Modern Computer Science Approaches in Biology: From Predicting Molecular Functions to Modeling Protein Structure

by
Oleksandr Narykov

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Computer Science
April 2022

APPROVED:

Professor Dmitry Korkin
Primary Advisor
Worcester Polytechnic Institute

Professor Randy C. Paffenroth
Committee Member
Worcester Polytechnic Institute

Professor Carolina Ruiz
Committee Member
Worcester Polytechnic Institute

Professor Gloria Sheynkman
Committee Member
University of Virginia

Professor Craig E Wills
Department Head
Worcester Polytechnic Institute

"So once you do know what the question actually is, you'll know what the answer means."

- Douglas Adams, *The Hitchhikers Guide to the Galaxy*

ABSTRACT

Computational machines have become an inseparable part of human lives during the last three decades. One of the crucial enabling technologies of this technological boom is Artificial Intelligence (AI), the field dedicated to simulating human-like behavior in machines. It takes many shapes and forms; however, a particular direction – Machine Learning (ML) – was incredibly impactful in the era of constant data aggregation. The goal of ML is an automated pattern inference and reasoning based solely on the input data. Becoming a household name, machine learning completely revolutionized natural sciences, providing aid to the physicists working on quantum mechanics, helping astronomers filter noisy data, as well as accelerating molecular and cellular discoveries made by chemists and biologists. One of the crucial aspects of everyone's lives affected by ML technology is the medical care. Perhaps most notable in this area, precision medicine provides the direct opportunity to improve patients' quality of life directly.

The field of precision medicine is dedicated to identifying reasons for different treatment responses from patients and designing the best-suited diagnostics and intervention strategy for each individual. In recent years, the available data pool was expanded by the emergence of high-throughput 'omics' experimental technics, making it intractable for conventional manual analysis by a clinician or a biomedical researcher. The omics field emerged in earlier 2000s when next-generation sequencing (NGS) methods that made studying individual genomes possible first emerged. The next big breakthrough happened in 2008, when the second generation of NGS came into play, drastically decreasing the costs of conducting experiments. However, genomics is not the only field that experienced the revolutionary leap. Other quantitative methods that

describe molecular processes taking place in the organism advanced rapidly: epigenomics, transcriptomics, proteomics, and metabolomics. Transcriptomics and proteomics are particularly interesting when studying diseases as they are providing a snapshot of the organism's current state, allowing us to search for the root cause of a particular ailment. Furthermore, transcriptomics provides information on an important regulatory process--alternative splicing (AS). AS increases the versatility of the organism's molecular arsenal and allows to build more complex systems using the same number of genes. This feat is achieved via combinatorically shuffling selected protein coding parts – exons – from the mRNA molecule prior its transformation into a protein. Thus, AS is a crucial intermediate stage between the gene expression and protein translation.

My work focuses on the computational analysis of biological data and encompasses structural genomics, transcriptomics, and proteomics. Individual projects range from elucidating disease etiology and uncovering molecular mechanisms of actions of the alternative splicing to searching for the protein expression-based treatment response biomarkers and studying the potential drug targets on the SARS-CoV-2 viral particle surface. Over the course of these studies I designed a machine learning model that estimates the AS effect on protein-protein interactions; developed a novel quantitative measure that gauges an impact that the alternatively spliced isoforms introduce to the biological system; predicted isoform stability using proteogenomic data and transfer learning; identified response biomarkers for the Gulf War veterans affected by one of the most complex known acquired syndromes for the acupuncture treatment; modeled protein complexes of SARS-CoV-2 virus and simulated its entire envelope in solvent using molecular dynamics methods.

This work brings together two important aspects of modern omics studies – transcriptomics and proteomics. It highlights an importance of computational methods development for the modern field of precision medicine.

ACKNOWLEDGMENTS

I extend my gratitude to the Ph.D. advisor Dr. Dmitry Korkin that provided me with guidance all these years. I thank all my committee members, Drs. Carolina Ruiz, Randy C. Paffenroth, and Gloria Sheynkman for sharp insights into research projects.

I would like to thank all lab members that I met during my studies – Katie Hughes, Nathan Johnson, Andi Dhroso, Hongzhu Cui, Suhas Srinivasan, Anastasia Leshchyk, Pavel Terentiev, Ziyang Gao, Senbao Lu, Nan Hu, Huaming Sun, Winnie Mkandawire, Kathryn Monopoli, Jocelyn Tourtelotte. You all are great friends and insightful colleagues. And, of course, I am grateful to all collaborators I encountered throughout these years – Efi Kokkotou, Siewert J. Marrink, Ben Jordan, Jamie Saquing, Weria Pezeshkian, Tsjerk Wassenaar, and Fabian Grünewald.

I thank my parents for their unconditional support and love. Their efforts to bring me up were unparalleled, and I appreciate them. I extend dedication to them during these difficult times when war came into our country. I am deeply grateful to all people that accompanied me through my journey during these challenging years and to all those who continue to support Ukraine and its defenders. I extend special ‘thank you’ to my bosom friends Dmytro Bogatov and Daria Bogatova, as I shared a lot of both happy and sad memories with this wonderful family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Data-Driven Precision Medicine as a New Paradigm in the Healthcare

1.1 The Emergence of Precision Medicine: History and Basic Challenges

Computational machines found their way into our daily lives during the last three decades and have become ubiquitous. Starting from a Personal Computer (PC) and going into the era of wearable devices and mobile gadgets, the role of automatization became more and more prominent in daily routines. People constantly use it to engage in remote communications, translate foreign language on the fly, get relevant recommendations, and rapidly obtain relevant information on any conceivable subject – something that recently you could encounter only in science fiction.

One of the most famous empirical observations in the field of Computer Science is Moore's Law. According to it, the number of semiconductors on silicon die doubles every year (Fig.1) (1). This simple projection introduced by Gordon Moore (co-founder and chairman of Intel Corporation) in 1965 still drives the chip manufacturing industry growth. It is closely tied to the exponential increase in clock speed, memory capacity, and related characteristics (2). These measures are directly reflected in the computational power available to the scientific community and act as hard constraints for the multitude of applications in data science, machine learning, and physical simulations.

However, in recent years certain GPU advances broke away from this projection, increasing the rate of operation by the rate of 25 in 5 years. For some special applications, such as training deep neural networks, speed up was by a factor of 500 (3). And GPUs are not a

unique case of diverging from Moore's Law. There was another monumental advancement in the genomics area.

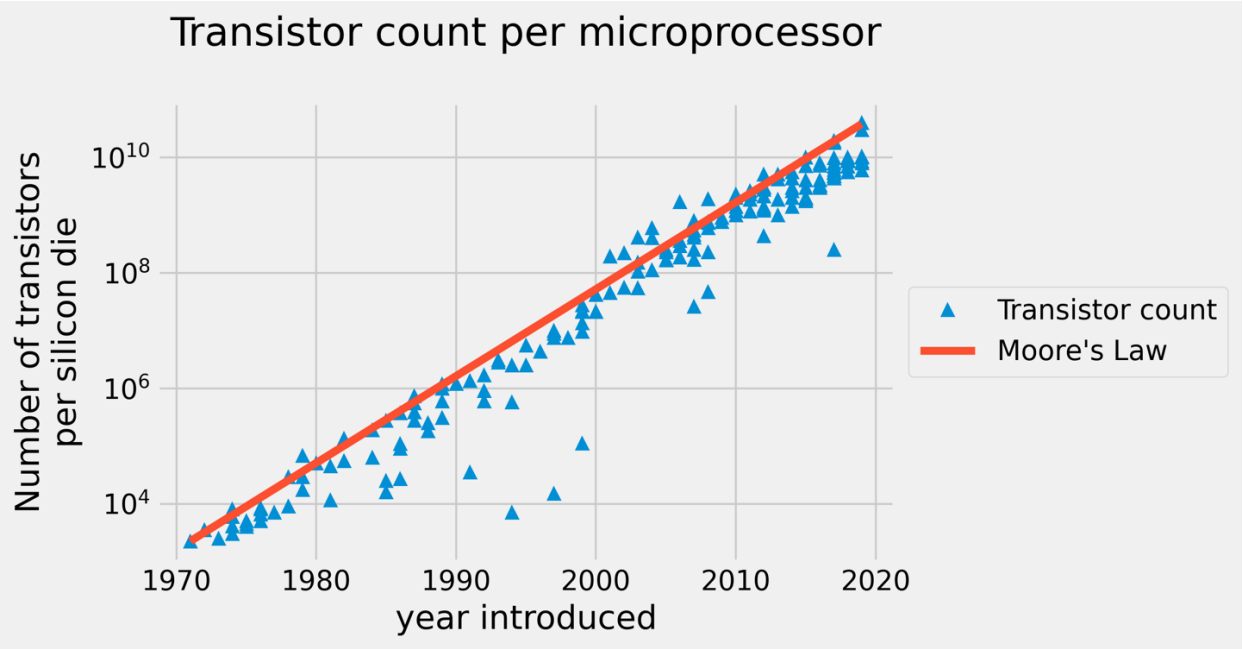

Figure 1.1 Moore's law and historical transistor count in microprocessors. Empirical observation from 1965 still holds relevance in modern Computer Science and the microprocessor manufacturing industry.

At the dawn of the 21st century, a novel, highly parallel approach for genome sequencing, Next Generation Sequencing (NGS), was introduced. It opened new horizons in studying the genetic roots of disease (4). For many years, the cost of genome sequencing steadily followed Moore's Law. The exponential decrease in expenses was promising; however, this projection predicted prohibitively high prices for obtaining complete genetic information from a given individual. Still, it inspired optimism and talks about personalized medicine, which envisioned prescribing therapeutic treatments best suited for a specific individual (5-7). Still, the individual cost was exceedingly high, and the amount of available data was insufficient to develop detailed clinical action items for specific patients.

The first decade of the 21st century brought forward astonishing advances in genome sequencing technology. In the year 2008, the second generation of NGS technology completely brokeMoore'ss Law (4). The expenses were drastically reduced, and the costs of obtaining a genome from a specific individual drastically dropped. For example, sequencing the entire genome had a cost of $100,000,000 in 2001 and was reduced to roughly $10,000,000 duringMoore'ss Law-like change in costs. But after the same span of six years, in 2013, the price of obtaining a complete genome was around just $8,000 — such a sharp decline in the costs allowed to conduct extensive population-wide genomic studies.

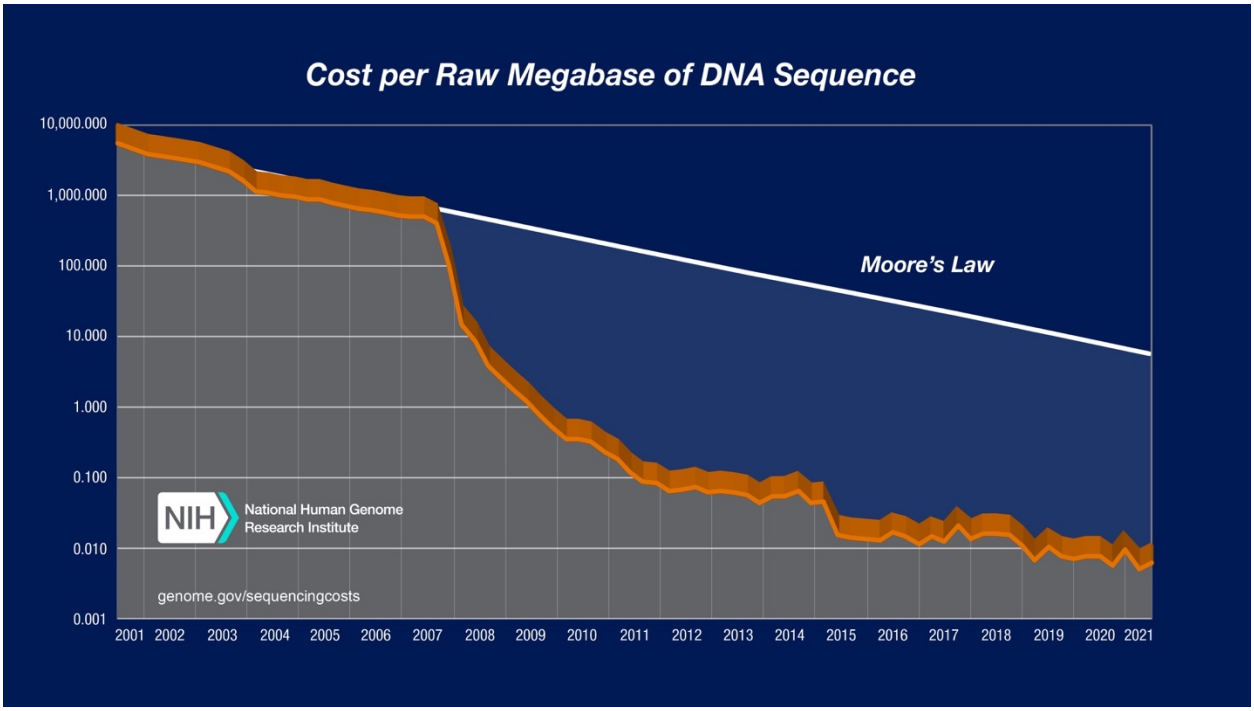

Figure 1.2 Sequencing cost per raw megabase of DNA. The sharp decline in required resources was achieved in 2008, breaking Moore's Law and further accelerating the exponential growth of sequencing capabilities. Adapted from (8).

After a few years of gathering information, a new paradigm emerged – precision medicine. In 2011 the National Research Council introduced this term to expand the definition of

personalized medicine to subpopulations based on susceptibility to specific diseases or treatment responses (9, 10). It was also tightly knitted with a novel view on the taxonomy of human disease rooted in a knowledge network and tightly linked to the molecular biology origin and detailed mechanism of action (9).

Precision medicine got the attention on the highest level in 2015 when Barack Obama, in the State of the Union speech, outlined a new national initiative with the focus on developing genomics methods in diagnostic and incorporating them in a health care setting (11).

The precision medicine spotlight brought forth a significant amount of institutional support. The National Institute of Health (NIH) launched the "All of Us Research Program" initiative in 2015 with the goal of sequencing genomes of 1,000,000 individuals and studying corresponding health conditions. The emphasis of this program is on advancing treatment, prevention, and diagnostics methods for oncology. On top of collecting top-notch scientific datasets, this initiative's participatory model focuses on providing access to cancer treatment in world-class institutions to all patients. Another oncology-centered initiative – Cancer Moonshot – was launched by Joe Biden with the aim to end cancer as we know it and began in 2016 (12). It is a data-intensive undertaking with the initial goals of accelerating scientific discovery in cancer, fostering collaboration among researchers and institutions, and improving data sharing (13).

Efforts dedicated to enriching genomics databases and understanding the disease basis directly from the blueprint of life were central to the progress of precision medicine. However, the gap between genotype and phenotype is still considerable; it may not even account for all necessary building blocks as environmental factors come into play. Because of this scientific community realized the importance of incorporating auxiliary data that either provide insight into

the current snapshot of organism functioning or elucidates molecular mechanisms of action. It led to the "'omic" boom and multimodal analysis in healthcare – so-called multi-omics. The information can come from various sources.

Among the omics technics, transcriptomics highlights the actual expression pattern of the genes and elucidates RNA regulatory processes (14-16). Proteomics is even more direct; it provides data on macromolecules that play the role of conduits for the majority of molecular functions. However, the range of the products it can detect is limited compared to transcriptomics (17-19). Interactomics studies seek to organize information on individual proteins into a network with the goal to advance understanding of the molecular mechanisms of action (20-22); it produced edgetics (23, 24), a research direction with the goal of bridging the gap between genotype and phenotype. Metabolomics keeps track of the products of life activity (25-27). Epigenomics elucidates the influence of DNA methylation due to the environmental factors on gene expression (28-30). Microbiomics adds data on microorganisms co-inhabiting host, their concentration, location, and influence on health conditions (31-33). Functional imaging (34-36) elucidates processes happening in the brain, which is immensely helpful for studying complex neurodegenerative disorders. Radiology (37-39) explains processes that occur on the scale of the entire organ.

Data gathering is only one of the facets of precision medicine. After obtaining large arrays of data, there is a need to conduct a comprehensive analysis. Perhaps, the most widespread tool for screening sequencing data is the Genome-Wide Association Studies (GWAS) (40). The goal of this population-wide analysis method is to identify genetic risk factors that make individuals susceptible to the specific disease and the underlying biological basis for the disease development. The omics data have similar analytical approaches, e.g., Epigenome Wide

Association Studies that focus on DNA methylation patterns (41). However, they are not a carbon copy of GWAS, as each type of omics data requires careful considerations for tissue specificity and appropriate cohort selection (42).

The major paradigm shift happened when high-throughput experimental technologies became not just a supplement to the classical hypothesis-driven research but transformed into a hypothesis-generating tool (43). With the amount of data constantly snowballing, it was evident that precision medicine entered an era of big data. The central concepts of big data are described by the fiveV'ss: Value, Volume, Variety, Velocity, and Veracity (44). Value is the most crucial concept; it corresponds to the significance of the insights that could be obtained from the data – namely, patterns, action items, and potential optimizations. Volume corresponds to the amount of available data that may be too huge to store and analyze with traditional database solutions efficiently. The Variety describes types of heterogeneous data – structured, semi-structured, and unstructured data. Velocity is the speed with which new data are gathered and processed. Veracity is the accuracy of the data and corresponding confidence. All these factors – Value, Volume, Variety, Velocity, Veracity – call for improved data analysis, visualization, and storage approaches.

The advances in data analysis for precision medicine would be slowed down without equally monumental efforts in data analysis. The modern communication mediums allowed for creating and coordinating humongous international research groups across multiple continents, e.g., a Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (45). It presented one of the most comprehensive and detailed analyses on cancer genomes, which allowed experts on the specific biological questions that composed it, to bring out insights in multiple areas – cancer

evolution (46), driving non-coding mutations (47), cancer signatures (48), somatic structural variations (49), and genomic basis of RNA alterations that take place in affected patients (50).

Conducting a high-quality data analysis and subsequently integrating massive amounts of the molecular data into the clinical decision-making process and informatics application, e.g., electronic health records, poses a significant challenge to the pathology laboratories attempting to bring precision medicine onboard (51, 52).

In the last years, the precision medicine field produced impressive results. It encompasses multiple technical and regulatory aspects, and the application areas differ drastically. Pharmacogenomics data was used to guide blood-thinning drug selection (warfarin or oral anticoagulants) and automatic dosage estimation (53, 54). Targeted therapies for cancer treatment to modulate aberrant pathways or protein activity to disrupt the sustenance of cancer cells emerged. One of the prominent examples is the precise disruption of the oncogenic kinases (55). Either with the usage of small molecule drugs, such as imatinib that brought to the table complete response for more than 90% of chronic myelogenous leukemia (56) or with monoclonal antibodies, as was the case with targeting tyrosine-protein kinase erB-2 (HER2) (57).

One of the advances on the technological frontier of personalized medicine is the adoption of lightweight wearable devices for constant monitoring, with is crucial for personalizing drug selection and dose parameters for patients undergoing treatment (58). It includes using non-intrusive sensors for analyzing tears, sweat, and saliva (59-61); battery-free devices that measure the oxygen level in blood and heart rate (62, 63).

1.2 Alternative Splicing – Zeroing in on Variations in Building Blocks of Life

According to the classical postulates of the Central Dogma of molecular biology, the molecular components of the living organisms proliferate in the following way: DNA is transcribed into an RNA molecule, which is further translated into the amino acid sequence, which forms a protein. However, according to the modern understanding, the complete picture of the transformations between these molecules includes additional transformations: DNA replication (DNA obtained from DNA), RNA replication (RNA obtained from RNA), and reverse transcription (DNA obtained from RNA).

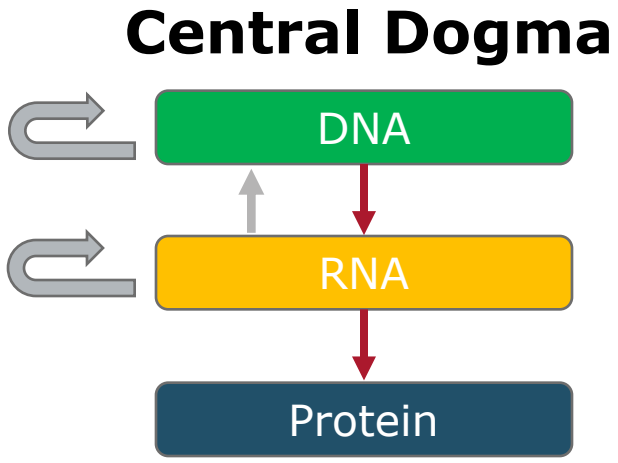

Figure 1.3. The central dogma of molecular biology. Main route of expression consists of transcription (DNA to RNA) and translation (RNA to amino acid sequence). Additional transformations include DNA replication, RNA replication, and reverse transcription.

Among these three types he molecules, proteins carry most of the functional burden. These macromolecules serve various purposes: structural support of the cell by forming cytoskeleton, passive and active transport of the molecules, conducting signals across CNS, replicating DNA, and regulating gene expression. Proteins play a crucial part in digestion, hormone production, immune resistance, and tissue growth. They vary in different organisms,

cells, tissue, or time points. These molecules' chemical reactions and movement underlie dynamic processes in living organisms.

However, it is impossible to carry out this wide array of functions based on the individual proteins. Interactions between biomolecules are a necessary component to exert their function. Among them, protein-protein interactions (PPIs) mediate most cellular processes, resulting in distinct phenotypes. The crucial role that PPIs play in human health and disease motivated multiple experimental and computational research projects dedicated to the annotation of the human interactome. Getting comprehensive information about possible interactions can help us elucidate disease etiology by understanding underlying cellular mechanics. However, factoring in protein-protein interactions significantly increase the complexity of the studied phenomenon.

