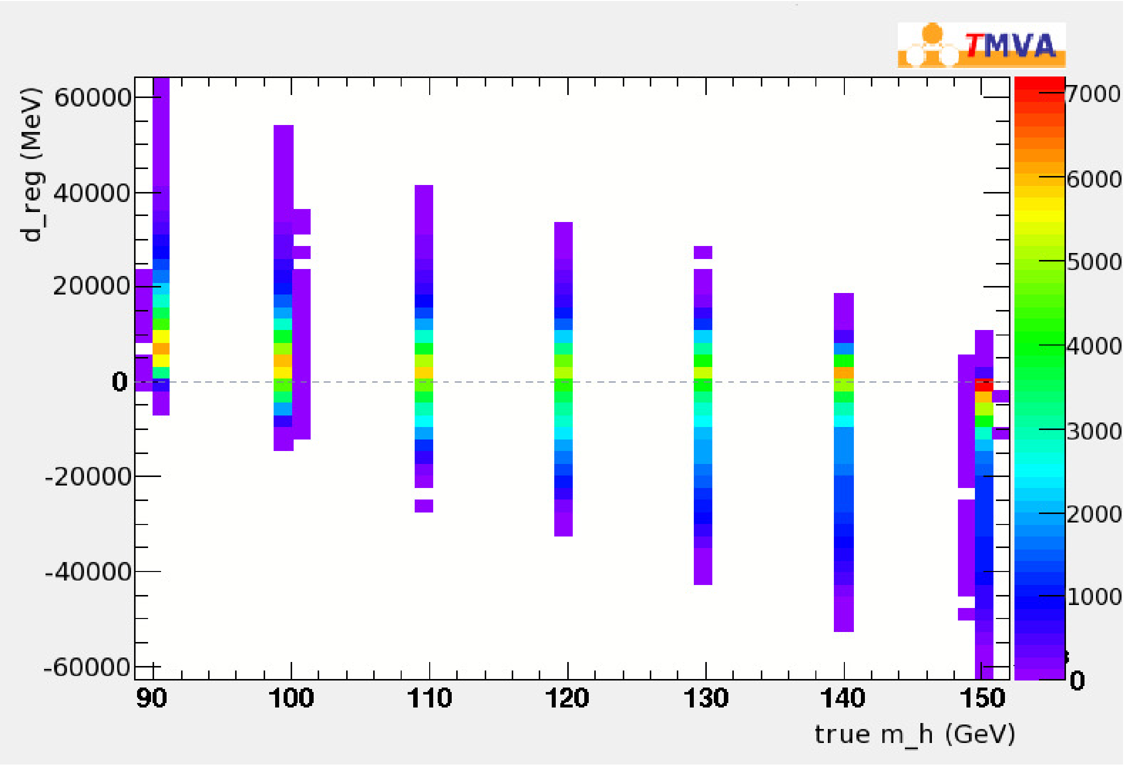


Figure 7: Plot of d_{reg} vs. $m_{h,true}$ for a BDT with GradientBoost.

DISCUSSION AND CONCLUSIONS

In order to visualize the actual performance of our trained MVA methods, we compare the reconstructed Higgs mass plot (Figure 8) to the PYTHIA MC generated $m_{h,true}$ plot, the $m_{h,reco}$ resulting from a single mass-reconstruction algorithm, and the current best NN-reconstructed Higgs mass plot all for events generated with an inputted $m_{h,true}$ of 120 GeV (Figure 9). We take this state-of-the-art in m_h reconstruction to be Barklow's [2], which uses the `TMultiLayerPerceptron` class in ROOT with multiple algorithmically reconstructed Higgs masses as inputs. This allows us to witness the development of m_h reconstruction methods from single algorithms to coarse, ROOT-based neural nets to powerful BDTs.

First, we notice that the PYTHIA generated $m_{h,true}$ plot has a width of 12.3 MeV, reflecting the narrow intrinsic Higgs mass width. Second, the m_h reconstruction using a single

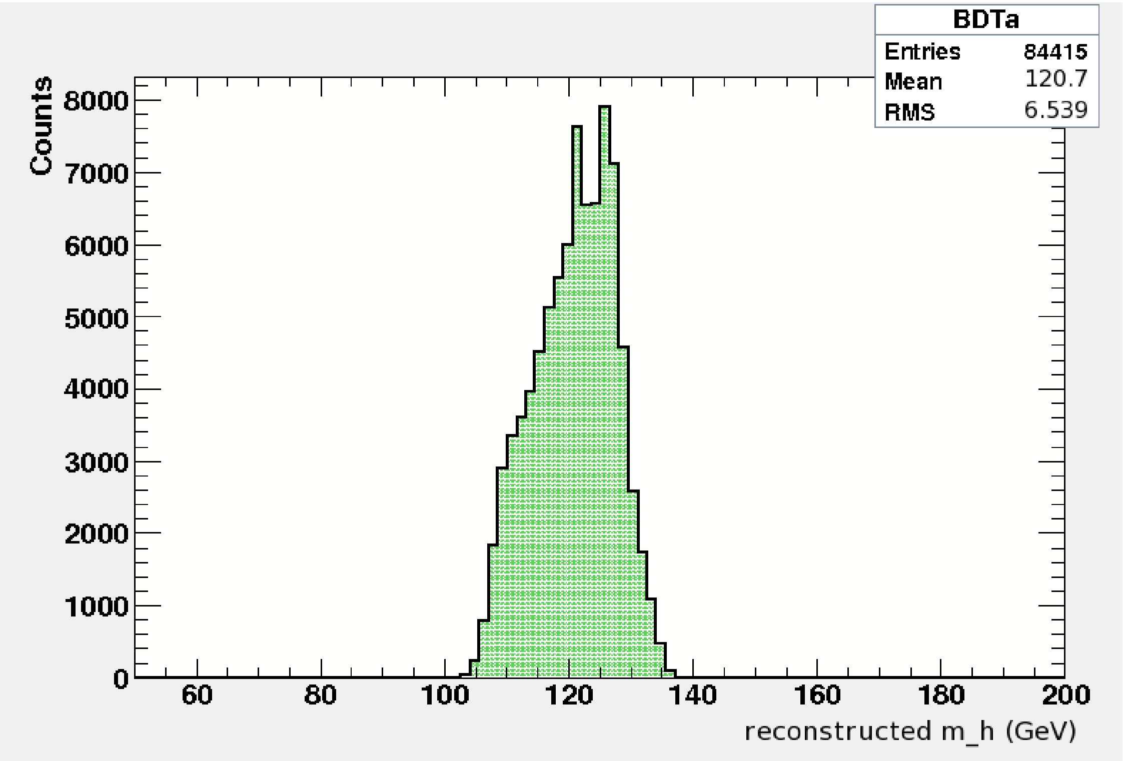
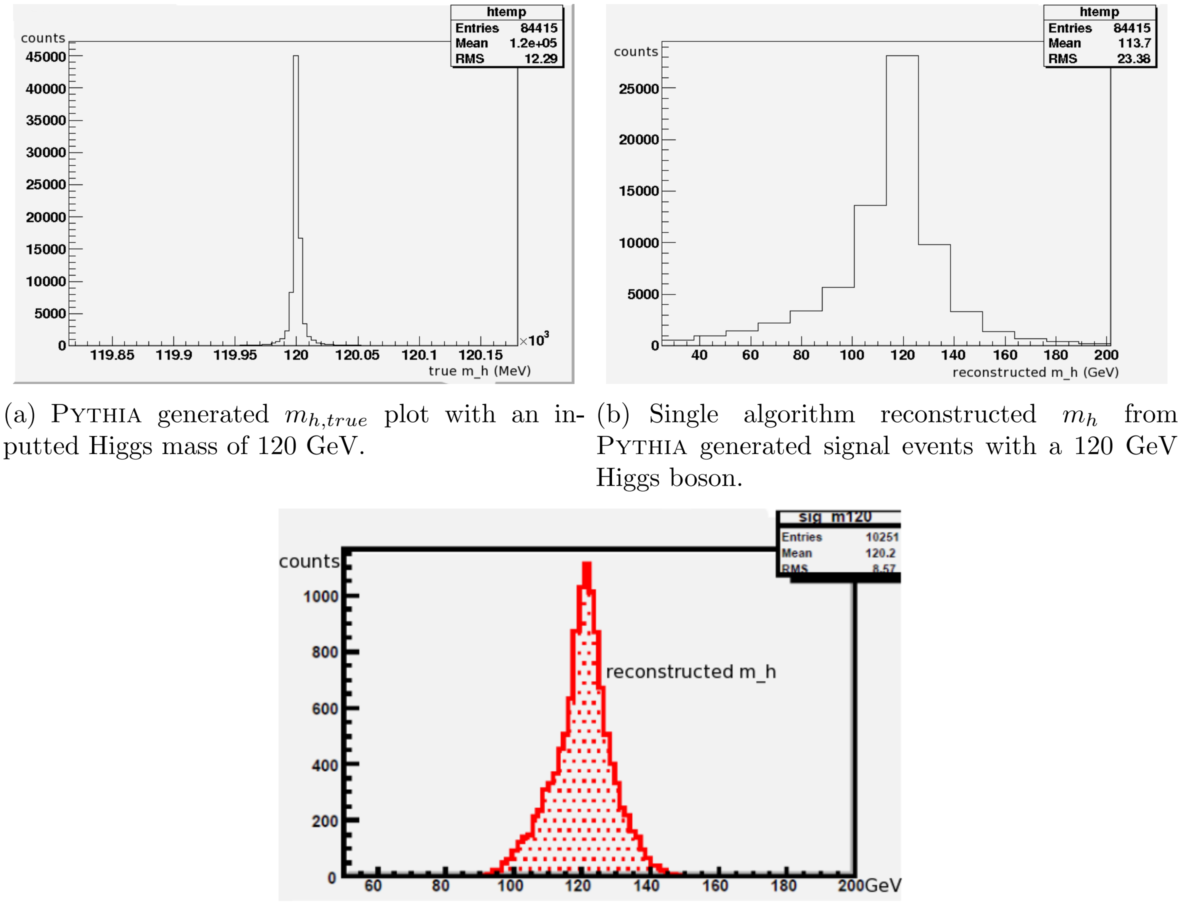


Figure 8: Plot of $m_{h,reco}$ for a BDT with AdaBoost.R2 using Pythia MC generated signal events with a 120 GeV SM Higgs boson.

algorithm shows a peak at $\sim$120 GeV, but a mean of 113.7 GeV and a width of 23.4 GeV. Third, the NN m_h reconstruction plot has its peak and mean at 120 GeV but a width of 8.6 GeV. Finally, our m_h reconstruction using a BDT with AdaBoost.R2 peaks around 120 GeV, has a mean of 120.7 GeV, and has a width of 6.5 GeV. We can therefore say that our MVA methods yield competitive m_h resolutions when compared to all previous approaches.

Although the $b\bar{b}$ Higgs decay signal was by no means extracted and observed, this reconstruction attempt was a success. Not only were gains made in Higgs mass resolution, but they were made as a proof-of-principle. Many elements of the development of the m_h-reconstruction method were crude and abbreviated — there is much room to improve our method. We have merely shown that it may be possible to construct a very effective m_h-reconstruction algorithm using MVA methods.



(a) Pythia generated $m_{h,true}$ plot with an inputted Higgs mass of 120 GeV.

(b) Single algorithm reconstructed m_h from Pythia generated signal events with a 120 GeV Higgs boson.

(c) Previous state-of-the-art NN-reconstructed m_h from signal events with a 120 GeV SM Higgs boson generated by ALPGEN (an MC particle physics generator best for low-p_T events).

Some of the improvements that can, and should, be made are in event generation, event processing, and the training and testing of MVA methods. Immediately, we should generate $gg \to h \to b\bar{b}$ events with more than seven distinct values for $m_{h,true}$ to reduce discretization in the final $m_{b\bar{b}}$ plots. As more results from the LHC are released, we could also generate events for additional Higgs decay modes as well as with beyond-the-SM physics. Another relatively easy improvement would be to optimize the MVA-method-specific parameters as well as select variables more gracefully. In both cases, the methods we use are crude and insensitive — algorithms may be used, for example, for both of these purposes. Additionally, Hoecker et al. [8] are continually working to improve TMVA and include more advanced MVA methods as well as combined MVA methods; these should be trained and tested as soon as they are released.

Finally, the largest potential improvement will be made from upgrading our event processing. Specifically, we will initially process events with a broader method of defining variables, allowing them to be selected from a pool of ~ 700. With this added information, the MVA methods may be able to find even subtler patterns in the events, increasing their ability to resolve the true Higgs mass. In conclusion, we are encouraged by our MLP and BDT results — they show that MVA methods can be very useful for m_h discrimination in ATLAS.

ACKNOWLEDGMENTS

This work was carried out at SLAC National Accelerator Laboratory from June–August 2011 as part of the Summer Undergraduate Laboratory Internship (SULI) program created, organized, and funded by the U.S. Department of Energy, Office of Science. I would like to thank my mentor Tim Barklow as well as Steve Rock, the director of the SULI program. Lastly, I would like to thank the ATLAS collaboration at the LHC for access to their resources.

REFERENCES

[1] M. Takahashi, "Standard Model Higgs Searches at the LHC," 2008, arXiv:hep-ex/0809.3224.

[2] T. Barklow, "Higgs → bb." `http://indico.cern.ch/getFile.py/access?contribId=40&sessionId=5&resId=0&materialId=slides&confId=116471`, Jan. 2011.

[3] B. Martin and G. Shaw, *Particle Physics.* Manchester Physics Series, Wiley, 2008.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics, New York: Springer, 2001.

[5] F. Hakl, M. Hlaváček, and R. Kalous, "Application of neural networks to Higgs boson search," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 502, no. 2-3, pp. 489 – 491, 2003.

[6] J. Bastos, "Tagging heavy flavours with boosted decision trees," 2007, arXiv:physics.data-an/0702041.

[7] T. Sjöstrand, S. Mrenna, and P. Skands, "Pythia 6.4 Physics and Manual," *JHEP*, vol. 05, p. 026, 2006.

[8] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," *PoS*, vol. ACAT, p. 040, 2007.

[9] M. H. Hassoun, *Fundamentals of Artificial Neural Networks.* The MIT Press, 1995.

SLAC-TN-12-016

X-Ray Emission Spectrometer Design with Single-Shot Pump-Probe and Resonant Excitation Capabilities

Katherine Spoth

Office of Science, Science Undergraduate Laboratory Internship (SULI)

State University of New York at Buffalo

SLAC National Accelerator Laboratory

Menlo Park, CA

August 12, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Dennis Nordlund at the Stanford Synchrotron Radiation Lightsource, SLAC.

Participant: ______________________

Signature

Research Advisor: ______________________

Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

X-Ray Emission Spectrometer Design with Single-Shot Pump-Probe and Resonant Excitation Capabilities. KATHERINE SPOTH (State University of New York at Buffalo, Buffalo, NY 14260) DENNIS NORDLUND (Stanford Synchrotron Radiation Lightsource, SLAC, Menlo Park, CA 94025)

Core-level spectroscopy in the soft X-ray regime is a powerful tool for the study of chemical bonding processes. The ultrafast, ultrabright X-ray pulses generated by the Linac Coherent Light Source (LCLS) allow these reactions to be studied in greater detail than ever before. In this study, we investigated a conceptual design of a spectrometer for the LCLS with imaging in the non-dispersive direction. This would allow single-shot collection of X-ray emission spectroscopy (XES) measurements with varying laser pump X-ray probe delay or a variation of incoming X-ray energy over the illuminated area of the sample. Ray-tracing simulations were used to demonstrate how the components of the spectrometer affect its performance, allowing a determination of the optimal final design. These simulations showed that the spectrometer's non-dispersive focusing is extremely sensitive to the size of the sample footprint; the spectrometer is not able to image a footprint width larger than one millimeter with the required resolution. This is compatible with a single shot scheme that maps out the laser pump X-ray probe delay in the non-dispersive direction as well as resonant XES applications at normal incidence. However, the current capabilities of the Soft X-Ray (SXR) beamline at the LCLS do not produce the required energy range in a small enough sample footprint, hindering the single shot resonant XES application at SXR for chemical dynamics studies at surfaces. If an upgraded or future beamline at LCLS is developed with lower monochromator energy dispersion the width can be made small enough at the required energy range to be imaged by this spectrometer design.

INTRODUCTION

Understanding the bonding processes that take place during the formation of chemical compounds can give us a greater ability to predict their chemical properties and reactivity. Some knowledge of these processes can be gained by studying changes in molecular orbitals. Many of these interesting reactions involve C, N, and O atoms. Core level soft X-ray spectroscopy techniques allow us to study these systems in an atom-specific way, meaning we can probe a specific energy transition by creating a core hole that can only be filled by a nearby valence electron [1]. The transfer of electrons during bonding processes occurs on a femtosecond time scale. Measurements on this scale are made possible using ultrafast X-ray pulses generated by the Linac Coherent Light Source (LCLS). A spectrometer that allows for Resonant Inelastic X-Ray Scattering (RIXS) or time-resolved X-Ray Emission Spectroscopy (XES) measurements to be taken in a single shot at the LCLS will allow us to study changes in the electronic structure of complex systems in greater detail than has ever been possible.

One of the many systems that would be interesting to study using this instrument is the dissociation of N_2 as occurs during the Haber-Bosch process through which ammonia is produced from nitrogen and hydrogen on a substrate of ruthenium or iron [2]. This process is widely used in the production of nitrogen-based fertilizers responsible for growing crops that feed much of the world's population and therefore is estimated to account for as much as one percent of the world's total energy consumption [3]. For this reason, it would be very beneficial if we could make improve this process to reduce its energy consumption.

The processes that take place during this and other surface reactions are difficult to observe experimentally. Between adsorption and desorption steps, charge and energy transfer takes place between the adsorbates and catalyst determining the intermediate steps in the reaction. These short-lived elementary steps can be observed using the time-resolved measurements made possible in ultrafast laser pump X-ray probe experiments using XES at the

LCLS.

Our goal is to determine the best design of a new XES spectrometer for the LCLS. We modify the design of a typical XES spectrometer to allow for single-shot measurements with a time delay or energy gradient along the sample footprint. Ray-tracing simulations of possible configurations are used to determine the best design, which has maximum throughput at the energy and imaging resolutions required to perform the desired experiments.

MATERIALS AND METHODS

Figure 1 shows an illustration of the X-ray emission process. In non-resonant X-ray emission, incoming X-ray photons excite an electron from the core of an atom in a sample, ejecting the electron into the vacuum. In the resonant case, this electron is not ejected but promoted to an unoccupied atomic or molecular orbital [4]. This is achieved by using a resonant incoming X-ray energy, which is equal to the difference between the core level energy and that of the excited state. In both cases, an electron from the valence level falls to fill the vacancy in the core, radiating an X-ray photon with energy equal to the difference between its initial and final states. The XES spectrum shows the intensity vs. energy of these emitted X-rays. From this spectrum, the energy and symmetry of the atom's orbitals can be determined, providing information about its valence band electronic structure.

Because each element has a unique core level energy, XES is element-specific. When bonds are formed with another element, the energies of an atom's orbitals transform. These changes in energy can be easily observed using XES, making this method chemical bonding-specific as well. Additionally, because the core holes created in this method can only be filled by valence electrons in the vicinity of the atom, XES is a useful tool in probing the local electronic structure of an atom.

In a typical XES spectrometer, which is shown in Figure 2, incoming X-rays excite core electrons in the sample causing outer electrons to decay and emit X-rays. The emitted rays are incident on a concave grating, which disperses the photons according to energy and focuses the light in this dispersive direction. At the detector, the focused rays' positions are related to their energy due to the their interaction with the diffraction grating.

Figure 3 illustrates how the sample footprint can be used to image laser-X-ray time delay. The sample footprint is defined as the area which is illuminated by incoming X-rays. In regular pump-probe experiments, the sample is excited by a laser pulse then probed using X-rays at a time delay. It is possible to create a varying time delay along the width x of this sample footprint by placing the X-ray probe at an angle to the laser pump such that the pump arrives first, followed by the probe at a time delay that increases across the sample footprint via the speed of light. If the laser pump incidence is normal to the sample, the time delay can be described by the equation TD= $\frac{x}{c}$. We can shorten this time delay by decreasing the incidence angle of the laser pump.

