>> I turn it over to Ed to introduce our speaker now.

>> Well thank you, Manish. Rommie Amaro is a professor and Shuler scholar in the Department of Chemistry and Biochemistry at the University of California San Diego. She received her Bachelor's of Science in Chemical Engineering and Ph.D. in Chemistry both from the University of Illinois at Urbana-Champagne. Rommie started her research program at UC Urvine in 2009 where she later moved to UC San Diego in 2012. Rommie is the director of the NIHP41 National Biomedical Computational Resource and Co-Director of NIH User One Drug Design Data Resource. She's the recipient of many outstanding awards including the NIH Innovator Award, the Presidential Early Career Award for Scientists and Engineers, the ACS Open [inaudible] Outstanding Junior Faculty Award, the [inaudible] Foundation Emerging Leader in Chemistry National Electorate, and the Cohen-Hinch Award. Rommie is one of the largest users of computational time on the NSF-funded Blue Waters Super Computer. On Blue Waters she is running some of the largest simulations in the word to fundamentally understand complex biological systems at the molecular level. She's not only pushing the boundaries of a scientific domain but also the science of scaling extremely large computer simulations on our nation's leadership cloud computing resources. I would like to thank her for taking the time out of her busy schedule to talk about her research and how NSF-funded cyberinfrastructure is enabling discoveries that would not otherwise be possible. Take it away Amaro.

>> Okay.

>> Rommie [inaudible].

>> Yeah, yeah.

>> There we go.

>> Okay. Well thanks for that really kind introduction and you actually touched on, of course, many of the things that I'm going to talk about today. So, just to reintroduce myself, my name is Rommie Amaro. I am a Professor Chemistry and Biochemistry at UC San Diego. As Ed mentioned, I direct the National Biomedical Computation Resource, which I'll tell you a little bit more about during this webinar today but essentially what we do with that resource is really try to develop new, multi-scale modeling algorithms and methods and tools in order to address sort of frontier problems in mostly biomedical space. I also co-direct, together wth Mike [inaudible], the Drug Design Data Resource, which is another NIH resource that is very focused on using a sort of, on improving methods in computer-aided drug design. And that also is a key area that is going to benefit from, and continues to benefit from, sort of [inaudible] computing but I've also weaved in some examples in demands in that space as well. So I tried to frame my talk really around sort of what I see as the emerging challenges in the space of computational biophysics, including some of computer-aided drug design that touches on computational biophysics. Okay, so this is sort of my introductory side and I think probably most people listening to the webinar, this will be really sort of something that they are quite familiar with already. I think that many of us sort of share this opinion that related to the convergence of

high performance computing together with advancements in data science and also extraordinary advances in data, new sources and kinds and types of data that's sort of, it's the convergence of all these things together that will really enable transformative discoveries at the intersection of experimental observational science and simulation sciences. And so, you know, sort of as I'm just showing here the really meteoric rise in computing power, the national super computers that have come about over the past couple of decades and really sort of taking us through to one of our current [inaudible], Blue Waters, I'll talk about some of the work we've done there. We also have worked down on the Triton Oak Ridge National Laboratory sort of leadership class machine and I'll also touch actually on, especially on the talking not just about the very, very large scale simulations but at the smaller scale at the individual domain level or protein level, how GPU computing and really tremendous advances in the algorithms [inaudible] GPU processors is sort of driving a whole new area of need for people like myself in computational biophysics. Okay, and so, again, I'm not going to get too much into sort of the technical details of what we do as computational biophysicists but, you know, really sort of at the broad, in the broader sense what we're trying to do is sort of use, sort of to numerically encode various sets of generally sort of physics-based and chemistry-based equations that sort of, that describe the structure and dynamic of biological systems and then integrate these equations or solve these equations iteratively on sort of the, on different type of architectures, either these massive architectures where we need tons of sort of inter-connected processors in order to compute on very, very large-scale systems but also where we can take advantage of, as I mentioned, sort of single processor graphical processing units but also across many of those, so at scale. And one of the reasons why we like to do this is because experiments, and I'll touch on this more later, one of the really, I think, exciting and sort of fun things about this field is really that it allows scientists to take and integrate and work with the data that we have from experiments but then extend it in key ways that sort of enrich our understanding of these systems beyond what is capable with experiments alone. And so what I'm showing here is the idea here, sort of the concept that I think is easy for many people to grasp is this idea, as I'm showing in this slide, of a computational microscope. And so, you know, we sort of take the, as I mentioned, sort of the data that we get from experiments but then we can bring them to life using physics-based simulation and not just in a way that's sort of video graphics but actually we sort of we simulate these systems in accordance with the theoretically rigorous laws of statistical mechanics and so, at the end of the day, we can extract out meaningful quantities from, like, sort of meaningful quantities that can be tested and validated and bench marked with experiments. That's one of the really sort of powerful approaches of these techniques. Okay, and so, again, so what I'm showing here is sort of an example of a protein that's involved in cancer. It's actually a protein that's just involved in sort of regulating sort of genetic expression so this is something called P53, the sort of wiggling and jiggling protein that you see there is the P53 molecule, which is known as the guardian of the genome. I don't want to get into too much of the details there but this is an important transcriptional protein that regulates a whole bunch of pathways related to cell death and cell birth arrest. And what we can see here, so this, so what I'm showing you sort of all the wiggling and jiggling of the atoms and one of the cool things that we can get, the important things that we can get out of these simulations is, as I mentioned, a sense of the flexibility or the dynamics of these drug targets. And so what, in yellow, [inaudible] is actually one residue that people are trying to develop drugs that would

