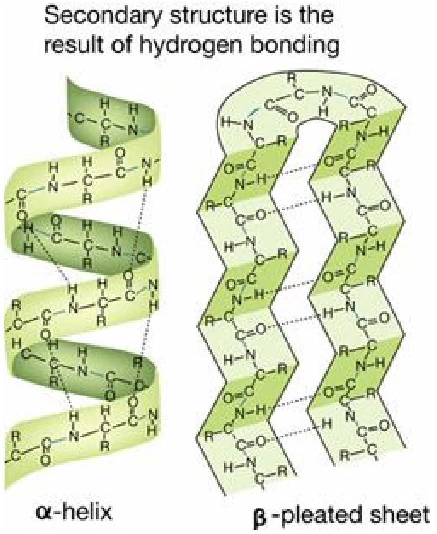
Protein Structure Determination: Zika Virus (Group 2)

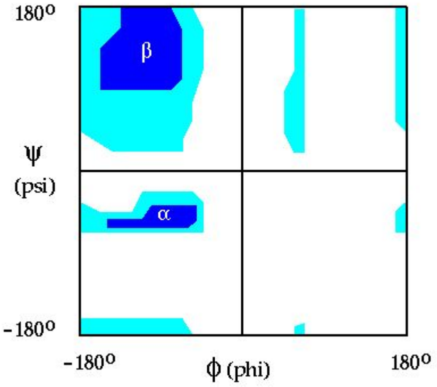
Background:

Protein structure determination is an important process and active area of research in biophysics and, more broadly, biology. During their production, proteins engage in a process called “folding” whereby they coil and fold into their final, three dimensional shape. In cells, protein folding is a complex, thermodynamically driven process that is dependent only upon the initial sequence of amino acids that compose the molecule (this is known as the “primary structure” of a protein). Primary structure gives way to more complex folding patterns, aided by internal hydrogen bonding, such as alpha helices, beta pleated sheets, and various loops and turns. This folding pattern yields another order of structure that is the protein’s three dimensional structure. Further interactions with other polypeptides (amino acid chains) yields a multi-subunit complex that possesses a quaternary structure. Ultimately, protein structure comprises many different levels of internal organization that contribute to the complexity of these molecular machines.

Alpha helices and beta pleated sheets can be thought of as structural building blocks. They form due to favorable energetics supplied by maximizing hydrogen bonding between various amino acids of the protein’s primary sequence. This can be demonstrated visually through the use of a Ramachandran plot, which shows stability as a function of the possible angles for backbone elements. Alpha helices and beta sheets also help minimize unfavorable hydrophobic interactions with solvents by “hiding” hydrophobic sides within the rest of the protein.



Knowledge of a protein’s overall 3D shape can yield clues on its binding patterns and targets and mechanism of function. Protein folding is a complex phenomenon *in vivo* that is not well understood, motivating a need to develop methods of discerning structures post-folding. Protein structure is largely determined through two approaches: X-ray crystallography and a more recent approach called cryo-electron microscopy.



Compared to X-ray crystallography, cryo-electron microscopy (cryo-EM) involves far less effort in the experimental setup due to the lack of need for crystallized samples. Protein crystallization is an arduous process that, in some cases, has taken decades to perfect for individual protein types. As a result, cryo-EM has begun to take a center stage in protein structure determination and we hence focus on it in this paper.

In our project, we outline the process of cryo-EM and then perform an image reconstruction procedure on a simulated dataset from Zika virus, yielding a 3D model of a protein, through two separate approaches.

The accuracy in reconstruction of both techniques are compared along with computational complexity. A discussion of the Dirac delta function and its application in cryo-EM image reconstruction as well as an overview of Zika virus' biology are included.

Experimental Setup:

For cryo-EM, electrons are used to study the structure of small molecules that cannot be examined with (visible) light microscopy. Electrons confer an advantage due to their small wavelength (electrons exhibit a particle-wave duality), a property that correlates with resolution.

The protein of interest is frozen in a block of ice in order to reduce radiation damage. The sample is then subject to a beam of electrons that scatter at various angles. A detector measures the scattered rays from various angles, allowing for the reconstruction of the protein's conformation through several mathematical procedures.

Since radiation damage has considerable limitations on the electron dose one can use for imaging, the signal-to-noise ratio in these images is low. Consequently, many images must be taken and then averaged in order to reduce the impact of noise on image quality. Even with new detectors, tens of thousands of images are needed to obtain a 3D reconstruction of the scattering potential at near-atomic resolution.

Modeling: Basic Assumptions

To introduce some formal notation:

Let $\varrho : R^3 \rightarrow R$ *model the electrical potential of the molecule at a given point.*
Let $F = (a\ b\ c)$ *be the coordinate frame of image* I, *where* $I = span(a, b)$ *and* c *the viewing direction.*
Furthermore, let $\{a, b, c\}$ *be an orthonormal set.*

While the cryo-EM process is overall quite effective, there are several assumptions that should be noted.

The physical model for cryo-EM uses classical physics and does not take into account quantum mechanical phenomena. However, it has been shown that classical descriptions perform very well in comparison to observed data.

The reconstruction procedure relies on many different image projections at various orientations with respect to the protein. While we assume that all image planes I_F are equally covered, in reality some orientations are more covered than others since proteins will not settle completely randomly in the ice but instead place their centers of mass toward the ground. This in turn results in varied structural resolution at different angles in experimental data.