In a similar way we can also vary the energy of incoming X-rays along the sample footprint for RIXS measurements, shown in Figure 4. At the Soft X-Ray (SXR) beamline at LCLS, a monochromator tunes incoming X-ray energy by dispersing the photons with a grating and selecting the desired energy using an entrance slit. By increasing the size of this slit, a range of energies (5 eV, for example) can be used to illuminate the sample with incoming energy related to position x on the footprint. The footprint width is dictated by the energy dispersion of the monochromator and the angle of the sample to the incoming beam. To achieve an energy difference of 2 eV at the SXR beamline a footprint with width of about 2 mm must be used.

The conceptual design of the proposed spectrometer is shown in Figure 5. Imaging over the sample footprint is in the non-dispersive direction of the grating; light in the dispersive direction, which corresponds to vertical on the sample, is dispersed by the grating according to its energy and focused. In the typical spectrometer design, rays are not focused in the non-dispersive direction. In order to have a meaningful picture on the detector that gives some information about the sample at increasing time delay or energy, the light in this non-dispersive direction must be focused. This is achieved by adding a concave mirror perpendicular to the grating. The result is an image on the detector in which many XES spectra can be observed as a function of incoming energy or pump-probe delay.

Due to the low cross-section of the X-ray emission technique, one of the most important aspects in the design of the spectrometer is to maximize throughput while maintaining an energy resolution small enough to study the systems of interest. Figure 5 includes a summary of the desired resolution. In the case of surface reactions involving adsorbates on catalysts studied here, 0.25 eV is the best resolution required. It is also useful in some cases to allow high throughput measurements at up to 1 eV resolution. For time-resolved experiments, we define a resolution of 10 μm on the source in the non-dispersive direction corresponding to a 30 fs time-delay (Figure 3). For the RIXS case with varying energy along the sample footprint we wish to have a 0.25 eV incoming energy resolution. Using the SXR beamline design and an incidence angle of 12°, we define this as 200 μm. For the sake of clarity, we refer to resolution in the dispersive direction as the energy resolution of the spectrometer and to resolution in the non-dispersive direction as imaging resolution, which for the RIXS case is in units of energy. Additionally, the spectrometer will be optimized for use at 520 eV, the oxygen K-line.

To analyze the possible designs of the spectrometer, we performed ray-tracing simulations using the SHADOWVUI extension of X-Ray Oriented Programs (XOP) [5]. We tested several illumination distances for the grating and for the non-dispersive focusing mirror. We also looked at the effect of changing the mirror's position, causing changes in the magnification at the detector, and investigated several possible mirror shapes.

RESULTS

The final design of the spectrometer has three main components: the imaging mirror, dispersive grating, and detector. The detector to be used has already been specified—a 4 x 4 cm multichannel plate used with a CCD camera to produce a picture of intensity vs. position in two dimensions. Therefore, our main focus has been analyzing the performance of the dispersive grating and the imaging mirror.

Dispersive Grating

The most basic spectrometer design consists of a dispersive grating and detector; this is a useful starting point for the new design. We choose three parameters of the grating that will dictate the geometry of the entire spectrometer: the radius of curvature R at the center, the ruling spacing d, and the incidence angle α. From these parameters one can mathematically determine the total length of the spectrometer and the energy dispersion at the detector. The energy dispersion describes the way that the vertical detector position changes with photon energy. For soft X-ray applications the incoming angle must be small because many of the X-rays will be absorbed by the grating material. This is limited by using grazing incidence near or below the material's critical angle. For a nickel coated mirror at 520 eV this angle is on the order of 4°; above this angle the reflectivity decreases sharply. This design uses an incidence angle of 5°, or $\alpha = 85°$ (α is defined relative to normal incidence).

We define the radius of curvature R and the ruling spacing d based on the spectrometer's desired length and dispersion. We wish to have a total length close to 1 meter and an energy dispersion of about 0.4 μm/meV. A reasonably small source size (about 10 μm) should have an energy resolution close to 0.25 eV, the high-resolution design goal. This fact dictates the energy dispersion required of the grating. The proper spectrometer length and energy dispersion are achieved with $R = 7.5$ m and a grating spacing of 800 rulings per mm.

From α, R, and d we can determine the correct position of the sample, grating, and detector such that the light coming off of the grating is focused at the detector. These positions are determined using the Rowland Circle geometry shown in Figure 6. For a grating with radius of curvature R there exists a circle of diameter R such that if the source of light is on this circle, then the diffracted image in all orders will be in focus at points on the circle. From these conditions, we can calculate the correct positions of the grating and detector relative to the sample (summarized in Figure 6). The spectrometer is operated in negative order to increase the distance between the sample and grating, which is necessary for the inclusion of the non-dispersive focusing mirror.

The grating's shape is that of a symmetric ellipse. This shape produces lower spherical aberrations than a spherical reflector would, allowing a longer section of the grating to be illuminated before similar errors to a sphere are produced. A longer illumination distance allows more rays to strike the grating, resulting in a higher overall throughput. The best shape of this ellipse is determined through calculations that show the energy dispersion at different offsets from the center of the grating. The relationship between the major and minor axes' lengths is constrained by the requirement that the curvature at the center of the grating be equal to 7.5 m. This allows us to manipulate the length of the major axis of the ellipse and see the effect on the energy dispersion. The best shape is that for which the dispersion is symmetric about the center of the grating. In this case, we use a major axis of about 55 cm for which the corresponding minor axis length is 4 cm.

Imaging Mirror

To allow for imaging in the non-dispersive direction, we consider a mirror reflecting this direction, focusing at the same spot as that defined by the grating. The position, shape, and grazing are chosen such that light from the source will be focused at the position of the detector. The mirror's position also determines the magnification at the detector.

The “perfect” shape of the mirror is dictated by the source distance, the detector (image) distance, and the incidence angle. For the same reasons as in the case of the grating, a very grazing incidence angle must be used; here we choose 4° again based on the reflectivity of a nickel coated mirror. SHADOW calculates the ideal shape of the mirror given its geometry (elliptical or spherical), the object and image focus distances, and the incidence angle. We analyzed the performance of three possible mirror shapes: a sphere, a symmetric ellipse, and an ideal ellipse. Figure 7 illustrates the distinction between the ideal and symmetric ellipses. What we call an ideal ellipse is the small section of the ellipse with foci at the sample and the detector for which light from the sample is incident at the specified angle. The symmetric ellipse is a section of the ellipse centered at one of the ellipse’s axes, thus its curvature is symmetric about its center.

The first case simulated in SHADOW was 1:10 imaging for an ideal elliptical mirror at 4° incidence. In addition to viewing the image spot at the detector we used a SHADOW post processor called FOCNEW to determine if the mirror was focusing as required. FOCNEW computes the waist size of the X-ray beam as a function of its position from an optical element and returns the position in which this size is smallest. Based on the parameters of the mirror, the best focus should be at the same location as the detector.

The result of analysis in FOCNEW was unexpected: the mirror’s focusing ability was extremely sensitive to source’s size and angular divergence. Specifically, if a source divergence of 0.015 rad is used the position of best focus deviates from the detector position for any source size larger than 10 μm. This sensitivity is related to the departure of an ideal point source towards a source that is closer to a plane wave. This issue is limited to the width of the source as this is the direction in which this mirror is focusing; changing the source height or divergence in the vertical direction has no effect on the focus. Further simulations in SHADOW showed that for larger source sizes up to 1 mm the use of a larger divergence (around 0.15 rad) corrected the focusing issue, however here the detector image was very distorted and asymmetrical.

To better understand how the imaging distortion affects the performance of the spectrometer we carried out a series of simulations in which for different magnifications we determined the maximum divergence that could resolve 200 μm on the source, conforming to the design goal. For example, in the 1:10 case, a spot 200 μm on the source corresponds to a 2 mm spot on the detector, thus we determined which source divergence produced a 2 mm spot. For these simulations, we used a point source 500 μm offset from the center, i.e. the edge of a 1 mm sample footprint. Our goal was to determine if a smaller magnification would decrease the effects of this focusing problem. The results can be seen in Table 1. The maximum divergence remained steady around 0.01 rad for magnifications between 1:10 and 1:2 but decreased in the demagnification cases. Based on these results there is no reason to use a smaller magnification than 10, since it would decrease the achievable resolution without improving the instrument's acceptance. A magnification higher than 10 is impractical, as the mirror would need to be placed too near the sample.

We also investigated the effect of the mirror's shape on focusing, with the goal of determining whether a spherical or symmetric elliptical mirror would perform better than the ideal ellipse. For the spherical mirror, SHADOW calculates the proper radius based again on the image and object focal distances and the incoming angle. The proper shape of the symmetric elliptical mirror can be determined using the radius that SHADOW calculates for the spherical mirror as the radius of curvature of the center of the ellipse. From this we know the Rowland circle radius and can find the major and minor axis lengths. In the 1:10 case using a 5 μm source, neither the spherical nor symmetric elliptical mirror was able to produce a source resolution of 200 μm for any divergence greater than 0.001 rad. Based on this observation, alternative shapes to the ideal ellipse were discarded.

From these results we can conclude that of the configurations that were investigated, the 1:10 ideal ellipse at 4° is the best choice for the non-dispersive focusing in this design.

Resolution Determination

The proper length of both the mirror and the grating now must be determined. Here, the balance between resolution and flux has a large impact—longer illumination distances on the optics will create high throughput but also cause imaging error through spherical aberrations present when concave optics are illuminated far from their centers. We optimized this balance by finding the longest illumination of the mirror and grating conforming to the previously listed energy and imaging resolutions required for the different applications (Figure 5). This is achieved by simulating the resolution at different illumination lengths on both the mirror and the grating. The results are shown in Table 2. We examine the case of a point source, a small source (50 μm offset point source) as in the time-resolved case, and a large source (RIXS case, 500 μm offset). The resolution is measured by finding the full width at half maximum (FWHM) of the spot on the detector in each axis (vertical corresponds to imaging, and horizontal to energy). These lengths can then be converted to time or energy resolutions using the relationships between sample position and energy in the non-dispersive direction, or the calculated energy dispersion in the dispersive direction. Usable combinations of illumination lengths are those for which the resolution is lower then the specified requirements.

According to this data, it appears as though the grating can be illuminated up to 16 cm and still have the appropriate resolution. However, at illuminations above 10 cm an asymmetry appears in the detector image due to spherical aberrations. This asymmetry is not easily identifiable in the numerical data because the FWHM was used to find the spot size, and the intensity of the asymmetry is low enough that it does not significantly alter the value of the FWHM. Figure 8 shows the detector footprint for a 10 cm grating illumination and for a 12 cm illumination. Here, it is easy to see the asymmetry that appears. One can determine whether it is beneficial to illuminate more than 10 cm of the grating by looking only at the contribution of the outer regions. If the main contribution of these regions is to the asymmetry in the image rather than to the peak then it is not useful to include them. The results of a simulation in which only the outer 1 cm of the grating was illuminated are shown in the graph in Figure 9. Around an illumination of 12 cm, once can see that the position of the image begins to deviate significantly from the center. Therefore, it is wise to exclude any length greater than 12 cm on the grating for measurements requiring high energy resolution. For applications in which a high resolution is less important, illuminating up to 18 cm of the grating allows greater throughput at resolutions up to 1 eV.

From these simulations, we determine that a grating illumination of 10 cm is acceptable for both of the spectrometer's applications. In the case of time-resolved XES, a time resolution around 30 fs on the sample is achieved for a mirror illumination of 1.8 cm. For the RIXS case, a resolution of 0.25 eV is achievable with 3 cm illumination of the mirror.

DISCUSSION AND CONCLUSIONS

Figure 10 depicts the final schematic of the design, along with a summary of the parameters determined from ray-tracing simulations that optimize its performance. With these parameters, the instrument is able to image sample footprints smaller than 1 mm. Under these conditions, time-resolved XES measurements can be performed as well as RIXS measurements in which a small spot size is possible. Figure 4 shows how this spot size is limited by the setup of the beamline at which these measurements are performed, providing a comparison between the SXR beamline of LCLS and beamline 13 at the Stanford Synchrotron Radiation Lightsource (SSRL). Currently, the energy spread per mm on the sample at SXR is such that to image an energy difference of 2 eV a sample footprint of width 2 mm or more would be required. This width can be minimized by placing the sample at normal incidence to the X-ray beam, however this would prevent the study of surface catalysis for which grazing incidence on the sample is needed. The setup at SSRL beamline 13 allows an acceptably small footprint size for these measurements, which allows steady state RIXS as well as picosecond time resolved measurements using the low-alpha operational mode of the Stanford Positron-Electron Accelerating Ring (SPEAR). However, LCLS's peak power and femtosecond resolution are ultimately needed for many of the prospective chemical dynamics studies considered here. At the SXR beamline, applications to surface reactions would be possible with an upgraded monochromator having lower energy dispersion (i.e. more eV per mm). A beamline at the future LCLS-II could also be designed with dispersion low enough to allow these measurements. The realization of one of these possibilities will make this spectrometer a powerful tool for the study of chemical systems.

ACKNOWLEDGMENTS

This research was completed at SLAC National Accelerator Laboratory between June and August 2011. Many thanks to my mentor, Dennis Nordlund, for his enthusiasm and guidance. I'd also like to acknowledge the funding provided to the SULI program by the U.S. Department of Energy, and thank the staff at SLAC who make this program such a success. Finally, thanks to my fellow SULI students for making this summer so memorable.

REFERENCES

[1] A. Nilsson, "Applications of core level spectroscopy to adsorbates," *Journal of Electron Spectroscopy and Related Phenomena*, vol. 126, no. 1-3, pp. 3 – 42, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S036820480200141X

[2] S. Dahl, A. Logadottir, R. C. Egeberg, J. H. Larsen, I. Chorkendorff, E. Törnqvist, and J. K. Nørskov, "Role of steps in N2 activation on Ru(0001)," *Phys. Rev. Lett.*, vol. 83, no. 9, pp. 1814–1817, Aug 1999.

[3] R. Schrock, "Nitrogen fix," *Technology Review*, May 2006. [Online]. Available: http://www.technologyreview.com/article/16822/

[4] A. Nilsson and L. G. M. Pettersson, "Chemical bonding on surfaces probed by X-ray emission spectroscopy and density functional theory," *Surface Science Reports*, vol. 55, no. 2-5, pp. 49 – 167, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167572904000573

[5] M. Sanchez de Rio and R. J. Dejus, "Status of XOP: an X-ray optics software toolkit," in *Proc. SPIE*, vol. 5536, 2004, p. 171. [Online]. Available: http://dx.doi.org/10.1117/12.560903

FIGURES

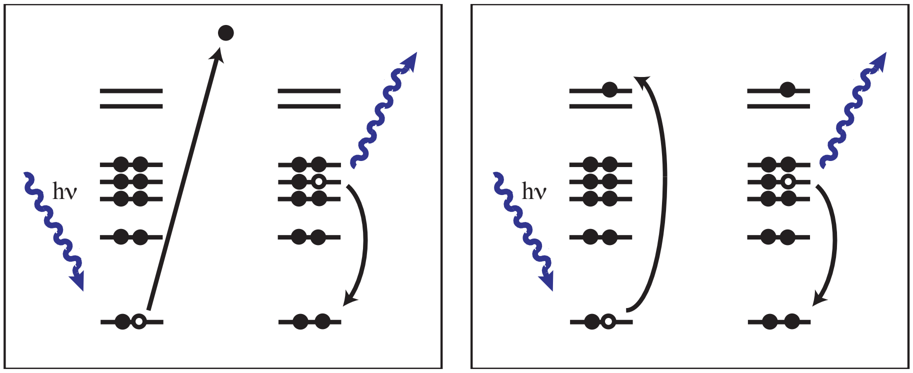


Figure 1: Illustration of the non-resonant (left) and resonant (right) cases of XES. Incoming X-ray radiation excites an electron from the atom's core level, ejecting it in the non-resonant case or promoting it to an unoccupied orbital level in the resonant case. A valence electron of the atom then falls to fill the core level vacancy, emitting a photon with energy equal to the energy difference between its initial and final states.

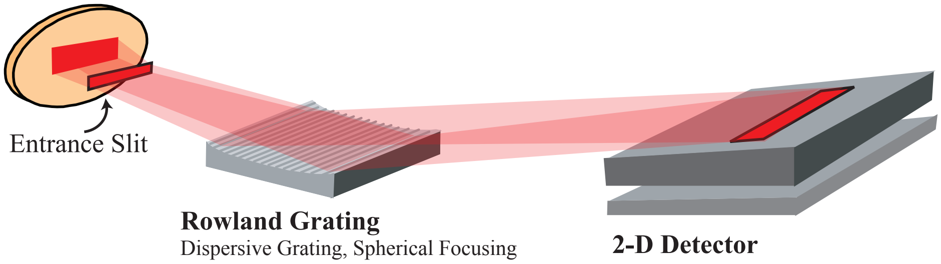


Figure 2: The basic schematic of a typical X-ray emission spectrometer, consisting of a dispersive grating and detector.

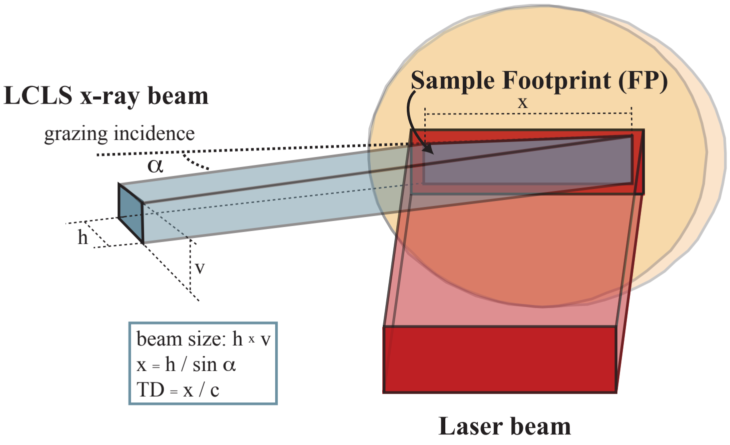
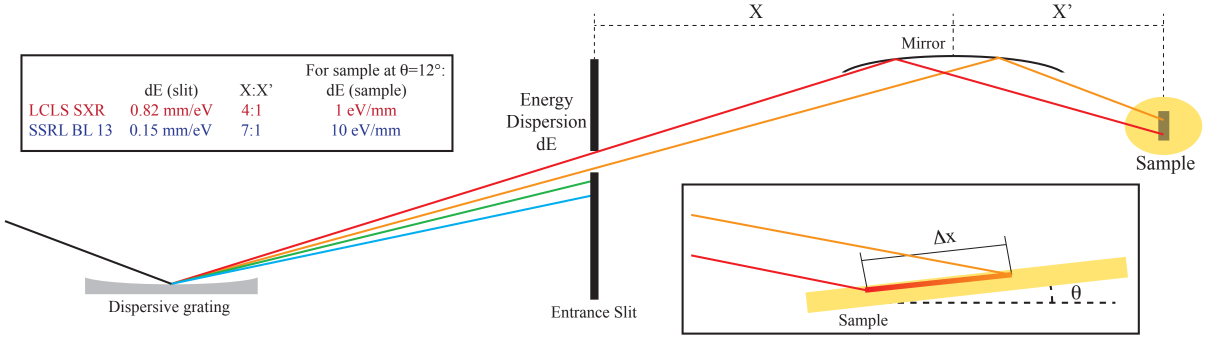


Figure 3: Illustration of imaging in the non-dispersive direction with variable time delay. For time-resolved imaging place the X-ray probe at an angle to the laser pump at normal incidence, causing the probe to arrive at a time delay linearly related to position x on the footprint.



	dE (slit)	X:X'	For sample at θ=12°: dE (sample)
LCLS SXR	0.82 mm/eV	4:1	1 eV/mm
SSRL BL 13	0.15 mm/eV	7:1	10 eV/mm

Figure 4: Illustration of the SXR beamline monochromator. Widening the entrance slit allows a range of wavelengths to illuminate the sample footprint. The footprint width is dictated by the monochromator's energy dispersion at the sample and the angle at which the sample is placed relative to the X-ray beam. The current setup at SXR allows 1 eV to be imaged per mm on the sample; we also provide a comparison to the monochromator at SSRL beamline 13 for which the dispersion at the sample is ten times smaller. An improved setup would have the ability to produce a higher energy difference at more grazing incidence on the sample.