bind to that. But what we saw in the crystal structures, if you look at all the experimentally available data, as I'm showing here on the left hand panel, there's over 100 crystal structures of this protein in the protein data bank. These are experimentally resolved high resolution structures of this drug target, and in that area that I showed you with that yellow dot, that yellow residue, basically it's sort of, there's a small opening but it's very, very small. It's not really like a pocket where things can bind. And what we see in the molecular dynamic stimulations that we performed on the [inaudible] super computers, we actually saw as I indicated on the last slide, you see a lot of mobility and we see a new pocket open that, in this area. And not only do we see this new pocket open, as I'm showing in orange here, but also it turns out that this site is druggable. And so in the middle panel I'm showing the results of a method developed by Sandra Vaja [phonetic] at Boston University, called computational solvent mapping. It's basically a computational method that allows us to assess where on the structures of biomolecules would ligands, or drug molecules, have a high propensity to bind. So what this is telling us is that not only do we see this pocket opening in the simulations, but also it might be druggable and so we then we can do things like virtual screening to try to, I'm showing the results of which that type of experiment in the slide, in the image on the right. And you can see how the molecule docked in there. So this new site was druggable. And so, one of the exciting things that we've seen is that this sort of idea of mobility of different drug binding sites, this is really sort of a core concept in drug receptors that, you know, the crystal structures that we can get experimentally, the do give us these beautiful images, the sort of Polaroid snapshots of these molecules but that in order to really understand their behavior, we have to get a better sense for their dynamics. Currently, if you want to understand the dynamics at the atomic level, really, the only way that you can do that at this point is with a very fine-grain temporal snapshot, is with these types of simulations. And so, one of the things that we did was we took this open pocket and we did a whole bunch of drug screening against that pocket and then we passed a predicted set of molecules to our experimental colleagues for testing. And what we found was a highly enriched set of compounds that actually reactivated this drug target and this turns out to be sort of something that is high interest for anti-cancer therapeutics. And so I, together with the other scientists involved in this initial discovery, cofounded a company called [inaudible], which I'm required to tell you I am a cofounder of, have equity interest in, and sit on the scientific advisory board of, it's over here in La Jolla in the JLABS incubator space and we now have developed multiple dozens of compounds that reactivate across different types of P53 cancer mutants so it's sort of a pRommiesing new avenue for potential therapeutics and this was really only made possible through these computational approaches. And the only the other thing that I wanted to sort of stress here, which I sort of have rushed through but just to give you a sense for it, is that many people had worked for sort of, you know, two decades or more on trying to develop compounds that would reactivate P53 and with this new model of this open pocket we were able to disclose and discover many more, many more compounds than had been disclosed in the previous decades before of all the research combined. So it really sort of suggested there was a power of a good model. And so as we and others try to apply these types of methods but across different, all sorts of different drug targets, what are the challenges that molecular computational biophysicists face? So you'll see throughout my talk I've sprinkled different challenge slides to try to really hone in one some of the major messages that I see as being places where

