All right, I’m very excited to introduce our speaker for today. We have Dr. Meghan Short. Dr. Short has completed fellowships at the Glenn Biggs Institute for Alzheimer’s and Neurodegenerative Diseases, and at Harvard’s Huttenhower Lab. Currently, Dr. Short is an assistant professor at Tufts University. Let’s give a warm welcome to Dr. Short.

Hi, everyone, Thank you for being here. Can you all hear me, okay? Sign in if you’re registered. All right, so, today, I’m going to talk about a project that I worked on as part of my postdoc down at UT Health San Antonio with the Glenn Biggs Institute for Alzheimer’s and Neurodegenerative Diseases, and I wanted to talk about this as a... None of the sort of methods that I’m gonna talk about in this talk are particularly new. This wasn’t sort of a methods development project. So the sort of main network method I’ll talk about is about a decade old at this point, at least, but what’s nice about it is that with increasing availability
of high dimensional biomedical data, it’s sort of seeing more use cases, and it’s not something that, at least, I learned about in my graduate program in biostatistics, but it’s something that I thought would be good to talk about today since it’s such a useful method. So let’s see if I advance. There we go.

So I’ll start just by giving a quick introduction. I know that when I was in grad school, I always wanted, I thought it was interesting to hear about people’s career paths as I was considering my own. So I started in biology as a field. I studied salt marsh ecology as an undergrad, and then by the end of undergrad, I was interested in getting more into sort of a human, more directly human-focused environment, and so I considered public health. I learned about statistics as part of my research in undergrad and wanted to continue with that so I participated in SIBS, which is a program that you may be aware of, and that was my first intro to biostat.

I was a graduate student at Boston University. I had fortune of working with the Framingham Heart Study,
which is where the data comes from that I’ll be talking to you about today, which is a really interesting study, and I’ll get more details on in the few slides. That was sort of my introduction to working with epidemiological data. After grad school, I continued on, again, to UT Health San Antonio, and then following that to postdoc at Harvard looking at developing methods for microbiome analysis. So if you have any interest in that, feel free to approach me, although I’m not gonna talk about that today, and then as of March this year, I started as an assistant professor at Tufts Medicine where I’m working on a variety of projects but a lot related to sort of omics data and aging and longevity. So I’ll start today’s talk with a bit of motivation for why network-based analyses we’re a good fit for looking at sort of the proteome in Alzheimer’s disease. So first of all, Alzheimer’s disease is a very prevalent condition. Many of you may be like me and know some family members or people who have been affected by it. It’s very common and expect it to be more so as populations age, and it’s a leading cause of mortality.
disability, and poor health among seniors, and one interesting feature of this disease is that precursors of it can appear years to decades before symptoms manifest. So those precursors can include indicators that are visible on brain MRIs, performance on neurocognitive testing, changes in gait, even changes in sense of smell, and cerebral spinal fluid markers, such as tau and amyloid. Because of this, there’s interest in being able to find plasma biomarkers for Alzheimer’s disease and related dementias. ADRD is an acronym we’ll be using sort of throughout. Because since there are indicators of sort of pre-disease development in years to decades before being able to detect those, either earlier or in a less invasive or expensive way, is very useful, and so when I say invasive, I mentioned CSF markers, such as how an amyloid can predict dementia, but that involves doing a lumbar puncture versus something like a blood draw, which is easier to do. Another good aspect of trying to find biomarkers
is that you can get a sense of biological processes that are involved in disease development, and that can hopefully lead to either preventative or therapeutic interventions.

What makes this difficult?

So in my case, I was looking at proteins. There are thousands and thousands to select from, and you get sort of this inherent trade off between trying to control a false positive rate for all these multiple tests that you may be performing, but if you effectively control the false positive rate, you’re going to likely end up with low statistical power.

There’s this trade off between... It’s sort of a needle in a haystack. Another thing that has tended to be true is that there is not very good replicability across studies. So one study may find 20 biomarkers and maybe one or two of them may replicate in a different study.