It is also assumed that all copies of the protein are identical within the ice sample. This allows us to model images of different copies as images from the same protein but simply oriented differently. Depending upon the protein of interest and the extraction procedure performed, this assumption varies in its appropriateness. For instance, structural proteins are often nearly identical conformationally whereas enzymes may possess multiple native states that they transition between during biochemical functions. In the case of the latter, steps may be taken to coax a sample into a desired state if necessary. Since we are using a model of the Zika virus capsid and membrane proteins (the protein shell that houses viral genetic information), it is relatively safe to assume that the sample is composed of identical virions.

Image Simulation: Radon Projections

Our image simulation was performed with two different processes: Radon projections and Fourier transforms.

Under the Radon projection technique, we model each pixel in our image as having the value of the line integral along the protein deposited there. Electron paths that have higher total electrical potentials have more matter obstructing the particle's path. This in turn results in different exposure levels for pixels on the detector.

Using the notation developed earlier, we obtain the following line integral describing a Radon projection for a given pixel in *I* with coordinate frame *F*:

$$I_F(x, y, z) = \int_{-\infty}^{\infty} \varrho(ax + by + cz)\, dz \qquad *\textit{ via change of coordinates from standard basis to } \{a, b, c\}$$

For an *n* x *n* image, one must perform a Radon projection n^2 times in order to cover the entire image. Multiplying by the computational complexity of a line integral *I*, we find that the computational complexity of simulating an image via Radon projections is O(*I* * n^2).

Image Simulation: Fourier Slice Theorem

In order to provide a conceptual foundation for the following method, we now introduce the Fourier transformation.

The Fourier transformation expresses functions in terms of the exponential basis $\{\exp_1,...,\exp_N\}$, where $\exp_i$ is:

$$exp_i = exp(-i\tfrac{2\Pi}{T}nx)$$

Using Euler's formula, this change of basis can also be thought of as a decomposition of a function into a sum of sinusoidal functions. Since the exponential basis is in fact an orthogonal basis, then we can represent a function f(x) as a sum of inner products with respect to basis elements multiplied by those basis elements (first converting our basis to an orthonormal one by normalization).. In other words:

$$v \varepsilon V,\ \beta \subset V$$

$$v \ = \ < v, \beta_1 > \beta_1 + . . . + \ < v, \beta_i > \beta_i$$

For the space of periodic functions (with period T), we define the inner product as:

$$< f(x),\ g(x) > \ = \ \int_{-T/2}^{T/2} f(x) * \overline{g(x)} dx$$

So we can express the inner product of f(x) and a member of the exponential basis as:

$$< f(x),\ exp(-\, i\tfrac{2\Pi}{T}nx) > \ = \ < f(x),\ cos(nx\tfrac{2\Pi}{T}) \ + \ sin(nx\tfrac{2\Pi}{T}) >$$
$$\Leftrightarrow \ < f(x),\ cos(nx\tfrac{2\Pi}{T}) > \ + \ < f(x),\ \ sin(nx\tfrac{2\Pi}{T}) >$$
$$\Leftrightarrow \int_{-T/2}^{T/2} f(x) * cos(nx\tfrac{2\Pi}{T}) \ dx \ + \int_{-T/2}^{T/2} f(x) * sin(nx\tfrac{2\Pi}{T}) \ dx$$

The values obtained are the a_n and b_n values specifically. Splitting f(x) into component even and odd functions (this can be done for any function) and taking into account the fact that integrals of compositions of even and odd functions yields zero over their period, we can define f(x) in terms of two separate orthonormal bases. Substituting these back into the equation for v in terms of the given orthonormal basis (and then splitting into the separate cases for the odd and even function components) we obtain:

$$f(x) \ = \ a_0 + \sum_{n=1}^{\infty} a_n\tfrac{2}{T}cos(nx\tfrac{2\Pi}{T}) \ + \ \sum_{n=1}^{\infty} b_n\tfrac{2}{T}sin(nx\tfrac{2\Pi}{T}) \quad where\ a_0\ is\ simply\ (f(x),\ \tfrac{1}{T} >$$

The expression of f(x) in terms of this Fourier basis is the *Fourier series* of f(x). Using these coefficient values, we can also obtain the original function f(x) by performing an *inverse* Fourier transform. The Fourier transform and its inverse operation are important mathematical tools deployed in a variety of tasks, from EEG analysis to image compression to heat transfer modeling.

In order to compute ϱ from a series of 2D images, we implemented a method of 3D image reconstruction using Fourier slice theorem. The theorem states that the Fourier transform of a N-dimensional space restricted to a N-1 dimensional subspace is equivalent to the Fourier transform of that N-1 dimensional space. Applied to cryo-EM image simulation, the theorem states that 3D Fourier transform of ϱ restricted to a viewing plane I is equivalent to the Fourier transform of I. In a more neatly put mathematical notation:

$$FT(\varrho)|_{\,I} \ = \ FT(I) \qquad where\ I\ is\ the\ image\ of\ \varrho\ \ with\ any\ given\ coordinate\ frame$$

Since the Fourier transformation is an invertible procedure, we can apply the inverse Fourier transform in order to recover the molecule ϱ from FT(ϱ). Hence a clear procedure emerges: compute the Fourier transform of I (acquired through Radon projections) for multiple coordinate frames, embed the Fourier slices in proper orientations relative to each other, and then perform the inverse Fourier transformation in order to recover the original electrical potential of the molecule (as guaranteed by Fourier slice theorem).