innovative cyberinfrastructure can contribute to the types of challenges that we're facing in the field of computational biophysics. So my first challenge slide, we definitely struggle with reproducibility and I'll explain more why that is in the coming slides. Also with scale ability and interoperability of different code. And the reason for that is that, as I sort of indicated in the slide, the slide where I showed the workflow of P53, these drug design experiments, the end-to-end sort of, the end-to-end experiment of starting with the crystal structure and working with a small molecule, you know, working with the small molecular library elucidating the dynamics of the receptors, doing the docking, doing the ranking, et cetera. This is really sort of a very complex experiment with a lot of decisions that has to be made. There's a lot of interaction and different steps [inaudible] computational scientist with different kinds of platforms and so one thing that people in our field have been trying to do, and that we're really mobilizing now I think better as a field, but it's still something we're working very hard to do well, is to develop automated workflows that help encode this 100-step process into something that is truly reproducible. Because, of course, what happens is if a domain scientist is sitting, a student is sitting writing scripts and so on and so forth, these little bits of information at the end of the day, they don't make it into the methods papers, you know, in the method of published works, et cetera, they are things that we somehow are not good at capturing in our sort of regular life. But right now, we're sort of undergoing a phase change where more people are using things, of course, like Jupiter notebooks but also encoding workflows within those that automate these experiments. And this allows us to increase re-use. Of course, it dramatically improves our reproducibility because we can track everything with Providence and we can also make sure that the execution is really maintained across multiple individuals who may be performing these experiments. And it also allows us to scale execution in interesting ways. So we use the Kepler framework and with this we can actually, I'm sure as most people on this call know, we can actually sort of port out these jobs to various types of platforms such as local compute cluster resources, which, of course, I have, most groups in our domain will have local resources. We also have access to the Exceed resources and that might be in different forms. It could be on Comet, it could be on Stampede, it could be on Bridges. You know, we utilize all of those and these workflows can now interoperate pretty seamlessly across any of those resources. And also, of course, Amazon web sources in the cloud and this is something that we've also developed through this. And so there are more and more people now who are trying to take advantage of these types of workflows. In this case we've developed this tool that we published last year, together with [inaudible] developers, of Kepler along with the developers of Amber Molecular Dynamics package in order to just make basically a generalizable, usable tool that others can also, can also sort of, you know, use in their groups. And this really mainly just runs the molecular dynamic portion of these experiments and then we have, we're also in the process of developing sort of the drug docking bits that will also plug in to the end-to-end workflow for drug discovery. So the other bits, I mean interoperability something, I guess I didn't mention that here, but one of the things that is certainly a challenge for people in molecular biophysics or computational biophysics and computational chemistry, also general, is code interoperability. And especially as we move towards increased automation of our workflows we need to or we are trying to sort of develop a more sort of cohesive community environment or ecosystem of codes and tools that can be, you know, sort of swapped in and for various pieces that may be necessary in the end-to-end workflow. So we

might want to use Amber for the [inaudible]. We might want to use a different code like Bromex for clustering. We might want to use Charm for FEP. There's all these different fits and trying to get all of these different kind of code bases to work together is something that is really a major challenge and [inaudible] has invested in this through an effort called [inaudible], which I'll also talk about momentary. Okay, another challenge though is, this is great, we can run these jobs but, you know, how reliable are they? How accurate are the results? And so this is why I wanted to just touch on the work of the Drug Design Data Resource. And so what I do here, and again, I direct this, this is an NIH resource that I direct together with Mike Bilson. What we're really running are blinded prediction challenges to drive method development in computer-aided drug design. So our central goal is to curate and use previously unpublished data sets that typically come from industrial drug discovery program. So, like GlaxoSmithKline, Yansen or Jansen depending on which side of the pond you're sitting on, Novartis, et cetera, and basically what they give us are sets of high quality x-ray crystal structures. So we have drug targets in complex with different drug molecules and so we know how the molecules bind in the active site. This is what we call the pose, and they also give us high quality binding affinity data, typically in the form of IC50s or KDs. And so what we really do then is we curate these challenge data sets, as we call them, and we get them out to the community. And these are, and then we allow the community to dock, to use their docking programs or their drug binding programs to figure out how, and then they make predictions about the pose of the molecule in the binding sites, so how well can they dock the drug. And they also try to guess at how well the molecules actually bind and they do this by ranking the set of compounds that we give them. And so what this allows people to do is really benchmark and test their drug design methods on data that their methods had not seen yet. And this turns out to be extremely valuable because most people are, of course, doing these sort of retrospective analyses and it turns out that the performance on retrospective data sets really does not align well with the performance on new data sets. And so it is, this is a resource where we're really sort of brokering that and hoping to sort of, again, evaluate CAD methods and sort of drive the field forward. So, as an example, so we have a few different, we run challenges every year and so you can see we have about, usually on the order of multiple dozen participants. And these are participant groups in computer-aided drug discovery from all around the globe and basically you can see we have different targets. So in any one year we might focus on, in this case, in 2015 we focused on the HSP90 drug target and then that [inaudible] target and then people try to do drug docking and ranking for those. Another year we used FXR data sets from Roche. And so people, you know, we're able to have these relatively small data sets but they are of high quality and people see how they do and then we write papers. Well one of the things that we realized is the through put of these grand challenges, as we call them, is unfortunately, it's not really high enough for us to make very deep conclusions about how these methods perform across the diversity of drug targets or small molecule compounds. And so what we really want to try to do is actually push towards greater statistical accuracy or statistical certainty here and one of the things that we did in order to do this, sort of a new approach that we've taken, is to collaborate with the protein data bank and so now we have this program, this type of challenge called the continuous evaluation of ligand-posed predictions and basically what we do is the PDB one week before or the Saturday before the, so PDB releases their structures every Wednesday. The previous Saturday they give the D3R access to the sort of small molecule information and