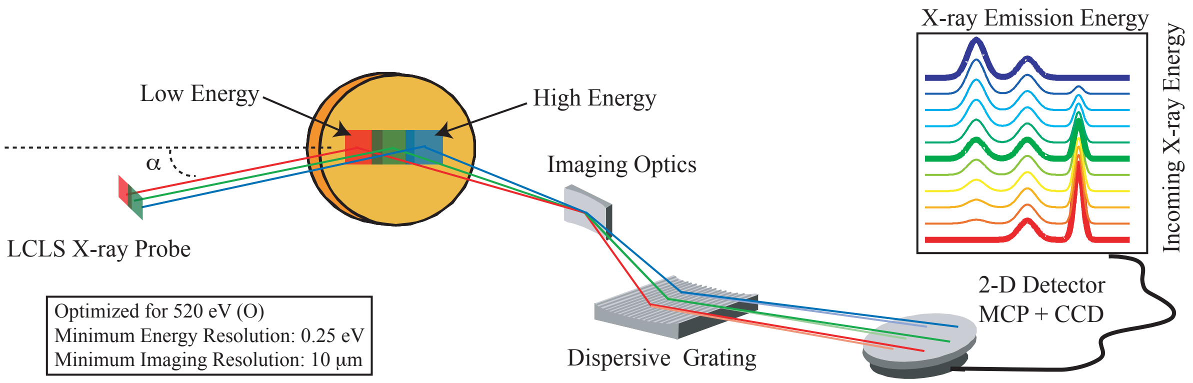


Figure 5: A schematic of the proposed spectrometer. The incoming X-rays vary in energy along the sample footprint x. X-rays emitted from the sample are focused in the non-dispersive direction by a concave mirror. An elliptical grating disperses the rays according to energy and focuses them in this dispersive direction, producing the detector image shown in the figure. The time-resolved case is very similar: a laser pump is added at an angle to the X-ray probe such that time delay varies across x. The 2D detector displays X-ray probe delay in the vertical rather than incoming X-ray energy.

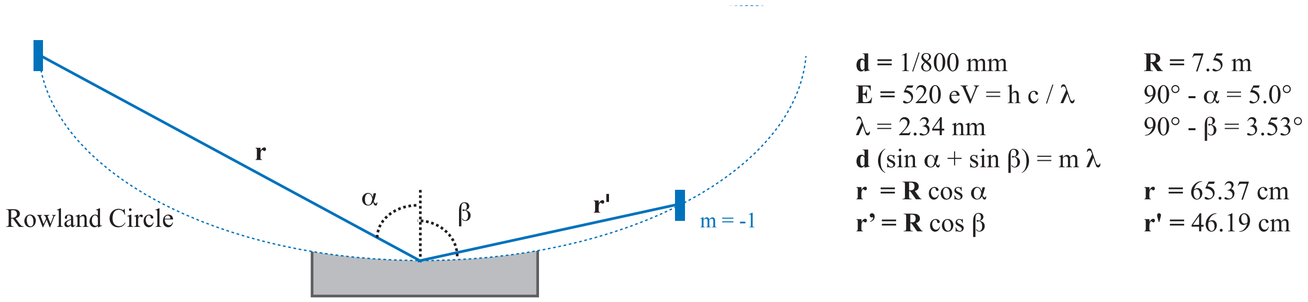


Figure 6: Illustration of the sample-grating-detector setup, showing the Rowland circle geometry and the details of calculations of the correct grating and detector positions. The Rowland Circle is a circle with radius one-half the radius of curvature of the center of the grating. If the source is located on this circle, the diffracted image in all orders will be in focus on the circle.

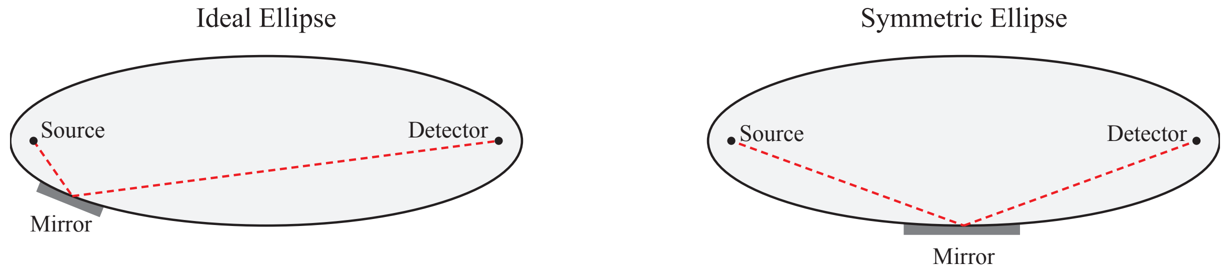


Figure 7: Distinction between symmetric and ideal elliptical geometries. In the proposed spectrometer design, the non-dispersive focusing mirror has an ideal elliptical shape, while the grating is a symmetric ellipse with focusing dictated by the Rowland condition

.

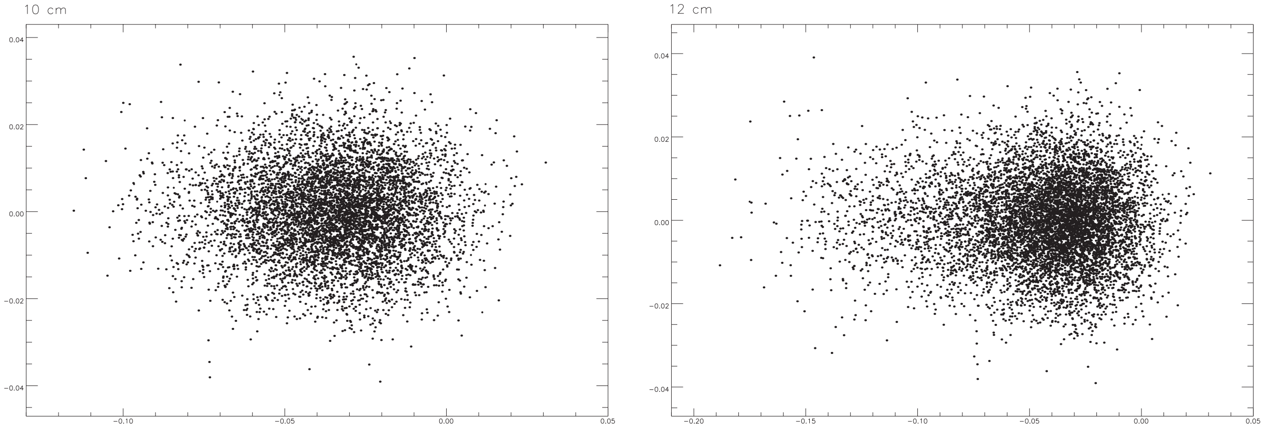


Figure 8: Detector footprints in SHADOW for a grating illumination of 10 cm (left) and 12 cm (right). These show the picture of the X-ray spot on the detector, with the dispersive direction being in the horizontal and the non-dispersive in the vertical. One can easily see the asymmetry that begins to appear in the 12 cm case due to aberrations resulting from the illumination of the grating far from its center.

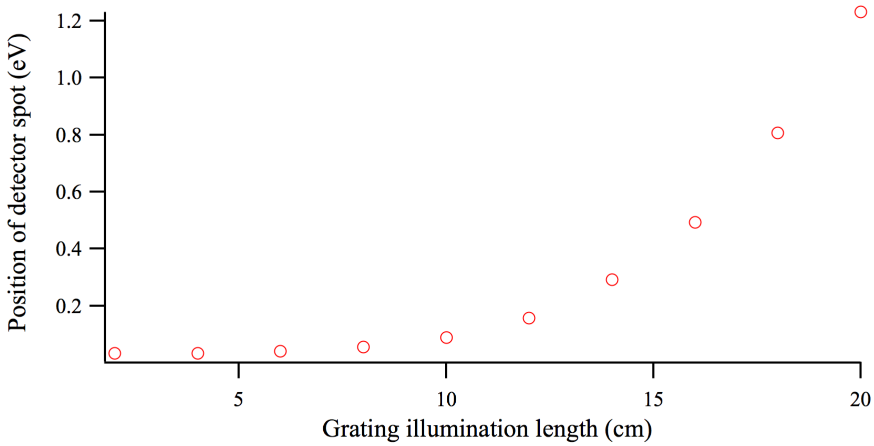


Figure 9: The contribution of only the outer 1 cm of the grating illumination to the detector spot size. If the position of the spot's center deviates significantly from its position when the center of the grating is illuminated, this section of the grating contributes mainly to the asymmetry in the image at the detector. The position is given in eV for ease of comparison to the design requirements. We see that for a resolution of 0.25 eV, we can illuminate 12 cm of the grating; for 1 eV resolution up to 18 cm can be illuminated.

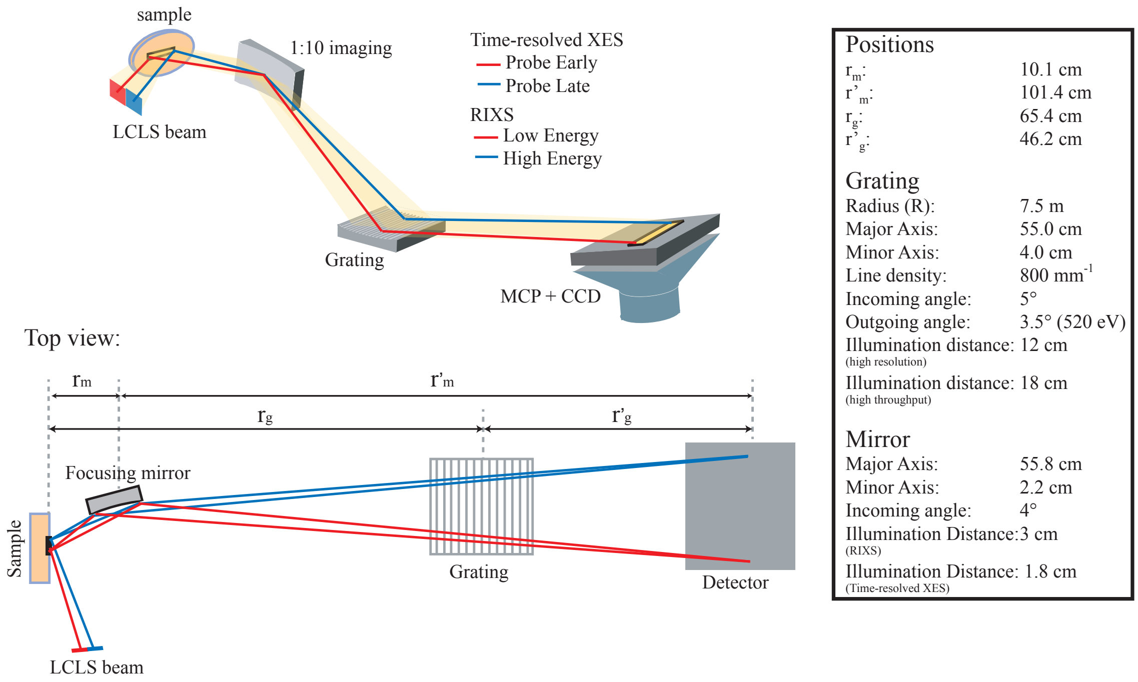


Figure 10: Schematic of the proposed design, including summary of the component parameters.

TABLES

Table 1: The maximum source angular divergence that allows 200 μm source resolution for a 5 μm source offset 500μm at different magnifications of the source. From this data we determine the best position of the non-dispersive focusing mirror to be that which allows 1:10 imaging.

Magnification	Sample Resolution	Source Divergence (rad)
1:10	2 mm	0.011
1:7	1.4 mm	0.011
1:5	1 mm	0.011
1:3	600 μm	0.011
1:2	400 μm	0.009
1:1	200 μm	0.007
2:1	100 μm	0.005
3:1	66 μm	0.0035
5:1	40 μm	0.0023
7:1	29 μm	0.0017
10:1	20 μm	0.0013

Table 2: The results of simulations in which different grating and mirror illumination lengths were tested to determine which combinations produce the necessary resolution for the instrument. In the time-resolved XES case, we used a point source offset 50 μm from the center to simulate a 100 μm source; for the RIXS the offset was 500 μm for a 1 mm source. Blue cells represent those combinations which yield an acceptable resolution. Yellow cells appear to have acceptable resolution but have grating illuminations at which there is significant image error.

Time Resolution (fs)

		Mirror length (cm)				
		3.0	2.4	1.8	1.2	0.6
Grating length (cm)	16.0	48.416	41.619	31.249	20.373	9.478
	12.8	48.345	41.796	29.511	19.476	9.775
	9.6	47.315	41.397	31.101	19.890	9.787
	6.4	47.042	40.950	31.216	19.916	9.647
	3.2	46.995	39.051	30.744	20.471	9.325

Energy Resolution (eV)

		Mirror length (cm)				
		3.0	2.4	1.8	1.2	0.6
Grating length (cm)	16.0	0.1230	0.1041	0.0859	0.0678	0.0335
	12.8	0.1230	0.1038	0.0857	0.0674	0.0333
	9.6	0.1231	0.1032	0.0860	0.0680	0.0340
	6.4	0.1227	0.1034	0.0860	0.0673	0.0333
	3.2	0.1215	0.1034	0.0839	0.0667	0.0333

SLAC-TN-12-027

Space Charge Correction on Emittance Measurement of Low Energy Electron Beams

Colleen J. Treado

Office of Science, Science Undergraduate Laboratory Internship (SULI)
University of Massachusetts Amherst
SLAC National Accelerator Laboratory
Stanford, CA 94025

August 12, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Cecile Limborg-Deprey at the Accelerator Division, SLAC National Accelerator Laboratory.

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Space Charge Correction on Emittance Measurement of Low Energy Electron Beams. COLLEEN J. TREADO (University of Massachusetts Amherst, Amherst, MA 01003) CECILE LIMBORG-DEPREY (Accelerator Division, SLAC National Accelerator Laboratory, Stanford, CA 94025)

The goal of any particle accelerator is to optimize the transport of a charged particle beam along a set path by confining the beam to a small region close to the design trajectory and directing it accurately along the beamline. To do so in the simplest fashion, accelerators use a system of magnets that exert approximately linear electromagnetic forces on the charged beam. These electromagnets bend the beam along the desired path, in the case of bending magnets, and constrain the beam to the desired area through alternating focusing and defocusing effects, in the case of quadrupole magnets. We can model the transport of such a beam through transfer matrices representing the actions of the various beamline elements. However, space charge effects, produced from self electric fields within the beam, defocus the beam and must be accounted for in the calculation of beam emittance. We present below the preliminary results of a MATLAB code built to model the transport of a charged particle beam through an accelerator and measure the emittance under the influence of space charge effects. We demonstrate the method of correctly calculating the emittance of a beam under space charge effects using a least square fit to determine the initial properties of the beam given the beam size measured at a specific point after transport.

INTRODUCTION

Accelerator physics studies the optimization of the acceleration and transport of charged particle beams. We define a beam through its distribution in phase space using the particles spatial coordinates and conjugate coordinates, or normalized momenta. The spatial coordinates, (x, y, z), correspond to the displacements of the particles from a synchronous particle, corresponding to the centroid of the beam and having a specified design velocity and trajectory. The z-coordinate, aligned with the design trajectory, represents the longitudinal position of the beam along the accelerator, thus leaving the xy-plane to represent the transverse plane. The normalized momenta, (x', y', z'), are taken with respect to the longitudinal position of the synchronous particle, s, rather than time [1].

In a particle accelerator, the beam is guided along a desired path by electromagnetic forces produced by components of the accelerator. These forces, known as Lorentz forces and defined as

$$\mathbf{F} = q\mathbf{E} + \frac{q}{c}(\mathbf{v} \times \mathbf{B}), \tag{1}$$

where q is the charge of a single particle, $\mathbf{E}$ is the electric field, $\mathbf{v}$ is the velocity of the synchronous particle, and $\mathbf{B}$ is the magnetic field, also confine the beam to a narrow region. To good approximation, all of the fields contributing to the Lorentz forces can be assumed as linear, and the evolution of the trajectory of the charged particle beam due to such linear Lorentz forces is known as linear beam dynamics. Such forces are produced by beam transport magnets, specifically quadrupoles and bending magnets, the two of which alone completely achieve the basic principles of linear beam dynamics, enabling beam transport systems to consist solely of linear components [2]. Thus, we can approximate the equations of motion governing the transport of the charged particle beam through the accelerator to first order as harmonic oscillators, as follows:

$$x'' + \frac{\kappa_x}{\gamma m c^2 \beta^2} x = 0, \tag{2}$$

$$y'' + \frac{\kappa_y}{\gamma m c^2 \beta^2} y = 0, \tag{3}$$

$$z'' + \frac{\kappa_z}{\gamma m c^2 \beta^2} z = 0, \tag{4}$$

where $\beta = \frac{v}{c}$, with c being the speed of light, $\gamma = \frac{1}{\sqrt{1-\beta^2}}$, m is the mass of a particle, κ_α represents the action of the transport element on the beam, and primes represent differentiation with respect to s [1].

An important concept in accelerator physics is that of the rms envelope. The rms envelopes represent the size of the beam and are described by

$$\tilde{x} = \sqrt{< x^2 >}, \tag{5}$$

$$\tilde{y} = \sqrt{< y^2 >}, \tag{6}$$

$$\tilde{z} = \sqrt{< z^2 >}, \tag{7}$$

for a centered beam [1]. More important yet is the beam emittance, which represents the volume in phase space occupied by the rms particle beam envelopes and is defined as

$$\epsilon_x = \sqrt{< x^2 >< x'^2 > - < xx' >^2}, \tag{8}$$

with similar definitions in the y and z planes. Under the influence of only linear forces, the emittance is unchanged by the motion of the beam, and a changing emittance indicates undesired nonlinear forces acting on the beam. Therefore, producing and maintaining a small transverse beam emittance is among the fundamental objectives in the design and operation of a high quality accelerator.

While it is possible to describe the transformation of the particle beam through transportation of individual particles, this method is often cumbersome and time-consuming. We demonstrate that it is possible and effective to instead directly transport only the second order moments of the beam. We surround the two-dimensional projection of the particle beam distribution with an ellipse, defined by:

$$\gamma_x x^2 + 2\alpha_x x x' + \beta_x x'^2 = \epsilon_x, \tag{9}$$

for the x projection, where α_x, β_x, and γ_x are the Courant-Snyder, or Twiss, parameters, defined below:

$$\alpha_x = \frac{< xx' >}{\epsilon_x}, \tag{10}$$

$$\beta_x = \frac{< x^2 >}{\epsilon_x}, \tag{11}$$

$$\gamma_x = \frac{< x'^2 >}{\epsilon_x}, \tag{12}$$

[1]. The ellipse has an area of $A = \pi\epsilon_x$, where ϵ_x is the emittance in x [2]. Similar ellipses and parameters exist in the y and z planes. Liouville's Theorem, which states that the density of particles remains constant along the beam transport, or

$$\frac{d}{ds} f(x(s), x'(s), y(s), y'(s), z(s), z'(s)) = 0, \tag{13}$$

where f is the density function of the phase space coordinates [1], ensures that the particles contained within the ellipse remain within the ellipse and that the ellipse's area, and thus the emittance of the beam, remains constant throughout the entire transport along the beam. Therefore, it is possible to describe the transformation of the beam through the transformation of its Twiss parameters alone [2]. As the emittance remains unchanged, measuring the transformation of the Twiss parameters is equivalent to measuring the transformation of the beam's second order moments; therefore, we can model the transport of the charged particle beam through the transport of its second order moments.

A charged particle beam is subject to defocusing electrostatic forces between the charged particles and to focusing magnetic forces, due to the motion of the charged particles and dependent on the velocity of the beam. However, at low beam energies, the electric forces are greater than the magnetic forces; therefore, the self-fields generated from the electric forces non-negligibly defocus, and thus destabilize, the beam. The equations of motion are no longer homogeneous, instead described by

$$x'' + \frac{\kappa_x}{\gamma mc^2\beta^2}x = K\frac{2\pi\epsilon_0}{qN}E_x, \tag{14}$$

$$y'' + \frac{\kappa_y}{\gamma mc^2\beta^2}y = K\frac{2\pi\epsilon_0}{qN}E_y, \tag{15}$$

$$z'' + \frac{\kappa_z}{\gamma mc^2\beta^2}z = \gamma^2 K\frac{2\pi\epsilon_0}{qN}E_z, \tag{16}$$

where $K = \frac{qN}{2\pi\epsilon_0}\frac{1}{\gamma^3\beta^2}\frac{q}{mc^2}$ [1] is the beam perveance, N is the total number of particles in the beam, and ϵ_0 is the permittivity of free space. These space charge effects, represented by the inhomogenous term, must therefore be accounted for in modeling the transport of the beam through the accelerator. The purpose of the project is to model the transport of a low energy, specifically between 8 and 100 MeV, electron beam through an accelerator, accounting for space charge effects when necessary, to accurately measure the emittance of the beam.

