# Communicating Research to the General Public

**The WISL Award for Communicating PhD Research to the Public** launched in 2010, and since then over 100 Ph.D. degree recipients have successfully included a chapter in their Ph.D. thesis communicating their research to non-specialists. The goal is to explain the candidate's scholarly research and its significance—as well as their excitement for and journey through their area of study—to a wider audience that includes family members, friends, civic groups, newspaper reporters, program officers at appropriate funding agencies, state legislators, and members of the U.S. Congress.

WISL encourages the inclusion of such chapters in all Ph.D. theses everywhere, through the cooperation of PhD candidates, their mentors, and departments. WISL offers awards of $250 for UW-Madison Ph.D. candidates in science and engineering. Candidates from other institutions may participate, but are not eligible for the cash award. WISL strongly encourages other institutions to launch similar programs.

Wisconsin Initiative for Science Literacy

The dual mission of the Wisconsin Initiative for Science Literacy is to promote literacy in science, mathematics and technology among the general public and to attract future generations to careers in research, teaching and public service.

**Contact: Prof. Bassam Z. Shakhashiri**

**UW-Madison Department of Chemistry**

**bassam@chem.wisc.edu**

**www.scifun.org**

# Leveraging Mass Spectrometry to Probe Protein Post-Translational Modifications in Pancreatic Disease

By

Dylan Nicholas Tabang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: June 21st, 2023

The dissertation is approved by the following members of the Final Oral Committee:

Lingjun Li, Professor, Pharmaceutical Sciences and Chemistry

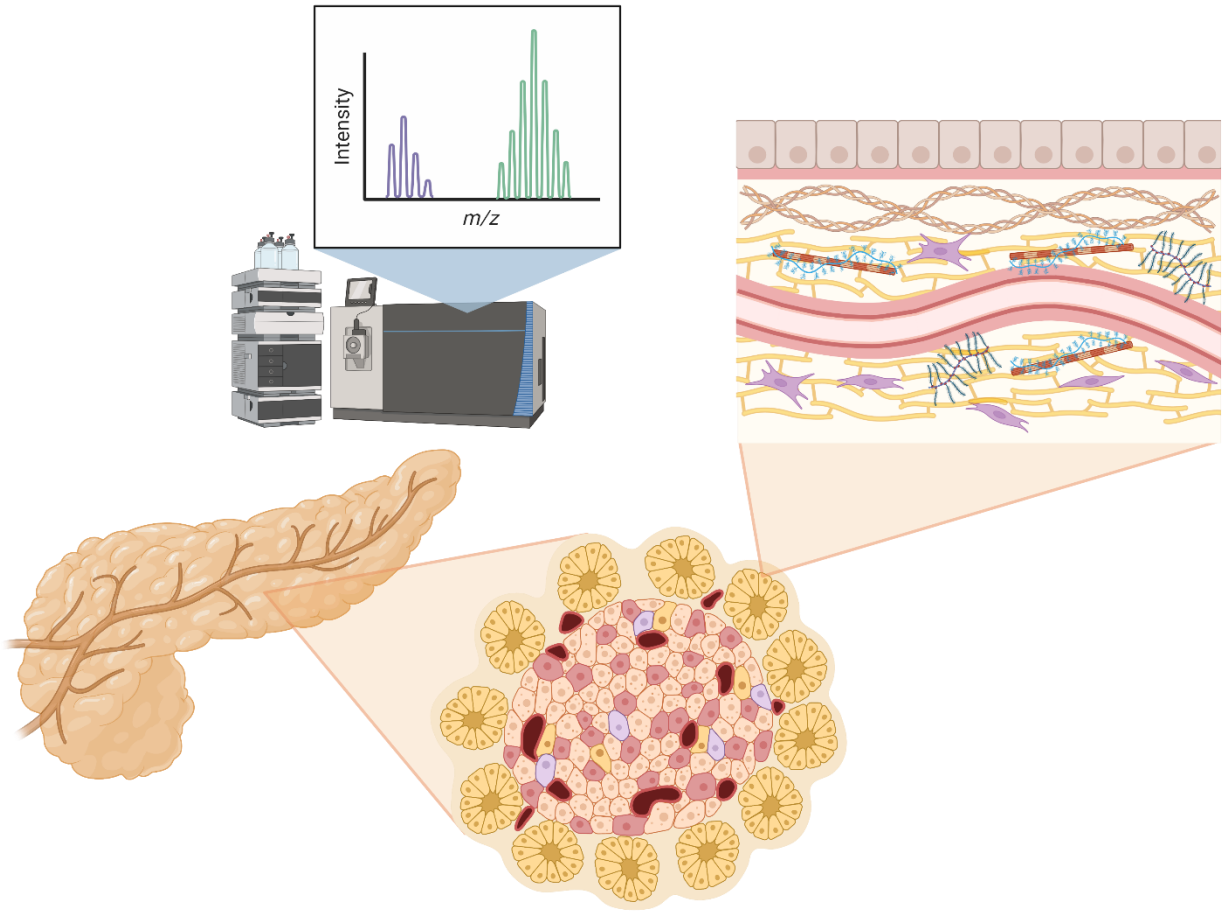
Ying Ge, Professor, Cell and Regenerative Biology and Chemistry

Kevin W. Eliceiri, Professor, Medical Physics and Biomedical Engineering

Jon S. Odorico, Professor, Surgery

# Chapter 2

*What is Written Is Not Always What is Printed: Using Mass Spectrometry to Identify Protein Post-Translational Modifications for Pancreatic Disease Research*




Written by Dylan Tabang (Lingjun Li Research Group) in collaboration with the Wisconsin Initiative for Science Literacy in order to describe this thesis for a broader audience.