protein sequence information that they're going to release in five days' time. We then send all of our participants through this automated web server. We basically send them this challenge packet where they can then have, essentially we have now drug docking challenges on a weekly basis because of this new source of data. And one of the really cool things about this is that, and we've had this running now, so this is now all done in workflows on sort of a, with just fully automated technology, what we've done is that in the last 64 weeks, so just over a year, we've been able to present to the community over 2000 novel docking sort of challenges. And we actually developed in-house some different, we have these in-house methods that we developed but there's also a whole bunch of different external participants who also are creating these automated methods. So this is, essentially, a lot of people have heard of [inaudible] for protein folding and protein structure prediction, this is essentially like [inaudible] for docking but it's really trying to take advantage of emerging data sources like the PDB and also sort of intelligent cyberinfrastructure or computing so that we can make that data available to the community in a way that they can use and then actually drives methods development. And I'll just say that in the one, in the just over a short year that we've been running this type of challenge, we've been able to amass an order of magnitude of more data and statistics than the whole of the nine years of the previous efforts combined of this research. And I think that this is really, this is really going to be sort of a, it's already proving to be sort of a really sort of a transformative change in how people in computer-aided drug design benchmark and test their methods and also develop them because the only way that this can happen is by capturing complex workflows because now you're talking about a turnaround time in a matter of days. So in our grand challenges where we basically give people this sort of [inaudible] data, they compute on stuff and then they send the answers to us, which was sort of the old format, we give people a couple of months to do that. Now we're only giving them a turnaround of days. So everything is automated in these workflows and what that allows us to do is to really more faithfully compare the methods. We have full reproducibility and we can take those same methods and actually, you know, extrapolate them to new data sets, which is also something that's very powerful and something that typically isn't done in our field. So what we're planning, so now that we've actually established this for docking and it's getting taken up by the community, our next thing to do is actually what we're trying to develop is something called CELLP+, which is basically going to be a similar type of exercise but actually for binding affinity predictions. And here we're going to have much, much more complex workflows and we are hoping to partner with the NSF sponsored Multi Scientific Software Institute in order to help us, really again, develop [inaudible] system of different code bases and platforms where people from all over the world can participate in this more easily. Okay, now, sort of switching gears again. Now, going back to the big scale things, sort of moving a little bit away from drug design again, so that's a picture of a ribosome. It's a blob. The idea here is, you know, what we focused on so far, what I've told you about, are efforts in computational biophysics that really focus at the molecular level or single protein level. Well the reality is that this protein is sitting in a much more complex biological environment in the cell and a major challenge for computational biophysics and for you know, sort of molecular level science [inaudible] really to fully appreciate, understand [inaudible] the complexity that is really in a cell. And so we are trying to, one of the things we're trying to do at the National Biomedical Computation Resource, is to develop methods that will allow us to cross a really broad range of scales from the molecular all