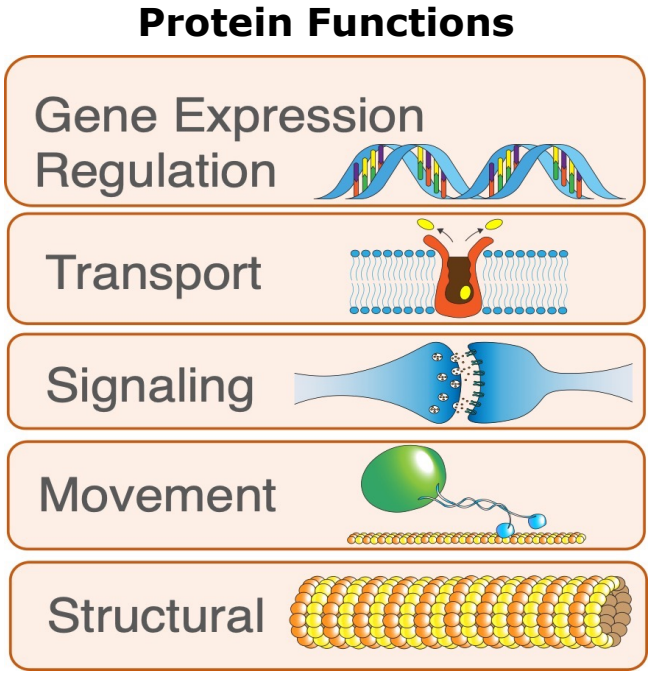

Figure 1.4. Examples of protein functions. Gene expression regulation (transcription factor attached to the DNA), passive transport (GLUT4 transporter), active transport (kinesin cargo transporter), neurotransmitter signaling (dopamine), structural support (microtubules).

Whenever we discuss precision medicine, it is difficult to underestimate transcriptomics role in elucidating molecular processes. It takes a unique place at the intersection between genomics and proteomics. Transcriptomics focuses on studies of mRNA, which is an intermediate state between DNA blueprints and functional macromolecules. Besides providing a glimpse into the protein landscape present in the organism, RNA molecules also may have a significant regulatory role, though this topic is outside of the scope of this work.

Unlike the studies focused solely on DNA sequences, transcriptomics provides a much more dynamic picture, highlighting tissue-specific processes and taking a snapshot of the actual state of the organism. It also allows us to get insights into proteome, a collective name for all proteins expressed under given conditions. And there is a large discrepancy in the size of the proteome and genome.

With the advancement of modern transcriptomics, scientists have realized that most genes in higher eukaryotes can produce more than one product per gene. It makes it possible for the organism to increase complexity without significant changes in the size of the genetic code. For example, humans and fruit flies have a comparable number of protein-coding genes, but humans have more specialized cells that constitute different tissues and form our organs (64). Even though the human genome has only about 20,000 distinct protein-coding genes, our organism developed some uncanny methods that can drastically increase variation across basic functional units that manage to support the growing complexity of the biological system. Over 90% of human multi-exon protein-coding genes can transcribe alternatively spliced mRNAs regulatory processes such as alternative splicing and post-translational modification, by some estimates, create over one million distinct proteoforms. The focus of transcriptomics is on the RNA

molecules that open up the ability to study alternative splicing events, which modulates more than one hundred thousand distinct transcripts.

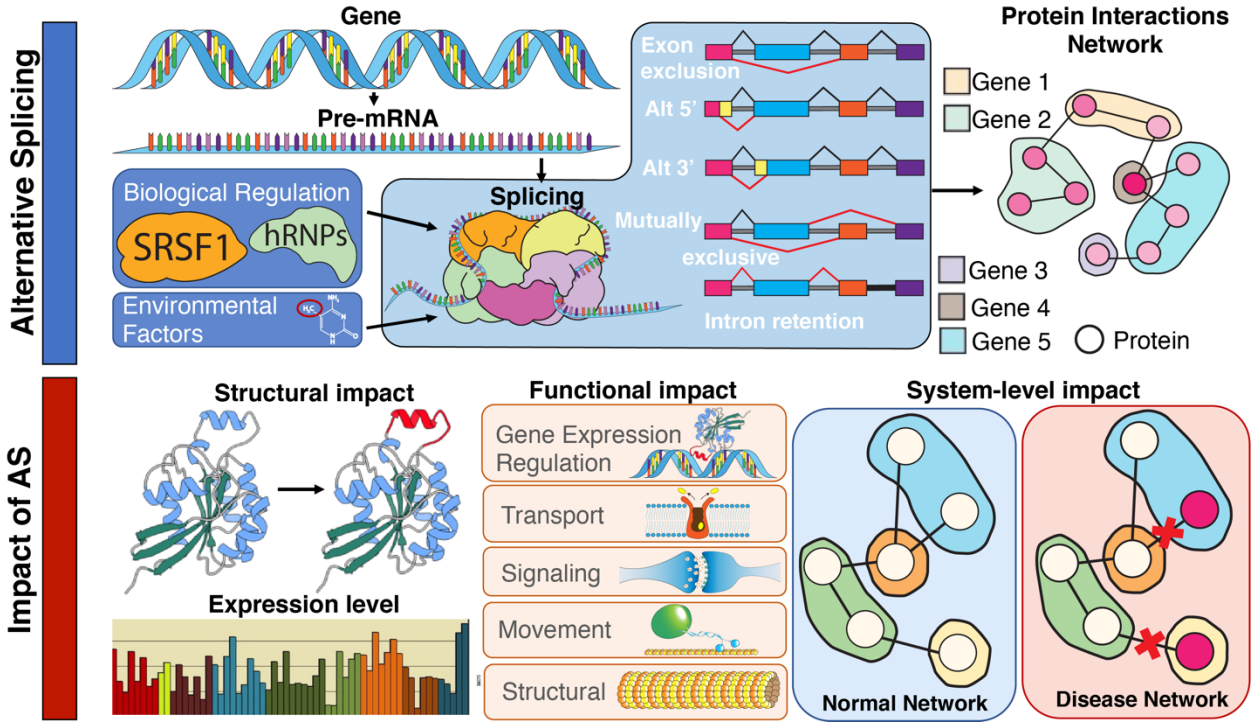

Figure 1.5. Alternative splicing regulatory process and its functional impact. Splicing takes place after the initial pre-mRNA molecule is transcribed from the DNA. The Spliceosome RNA-protein complex is then responsible for the removal of the intronic regions. This process may produce either a reference isoform or an alternative one; then, we discuss alternative splicing events occurring. Biological factors can regulate this process, e.g., Serine And Arginine Rich Splicing Factor 1 and heterogeneous nuclear ribonucleoproteins, an environmental influence that causes DNA methylation, and small molecules – alternative splicing modulators. AS events significantly diversify the protein interaction network and lead to the combinatorial increase in the number of potential protein-protein interactions. They have the potential to dramatically impact protein structure, cutting out parts of the functional domains and affecting the folding process. The expression level of the gene products can also be altered due to the AS. The cumulative effect of these changes leads to the diversification of protein specialization that we observe in different tissues and modifying protein-protein interactions.

Splicing is one of the biological regulatory processes that transform pre-messenger RNA molecules obtained via reverse complimenting the original DNA sequence into mature messenger RNA, which can then be translated into protein. It removes non-coding non-coding subsequences (introns) and joins coding regions (exons).

This variation in cell functions is made possible because of proteome diversification. The main regulatory mechanisms that allow us to achieve such an increase in complexity are pre-mRNA alternative splicing and post-translational modification. In this work, we will focus on the first one.

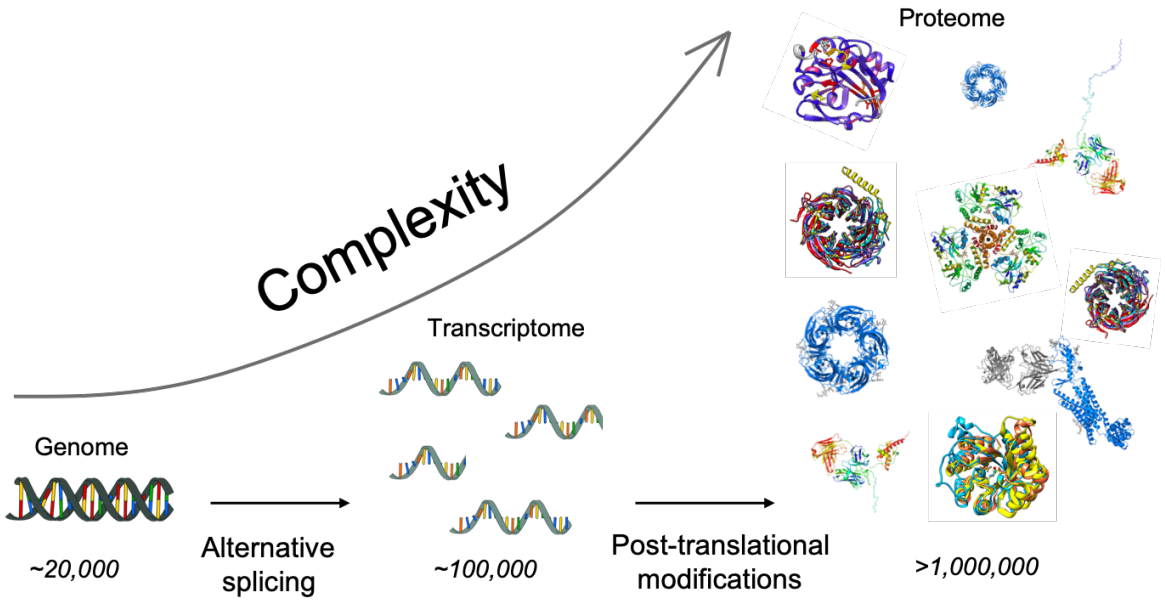

Figure 1.6. Growth of gene product diversity due to alternative splicing regulatory process and post-translational modifications. Estimated 20,000 protein-coding genes produce more than 100,000 distinct transcripts via alternative splicing, which are further differentiated by the post-translational modification mechanisms, resulting in more than 1,000,000 proteoforms.

Pre-mRNA splicing is a crucial step in mRNA maturation, and alternative splicing due to either natural or disease-causing variation in transcriptome is a process by which the same gene can result in different gene products through selective inclusions and exclusions of the gene's exons and introns (65). It creates different combinations of splice sites, allowing for the production of distinct proteins from a single gene. It can be induced by biological regulators, such as SRSF1 and heterogeneous nuclear ribonucleoproteins, or environmental factors, such as DNA methylation. Alternative splicing gives rise to the combinatorial increase in complexity of the gene products and leads to the significant expansion of the PPI network.

The primary source of experimental data for high-throughput alternative splicing studies is an RNA-Seq, next-generation sequencing technology that quantifies the amount of mRNA material present in a biological sample. It fragments input RNA material and converts it to the cDNA fragments (reads), which are aligned to the reference genome or assembled into a new one using *de Bruijn* graphs. Increasing the number of reads and reads length helps to make precise estimates. Gene expression profiling experiments may require 5-25 million reads. The ability to describe alternative splicing events requires much higher depth and, depending on the application, may take up to 200 million. Original RNA-Seq technology required a significant amount of biological material. In most experiments, all sample material belongs to the specific tissue and contains a mixture of cells that constitute this tissue. Recent advances in technology led to the increase in resolution level – an ability to quantify expression levels of the individual cells. It brings a clear advantage - the ability to study separate cell types. But it comes with considerable drawbacks – scarcity of biological material in each experiment increases the noise level in data and restricts sequencing depth. The latter comes in the way of conducting alternative splicing studies.

Analysis of mammalian tissues in a study (66) shows high conservation of tissue-specific gene expression patterns among mammals evolutionally diverged from <30 million years to >300 million years. Alternative splicing patterns are conserved in brain, muscle, heart, and testis tissue and vary in other parts of the organisms in different species, as shown in (66). This finding suggests that alternative splicing is more frequently affected by changes in species biology. Understanding splicing patterns can help pinpoint the differences between lineages and bridge the gap between model organisms and humans.

While many alternative splicing events naturally occur in different tissues, cells, and under different cellular conditions, a growing number of alternatively spliced genes have been associated with genetic disorders, including cancer, neurodevelopmental and heart diseases, and others (67-69). Alternative splicing has been shown to alter the protein function (70). The range of functional variation between the alternatively spliced isoforms may vary drastically: from a complete loss of original function due to misfolding and removal by the cell degradation mechanism of the corresponding alternatively spliced isoform to a subtle difference in the protein functioning, or perhaps the gain of a new function, due to acquiring by the isoform of a new exon that encodes a new functional protein domain.

These findings highlight that though genetic information serves as a blueprint for our organism, knowing the sequence information alone is not enough to assess underlying cellular processes crucial for its normal functioning accurately. Alternative splicing events can significantly influence the resulting phenotype, leading to an increase in drug resistivity or a disease occurrence. Acknowledging this fact, multiple companies and institutions develop splice modulator drugs for cancer treatment (71-73).

Ideally, tools available for studying alternative splicing would directly identify the entire molecules present in the cells. This approach is called *long-read* sequencing. Unfortunately, the current state of technology has significant limitations, described in the scRNA-seq section, that significantly restrict the range of biological questions it can help answer. It impaired widespread adoption, and the *short-read* sequencing approaches that heavily rely on reference sequences remain prevalent. Short-read data constitute most RNA-Seq samples in the major databanks, such as TCGA (74), GTEx (75), and ICGC (76).

Neither reference genetic sequence data nor short reads obtained during experiments directly include comprehensive information on the isoform configuration. This makes computational prediction of the alternative splicing isoforms the most efficient and widespread class of methods that allow researchers to tap into vast biological databanks that house short- and medium-length reads RNA-Seq samples.

1.3 The Role of Machine Learning in Precision Medicine

As the amount of publicly available biological datasets keeps rapidly increasing with the advances in sequencing technologies, medical imaging, and screening assays, structural information remains the most comprehensive description of proteins and other biological elements, as it exposes interaction interfaces, molecule shape, and surface area.

The need to analyze the vast arrays of sequenced genomics data that constantly keeps growing leads to the tighter coupling between healthcare and such computational fields as big data and Artificial Intelligence (AI) (77).

The AI field combines various computational methods that have partial attributes of human reasoning or the ability to mimic human behavior – like solving problems on incomplete data, automating the inference process, and optimizing giving functions. Its history comes back to the middle of 20th century, along with the emergence of the first computers. The hallmarks of that period are the formulation of the Turing Test in 1950 (78), the very first checkers program in 1952 (79) and the perceptron model in 1957, which laid the foundation for a large number of modern machine learning algorithms (80). In 60s, AI applications found their way to the assembly lines of General Motors in the form of robotic arms (81). At the time, there were two main contesting points of view – a top-down and bottom-up approach (82). The former was

dedicated to taking high-level functions and adopting them to the problem; the latter attempted to reconstruct intellectual faculties by imitating neural activity. 70s brought forward the initial architectures of multilayer neural networks and the development of backpropagation for weight updates (83), along with the development of the connectionism theory.

After initial expectations for the AI capabilities were dampened due to the complexity of the undertaking, the field entered a "Winter of A" state for a prolonged time, with limited advances (84). In 90s and earlier 2000s we see the development of statistical learning theory (85-88) and feature-based machine learning algorithms (89-91). These methods generally require an additional step – feature engineering – which requires researchers to come up with the information encoding strategies for the complex input data, e.g., histograms of oriented gradients for images and video classification, Fourier spectrum for audio recognition, and a bag of words for the Natural Language Processing (NLP) applications. Neural networks were developing in parallel and were dealing with unique computational challenges, e.g., vanishing gradient (92). This problem was especially noticeable with the increase of hidden layers in the architecture. After subsequent backpropagation steps, weight changes are reduced, and the time necessary for the training phase becomes prohibitively long. However, in 2012 a spectacular breakthrough happened with the creation of the AlexNet model with the ingenious usage of GPU hardware (93).

In recent years the term AI became almost synonymous with Machine Learning, a subclass of AI algorithms with the goal of the automatic inference of patterns and correlations from data arrays.

Big data availability and AI's ability to analyze patterns open doors to the incorporation of essential healthcare information that exists outside of the healthcare system (94). Lifestyle,

nutrition, exercise, and social activity can be integrated along with the high-dimensional molecular data and clinical information. There are multiple modalities that could be efficiently exploited with the usage of AI.

Current advanced methods, such as ChIP-seq (95, 96), ATAC-seq (97), and Hi-C (98), capture chromatin conformation, providing insights on 3D positioning and interactions, which is indispensable for epigenetics studies. However, it is not precise enough to obtain a comprehensive molecular structure. Therefore, proteins remain the primary targets in the structural biology field, as they are the most stable class of gene products, and in this work, I would focus my attention precisely on them.

Structural information can be used to infer the functional role and significance of the protein products, assess them as a drug target via small molecule docking and investigate their behavior in physically plausible settings using molecular dynamics simulation.

One of the most challenging problems of structural bioinformatics is 3D structure prediction from the sequence information. Here, the difficulty lies primarily in the complexity of interactions between amino acids residues and in the non-trivial way in which molecule is assembled. Besides physical forces that originate from the molecule itself, molecular scaffolds (99) introduce further changes, and the environment plays a certain role. In order to facilitate progress in this area, an annual competition – a critical assessment of methods of protein structure prediction (CASP) (100-102) – is held; and recent deep-learning based models demonstrated one of the best results ever in recent years.

However, informative, structural information characterizes distinct molecules and does not provide ready-made recipes for studying complex processes. Molecular dynamics simulation is an incredibly computationally expensive endeavor, and the current capacities of

supercomputers are not even close to enabling whole-cell simulations, not to mention that even microsecond-long runs of such systems may not reveal helpful information pertaining to the disease phenotype.

To bridge this gap, the biological field embraces a system approach that aims to uncover underlying regulatory mechanisms of the cells (in particular) and the entire organism (in general). We are gaining more evidence indicating that even though certain phenotypical outcomes (e.g., mendelian disorders) are governed by a singular gene, in the general case, this process is far more complex. It brings together distinct effectors that exhibit influence on multiple other elements: genes, RNA products, proteins, metabolic compounds, and microbiome. Operating across even a single modality and inferring mutual influence is a challenging task due to the number of regulatory elements.

To get traction from complex interdependent data, scientists developed multiple network-based models. This approach demonstrated viability across multiple applications such as disease module identification, gene regulatory network (GRN) inference, comorbidities detection (103), and drug repurposing (104). It allowed researchers to establish causal links between multiple regulatory elements.

However, due to the highly heterogeneous nature and high dimensionality of biological data, it is challenging to incorporate relevant information into knowledge-based models, so researchers resort to either the statistical approaches (105, 106) to drastically reduce number of variables of interest or more complex manifold learning methods that allows combining of individual values into meaningful joint representation which enhances machine learning methods' capability to detect relevant patterns.

1.4 Conclusions

Precision medicine aims to identify factors that contribute to different treatments outcomes and design an intervention strategy best suited for each population group. To achieve this goal, it leverages a wide array of techniques that are collectively known as "omic." While genomics describes a "blueprint of life" information encoded in DNA, the rest of the omics methods (transcriptomics, metabolomics, proteomics, epigenetics) provide information on the dynamic landscape of molecular activity taking place in the organism at the given moment. It allows to get insights into the fundamental molecular basis of the disease and not merely correlate ailments with the mutation presence.

An important role in data analysis play machine learning methods. Whether we are talking about clustering cell types, identifying relationships between expressed genes and phenotypical characteristics, or assessing pharmacological properties of novel drugs, data dimensionality is a major factor. No amount of manual analysis by biologists and technicians can reliably extract vital information from the biological samples. In this setting, the ability of machine learning methods to extract patterns automatically is a boon. However, outsourcing the most complex part of the analysis to the machines comes with caveats. Biological data are exceedingly complex and often intractable from the human perspective. This creates additional difficulty in the model validation step. Another drawback from the ML point of view is the small number of training samples that come from the costly experiments, along with the high degree of technical variation.

Fortunately, most of the difficulties can be addressed either by advancing computational methods or leveraging biological expertise. This makes precision medicine an extensively

interdisciplinary field that provides fertile ground for the collaboration of scientists with diverse backgrounds.

Chapter 2. Deep Learning Methods as the Most Recent Advancements of Data-Driven Approaches

2.1 Deep Learning Methods: History and Development

Deep learning (DL) is the recent rebranding of multi-layer perceptrons that exhibit powerful learning capacities after advances in hardware and the machine learning field managed to resolve the vanishing gradient problem (92) and enabled construction of the large-scale models, such as AlexNet (107), BERT (108), GPT3 (109). The main advantage of the deep learning methods is their ability to learn compact representation relevant to the problem from the raw data (110). This puts DL methods into the category of representation learning algorithms, with the data point representation being derived in several hierarchical levels (111).

As the construction of a compact representation of the raw data is the key feature, advanced unsupervised or semi-supervised learning strategies such as self-supervised learning (112) hold great potential for distilling relevant physical and spatial properties of the structural data or underlying relationships of reconstructed networks into reusable and shareable libraries. This would address the problem of technical base accumulation, as the current situation in the computational biology field mandates researchers to re-implement the training stage for the majority of applications. There are already steps, like an autoencoder(113, 114)-based TorchDrug (115).

Another significant deep learning advancement is transformer-based network architecture (116). Originally incepted for machine translation tasks, this type of neural networks excels in sensing even minute changes in sequence information by focusing on the context. The model trained via this approach has a high potential for being reused via transfer learning, serving as a

base for multiple specialized tools, similar to the BERT model that gave rise to Alberto (117), ALBERT (118), BERTAC (119), and BioBERT (120). The latter one is used for biomedical text mining. Derived sequence representation can be used as the auxiliary information for both structural and network biological applications.

Deep graph neural networks (121) are able to learn from the global graph topology, which makes them extremely suitable for developing gene regulatory networks and similar models across multiple omics modalities. Such models can be applied to disease diagnostics or drug development problems (122).

2.2 The Architecture of the Deep Learning Networks

2.2.1 Convolutional Neural Networks (CNN)

Convolutional neural networks are inspired by the neuron composition in the human visual cortex (123) and use consecutive filtering to uncover local correlations. When applied to the image recognition problem, this class of neural networks indeed learns representations that closely correspond to the image primitives, such as corners, straight lines, and circles (124).

CNN architecture consists of subsequent convolutional and subsampling (pooling) layers. Unlike in traditional neural network architecture, neurons in convolutional nets are connected only to the small number of neurons in the next layer and do not form fully connected structures. In image-based problems it corresponds to the local receptive field and is defined by convolution size. This feature reduces the chances of overfitting and training time.

Sub-sampling (pooling) layers further reduce the size of the network by applying local averaging or maximum filters across neighboring neurons.