## Preface

Having done the bulk of my graduate studies in bioanalytical chemistry during the COVID-19 pandemic, I could not appreciate more the efforts to communicate science to the public. The public health crisis brought the efforts of both “bench” scientists and clinicians to the forefront of public attention on a global stage. This much exposure to the inner workings of the scientific method also opened the door to misinformation about the scientific process. I have gained a better understanding that science is more a journey towards a better understanding of our world than arriving at concrete truth as if it were a destination.

I use this opportunity to summarize some of my dissertation research for the public, in my own words, with pride in the work that I have done. It is my hope that my contributions to the field can inspire new lines of research towards improving the lives of those who suffer from pancreatic disease. Furthermore, I hope that future scientists can likewise make their work more accessible to the public, who support many research efforts through their tax contributions toward federal grants. I applaud the work of and thank the Wisconsin Initiative for Science Literacy for their efforts to help do so.

## Introduction

Molecular biology studies cells, the basic component of life, and the molecules that comprise them. Cells make up tissues, organs, and at the highest organizational level, organisms. Everything about organisms is dictated by the “central dogma” of molecular biology, where genetic information encoded as DNA is transcribed and sent throughout cells in the form of RNA. RNA information is translated into protein information through the sequence of amino acids that comprise the protein. Though the exact sequence of a protein can be predicted by RNA

and DNA information, biological reality is not as simple. What is written in the instructions of the DNA sequence is not always what is printed in the protein sequence.

A protein's amino acid sequence is key to understanding its structure and the jobs it can do. There are twenty common amino acids which vary widely in their properties. Different sequences of the same amino acids lead to vastly different protein properties and functions. For example, hemoglobin is a key human protein, made of four subunits, that transports oxygen from the lungs through the body. Scrambling the amino acids that comprise hemoglobin into a different sequence prevents the four subunits from joining and removes its ability to transport oxygen. This is similar to taking apart a truck made of Lego blocks and using the same pieces to make something new. From the pieces used to make a four-wheeled vehicle, for example, one could perhaps also make a motorcycle. Though the constituent parts are the same, the overall structure and function of the build will be different. Beyond the twenty common and natural amino acids, other processes can also change protein properties.

After proteins are translated and synthesized inside of cells, they can still be modified. These post-translational modifications (PTMs) can be essential to the actual functions performed by proteins that are only made possible after the stitching together of amino acids. One major category of PTMs is the addition of chemical groups to specific amino acids. These chemical groups can range in size from small, like adding several atoms' worth of mass, to large, as in the addition of another protein (several hundred atoms) to a protein. Again, the exact chemical nature of the added group is key to understanding how it modifies proteins. Phosphorylation (addition of $-PO_3$), for example, adopts a negative charge in the conditions of the human body. This switch from neutral to negatively charged can help some proteins bind to other proteins in different signaling pathways, modifying their functions as well. Phosphorylation is reversible

and dynamic, making it an important PTM for regulating different pathways and biological processes. Like phosphorylation, glycosylation is another PTM that introduces a chemical group. Glycosylation, however, is a broad term that encompasses the addition of any glycan, which is composed of at least one monosaccharide unit. Glycans can be composed of several different types of monosaccharide units, each with their own structures and properties. This modification, therefore, is complex due to the number of possible structures that can be added to proteins in different combinations. Glycosylation covers proteins and cells just like trees covering hills; there are many different types of trees with their own appearances, like redwoods and palm trees, but each of these things is still a tree. In general, glycosylation adds to a protein a mass more than that of a phosphate group but less than the addition of another protein.

Another type of PTM is proteolysis, or cleavage of a long protein sequence into shorter parts called peptides, which on their own can also be functional. One important example of this is the synthesis of insulin, an important molecule in diabetes that is made by the pancreas. First, the ends of the insulin precursor protein are removed, leaving a peptide structure which is linked to itself in a loop. This structure is then cut again, where the middle of the structure (called the C-peptide) is removed. The end of the processing creates functional insulin, which consists of two peptide chains linked together. Tests that measure the C-peptide can tell how well someone is producing insulin, which is important for patients with diabetes.

Post-translational modification is essential in processing proteins towards their functional forms. While cleavage is not reversible, modification with a chemical group often is reversible. Understanding dynamic modifications of proteins and the changes they impart is key to understanding human health and biology. Only in approximately the last thirty years have we had the ability to probe protein sequences and their structures down to the atomic level. Moving

beyond the Human Genome Project in the 1990s into this century's Human Proteome Project, our understanding of proteins has improved vastly in the last decade. As we have learned, the genetic instructions does not transfer perfectly to the resulting protein information. Various forms of the same protein can result from the same genetic sequence. These protein variations, called proteoforms (**Figure 1**), result from the same set of genetic instructions, but can carry different sequences and functions. Changes in the transcribed RNA may lead to different amino acid substitutions when a protein is synthesized, for example.