the way to the cell and even to tissue and whole organ level studies. Okay. And why we want to do this is because, you know, at the end of the day if we're trying to understand biological function or disease, the underlying molecular mechanisms of disease, we really have to have a picture. So drug action at its essence, chemical action happens at a very, very small atomic and molecular scale. And what we really, but what we aim to do is to understand how these small-scale changes at the molecular level, how they play out in their broader context and sort of, you know, and result in the immersion behavior at the system level. So the only way to really do that successfully is to have an arsenal of multi-scale methods that bridge these gaps that we have within, you know, across different regimes, right. So we have a lot of tools for studying the molecular structure function, you know, on the angstrom to sort of nanometer-length scales. We have tools to look at subcellular type behaviors here. We're looking at milliseconds, [inaudible] to microns sort of distances. And then we need to understand, also, how these subcellular, sort of, bits all come together and work in the, function and cell, so forth. So what I'm going to talk about now is how we're trying to push what we're doing to sort of push molecular simulation into the boundaries of the subcellular and, you know, ultimately I think probably what we'll see is sort of [inaudible] studies and how we want to make these different sort of what is now very cutting edge, we think this is going to be routine within five and hopefully ten years' time. So I mainly, you know, I want to stick in this talk, I know I don't have too much time. I want to talk mostly about sort of the cyberinfrastructure challenges and sort of, and really focus on that because of the audience. I did just want to also mention, of course, that there are a number of algorithmic challenges that people are really working on and I didn't want to go into details but I just wanted to point out sort of, I think, a nice perspective article that I work with a colleague, Adrian Mulholland at Bristol that sort of goes more into some of the algorithmic challenges of the multi-scale problem. But what I'll touch on here today, I mean one of the things that we find particularly exciting is this idea or this concept that because we have just really enormous advances in our, in acquisition of biological structural data, what we want to do is develop methods that will allow us to integrate these different scales of data. Right? And so we have things like x-ray crystallography and electron crystallography that give us very high resolution detail of sort of single, sort of single receptors or proteins in their native environment. You have, of course, everybody, cryoelectronmicroscopy is really on fire, you know. Everybody is trying to get their hands on these Titan microscopes and these now are able because of a number of different technological advances, they are able now to, in some instances, give us near atomic resolution for [inaudible] molecular complexes, which is fantastic. We also have approaches like cryoelectron tomography where actually they can vitrify whole cells and then take very, very thin slices of those cells to see how these macromolecular machines are really situated inside their subcellular environment but in intact cells, which is so exciting. And then, in a different imaging modality we have sort of what we call the resin-embedded methods where we're able to, again, sort of increase our time and length scale so now we're looking at these different sort of cellular and subcellular systems or complexes and then serial block EM where we actually have sort of whole tissue samples that sit inside these microscopes and these microscopes will basically image the top plane of the sample and then shave off a very thin slice, image another plane, shave off another slice. These things run for, like I said, weeks at a time and they get routine data sets on the order of 1.2 trillion pixels. And so in these data sets here at the end of the tissue scale, we have hundreds of

thousands of these individual structures but in its sort of full biological complexity. So this challenge is, there are a lot of challenges in all of these different modalities of structural acquisition and I want to focus mostly on the computational bits but just touching on some of these. Data complexity; data complexity is just a sort, of its just a challenge that we're facing in computational biophysics going forward. This relates to things like, for example, the development of better algorithms to perform better imaging, segmentation, and refinement of those big data sets in the tissue and so forth. Also, how do we extract, how to really maximally extract signal from rich data sets like cryoEM and there is, again, a lot of people working in that area also. And then from the computational biophysics side, in terms of simulation, a key challenge has been data integration. So how do we piece together all of the different components to bring molecular structure into cellular environments. And so one way that we're trying to do that at our center is through a tool called cellPACK. And cellPACK basically can, so it's a cell-centered, data centric modeling framework that basically can ingest all different types of data; essentially, any kind of data that a scientist can feed in. This could be structural data, you know, at the individual protein level, it could be tomography data, you know, taken at larger scale. It could be fluorescence data, it could be cardiomic data, or just anything that is mined from the literature. And then the users create different recipes that define compartments, so the interior of this particular system and the exterior of the surface ingredients or compartments, and then you can find, you can define different ingredients. So what are, what is inside the interior you, in many cases we know the predicted [inaudible] of the different constituency within, for example, cell or the particle. And we could also further constrain that if we know something about where they're placed and their localization in terms of structure. And we can then create these different recipes for different parts of the system and then what cellPACK does is it creates not just one model but an ensemble of models that sort of obey these different rules. And so one of the really exciting things about this is that, you know, typically when you're trying to deal with all these types of data even to construct one model with this type of detail, you know, has been a path taken with an incredible amount of time, almost taking really years to work it up. But now what we can do with this framework is in a matter of, really, minutes, we can create ensembles of molecules of different models and why that's important is because it allows us to sort of get a hint of what's happening with biological heterogeneity, okay, because it's not just, you know, if you have one instance of a system that's great and you can get a lot of insight but if really biologically there's many copies and they may all have particular features but sort of express them in different ways. So this goes toward that. Another major issue or challenge with computational biophysics, especially at the molecular level, are the development of membranes; how we build membranes, how we can simulate membranes. So, you know, there's a lot of people working in this space. I'm just showing you an example of one tool that we built that interfaces with cellPACK in order to create. So, for example, if you had this particle and you wanted to map a membrane around it, we have this tool now called lipid wrapper that basically will make these atomic level membranes in sort of a reproducible, in a reproducible way. And what this allows us to do is really move from single protein studies of computational biophysics to much more complex systems. For example, here what I'm showing is an example, sort of a science example in influenza where we had cryoelectron tomography data from Alistair Stephen at the NIH and we have individual crystal structures for each of the different bits of influenza and then we can bring, we can use