MATERIALS AND METHODS

One can analytically transport the beam through solving the differential equations of motion above. However, it is more convenient, and often as effective, to transport the beam through a matrix formalism similar to that of linear optics used in the transport of optical beams. A transfer, or transformation, matrix represents the linear component of the force of a beam transport element, whether it be a quadrupole, bending magnet, or drift space, which is a region between magnets in which no external electromagnetic forces act on the beam. The transport of particle coordinates through that element of the accelerator is given by the product of the coordinate vector, (x, x', y, y', z, z'), and the corresponding transfer matrix of the accelerator element, as below:

$$\begin{pmatrix} x \\ x' \\ y \\ y' \\ z \\ z' \end{pmatrix} = \begin{pmatrix} C_x(s) & S_x(s) & 0 & 0 & 0 & 0 \\ C'_x(s) & S'_x(s) & 0 & 0 & 0 & 0 \\ 0 & 0 & C_y(s) & S_y(s) & 0 & 0 \\ 0 & 0 & C'_y(s) & S'_y(s) & 0 & 0 \\ 0 & 0 & 0 & 0 & C_z(s) & S_z(s) \\ 0 & 0 & 0 & 0 & C'_z(s) & S'_z(s) \end{pmatrix} \begin{pmatrix} x_0 \\ x'_0 \\ y_0 \\ y'_0 \\ z_0 \\ z'_0 \end{pmatrix}, \tag{17}$$

where $(x_0, x'_0, y_0, y'_0, z_0, z'_0)$, is the initial coordinate vector, (x, x', y, y', z, z'), is the coordinate vector after transport, C_α and S_α are the solutions of the differential equations of motion for a given transport element, and C'_α and S'_α are their derivatives with respect to s [2]. It is also possible to calculate the transport of the second order moments of the beam through the same transfer matrices; however, instead of directly transporting the moments as we do the coordinates above, we create a correlation matrix with indices consisting of the second order moments of the coordinate vector. We define the correlation matrix as:

$$\sigma = \begin{pmatrix} <x^2> & <xx'> & <xy> & <xy'> & <xz> & <xz'> \\ <xx'> & <x'^2> & <x'y> & <x'y'> & <x'z> & <x'z'> \\ <xy> & <x'y> & <y^2> & <yy'> & <yz> & <yz'> \\ <xy'> & <x'y'> & <yy'> & <y'^2> & <y'z> & <y'z'> \\ <xz> & <x'z> & <yz> & <y'z> & <z^2> & <zz'> \\ <xz'> & <x'z'> & <yz'> & <y'z'> & <zz'> & <z'^2> \end{pmatrix} \tag{18}$$

[3], and we transport the second order moments through multiplying the transfer matrix by the product of the correlation matrix with the transpose of the transfer matrix, as below:

$$\sigma = M\sigma M^T, \tag{19}$$

[1], where M is the transfer matrix.

In our case, we focus primarily on the use of quadrupoles, rather than bending magnets, as transfer components. Quadrupole electromagnets, which are focusing in one plane and defocusing in the other, can be placed in an alternating focusing and defocusing pattern along the beam line to provide the appropriate restoring forces necessary to confine the charged particle beams in the transverse plane. We therefore model the beam line as a system of quadrupoles and drift spaces, applying the action of each element to the beam as it travels through the accelerator.

We develop a MATLAB code to model the transport and measure the emittance of low energy electron beams, specifically less than 100 MeV, in the accelerator. We do so through the use of transfer matrices and the performance of quadrupole scans to transport both the particles of the beam individually and the beams statistics as a whole, from which we calculate the emittance. We discretize the beamline into separate elements, including the quadrupole magnets and drift spaces, the actions of which we represent with different transfer matrices. In the case of a quadrupole that is focusing in the x-plane and defocusing in the y-plane, the element is represented by the following 6×6 matrix:

$$M_{quad} = \begin{pmatrix} \cos(kl) & \frac{1}{k}\cos(kl) & 0 & 0 & 0 & 0 \\ -k\sin(kl) & \cos(kl) & 0 & 0 & 0 & 0 \\ 0 & 0 & \cosh(kl) & \frac{1}{k}\sinh(kl) & 0 & 0 \\ 0 & 0 & k\sinh(kl) & \cosh(kl) & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & l \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{20}$$

where $k = \sqrt{|K|}$, with K being the focusing strength of the quadrupole magnet, and l is the length of the quadrupole [1]. We represent a drift space through the following transfer matrix:

$$M_{drift} = \begin{pmatrix} 1 & L & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & L & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & L \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{21}$$

where L is the length of the drift space [1]. We systematically move along the beamline, separately multiplying the particle coordinate vector by the corresponding transfer matrix that the beam is currently moving through in order to computationally transport the coordinates of the beam. Simultaneously, we transport the statistics of the beam through the same approach, comparing the transported moments with the second order moments of the transported coordinates to confirm their consistency. We model the action of the entire accelerator on the beam through the matrix product of each successive beamline element's transfer matrix. However, in cases of low energy or high charge, space charge effects on the beam become non-negligible and must be accounted for in the beam transport in order to accurately calculate the beam emittance. We account for space charge effects by applying a kick to the beam at regular, short intervals and represent these effects through a matrix similar to the transfer matrices above:

$$M_{SC} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{\Delta s}{f_{sc,x}} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{\Delta s}{f_{sc,y}} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \frac{\Delta s}{f_{sc,z}} & 1 \end{pmatrix} \tag{22}$$

where Δs is the short interval over which the space charge effects are applied, and f_{sc} is the thin lens focal length due to space charge, defined as:

$$\frac{1}{f_{sc,x}} = \frac{K}{2}[\frac{1}{5}\frac{1}{< x^2 >}]^{3/2} R_D[\frac{< y^2 >}{< x^2}, \frac{< z^2 >}{< x^2 >}, 1] \tag{23}$$

$$\frac{1}{f_{sc,y}} = \frac{K}{2}[\frac{1}{5}\frac{1}{< y^2 >}]^{3/2} R_D[\frac{< z^2 >}{< y^2}, \frac{< x^2 >}{< y^2 >}, 1] \tag{24}$$

$$\frac{1}{f_{sc,z}} = \frac{K}{2}[\frac{1}{5}\frac{1}{< z^2 >}]^{3/2} R_D[\frac{< x^2 >}{< z^2}, \frac{< y^2 >}{< z^2 >}, 1] \tag{25}$$

where K is the beam perveance, defined above, and R_D is the Carlton elliptic integral of the second kind, defined as

$$R_D(x, y, z) = \frac{3}{2}\int_0^\infty \frac{dt}{(t+x)^{1/2}(t+y)^{1/2}(t+z)^{3/2}}, \tag{26}$$

[1]. We thus account for space charge effects by transporting the beam over small subsections

of the beamline through the transfer matrix approach presented above and then applying the space charge effect matrix after each subsection. In other words, we perform the matrix multiplication described above to yield the transported second order moments, which we use, along with the small longitudinal displacement over which we calculated the transport, in the calculation of the space charge matrix. We then transport the beam, either the coordinates or the second order moments directly, through the space charge matrix using the same formalism as without space charge. We mathematically express this method below:

$$U = M \times U_0, \tag{27}$$

$$U_{SC} = SC \times U = SC \times M \times U_0, \tag{28}$$

where U_0 is the initial coordinate vector, M is the transfer matrix, U is the coordinate vector after transport through the transfer element, SC is the space charge effect matrix, and U_{SC} is the coordinate vector after application of space charge effects, and

$$\sigma = M \times \sigma_0 \times M^T, \tag{29}$$

$$\sigma_{SC} = SC \times \sigma \times SC^T = SC \times M \times \sigma_0 \times M^T \times SC^T, \tag{30}$$

where σ is the correlation matrix. This process of repeatedly transporting the beam is equivalent to representing the action of the entire beam as a single transfer matrix, formed from the product of each successive transfer matrix, as expressed below:

$$M = M_{sc,n} M_n M_{sc,n-1} M_{n-1} \cdots M_{sc,2} M_2 M_{sc,1} M_1. \tag{31}$$

From this application of space charge effects, we obtain the true transport of coordinates and moments and thus the correct emmitance of the beam under space charge effects.

The hope of this project is to develop a code that enables us to reconstitute the initial properties of an actual beam, from which we can then calculate the emittance. Because the divergence of the beam is not easily accessible, we measure the xrms value at a desired point on the beamline. From the method used in individual particle transport, we know that

$$\begin{pmatrix} x \\ x' \end{pmatrix} = M \begin{pmatrix} x_0 \\ x'_0 \end{pmatrix} \tag{32}$$

where

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \tag{33}$$

is the x-related portion of the matrix formed from the product of consecutive beam transport matrices. From the above equation, we see that

$$x = m_{11}x_0 + m_{12}x'_0, \tag{34}$$

$$x^2 = {m_{11}}^2{x_0}^2 + 2m_{11}m_{12}x_0x'_0 + {m_{12}}^2{x_0}'^2, \tag{35}$$

which we can express as

$$< x^2 >= a < {x_0}^2 > +b < x_0x'_0 > +c < {x_0}'^2 > . \tag{36}$$

We know the coefficients, a, b, and c, from the indices of M, and we strive to determine the initial second order moments of the beam. Thus, we have three unknowns but only one equation. In order to solve for the unknowns, we introduce more equations by varying the beam elements and measuring the corresponding beam sizes. In our case, we varied only the focusing strength of one quadrupole of the system to calculate the xrms values as a function of varying focusing strength. Using MATLAB's lsqcurvefit function, we least square fitted the coefficients of the above equation to the squares of the measured xrms values, from which we determined the initial properties and calculated the emittance of the beam. If we exclude the space charge effects matrices from the calculation of M, we should expect to see a difference between the corresponding emittance and the emittance calculated from an M that includes space charge, with the emittance including space charge being the correct value. We also benchmark our code through comparison with simulations in ASTRA [4], a macroparticle based space charge code.

RESULTS

Through running our code over drift spaces and quadrupoles, we obtained the same results from individual particle transport as from second order moment transport, thus confirming the consistency between the two methods. Both methods yielded the same second order moments through transport along the beam, and therefore the same emittances, enabling us to continue with the faster method of direct second order moment transport. We ran our code over a system of four quadrupoles with length 0.2062 m and various focusing strengths placed at 2.003, 2.410, 3.216, and 3.6122 meters along a 5.5 meter beamline, comparing our results with those from ASTRA simulations. In the case of a quadrupole system with focusing strengths of -3.416, 2.3, 5,801, and -7 respectively, we found small differences between our distributions transported without regard to space charge and those created in ASTRA simulations (Fig. 1), confirming the necessity of including space charge effects in the transport of a charged particle beam.

Then, using a separate quadrupole system, with quadrupoles of strength -9.059, -7.865, and -1.342 for the first, third, and fourth quadrupoles, respectively, we varied the focusing strength of the second quadrupole. We transported the beam at various charges through the system, using seven different focusing strengths, between 9.5 and 11.5, of the second quadrupole and calculated the beam envelopes at 4.3 meters for each focusing strength. ASTRA simulations of the xrms values as a function of focusing strength for the system described above are shown in Fig. 2. We fit the curves of this ASTRA simulation using the transfer matrix, M, of the beam that excludes space charge effects, thus calculating different normalized emittances for beams of different charges, as we expected (Fig. 3). We then compared this simulated data with the results from our transport code, run over the same setup [Fig. 4]. Using the beam sizes calculated from our transport, we performed the least square fit using both the transfer matrix produced from the product of all beam transfer matrices, including space charge, and the transfer matrix produced without space charge. Similar to the ASTRA data and expected results, we saw a constant emittance from charge to charge when including space charge effects in the matrix, and we saw a changing emittance from charge to charge when excluding space charge effects, which we show in Fig. 5.

We also transported the beam over a single drift space of one meter to better compare our results with ASTRA, slicing the beam along the z-coordinate and transporting each slice individually for various beam energies and charges. We introduced a form factor to the charge of the beam, which we also later introduced in our quadrupole scans, in order to fit the xrms values of the beam to the ASTRA data. In the case of the 100 MeV beam, we were able to closely fit the xrms values to the ASTRA data by applying a factor of .045 to the charge of each slice of the beam. We sliced the beam because in lower energy beams, the spread of the energy is large, meaning the linear approximation applied in calculating the space charge may no longer be valid. We expected the slices to be consistent, in that we would need to apply the same factor to each slice, which was true in the case of the 100 MeV beam. While each slice and each charge did not exactly fit the ASTRA data for the 100 MeV beam, our data came very close using the .045 form factor, as seen in Fig. 6. We then modeled this drift in a single slice, the results of which are shown in Fig. 7, finding that the necessary form factor was the previously mentioned factor multiplied by the number of slices, which in our case was ten. Thus, in modeling the beam without slicing, such as in our quadrupole scans, we used a form factor of .0045.

We also studied an 8 MeV beam, for which our results were much farther off from the ASTRA simulations. The necessary factor used to fit the data was much greater, at 0.45, and the data still did not fit the expected results well, as seen in Fig. 8.

DISCUSSION AND CONCLUSIONS

Our initial results indicate a problem in our transport code that is as of yet unclear. Our transported data for the second quadrupole variation does not completely correspond to the ASTRA data for beam charges of 0, 250, and 500 pC. Although we were able to demonstrate the necessity of the inclusion of space charge effects in the calculation of the transfer matrix and thus in the reconstitution of the beam emittance by obtaining a constant emittance with respect to changing charge, our calculated emittance differs slightly from that found in the no space charge case of the ASTRA simulation. This is most likely due to that our transported beam does not correspond exactly to the ASTRA simulated data. As is clear from Fig. 2 and Fig. 4, the beam sizes calculated in our transport code are somewhat different from those calculated in ASTRA, even including the form factor derived from the results of our slicing and drift runs. The form factor was introduced because the distribution of the beam after acceleration is not exactly Gaussian, instead having projections closer to parabolic distributions. Thus, we apply the form factor to the charge in order to account for the different beam volume in the calculation of the chare density of the beam, which we use in calculating the space charge effects. It is also important to note that our our calculation of the space charge includes an approximation of the R_D elliptic integral, as follows:

$$R_D(Z^2, Y^2, X^2) = \frac{3}{XZ}\frac{1}{X+Y}(1-\xi(\frac{\gamma Z}{\sqrt{XY}})), \tag{37}$$

$$R_D(Z^2, X^2, Y^2) = \frac{3}{YZ}\frac{1}{X+Y}(1 - \xi(\frac{\gamma Z}{\sqrt{XY}})), \tag{38}$$

$$R_D(X^2, Y^2, Z^2) = \frac{3}{XYZ}\xi(\frac{\gamma Z}{\sqrt{XY}}), \tag{39}$$

where $X^2 = 5 < x^2 >$, $Y^2 = 5 < y^2 >$, $Z^2 = 5 < z^2 >$, and $\xi(s)$ is defined below:

$$\xi(s) = \frac{1}{1-s^2}(1 - \frac{s}{\sqrt{1-s^2}}\cos^{-1}(s), s < 1 \tag{40}$$

$$\xi(s) = \frac{1}{1-s^2}(1 - \frac{s}{\sqrt{s^2-1}}\cosh^{-1}(s), s > 1 \tag{41}$$

The gamma factor in the $\xi(s)$ approximation was added in an attempt to better fit our data to ASTRA. This gamma factor corresponds to the Lorentz boost, which appears throughout our equations. We should not have had to introduce this factor here, however, which suggests we may be missing a gamma factor somewhere else in our equations. Once this missing factor is found, our data should fit more correctly. We also encountered problems in our results when running over a simple drift space. We focused on applying a factor to fit the xrms data, but in doing so, our x'rms and zrms data does not match the ASTRA simulated results. Because space charge effects depend on the entire volume of the beam, applying a factor to fit the beam size in x should have affected all of the rms values in a similar way. That the form factor does not fit the x'rms and zrms values to the ASTRA data suggests another aspect of our code is negatively affecting our data. Similarly, our results from the 8 MeV beam are significantly different from ASTRA data. As of yet, more benchmarking with the 8 MeV code is required to understand the problem. We plan to continue with our code, benchmarking it with ASTRA to determine where the problem lies. We can also perform a convergence study to determine the best number of steps to use in calculating our space charge, which may also affect our results. Upon fixing our code to correctly transport the beam, we will be able to calculate the correct emittances under space charge effects.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Cecile Limborg-Deprey, for volunteering her time and effort this summer. I am extremely grateful for her kindness and help throughout the past weeks and for providing me with such an interesting and exciting project to work on. I also thank the Department of Energy and SLAC for sponsoring my research through the SULI program. I especially thank my fellow SULI students, who made this summer unforgettable.

REFERENCES

[1] C. K. Allen and N. D. Pattengale, "Theory and technique of beam envelope simulation: Simulation of bunched particle beams with ellipsoidal symmetry and linear space charge forces," Los Alamos National Laboratory, Los Alamos, New Mexico, Tech. Rep., August 2001.

[2] H. Wiedemann, *Particle Accelerator Physics: Basic Principles and Linear Beam Dynamics*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 1993, ch. 4-5, pp. 75–158.

[3] C. K. Allen, "Bunched-beam rms envelope simulation with space charge," January 2005, los Alamos National Laboratory Seminar Presentation.

[4] *ASTRA A Space Charge Tracking Algorithm User's Manual*, July 2008.

FIGURES

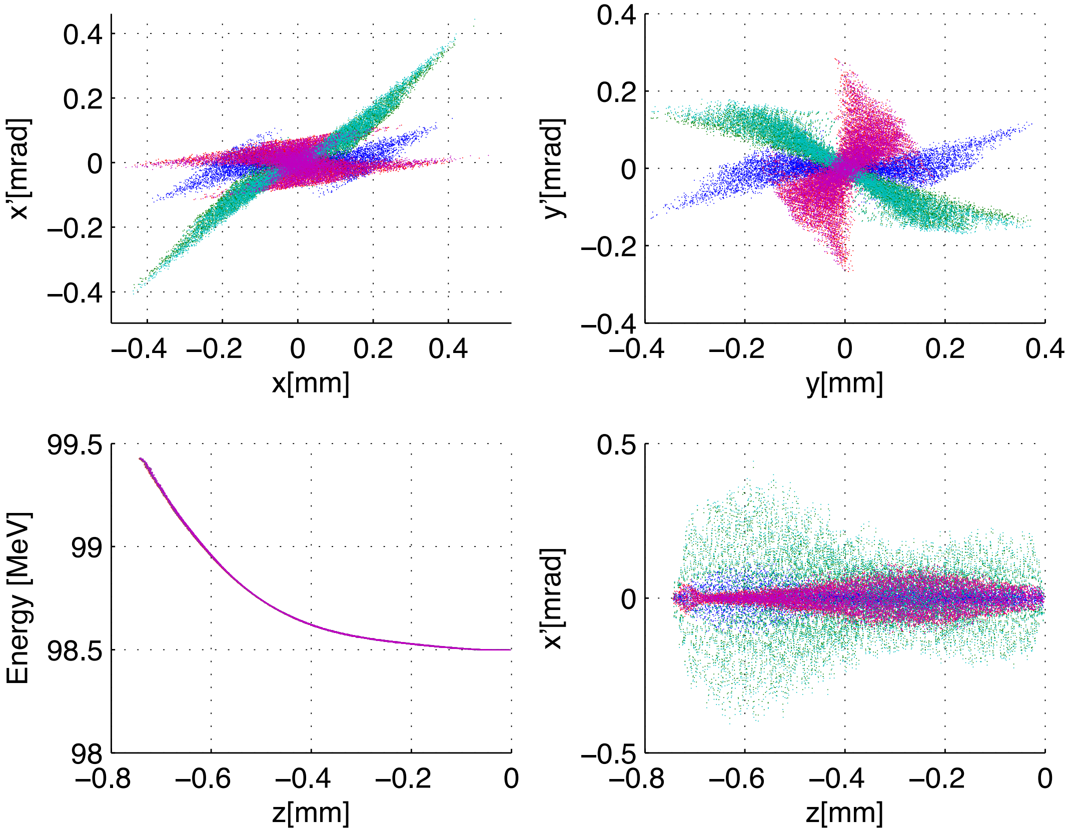


Figure 1: Overlay of transported distributions on ASTRA simulations at $z = 1.9$ m (blue), $z = 2.1062$ m (green, cyan), or immediately following the first quadrupole, and $z = 5.5$ m (red, magenta), or the end of the beamline, for 100 MeV beam. Transported distribution does not take space charge into account. Slight differences between ASTRA and modeled distributions confirm effects of space charge on transport of beam.