Protein post-translational modification, however, is the largest contributor to the variety of proteoforms that have been observed. The Human Genome Project found that humans carry approximately 20,000 genes. This number pales in comparison to the number of observed proteoforms, of which over 6,000,000 have been identified. The number of proteoforms observed continues to increase with advances in how we measure them and in what systems we are measuring. As proteins are the main drivers of action within cells, studying proteoforms in human health and disease continues to be an important area of research.

**Mass spectrometry-based proteomics**

Proteins are made of amino acids which are in turn made up of atoms. Each of these things has mass, which relates to its respective amount of matter. Proteins, amino acids, and atoms can also carry an electric charge, becoming an ion, based on their relative number of positively charged protons and negatively charged electrons. Mass spectrometry (MS) is an analytical technique to measure an ion's mass-to-charge ratio, abbreviated *m/z* (calculated as mass divided by charge). The 2002 Nobel Prize in Chemistry was awarded in part to John Fenn for advancing the field of MS, highlighting the transformative capacity of this field in the current century. Knowing the information gained using MS, along with how an ion can fragment into its

components in the stage of measurement known as tandem MS, can help identify unknown molecules. Fenn is credited with the development of electrospray ionization, in which molecules dissolved in solution are vaporized after application of high voltage. This ionization keeps molecules mostly intact before they enter the instrument for further tandem MS experiments which can be rigorously controlled.

One issue with MS-based experiments of proteins is instrumental limitations. For large proteins comprising thousands of amino acids and thus weighing hundreds of thousands of Daltons (kDa), there is a large distribution of charge. This decreases the signal for a given protein and distributes it over several peaks that are close to one another. Depending on the resolution of the mass spectrometer, these signals may merge into a single peak, which can make determining the exact *m/z* difficult. This concept is illustrated in **Figure 2**. Just as a blurry camera can reduce a photo of a complex scene into messy blobs, a low-resolution MS measurement can fail to capture all the details of a complex mixture of proteins. Instrumental advances over the last twenty years have improved resolution to a point where this is no longer an issue, though analyzing intact proteins today is a growing area of research.

To combat instrumental limitations, the field has mostly adopted a “bottom-up” approach, where proteins isolated from a biological sample, like biofluids or tissues, are cleaved into their component peptides. These peptides, which now contain far fewer amino acids than their precursor proteins, are then analyzed using MS. A diagram of the bottom-up proteomics workflow can be seen in **Figure 3**. The enzyme trypsin, a protein involved in digestion, cuts up proteins into easily managed and measured peptides. Though the approach of breaking up the protein puzzle into its component peptide puzzle pieces to then have to infer the composition of the whole puzzle seems counterintuitive, it is currently the best paradigm for MS-based protein

analysis. Another way to put this is trying to identify an animal only given pictures of its ears, feet, tail, and fur. These things may be easy to identify in isolation, though identifying the original animal where each of these components came from is much more difficult. The resulting complex mixture of peptides far outnumbers the original number of proteins isolated from a sample. A liquid chromatography separation step, where peptides are separated based on their properties like charge or their interactions with water, is often included before MS analysis to improve the coverage of peptides measured and identified. Peptide fragmentation in tandem MS can be predicted. Fragment peak distributions reveal the sequence of the peptide and can also uncover the existence of PTMs and where they reside on a peptide, as seen in **Figure 4**.

Modern MS methods have developed over the last three decades towards clinical use. MS remains the gold-standard method for the identification of unknown compounds. Using compound identification, MS can help identify unknown microbe pathogens from patient samples to help diagnose and treat infections. MS analysis can also be performed on small amounts of samples, down to several microliters – there are one million microliters in one liter – of blood released in a finger prick test. One challenge for MS achieving wide-spread use in the clinic, however, is the issue of throughput. MS analyses often require an hours-long separation which is not viable for analyzing hundreds to thousands of patient samples per day. Nevertheless, MS analyses have been helpful for improving diagnostic capabilities and for identifying biomarkers for diseases. MS also outperforms immunoassays, which use antibodies that specifically target and help detect certain compounds and are also commonly used in the clinic, for some analyses. Therapeutic drugs can readily be detected, as well as vitamin D and other important molecules in metabolic disorders. The latest work in clinical MS continues to improve sample throughput and make MS more attractive towards widespread use.

**Studying post-translational modifications in the pancreatic extracellular matrix**

Besides the strength of MS in analyzing biofluids like blood, it is also a major technique for understanding the components of cells and tissues. MS can paint a picture of a sample's protein composition even with just micrograms of protein. This is a typical amount extracted from milligrams of cells and tissues. Understanding how protein composition is different among different cell types or disease conditions is a key step in disease research. This guiding principle led me towards studying PTMs in the context of pancreatic disease research.

The pancreas is a key organ for human life. Located in the gut, it is responsible for the dual roles of digestion and hormone production. Most of the cells in the pancreas synthesize the digestive enzymes that break down our food into nutrients. These enzymes travel from the pancreas through ducts into the gastrointestinal tract where they can then digest food. The cells that produce digestive enzymes and the ducts used for transport comprise most of the pancreas (approximately 97%). The rest of the organ is devoted to hormone production. Pancreatic cell clusters called islets of Langerhans produce hormones that are crucial to bodily regulation. The two most important hormones are insulin and glucagon, which are synthesized in special cell types respectively called alpha and beta cells. Insulin and glucagon are hormones that regulate how much sugar, specifically glucose, remains in the blood. When this system malfunctions, too much glucose remains in the blood, leading to diabetes and further complications.