So there’s a lot of noise that ends up coming through. The approach that I took in this project was to use network analysis to analyze the protein data, and the motivation there is to try and capture subtle but consistent variation in groups of proteins.
I'll refer to them as modules during this talk. In then just a few things, so first of all, it reduces the dimensionality of the statistical testing problem that you have. So rather than testing each protein individually and having to adjust for all of those multiple tests, you can sort of reduce the space to a smaller number of tests where the proteins within each group being tested are inter-correlated with one another, and unlike other dimensionality reduction methods, something like a principle components analysis that you may have maybe familiar with, the network method has sort of a benefit of looking not just at, say, correlations or relationships between pairs of proteins, but, also, at sort of the correlational neighborhood of what common neighbors those proteins share in the network. Another benefit of or sort of way that we try to get around some of the pitfalls of proteomic analysis is by focusing on biological pathways instead of on individual proteins themselves. So within groups of proteins that we find to be of interest.
or possibly associated with dementia outcomes, we use a tool called over-representation analysis, which I’ll talk about later, but it essentially tries to pinpoint biological pathways that may be overrepresented by the proteins that are found to be associated with the outcome, and the hope there is to find, to get sort of insights that are more robust across studies and, hopefully, address some of the issues with replicability. Okay, so that’s sort of the motivation for this study, and, now, I’ll talk a little bit about the data. The data for this study comes from the Framingham Heart Study, which has been going on for a very long time. It started in 1948 in a town of Framingham, Massachusetts, and at the time they enrolled, they reached out to two-thirds of the population of the town to try and enroll them in this epidemiological study. It was one of the first ones of its kind, and people would come in for exams every few years, and they would take all of this information about them, and then follow them for outcomes.
Cardiovascular outcomes was really the sort of outcome of interest when it first started.

Over the years, they’ve then enrolled offspring of the original cohort participants as well as grandchildren and third generation, and then as sort of the demographics of Framingham have changed over the years, if you’re only enrolling descendants of people who live there in 1948, you’re not gonna capture that.

So they also have been enrolling omni cohorts to reflect sort of more diverse populations (indistinct).

Again, they were sort of aiming towards identifying risk factors and etiologies of cardiovascular disease, but as those populations age, brain health and cognition is also an important outcome, so they’ve measured sort of cognitive outcomes and incidents of dementia as well, and, of course, those things are also related to cardiovascular.

For our study in particular, we were using the offspring cohort, and at their examination cycle five, which was in the early 90s, they collected blood samples, and froze the plasma from those samples, and years later, when they sort of had...
these broader proteomic analysis assays available,

they measured the plasma proteome,

I’ll talk about the methods for that on the next slide,

but they did this in about 1,900 participants

who were approximately aged 55 when the blood was drawn.

So this is sort of a middle-aged cohort,

generally, cognitively healthy

and a little more than half women.

The main outcomes that we looked at in this study

are MRI-based measures, so brain MRIs were taken

about 10 years or so, five to 10 years

after the initial blood draws, and those had...

The sort of outcomes that I looked at there are

total brain volume as well as the volume of the hippocampus

and then a measure called white matter hyperintensities,

which is sort of a measure of vascular injury in the brain,

and a reason to look at those outcomes is that

I mentioned there are sort of precursors of dementia

or risk factors for dementia that can be identified on MRI,

those are some of the big ones.

Especially since we had a middle-aged cohort,
you may not see a lot of incident dementia,

and so being able to detect proteins
that are associated with some of those precur-
sors is a way of getting at this issue.

We did also look at incident dementia.

So we had about 20 years of follow-up,
which is one of the strengths of this,
looking in this particular sample,
and we had 128 incidences of dementia
of which 94 of them were classified
as Alzheimer’s type dementia.