The cost function for CNNs can be defined as the following:

$$J(W, b; x, y) = \frac{1}{2} ||h_w(x) - y||^2,$$

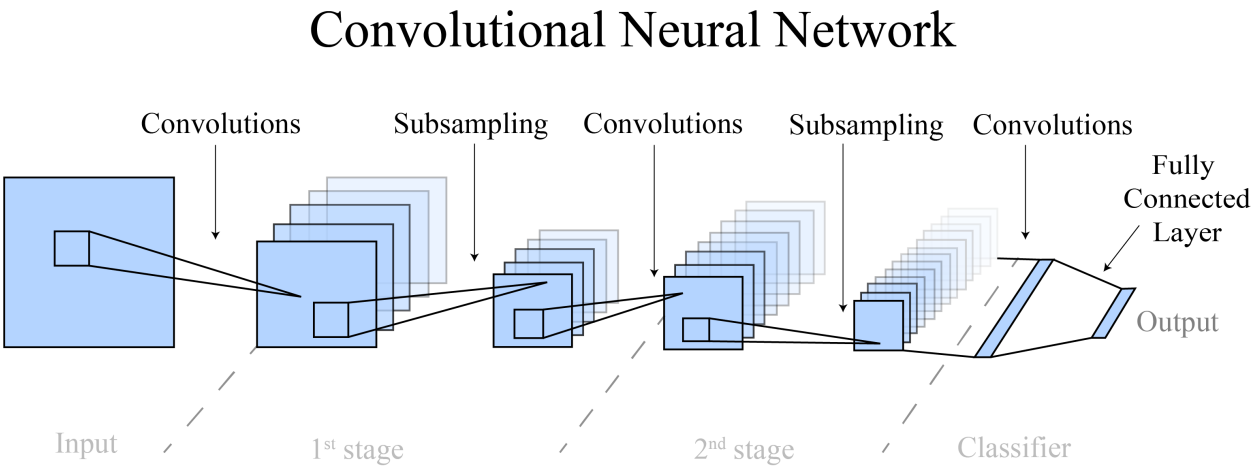

Figure 2.1. Convolutional Neural Network architecture. It consists of the interchanging convolutional and subsampling layers. Convolutional layers serve the purpose of combining input information with the goal of extracting features. In the general case, it has trainable parameters. The subsampling layer serves a single purpose – reducing the dimensionality of the previous layer outputs. Based on (125).

where x is the data point, y is the corresponding label, W is the network's weight matrix, b is the intercept term, and h_w is the transformation applied by the neural network. The error term δ^l of the layer l is denoted as

$$\delta^l = ((W^l)^T \delta^{l+1}) f'(z^l),$$

where f' is the derivative of the activation function. Gradients can be computed using the following formula:

$$\nabla_{w^l} J(W, b; x, y) = \delta^{l+1} (a^{l+1})^T$$

$$\nabla_{b^l} J(W, b; x, y) = \delta^{l+1}$$

with the term a^l corresponding to the input of the l-th layer.

Error of sub-sampling layer is defined as

$$\delta_k^l = upsample\left((W_k^l)^T \delta_k^{l+1}\right) f'(z^l{}_k),$$

where k corresponds to the layer-specific filter number. Convolution between input of the i-th layer with the respect to the k-th filter, ${a^l}_i * {\delta^{l+1}}_k$, contributes to the calculations of the pooling layer's gradient in the following way:

$$\nabla_{w^l} J(W, b; x, y) = \sum_{i=1}^{m} {a^l}_i * rot90({\delta^{l+1}}_k, 2)$$

$$\nabla_{b^l} J(W, b; x, y) = \sum_{a,b} ({\delta^{l+1}}_k)_{a,b}$$

The training phase constitutes a standard backpropagation algorithm: after obtaining predictions $\hat{y}$ from the forward run, find the error term for the output layer and consequently update previous layer's weights by applying a gradient descent algorithm based on previously calculated gradients.

Residual Neural Networks (ResNet)

ResNets are a special case of DNNs that introduces a weight aggregation strategy that combines the output of the current layer with the output generated by one of the previous layers, or with the original data. The updated inputs are calculated as

$$h(x) = F(x) + x$$

The training phase remains primarily unchanged; gradients for backpropagation are calculated solely based on the latest layer error terms, while the layer inputs for the forward run are calculated as $h(x)$.

By propagating residual values from the previous layer, ResNets mitigate a vanishing gradient problem, allowing for the construction of multilayer system and making a design of extremely deep networks possible (126).

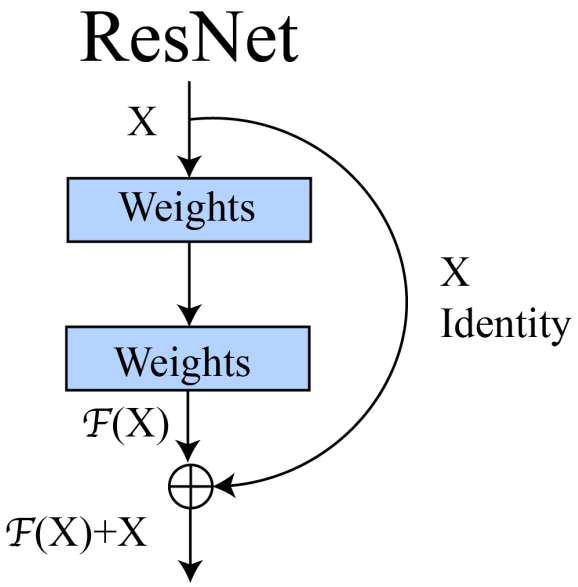

Figure 2.2. Residual neural network architecture. Skipping subsequent layers is a strategy that aids in countering the vanishing gradient problem. Backpropagation in such a network allows giving feedback to the previous part of the neural network without diminishing gradient values. Based on (127).

2.2.2 Long-Short Term Memory (LSTM)

LSTM is the subtype of neural network that is useful for processing sequential data. Unlike regular feed-forward network types, LSTM can retain knowledge distilled from the previous runs (128). Instead of discarding this information or using a fixed storage size, it learns for how long it should retain context depending on the previous inputs.

The key parts of the LSTM cell are memory cell C, forget gate, and output gate. A signal from the input cell is being controlled by these three components. Let x be an input vector, h - an output value, and t – the time stamp.

Gate values from Fig.3.3 are denoted as f_t, i_t, o_t and correspond to the forget gate, input gate, and output gate. They can be obtained by following the next equations:

$$f_t = \sigma\left(W_f x_t + w_f h_{t-1} + b_f\right)$$

$$i_t = \sigma(W_i x_t + w_i h_{t-1} + b_i)$$

$$o_t = \sigma(W_o x_t + w_o h_{t-1} + b_o)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \sigma(W_c x_t + w_c h_{t-1} + b_c)$$

$$h_t = o_t \otimes \sigma_h(c_t)$$

where W, w, b correspond to the weights of the input, recurrent output, and intercept. $\otimes$ is the elementwise multiplication.

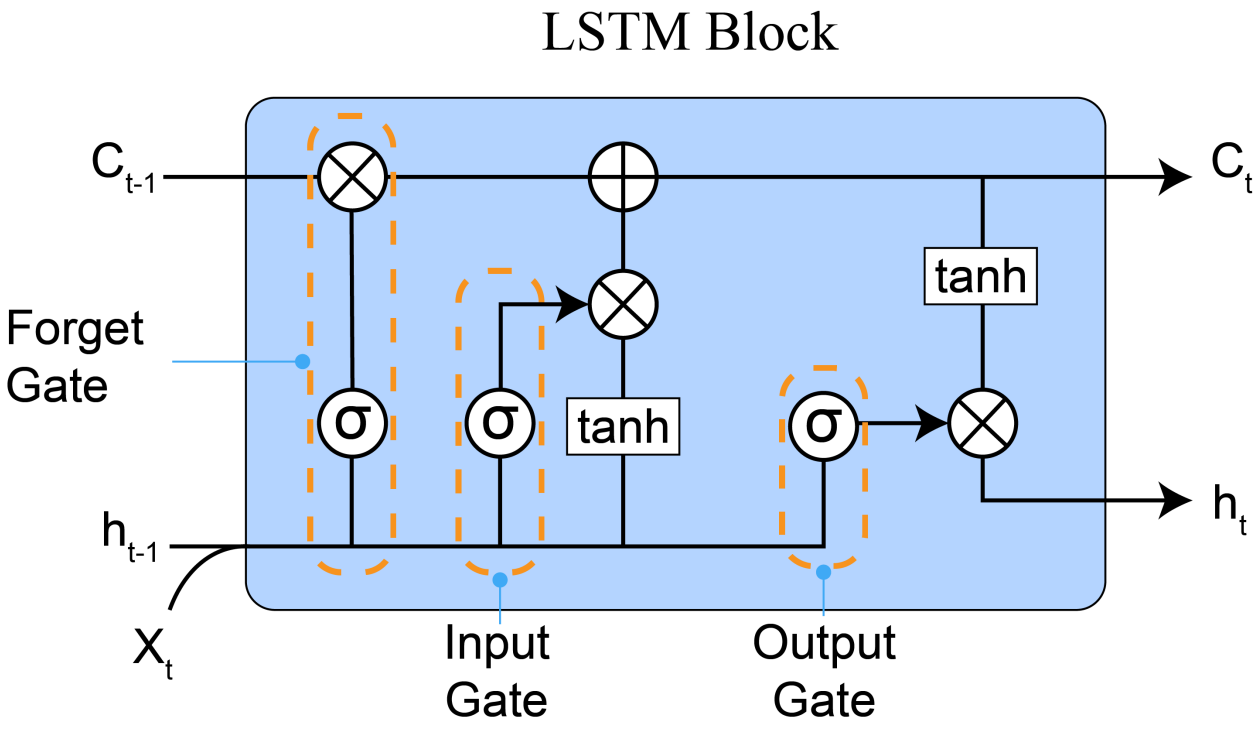

Figure 2.3. Long-short term memory (LSTM) block with memory gates. σ corresponds to the sigmoid function, $tanh$ - to the hyperbolic tangent, operator + depicts summation, operator × represents multiplication. Based on (128).

2.2.3 Generative Adversarial Networks (GAN)

GANs adversarial modeling works by learning underlying data distribution from data samples. This method conducts minimax game $\boldsymbol{V(D,G)}$ between two players with the competing objectives – a discriminator $\boldsymbol{D(x;\theta_d)}$ that classifies samples $\boldsymbol{x}$ into either ground truth or synthetic category and a generator $\boldsymbol{G(x;\theta_d)}$ with the objective of decreasing discriminator's performance:

$$\min_{\boldsymbol{G}} \max_{\boldsymbol{D}} \boldsymbol{V(D,G)} = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log \boldsymbol{D(x)}] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[\log\left(1-\boldsymbol{D}\left(\boldsymbol{G}(\boldsymbol{z})\right)\right)\right]$$

Discriminator $\boldsymbol{D}$ is trained to maximize the probability of assigning a correct label to the data point, while generator $\boldsymbol{G}$ is trained to minimize it.

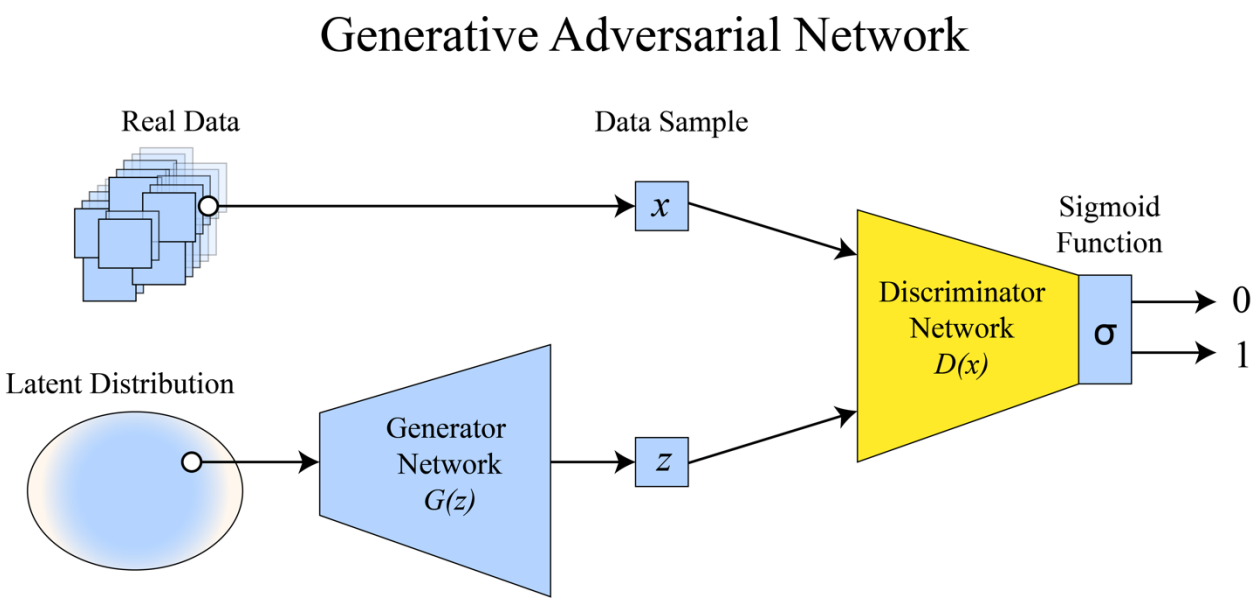

Figure 2.4. Generative adversarial network (GAN) architecture. Two major components of GANs are a discriminator network $\boldsymbol{D(x)}$ and a generator network $\boldsymbol{G(z)}$. *The discriminator* learns to distinguish real data from the generator network outputs and optimizes this task. The generator network estimates a prior distribution of the real data and is capable of sampling surrogate data from it. Its goal in the adversarial setting is to counter the discriminator network by making sampled data points indistinguishable from those in a real training set. Based on (129).

One discriminator training epoch is conducted by feeding all ground truth samples and generated samples in $\boldsymbol{m}$ mini-batches, $\boldsymbol{x^i}$ and $\boldsymbol{z^i}$ correspondingly. During each run subsequent discriminator is updated by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^i) + \log \left(1 - D\left(G(z^i) \right) \right) \right]$$

After each $\boldsymbol{k}$ epochs of the discriminator training generator is updated by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[\log \left(1 - D\left(G(z^i) \right) \right) \right]$$

Successfully trained GANs are able to generate life-like data examples, though this family of algorithms suffers from non-convergence and vanishing gradient problem in a case when discriminator $\boldsymbol{D}$ becomes too successful (130).

2.2.4 Autoencoders

Autoencoders belong to the unsupervised learning class of the algorithms (113, 131-133). They leverage neural networks' ability to work as the universal approximator (134). Their training phase does not differ from feed-forward neural networks, but their unique point lies in the layers' architecture.

The simplest autoencoder consists of three layers: input layer, hidden layer, and output layer. What differentiates it from the conventional neural network is the hidden layer size, which is much smaller in comparison to the other layers. Its ground truth values are the same as the

input values, and the loss function $\mathcal{L}$ is defined as the discrepancy between predicted values $\hat{x}$ and the original data points x:

$$\mathcal{L}(x, \hat{x}) = ||x - \hat{x}\,||^2$$

The small size of the hidden layer eliminates the possibility of learning a trivial mapping $\hat{x} \rightarrow x$. This forces the neural network to learn a compact encoding for the input samples based on patterns present in the data.

One of the popular modifications to the base algorithm, stacked autoencoders, introduce additional hidden layers that increase the number of transformations performed for the encoding and decoding representation from the smallest hidden layer h.

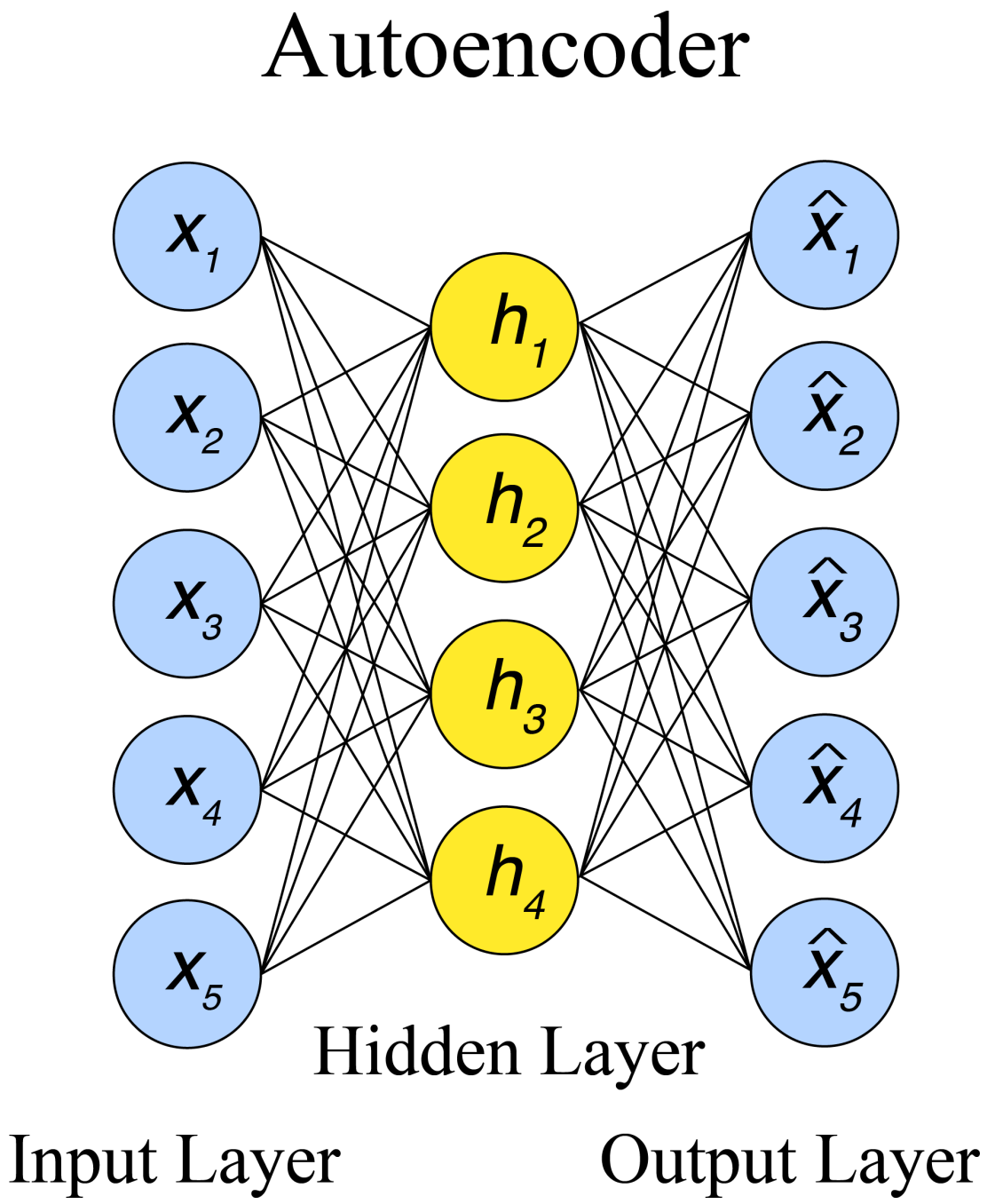

Figure 2.5. The architecture of the autoencoder with a single hidden layer. The crucial point of this neural network is that hidden layer has a smaller dimensionality than an input layer, making trivial one-to-one mapping between inputs and outputs impossible. It forces the network to compress information and learn patterns present in the studied dataset. Based on (128).

2.2.5 Transformers

The seminal paper named 'Attention is all you need' introduced a novel approach to sequence-to-sequence learning (135). It relies on the parallel run on the input and output, also known as sequence to sequence (Seq2Seq) learning. The output sequence is shifted to the right by one timestamp in order to avoid learning of the trivial mapping.

The core part of this algorithm is the attention mechanism that maps key-value input to the output sequence according to the following equation:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,$$

where Q is the set of queries, K correspond to keys and V to values. d is the number of dimensions.

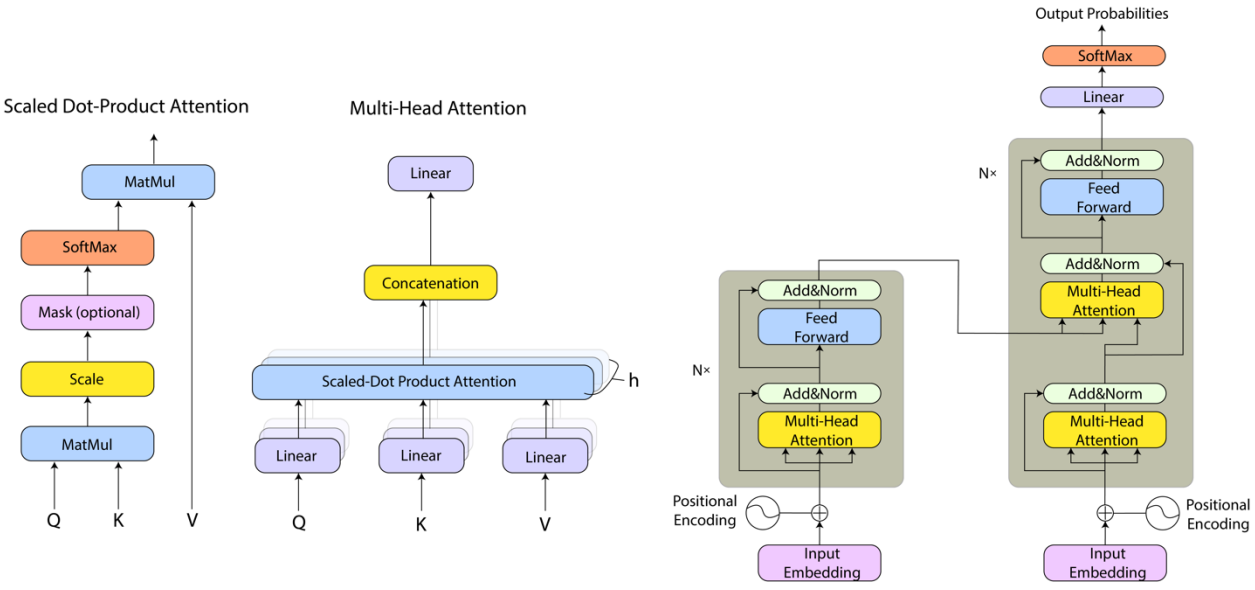

Figure 2.6. Transformer architecture. In a key part of "encoder-decoder attention" layers, the queries originate from the previous decoder, and the encoder outputs memorized K, V. The queries, keys, and values concepts are inspired by the information retrieval problem; in the transformers architecture, K, Q, and V are incorporated into the scaled dot product, which is the basic building block of this type of neural network. In the context of global attention, keys serve as a static knowledge field when queries help to determine the most relevant items from it. Then corresponding values associated with keys are scaled in proportion to the estimated relevance. Adapted from (116).