Using the fast Fourier transform (FFT) for the above procedure, the algorithmic complexity of converting between the Fourier basis and the space of "delta functions" (discussed later) is O(n * log(n)).

3D Image Reconstruction:

Now that we can obtain a set $\{FT(I_1)...FT(I_N)\}$ of two-dimensional Fourier slices, the task of creating a 3D model becomes a primary concern. This process of 2D → 3D image reconstruction is done through the backprojection algorithm. Given a set of images and their associated orientations, the algorithm returns an approximation of the original molecule, ϱ.

Since EM images do not carry any information on their associated viewing direction, we must find a method to generate the orientations $\{F_1,...,F_N\}$ necessary for backprojection. To this end we use the common lines approach, otherwise known as *angular reconstitution*, to embed these Fourier slices in three dimensional space.

The intuitive reasoning behind this approach is that all Fourier slices are (ideally) cuts of the same Fourier structure from various directions. As a result, the planes will share "common lines," containing identical values along their spans, since the Fourier slices intersect each other. For a concrete example, imagine the intersection of two planes in a three dimensional space. The line formed by the intersection of the two planes contains points located in both planes.

$Let\ l_{12}\ be\ the\ common\ line\ in\ plane\ 1\ that\ lies\ along\ the\ intersection\ with\ plane\ 2.$
$Let\ l_{21}\ be\ defined\ similarly.$
$Then:$

$$<\ l_{12},\ l_{21}\ >\ is\ an\ absolute\ max\ for\ all\ possible\ lines\ in\ planes\ 1\ and\ 2$$

Given the set of pairwise common lines for all slices, we can find an embedding for the slices such that common lines between any two planes are correctly aligned. This process restricts the orientation of images relative to each other to a single possible outcome that in turn allows us to measure angles between the slice planes. Establishing one plane as the XY axis, we can then obtain the viewing direction F_i for a given Fourier slice FT(I_i) relative to the axis using the derived angles between planes.

Now that the associated viewing directions for a 2D Fourier slice set can be obtained, we turn our attention to the backprojection algorithm.

The basic premise of backprojection is the notion that, if a given point has been projected onto a series of images, then backtracking along that projection should yield the possible locations for the original point. Given a set of images, the original object location will have a higher weighting than other regions as it will be the point where lines from every image, caused by "smearing" the images, intersect.

The backprojection of a given image I_F is:

$b(a,\ b,\ c)\ =\ I(a,\ b)\circ rect(c) \qquad where\ rect(c)\ =\ \{1\ if\ -D \le c \le D,\ 0\ else\}$

D is the bounding box for our function ϱ representing the molecule. The effect of multiplication by *rect(c)* is essentially a "stretching" of I_F into the *c* dimension from -D to D. This is done for all Fourier transforms of the image set and the results are then summed together. A filter, denoted *H(x, y, z)*, is then applied. The result is an approximation of FT(ϱ) (by Fourier slice theorem) so we must apply an inverse Fourier transform to recover the molecule. In other words:

$$\varrho(x,\ y,\ z)\ \approx FT^{-1}((\sum_{i=1}^{N} b_i(ax,\ by,\ cz)) \times \frac{1}{H(x,y,z)})$$

In order to reduce computational complexity, we can apply the convolution theorem to this equation in order to solve for the b_i by multiplying the Fourier transforms of the convolved functions. Specifically, the theorem states:

$FT(f\circ g)\ =\ FT(f)\ \times FT(g)$

Hence:

$b_i = FT^{-1}\{FT(I(a,\ b)\ \times FT(rect(c))\}$

Dirac Delta Function:

Dirac function, or $\delta(x)$, is a generalized function introduced by the famous physicist Paul Dirac. It is defined in R^n and remains 0 for all of the domain except for $\delta(0) = \infty$. Specifically, the integral of the Dirac delta function in R^n is one.

As one can see, the Dirac delta function is not a function under the normal definition. If we calculate its Riemann integral in R^n, the result should be zero. Actually, it is a typical example of what we call "generalized function", which is strictly defined below:

Definition: For a Cauchy function sequence $\{f_n\} \subset C^\infty(a,b)$, if for any $\phi \in C^\infty(a,b)$ limit $lim_{n\to\infty}\int_a^b f_n(x)\phi(x)dx$ always exists, then we have a generalized function *f* or $f = \{f_n\}$ such that $<f,\phi>= lim_{n\to\infty}\int_a^b f_n(x)\phi(x)dx$. The set that consists of such generalized functions is denoted as $D'(a,b)$.

We can easily learn from history that the Dirac function $\delta(x)$ was first expressed as the limitation of a function that describes rectangular pulse when $\varepsilon \to 0^+$: $Q_\varepsilon(x) = 0, |x| > \varepsilon$, $Q_\varepsilon(x) = \frac{1}{2\varepsilon}, |x| \le \varepsilon$.