We also had a replication cohort.
I mentioned the importance replication,
and so we worked with collaborators
at the University of Washington and their
cohort study called the Cardiovascular Health Study,
which has sites, I think, four different sites
around the US and has measures of the same proteomic plat-
form and same outcomes that we’re looking at in
the study.

The assay that we used to measure proteins
is called SOMAScan.
It’s by this company called SomaLogic.
They use these single-stranded DNA aptamers
that are designed to specifically bind
to different proteins, and you can sort of tag
them that way and measure their concentrations.
In our sample, the assay had 1,300 proteins,
which is even sort of becoming dated now.
I think the latest version
253 00:12:46.470 --> 00:12:48.300 has something like 7,000 proteins.
254 00:12:48.300 --> 00:12:50.580 So there’s a lot that can be measured with this,
255 00:12:50.580 --> 00:12:55.203 but there is some sort of bias towards, I think,
256 00:12:56.850 --> 00:12:59.190 molecules that sort of have some evidence
257 00:12:59.190 --> 00:13:01.080 of being important in cardiovascular disease.
258 00:13:01.080 --> 00:13:04.743 So it’s not an entirely sort of agnostic choice
259 00:13:05.793 --> 00:13:07.893 of proteins,
260 00:13:10.620 --> 00:13:14.730 Okay, so that’s a description of the data,
261 00:13:14.730 --> 00:13:16.920 and, now, I want to dig in a bit
262 00:13:16.920 --> 00:13:19.083 to the network methods that we used.
263 00:13:20.010 --> 00:13:24.390 So this is sort of a graphical abstract
264 00:13:24.390 --> 00:13:26.253 from their original paper,
265 00:13:28.080 --> 00:13:29.460 describing this weighted gene
266 00:13:29.460 --> 00:13:31.413 correlation network analysis method.
267 00:13:32.310 --> 00:13:34.410 So that’s what WGCNA stands for.
268 00:13:34.410 --> 00:13:37.159 I put gene in parentheses because they’ve started
269 00:13:37.159 --> 00:13:40.290 dropping that from the name when it gets used elsewhere
270 00:13:40.290 --> 00:13:42.330 because, originally, it was developed
271 00:13:42.330 --> 00:13:45.180 for gene expression data, but it’s been found to have use
272 00:13:45.180 --> 00:13:48.240 in other high dimensional data sets as well,
273 00:13:48.240 --> 00:13:52.380 and so in our case, we’re using it to analyze proteins,
274 00:13:52.380 --> 00:13:56.463 but the language here makes reference to gene expression.
275 00:13:57.450 --> 00:14:00.621 So just broadly, what this method does
276 00:14:00.621 --> 00:14:04.050 is you get a co-expression network,
277 00:14:04.050 --> 00:14:07.710 and I’ll sort of give details on the next few slides,
but the idea is that the network is based on co-occurrence or correlation in your sample. So there’s not really information coming from outside.

You’re not even considering your outcome at all. It’s just looking at the space of the proteins and which proteins are correlated with one another.

Once you’ve identified this sort of network matrix, you use a hierarchical clustering algorithm to define modules.

It’s a little small here, but I’ll show a a bigger example. Basically, you have a dendrogram, and you see that if sort of proteins are on this x-axis of this figure here. I’ll do the mouse for people who are online.

You get these sort of bands or groups of proteins that are highly correlated with one another and not correlated with other proteins. So that is where those sort of protein groups come from.

Once you have those, you can use a numerical summary of each protein group as sort of a feature or a predictor in a regression or some sort of analysis to try and relate the modules or groups to external information.

So that’s how we relate our protein groups to dementia outcomes in this study.
There's also the possibility of looking at relationships between modules. So I mentioned the modules in the network are highly inter-correlated within the proteins within themselves, but there may also be some correlation between modules, and that could be important to look at as well, and then within modules, you may have tens or hundreds of proteins, and so trying to figure out which proteins within those modules are driving any associations you see is sort of a final step that can be useful for getting sort of biological meaning out of these associations.