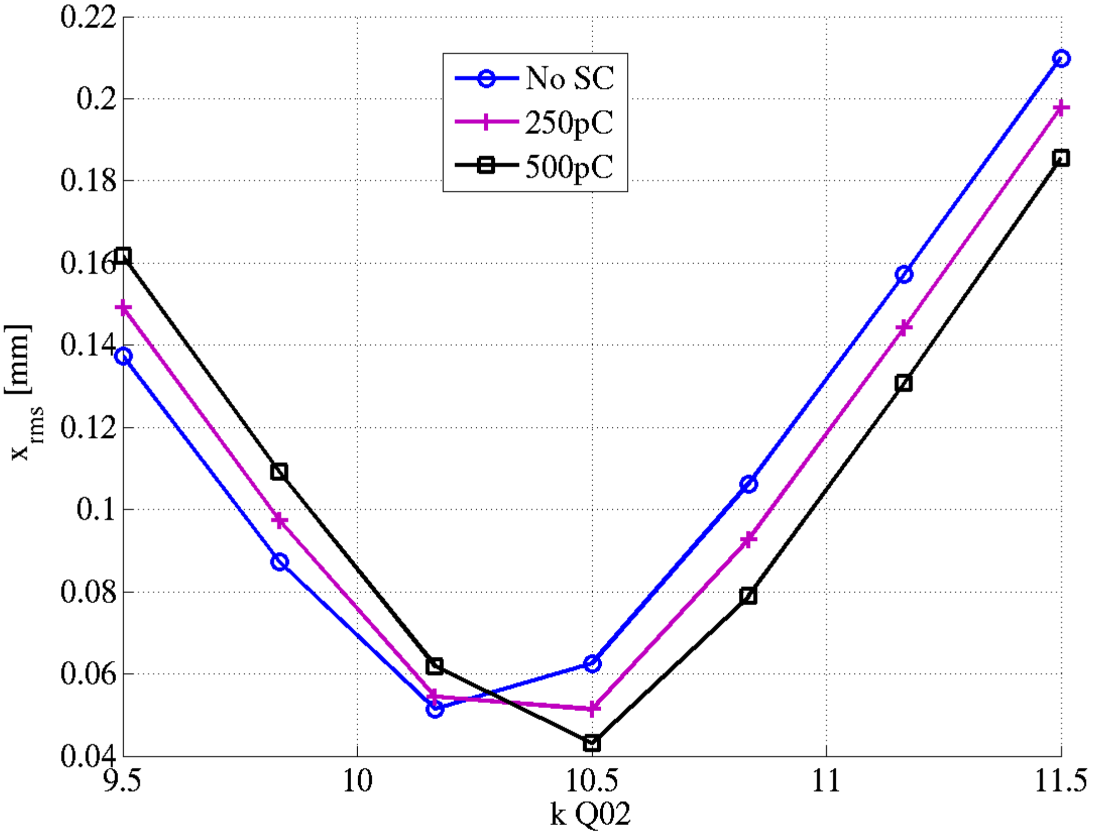


Figure 2: Beam envelope in x (xrms) of 0, 250, and 500 pC charged beam from ASTRA simulations as function of focusing strength of second quadrupole. Space charge effects clearly influence beam size.

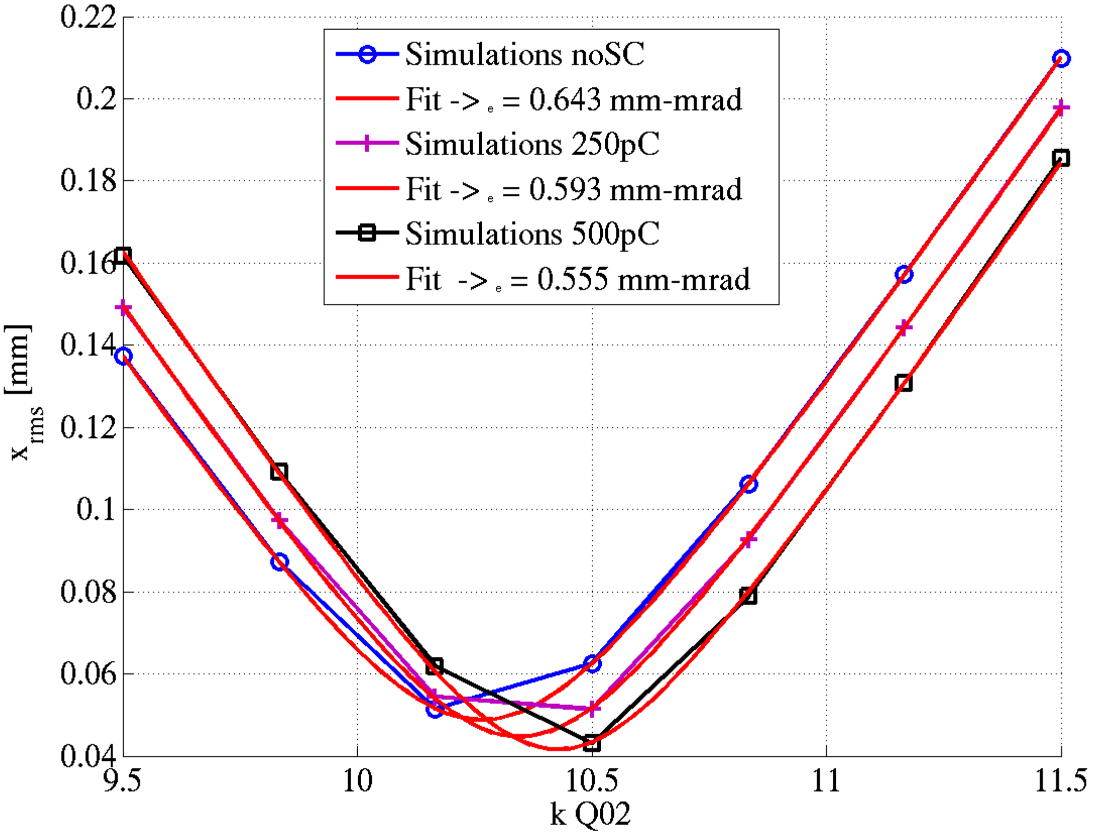


Figure 3: Least square fit of ASTRA simulations of beam envelops in x (xrms) as function of focusing strength of second quadrupole. Matrix used to least square fit does not include space charge, resulting in changing emittance with increasing beam charge.

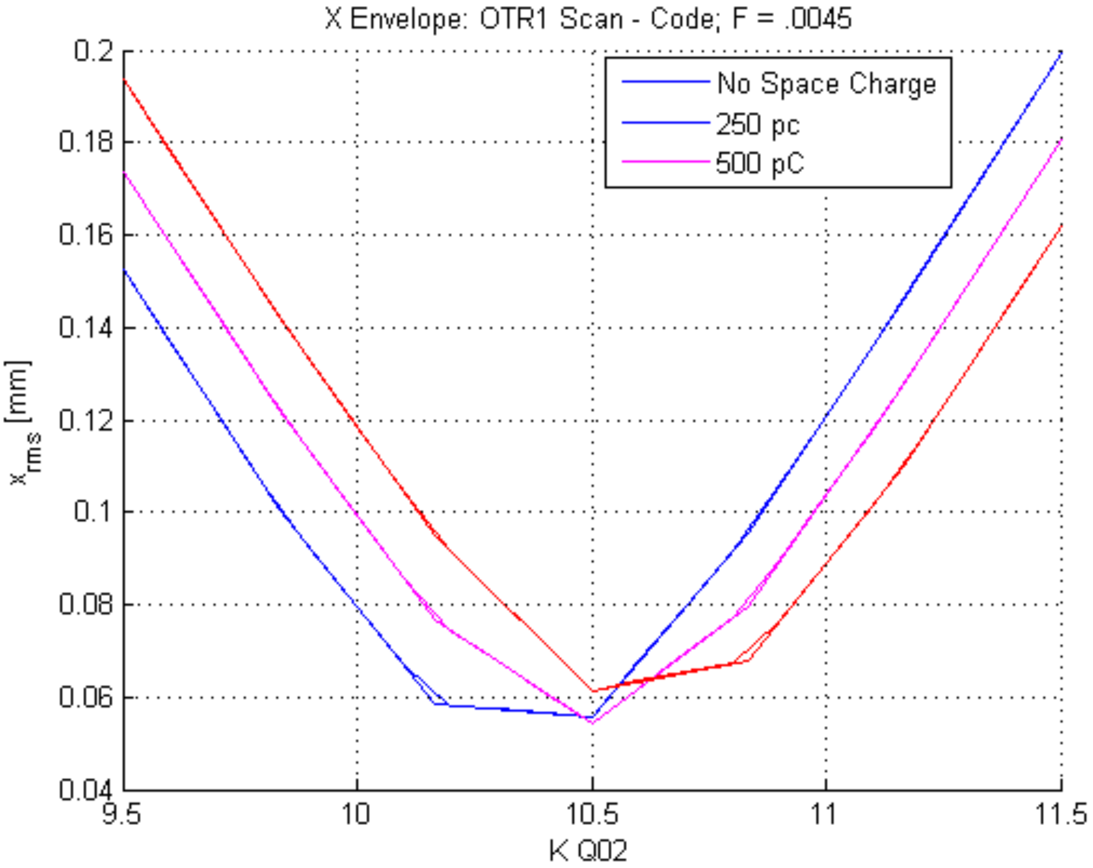


Figure 4: Beam envelope in x (xrms) of 0, 250, and 500 pC charged beam from MATLAB transport code as function of focusing strength of second quadrupole. Form factor of .0045 introduced to charge to better fit results to ASTRA simulations, although beam sizes still differ from expected values. Regardless, fitting with matrix including space charge still results in constant emittance from charge to charge, although calculated emittance differs slightly from expected value.

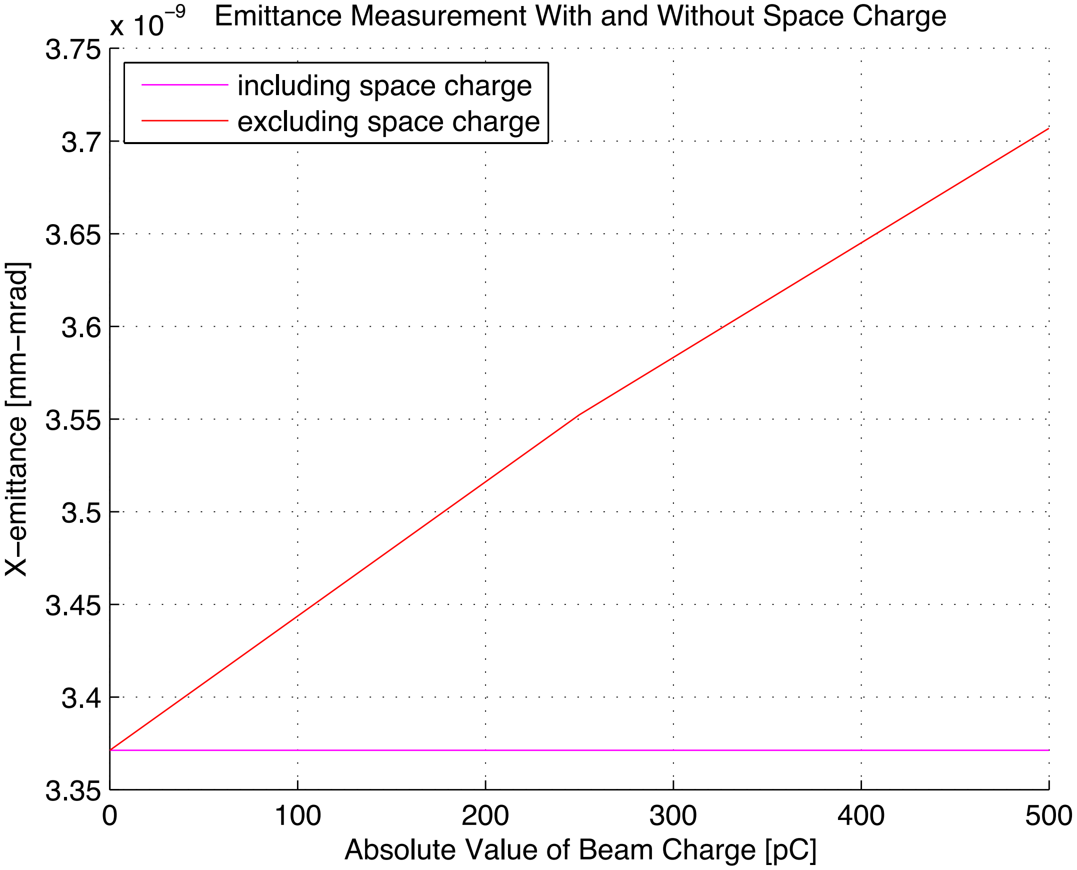


Figure 5: Beam emittance, calculated from least square fit to beam size, as a function of charge. Excluding space charge effects from calculation of transfer matrix, indices of which used as coefficients in least square fit, results in changing emittance with increasing charge.

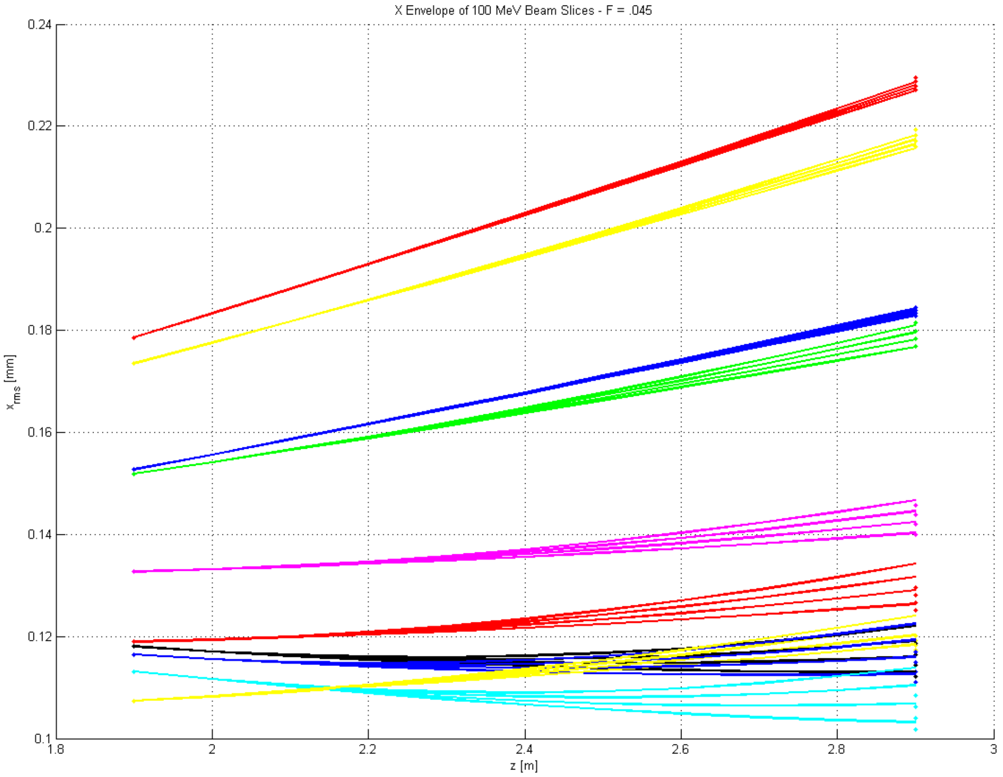


Figure 6: Beam size of 10 z-slices of 100 MeV beam with 250, 500, 750, and 1000 pC charge, respectively, along 1 meter drift. Points represent ASTRA simulated beam sizes at start and end of drift space. Lines represent beam sizes calculated in transport code. Form factor of 0.045 applied to charge for each slice. With increasing charge, difference between transported beam size and ASTRA simulated beam size at end of drift increases.

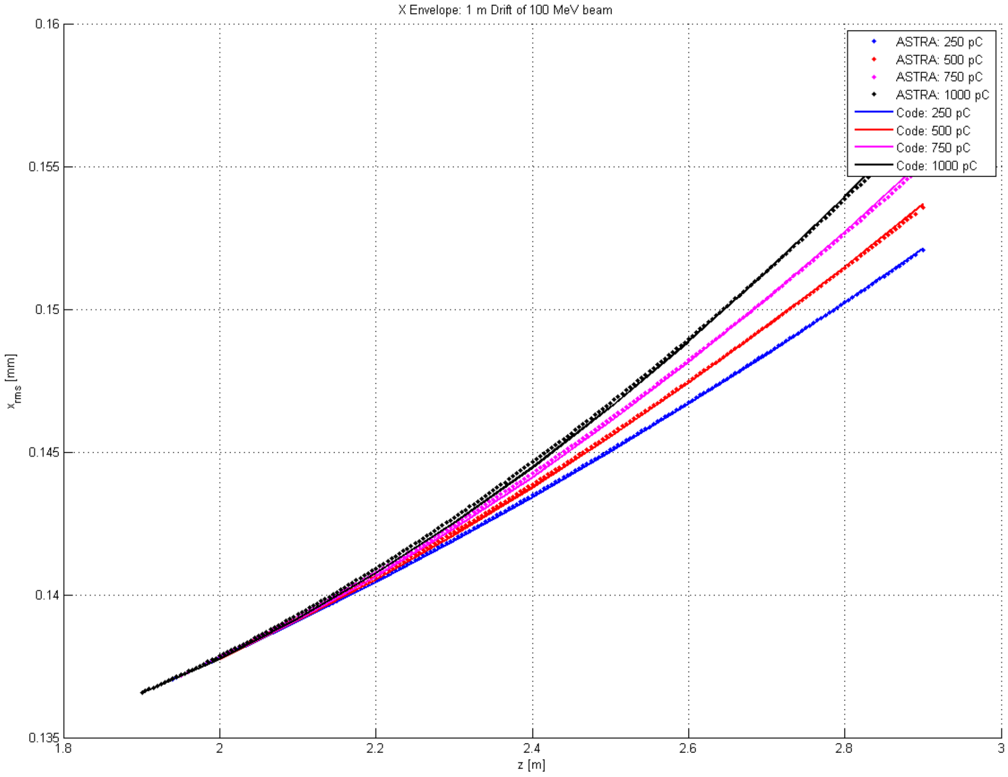


Figure 7: Beam size of 100 MeV beam without slicing and with 250, 500, 750, and 1000 pC charge, respectively, along 1 meter drift. Using 0.0045 form factor on each charge, calculated beam sizes correspond well with ASTRA simulated curves.

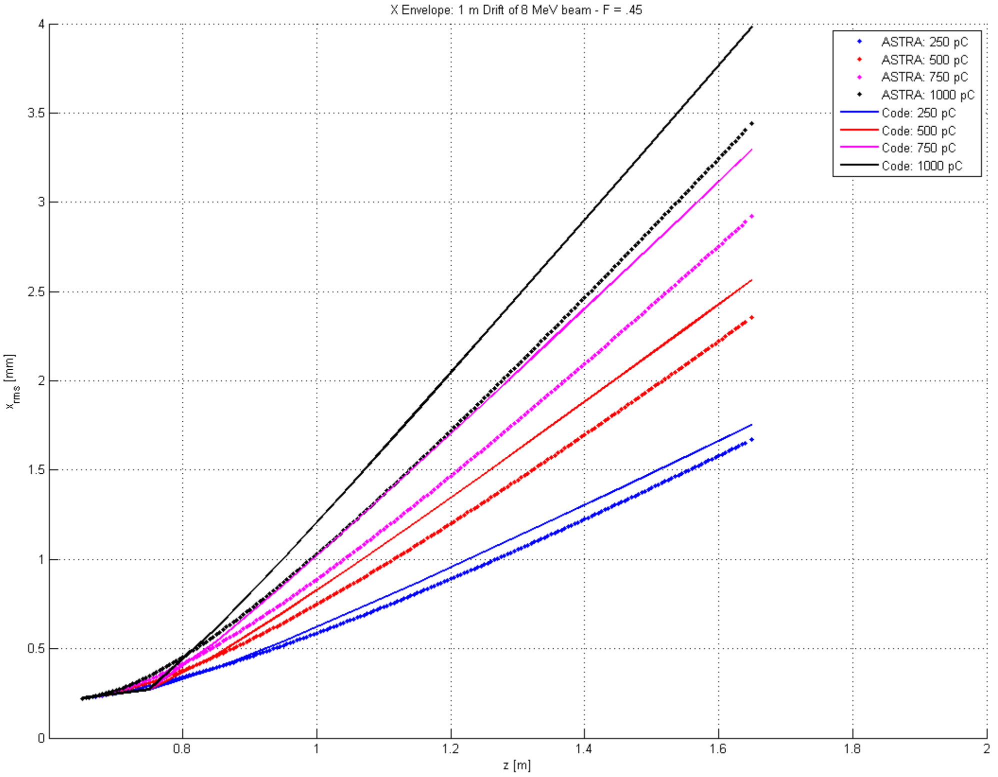


Figure 8: Beam size of 8 MeV beam without slicing and with 250, 500, 750, and 1000 pC charge, respectively, along 1 meter drift. Form factor of 0.45 applied to charge of beam. Clear and significant differences between ASTRA simulated data and data calculated from transport code, with differences in beam size increasing with increasing charge. ASTRA simulates curved growth of beam size along drift, while calculated data models growth of beam size linearly.