The key feature of this architecture allows DNN to learn the relevant context for each input position. Such an approach reduces the complexity of the learning problem and caters to the individual needs of the corresponding language pairs.

2.2.6 Deep Graph Neural Networks (DGNN)

Deep graph neural networks are usually implemented as CNNs that consist of two parts: graph convolution layers and 1-D convolution layers. This family of algorithms propagates information on network nodes based (121)

The most crucial part of the algorithm is the implementation of the graph convolutional layer. Depending on the input data format, convolutions can be described via various means. In (121) proposed approach is the usage of the Weisfeiler-Lehman subtree kernel (136), which is widely used for graph isomorphism checking:

$$k(G, G') = \sum_{t=0}^{h} \sum_{v \in V} \sum_{v' \in V'} \delta\left(c_v^t, c_{v'}^t\right)$$

where c_v^t is the color of vertex v during t-th iteration, h is the number of iterations, and δ is the delta function.

Alternative approaches use a physics-based Message Passing algorithm that estimates the minimal energy of the system (137).

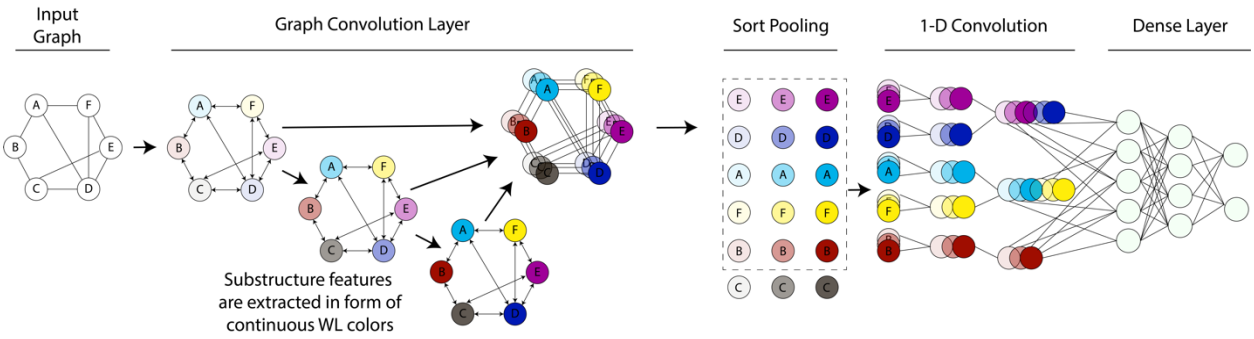

Figure 2.7. The architecture of deep Graph Neural Networks (GNN). Graph convolutional layer takes a central part in architecture and often makes use of graph kernels to extract features. Sort pooling layer combines information on values associated with each node from distinct convolution layers. 1D-convolution kernel slides along sorted vertices of the graph. The dense layer is a final step used for classification. Based on (121).

The advantage of deep GNNs is the ability to incorporate topological information into the learning model. However, the selection of a convolutional filter is a non-trivial question and largely depends on the available input data and specific task.

2.2.7 Self-supervised learning

Self-supervised learning approaches gained traction with the increase of available unlabeled data. In addition to extracting information directly from data points distribution, this family of algorithms is capable of leveraging advantages provided by the contrastive (supervised) learning approaches via adopting self-defined pseudolabels and even performing multitask learning.

The first step in a self-supervised pipeline is related to data augmentation (138-140). The common approaches used in deep learning include dropout regularization (107), batch normalization (141), transfer learning (142, 143), one-shot and zero-shot learning (144).

Second step involves an encoder used to obtain a compact representation of the input data. The choice of architecture is mostly depending on input data structure. For the image classification problem CNNs are traditionally used.

Pretext task generation is an important step that assigns pseudolabels to data according to the predefined tasks. For the image classification, pretext tasks often involve color transformation, geometric transformation, scrambling, or future prediction (in the case of sequential information such as video, audio, or text) (130).

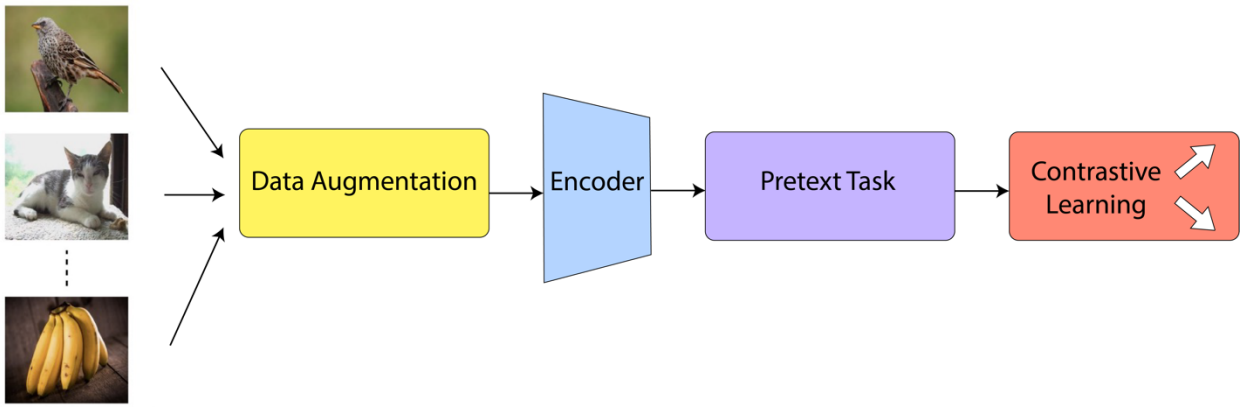

Figure 2.8. Contrastive learning pipeline for the self-supervised algorithm. It involves the creation of additional training samples based on the known but unlabeled data points. The encoder creates compact embeddings for each of the images. The pretext task is used to assign pseudolabels for each of the feature vectors, and then the contrastive learning step is responsible for the pattern detection. Adapted from (130).

Contrastive learning, a common subroutine in modern self-supervised algorithms, acts upon pseudolabels defined during the pretext task step. The goal of this step is to transform feature representation in such a way that similar samples stay close to each other in the data space while the distance between distinct samples increases (145, 146).

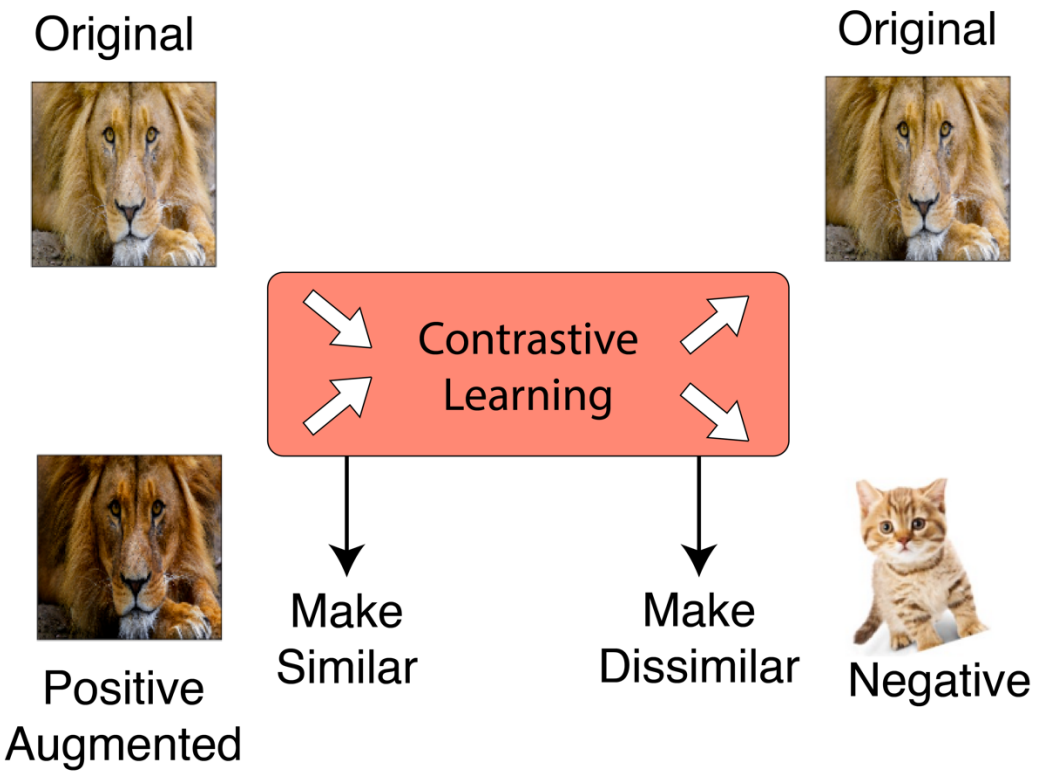

Figure 2.9. Unsupervised contrastive learning. This metric learning approach changes distances between data points in a way that images with the same pseudolabel based on the pretext task are minimized and distance between images with different pseudolabels is maximized. In addition to the described features of unsupervised contrastive learning, supervised variation makes it possible to combine multiple original images into a single category, e.g., cats, dogs, etc. Adapted from (130).

2.2.8 Deep learning in structural and network biology

Deep learning methods demonstrated a remarkable ability to derive spatial invariants, such as SO(3) rotation group (147) from the image and video data. Therefore, a natural area of application of the neural network methods in the area is related to the special information. Indeed, the recently published AlphaFold model (148, 149) for the sequence-based protein structure prediction took the lead at CASP13 competition. Such results signified an astonishing leap in an *ab-initio* protein modeling problem (150); however, researchers caution from declaring the problem of protein folding solved(151).

TorchDrug (152) is a powerful framework for drug discovery that includes molecular datasets, knowledge graphs and integrates them with a plethora of deep graph learning networks such as GraphAF (153), ChebNet (154), InfoGraph (155).

A pre-trained model for the regulatory genomic data, GeneBERT (156), incorporates omics data across multiple modalities – sequence information, regulatory region information, and ATAC-seq datasets. This model is based on transformers deep learning architecture and solves a wide range of problems, including those relevant to structural biology – transcription factor binding sites classification, disease risk estimation, and RNA splicing site prediction.

2.3 Methods to mitigate inherent biological bias

Machine Learning (ML) methods strive to extract information and complex relationships from the datasets to solve a multitude of real-world problems: disease diagnostics, fraud detection, machine translation, and image recognition. During the last decade, they produced with the improvement in hardware technology and increase in computational power, we witnessed an advent of ML models in various scientific fields such as physics (157, 158), chemistry (159), biology (160), and finances (161). Models obtained via data-driven approaches are becoming ubiquitous commodities accessible with a touch on your smartphone.

Such success is attributed to the various aspects of ML models: the ability to quickly identify patterns, even in high-dimensional datasets, little need for human intervention (162), capacity to approximate solutions to NP-complete problems (163). However, many models are unable to achieve a stellar performance reported by the creators after the transfer from the testing environment into the real world (164, 165). The main reason behind this is twofold. The first reason, ML models are based on the assumptions about data (166) that shape the algorithm and

influence outcomes. This fact does not pose a problem on its own, as they remain constant during the training and exploitation stage, and the assumptions are necessary to make predictions. The second reason, the structure of the real-world data is often different from that of one of the training sets. Because of this ML model may learn biases along with the relevant patterns and miss relationships it did not observe during the training stage. Recent results also suggest that ML models trained on high-dimensional datasets (more than 100 features) work almost exclusively in extrapolation mode, i.e., most of the real-world data points lie outside of the convex hull defined by the training set (167). This fact highlights the importance of learning underlying rules that govern relationships between samples.

Semi-Supervised Learning (SSL) methods attempt to mitigate this issue by extending the model to the unlabeled samples (168-170). However, traditional SSL approaches rely on the assumption that both labeled and unlabeled samples are drawn from the same distribution, which is not always the case.

As the machine learning methods strive is to increase generalization, accounting for distribution mismatch in data it would be applied to is another challenge we have to address. In machine learning, this problem is known as domain adaptation (115, 171-173).

2.3.1 Transfer learning for biological applications

Transfer learning methods are extremely useful in settings with limited information when new labeled data to obtain, as they address generalization improvement based on underlying data structure or previously trained models.

One of the rapidly developing biological fields – RNA-Seq analysis – is heavily dependent on data modeling stages. These algorithms operate in extremely high dimensions and

have to make a distinction between technical noise and biological variability. Transfer learning algorithms applied to this problem ensure higher cell type sensitivity for the clustering (174, 175) and classification (176) applications, data imputation (177), denoising (178), and batch effect correction (179).

Transfer learning could be used to guide the ML algorithm to learn joint tasks across different modalities, such as electroencephalographic (EEG) and electromyographic (EMG) data (180), scRNA-Seq, and scATAC-Seq (181).

Disease diagnosis is a classical ML application in the biological domain. Recently a number of algorithms have been introduced that benefit from a large amount of available medical images for diagnostic applications, including cancer, cardiovascular diseases, and neurological disorders (143, 182-186). Other clinical advances include survival prediction (187) and drug sensitivity estimation (188).

There were attempts to estimate RNA expression levels directly from the biological material slice images (189), potentially curtailing the run of the expensive and time-consuming experiments. Translational studies can significantly benefit from the transfer learning by using identifying relevant regulatory mechanisms between model organisms and humans (190).

Transfer learning approach for alternative splicing functional effects prediction

Alternative splicing is an RNA regulatory process that is responsible for the emergence of multiple distinct gene products from the same gene. It can introduce significant structural modifications due to the inclusion or exclusion of exonic and intronic regions (191), induce functional changes by perturbing protein-protein interaction networks (192) and diversification of gene interaction capabilities (64).

Currently, the amount of experimentally validated information on alternatively spliced isoforms interaction is exceptionally scarce, numbering in under 2,500 interactions (64). Human Genome Project (193) estimates the number of unique genes encoded in human DNA as ~20,000-25,000. Even if we exclude non-coding genes, the number of potential interactions will reach hundreds of millions. On average, each gene has seven distinct isoforms, bringing the total number of transcripts up to 150,000 (194), further increasing potential interactions number by two orders of magnitude. For instance, one of the latest comprehensive PPI databases, Human Reference Interactome Map, which combines experimentally validated interactions with literature curated references, has information on ~64,000 interactions from ~9,000 proteins. (195). Based on these estimates, comprehensive experimental coverage for isoforms interaction would not be feasible for an extended time period. Coupled with the fact that PPIs are highly relevant for the studies of cancer (196, 197), neurological disorders (198-200) and cellular regulation mechanisms (198, 201), it emphasizes the importance of having a reliable computational method for estimating the functional impact of alternatively spliced isoforms on PPI network.

2.3.2 Domain Adaptation Problem

Domain Adaptation is a subtype of transfer learning method used when classification or regression problem remains unchanged, but data come from multiple generative models (Fig. 5). This situation often arises in a real-world problem. For example, image recognition tasks can be affected by the hardware used to take photos, and this approach can be used to fine-tune the ML model for a specific smartphone model.

Data granularity classification

Depending on the hypothesis about the underlying structure of data, we can classify domain adaptation methods into four categories, sorted in the order of complexity increase: 1) Single source, 2) Multiple sources, 3) Multiple sources multiple targets, 4) Unsupervised adaptation. Splitting data into a larger number of domains increases the model's ability to adapt; however, it also increases the required amount of data, as from each domain should be obtained a number of samples sufficient to meaningfully describe it.

In this work, the problem of the single source domain adaptation (DA) is defined as a construction of a machine learning model based on data coming from source domain S that is intended for making predictions on target domain T, and their distributions are related but not identical (202): $P_S(X,Y) \neq P_T(X,Y)$.

The problem of the multiple source domain adaptation consists of building a single machine learning model based on training data coming from M source domains $S_1, S_2, \dots, S_M$ that is intended to make predictions on target domain T, and their distributions are related but not identical: $P_{S_1}(X,Y) \neq P_{S_2}(X,Y) \neq \cdots \neq P_{S_T}(X,Y) \neq P_T(X,Y)$. (203)

The problem of the multiple sources multiple target domain adaptations consists of building a single machine learning model based on training data coming from $\boldsymbol{M}$ source domains $\boldsymbol{S_1, S_2, \dots, S_M}$ that is intended to make predictions on $\boldsymbol{K}$ target domains $\boldsymbol{T_1, T_2, \dots, T_K}$ and their distributions are related but not identical: $\boldsymbol{P_{S_1}(X,Y) \neq P_{S_2}(X,Y) \neq \cdots \neq P_{S_T}(X,Y) \neq P_{T_1}(X,Y) \neq P_{T_2}(X,Y) \neq \cdots \neq P_{T_K}(X,Y)}$.

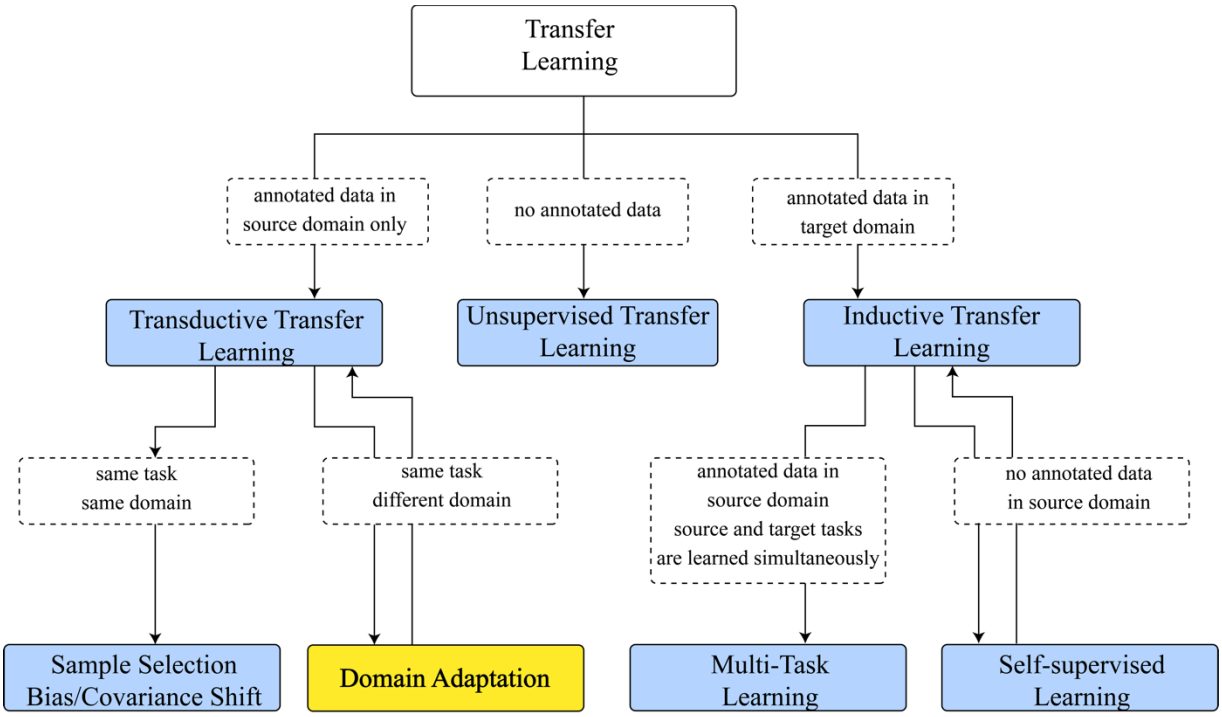

Figure 2.10. Transfer learning methods classification. Domain Adaptation is a subtype of transductive transfer learning that aims to generalize learned task to the data from different domains. Adapted from (204).

The common assumption in the domain adaptation problem is the covariance shift which stipulates that only distributions of the input features differ in the two domains: $P_S(X,Y) = P_T(Y|X)$ but $P_S(X) = P_T(X)$(202). Violating it results in the particularly challenging variant of the DA problem, unsupervised domain adaptation, which arises when there are no available labeled data points from the target domain (205).

Learning approach classification

Domain adaptation methods can be further systemized based on the learning approach (Fig.6). The traditional approach to this problem can be separated into data-centric, model-centric, and hybrid methods. Data-centric methods aim to determine the most critical data points for different domains and weigh them accordingly (206, 207), rely on features pseudo-labeling (208, 209), or pre-training methods (210).

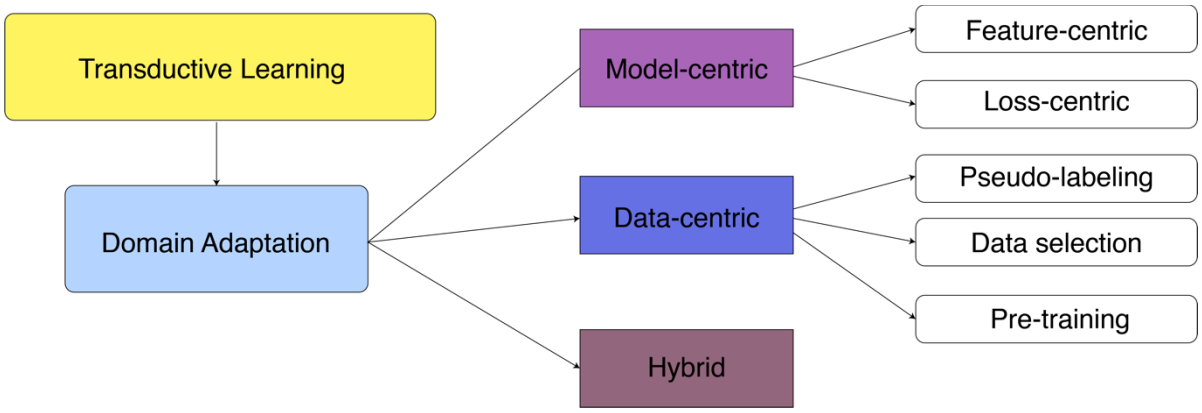

Figure 2.11. Domain adaptation methods categorization based to the learning approach. Two main categories differ based on whether they are focused on amending the model or on adopting the data. Hybrid methods include various combinations of approaches from the aforementioned categories. Adapted from (211).