This is sort of a more graphical abstract from our study, and I'll sort of go through bit by bit the different pieces of the analysis. So, again, this WGCNA step is sort of the first step of getting from this protein expression matrix where you have sort of your proteins by participants, and using the sort of correlations in your sample to come up with these modules of co-expressed proteins. The first step in doing that is to make a pairwise correlation or similarity matrix.
So if you have $n$ proteins, then that becomes an $n \times n$ matrix where each cell is describing the similarity or correlation between protein $i$ and protein $j$ in your sample.

You then use this to create what’s called an adjacency matrix, which is a more networky way of describing the association between proteins, and then a topological overlap matrix, which then takes into account not only the correlation between proteins but their shared neighborhood, and then, again, that is what is used to cluster the proteins.

So to get into a bit more detail about sort of the network construction, you described the network as an $n \times n$ matrix with the number of nodes or genes, proteins, et cetera, and in our case, we use to describe the similarity, just a simple correlation, absolute value of the correlation, between a given node $i$ and $j$. The adjacency is then a measure of whether or how strongly the nodes are connected in the network. So the idea being that nodes that have very high correlations...
are particularly interesting.
Nodes that have moderate to low correlations are probably not informative.

is sort of the underlying idea, and if you look at this figure here, the correlation or similarity is on the x-axis, and then the adjacency is on the y, and so if you use

what’s called an unweighted network approach, you pick a threshold value, here, it’s 0.8, and you say that anything with a similarity less than 0.8 is considered to not be a connection in the network, and everything greater than 0.8 is considered to be a connection. So it’s sort of a binary yes or no.

What WGCNA does that was novel was to introduce a weighting where sort of the downside of this unweighted metric is that if you have a correlation of 0.79, that could be useful to know, but it counts as a zero.

So you’re losing information, and so what the weighted network does is it uses a sort of power transformation to get from sort of the straight correlation shown in this red line,

and sort of depending on this power value that you use,
you weight more or less towards the higher correlations in your network, and when you fit this model or when you sort of build the network, your choice of data is sort of one of the parameters that you choose going in, and there’s ways to sort of measure which gives the best fit to the data. So then once you have your sort of unweighted or weighted adjacency matrix, then is the part where you account for shared neighbors. So this is this topological overlap matrix that is created, so, basically, this measure omega of connectedness. The equation, I don’t find super sort of intuitive, but the components are... This is the sum, so u are, basically, all of the nodes other than i and j that you’re looking at the connectedness between, and so you’re summing up the sort of common connection strength between i and u and j and u as a product. So if I and J both have a strong connection to this other node, then that’s adding to this term, and then these k terms here are just the individual connections between, no,
each sort of the node i of interest and other nodes in the network, but I find sort of the easiest or most intuitive explanation from this original paper shows that for the unweighted case, omega is equal to one if the node with fewer connections has all of its neighbors, also, has connections of the other node. So the connections of node i are a subset of the connections of node j, and, also, i and j are directly connected. So that’s sort of the most interconnected that those two nodes can be, and then the least interconnected they can be if they are not connected to one another, and they don’t share any neighbors. So that would be sort of the zero case. So this a value can either take on the unweighted or the weighted case, and in our sample with WGCNA, we’re using those sort of weighted network connections that just adds more information into this topological overlap matrix. Okay. So, now you have the topological overlap matrix, again, this measure of sort of interconnectedness accounting for shared neighbors, then you can use hierarchical clustering
to divide those proteins into groups based on their similarity, and this is the results from our analysis.

So sort of on the x-axis, you have the different proteins, you have the dendrogram, which represents the hierarchical clustering of the topological overlap matrix, and then you have this dynamic tree cut algorithm which then defines these clusters which are shown in colors on the bottom based on the tree. You see this huge branch down here. That's gonna be this black cluster. There's this other cluster over here in green, and so there's, again, a few more parameters that you can use to decide how those cuts are made, and, in some cases, you can sort of merge branches that have correlation with one another, and my general advice is to try different values and see how robust the network is. It tended to be pretty consistent. In our case, where we saw four modules pretty much regardless. I think if we merged,
if we really cranked up one of the merging parameters,
we would get to three, but other than that it sort of stayed put.
Okay.