SLAC-TN-12-028

Longitudinal Bunch Pattern Measurements through Single Photon Counting at SPEAR3

Hongyi (Jack) Wang

Office of Science, Science Undergraduate Laboratory Internship (SULI) Program

UCSD Electrical Engineering Class of 2013

SLAC National Accelerator Laboratory

Stanford, CA

08/03/2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science undergraduate Laboratory Internship under the direction of Jeff Corbett at SSRL, SLAC National Accelerator Laboratory.

Participant: --
Signature

Research Advisor: --
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Table of Contents:

Longitudinal Bunch Pattern Measurements through Single Photon Counting at SPEAR3. HONGYI WANG (University of California, San Diego, La Jolla, CA 92093). Jeff Corbett (SSRL Operations, SLAC National Laboratory, Menlo Park, CA 94025)

Abstract

The SSRL is a synchrotron light source that provides x-rays for experimental use. As electrons are bent in the storage ring, they emit electromagnetic radiation. There are 372 different buckets which electrons can be loaded into. Different filling patterns produce different types of x-rays. What is the bunch pattern at a given time? Which filling pattern is better? Are there any flaws to the current injection system? These questions can be answered with this single photon counting experiment.

Introduction

The Stanford Synchrotron Radiation Lightsource (SSRL) is a division of SLAC National Accelerator Laboratory. SSRL provides to users electromagnetic radiation in x-ray, ultraviolet, visible, and infrared spectrums for a variety of experiments ranging from basic materials science to advanced protein crystallography and biological research. The various portions of the electromagnetic spectrum are delivered to 14 different beam-lines by means of mirrors, collimators, and aggressively-cooled grating and crystal monochromaters. The electromagnetic radiation is given off as synchrotron radiation by electron bunches stored in the SPEAR3 ring. The electron bunches can be loaded into 372 different buckets, which are radio-frequency power potential wells in the storage ring. Each electron bunch produces synchrotron radiation as they are radially accelerated in bend magnets or more advanced wiggler and undulator magnets.

The circumference of the SPEAR3 storage ring is 234.12 m. The electron bunches therefore revolve with a turn time of 780 ns, corresponding to a 1.28 MHz repetition rate. When all 372 buckets contain electrons, experimenters at the photon beam lines see a photon-beam pulse repetition rate up to 476 MHz. With a maximum of 500 mA total of circulating electron beam current in SPEAR3, the total synchrotron radiation (SR) power ranges up to 450 kW; however, after filtering and collimation, a typical beam line only receives a few mW of photon beam power. For most applications, SR Users are interested in average beam power and not the pulse-by-pulse time structure generated by the individual bunches separated by 2.1 ns. Nevertheless, the 372 different buckets can be filled with different electron 'bunch' patterns.

Dark-space or 'gaps' are normally left between trains of electron bunches, to prevent the trapping and accumulation of positively-charged ions in the SPEAR3 vacuum chamber. As electrons travel through the storage ring, there are some positively charged particles flowing around the tube, despite vacuuming efforts. If the buckets are filled without gaps, the continuous electron beam will attract positively charge particles inside the vacuum chamber towards the beam. As a result, the positively charged particle will be attracted towards the electrons and continuously deflect the electron beam. The trapped ions can lead to electron beam instability and particle loss, due to elastic and inelastic collisions. For this reason, some buckets are left unfilled, to give positively charged particles time to escape the electric field from the beam and thus reduce deflection. A second reason for leaving 'gaps' in the electron bunch pattern is to produce different types of time-dependent radiation for various experimental needs. Certain measurements prefer single isolated photon pulses to image time-dependent sample dynamics. Often a laser is used to 'pump' the sample while the x-ray beam is used to 'probe' the response. Since the fast-gating of the detectors requires up to several hundred nanoseconds, the probe pulse must be isolated from the rest of the bunch pattern. Of significance to this study is the fact that the probe pulse must be very well isolated with charge in adjacent buckets down by a factor of 10^{-5} or less. Some fill patterns are better than others, due to reasons that will be discussed later.

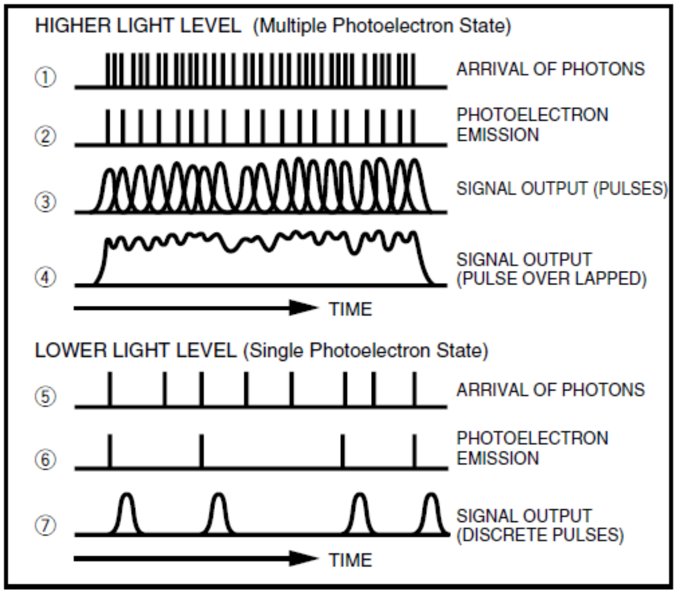
The fill pattern in the SPEAR3 ring is established by a graphical-interface control system that communicates with a complex timing system from the SSRL control room. From the control room, operators inject beam into the individual SPEAR3 buckets, and can monitor the electron bunch pattern as detected from a capacitive pick-up electrode with relative ease; however, this does not have the dynamic range required for isolated-bunch timing experiments. Often it is difficult to detect timing flaws with the current injection system, which can be attributed to frequency drift, electronic jitter or temperature-dependent phase-errors that accumulate in the long-haul timing cables extending between the SPEAR3 control room and the injector. All of these effects can lead to inadvertent injection of charge in buckets adjacent to the intended target bucket. These are the causes of some of the questions investigated with single photon counting experiments carried out in the SSRL synchrotron light monitor diagnostic room, called the light-shack.

By utilizing a single photon counting technique, the electron bunch pattern in the SPEAR3 storage ring was able to be mapped out. From the user side, it was determined which buckets were filled, which were left empty, and in relative terms how many electrons were in each bucket. This information can be used to confirm with the control room to ensure good beam delivery to each beam-line. Although there are other ways to measure the bunch pattern, as previously mentioned, single-photon yields the highest dynamic range. For other applications, such as laser-SR beam cross-correlation to measure individual bunch length, photon production from the cross-correlation experiment is so slow that single-photon counting is the only way to measure the result. Many other single-photon counting applications occur on the SSRL beam lines and throughout both applied physics and low-light intensity measurements, such as those found in astronomy.

Single-Photon Counting Technique:

The technique of mapping the location and relative radiation power of each electron bunch is known as Single Photon Counting. Basically, this requires counting the number of photons hitting the detector from each electron bunch; as the name of the technique suggests, the number of photons hitting are in the single photon range. Therefore, the synchrotron radiation light must be of such low intensity that only one photon strikes the detector in a single counting interval. To do this, a photomultiplier tube (PMT) is used as the detector. A photomultiplier tube consists of a photocathode, electron multipliers, and an anode. When a photon hits the photocathode, electrons are emitted under the photoelectric effect. The number of electrons is then multiplied through cascade electron emissions from the dynodes, or electron multipliers. These electrons are collected by the anode as a voltage pulse output. The output of the PMT is an analog TTL signal for each photon detected, and a fast digitizer is needed to translate and store the output signal in the desired data format. With single photon counting, each pulse in the output signal represents one photon hitting the detector. Since each photon is generated by an orbiting electron bunch in the ring, the timing of the pulses can be tracked, to extrapolate the position of each electron bunch in the storage ring relative to each other.

At nominal electron beam currents of several hundred milliamps in SPEAR3, the photon entering the diagnostic room spans a range of approximately 420-1000 nm, and has a beam power of about 100 uW, which contains way too many photons than we need. Band-pass filters and wide-band neutral-density filters, or "sunglasses", are therefore used to control the amount of power incident on the single-photon counter. Filtering the synchrotron radiation to the single photon range is important, because at high light levels, multiple photons will hit the detector at once. This runs the risk of overloading the detector, as too many electrons may be emitted at one time. Also, if photons hit the detector too frequently, the output pulses will overlap with each other (Figure 1). As a result, there will be a waveform of the train of photons, but no information about each individual photon arrival. As we are interested in studying the individual buckets in the storage ring, as well as the overall distribution, photon counting will give us more accurate information at low light level.


HIGHER LIGHT LEVEL (Multiple Photoelectron State)
①
ARRIVAL OF PHOTONS
②
PHOTOELECTRON EMISSION
③
SIGNAL OUTPUT (PULSES)
④
SIGNAL OUTPUT (PULSE OVER LAPPED)
TIME
LOWER LIGHT LEVEL (Single Photoelectron State)
⑤
ARRIVAL OF PHOTONS
⑥
PHOTOELECTRON EMISSION
⑦
SIGNAL OUTPUT (DISCRETE PULSES)
TIME

1 Figure: Output Pulses from Photomultiplier Tube at Different Light Levels
At high light level, signal output pulses overlap, and single photon arrivals cannot be resolved.
At low light levels, signal outputs are discrete pulses, and single photon arrival can be resolved.

With single photon counting, the beam is filtered down to the ‘single photon’ range to avoid pulse overlaps. This also means that only a few photons hit the detector in one turn, and only limited information about the electron bunches can be gathered in one rotation in the ring. For typical single-photon counting applications, one wants to detect only about 1 photon in every 10 counting intervals, to avoid overload. To solve this problem, a time-correlated integration technique was used. Since only the relative position of electron bunches was of interest, the signal could be integrated from the detector, as electron bunches revolve for multiple turns. Although only a few photons are let through each pass over many turns in the ring, each photon has an equal probability of hitting the detector. While only several pulses are collected in one turn, over millions of turns the full electron bunch profile can be mapped out. Each turn takes 780 ns, so integrating over millions of turns would still allow a reasonable short sampling time.

Materials and Methods

The single-photon counting experiments were conducted in the SSRL Light-Shack. The Light-Shack is a special diagnostic beam-line that only collects radiation in the visible, near IR spectrum (420nm-1000nm). For this experiment, harmful radiation in the x-ray range is not needed, since visible spectrum photons convey the same message about the electron bunch pattern. Although x-ray radiation is filtered out, other properties of the radiation from this diagnostic beam-line is the same as other experimental beam-lines. Therefore, the light-shack is an ideal location in which to conduct this experiment.

When the light comes out of the beam-line, it is guided onto an optical bench. Through a series of focusing mirrors and focusing lenses, the beam is then directed into a light-tight box where the detector sits. In order to filter the incident photon beam down to single-photon counting regime, a combination of neutral density filters and a 10 nm band-pass filter centered at 550 nm were used. Through measurements with a Newport Power Meter, the average power of the beam was measured (Table 1). An estimate of number of photons per turn was calculated using the equation below:

$$\text{Beam Power} = \frac{N \times h \times c}{T \times \lambda}$$

(N represents the number of photons, and is the only unknown variable)

Neutral Density Filter Level	Measured Beam Power(µW)	Number Of Photons in Beam Per Turn (780 ns)
0	206	993123456.8
1	20.4	98348148.15
2	2.58	12438148.15
3	0.458	2208012.346
4	0.0712	343254.321
5	0.01	48209.87654

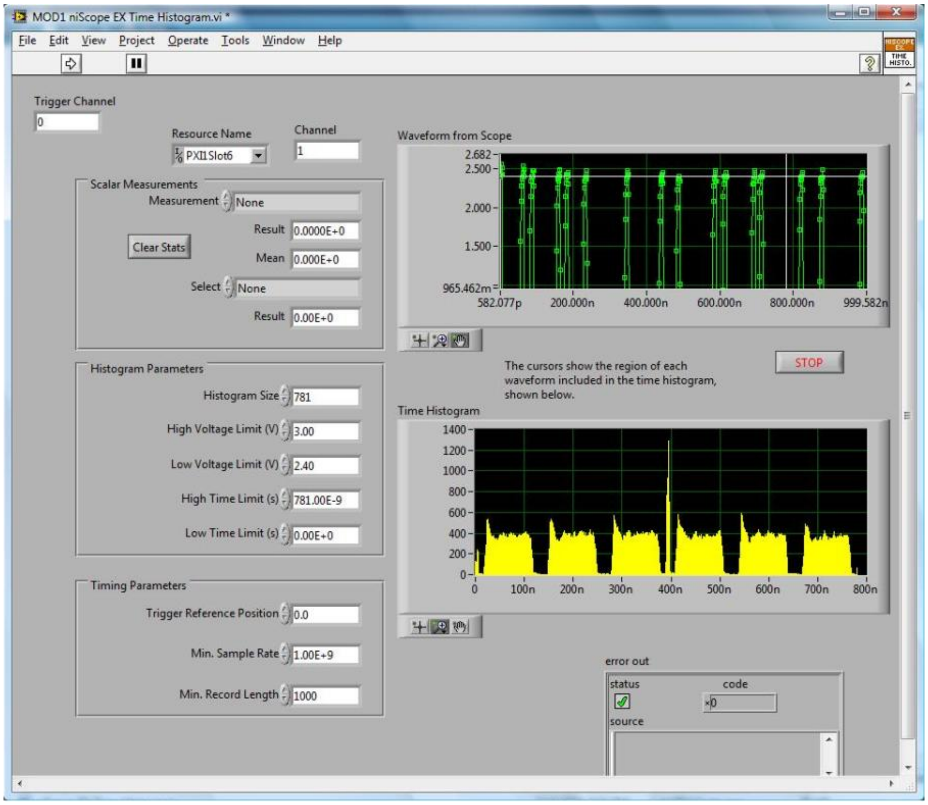
Table 1: Measured beam power from power-meter and corresponding number of photons at each filter level. No band-pass filter used. SSRL Running at 350 mA.

Extrapolating from the data in Table 1, a level 8 neutral density filter in series with a 10 nm band-pass filter only lets through 1 photon out of every 10 turns. Ideally, the neutral-density filters should correspond to beam power reduction in powers of 10. Our results roughly follow the pattern, some deviate may be caused by noise and sensitivity of detector. ND8 therefore reduces beam power by a factor of 10^7. This filtering achieves the desired single-photon range, but slight adjustments are made in accordance with sampling time: fewer photons hitting means a longer sampling time is necessary to produce the same result. The detector was sealed light-tight to block out ambient room light noise, and experiments were run with room lights turned off.

Figure 2: (Left) Hamamatsu Photonics H-7360 Photomultiplier Tube.
(Right) National Instruments 5154 2Gs/s Digitizer

The detector chosen was a Hamamatsu Photonics H-7360 Counting Head. This detector contains a photomultiplier tube that is sensitive to single photons. The detector can be biased with a DC voltage between 3.4 V and 5 V. When a photon hit the detector, a TTL pulse was generated from the anode. This TTL pulse was then sent to a National Instrument 5154 2GSample/s Digitizer for analysis. The digitizer was triggered from the constant 1.28 MHz SSRL ring-clock, a TTL pulse provided from the control room synchronous with the revolution time of the individual bunches. For each ring-clock trigger, the TTL pulses were saved as a ‘time-of-flight’ record for the single-photon arrival times. A Ortec Counting Card was intended to be used originally. It is also a digitizer and produces very high resolution histograms. The Ortec Card was tested with DG535 generators, which generated simulated TTL pulses, and the setup was ready to take actual data. However, the counting card was taken to a different beam-line due to a conflict in experiments, and the National Instruments Digitizer was used instead.

After preparing the detector, software was developed to communicate with the detector output and extract the desired data from the analog TTL signals. Using a Lab-View Module that makes time histograms, we modified the program to work for our specific experiment. With the help with a experienced LabView programmer, we connected all terminals of the block diagram and correct input to parameters were given. Our final program keeps track of all TTL pulses from the detector and makes a time histogram evaluated over the rotating period of the ring, which is 780 ns. Output is stored in a text file. Since each TTL pulse indicates the presence of an electron bunch in the storage ring, a histogram that keeps track of all TTL pulse generates a full longitudinal electron bunch pattern over the course of time. Figure 3 is a screen shot of the user interface of the program.



Trigger Channel: 0

Resource Name: PXI1Slot6

Channel: 1

Scalar Measurements

Measurement: None

Result: 0.0000E+0

Mean: 0.000E+0

Select: None

Result: 0.00E+0

Histogram Parameters

Histogram Size: 781

High Voltage Limit (V): 3.00

Low Voltage Limit (V): 2.40

High Time Limit (s): 781.00E-9

Low Time Limit (s): 0.00E+0

Timing Parameters

Trigger Reference Position: 0.0

Min. Sample Rate: 1.00E+9

Min. Record Length: 1000

The cursors show the region of each waveform included in the time histogram, shown below.

error out: status ✓ code 0

Figure 3: User Interface of Lab View Program
The software contains a Scope function (top right display) that keeps track of TTL pulses after a single trigger. The time-correlated Integration is done in the histogram (bottom right display) that integrates the data from Scope.

Program Parameters:

Parameter Name:	Description:	What was used in experiment:
Histogram Size	number of bins contained in the histogram, histogram resolution.	781(2 bins = 1 bucket)
High Voltage Limit/Low Voltage Limit	This defines the range which the program samples data and includes in the histogram.	2.4V-3.0V (signal from detector falls within this range)
High Time Limit/Low Time Limit	This defines the x-axis of the histogram	0ns-781ns(represents the turn time of SSRL)
Trigger Reference Position	defines the position of the trigger relative to data. At 0, all data collected are after the trigger. At 100, all data collected are prior to trigger.	0 (all data collected after trigger)
Min Sample Rate	This is the actual sample rate. (1 Ghz maximum)	1 Ghz (produces the highest resolution)
Min Record Length	This defines the time length which the detector looks for data after a trigger. It specifies the number of sample points the scope looks for. It interacts with sample rate to define the time length which the detector looks for data after a trigger (Example. At 1Ghz, 700 for record length = samples for 700ns after trigger).	781 (samples 781 data points at 1 Ghz after trigger, which is 781 ns)

Table 2: Lab View Program Parameters
This table contains an description of all parameters associated with our software,
as well as the values we used for our experiment.

Results

Using our experimental setup, we were able to get very good data. A few different bunch patterns were detected and analyzed.

The Following results were taken during normal SSRL operations, when the pattern was run for user experiments.

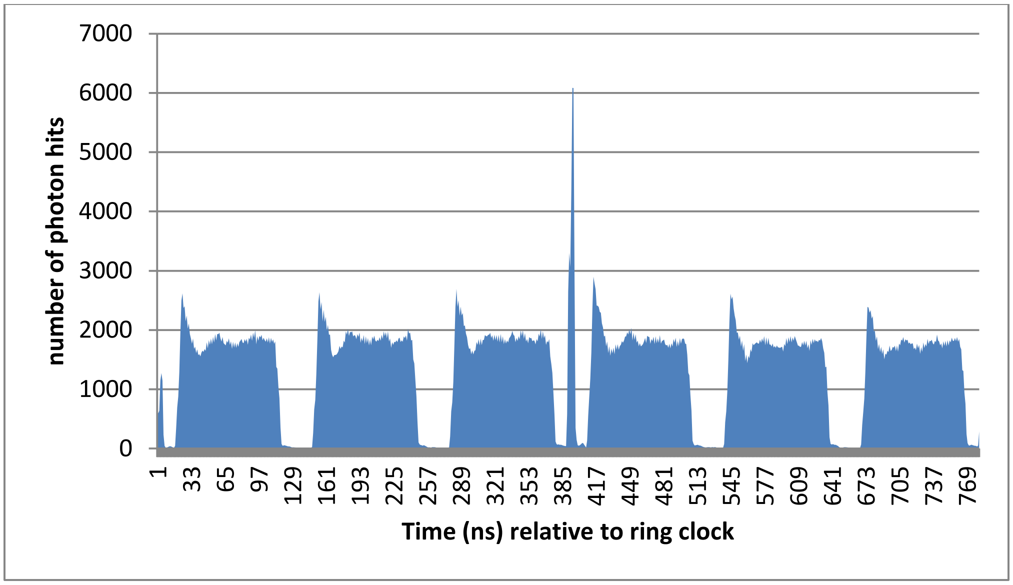


Figure 5: 6 Electron Bunch Train with 1 Mondo Bunch
45 bunches per train with 17 bunch gap
5min Sampling Time
Taken at 4:09pm 07/22/2011

This pattern is what the users see from their beam-line for their experiments. The bunch trains are used as a continuous radiation source. The Mundo bunch is an isolated bunch with high current used for pump-probe experiments.