Engineers use it to describe unit momentary pulse and decided it should have the following properties:

$\delta(x) = 0, x \neq 0$, $\delta(x) = \infty, x = 0$, $\int_{-\infty}^{\infty} \delta(x)dx = 1$.

Taking the last property more universally: $\int_{-\infty}^{\infty} \delta(x)\phi(x)dx = \phi(0)$ *for any continuous function* $\phi(x)$.

In fact, these properties are all easy to prove with our definition of generalized function and $\delta(x) = \{Q_{\frac{1}{n}}(x)\}$.

The Dirac delta function is useful with its last property. For any continuous function $f(x)$, we can pick its value at a random point $x = x_0$ by $f(x_0) = \int_{-\infty}^{\infty} f(x)\delta(x_0 - x)dx$. In other words, we restrict $f(x)$ at x_0 with the Dirac function. Things are similar for the continuous function in the high-dimensional space: $f(x_1 x_2 ... x_n) = \int_{R^n} f(x)\delta((x_1 x_2 ... x_n) - x)dx$.

Moreover, if we want to restrict a function on a certain line, curve or plane, we can also make it with the Dirac delta function. The trick is to form the integral:

$$\int_{R^n} f(x)\delta(p(x))dx$$

Where $p(x)$, a function on variable x, equals zero if only if x is on the certain line, curve, or plane we want to restrict the function to.

In the backprojection algorithm, we apply the Dirac delta function in the back projection to propagate the FFT data of a certain point back into its original place. Let us first generalize what the delta Dirac function does to vectors in R^3.

For any vector, V in R^3 we can write this in as a function going from $R^3 \rightarrow R$. For example, V = [2, 3, 1] can be written as V(1) = 2, V(2) = 3, V(3) = 1. This corresponds to three special functions, namely

X	Y	Z
X(1) = 1	Y(1) = 0	Z(0) = 0
X(2) = 0	Y(2) = 1	Z(1) = 0
X(3) = 0	Y(3) = 0	Z(2) = 1

So, V = 2X + 3Y + Z. We have now written V as a sum of three “delta functions”.

The idea that this example shows is that the discrete Fourier transform is a change of basis from the delta function to the exponential basis. $R^n \leftrightarrow fncs\{1, 2, \ldots, N\} \rightarrow R$ and $[1,0, \ldots,0], [0,1, 0, \ldots,0], \ldots, [0,0, \ldots, 1] \leftrightarrow \delta_1, \delta_2, \ldots, \delta_n$.

Therefore $\delta_i(j) = 1$ *if* $i = j$ *and* $\delta_i(j) = 0$ *everywhere else*

Note: In the δ_i basis, an arbitrary function $V : \{1, 2, \ldots, N\} \rightarrow R$ is written

$$V = V(1)\delta_1 + V(2)\delta_2 + \ldots + V(n)\delta_n$$

From this fact, we can show that the delta function is the underlying reason for why the back projection algorithm works.

$$f \star d = f \ \forall f$$

Fact: $f \star d \ = \ f \ \ \forall f$

Now we shall derive why back projections works. We know that $B(a,b,c) \ = \ I(a,b) \times rect(c)$:

$B_i = \ I_i \ \star \ \ (rect(c) \times \delta(a,b))$

$F\{B_i\} = \ F\{I_i\} \ \times \ \ F\{(rect(c)\} \times F\{\delta(a,b)\}$ *by convolution theorem*

$F\{B_i\} = \ F\{I_i\} \ \times \ \ F\{(rect(c)\}$ *and we know that* $F\{\delta(a,b)\} \ = \ 1$

$F\{B_i\} = \ F\{P\}|p \ \times \ \ F\{(rect(c)\}$

$\Sigma F\{B_i\} \ = \ \Sigma F\{P\}|p_i \ \times sinc(c)$

Now assume $P(x,y,z) \ = d(x,y,z)$, then all images are the same $\Rightarrow$ $If \ = \ F^{-1}\{F\{P\}|p_f\}$

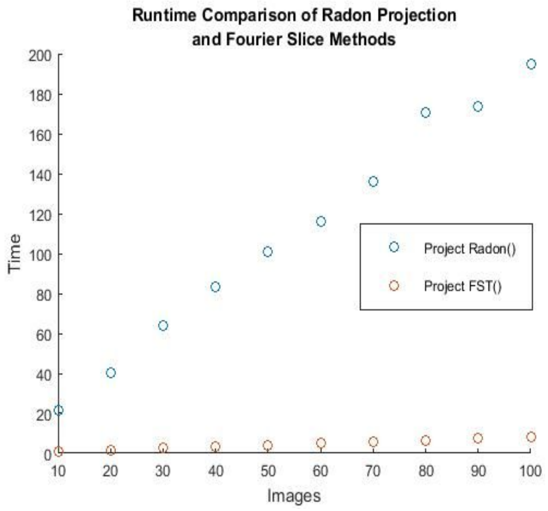
$\Sigma F\{B_i\} \ = \ F\{P\} \ \times \Sigma \, sinc(c)$