So the next step is trying to get a numerical summary measure of the groups of proteins that we’ve identified from our network.
So from these modules of co-expressed proteins,
we then use, basically, a principle components analysis to get what we call an eigenprotein or it was called an eigen gene in the original paper.

What it is is, essentially, a weighted sum of the values of each of the proteins in the module, and the weights correspond to sort of how well correlated that protein is with the overall module.

So if a protein has a high weight in the module, it means that it’s sort of the most interconnected in the module or sort of best represents the overall module.
So each person is going to have an eigenprotein value for each module,
and when we look at the sort of weights within each of the modules, so just to sort of orient us,
or eigenproteins, and then each sort of bar
on the y is a different protein.
In this case, we’re only including
proteins that fall into one of the four modules.
There were, also, if you notice on the last slide,
plenty of proteins that didn’t fall into any module
and were sort of the extras, so to speak,
and if you were to expand this down
and include more rows with those,
that would sort of show those, but for purposes of this,
we’re just including ones that fell into at least one of the four,
each of these bars represents a correlation
between the individual protein
and the overall eigenprotein.
So for these blocks of red,
it’s sort of the higher weighted proteins
that are within in this example module one,
module two, three, and four, and then you can see,
if you look sort of laterally from these proteins,
it’s the correlation of these proteins
with the other modules.
So the idea being we wanna see sort of blocks of red,
and then not a lot of correlation
between the blocks and other modules,
which is what we see.
All right, now that we’ve constructed our network,
and we’ve come up with numerical summary measures for each of the protein groups that we’ve identified, that is sort of the input or the predictor for these associations with outcomes. So for the MRI measures, which, again, our total brain volume, hippocampal volume, and white matter hyperintensities, we use just a simple or, you know, linear regression with covariates, and then a Cox proportional hazards regression, these eigenproteins are, they’re sort of one for each module. So we’ll run a separate regression analysis for modules one, two, three, and four. APOE is a gene that confers a lot of risk for Alzheimer’s disease. So it’s associated with the outcomes, and we include it as a covariate, and then a measure of time lag between when the blood was sampled and when the MRI was taken to account for any differences between people or the time difference, for dementia, it’s slightly simpler regression equation.
We only adjust for age, sex, and APOE status. All right, so next, I will show the results in the Framingham Heart Study. So from the four modules that we tested, there were two that we identified to have some association with outcomes. The first is module two. I gave it sort of a name clearance and synaptic maintenance, and I’ll talk about how I arrived at that name for the module in a bit. It has 165 proteins in it. Some of the half weighted proteins sort of give an idea of which ones are sort of most highly weighted or sort of most correlated with the eigen protein. I’ll talk about how we got to these in another slide as well, but, basically, this is from that over-representation analysis where you’re trying to identify biological pathways that are important or overrepresented by proteins in those modules. So we have the Axon guidance pathway. was most strongly associated with this module, and then in terms of relating to outcomes, total brain volume was the only significant association that we saw. So since this is a linear aggression,
effect greater than zero means a positive association.

So we see that for larger values of the eigenprotein for module two, we saw larger total brain volume. So it’s sort of a protective effect since brain atrophy is what is the risk factor for dementia.

We did not see a significant effect after correcting our p-values using a Bonferroni correction. You’ll notice that the confidence interval excludes one, which would be the null value, and that’s just because that’s based on the non-Bonferroni corrected value, but after testing for or adjusting for the four modules that we tested, we didn’t see a significant association.

It is nice at least that the direction of effect is what we would expect, based on our total brain volume association, which is that higher values of M2 correspond to sort of a lower incident dementia occurrence.