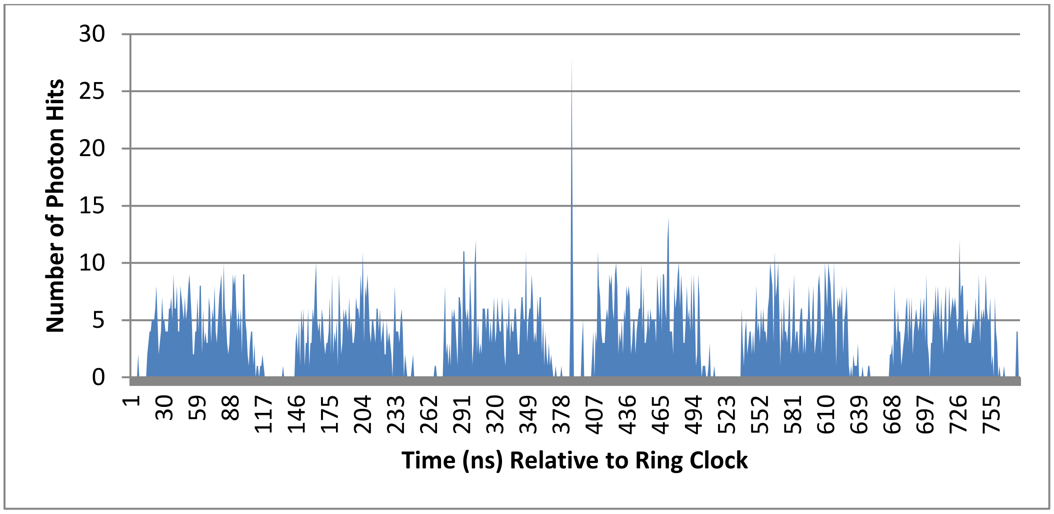


Figure 6: 6 Electron Bunch Train with 1 Mondo Bunch
45 bunches per train with 17 bunch gap
5 sec Sampling Time
Taken at 5:40pm 7/20/2011
Results are fairly noisy due to short sampling time.

The following results were taken during accelerator physics beam time. These bunch patterns are used for SSRL facility diagnostics.

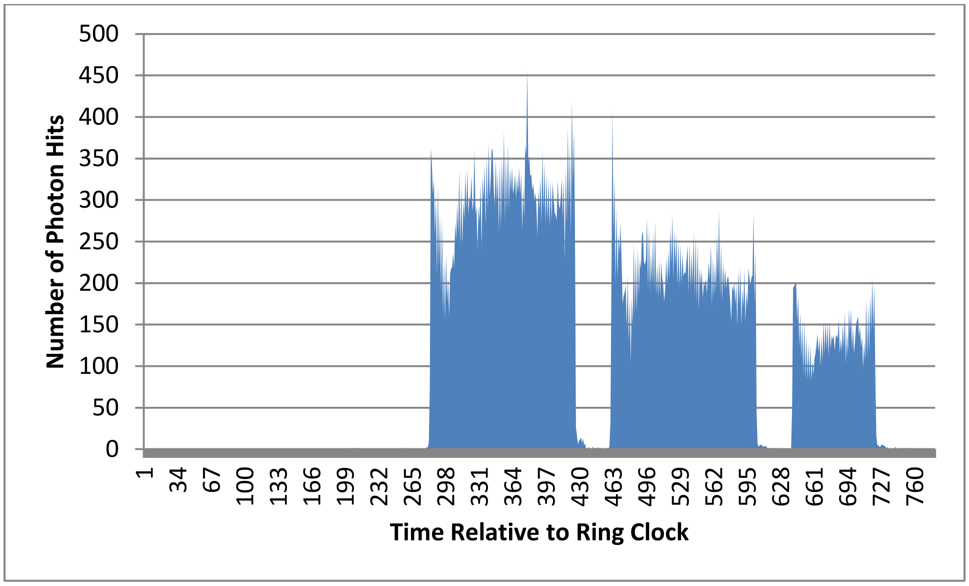


Figure 7: 3 Electron Bunch Train
Taken 8:20pm 07/25/2011

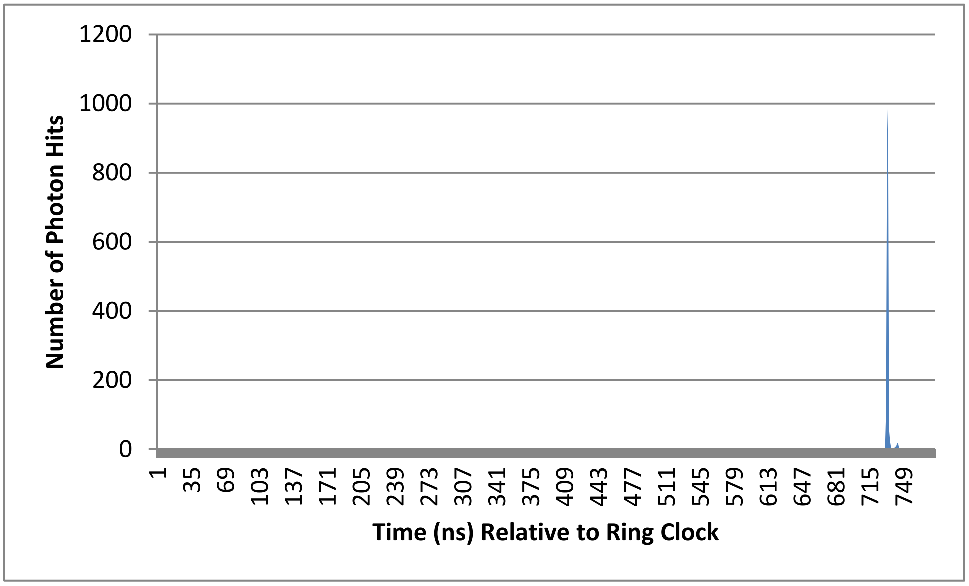


Figure 8: Single electron bunch at 5mA. Roughly 5% of charge is spilled into adjacent bunches. There is a small peak 6 buckets away, but we suspect that is the ringing effect from the PMT signal.
Taken: 5:52am 07/26/2011

Discussion/Conclusions

The results of the single-photon counting measurements revealed much about the electron bunch distribution in SPEAR3. As seen in Figure 5, a short sampling time produces a roughly accurate electron distribution in the ring. However, when compared to Figure 6, the electron bunch pattern is much more accurate as the sampling time increases, as signal to noise ratio increases linearly to the square-root of N (N being the number of samples).

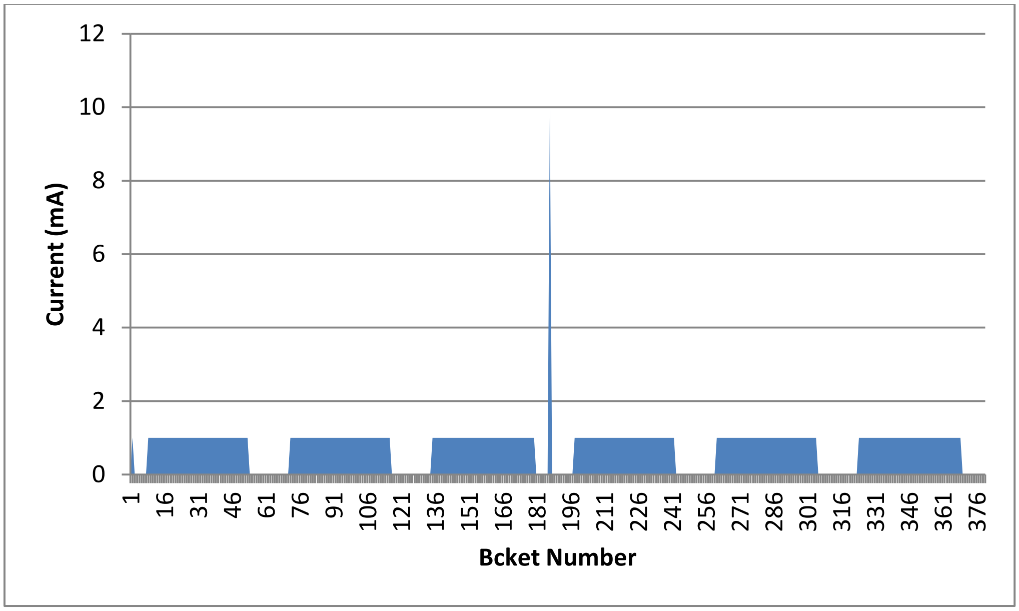


Figure 9: Ideal current distribution inside SSRL Storage ring.

Figure 9 represents the idea electron bunch distribution in the SSRL Storage at the time Figure 5 data is taken. While both distributions have roughly the same shape and timing, there are still many minor, yet important, differences.

The first noticeable difference is the unevenness of the bunch train in the experimental data. Noise, which in this case is ambient light and dark current from the PMT, contributes partially to the oscillation. However, the big spike at the beginning of each bunch train is more than just noise. The current injection system at SPEAR3 injects electron bunches into the assigned bucket without first measuring the current in that bucket. In other words, the system assumes what is left in the bucket and adds a predetermined amount of electrons into the bucket. If the remaining energy inside the bucket differs from assumption, then the total energy inside the bucket will differ from the target current after an injection. As seen in Figure 5, this "blind" injection resulted in more electrons at the beginning buckets of each train, as compared to the rest of the train. This result reveals a flaw in the top-off injection system at SSRL. The remaining current inside each bucket should be measured first, before injecting additional electrons. Such a feedback mechanism can help eliminate this problem.

Another important difference between the ideal and measured bunch patterns is evident in the "spilling" of electrons into adjacent buckets. Although electrons are injected into a bucket, oscillation and collision with other particles may knock an electron into an adjacent bucket. This result can be seen in the experimental results, as unfilled buckets near filled buckets have a small amount of current. This effect can be seen in Figure 8. For this specific bunch pattern, only one bucket is filled. However, in the results, observed electrons in the adjacent 8 buckets. Most of the charge was spilled into the 2 adjacent buckets, with roughly 5% of the Mundo bunch spilled in each bucket. This "spilling" effect is a known phenomenon, but the percentage of spilling must be monitored to ensure good beam quality.

Acknowledgements

Thanks to Jeff Corbett for being a wonderful mentor. He designed the experiment, supervised each step, and was of tremendous help to me. I worked in the same lab as Mitch Miller, and Jonathan Kamp. We collaborated on our projects and they helped me through many parts of my project. Dennis

provided the counting card, which was essential to our experiment. Doug helped me with modifying the Lab View program. He is an experienced programmer and my program would not have worked if it wasn't for him. And thanks to the SULI program and DOE for this wonderful opportunity here at SLAC National Laboratory.

References

[1]Advanced Measurement Technology, Inc. "Ortec Model 9353 100 Picosecond Time Digitizer, Hardware and Software User's Manual"

[2]Hamamatsu Inc. "Photon Counting Using Photomultiplier Tubes".

[3] Michael Wahl, PicoQuant GmbH. "Time-Correlated Single Photon Counting".

[4] A. Jeff* (CERN, Geneva, Switzerland & University of Liverpool, UK), M. Andersen, S. Bozyigit,

E. Bravin, A. Boccardi, T. Lefevre, A. Rabiller, F. Roncarolo (CERN, Geneva, Switzerland);

C.P. Welsch (Cockcroft Institute, Daresbury, UK); A.S. Fisher (SLAC, Menlo Park, USA). "DESIGN FOR A LONGITUDINAL DENSITY MONITOR FOR THE LHC".

[5] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov. "Single-photon sources and detectors"

SLAC-TN-12-029

Optical Design of a Broadband Infrared Spectrometer for Bunch Length Measurement at the Linac Coherent Light Source

Kiel Williams

Office of Science, Science Undergraduate Laboratory Internship (SULI) Program
Guilford College, Greensboro, North Carolina

SLAC National Accelerator Laboratory
Menlo Park, California

20 August, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, U.S. Department of Energy's SULI Program under the direction of Dr. Josef Frisch at the Linac Coherent Light Source (LCLS).

Participant: ______________________________
Signature

Research Advisor: ______________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Table of Contents

ABSTRACT

The electron pulses generated by the Linac Coherent Light Source at the SLAC National Accelerator Laboratory occur on the order of tens of femtoseconds and cannot be directly measured by conventional means. The length of the pulses can instead be reconstructed by measuring the spectrum of optical transition radiation emitted by the electrons as they move toward a conducting foil. Because the emitted radiation occurs in the mid-infrared from 0.6 to 30 microns a novel optical layout is required. Using a helium-neon laser with wavelength 633 nm, a series of gold-coated off-axis parabolic mirrors were positioned to direct a beam through a zinc selenide prism and to a focus at a CCD camera for imaging. Constructing this layout revealed a number of novel techniques for reducing the aberrations introduced into the system by the off—axis parabolic mirrors. The beam had a recorded radius of less than a millimeter at its final focus on the CCD imager. This preliminary setup serves as a model for the spectrometer that will ultimately measure the LCLS electron pulse duration.

INTRODUCTION

The Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory is a free-electron laser (FEL) capable of producing some of the shortest x-ray pulses ever observed. The LCLS produces electron pulses less than ten femtoseconds in duration. These pulses then emit soft and hard x-rays as magnetic fields force the electron bunches to oscillate back and forth [1].

Due to the short duration of the LCLS electron pulses, a precise measurement of their length has proven difficult to obtain [2]. Improved information on the pulse length would enable researchers to better understand the time scales over which the processes they observe are happening. The dissolution of chemical bonds and the folding of different types of protein are known to happen over scales similar to the length of LCLS electron pulses, and more knowledge of the pulse length could be translated into more a knowledge of these fundamental reactions.

An indirect measurement of the electron pulse length can be obtained by observing the optical transition radiation (OTR)[1] given off by the electron pulses as they strike a conducting material [3]. The OTR can then be directed through an optical layout and toward a thermal detector that records the constituent spectrum.

Due to the broadband nature of the spectrum of OTR emitted by the electrons, the internal configuration of this spectrometer differs significantly from that of most modern spectrometers. Most modern spectrometers focus on too narrow a spectral band to appropriately analyze the electrons' full OTR band. Conventional glass optical materials also fail to transmit

[1] The induced mirror charge on the conductor and the approaching electron together behave like a collapsing electric dipole as the electron approaches, with the emitted OTR falling in a wide spectral band primarily within the infrared [4].

over wavelengths comparable to that of the electron bunch OTR (0.5-30 microns). Additionally, typical diffraction gratings fail to examine a wide enough spectral band to sufficiently analyze the OTR. Triangular prisms can decompose light across a wide range of frequencies into its constituent wavelengths, permitting such prisms to act as appropriate substitutes. Figure 1 illustrates the relevant geometry of this type of prism. The total deviation angle of a light ray γ through such prisms is given by

$$n \sin \frac{\gamma + \varepsilon}{2} = n_0 \sin \frac{\varepsilon}{2} \qquad (1)$$

where n and n_0 are the indices of refraction for the outside medium and the inside of the prism respectively, and ε is the apex angle of the prism (aberrational effects through the prism are minimized for this configuration). Furthermore, when the prism is in this aberration[2] minimizing configuration it can be shown that the relation

$$\cos \alpha = n_1 \sin \frac{\gamma + \varepsilon}{2}, \qquad (2)$$

where α is angle formed by the incident light ray and the surface of the prism, also holds.

Expanding beam optics can be used as a means of bringing a beam parallel to a chosen axis. One concave and one convex lens can accomplish this in a simple case. A diagram of this setup is illustrated in Figure 2. The magnitude of magnification *M* obtained by such a setup is given by the relation

$$M = \frac{f_2}{f_1}, \qquad (3)$$

where f_2 is the focal length of the convex lens, and f_1 is the focal length of the concave lens.

[2] In this paper an aberration is taken to be any optical effect that degrades the circularly-symmetric Gaussian profile of the beam.

For a Gaussian mode TEM_{00} beam, analysis of the analytic expression for the beam profile admits numerical formulae for the half-width and $1/e^2$ width of the beam in terms of one another. Frequently the half-width can be measured with greater ease and converted to the $1/e^2$ width based on the formula

$$\omega_0 = \frac{M_H}{\sqrt{2\ln 2}} \simeq 0.84932\ M_H \quad , \qquad (4)$$

where ω_0 is the $1/e^2$ radius of the beam and M_H is the full-width of the beam at half of its maximum power.

MATERIALS AND METHODS

Preliminary construction of the optical layout utilized a 632.8 nanometer wavelength helium-neon mode TEM_{00} laser of minimum peak power 0.8 mW and a $1/e^2$ diameter of 0.48 millimeters. A series of reflecting mirrors and lenses were inserted to expand and direct the beam. A plano-concave and plano-convex lens were placed the difference of their focal lengths apart to enable their operation as a Galilean beam expander. The projected final beam diameter was calculated according to Equation 3, and the final diameter of the beam was measured by ruler as it emerged from the convex lens onto a portable screen. Gold-coated plane mirrors then directed the beam towards a collection of gold-coated off-axis parabolic (OAP) mirrors, with irises aligning the beam along each portion of its path. The first two OAP mirrors were separated by the sum of their focal lengths, allowing the beam to further expand as it passed from the first to the second. A zinc selenide (ZnSe) triangular prism of apex angle 10 degrees was placed approximately 50 mm in front of the second OAP mirror. A prism made of thalliumbromoiodide will be used in the final spectrometeter, but due to its toxic properties and great fragility ZnSe was used for this preliminary setup.

A third gold-coated OAP mirror was placed 71 mm in front of the prism. Equation 1 dictated based on the prism apex angle that aberrational effects would be minimized for a total deviation angle of approximately 16.9 degrees. The third OAP mirror was then positioned 21 mm to the side of the prism in such a way that it would intercept the incoming beam at this optimal angle. The orientation of the ZnSe prism was then adjusted until the beam reflected off of the positioned third OAP mirror. Equation 2 dictates a backscatter angle α of 76.9 degrees for optimal aberration reduction. Fine adjustments to the tilt of the ZnSe prism were performed until a backscatter angle closely echoing this theoretical optimum was measured. The backscatter angle was measured by blocking the backscattered beam with a screen, measuring the distance from the prism to the screen, and comparing this distance with the length between the second OAP mirror and the backscatter-blocking screen. The backscatter angle could then be extracted from the situational geometry. The third OAP was also adjusted to reduce aberrational effects as it intercepted and reflected the beam.

The focal point of the third OAP mirror was located as the laser reflected from it using a screen and an adjustable slit capable of accurately recording widths on the order of the hundreds of microns was positioned there. A power meter was used to detect the drop-off in power of the beam as the slit was closed, recording the width at which the power dropped to half its initial peak. This width was converted to the $1/e^2$ diameter of the beam using Equation 4. The slit was then repositioned to the shared focal point between the first and second OAP mirrors. Two adjustable slides were attached to the base of the slit and it was moved to the focal point of the first OAP mirror by sliding it alternatively towards and away from the OAP mirror. The point at which sliding the slit perpendicular to the beam reversed the direction the image moved on a mounted screen was scanned for as a means of precisely locating the focal point of the first OAP mirror. Each time the beam image switched the direction it moved relative to the horizontal slide of the stage it was revealed that the focal point had been passed. This process could then be iterated to bring the slit closer to the focal point. Following these first-order positioning corrections the slit size was adjusted and it was moved in a way that maximized the amount of passing light. The horizontal and vertical knobs on the first OAP mirror before the slit were also adjusted as a means of eliminating beam aberrations introduced by this mirror into the optical system. A measurement of the beam waist using the same power-based method as for the prior focal point after the third OAP mirror was taken and compared with simulated data [6]. A measurement of the beam waist at its origin from the laser was also performed in this manner.

A CCD camera fitted with a 30 dB circular attenuator was then placed at the final focal point to profile the beam at the end of the optical layout and connected to a nearby computer terminal. Placing the CCD camera so that its surface formed an angle of approximately 44 degrees with the incoming beam allowed for further beam aberration minimization based on the results of previous optical simulations**.** Figure 3 illustrates this configuration of optical components. Table 1 records the focal length of each lens and OAP mirror.