Treatment for diabetes differs based on the cause of high blood sugar. In type 1 diabetes, beta cells that produce insulin die and insulin production ceases. Patients with type 2 diabetes, on the other hand, may still have insulin present though blood sugar remains high. Type 1 diabetes patients become dependent on receiving insulin from outside the body, though procedures involving islet transplants to regain insulin self-sufficiency are becoming more widespread. Islets

must first be isolated from another patient. Making sure that these transplanted islets can survive in their new environment and actually produce insulin continues to be a challenge, however.

I have sought to use MS to study the pancreas and pancreatic disease through my collaboration with the University of Wisconsin School of Medicine and Public Health. Protein compositions differ among all organs and are affected by age, gender, and health status. I first wanted to establish a baseline composition of pancreas proteins and their PTMs in healthy patients. This information will be helpful for future work in engineering islet scaffolds and mimics grown in a lab from stem cells, which must have similar protein compositions to pancreas from patients to function correctly. Our research labs also wanted to gain a better understanding of the pancreas extracellular matrix (ECM), which consists of everything in the space outside of cells. This space contains a complex network of proteins which help cells communicate and adhere to one another. More importantly, the ECM is key to transporting molecules between cell types which we believe are essential for islet survival and proper function. A diagram of the pancreas, islets of Langerhans, and the ECM is depicted in **Figure 5**.

To accomplish our goal of getting a better picture of the pancreatic ECM and the PTMs therein, I used a bottom-up proteomics approach that is summarized in **Figure 6**. I obtained pieces of pancreas tissues from deceased donors that were not immediately used for organ transplantation and instead diverted for research use. Proteins from the tissues were extracted using the detergent sodium dodecyl sulfate (SDS), a common ingredient in dish soap. SDS is an important part of soap because it can dissolve the membranes that keep bacteria and viruses intact. In the same way, SDS can help rip apart cells in tissues, releasing their protein contents. Though SDS is helpful in protein extraction, it must be removed prior to MS analysis because it can mask signals from the peptides of interest. I removed SDS by using a mixture of ethanol,

acetone, and acid. This mixture causes proteins to pellet into a solid precipitate while SDS remains dissolved. Once the pellet is washed several times, I then proceed to protein digestion. Digestion is done by adding the enzyme trypsin to the liquid sample and letting the reaction proceed overnight. I then purify the peptides using a specialized separation material.

Peptides containing PTMs make up only a small portion of the total peptides, making MS analyses difficult. To overcome this challenge, we can use the properties imparted by the chemical modifications on the peptides to selectively separate them from their non-modified counterparts. Phosphorylation introduces an extra negative charge to peptides. Using a positively charged material can enhance the detection of phosphorylated peptides. Another PTM that plays a major role in the ECM is glycosylation. Though there are many possible structures for glycosylation, they in general tend to make peptides more hydrophilic, which increases its tendency to dissolve in water. Using a material that can separate water from other liquids in the sample solution can enhance the detection of glycosylated peptides. The strategy I employed takes advantage of the properties of both phosphorylated and glycosylated peptides to enhance their detection. Using a strategy like this increases how much we can learn about the proteins in a sample using one workflow instead of separate steps for either modification. This is an important consideration when working with small sample amounts, like from biopsies. Now that peptides with PTMs are enriched, they are ready for MS analysis. A high voltage ionizes the samples, which are then analyzed by the mass spectrometer. I then upload all the thousands of mass spectra into a data analysis program. This software assigns spectra to a sequence and the modification sites based on how different sequences are predicted to fragment. The software finally matches identified peptides to their precursor proteins as they existed in the sample.

Analyzing the proteins which are glycosylated and phosphorylated shows the importance

of these modifications in the ECM. This is demonstrated in **Figure 7**, which depicts numbers of ECM proteins and modified proteins identified by the software. Furthermore, the bubble plot depicts levels of co-modification (proteins containing both glycosylation and phosphorylation) of proteins identified, the majority of which are ECM proteins. Both modifications play roles in mediating and stabilizing protein structures, which is evident in their presence on collagen proteins. These proteins adopt a rigid fibrous, brush-like structure which provides resistance to stretching. This is important for maintaining the structural integrity of pancreas tissue. Collagen proteins identified also contained certain modification sites where many types of glycosylation were seen at the same time. These various proteoforms would be slightly different in their overall structure, leading to slightly different roles overall. Type V collagen, one of many types of collagens each with their own specific roles in the body, was shown in this work to have both glycosylation and phosphorylation. Previous studies in the literature showed that this specific collagen is involved in insulin secretion and transport, suggesting a major role of this protein in the pancreas.

The findings of this work are interesting in that we now have a better picture of modifications of these proteins as they exist in the healthy pancreas. Any changes to these protein compositions during disease progression may shed light on new biomarkers or proteins to target with new therapeutics. Based on the importance of these modifications as shown in our data, I anticipate future proteins of interest will also be part of the ECM.