integrative modeling to place those [inaudible] components into the locations specified by the cryoelectron tomography, these are the tools we use to do that. And we were able to fully reconstruct the influenza virus in full atomic detail. And so there's many different reasons why we want to do this but, again, because I want throughout talk more about challenges, I'm just going to sort of push forward on this particular slide. So what we did here is we built this atomic level system of influenza, it's fully [inaudible] has approximately 160 million atoms. I think it's still probably the largest system ever simulated at the atomic level from a biophysical standpoint. On Blue Waters we were able to get about four and a half to five nanoseconds per day using about 114,000 CPUs and we collected 160 nanoseconds of total simulation. This resulted in about 25 terabytes of data and was a collaboration of [inaudible] P41 resource in Urbana. And so this was really quite exciting. So what we actually, so what I'm showing you here is what the dynamics. So this is the influenza virus and I'm spinning it but you can also see very faintly sort of the wiggling and jiggling of those atoms. That's the actual, the wiggling and jiggling, as predicted by the physics-based molecular dynamic simulations. And so this is telling us something about how, you know, the structure and dynamics of this particle. Now, one of the sort of, you know, challenges is simulating these big systems for long time scales because, obviously, it's an enormous amount of computing time to simulate even just hundreds of nanoseconds. So, you know, I've been a big supporter of these very large machines because I see that, you know, this is, the influenza virus is just one example but now that we have this tool like cellPACK and other tools that allow us to build and construct these systems more readily, you know, I think that moving forward there's going to be many more people who also, you know, are working at this scale and it's going to take large-scale computing resources in order to make this happen. So challenges in this space are accessibility. Accessibility to increasingly large data sets and this actually could be experimental data sets so how we interact with them as computational scientists, where they fit, what we can see, how we process on them. Also, accessibility to computing on the big machines like Blue Waters and Titan. This is something that, you know, by design it has to be sort of a competitive sort of enterprise but, you know, it is also challenging because it takes an enormous amount of effort to sort of even build one of these systems and sort of getting in sync with the building of the system and the sort of the proposaling to actually get the computing time and then it's a risk that you may or may not get it. These are all things that sort of add, I think, to some dimension of challenge for computational biophysics community and then also, as I'll touch on next, we also need many, many people and I'm sure I speak for the many, many people [inaudible] in our field, we really now need access to very, very large farms of GPUs. And this is because of new visualization and mostly, actually, analysis that I will tell you about next. So, skip this slide because I'm running out of time. So not only can we run these very large scale simulations like I showed you and these are sort of the brute force way of looking at the dynamics but there's a new sort of, a method that has emerged called Markov's state model theory that basically allows us to create sort of a network of states based on many small, many individual simulations. We can sort of [inaudible] our, in a sense what we're doing here is [inaudible] our large system into many individual copies and then we can create a network of states and we can understand, we can see the transitions between the states. And what this allows us to do, essentially, is to extract long time scale dynamics from many short time scale simulations. And I guess in this case of the flu, we have, because we have so many copies of, for example, this peak protein, this is the

hemagglutinin protein. We have so many copies in this one system that we can basically treat them independently and understand their dynamics. But for many people, what they do is they're just running many individual copies independently on graphical processing unit. So they have one drug target and they're running it off to like, you know, thousands of GPUs in order to understand maybe an activation mechanism or what happens upon ligand binding. More and more and more and more people are doing that. I think this is definitely the future for the analysis of molecular simulation and this really is going to drive, I think, leads for computing infrastructure going forward and I think that is something that we've tried to make clear. And, again, what this allows us to do is even from that 160 nanosecond simulation that I showed you of the flu virus, we can actually determine through Markov state model theory, we can actually determine through the rigorous kinetics of, for example, the loop opening and closing, which happens on the order of tens to hundreds of nanoseconds. And why we care about this, again, is because this is the [inaudible] is the target in flu, this is the drug molecule Tamiflu, which probably a lot of you have heard of or maybe even taken, it's one of our only orally available drugs against the flu. We're trying to understand, you know, basically how new pockets open up, what their propensity is to open and close in different strains of the flu, et cetera. The biology, sort of the driving science here, is actually quite exciting for influenza and for potential therapeutics. I also want to touch on though, it's not just biomedical science but it's also chemical science of the intersection also of biology and chemistry and atmospheric chemistry. There's this fantastic MSF center for understanding the impact of sea spray aerosols chemistry of the environment something called CASE. And so a lot of what I talked about in the face of tissues and cells, similar methods can now be brought to try and to understand how these sea spray aerosol particles, which basically form off of crashing waves, how they go up into the atmosphere and impact chemical reactions that are happening in our atmosphere and affecting the chemistry of the environment. And whereas typically we were studying the sort of very short time scale sort of reaction dynamics, CASE is now really sort of increasing the complexity of their modeled systems and the reason for that is that there's all sorts of different biological components that are transferring out of the ocean into these aerosol particles that are sort of floating up into the air that we breathe and into the air in the troposphere. And so, again, what we can do is we can use, again, similar systems like cell packs, we can use that sort of framework but not in the biomedical space but in the space, again, of atmospheric chemistry. So here we're really trying to develop data-driven simulations that are going to help bridge biological and chemical complexities so our input data sets are now techniques more related to analytical atmospheric and physical chemistry and we have these, this is an example of the sea spray aerosol particle and it has biological components in it. It also has a whole bunch of different ionic species and different, different sort of fatty acids and liposaccharides and so forth, and then there's this sort of monolayer of fat and other bits that create it's sort of external, the part that basically interfaces between the bulk solvent of the particle and the air. And so we can create various different particles and very finely tweak different components in order to match what they are measuring experimentally. And what this allows us to do is really get like unprecedented insight into the behavior and dynamics of particles in atmospheric chemistry to try to understand how it matches to reaction dynamics. And, of course, again I said that we wanted to make this more data driven, and this is actually a slide that I had borrowed from a collaborator at [inaudible] here also at UC San Diego, and we're using her Kepler