Precision alignments of the OAP mirrors following their initial placement were performed by first rotating them 180 degrees and observing that the beam profile fell on the center bolt on the rear of each mirror. It was verified that each OAP was at an appropriate height by doing this in turn to each of the three OAP mirrors in the system. The horizontal and vertical fine adjustment knobs on the OAP mirrors then enabled the beam to be gradually directed through the last half of the system and to the center of the CCD camera.

Positioning a small screen immediately after each of the three OAP mirrors then allowed an assessment of the extent to which they introduced damaging aberrations into the beam profile. The methods used to reduce the observed aberrations in the OAP mirrors included small rotations along the optical axis of the mirrors and minor height adjustments to better center the incoming beam, and turning the OAP mirrors while they rested in their holders to alter the angle at which they deflected the incoming beam. Adjustment of this tilt enabled the beam to maintain its six-inch height in addition to the correction of aberrational effects. Adjusting the micrometer on the slit after the first OAP mirror and observing the laser profile on a screen also guided aberrational corrections. Figure 4 illustrates the final layout of the aligned optical components.

A set of images of the beam was collected immediately after concave lens at the beginning of the optical layout with the CCD camera to ensure its linearity. Images of the beam at different widths as it grew larger through the beam expander were collected by shifting the position of the camera. These images were then loaded into Matlab^{TM} and the intensities across a center cut of the beam were summed together. By comparing this summation for the beam at different widths, the degree to which nonlinear digital effects applied by the CCD camera impacted images of the beam and its accompanying intensity at points throughout the optical system was determined.

RESULTS

At its origin the beam $1/e^2$ diameter was measured as .28±.02 mm using the slit micrometer. Following the passage of the beam through the Galilean expander it attained a diameter of 6±1 mm based on an approximate ruler measurement (the size of the beam at this point in the system made a micrometer measurement impractical). This measurement is in line with that predicted by Equation 3. Additionally, CCD camera images taken at the conclusion of the optical layout reveal the beam to have an approximately Gaussian structure when the slit was left fully open. Summing the Matlab^{TM} intensity indices of different cross-sections of the beam taken with the CCD camera revealed an image-to-image variation of less than five percent. Figure 5 shows an image of the beam at this focal point of the third OAP mirror under 50 dB of total attenuation (the additional 20 dB of attenuation resulted from a second attenuator placed immediately in front of the CCD camera).

In the final optical configuration a total deviation angle of 15±1 degrees through the prism was measured. Measure of the backscatter angle after adjustment revealed a backscatter angle of 73 degrees to within an error of less than five degrees. This falls within five degrees of the theoretical optimum angle given by Equation 2 to reduce aberrational effects. Small variations in this angle were not observed to have a major impact on the reduction of aberrational effects.

A beam profile diameter of 173±5 microns was recorded at the focal point of the first OAP mirror using the slit micrometer. Using the slit to measure the beam profile at its focus after the third OAP mirror yielded a measurement of approximately 237±5 microns for the $1/e^2$ beam diameter at this point in the optical layout. Performing fine adjustments to the orientation of each OAP mirror and positioning the slit carefully between the first and second OAP mirror eliminated many aberrations throughout the system. Initially, the beam appeared to diminish from the vertical direction as it passed through the horizontal slit following the first OAP mirror. By empirically adjusting the horizontal and vertical tilt of the first OAP mirror this aberration has been mitigated. The beam continues to focus in two different planes as it approaches the focal point of the third OAP mirror.

DISCUSSION

Significant reductions in the severity of optical aberrations throughout the final system proved attainable through hand adjustment of the constituent optics. Following the initial placement of the optical equipment the beam focused horizontally and vertically in different planes in the space following the third OAP mirror. Rotating the third OAP mirror in a way such that the beam fell precisely at its center helped to reduce the problem somewhat. Rotating the tilt of the angle at which the second OAP mirror intercepted the incoming beam proved helpful as well. Errors in the angle of the tilt of OAP mirrors about their optical axes of less than one degree can greatly distort the final image by bending it into an arc-like pattern. Adjustment of this tilt itself proved one of the more effective means of determining the correct tilt angle for the OAP mirrors due to the observable aberrations introduced by even small initial errors.

Simulations predict a beam diameter of 65 microns at the focal point of the first OAP mirror, and a required tilt of 0.4 degrees of the first OAP mirror to attain the measured $1/e^2$ diameter of 173±5 microns. Similar simulations project a spot size of 38 microns at the focal point of the third OAP mirror, with a required tilt of approximately 1 degree to attain the measured $1/e^2$ diameter of 237±5 microns. This is less consistent with expected human alignment errors, and suggests deeper aberrational effects occurring in the second and third OAP mirror, and potentially within the prism as well.

One aberration of particular note is that which occurs immediately after the closure of the slit in which the beam would appear horizontally oriented despite the vertical orientation of the slit. Noting that this aberration coupled the horizontal and vertical propagation of the beam, the horizontal and vertical tilt of the first OAP mirror were gradually changed until the beam was oriented vertically as it exited the slit. Because altering the vertical and horizontal mirror tilt iteratively proved effective in eliminating the problem, it is suggested that the way the OAP mirror couples the propagation of the beam in the horizontal and vertical direction causes the beam to twist in some manner as it passes through the slit. Noting patterns like these will become a progressively more important means of correcting aberrations in later versions of the spectrometer as lasers in the near- and mid-infrared are incorporated into the design.

Future modifications to this preliminary layout will incorporate a pyroelectric thermal line detector that will ultimately be used to capture the spectrum of the LCLS beam [7]. The orientation of the beam as it reflects from the third OAP mirror will require a greater level of attention due to the sensitivity of the detector. In conjunction with this, additional lasers emitting in the near- and mid-IR spectrum will be added to enable calibration of the sensor and test how the optical layout transmits signals through portions of the IR spectrum.

ACKNOWLEDGEMENTS

This work received support from the SLAC National Accelerator Laboratory and Department of Energy, Office of Science. I would like to thank my mentor Josef Frisch for the assistance he provided throughout this project. I also wish to thank Alan Fisher, Mark Petree, George Burgueno, and my research partners Julie Cass and Gilles Dongmo-Momo for their invaluable assistance.

REFERENCES

[1] Z. Huang, A. Baker, C. Behrens, M. Boyes, J. Craft, F.-J. Decker, Y. Ding, P. Emma, J. Frisch, R. Iverson, J. Lipari, H. Loos, D. Walz, "Measurement of Femtosecond LCLS Bunches Using the SLAC A-Line Spectrometer." Paper presented at the 2011 Particle Accelerator Conference, New York, New York, March 28-April 1, 2011.

[2] M. Zelazny, S. Allison, S. Chevtsov, P. Emma, K.D. Kotturi, H. Loos, S. Peng, D. Rogind, and T. Straumann, "Electron Bunch Length Measurement For LCLS at SLAC." Paper presented at the 2007 International Conference on Accelerator and Large Experimental Physics Control Systems, Knoxville, Tennessee, October 15-19, 2007.

[3] J. Frisch, R. Akre, F.-J. Decker, Y. Ding, D. Dowell, P. Emma, S. Gilevich, G. Hays, P. Hering, Z. Huang, R. Iverson, R. Johnson, C. Limborg-Deprey, H. Loos, E. Medvedko, A. Miahnahri, H.-D. Nuhn, D. Ratner, S. Smith, J. Turner, J. Welch, W. White, J. Wu, "Beam Measurements at LCLS." Paper presented at the Beam Instrumentation Workshop, Lake Tahoe, California, May 4-8, 2008.

[4] C. Behrens, A. Fisher, J. Frisch, S. Gilevich, H. Loos, J. Loos, "Design of a Single-Shot Prism Spectrometer in the Near- and Mid-Infrared Wavelength Range for Ultra-Short Bunch Length Diagnostics." Paper presented at the 10th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Hamburg, Germany, May 16-18, 2011.

[5] C. Behrens, "Design of an Infrared Prism Spectrometer for Ultra-Short Bunch Length Diagnostics." Presentation at a meeting of the SLAC National Accelerator Laboratory Accelerator Physics and Engineering Department, Menlo Park, California, November 30, 2010.

[6] J. Cass, "Simulations and Analysis of an Infrared Prism Spectrometer for Ultra-Short Bunch Length Diagnostics at the Linac Coherent Light Source." Produced for the Department of Energy Student Undergraduate Laboratory Internship program at the SLAC National Accelerator Laboratory, Menlo Park, California, 2011.

[7] G. Dongmo-Momo, "Infrared Spectroscope for Electron Bunch-length Measurement: Heat Sensor Parameters Analysis." Produced for the Department of Energy Student Undergraduate Laboratory Internship program at the SLAC National Accelerator Laboratory, Menlo Park, California, 2011.

TABLES AND FIGURES

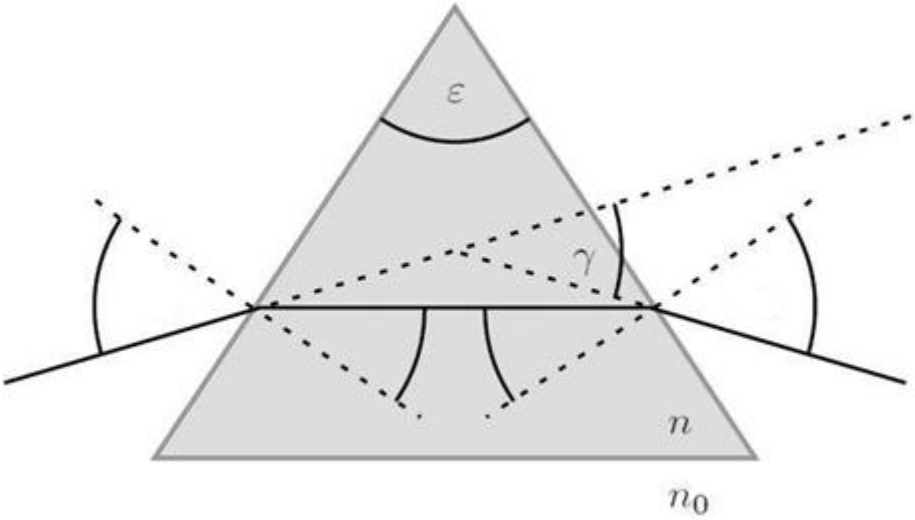


Figure 1: Illustration of a prism in an aberration-minimizing symmetric configuration. Note that all incoming and outgoing angles are equal to their symmetric partners. Courtesy C. Behrens.

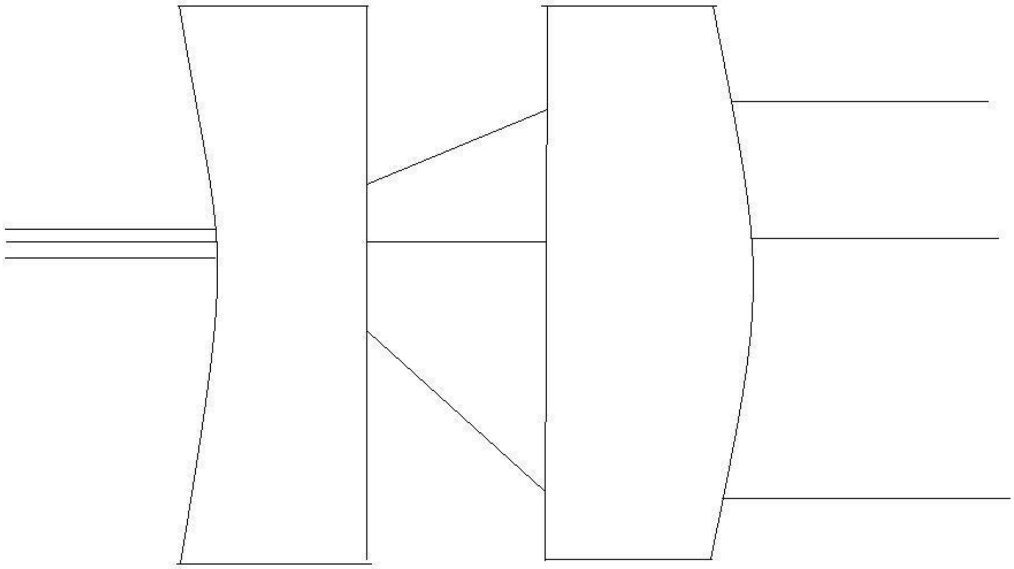


Figure 2: Illustration of the lens configuration that expanded the beam in this optical layout. Note that the lenses are separated by the difference of their focal lengths.

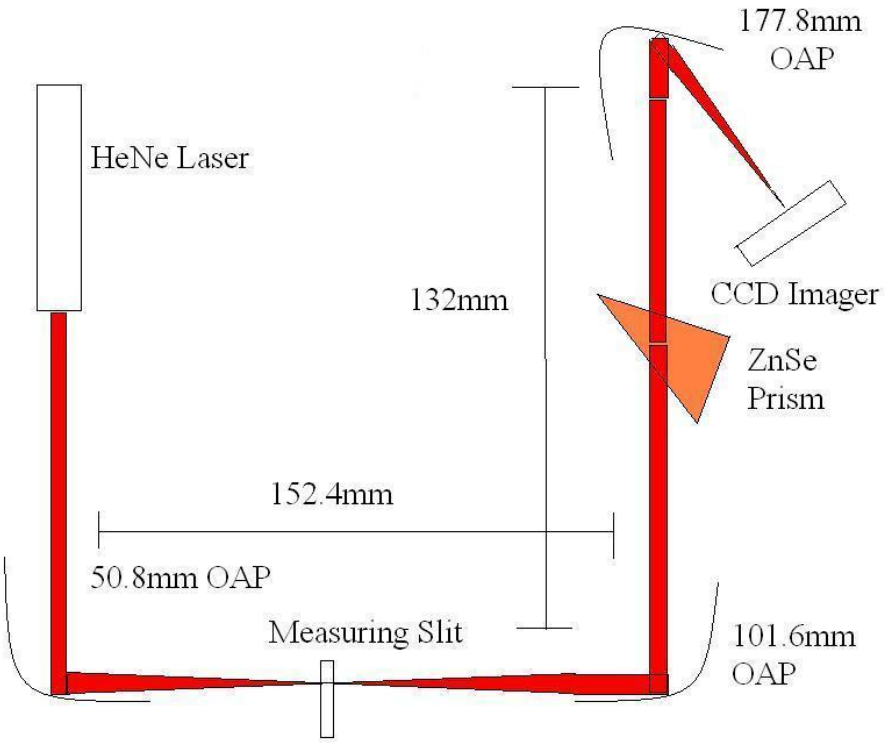


Figure 3: Illustration of the optical layout directing the helium-neon beam to a final focus. Note that this is one of several placements used for the measuring slit. The lens were placed before the first OAP mirror and are not shown.

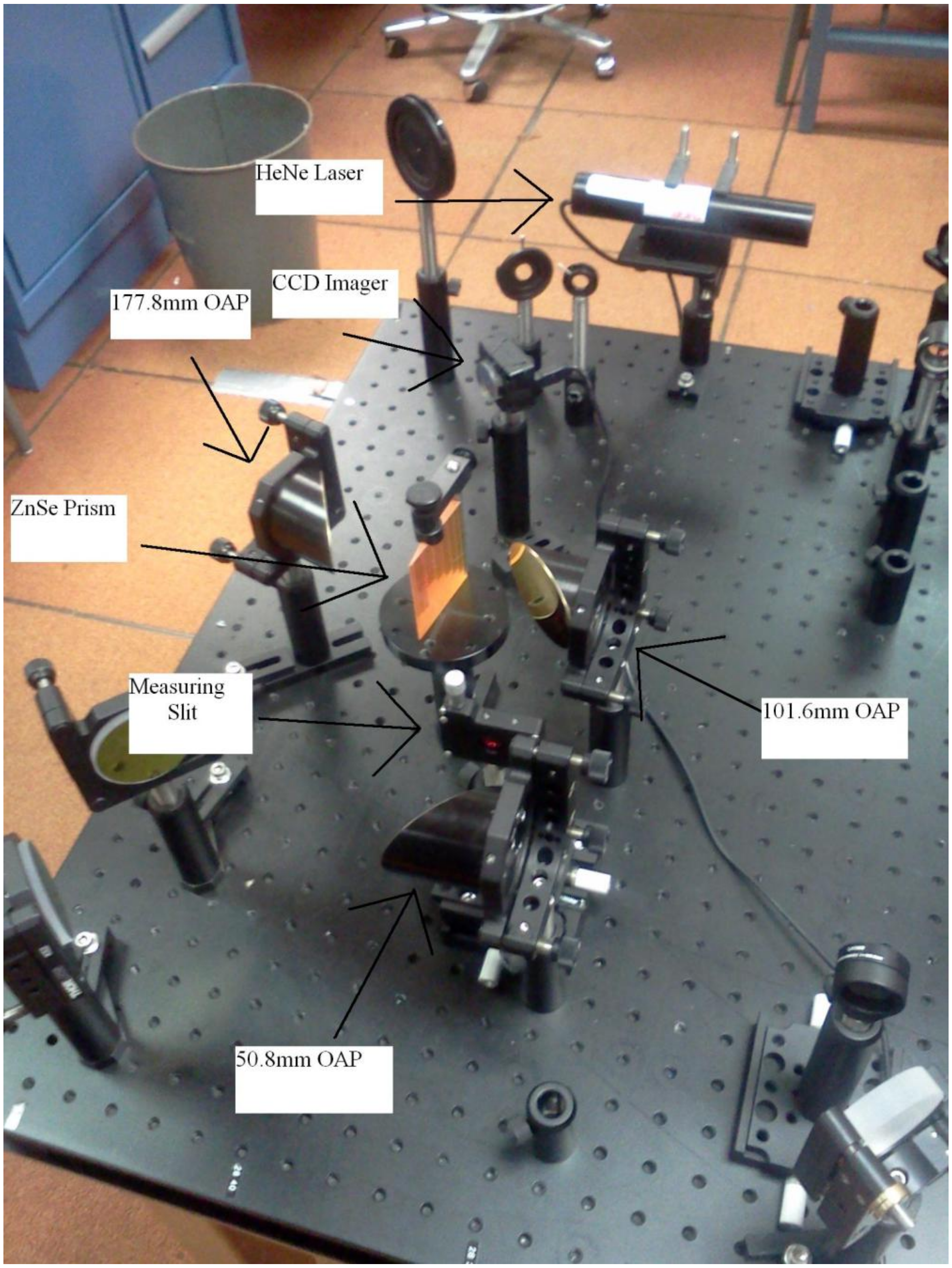


Figure 4: Final layout of aligned optical components. The ZnSe prism and gold-coated OAP mirrors are all visible.

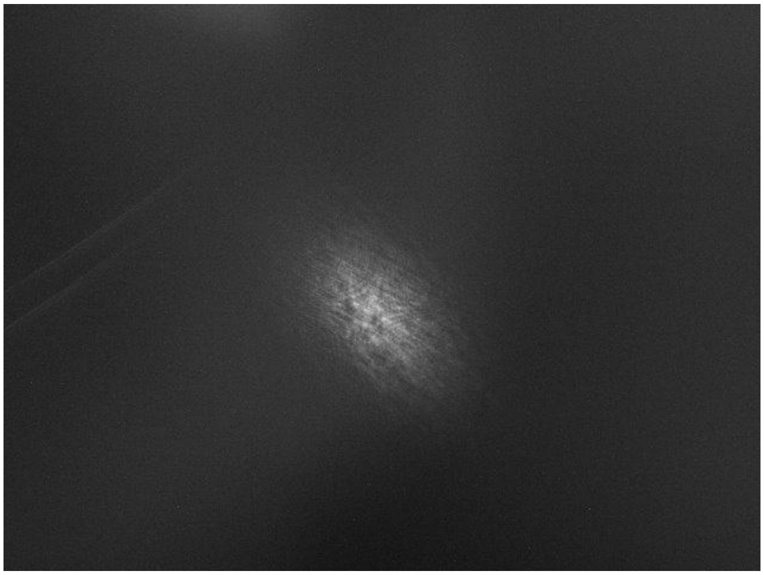


Figure 5: CCD camera image of the beam under 50 ODs of attenuation taken at the focal point of the third OAP mirror. Note its approximately Gaussian composition and the absence of severe aberrational effects.

Table 1: Focal Lengths of Relevant Optical Equipment

Focal Lengths	
Optical Equipment	Focal Length ***mm***
Plano-Concave Lens	-30
Plano-Convex Lens	500
OAP 1	50.8
OAP 2	151.6
OAP 3	177.8