**Research impact and outlook**

Pancreatic disease affects millions of people each year. **Figure 8** depicts two of the most common pancreatic diseases, namely diabetes and pancreatic cancer. Diabetes involves malfunction of insulin regulation of glucose, which cannot enter cells. Untreated diabetes may

lead to fatal complications, including heart disease, stroke, and kidney disease. Type 1 diabetic patients must administer insulin once their own islets stop producing it, though type 2 diabetic patients may still have insulin activity but require medication to control blood glucose. In pancreatic ductal adenocarcinoma (PDAC), uncontrolled cell growth leads to tumor growth in the pancreatic ducts. Eventually, the tumor starts to metastasize and spread to other organs. Commonly, pancreatic tumors spread to the liver and lungs. The five-year survival rate for PDAC still remains around 10% due to the difficulty in diagnosing the disease at a stage where surgery and chemotherapy are most effective.

The impact of pancreatic disease is immense. Bioanalytical chemistry, and especially MS, plays a vital role in improving our understanding of how pancreatic diseases progress. Basic research that I perform in a laboratory is performed with the hope of advancing the field for applied research done in the clinic. One major area of research is biomarker discovery. The best biomarkers are easily detected in an easily collected sample, like blood, and are specific to a disease. Blood biomarkers for pre-diabetes can help patients determine whether they need lifestyle changes to prevent diabetes onset.

For pancreas cancer, early detection and treatment is essential to improve patient survival. Current research has focused on glycan biomarkers for PDAC, though individual differences in cancer among patients makes the discovery process difficult. MS techniques have also made their way into the operating room to help differentiate cancerous tissue from non-cancerous tissue. This extra step can ensure that surgeons remove all cancer during surgery to prevent it from recurring.

MS has helped improve our understanding of human health and disease immensely in the past three decades. Though earlier work during the Human Genome Project has cracked the

genetic code and how its instructions manifest as observable characteristics, biology is not so simple. We have instead seen how the proteins coded by genes are changed through various processes, including post-translational modification. There is perhaps no better tool to study proteins and their PTMs than MS. As new measurement methods develop, our knowledge of diseases and how to treat them will continue to grow. It is my hope that my contributions to pancreas research will advance our understanding towards developing new treatments for improving the quality of life for all those who suffer from pancreatic disease.

**Acknowledgments**

Research described here was supported in part by grant funding from the National Institutes of Health (R21AI126419, R01DK071801, RF1AG052324, U01CA231081, and 1F31DK125021-01), and Juvenile Diabetes Research Foundation (1-PNF-2016-250-S-B and SRA-2016-168-S-B). Data discussed here were also in part obtained through support from a National Center for Advancing Translational Sciences UL1TR002373 award through the University of Wisconsin Institute for Clinical and Translational Research. Figures in this chapter were in part created using Biorender. I would also like to acknowledge the generous support of the University of Wisconsin Organ and Tissue Donation Organization who provided human pancreas for research. Our research team would like to give special thanks to the families who donated tissues for this study.

**Further reading**

Banerjee, S., Empowering Clinical Diagnostics with Mass Spectrometry. *ACS Omega* **2020,** *5* (5), 2041-2048.

Smith, L. M.; Kelleher, N. L., Proteoform: a single term describing protein complexity. *Nature*

*Methods* **2013,** *10* (3), 186-7.

Tabang, D. N.; Cui, Y.; Tremmel, D. M.; Ford, M.; Li, Z.; Sackett, S. D.; Odorico, J. S.; Li, L., Analysis of pancreatic extracellular matrix protein post-translational modifications via electrostatic repulsion-hydrophilic interaction chromatography coupled with mass spectrometry. *Molecular Omics* **2021,** *17* (5), 652-664.

Tabang, D. N.; Ford, M.; Li, L., Recent Advances in Mass Spectrometry-Based Glycomic and Glycoproteomic Studies of Pancreatic Diseases. *Frontiers in Chemistry* **2021,** *9* (579).