Model-centric methods either assume the existence of the underlying feature manifold shared across domains and focus on feature transformation (212-214), or incorporate knowledge about domains into inference procedure, e.g., by applying regularization terms in order to select feature subsets that are robust across different domains (215).

Strong sides of the data-centric approaches 1) are relative independence from the underlying ML algorithm, which allows reusing preprocessed data, 2) and predictable training time, as once data was preprocessed, no additional modifications are made to the base algorithm, 3) feature and model consistency, learned relationships do not have to be further modified, which can improve model interpretability. These advantages made data-centric approaches de-facto standard in the computational linguistics field (108, 216). However, data-centric methods require a significant amount of labeled training samples, which restricts their area of applicability.

Feature-centric methods depend on finding common descriptions for the samples from different domains, which is also referred to as feature alignment. The straightforward approach to this problem would dictate that we can find a subset of representative features that behave

similarly across domains, but such an approach results in tremendous information loss. Due to this limitation, feature alignment algorithms were developed.

Feature alignment

Feature alignment class of methods rely solely on the features data structure to define a transformation that keeps

Structural correspondence learning (217) defines feature alignment as an algorithm that determines pivot features – a robust subset of features relevant to ML problems that behave similarly across different domains. For each feature in a source space, a one-class anomaly detection model is created and then features are filtered based on classification results. Those results were improved by the Spectral Feature Alignment algorithm (218). This approach first defines the distance between features as mutual information

$$\boldsymbol{I(X^i; D) = \sum_{d \in D} \sum_{x \in X^i, x \neq 0} p(x,d) \log_2 \left(\frac{p(x,d)}{p(x)p(d)} \right)},$$

where $\boldsymbol{D}$ is domain variable, $\boldsymbol{X^i}$ is a single feature. The higher the mutual information between the feature and domain label, the more domain-specific it is. Based on the mutual information score, domain-independent features are identified. After this, a bipartite graph is constructed between the two obtained feature subsets where domain-specific and domain-independent features always share an edge. Then spectral clustering algorithm is applied to this graph. This subroutine constructs the Laplacian matrix:

$$\boldsymbol{L = D^{-1/2} A D^{1/2}},$$

where $\boldsymbol{A}$ is the weighted adjacency matrix of the bipartite graph, $\boldsymbol{D}$ is the diagonal matrix with $\boldsymbol{D_{ii} = \sum_j A_{jj}}$. First $\boldsymbol{k}$ largest eigenvectors of $\boldsymbol{L}$ form the matrix $\boldsymbol{U}$, which defines transformation

over the features that adjusts domain-specific predictors based on the domain-independent ones that are present in the same cluster.

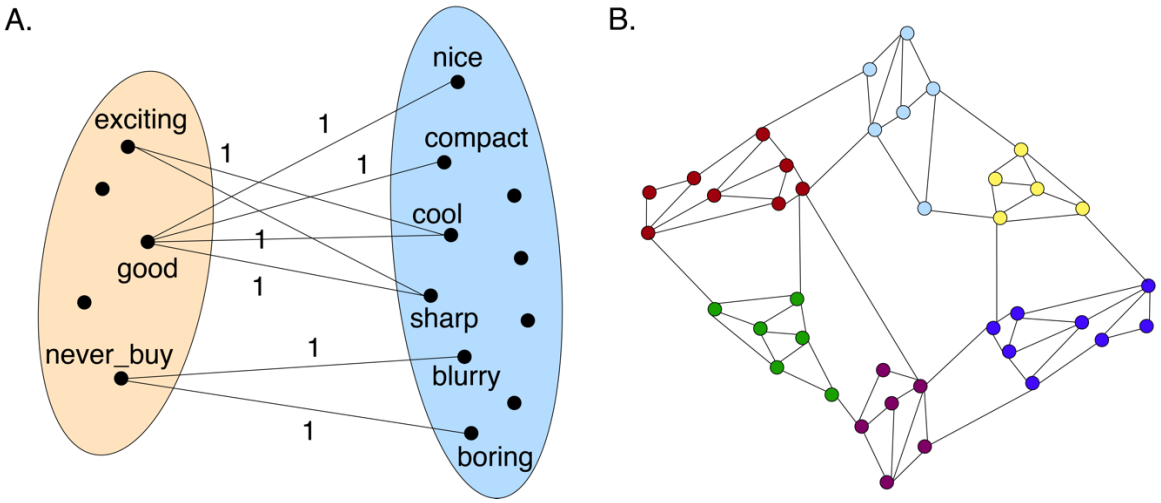

Figure 2.12. Spectral clustering-based feature transformation. (A). The bipartite graph describes the correspondence between domain-specific and domain-independent features. (B). Example of seven identified clusters in a graph based on spectral clustering. Based on (218)

Another notable subclass of feature alignment methods is subspace alignment. Sampling geodesic flow algorithm (219) and its improved version that uses kernel for improving computational efficiency (220) infer high-dimensional manifold based on the geodesical path between source and target domains. These algorithms map domains into lower-dimensional Grassmann manifold (221) as two distinct points (one per domain) using PCA (222). Then it calculates the geodesical path between them and samples a set of points along with it that correspond to the subspaces. Original features from the source domain are subsequently projected into selected subspaces, and the results are concatenated into a single high-dimensional vector. The advantage of this algorithm lies in the ability to construct a comprehensive feature space manifold that bridges the gap between two domains while retaining most of the information. The drawback comes with the extreme increase of dimensionality.

This class of methods is useful for the unsupervised domain adaptation problem as it does not depend on labeled examples (223). The same reason becomes a disadvantage when applied to the supervised or semi-supervised setting, as the obtained transformation does not account for the feature informativeness in regard to the supervised part of the problem.

Loss-centric methods. This category of methods modifies the loss function of the learning algorithm. The earliest iterations do this implicitly by the means of reweighting training samples. Another prominent class of the loss-centric methods that gained traction with the advent of deep learning methods is Adversarial Learning.

Reweighting

The representative algorithm for reweighting methods is Kernel Mean Matching (KMM) (224) which is based on the Maximum Mean Discrepancy (MMD)(225) measure. MMD is defined as the following:

$$\boldsymbol{MMD[F,p,q] = \sup_{f \in F}(E_{x\sim p}[f(x)] - E_{y\sim q}[f(y)]),}$$

where $\boldsymbol{F}$ is a class of functions $\boldsymbol{f: X \to \mathbb{R}}$ and $\boldsymbol{p, q}$ are probability distributions that correspond to different domains. The algorithm uses an unbiased empirical estimate of this measure

$$\boldsymbol{MMD_u^2[F,p,q] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h(z_i, z_j),}$$

where $\boldsymbol{Z = (z_1, z_2, \dots, z_m)}$ is $\boldsymbol{m}$ independent identically distributed (i.i.d.) random variables, $\boldsymbol{z_i = (x_i, y_i)}$. With $\boldsymbol{h(z_i, z_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)}$ this expression becomes a one-sample U-statistic (226).

Adversarial methods

Theoretical results on domain adaptation suggest existence of feature representation that makes samples indistinguishable based on the domain of origin (227). Adversarial approach for the domain adaptation attempt to achieve this goal by introducing a modified loss function (228) inspired by the Generative Adversarial Networks (GANs) (229) that iteratively minimize the discrepancy between original and synthesized data distributions. Though domain adaptation methods are not limited to the classification problem, we would present an overview specifically for the classification. Proposed approaches could be easily adapted to the regression by modifying loss function, and general ideas remain the same.

The goal of the classical adversarial modeling approaches is to create a generative model that produces non-trivial samples indistinguishable from the ground truth examples. To achieve this goal adversarial approach employ minimax game $\boldsymbol{V(D,G)}$ between two players with the competing objectives – a discriminator $\boldsymbol{D(x;\theta_d)}$ that classifies samples $\boldsymbol{x}$ into either ground truth or synthetic category and a generator $\boldsymbol{G(x;\theta_d)}$ with the objective of decreasing discriminator's performance:

$$\min_{\boldsymbol{G}} \max_{\boldsymbol{D}} \boldsymbol{V(D,G)} = \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_{data}(x)}}[\boldsymbol{\log D(x)}] + \mathbb{E}_{\boldsymbol{z}\sim \boldsymbol{p_z(z)}}\left[\boldsymbol{\log\left(1 - D\left(G(z)\right)\right)}\right]$$

Discriminator $\boldsymbol{D}$ is trained to maximize the probability of assigning a correct label to the data point, while generator $\boldsymbol{G}$ is trained to minimize it. The stopping criteria that signify a successful approximation of sample distribution $\boldsymbol{p(x)}$ are the 'loss' of the discriminator $\boldsymbol{D}$ – its inability to differentiate generated samples.

In the domain adaptation context, the goal of the adversarial method is to make sample distributions $\boldsymbol{S_i(f) = [G_f(x;\theta_f)|x\sim O_j(x)]}$ comparable. Deep adversarial neural network

(DANN) algorithm approaches this problem by learning transformation for features $\boldsymbol{f}$ that minimizes discrepancies between distributions $\boldsymbol{S_i}$ with respect to the original classification objective (228). Under covariance shift assumption (230) solution for the problem of interest would have the same performance on the source and the target domains. In order to identify transformation, DANN introduces a feature regressor $\boldsymbol{G_f(x;\theta_f)}$ that simultaneously maximizes the original classification problem learning objective $\boldsymbol{L_y}$ and minimizes learning objective $\boldsymbol{L_d}$ that distinguishes samples based on the domain of origin $\boldsymbol{O_j}$. It results in the following loss function:

$$\boldsymbol{E(\theta_f, \theta_y, \theta_d) = \sum_{i=1..N} L_y\left(D_y(G_f(x_i;\theta_f);\theta_y)\right) - \lambda \sum_{i=1..N} L_d\left(D_d(G_f(x_i;\theta_f);\theta_d)\right),}$$

where $\boldsymbol{D_y}$ is discriminator for the original classification problem, $\boldsymbol{D_d}$ is a discriminator for the domain classification, $\boldsymbol{\lambda}$ is the regularization parameter, $\boldsymbol{\theta_f, \theta_y, \theta_d}$ are model parameters for the $\boldsymbol{G_f}$, $\boldsymbol{D_y}$, and $\boldsymbol{D_d}$ classifiers correspondingly.

The strength of this approach is the extreme generality. In fact most of the neural unlabeled domain adaptation methods are based on domain adversaries (228, 231). DANN-based approaches are also highly scalable. However, they model solely feature representation that is shared across the domains. In the case when samples can be accurately classified across the domains, this class of algorithms suffers from the vanishing gradient (232).

Pseudo-labeling involves leveraging a large pool of unlabeled samples and iteratively predicting their labels, subsequently feeding them back into model. This can be achieved via bootstrapping methods: self-training (233, 234), and co-training (235). Self-training was applied

with limited success in the ALT-IN algorithm(236) for predicting alternative splicing-induced changes in protein-protein interactions.

Data selection

Pre-training approaches can take multiple shapes. The most straightforward case is the usage of a single base model trained on a large amount of data, e.g., GoogleAI or BERT (237-239), and fine-tuning it for the specific task. This approach encourages model sharing among the researchers and saves time and energy required for getting baseline results. This framework can be further modified by stacking pre-training runs together, subsequently increasing domain specialization. Multitask learning (240, 241) can be employed as the strategy to force the algorithm to learn additional objectives, presumably, related to the main problem, forcing the algorithm to account for it.

2.4 *Conclusions*

Deep learning methods are able to distill compact representation from the raw data, either with respect to the specific task or in an unsupervised manner. This is especially valuable for the biological setting because data in a vast number of applications have extremely high dimensionality that cannot be cracked by the traditional statistical approaches without discarding a large portion of information. On top of that, the relationship between variables remains highly confounded, which reduces the applicability of the linear methods and increases feature engineering difficulty. So, traditional ML methods make way for the DL application that a poised to solve these issues. They offer a promising approach to integrating multiple modalities, which is incredibly important for structural and network biological applications.

Domain adaptation is the widely used subclass of the transfer learning algorithms that aim for achieving the stable performance of the machine learning algorithms across statistically non-identical samples. There exist a vast number of methods to achieve this goal, and quite a lot of them are applicable to the DL setting.

Currently, a plethora of methods dedicated to the maximization of publicly available data usage is being proposed. Omics databanks get new information in large quantities, and our goal should be to transform it into valuable pieces of information. And we have to move quickly, as the latest experimental technologies emerge and start to push the older counterparts into oblivion, so it is essential to make the best use of freshly derived information to keep high-profile attention on precision medicine.

Chapter 3. Alternative Splicing and the Prediction of its Properties

3.1 Background

3.1.1 Introduction

Proteins are the large molecules formed by the amino acids sequence and encoded by genes. These molecules' chemical reactions and movement underlie dynamic processes in living organisms play a crucial part in digestion, hormone production, immune resistance, and tissue growth. They vary in different organisms, cells, tissue, or time points. The collective name for all proteins expressed under given conditions is called proteome. There is a large discrepancy in the size of the proteome and genome.

3.1.2 Alternative splicing and Protein-Protein Interactions

Protein-protein interaction network, or interactome, is a Facebook of proteins, the basic building blocks that underlie the cell's basic functioning. Whether we study complex genetic disorders, stem cell specialization, or epigenetic effects on an organism, those blocks form a basis for the higher abstraction levels. Protein-protein interactions (PPIs) are a glue that mechanistically ties those blocks together.

Protein-protein interactions (PPIs) underlie many key mechanisms of cellular functioning (242). With thousands of PPIs simultaneously occurring in every cell of an organism, an average protein is expected to interact with two or more other proteins forming large molecular assemblies, transporting proteins, facilitating a chemical reaction, protecting the organism from pathogens, and carrying out other basic functions (68, 243, 244). Increasing understanding of PPIs gives us an insight into molecular mechanisms of cellular pathways and regulatory

mechanisms which form the basis for the description of complex genetic disorders such as cancer, diabetes, autism, and schizophrenia (245). This knowledge is crucial for the target discovery in drug design.

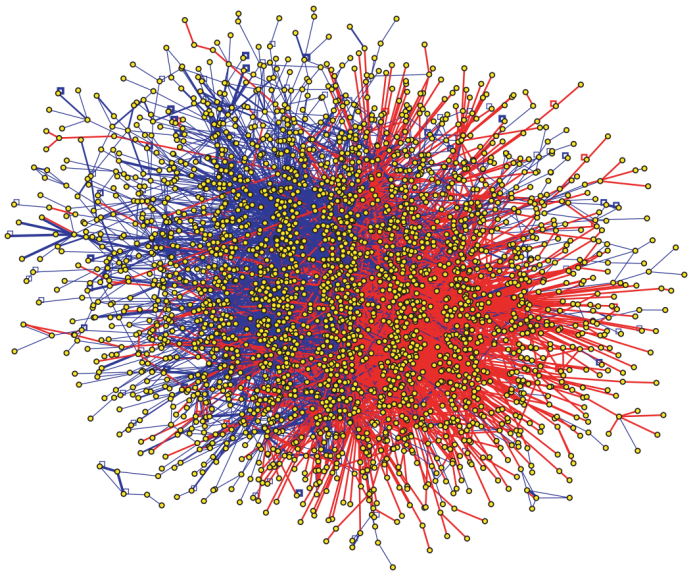

Figure 3.1. Human interactome snapshot from the [Rual, 2005]. Graph nodes (yellow) correspond to the distinct proteins, edges (blue and red) denote existing interactions. This snapshot presents a combination of literature-curated interactions (blue) and high-throughput experiment results (red). Adopted from (246)

Throughout the past two decades, there have been efforts in characterizing the experimentally confirmed PPIs by describing the structure of molecular complexes and interaction interfaces formed through the PPI (247, 248), determining a protein function that is controlled by the interaction (249), and understanding the evolutionary principles shared between the homologous interactions (250, 251). Large-scale characterization of protein-protein interactions (Fig. 3.1) using high-throughput interactomics approaches, such as yeast-two-hybrid and tandem-affinity purification/mass spectrometry methods (245, 252), has provided the scientists with new insights into the cell functioning at the systems level and allowed to better understand the molecular machinery underlying complex genetic disorders (67, 253, 254).

Although a comprehensive understanding of PPIs in an organism has important practical applications, our current collection of PPIs is still far from being complete. A combination of high-throughput experiments results and small-scale interactomics studies, like co-immunoprecipitation and surface plasmon resonance, reported in the literature, yields a collection of ~25,000 high-quality interactions (255). According to some estimates, the number of PPIs in a human cell is over 600,000 (255, 256). Currently, the level of diversity provided by alternative splicing is mostly overlooked in PPI interactomes that stick to the formula "one gene – one protein". More recently, several studies have been published that focus on studying the interaction-rewiring, edgetic, effects of genetic variations caused by genetic diseases (257, 258). The edgetic effects on the whole protein interactome of other types of variation, such as copy-number variation, epigenetic variation, and transcriptional variation, or alternative splicing, are far less studied (64, 68).

Considerable effort efforts were dedicated to filling the gaps in PPI datasets. They collectively called PPI prediction methods. These applications often define PPI prediction problem as a classification task and leverage supervised learning approaches, including deep learning, where the training set includes labeled variants for which the function is known and is experimentally validated. The supervised learning approach is designed to benefit from the labeled training set in order to provide an accurate prediction, however, the labeling (*i.e.*, functional annotation) may not be feasible for large datasets required by many supervised methods.

Recently, a high-throughput interactomics study has demonstrated a widespread interaction rewiring by the alternatively spliced gene products (64). In some cases, new interactions were shown to be formed. The experimental approach covered ~10% of human

protein-coding genes and provided data on 2,503 interactions including alternative isoforms from five healthy human tissues: brain, heart, liver, placenta, and testis. In spite of being very accurate, these large-scale experiments are time-consuming and expensive. When we include alternatively spliced variants as an additional variable into the search space for potential interactions, this results in up to ~200 million potential PPIs that need to be assayed (255). Up to date, no reliable computational approaches that predict the edgetic effects of alternatively spliced variants have been introduced. We found that existing sequence-based PPI prediction methods have difficulties providing viable characterizations of rewiring events. Thus, there is a need for a cheaper and faster, *in-silico*, approach that is AS-specific.

As of today, our understanding of the functional implications that alternative splicing may have on molecular regulatory processes is quite limited. It has been observed that AS often has a direct influence on protein machinery: the isoforms frequently behave like separate proteins rather than follow the functional designation of the main splice variant (64,) but large-scale effects it has on organism remain unseen. Cell specialization is largely dependent on alternative splicing, but we are not able to give a clear picture of it because even cutting-edge scRNA-Seq currently unable to provide the sequencing depth and coverage adequate for the whole-transcriptome analysis of AS events. Alternative splicing is a dynamic process influenced by a variety of factors. In order to pinpoint disease-related events one needs to be able to dissect it in different patterns related to the cell cycle, circadian clock, tissue genesis, or other biological distinct groups that may influence the alternative splicing.

3.1.3 Experimental methods of studying alternative splicing

RNA-Seq

The primary source of experimental data for high-throughput alternative splicing studies is an RNA-Seq, next-generation sequencing technology that quantifies the amount of mRNA material present in a biological sample. It fragments input RNA material and converts to the cDNA fragments (reads), which are aligned to the reference genome or assembled into a new one using *de Bruijn* graphs. Increasing the number of reads and reads length helps to make precise estimates. Gene expression profiling experiments may require 5-25 million reads.

The ability to describe alternative splicing events requires much higher depth and, depending on the application, may take up to 200 million. Original RNA-Seq technology required a significant amount of biological material. In most experiments, all sample material belongs to the specific tissue and contains a mixture of cells that constitute this tissue. Recent advances in technology led to the increase in resolution level – an ability to quantify expression levels of the individual cells. It brings a clear advantage - the ability to study distinct cell types. But it comes with considerable drawbacks – scarcity of biological material in each experiment increases the noise level in data and restricts sequencing depth. The latter comes in the way of conducting alternative splicing studies.

Single-cell RNA-Seq

Single-cell RNA-Seq (scRNA-Seq) methods are the cutting-edge techniques for transcription level quantification that provide valuable insight into molecular processes occurring in individual cells.

However, scarcity of biological material limits transcriptome coverage and restricts sequencing depth. For example, full-length scRNA-Seq protocols that use a switching mechanism at 5'End of RNA template (SMART) can reach ~40% coverage, when high-throughput 3'-end approaches provide roughly 10% and are not feasible for the alternative splicing events detection (259, 260). These are intrinsic technological limitations as an abundance of PCR duplicates make it unfeasible for standard protocols to increase the depth to more than 1,000,000 reads (260, 261).

There are three main categories of scRNA-Seq protocols: long-read technologies, Smart-based and Unique Molecular Identifier (UMI)-based. *Long-read* approaches such as PacBio and Nanopore allow capturing an entire transcript molecule that is exceptionally suitable for isoform detection tasks as we would not miss any types of alternative splicing events. This family of methods, though, has significant downsides – low sequencing depth, the limited number of cells (four to six) (262, 263) and a high rate of sequencing errors. *UMI-based* methods, such as inDrop (264), Drop-seq (265) and MARS-seq, provide results based on short reads and unique molecular identifiers (UMIs) attached to them for mitigating amplification bias. These protocols can process a high number of cells, accurately quantify expression level but have medium sequencing depth, and are difficult to use for isoform quantification purposes due to the technical noise (266). *Smart-based* protocols (SMART-seq, SMART-Seq2) provide a convenient middle ground performance between long-read and UMI-based approaches in expression quantification and sequenced number of cells with the exceptional sequencing depth and low error rate.

3.2 Alternative splicing prediction – literature overview

Due to the alternative splicing regulatory mechanisms, different exons may be either included or excluded from the final mRNA product. The quantification of reads helps determine

the percentage of splicing inclusion (PSI, Ψ) for each alternative exon. As separate reads are obtained from the experiment, they are aligned across the reference genome in order to get a cumulative signal. Then this signal has to be deconvoluted into separate transcripts. Such quantification relies on splice junctions – reads that simultaneously belong to the two distinct exons. Alternative splicing events such as intron retention, alternative 5' donor site, and alternative 3' acceptor site introduce significant changes to the known gene structure (Fig.3.2).