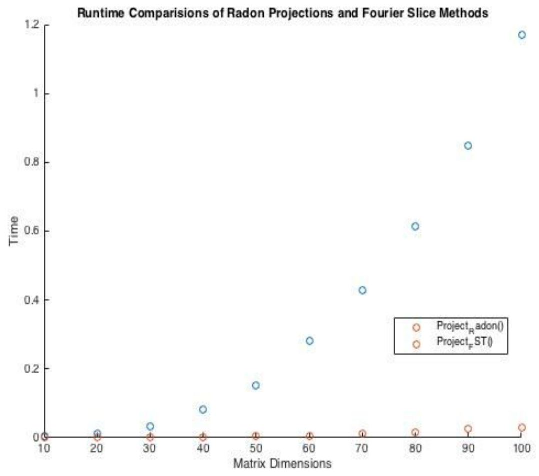
$F^{-1}\{\Sigma F\{B_i\} \ \times \ \frac{1}{\Sigma sinc}\} \ = \ F^{-1}\{F\{P\}\} \ = \ P$

So for general P : $P(x,y,z) \ = \ \int_{R^3} P(x,y,z) \ \times \delta(x-X, \ y-Y, \ z- \ Z) dX dY dZ$

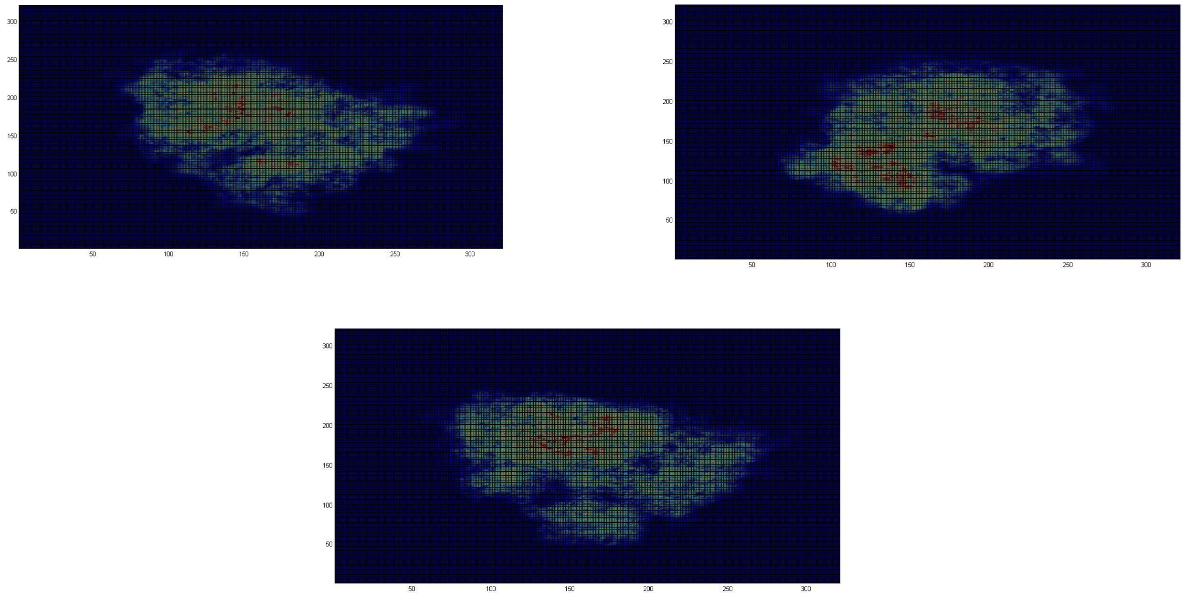
Results:

As the graph on the left demonstrates, the project_FST() function handles Fourier transform image set creation far more effectively. But use of the Fourier slice theorem to generate a set of projections with FFT applied requires knowledge of ϱ, a piece of information that researchers do not have access to (and are in fact attempting to solve for).





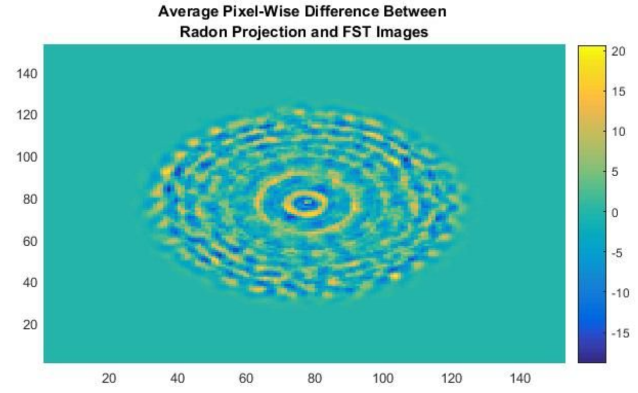
The graph on the right shows runtime growth as we modify the size of the image input itself (as opposed to increasing the number of required images). Clearly the computational complexity of Radon projections is greater than that of the Fourier slice method, a result in line with the theoretical complexity analysis earlier.