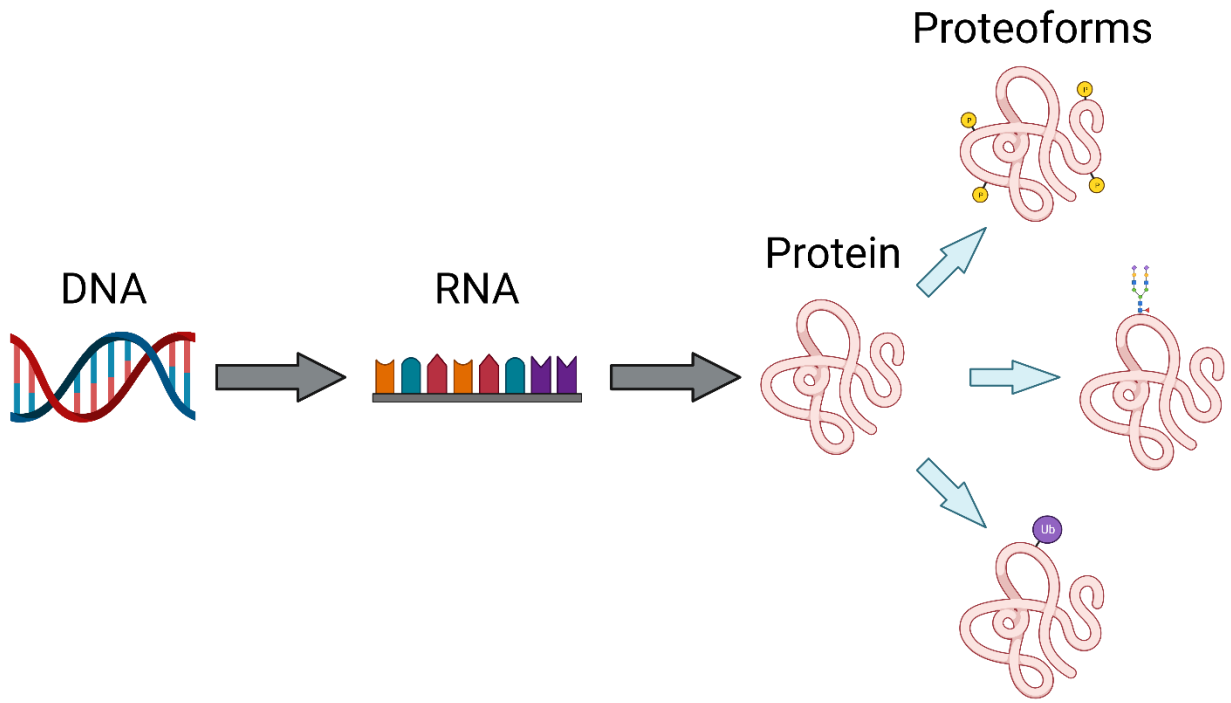



**Figure 1**: The central dogma of molecular biology states that our genetic information encoded in DNA is transcribed into RNA which is read as instructions for synthesizing the sum of all our proteins. This process, however, is not as simple as the written instructions perfectly being followed into the resulting proteins. Instead, many various proteoforms can result from the same genetic instructions though different processes, including post-translational modification.

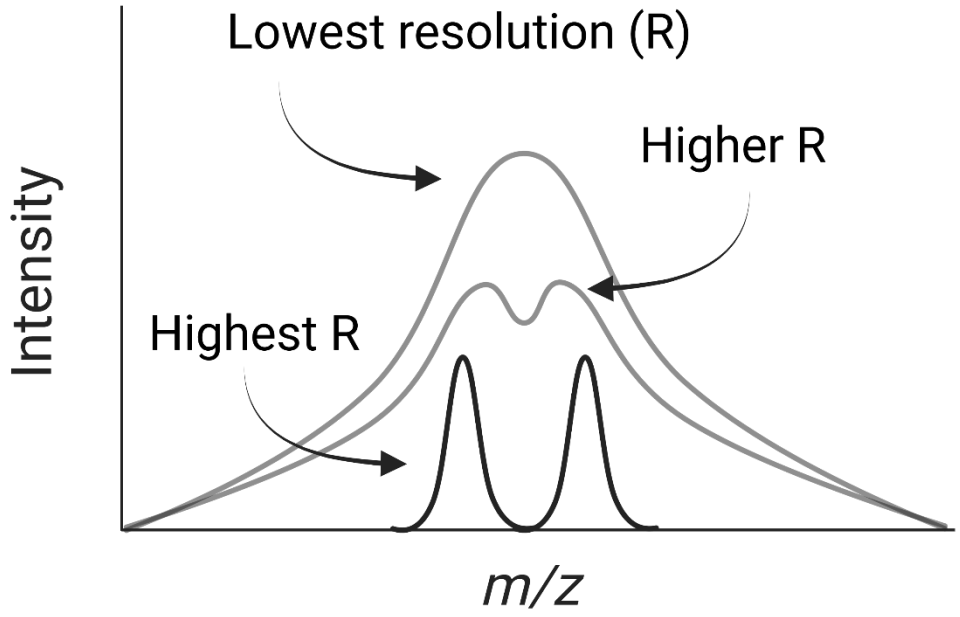



**Figure 2**: Illustration of resolution (abbreviated as R) and its effect on peaks in a mass spectrum. Higher resolution reveals more peak details. At the highest resolution shown, the two peaks are separated at their bases. At the lowest resolution, these peaks cannot be resolved and are thus merged into one peak.

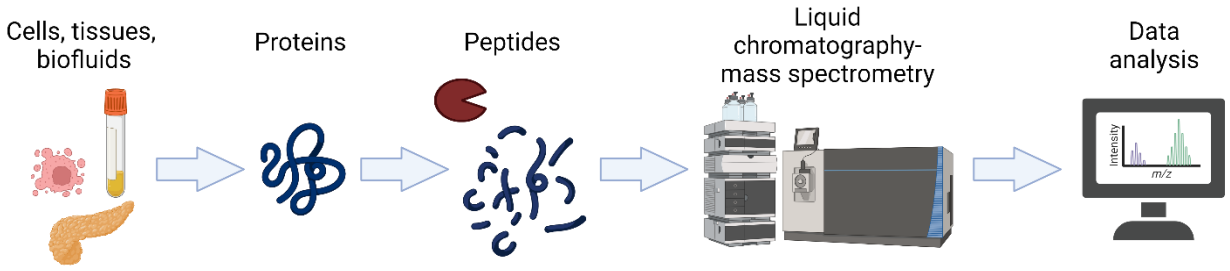



**Figure 3**: The "bottom-up" proteomics workflow involves extracting proteins from various sources and cutting them up into more easily measured peptides. These peptides are separated based on their properties using liquid chromatography and measured using mass spectrometry. Data analysis includes determining the peptides' identities and their original proteins and comparing the amounts of peptides among different samples and groups.