Delineating gene structures and, subsequently, alternative splicing isoforms is dependent on precise detection of the splice junctions. After this stage is complete and initial alignment is obtained, we have to deconvolute cumulative signal into multiple channels, namely, alternative splicing variants. At this stage, as it is impossible to differentiate between molecules each read originated from, the typical strategy is to employ statistical and machine learning-based methods.

Predicting alternative isoforms is a multi-stage process that has to overcome multiple technical challenges, including incorrectly detected splice junctions, technical noise, and data scarcity, in the case of scRNA-Seq samples.

3.2.1 Refining splice junctions

Sequence alignment is one of the standard steps in RNA-Seq pipelines. While modern alignment algorithms (267-269) do not require explicit annotation of exon coordinates and can perform *ab initio* alignment, thus, being able to detect novel splice junctions based on the evidence. Despite employing sophisticated strategies, such as de Bruin graphs (270), hierarchical indexing (267), or Maximal Mappable Prefix (269), aligners struggle to avoid false positives. It is well within expectations, as the probability of mapping a random short read to the large reference genome is high. Technical noise, such as intergenic and intronic unspliced RNA,

spurious RNA fragments from the library preparation further contribute to the challenges in splicing junction identification.

Traditional strategies for dealing with false positives are based on two measurements: the number of samples in which a given splice junction was detected, and the number of reads aligned to the specific splice junction. A drawback of such an approach lies in the difficulty of determining the cutoff threshold (271).

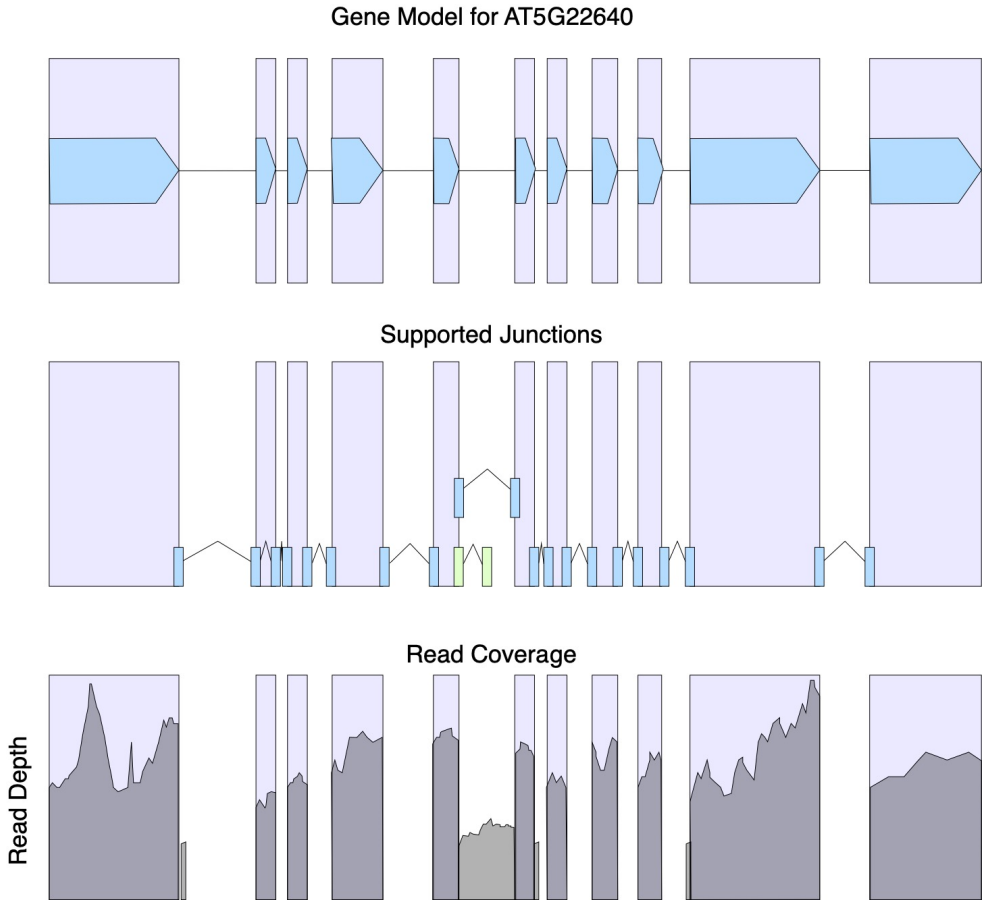

Figure 3.2. Example of splice junction detection for the model read coverage of AT5G22640 gene. *Top*. AT5G22640 gene model with experimentally validated splice junctions. *Mid*. Grey splice junctions correspond to the known gene structure; novel splice junctions are denoted green. *Bottom*. Short RNA-Seq reads mapped to the reference model. Based on (272).

ASPIC (273) is the heuristic approach based on the progressive alignment of the transcriptional data to the genomic sequence that solves multiple EST factorization compatibility problems. This algorithm performs a multiple sequence alignment of available transcript data (including EST and full-length cDNA) to the relevant genome sequence. Its advantage is the ability to incorporate multiple sources of genomic information. The main limitation is the relatively rigid model, as ASPIC detects the set of introns that minimizes the number of splicing sites (274). Such an approach does not account for the subtle changes in splicing regulatory factors in individual samples.

SpliceGrapher (272) is graph assembly-based method for splice junction prediction. Acceptor site from each newly predicted exon becomes the next primer that is compared with all other exons. The advantage of this algorithm lies in the ability to incorporate expression sequence tag (EST) data along with RNA-Seq information. As a major drawback, it has exceedingly high missing scores (275).

Multiple feature-based methods that use SVM (SpliceMachine (276), DM-SVM (277)), semi-supervised ensemble methods (STED (278)), dbSNFP (279)), and Hidden Markov Models (AUGUSTUS (280)) were introduced. Overall, validation studies on experimental data (275, 281) conclude the combination of multiple tools can filter out low-probability splice junctions without losing true spliceosomes. The common drawback of this category of methods is the loss of information from the training data, as the feature extraction process is prone to missing complex dependencies.

Current state-of-the-art methods for the splice junction detection, such as SpliceAI (282), DeepSplice (271) and SpliceVec (283), take advantage of the deep learning methods' ability to find complex dependencies from the input data. It allows them to simplify input data. For

example, SpliceAI can predict splice junctions based solely on the sequence information. Feature learning step in the deep neural networks helps capture even minute changes like single nucleotide variants (SNVs) and adjust splice junctions accordingly. The downside of such methods is training data bias. Thus, benchmarking studies (165) found a significant discrepancy between performance on the general populations' datasets they were trained on, such as GTEx (75), and on smaller clinical datasets. They suggest that these tools facilitate detection of the non-pathogenic neutral splicing variants but exhibit limited applicability in the clinical setting.

Because genetic sequence information is sufficient for the current state-of-the-art splice junction prediction algorithms, there is not a lot of need to specifically adjust them to the scRNA-Seq data.

3.2.2 Alternative isoforms prediction

The alternative isoform prediction problem bears a resemblance to the demultiplexing problem, as we have entire read counts corresponding to the gene level assembled (Fig. 2), which is also known as a gene expression value.

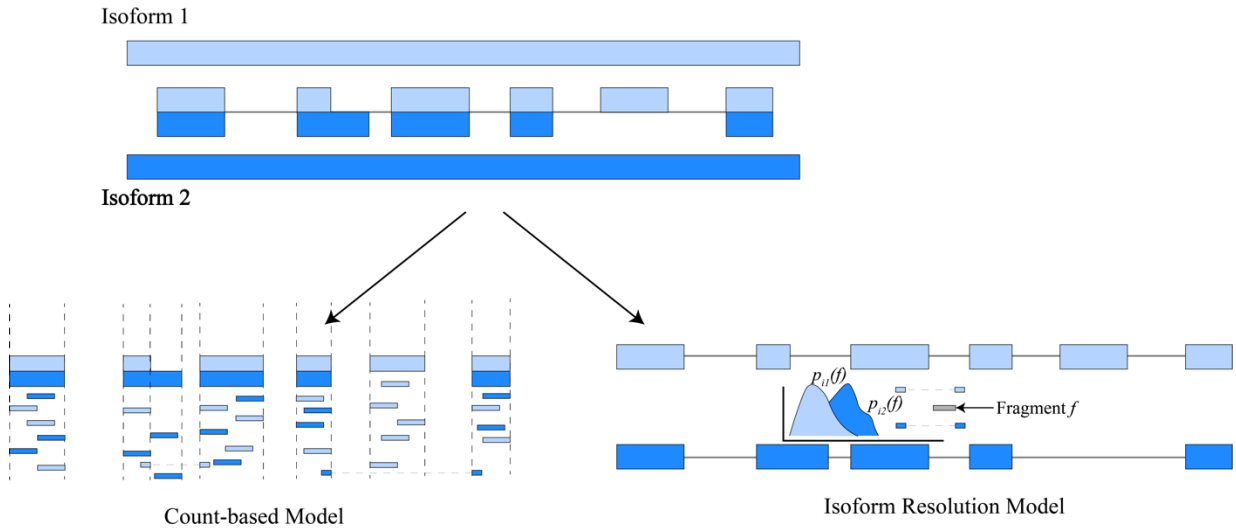

Figure 3.3. Two different approaches for sequence mapping – count-based model and isoform resolution model. Count-based approaches keep track only on the number of spliced exons while isoform resolution model attempts to assign expression levels for each individual transcript. Adapted from (284).

In order to get transcript-level information, we would have to determine the number of reads that come from each particular isoform. But unlike with signal demultiplexing, we cannot make a distinction between signal sources, and instead, we have to use computational models. Two large classes of computational models are count-based models and isoform resolution models (Fig. 3.3).

Count based models

This class of models relies on the approaches used to quantify transcripts with a single isoform. They are commonly used in differential gene expression studies. To accommodate the inclusion of the transcript-level information, those methods use smaller counting units, like exons. Those units can be exonic regions (possibly, truncated), like in the case of DEXSeq (285), or splice junctions, in the case of MATS (286). This family of methods does not provide a direct estimation of each transcript abundance, but obtained results can be used for the differential studies. It was shown that exon count units accurately reflect information on the alternative splicing as long as no isoform can be constructed from other isoforms (287).

Estimating differential alternative splicing between samples remains an essential problem for comparative studies. One of the recent models, DARTS (288), infers differential alternative splicing based on the deep neural network model.

Isoform resolution models

Isoform resolution models predict optimal isoform composition and directly provide abundance for each separate alternatively spliced variant. One of the pioneering tools from this category, Cufflinks, maximizes the likelihood of the isoform proportion vector q in respect to the observed set of aligned reads **(289)**.

RSEM (290), Kallisto (291) and Salmon (292) are lightweight quantification models that are able to function with or without prior alignment. Salmon makes additional corrections for the biological GC bias in their generative model, achieving even more precise results.

Single cell-specific methods

Despite high technical noise and lack of coverage in scRNA-Seq data, a number of methods for alternative splicing events detection were developed: SingleSplice, BRIE (Fig.3.4), Expedition, (293-295), SatuRn (296), SCATS (297). They use different approaches to overcome scRNA-Seq limitations, which result in distinct isoform expression metrics. The major drawback of those methods is the inability to provide one expression value per isoform (266), which is only useful for differential studies.

Previous studies of alternative splicing detection in scRNA-Seq predominantly cover SMART-based protocols, though some techniques attempt AS events detection in UMI-based data (297). A simulation-based study (298) on isoform quantification of SMART-based protocols concludes that AS events detection does not exhibit a significant drop in performance in comparison to the bulk RNA-Seq methods. Still, in more recent work, authors accounted for dropout rates and technical noise and concluded that current methods produce highly confounded results (299). Another study (300) suggests that single-cell RNA-Seq methods capture of

alternative splicing are prone to labeling the appearance of two separate isoforms as mutually exclusive due to the low level of coverage, which suggests that previous observations of bimodal pattern splicing patterns among supposedly homogeneous cells (294, 298, 301) is a technical artifact and not a true biological observation. UMI-based protocols (e.g., inDrop (264), DropSeq (265)) provide even shallower sequencing depth, so these results also carry similar implications for datasets obtained via those methods.

Challenges

Biological variability in study subjects further increases data analysis difficulty. Cells used for library construction are destroyed at random time points so that they may be at the different phases of the cell cycle, carry distinct molecular signatures of the circadian clock or simply vary in size (295, 302). In order to mitigate those effects, statistical models have been developed (302) along with the concept of pseudotemporal ordering (303). There are also methods for batch-effect correction that remove technical artifacts (304).

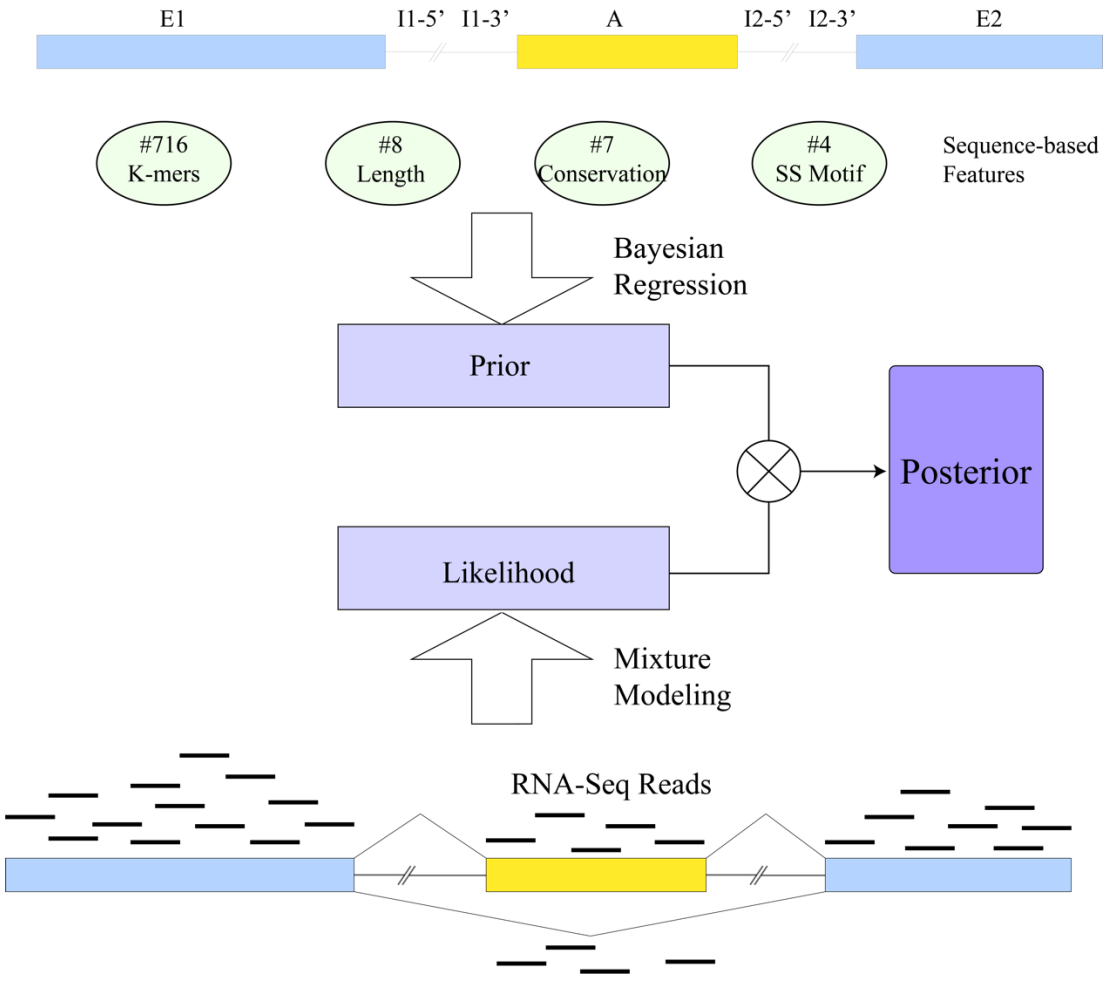

Figure 3.4. Workflow of BRIE, isoform resolution model for scRNA-Seq. Prior distribution of the percentage spliced in (PSI) exons for each individual exon is via Bayesian regression based on such features as K-mers, sequence length, conservation, and splice site motifs. It is then adjusted according to mixture modeling likelihood based on the RNA-Seq reads in order to obtain posterior distribution of PSIs. Adapted from **(295)**.

Data scarcity currently remains the most significant challenge for scRNA-Seq isoform quantification. It leads to amplification of such effects as the read quality, unwanted biological noise, PCR amplification bias, and dropouts. Those effects become so profound that some studies are unable to differentiate between biological signal and technical noise **(305)**. The most significant challenge data scarcity introduces is the quantification of the expression level for each separate isoform. Currently, this problem remains unsolved. Nearly all genes in a human cell undergo alternative splicing, a process that can obtain a diverse pool of isoforms from the same

gene through selective inclusions and exclusions of the gene's exons and introns (65). Single-cell and bulk RNA-sequencing data allow scientists to unveil a genome-wide gamut of post-transcriptional variation caused by alternatively spliced isoforms that could be specific to tissue and cell type, developmental stage, disease phenotype, and many other factors and conditions (306-308). Alternative splicing has also been shown to alter a plethora of protein functions (70). The range of functional variation between the alternatively spliced isoforms may vary drastically: from a complete loss of original function, due to misfolding of the alternatively spliced isoform and its removal by the cell degradation mechanisms to a subtle difference in the protein function and to even gain of a new function, due to the alternative isoform's new exon encoding a new functional protein domain (64, 309-312). Unfortunately, large-scale functional regulation by alternatively spliced isoforms remains poorly understood because of the lack of system-wide experimental studies. Recently, a high-throughput interactomics study demonstrated a widespread interaction perturbation caused by alternative splicing (64). In some cases, new interactions emerge driven by new exons in the alternatively spliced isoforms. Despite being very accurate, these large-scale experiments are time-consuming and expensive, which affects their coverage. Thus, there is a need for a cheaper and faster *in silico* approach.

3.3 Predicting protein-protein interaction rewiring

3.3.1 Introduction

Nearly all genes in a human cell undergo alternative splicing, a process that can obtain a diverse pool of isoforms from the same gene through selective inclusions and exclusions of the gene's exons and introns (65). Single-cell and bulk RNA-sequencing data allow scientists to unveil a genome-wide gamut of post-transcriptional variation caused by alternatively spliced isoforms that could be specific to tissue and cell type, developmental stage, disease phenotype, and many other factors and conditions (306-308). Alternative splicing has also been shown to alter a plethora of protein functions (70). The range of functional variation between the alternatively spliced isoforms may vary drastically: from a complete loss of original function, due to misfolding of the alternatively spliced isoform and its removal by the cell degradation mechanisms to a subtle difference in the protein function and to even gain of a new function, due to the alternative isoform's new exon encoding a new functional protein domain (64, 309-312). Unfortunately, large-scale functional regulation by alternatively spliced isoforms remains poorly understood because of the lack of system-wide experimental studies. Recently, a high-throughput interactomics study demonstrated a widespread interaction perturbation caused by alternative splicing (64). In some cases, new interactions emerge driven by new exons in the alternatively spliced isoforms. Despite being very accurate, these large-scale experiments are time-consuming and expensive, which affects their coverage. Thus, there is a need for a cheaper and faster in silico approach.

I introduce a supervised and semi-supervised machine learning approaches that predict if an alternatively spliced isoform will disrupt a protein-protein interaction originally formed

between a reference isoform and another protein (Fig. 3.5). Machine learning has previously been used in bioinformatics applications for the characterization of functional effects caused by genetic and post-transcriptional variations (64, 257, 313-315).

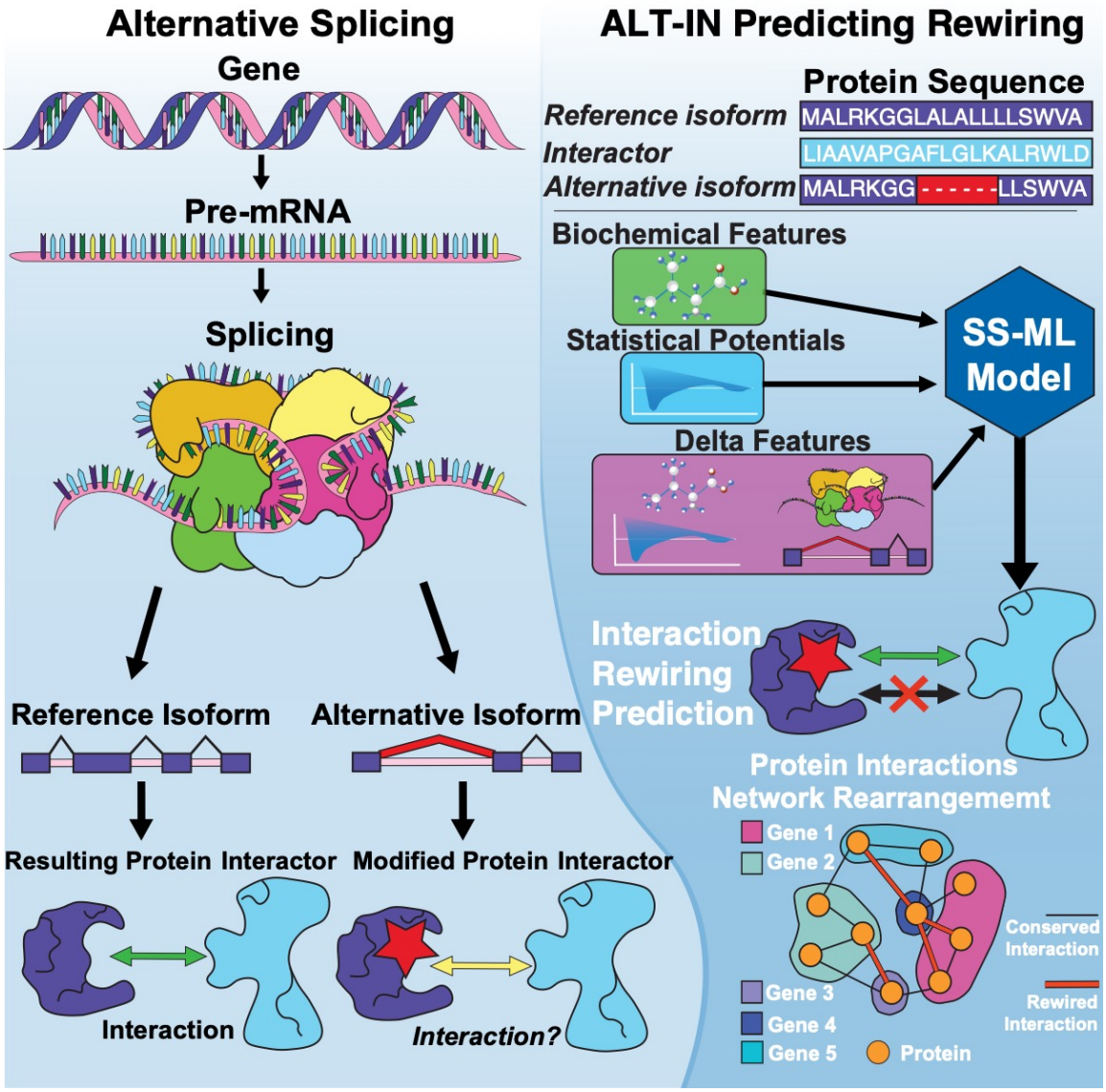

Figure 3.5. High-level overview of the protein rewiring task. Left panel describes potential rewiring event that may take place due to the alternative splicing. Right panel depicts out approach to the problem that combines information from the reference isoform, interactor partner, and alternative isoform by extracting three main groups of features: biochemical fingerprints, statistical potentials for domain interaction, and delta features that represent difference between reference and alternative transcripts. Then semi-supervised model is trained on the extracted features, that later can be used to assess changes in PPI networks.