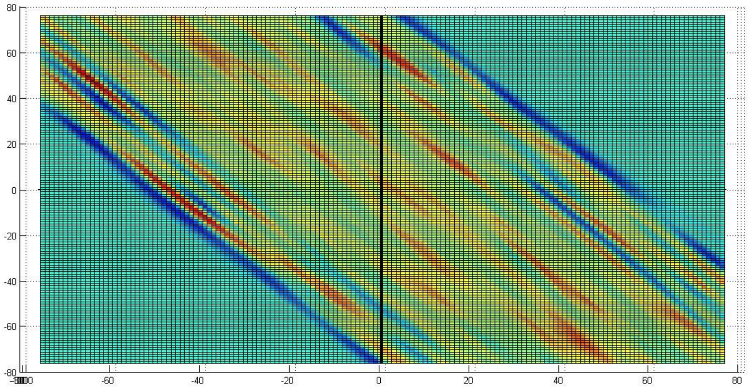
Common lines algorithm fed three images of a ribosomal subunit from unique viewing directions. Images 1, 2, and 3 (counterclockwise from top left)

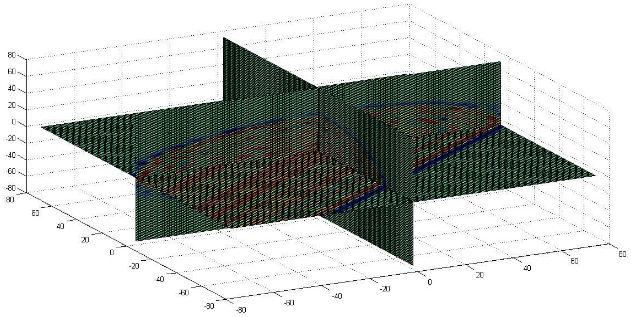
We also examined situations in which the common lines approach operated successfully or failed. As the common line between any two images must go through the origin (shifted to image center), we used the angle between the common line and the x-axis to describe the common line. From the common line code, we know that the common line between image 1 and image 2 is [0 0], and the common line between image1 and image 3 is [0,0] too. Note that the shape of the ribosomal subunit is asymmetrical.

It is crucial to set the given images to have unique common lines because the algorithm otherwise fails. Zika is a symmetric virus and as a result our common line function judges common lines between every two images but will ultimately find many such lines per image. As a result, images from different viewing directions will appear the same to our algorithm and misinterpret their actual orientations relative to each other.



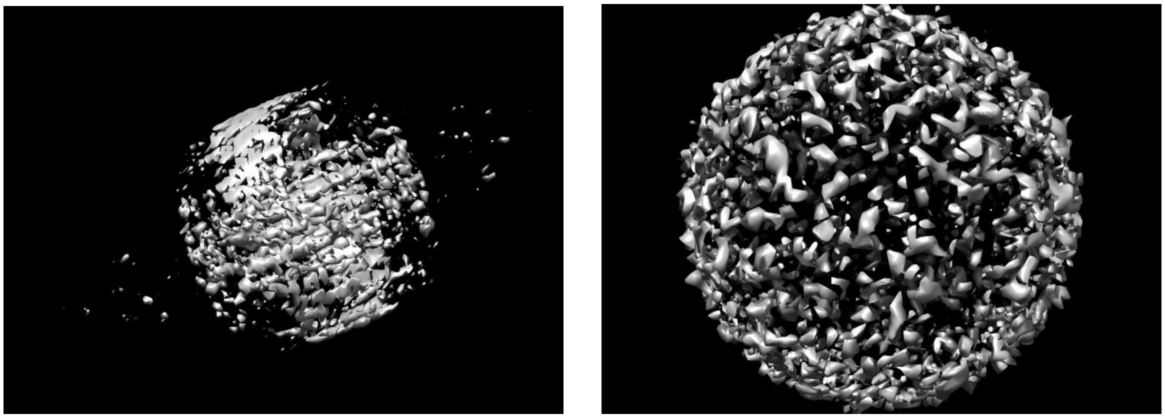
The graph to the left demonstrates the pixel-wise error generated by Radon projections when compared to slice extraction through Fourier slice theorem followed by IFFT. Images with identical viewing frames were compared.





The above left image displays the result of the backprojection algorithm with a single image as input. This image is a two-dimensional slice of the full effect of back projection that smears an image into three dimensions (shown left). A single image 3D reconstruction shown left helps demonstrates the effect of smearing in three dimensions. As we apply more projections to our reconstructed model, we obtain a roughly spherical shape (below left). As we increased our noise filter, we found a reduction in overall image quality as expected. To reduce the interference of the noise, one can try the following possible routes:

1. Do more sampling and collect more images. This allows for better angular resolution.
2. smoothe the image before backprojection: the noisy parts of the image should be distinct from the normal parts, so we can try to make an interpolation before applying algorithm.



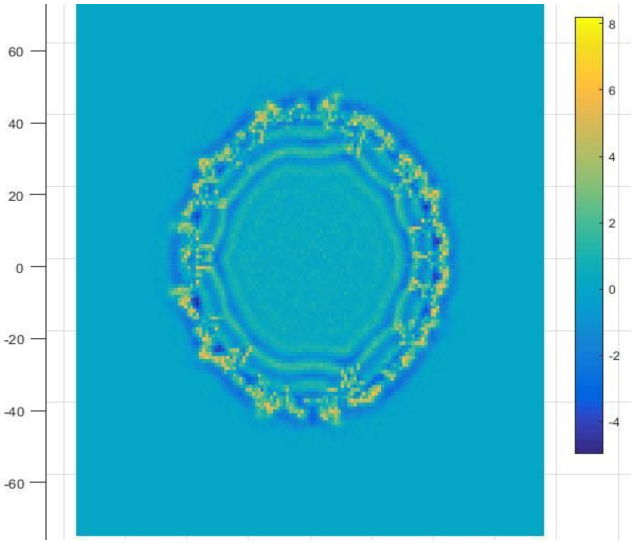
Attempted reconstruction using 10 images (left) versus a reference 3D structure (via Chimera program)