The second module that we found to be associated with total brain volume was this M4, which I will call sort of an inflammation-related module. It had 42 proteins in it.
The highlighted pathway there was cytokine-cytokine receptor interactions, so these sort of immune signaling molecules, and in this case, the association was in the opposite direction where higher values of this module for eigenprotein are associated with lower total brain volume. So it’s sort of a risk conferring module and, again, similar to what we saw here, not a significant, sort of an annoyingly borderline association between this and dementia, but, again, the direction of effect is what we would expect based on our observed association with brain volume.

I’ll just mention that I standardize the eigenprotein so that the effect sizes correspond to a standard deviation increase in eigenprotein. So it’s a little bit...

One sort of drawback I would say of these methods is the interpretation. Since a standard deviation increase, in this case, depends entirely on the sample that you’re using, it’s really just sort of a direction of effect more than anything. So to try and get at some of, get a better understanding of how these modules relate to our data or sort of what may be responsible
for some of the associations we see, this is a map of the correlations between different demographic variables and each of the modules, and I mentioned that we have a replication cohort as well, the CHS. So these two bars, sort of the two columns, show the two different cohorts that were included. So I put blue arrows to show the covariates that were included in our regression model, and you can see that there are some correlations between, say, sex and the modules, not really anything with APOE carrier status, maybe some education associations, and some associations with age. So it’s good that we adjusted for those in our models. However, you can also see there are a lot of other factors, cardiovascular risk factors, such as systolic blood pressure, BMI, fasting glucose that have associations with these modules. So we wanted to see if any of those could perhaps explain the associations that we saw. So I’m repeating sort of our standard model here was what I showed results from previously. The expanded model that we considered included a bunch of these risk factors,
basically, something representing BMI, hypertension, sort of lipid dysregulation, and diabetes, and I also included smoking as well, and we also included a measure of kidney function, which can also be an indicator of cardiovascular disease. So for module two, I'm repeating the sort of effects we saw from the standard model here, and when you adjust for the expanded set of covariates, your effect is attenuated by half, and it's no longer significantly associated. So with that says, it's either you have a sort of confounding issue where the association you're seeing between these proteins and total brain volume is really just in effect of sort of poor cardiovascular health or better cardiovascular health or you may think that it might be some sort of mediation effect where perhaps the risk associated between the proteins and the sort of total brain volume could be mediated by some poor cardiovascular health outcomes, and then for module four, again, this sort of inflammation module, we don't see any real effect attenuation. Regardless of whether you adjust
for cardiovascular factors or not, it’s still associated with total brain volume, which suggests it’s sort of different mechanism or lack of compounding between or based on cardiovascular health. Okay, so I mentioned in the sort of initial graphical abstract that once you find protein modules associated with your outcomes of interest, it can be good to look within the proteins of those modules to try and find sort of subsets or specific proteins that may be driving the associations. So for modules two and four, where we found associations with brain volume, we wanted to see if we removed proteins one at a time based on their sort of increasing weight, so remove the lowest weighted proteins in the modules first, what sort of happened to the strength of the associations. So these are both associations with total brain volume. It’s sort of the p-value on the y-axis, and you can see that as you remove, say, from module two, the first 20 proteins or so, you’re really not seeing a difference in the effect of the overall module with total brain volume, which suggests that those proteins
aren’t really impacting the association,
whereas beyond that point, once you start removing proteins,
the association becomes less strong,
and so that’s suggesting that those proteins may have more of an impact on sort of the overall module,
and so for both of these modules, we identified the spot where sort of the based on the lowest p-value, which proteins were sort of the most important in the module. I wanna emphasize that we didn’t use this to... So for things like dementia, if you were to run this, since we didn’t see a strong association or a significant association beforehand, we didn’t sort of use that to try and find a subset that we’re significantly associated because I would call that cheating. Okay, so the last piece that I’ll talk about in terms of teasing apart associations or sort of understanding protein within the modules is this functional enrichment or over-representation analysis within the modules. So based on the ones, sort of the significant modules or significantly associated modules with the outcomes, there is this software called STRING
that does a few different things, but what I used it for is doing an over-representation analysis of biological pathways.