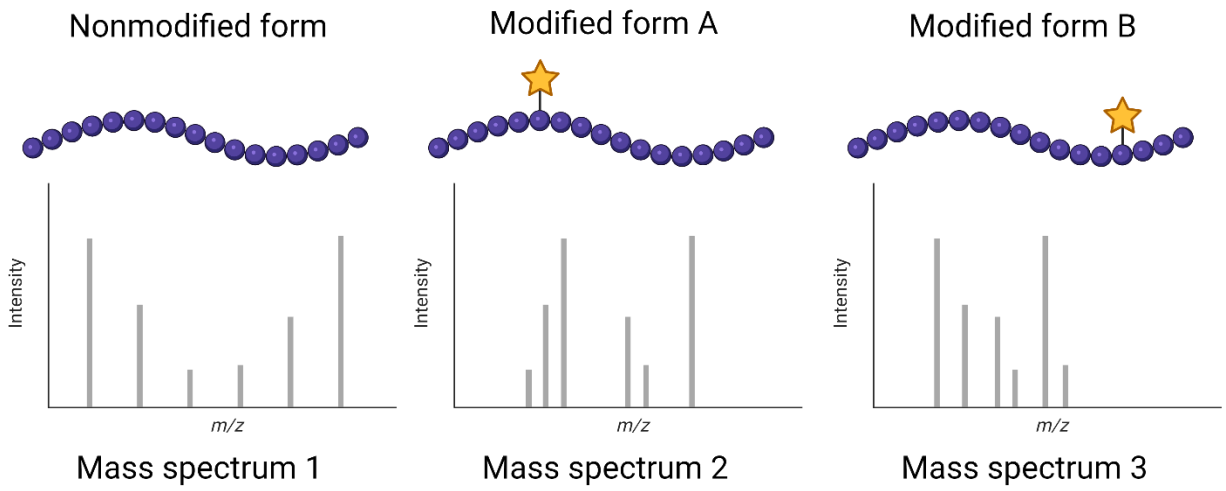



**Figure 4**: The resulting mass spectra of different modified forms of peptides appear different from one another and can be used to distinguish the same modification but on different sites. Fragmentation behavior can differ widely based on the properties of the modification. Generally, the more fragmentation information that can be determined from a mass spectrum, the more confident the assignment of modified site.

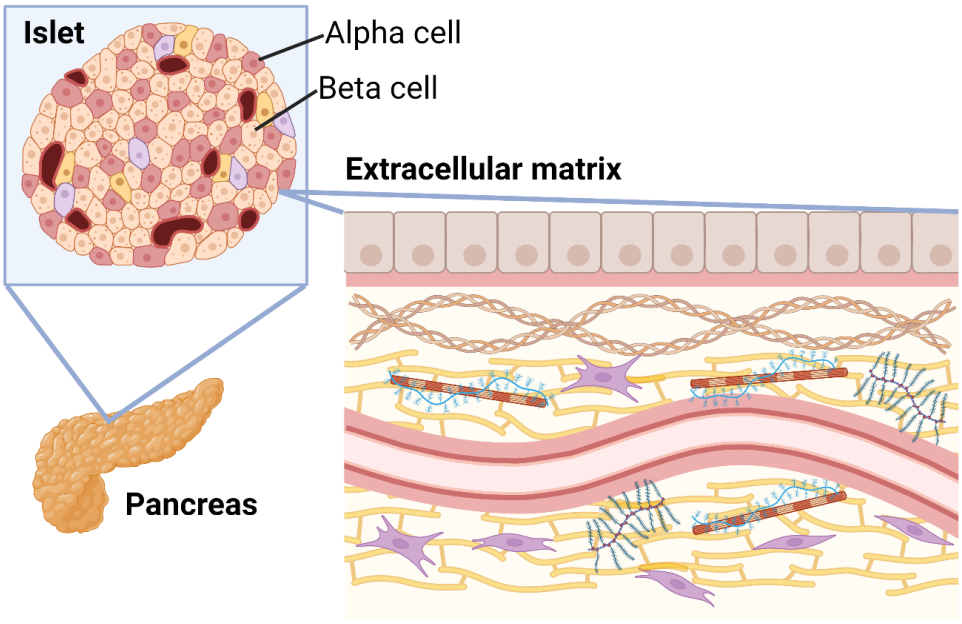



**Figure 5**: The pancreas is an essential organ with two main roles, one of which is producing hormones in the islets which regulate bodily processes, including blood sugar regulation by the hormones insulin and glucagon, produced by beta cells and alpha cells, respectively. Pancreatic islets are small cell clusters dotted throughout the pancreas but only comprise up to 3% of the organ. The space outside of cells is called the extracellular matrix, a network of molecules important to structure and regulation throughout the organ.