Zika Biology:

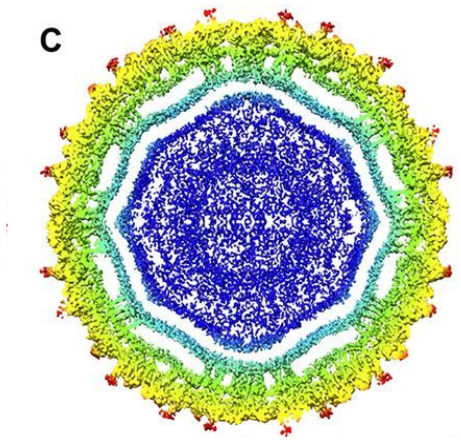
Zika virus has appeared several times in human history, including an outbreak in French Polynesia several years ago. Only recently has it become a major priority, however, with an enormous outbreak in Brazil acting as ground zero for a potential worldwide pandemic. Although the disease's symptoms are relatively weak in comparison to related strains such as Dengue virus or West Nile virus, it has been linked to massive cranial deformities in newborns and may potentially cause Guillain-Barre Syndrome in adults[1], an autoimmune disease that results in nerve cell damage in the peripheral nervous system that in turn causes paralysis. Public health officials are also noting that, given how recently the virus has emerged on the global scene, its full impact on human health is unknown[2].

Zika virus is a (+) sense ssRNA virus, meaning that its genome is composed of a single-strand RNA stretch that can be directly translated into protein products by acting as a template for ribosomes (RNA-protein complexes that act as protein production sites). In order for Zika to reproduce, its genome must be duplicated. Zika accomplishes this by producing a polymerase that generates a (-) sense RNA strand complementary to the viral genome. The newly generated strand can then act as a template for genome duplication. Hence, unlike viruses such as HIV, Zika can reproduce within the host cytoplasm without resorting to splicing itself into nuclear DNA. The virus can then take advantage of ESCRT proteins to transport itself to the cell membrane, where it then exits the cell.

Viruses surround themselves with a protein shell called a *capsid* in order to protect their genomes from harsh environments. This capsid is often quite structurally rigid and forms ordered geometric patterns. In Zika virus, this capsid shell is in turn enveloped by a plasma membrane studded with viral proteins. The membrane, obtained from its host, can help hide the virus from the host immune system.



1. *Capsid protein section (likely location)*



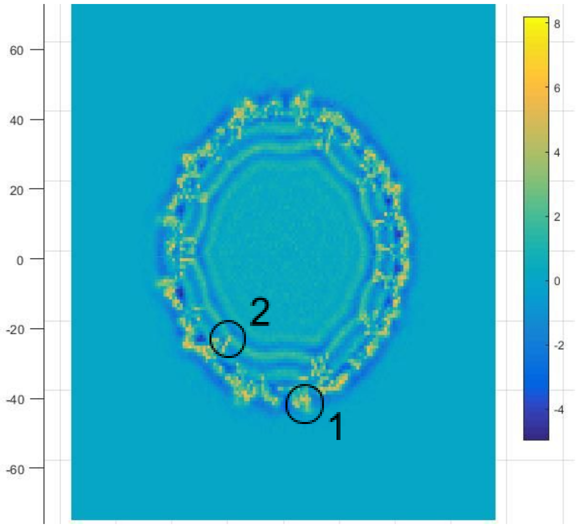
High-resolution image with capsid shell (blue)

Cells use a process known as receptor-mediated endocytosis to selectively transport molecules across the cell membrane and into the cytoplasm. Zika virus takes advantage of this procedure to enter the host. On its surface, Zika displays E proteins that act as binding targets for transmembrane host receptors (the receptor stretches through the entirety of the cell's plasma membrane). This is specifically accomplished through the use of glycosylation sites on the E protein that allows for the attachment of carbohydrates.

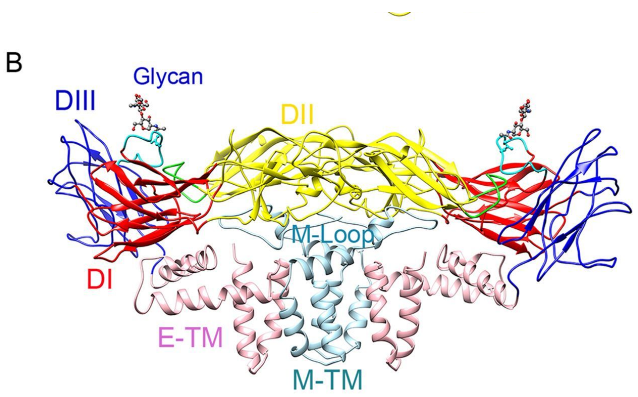
[1] http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2816%2900562-6/abstract
[2] Interview with Alameda County Public Health Department official

Since these carbohydrates mediate host membrane binding, it is not unreasonable to ask if glycosylation sites are highly variable due to viral evolution. After all, if a virus consistently targets a receptor then a host species may adapt evolutionarily by modifying the receptor to no longer recognize the virus. For instance, HIV resistance in humans is conferred by a mutation of the CCR5 receptor that stops the virus from binding.