framework to try to better encapsulate our end-to-end workflow, not only computationally but also with input from experiments. And so this sort of, this is sort of a hand drawn workflow that we drew of all the different things, sort of steps in our experiment. We are using GPU-enabled molecular dynamics here and sort of the end-to-end workflow, large-scale PCA and clustering. We have model optimization. We need continuous data access integration and transformation and this is sort of feeding back to experimental design, right? So some of the results that we get from the computational world we can use to create sort of new hypotheses that have been tested and so that we have this experimental feedback. And, again, when we do this in the framework with workflows, we sort of have this built-in scientific communication and reproducibility that we otherwise typically do not have. And this, like when we're doing all of the different Markov state models and simulations and analysis, it uses a lot of different core hours of course, on the various type of architectures. And so being able to map particular problems to these particular spaces is something that we, you know, we have to sort of grapple with and to do that in a reproducible way. And then as Bill kind of noted, and I think this is really the case not only in biomedicine, as I showed, but also with atmospheric chemistry, is that real problem solving [inaudible] scientists now a days I think, in our case is really happening at the application integration level. And so, you know, we want to combine or to use sort of more [inaudible] tools. In this case, we were using Kepler and I sort of touched on the Amber GPU workflow that we developed for biomolecules. We used that here also in our MD workflow component of the data generation stage. So we have really, we're using these frameworks now to sort of try to encapsulate everything from data acquisition to data generation and data analysis and storage and to sort of take better care of all of the different components of the experiment to do it in sort of a way that's going to sort of maximize discover and reproducibility and rigor and so forth. Okay, and then of course, I really didn't mention this but I have to basically say this, for sure, I have to remember to say this, that, you know, a major challenge is really student, post doc and scientist training. Not only just in the particular domains but you see that at this, for computational biophysicists, whether it's biomedicine or something else or materials design, et cetera, they really need to have exposure to and training in so many different really disciplines and it has to be training that is not so superficial in the end. You know, to really be able to do something meaningful again in sort of a reproducible way that's scientifically grounded. It's actually quite challenging and so we've appreciated, certainly, all the efforts that various NSF centers have made, and others, to try to train the next workforce that is really sort of more data science oriented and data science centric. All right, and then I just have a couple of acknowledgement slides and then I'll be happy to take questions. I think I still have about ten minutes for questions. I touched on a lot of different work and so there's, of course, a lot of different people to think including people in my group, a lot of funding from the National Institute of Health, always, always have had such tremendous support from the National Science Foundation in terms of computing architectures. And now, also, through the NSF Center for Aerosol Impacts on Chemistry and the Environment, [inaudible] Chemical Innovation Center that was just recently renewed that's also now really turning into sort of a major effort going forward and I think is a really beautiful example of driving elements from the domain science into sort of the cyberinfrastructure of the future. So, with that, I'm happy to take questions.

>> Thank you very much. We will take questions for Dr. Amaro. I am going to briefly switch to a screen that will show the email address to which our remote participants can send questions, and then I will switch back so that if Dr. Amaro if you would like to refer back to your slides you'll be able to do that.

>> Sure.

>> [inaudible] back here.

>> Okay. Thanks Rommie. That was a wonderful and wide-ranging talk that touches upon simulation and computation at all scales. So I just have one question just to kick off the question and answer sections of this presentation, so you talked about challenges. Specifically, you talked about accessibility of data sets, the need for hardware, big machines as well as GPUs and you also talked about students and scientist training. I was just wondering if you could say something about the need for a [inaudible] as well because I know you had wonderful interactions with the Blue Waters team as well and maybe you could touch upon, you know, what were the challenges you faced in importing your code in Blue Waters and how important were the availability of experts in your research?