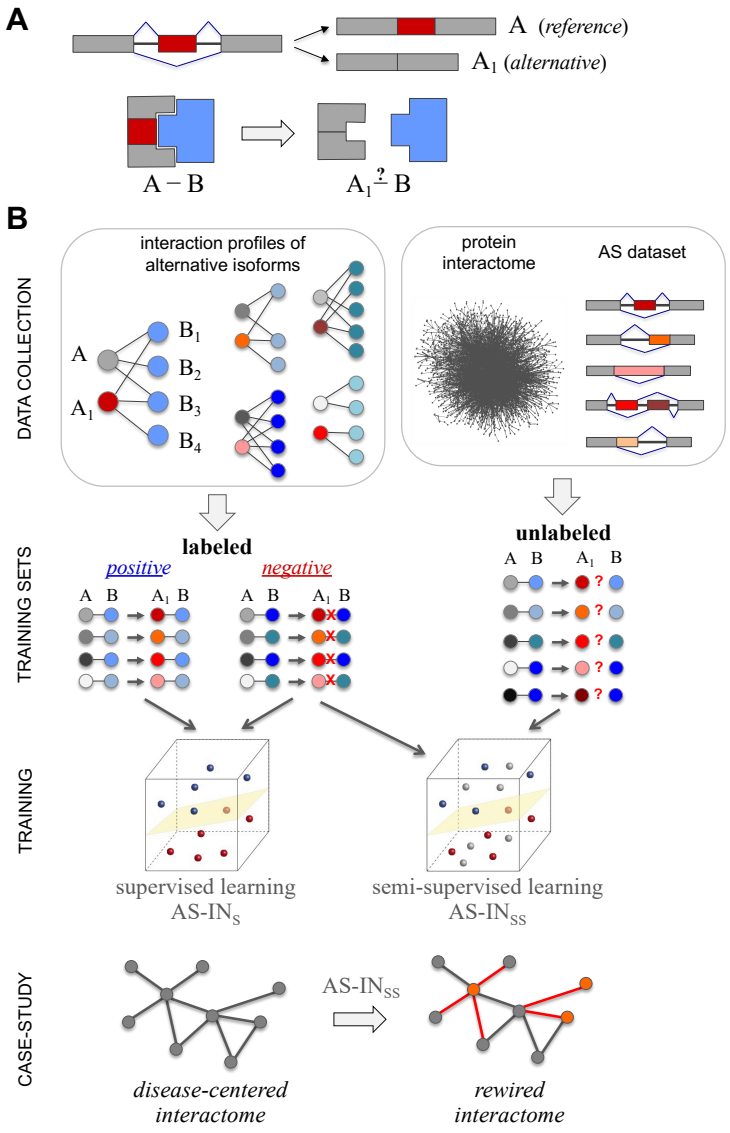

Figure 3.6. Overall ALT-IN approach. A. Characterization of edgetic effects of AS on PPI formulated as a binary classification problem, where A is a reference isoform, A_1 is an alternative isoform, and B is an interaction partner of the reference isoform. B. Four basic steps of computational study.

The characterization of functional effects can often be defined as a classification task (e.g., if a mutation alters a protein function or not), and thus be tackled by supervised learning approaches, including deep learning. By design, the supervised learning approach benefits from a labeled training set (e.g., experimental functional annotation of genetic variants) to provide an

accurate prediction; however labeling may not be feasible for large datasets, which hinders the ability of many supervised methods to generalize well. As an alternative, semi-supervised learning can be introduced, where, in addition to a small labeled training set, the classifier can benefit from the knowledge of a large unlabeled dataset, i.e., a dataset consisting of alternatively spliced isoforms with unknown functional effects (Fig. 3.6) (316).

Predicting the functional impact caused by alternative splicing has only recently been approached by machine learning, with protein-protein interactions being the main focus due to their functional importance and abundance in the cell (317, 318). One of the two currently existing methods is limited to proteins with annotated protein domains (317), while the second method leverages a deep learning approach trained on data produced by a generic PPI prediction method (319). Most importantly, the performance of each of the two methods leaves substantial room for improvement in terms of methods' accuracy and coverage.

3.3.2 Datasets and feature statistics

The first dataset (D1) used to generate the training set for the supervised learning classifiers included 1,837 protein-protein interactions (PPIs) from 638 genes with 881 alternatively spliced isoforms. The number of isoform products for each gene ranged from 2 to 7, with an average of 2.3 isoforms per gene. Overall, the dataset contained 1,379 positive and 452 negative samples. The second dataset (D2) was composed of known human PPIs (64, 246, 252, 320-322) and included 5,460 unique known interactions mediated by a total of 1,203 unique proteins (reference isoforms), 1,082 of which had at least one alternative isoform in addition to the reference isoform. In total, 4,884 unique alternative isoforms were identified, and 42,652 new, unlabeled, triplets (A1, A2, B) were formed, where isoform A1 interacts with its partner B, but it was unknown whether another isoform A2 interacted with the same partner B. For dataset

D2, the number of isoforms for each gene ranged from 1 to 92, with an average of 8.7 isoforms per gene. The number of interactions per gene ranged from 1 to 800, with a similar average of 35.4 interactions per protein. We investigated possible bias for rewiring interactions based on the number of isoforms using the Mann-Kendall trend test (323-325) and found no statistically significant trends in D1 and T2D case study dataset (Fig. 3.7).

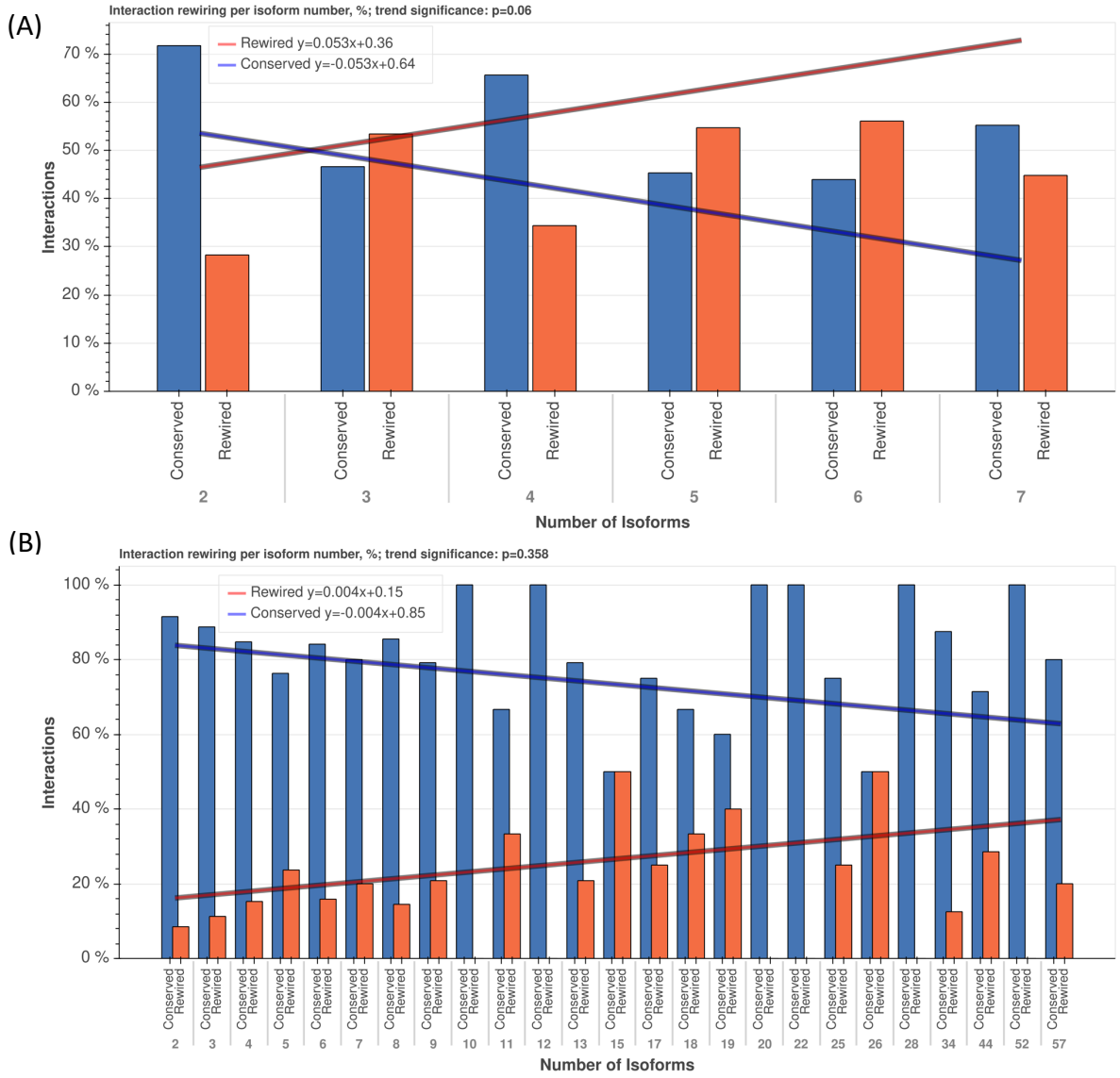

Figure 3.7. Percentage of interaction rewiring depending on isoform number. (A) – statistics on interactions rewiring in experimental dataset D1. (B) – statistics on predicted interactions rewiring for the type 2 diabetes case study. We investigated trends for the rewired and conserved interactions using Mann-Kendall trend test and found no statistically significant trends.

Of the three groups of features generated for each data point, the features corresponding to the occurrence frequency of the SCOP domains were substantially sparse. This phenomenon was the result of some proteins lacking any SCOP domains predicted by SUPERFAMILY. On the other hand, not all SCOP families were equally represented across the proteins from either D1 or D2 datasets. Of the 356 proteins in D1, 260 had between 1 and 8 SCOP domains predicted by SUPERFAMILY, with a mean of 1.4. Similarly, of the 4,028 proteins in D2, 2,917 proteins had between 1 and 25 SCOP domains annotated by SUPERFAMILY with a mean of 2.5.

Another interesting question was whether any of the delta features (the third group of features—see Methods for more details) could be used to provide an accurate separating boundary for the classifier. For instance, if an alternatively spliced isoform altered more than k residues of the reference isoform, then the alternative isoform would be predicted to eliminate the original interaction. There was a wide range of changes for each feature type, with the values seemingly independent of whether or not the alternative isoform disrupted the original interaction (Fig. 11A, B, C). The changes in SCOP domain architecture in the alternative isoform juxtaposed with the reference isoform can be grouped into three categories: no change, deleted domain, or modified domain. For D1, there were 874 (90%) reference isoforms with no change, 99 (10%) isoforms with at least one SCOP domain deleted, and 374 (38 %) with at least one modified SCOP domain. For D2, there were 11,456 (33%) reference isoforms with no change, 23,120 (67%) with at least one domain deleted, or 16,390 (47%) with at least one domain modified.

We next used unsupervised learning followed by low-dimensional embedding with t-SNE (326) to analyze the internal data structure of the two modeled classes, the conserved and perturbed PPIs. By doing so, one would expect to see two clusters of data that are also separated

in a low-dimensional space; the lack of grouping in a low-dimensional space often reflects the absence of intrinsic structure in the dataset or an inability of the method to detect one. However, unlike the case of two well-defined and often well-separated classes, we found that the class of conserved interactions was represented by a union of several sub-clusters corresponding to different types of physical interactions (Fig. 3.11D), which were characterized by a wide range of biochemical, structural (Fig. 3.12), and statistical factors. The perturbed interaction class was defined even more loosely, as a set of conditions sufficient for disrupting each particular interaction type. Specifically, we observed three clusters formed predominantly by the conserved interactions; the rest of the data is grouped into mixed clusters each including both the disrupted and conserved interactions, with no obvious hyperplane separating conserved and perturbed interactions. These results may be explained by the high level of noise in the t-SNE input and data scarcity: because t-SNE attempted to conserve overall distance between samples in data space, the method was unable to capture important dependencies that helped separate the data into classes. Indeed, after weighting the input data from the original dataset by the Random Forest model feature importance, we observed a higher number of well-defined sub-clusters and increased separability among the sets of conserved and perturbed interactions (Fig. 11E).

3.3.3 Methods

Our approach, ALTernatively spliced INteraction prediction (ALT-IN) Tool, is designed to determine the rewiring, or edgetic, effects of alternative splicing. This can be formulated as the following binary classification problem (Fig. 3.6A): Given a known reference isoform A1 that is involved in a protein-pr0otein interaction A1−B with another protein B, will an alternatively spliced isoform A2 preserve the interaction with B or disrupt it? Triplets (A1, A2, B) where A2 preserves the interaction with B, given A1 and B interact, are labeled as members

of the negative class. Alternatively, triplets (A1, A2, B) where the alternatively spliced isoform A2 disrupts the interaction with B are labeled as members of the positive class. Each of the developed methods presented in this work is a feature-based approach (Fig. 3.6B). Specifically, the features encode information concerning the known interaction A1−B and information about the differences between isoforms A2 and A1 of the same gene that may contribute to the disruption of the interaction.

Data retrieval and processing

For training and evaluation of the supervised machine learning classifiers in this work, we use an experimentally obtained human interactomics dataset (D1) developed for the high-throughput analysis of alternative splicing (AS) effects (64). The second dataset (D2) of unknown interaction-rewiring effects by alternative isoforms is a source of the unlabeled data in the training of the semi-supervised classifier. The unlabeled dataset D2 is constructed using a set of protein-protein interactions (PPIs) retrieved from five high-throughput human interactomes (64, 246, 252, 320, 321) and a list of AS isoforms for all proteins that are involved in the above PPIs (327).

Experimental data availability is among the key challenges for the development of supervised learning methods studying alternative splicing effects on PPIs. To the best of our knowledge, there has only been one large-scale experimental study (64). A naïve approach to increasing the experimental dataset would be to merge all existing interactomics datasets and describe different protein products associated with the same gene as isoforms (318). However, this approach is inherently biased toward existing PPIs and cannot differentiate between the absence of interaction and missing interaction information. On the contrary, high-throughput interactomics experiments performed at the AS isoform level (64) account for both the presence

and absence of a PPI mediated by the isoforms of the same gene. Data limitation also influences the selection of machine learning algorithms because some supervised models (e.g., neural networks) require a substantial amount of training data (328). Some approaches try to minimize this effect by leveraging data augmentation, i.e., by applying transformations to the training set (329, 330). However, applying biologically meaningful transformations is only possible when one understands the resulting effects. Protein-protein interaction is a complex molecular mechanism that can be altered by even a single residue change in one of the interacting proteins (257, 331) and is not suitable for this type of data augmentation. Another challenge that the small dataset size brings is the model evaluation. Data scarcity makes the traditional split into training and testing subsets impractical because the training set may not be representative, thus such evaluations may not provide enough information on the model's generalization power. Another traditional approach, 10-fold cross-validation, is prone to fail in isolating highly correlated information in the training dataset. In addition, one has to ensure that both, our model and the data transformations applied to the test set, are completely oblivious of the incoming data. Therefore, we propose a validation protocol that prevents the information leakage from the testing set to the trained model at every step of the training process (for more details, see subsection Quantification and Statistical Analysis and Fig. 3.8).

To obtain dataset D1, we remove reference isoforms and the corresponding alternative splicing variants that satisfy one of the following two criteria: 1) There is no interaction between any isoform of a particular gene and a specific protein; or 2) There is no alternatively spliced isoform for a particular reference isoform, i.e. the corresponding gene has one isoform in total. Then the remaining dataset is organized as a set of triplets (A1, A2, B), where A1 is a reference isoform, A2 is an alternatively spliced isoform, and B is the interaction partner of A1. After

applying this procedure to 2,501 interactions from (332), we obtain the final dataset, D1, which consists of 973 triplets. This dataset is randomly split into 10 subsets for a 10-fold cross-validation protocol. The folds are then modified to ensure that all isoforms related to a gene are either in the testing or training split, as described in the group cross-validation protocol below.

To obtain dataset D2, we first remove RNA-protein interactions as well as interactions from the homo-oligomeric complexes in the original set of human macromolecular interactions (64, 246, 252, 320, 321), leaving only PPIs between two different proteins. Then, to compile a list of AS isoforms for all proteins that are involved in the above PPIs, we download the protein, gene, and isoform mappings from Ensembl (GRCh38 version 91) (327). Lastly, all protein-coding isoforms associated with each reference protein that participates in the PPIs are selected as the final set of AS isoforms.

Statistical Potentials

Large-scale characterization of protein–protein interactions (PPIs) using high-throughput interactomics approaches, such as yeast-two-hybrid and tandem-affinity purification/mass spectrometry methods, have provided the scientists with the new insights of the cell functioning at the systems level and allowed to better understand the molecular machinery underlying complex genetic disorders. Structural studies of PPIs have revealed that a PPI is often carried out by smaller structural protein subunits, the protein domains. Roughly two-thirds of eukaryotic and more than one-third of prokaryotic proteins are estimated to be multi-domain proteins, and thus it is not surprising that ≈ 46% of structurally resolved interactions are domain–domain interactions. A high-throughput breakdown of the interactome at this, domain-level, resolution is a much more experimentally challenging task, currently unfeasible at the whole-system level and requiring computational methods to step in.

Here, we present a simple knowledge-based domain interaction statistical potential (DISPOT), a tool that leverages the statistical information on interactions shared between the homologous domains from structurally defined domain families. The knowledge-based potentials are extracted from our comprehensive database of structurally resolved macromolecular interactions, DOMMINO. Our statistical potential can be integrated into PPI prediction methods that deal with multi-domain proteins by ranking all possible pairwise combinations of domain interactions between two or more proteins. We want to stress that although DISPOT potentials provide some insight into PPI, it is not a classification method, and data provided by it should be used in conjunction with additional information, e.g. a specific pathway.

The development of DISPOT is driven by several observations. First, an average interaction between a pair of proteins is not carried out by all domains constituting each protein, but only by a select subset. Indeed, each domain has its unique structure and biological function and may not be designed to interact with a particular domain from another protein. Second, the domain–domain interactions often share homology: when two homologous domains interact with their partners, these partners frequently also share the homology with each other. Thus, one can introduce the domain–domain interaction propensity in terms of the frequency of domain–domain interactions between the two domain families. Lastly, the propensity of domains to interact is expected to vary across different families, thus allowing to provide the finer resolution of the PPI network.

The quantification of the odds for a domain from one domain family to interact with a domain from another family is defined in this work as a knowledge-based statistical potential. Statistical potentials are widely used in biophysical applications, often for characterizing the residue contacts between the protein chains. One of the main applications of the residue-level

statistical potentials is in protein docking. Our domain–domain statistical potential complements the residue-level potentials by considering structural units from the higher-level of protein structure hierarchy and requiring no structural information about the protein domains. Specifically, the input for DISPOT includes the protein sequences of the two proteins interacting with each other.

First, the domain architecture of each protein is obtained. To do so, a region of the protein sequence is annotated to a family of homologous domains. For the definition of domain families, we leverage the structural classification of proteins (SCOP) family-level classification. SCOP represents a structure-based hierarchical classification of relationships between protein domains or single-domain proteins with 'family' being the first level of SCOP classification and 'superfamily' being the second level. Protein domains from the same SCOP family are evolutionary closely related and often share the same function. Since a protein with no structural information cannot be directly annotated by SCOP, we use SUPERFAMILY, a Hidden Markov Model (HMM)-based approach that maps regions of a protein sequence to one or several SCOP families or superfamilies. SUPERFAMILY allows us to cover a substantial subset of known proteins: the HMM coverage at the protein sequence and overall amino acid levels for the UniProt database were reported at 64.73% and 58.78%, respectively, in 2014.

Second, for each pair of SCOP families we count a number of non-redundant PPIs between the members of these families that have been experimentally determined. Our source of data is DOMMINO a comprehensive database of structurally resolved macromolecular interactions. It contains information about interactions between the protein domains, interdomain linkers, terminal sequences, and protein peptides. In this work, we use exclusively domain–domain interactions because the data about this type of interactions is the most abundant. To

remove redundancy in the data, we use ASTRAL compendium, which is integrated into the SCOPe database (333). From ASTRAL, we obtain a set of domains, where each domain shares <95% sequence identity to any other domain in the set. This set is then used to determine pairs of redundant domain–domain interactions in the original DOMMINO dataset. Two domain–domain interactions are determined as redundant if both corresponding pairs of domains share 95% or more sequence identity. For each pair of redundant domain–domain interactions, one interaction is randomly removed. The process continues until no pair of redundant interactions can be detected.