Comparison of flavivirus E protein glycosylation sites has revealed highly variable patterning and local residue structure, indicating evolutionary pressure to diversify the region. Even among Zika strains, these sites vary in their location and structure[3]. Furthermore, the absence of a single glycosylation site near the N-terminus of E protein (proteins contain N-terminal and C-terminal ends) was found to enhance Zika's infectivity in mosquito cells, leading researchers to posit that this specific mutation may have enabled viral outbreak in West Africa. Statistical tests have shown a high degree of correlation between the presence or absence of the site and the use of a specific mosquito species as a vector during outbreaks in 20th century West Africa[4]. This change shows a concrete example of viral adaptation through modification of its E protein, a process critical to continued survival as hosts respond.



1. E protein on surface 2. M protein Transmembrane domain



High resolution EM structures of E and M proteins

Upon E protein binding, the intracellular portion of the membrane changes conformation (this ability is called *allostery*) to allow an adaptor protein to in turn bind to the receptor. The adaptor protein in turn allows for the recruitment of *clathrins*, specialized proteins that mediate vesicle formation. Clathrins will self-polymerize to form curved lattices around the binding site that encourage membrane deformation[5]. By promoting favorable interactions with the lipid bilayer that comprises the membrane, clathrin

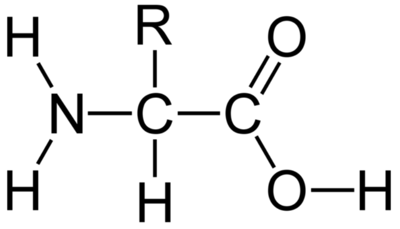
[3] http://science.sciencemag.org/content/early/2016/03/30/science.aaf5316.full#ref-32

[4] Ibid.

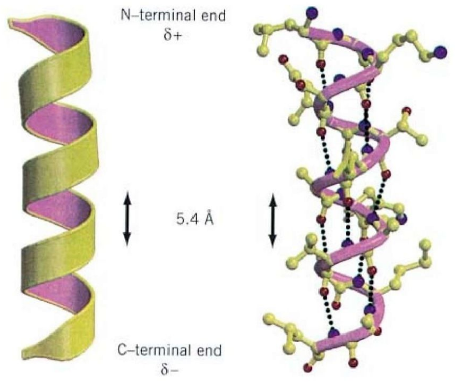
[5] https://www.ebi.ac.uk/interpro/potm/2007_4/Page1.html

encourages the bending processes by compensating for the energy associated with membrane deformation.

As more clathrin proteins are recruited to the region, they pinch the local plasma membrane inwards, creating a pocket that the virion occupies. Another protein called *dynamin* will then mediate severing of the bulge, resulting in a vesicle surrounding the virus on the interior of the cell. The cell is now compromised and Zika will position itself for reproduction inside the rough endoplasmic reticulum, an organelle (a specialized cellular structure) with a high density of ribosomes specialized for protein production and transport.



The E protein itself is attached to the virion partly through the action of M (membrane) proteins. M proteins are transmembrane protein, thus sticking themselves into the plasma membrane that surrounds the virus, and form dimers with E protein in the fully active form of Zika. The transmembrane domain of the M protein is a cluster of alpha helices[6], a conformation that greatly stabilizes the protein's interaction with the plasma membrane and is thus favored for transmembrane (TM) domains. The alpha helix is a spiral that points amino acid side chains (the unique chemical attachment that differentiates amino acids from each other) outwards. This is the R group in the image above. For transmembrane regions, these side chains tend to be hydrophobic in order to encourage favorable interactions with the highly hydrophobic fatty acid tails that compose the interior of the plasma membrane. A useful comparison is imagining the interaction between water and oil versus two oil samples. The alpha helix structure itself allows for polar components of the amino acid backbone (namely the hydrogen and oxygens) to engage in bonding with each other, further stabilizing the structure. Since the M-TM domain is composed of a cluster of alpha helices, its density as visualized in the cryo-EM image reconstruction stands out well against the otherwise low density plasma membrane. Recall that each pixel contains a line integral of electrical potentials given by ϱ, hence atoms such as nitrogen and oxygen with *non*bonding valence electrons (electrons in the outermost occupied energy level) will yield higher electrical potentials and thus densities in our heat mapping. As a result, it is quite easy to identify M proteins in cryo-EM images of Zika with sufficient resolution as shown in the earlier image.



A relevant connection to the larger emergence of cryo-EM as a preferred method to X-ray crystallography is the difficulty of studying transmembrane proteins using the latter approach. As mentioned, X-ray crystallography requires a large amount of protein in order to produce crystals for the experimental setup.

[6] http://science.sciencemag.org/content/early/2016/03/30/science.aaf5316.full#ref-32

Transmembrane proteins are not expressed in high volumes and consequently less than 0.1% of determined structures were membrane proteins despite their accounting for 20-30% of all human proteins (as of January 2013)[7]. Given that membrane proteins are massively important clinical targets (e.g. antihistamines target a type of transmembrane protein called a GPCR) and also crucial for studying cell biology, this disparity is indicative of trouble with current methods as opposed to lack of interest.

[7] http://www.isotope.com/applications/subapplication.cfm?sid=Membrane%20Proteins_14