>> Oh sure. Oh yeah, that's a great, yeah. And I did not, I'm sure I did not do justice [laughter] so thank you for letting me thank those people too. Right, so on Blue Waters and also on Titan, we used the [inaudible] code, which is developed at the Center for Theoretical and Computational Biophysics in Urbana. And they developed that mainly and the main developer there is someone named Jim Phillips, who is a fantastic computational scientist and physicist, sort of the interface of sort of domain science but like with a very special expertise in terms of like porting the code to the various different architectures. And, you know, so Jim and also, of course, the cyberinfrastructure contacts at each of the sites, of course, none of this would be possible without their very, very hard work. So when we were porting the flu virus to, at first, Blue Waters, it was the largest simulation that had been run and so it was very exciting I think for the developers of [inaudible] to also see how their code scaled, that we would not have been able to get up and running on that machine certainly without their hard efforts. And so there it was really sort of a triangulated team, right? So it was our team with the system that we built. So we're sort of acknowledging that our group focuses more on the tool development that allows us to build the systems and then with the collaboration with Jim and the [inaudible] team in order to actually be able to run the molecular dynamics but then on the Blue Waters super computer. So it was really the three groups together that sort of, that worked it out. And this is the same, this sort of same approach is also what we use on the other big machines like Titan and so forth but it's just with a different cyber infrastructure contact. So I'm not actually the person actually porting the cord, gratefully, the [inaudible] code to these big machines, we rely on their expertise and they, of course, are fantastic in helping us get up and running. Absolutely.

>> Thank you Rommie.

>> Yeah.

>> I think we have time for perhaps one question.

>> So I was really, this is Bill Miller, another colleague at OAC. Rommie, I was very interested in this question of workflows and how that turns into sort of a computational strategy because once upon a time perhaps, there were less kinds of resources and you had, you know, fewer choices but also, maybe, the questions were less complex. And now, what I saw-

>> Mmhmm.

>> from your talk is that, you know, you have many different ways you could go about the various computing pieces and many different ways you could go about the modeling pieces so they sort of, the whole science of designing an experiment in this regime seems much more complicated than it might have been once upon a time. So I wonder if you could just, just your own personal experience about that challenge would be very interesting to hear.

>> Yeah. Well, I think that's really one of the main messages that I was trying to sort of convey, right, is that by the nature of everything combined, everything is really becoming more complex and I think that the really outstanding challenges are, they're really at the intersection of, you know, independently complex areas. And the work, you know, the workflow integration, integration of workflows into these approaches allows us to really sort of start thinking about [inaudible] a really great slide even though it's a little bit busy but it really sort of allows us to develop these approaches that are, again, robust, right? And so, yeah I agree with you, I think, I don't think this is a trend that is going to stop, right? I think that it's just going to increase sort of the dependence that we have on, again, on integrating different areas of science, different techniques in order to answer a question that is outstanding or where there's, you know, discovery to be made. That's just, it's only going to continue to become more complex. So, I mean, I think that, again, that's one of the real advantages with workflows is that we can sort of codify the approaches and then sort of iterate these processes and really carefully tweak individual components of that also to really tease out at a very fine level, you know, what is impacting various results or, you know, in the case of sort of method development. But then in the case of sort of experimental integration, sort of using the work flows to bring in new sources of data, maybe contemporaneous sources of data, that are used to actually drive the simulation as it goes. That's another sort of very cool aspect that I didn't touch on at all but that is coming into play now for molecular simulation.

>> I just wanted to add, you do have the presentation ball here so if you had a slide that you wanted to share, are you able to see the screen sharing option?

>> Oh sure. I'm sorry, I did that whole thing without sharing my screen [laughter].

>> No problem.

>> I was looking at, I'm sorry, I was looking at this slide where, you know, where you have, where workflows are really sort of enabling us to, you know, in this case we have, this is our actual experiment like in our lab notebook while we're thinking about it, right? So how we draw it up on the back of a napkin or whatever, but actually at the end of the day, what the workflows allow us to do is to create really this codified framework that allows us to very precisely control and execute each of these and explore each of the different bits in this very complex strategy.

>> Okay. I think we are just about out of time but I would like to think Dr. Rommie Amaro again for her presentation and I'd like to thank all of you for attending. Again, there will be no OAC webinar in November because of Supercomputing 2018 but we are planning to resume the series in December, speaker to still be determined. And with that, thanks again and we'll close.