So the idea is that there are annotation databases for proteins that sort of group them into biological functions or pathways that they’re involved in, and the idea is that if you have a module that has more proteins than you would expect from a given pathway, then that’s sort of the over-representation piece, and it indicates that that biological pathway might be important in whatever functions the module is carrying out. So this is just a screen grab of one example. So this is from module four. So you can see the annotation database is over on the left. So KEGG is one of them. Gene Ontology is another, and so you have these sort of observed proteins, and then the background is sort of the total number of proteins that are in the pathway, and the idea being that if you were to grab, I don’t know, however many proteins out of the background, like how many would you expect to be in this module.
due to chance, and do we have sort of over-representation compared to what we would expect?

And so for module four, the cytokine-cytokine receptor interaction was the strongest overrepresented pathway, and then you can sort of look at these others that have some sort of false discovery rate greater than 0.05, have some sort of false discovery rate greater than 0.05, and so I found the KEGG pathways, personally, to be the most informative. Gene Ontology tends to be a lot more specific, which may be more useful for targeting certain sort of therapeutic processes or something like that, but so depending on the scale that is important to you, you can sort of use different annotations. Okay, so the last thing I wanted to talk about, with the Framingham data in particular, was sort of getting back to our motivation for doing a network analysis in the first place. So the sort of contrast or comparator would be to do individual protein analyses where you’re running a regression model for each protein that you’re analyzing, and so we did that as a point of comparison. So for total brain volume, there were like a dozen proteins that were associated with total brain volume.
One was associated with hippocampal volume, and two were associated with Alzheimer’s disease at an FDR value of less than 0.1. So what was interesting, especially with the brain volume results, and, again, that was where we had seen associations with these modules, some of the proteins that were significantly associated were from module two and module four and others weren’t.

So what I get from that is a few things. One is that some proteins that are associated with the outcome are sort of individually associated but not sort of detectable within sort of a larger network of proteins that are associated with that outcome, and then the other is that for those that are within the modules, we would only be getting information about sort of a few of the proteins in the modules, whereas, as we see here, the associations tend or continue to get stronger with sort of looking at the broader network around sort of the most highly weighted proteins. So you’re getting a bit more information about proteins that may be associated with total brain volume.
and maybe at some of the biological processes compared to if you’re looking at things individually,
but, again, because you’re seeing associations that you don’t catch with the modules, it’s sort of important to look at both, and you get sort of complimentary information from the two approaches. So a caveat, I mentioned issues with lack of replication. We replicated this analysis in the Cardiovascular Health Study, and we did so by taking the same module, so module two and module four, taking the same weights from those proteins and applying them to the protein concentrations in the Cardiovascular Health Study. So we didn’t do a network reconstruction or anything in the different study. We were just seeing if these modules replicated in their associations with outcomes in a different cohort. So in this case, it’s really not seeing much in terms of association with both total brain volume and also looked at dementia out of interest since things were sort of close in our cohort, but, really, we’re not seeing much in terms of associations. Part of the reason for that,
so there are not that many cohorts that are available that have a large proteomic panel with the same proteins that we were looking at as well as MRI and incident dementia outcomes, and, in this case, the demographics of the cohort are fairly different from (indistinct) Framingham. So about 20 years older on average. I’m just including the sort of first few rows of our table one, but you can see differences in education, systolic blood pressure, and the same is true of a lot of the other cardiovascular risk factors. So it’s a very different cohort, and digging a bit into the literature about sort of proteins over the life course, it’s not too surprising that we don’t see the same associations, but it does sort of it’s a good cautionary message about drawing conclusions too far based on sort of one set of data or one set of demographics. Just to put these results in context, so our module four included a lot of immune-related signaling molecules like interleukins, TNF receptor proteins, which are both types of cytokines, and have been associated with Alzheimer’s disease previously,
in particular, interleukin-1 beta was in our module four, and it had been found to be elevated in 80 cases in a meta-analysis. However, other biomarkers that have been sort of validated in other cohorts were not identified in our module. In module two, we saw Axon guidance pathway proteins including ephrins, netrins, and semaphorins, which have been associated with AD in previous work, and complement cascades are also have been associated with AD probably for the reason of inducing these immune cells called microglia in the brain to, basically, eat up cells in response to amyloid deposition. So there’s some biologically plausible mechanisms that could be associated with these modules that could be associated with these modules in Alzheimer’s disease, and the last thing I’ll say is talking about some sort of other ways of approaching this problem, so as I mentioned, the CHS cohort has different underlying characteristics, and so it may well have a different network structure. So one thing that could be good to do is to look at sort of consensus modules across the cohorts where you construct networks in each cohort,
and then look at where the overlaps are, and you can get sort of a more, hopefully, more robust network across cohorts, and then there are other network-based approaches. That can incorporate external information. So, again, our network approach was just based on correlation in our dataset, whereas other methods use sort of those annotation databases and that sort of thing to construct the networks. So that’s another approach, and then the last thing I’ll say is that I’m sort of still using this kind of method now in work with longevity and aging and trying to apply it to metabolomics, so metabolites data in cohorts related to those outcomes. So thank you all for being here. Thank you, my collaborators. This is the folks down at UT. I’ll say that (indistinct). Thank you.