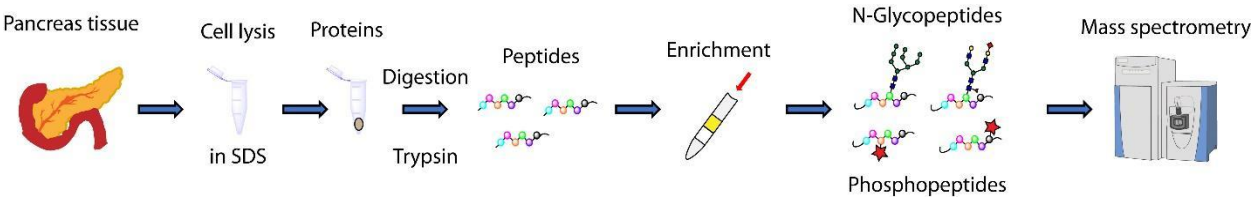



**Figure 6:** Bottom-up, or enzyme digestion-assisted, proteomics workflow. Pancreas proteins are extracted after lysis in the detergent SDS, which releases all of the cells' contents. Proteins are digested into peptides with the enzyme trypsin. These peptides undergo an enrichment which enhances detection of glycopeptides and phosphopeptides. Peptides are analyzed using a mass spectrometer. The resulting data identifies proteins from the sample.

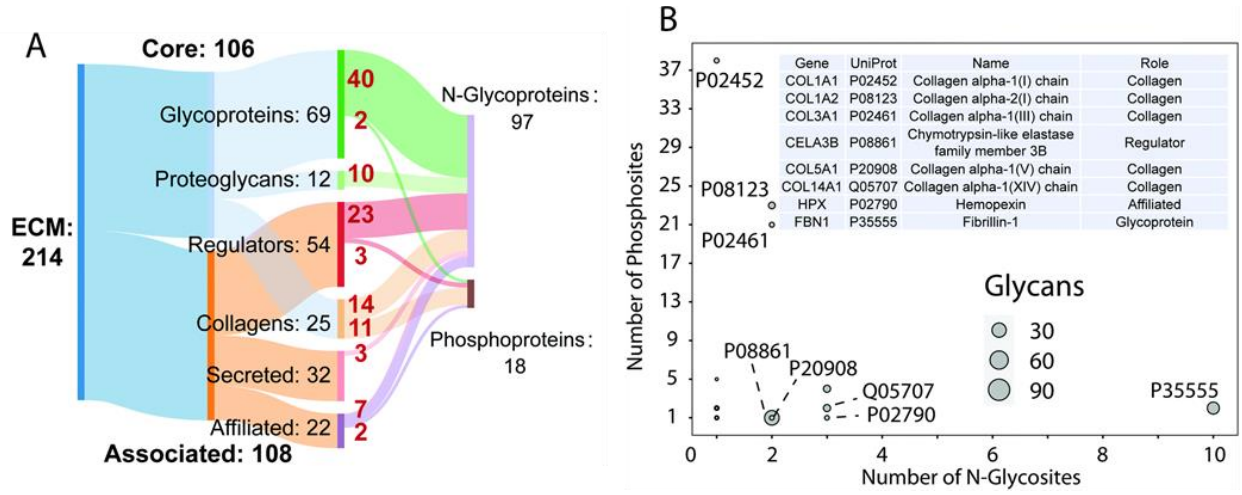



**Figure 7**: In a mass spectrometry-based proteomics analysis of the pancreatic extracellular matrix, we sought to identify protein glycosylation and phosphorylation. These two modifications are the most common biologically and play important roles in cellular signaling and structure. These modifications were identified in all categories of extracellular matrix proteins (panel A). Furthermore, these modifications can also occur on the same proteins (panel B). The most co-modified proteins are structural proteins in the extracellular matrix.

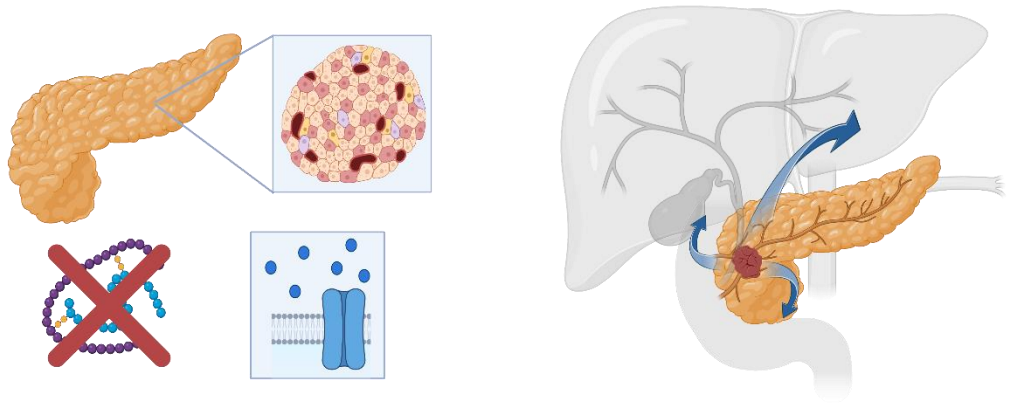


**Figure 8**: Two important pancreatic diseases are diabetes (left) and cancer (right). In diabetes, the pancreas cannot regulate blood sugar correctly. Insulin is either not present or cannot regulate how much glucose is taken up by cells, leading to more glucose staying outside of cells, causing various complications like heart and kidney disease. In cancer, uncontrolled cell growth leads to tumors. Most pancreatic cancers originate in the ducts and can spread throughout the organ, potentially leading to diabetes, and to other organs, like the liver and lungs. Both cancer and complications from diabetes can be fatal.