Thank you for wonderful presentation. We’re open for questions. So let’s start with people in the room. Any questions?
Got one over here. Perfect, thank you.

Yeah, so my research interest is about the cancer, and, also, we're interested in your study.

So I've got some technical issues about this project.

So the first issue that, how do you do the normalization in your process?

Yeah, great question. So yeah, I totally glossed over all the pre-processing stuff.

So before doing the network construction, I log transformed the protein concentrations to reduce stiffness.

I sort of standardized within those batches, and then after that, I did a rank normalized or inverse normal rank transformation to sort of-

(audience speaks indistinctly) What's that?

What's that?

Basically.

Yeah, yeah, yeah.

So that was sort of the data pre-processing.

So I think I, you know, I've thought about sort of the pros and cons of those things as well and I think my biggest qualm
with the way that I did it is sort of interpretability, because, yeah, sort of what does it mean to be at one quantile versus another where you have this huge dynamic range of protein concentrations?

Another question is that I know that in your project, the modules identification is very important. So I wonder, you have talked a little bit about how to answer the modules, but so can you explain a little bit more about how you gonna bring modules from the data?

I'm not sure, can you say a little bit more?

Yeah, so in your previous pages, I think you talked a little bit about the clustering of the modules so that we know that there are four main modules. So what is the name of that algorithm and how it basically work?

Yeah, so the clustering itself was done using algorithm called H+.

To be honest, I'm not too sure about sort of the details of it.
It can use any dissimilarity measure, which, in our case, comes from the TOM matrix, but-

So this is the algorithm that we separate the whole proteins into four different modules so that we can analyze it one by one.

Yeah, yeah, yeah, yeah. Yeah.

So I also noticed that in the weighted protein expression network analysis, you talk about the beta values. That you use like the soft threshold. Yeah.

To make the genes to be more important if that is the thing that you wanna analyze.

So in this process, I want to know how you would make sure the value of the data in this process.

So sorry, we have to end 'cause it’s 12:15.

I know others have classes and everything.

Maybe you guys can discuss a little bit.

Yeah, (indistinct), yeah.

Maybe if you have time.

Please, if you’re registered,

make sure you signed in on a sign in sheet.

There’s three of ’em.

You only have to sign on one of them.

and then one-fourth page reflections will be due.
before the next speaker’s time to speak.
(indistinct talking)