Third, for each domain family from each protein, a statistical potential is calculated (Fig. 3.8). There are two types of statistical potentials introduced in this work: (i) calculated for a domain from a specific domain family and (ii) calculated for a pair of domains, one domain from each of the two interacting proteins. The statistical potential P_i for a single domain D_i is calculated based on the total number of interactions NDiNDi extracted from the non-redundant DOMMINO dataset for the specific SCOP family this domain belongs to. The statistical potential P_{ij} for a pair of domains, D_i and D_j, is calculated based on the total number of occurrences N_{ij} of the interactions between all domains from the same two SCOP families as D_i and D_j. Those numbers are then transformed into probabilities as follows:

$$P_i = \frac{1}{Z_1} \ln \frac{N_{D_i}}{N_{mean}} \qquad Z_1 = \sum_j \ln \frac{N_{D_i}}{N_{mean}}$$

$$P_{ij} = \frac{1}{Z_2} \ln \frac{N_{D_{ij}}}{M_{mean}} \qquad Z_2 = \sum_{k,l} \ln \frac{N_{D_{kl}}}{M_{mean}},$$

where N_{mean} is the average number of interactions for one domain and M_{mean} is the average number of interactions for a pair of domains present in the database.

DISPOT potentials are derived following a standard strategy for calculating a statistical potential. The statistical potentials for the atomic contact pairs are traditionally derived based on Boltzmann relation:

$$P_{ij} = -k_B T \ln \frac{p_{ij}(r)}{p_{ij}^*}$$

where k is the Boltzmann constant, T is the system's temperature, p_{ij} is an experimentally observed density of atom pairs from different partners in a complex at distance and p_{ij}^* is corresponding density in the reference state. Since we do not work with the atomic-level physical interactions, we replace the Boltzmann constant from DISPOT equations and substitute temperature with the inverse of normalization constant Z. In addition, p_{ij} and p_{ij}^* are substituted with the number of interactions between domains in DOMMINO database.

DISPOT can also provide integrated protein-level statistics. There are multiple ways to combine the domain-level statistics into a protein-level statistics. Two simple approaches to integrate domain–domain interactions for a given PPI in terms of a standalone (single protein) and interaction (protein pair) potentials are:

$$P_{M_u} = \max_i P_i \text{ and } P_{M_{uv}} = \max_{i,j} P_{ij}$$

respectively, where i and j correspond to the domains from protein u and v. The rationale behind these definitions lies in the assumption that a single strongest domain–domain interaction is the one of the most important defining factorsfor the PPI. These definitions of cumulative potentials were tested in terms of their ability to predict a PPI using several experimental sources. First, we obtained the coverage landscape by the cumulative potentials on the experimental protein–protein interactomes one obtained using high-throughput yeast-two-hybrid screening (HI-I-05)

and another one obtained using curated literature-based search (LitBM-17, http://interactome.baderlab.org/data/LitBM-17.psi). As expected, while this naïve method was able to recover 2944 PPIs in HI-I-05, it missed 1188 PPIs even using a lenient threshold of −20 (Fig. 3.8). Similarly, the cumulative potential was able to recover only 1718 PPIs while 1453 PPIs were not recovered. We then apply the same pairwise cumulative potential to the large-scale mass spectrometry study. Specifically, we study the correlation between the hu.MAP probability score and cumulative pairwise score among KEGG (334) pathways and GO clusters produced by GeneSCF (335) on 13 855 genes with SUPERFAMILY annotation. While the number of highly correlated pairs was substantial, the number of pairs with very little correlation still prevailed. Finally, the analysis of the cumulative single potential for a protein showed that it can obtain a diverse range of values and this property seems to be independent of how many domains this protein has. Similar behavior was observed when looking at the other basic cumulative measures.

Overall, we have analyzed and summarized interactions from 3619 SCOP family pairs that were extracted from 352 199 PPIs. In total, domains from 1384 SCOP families were characterized that form domain–domain interactions in 1384 ‘homo-SCOP’ interaction pairs (i.e., both domains are annotated with the same SCOP family) and 2235 ‘hetero-SCOP’ pairs. The analysis of the calculated statistical potentials showed a wide diversity across different families.

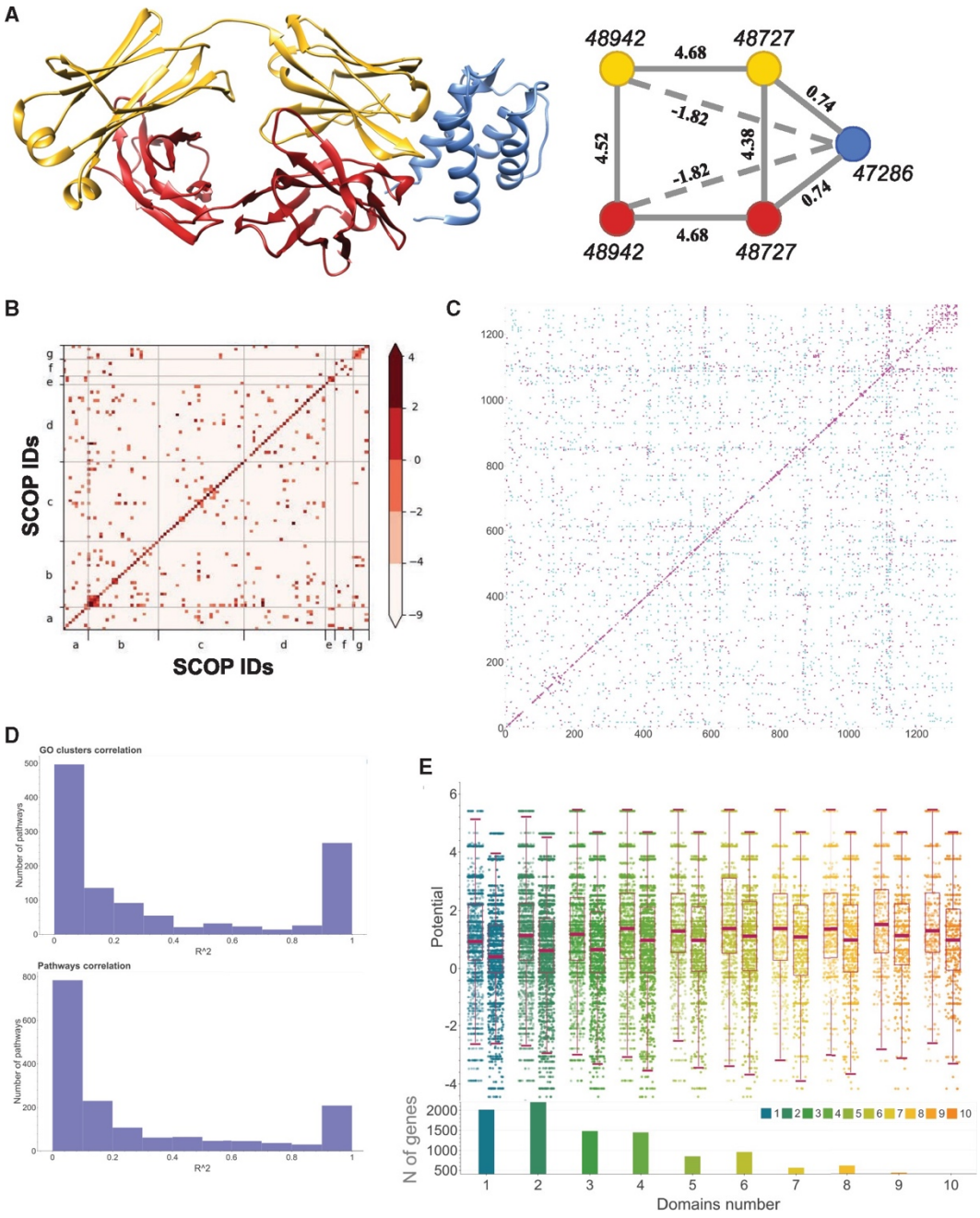

Figure 3.8. DISPOT statistical potential and its application. (**A**) A crystal structure (left) of the protein complex between CNTO607 Fab human monoclonal antibody (yellow and red colors denote two different chains) and interleukin-13 (IL-13, shown in blue), and the corresponding domain–domain interaction network (right). Shown in italics are SCOP family IDs, and in bold are DISPOT values for the corresponding interactions. Nodes colored with the same color belong to the same chain. Solid lines connecting nodes correspond to the physical interactions, while dashed lines connect nodes corresponding to the protein domains that do not physically interact. (**B**) A heatmap showing DISPOT values calculated for each pair of SCOP families, where only potentials for pairs of SCOP families with five and more non-redundant interactions are plotted. The families are grouped based on the SCOP class (a–g) and are ordered within each fold based on their IDs. (**C**) A contact map showing the correlation between

experimentally obtained human interactome HI-I-05 and DISPOT-based PPI prediction. A prediction that calls a PPI correctly is shown in magenta, while PPIs that were missed are shown in cyan. (**D**) Correlation calculated using R^2 correlation coefficient between the hu.MAP interaction probability score and DISPOT statistical potential for KEGG pathways (bottom) and GO clusters (top). (**E**) Distribution of the protein-level DISPOT statistical potentials grouped by the number of SCOP domains in a protein defined using SUPERFAMILY.

Finally, we would like to make a cautionary note of using the developed tool. DISPOT was designed not as a PPI prediction tool, but rather a tool that provides additional information on the likelihood of specific domain–domain interactions in a given physical PPI. The main reason is the fact that structural coverage of the PPI space is still far from being full, which leads to the presence of a high number of false negatives if one was to use DISPOT as a standalone predictor. This intuition has been supported by our evaluation of DISPOT against the two interactomics golden standards. Thus, if a researcher wants to employ DISPOT in a PPI prediction method, we recommend adding the DISPOT potentials as features to the overall feature vector, that will include other parameters, such as secondary structure, evolutionary conservation of the sequence, predicted residue hydrophobicity, etc.

Feature engineering

The question we are answering in this work, if the alternatively spliced isoform A2 will retain an interaction originally established between the reference isoform A1 and its interaction partner, is somewhat similar to a PPI prediction task. However, here we can leverage additional information on alternative splicing and knowledge regarding the original interaction. This naturally imposes structure on the features we generate. So far, we are using three groups of features: (1) biochemical features of the reference isoform and its interaction partner, (2) domain interaction knowledge-based statistical potentials, and (3) so-called “delta” features. The first group of features is inspired by PPI prediction methods (319, 336). The second group represents

set of features derived from our DOMMINO database of macromolecular interactions (337) The rationale behind using this group of features is the following: given that an average protein includes multiple protein domains (242), it is important to know which domains are directly involved in a particular PPI. The third group, the "delta" features, includes selected characteristics of alternative splicing events. Specifically, the features are designed to capture those differences between the reference isoform and its alternatively spliced variant that may result in a loss of interaction.

The biochemical features provide a general outline of the different properties of the known interaction. These features include molecular weight, number of residues, average residue weight, charge, isoelectric point, A280 molecular extinction coefficient for both reduced and cysteine bridges, and several other characteristics (Table A1, Appendix).

There are four subgroups of the delta features. The first subgroup includes features that describe the difference of the biochemical characteristics between the reference isoform A1 and alternatively spliced isoform A2. The second subgroup includes the difference between the statistical potentials of A1 and A2. The third subgroup is a set of simple sequence features that can be computed with a basic sequence alignment, but nevertheless may provide important knowledge. For instance, an exon skipping event that results in a large portion of protein missing is usually more detrimental to the protein-protein interaction than several exon skipping events, each missing only a small portion of the protein. Similarly, the modifications in N- or C-terminus are less likely to result in interaction perturbation compared to an equally sized modification occurring in one of the protein's domains. The last subgroup of features is concerned with the SCOP family domain information defined by the SUPERFAMILY tool (338) to determine if the alternative splicing affects specific protein domains.

Learning Under Privileged Information (LUPI)

Learning under privileged information (LUPI) is a machine learning paradigm that accounts for valuable features which are either impossible or too expensive to obtain when using the trained model for predictions (privileged information). The LUPI paradigm accounts for this privileged information by adjusting the decision boundaries inferred by the algorithm over the regular features (339-341). Although LUPI is a generic machine learning paradigm and is not specific to selected classifiers, current general-purpose implementations are limited to the SVM variations (339, 340, 342, 343). The only available alternative - neural networks implementations – are primarily focused on computer vision tasks (344-346). In the current work, we are using SVM+ LUPI classifier and AdaBoost built on top of this classifier. Our list of privileged features consists of the FoldX interaction energy and supplementary terms (e.g., electrostatics) (347, 348), OPUS-PSP score (349), GOAP potential (350), NACCESS2 accessible surface area (351), Geometric score (352) and Dfire2 score (353), as well as statistics on the binding sites (354) (Table S4, Supplementary Data). Many of these features are based on the structural information of the proteins and therefore were calculated only for the labeled training set.

Machine learning approaches

Six machine learning classifiers were trained, and their performance compared, including four supervised learning methods: support vector machines (SVM) with two kernels, random forest, and AdaBoost, as well as a LUPI approach using SVM+ and a semi-supervised learning method using an iterative self-learning random forest approach.

Support vector machines (SVM) belong to a family of widely used kernel methods (19). It is also among the most well-established and popular machine learning approaches in bioinformatics (20, 21). In our experiments, two kernel functions were explored: linear kernel

and radial basis function (RBF) implemented in libsvm library (355) For the SVM models, the parameter optimization was performed using grid search. Optimal values gamma = 0.005 and C = 9 were obtained after the search in the range from gamma = 0.001 to gamma = 1 with a step 0.002, and from C = 1 to C = 100 with a step 1.

Random forest (91) is an ensemble classifier, which combines multiple supervised learning classifiers to get a prediction. It uses the ideas of bagging and random split decisions to predict a class of untrained vectors. In bagging, a random selection of the examples in the training set is used to build each decision. Because of the data heterogeneity in our problem and the necessity of addressing missing values for the domain-based information the RF classifier is a good fit. In this work, the random forest models were trained using the scikit-learn 0.19.1 package (356). Parameters obtained by nested cross-validation include Gini criterion, a minimum number of samples required to split node = 2, and an unbound maximum depth.

AdaBoost (Adaptive Boosting) is an ensemble classifier that produces accurate prediction rules via combining several weak learners into a weighted majority hypothesis, adaptively changing weights based on the accuracies of individual components (357, 358). This model is widely used in bioinformatics applications (359-364). The algorithm has an iterative nature, updating weight vector $W = \langle w_1, \dots, w_N \rangle$ that corresponds to the N labeled examples $(x_1, y_1), \dots, (x_N, y_N)$ for each iteration t based on weak learner's weighted error $\varepsilon_t = \sum_{i=1}^{N} p_i^t |h_t(x_i) - y_i|$. Here, $h_t(x_i)$ is a weak learner's hypothesis, given distribution $p^t = \frac{w^t}{\sum_{i=1}^{N} w_i^t}$. In modern implementations weights are updated according to the formula $w_i^{t+1} = w_i^t e^{-\alpha_t |h_t(x_i) - y_i|}$, where $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ (365).

SVM+ is one of the standard LUPI algorithms based on the SVM classifier (366). The key advantage achieved by this algorithm is the much faster convergence of the learning process, as a function of the number of training examples, n. LUPI with its fast training convergence is well-suited for our problem, since the number of labeled isoform pairs is limited for a traditional supervised learning approach. Formally, the learning problem using privileged information is defined as follows: given a set of training triplets $(x_1,x_1^*,y_1),...,(x_n,x_n^*,y_n)$, where xi is a normal feature vector and xi* is a vector with privileged information, map each vector $x_i \in X$ to a vector of another space, $z_i \in Z$, each privileged vector $x_i^* \in X^*$ into another space, $z_i^* \in Z^*$, and find in Z a separating hyperplane that minimizes the cost:

$$R(w,b,w^*,b^*) = (w,w) + \gamma(w^*,w^*) + C\sum_{i=1}^{n}[(w^*,z_i^*) + b^*], \text{ subject to: } y_i[(w,z_i)+b] \geq 1 - [(w^*,z_i^*)+b^*], \quad i = 1,...,n$$

The minimization problem then is solved using one of the current efficient SMO optimizers(367).

We also explored a straightforward modification of AdaBoost with SVM+ as a base learner, using the same training algorithms as in the study (365).

Finally, another machine learning concept was explored, semi-supervised learning (257, 316, 368). One of the main bottlenecks of supervised learning is the cost of labeling data. The idea behind semi-supervised learning is to utilize a large amount of unlabeled data to improve the results of the corresponding supervised learning algorithm. There are a number of existing approaches to combining labeled and unlabeled information that try to exploit the underlying structure of the unlabeled data. In most cases, the learning algorithm attempts to find clusters to modify the decision boundaries. Here, we implement a simple semi-supervised learning

approach, iterative self-learning random forest, that has previously shown to outperform more advanced semi-supervised learning methods on protein-protein interaction data represented by heterogeneous features (257). The algorithm starts with a labeled training dataset and a pool of unlabeled feature vectors (Fig. A1, Appendix). At each step, the algorithm initially trains a supervised learning classifier on the labeled training set. Then, it evaluates the model using a grouped 10-fold cross-validation over the training set. Next, the algorithm applies the obtained classifier to the remaining unlabeled dataset, predicting their labels, selecting several examples, and adding them back to the training set, and retraining the supervised classifier. Selection of the newly labeled examples is based on the confidence score provided by the random forest algorithm. After multiple iterations, the model with the best evaluation score is selected when the current iteration's F1-score decreases by more than 0.03; at this point, the model for the step before the subsequent decline in performance is returned.

Feature selection protocols

To improve the performance of the classifiers, three feature selection methods are explored including LASSO, recursive feature elimination (RFE), and principal component analysis (PCA) (369). We also analyze the importance of individual features. To calculate the feature importance, we use the mean decrease of impurity in the random forest model, our top-performing supervised classifier. This is a tree-specific metric and is directly related to the Gini impurity, calculated at each tree node (91). The same feature is present in multiple trees in a random forest model, thus the average decrease in impurity integrates the feedback from all trees that contain this feature.

LASSO is a regression model with l1 regularization. Because of the l1 penalty, a solution for the regression naturally contains zero coefficients for many features, thus discarding them

from the model. RFE is another widely used feature selection algorithm that consecutively removes one feature from the model and evaluates the results using cross-validation. The optimal number of features in RFE is also determined by cross-validation. The last method is a feature generation method, PCA. It is a technique that performs orthogonal transformation of the feature set to obtain linearly uncorrelated components.

The number of selected principal components is determined by the 98% explained variance cutoff threshold. Feature selection methods produce varying results for SVM and fail to improve the performance of the random forest classifier, which, in turn, shows the most accurate performance among all supervised methods in our study. This result is expected, since the total number of features is significantly smaller than the number of samples, so the random forest model does not overfit, and the influence of less informative features' is limited due to the random subspace selection.

Quantification and Statistical Analysis

Data scarcity makes efficient training of machine learning models and their subsequent evaluation difficult. Obtaining a fully independent evaluation dataset is equally critical because failure to isolate training and testing data may result in overly optimistic scores, even if the obtained model does not generalize well (164, 370-373). Therefore, we propose two ways of assessing our approach. The first is a validation protocol that takes into account the nature of the AS data, while preventing information leakage from the testing set into the model through the entire training process. The protocol includes four steps: (i) split on test/train for cross-validation (CV) iteration; (ii) feature selection; (iii) hyperparameter tuning; and (iv) evaluation of each CV iteration. Our evaluation protocol is based on the concurrent usage of leave-group-out cross-validation (LGOCV) (374) and nested CV (375) protocols (Fig. 3.9). For the cross-validation

protocol, we employ one of the most widely used 10-fold cross-validation. Secondly, we compare the performance of our methods with the state-of-the-art ab initio PPI prediction tools, including TRI_Tool (M1) (336), LR_PPI (319) with negative set 1 (M2), and LR_PPI with negative set 2 (M3). One can apply each of the ab initio tools to predict if a PPI between A2 and B exists, independent of knowing whether or not A1 and B interact, while similar to our case, the existing AS-based method requires a PPI to exist between A1 and B. Lastly, we compare our methods with the only two methods that leverage AS information, albeit for slightly different problems (317, 318).

Nested cross-validation protocol (375) is used to avoid overfitting during hyperparameter tuning (Figs. 3.9, 3.10). It is a crucial step for the learning methods whose performance is heavily dependent on their hyperparameters. Each iteration of the nested cross-validation protocol includes two loops, the outer loop, and the inner loop. For each outer loop, the dataset is divided into the training and testing sets, as in regular cross-validation, and this loop is used to generate performance assessment measures for a classification algorithm. The inner loop is executed within each iteration of the outer loop. Specifically, the training dataset from the outer loop is further split into the initial training and parameter tuning subsets followed by another cross-validation step to find the optimal hyperparameters. This procedure allows for the evaluation step to be explicitly detached from the hyperparameter tuning step.

Regular cross-validation performs well if each data point is completely independent of others. Unfortunately, this may not always be the case for our dataset, as multiple isoforms are the products of the same gene. If one subset of related isoforms is present in the training set and another subset is present in the testing set, then the model is provided with an unfair advantage during evaluation. We expect our model to generalize well and handle novel isoforms with no

prior information about them. Therefore, the original 10-fold cross-validation is modified into a leave-group-out cross-validation (LGOCV) (374). Specifically, we group all isoforms that are products of the same gene, and each group is then allocated exclusively either to the training set or to the testing set. This LGOCV protocol is more stringent than the regular CV protocol, thus it is expected to reduce the reported accuracy of the method.

The performances of the supervised and semi-supervised learning methods are assessed using three evaluation protocols: (1) the nested leave-group-out cross-validation; (2) comparison with the state-of-the-art ab-initio PPI prediction methods; and (3) comparison with the domain-based prediction (317). For each method, standard assessment criteria are computed, including accuracy (Acc), recall (also called sensitivity, Rec), precision (Pre), f-measure (F1-score), Matthews correlation coefficient (MCC), and area under the curve (AUC). The area under the curve can be computed with the help of the Gini coefficient (G_1):

$$AUC = \frac{1+G_1}{2} G_1 = 1 - \sum (X_k - X_{k-1}) (Y_k + Y_{k-1}),$$

where $X_i$is a true positive rate (TPR), and Y_i is a false positive rate (FPR) for the threshold i. A pair (X_i,Y_i) defines a point on the receiver operating characteristic (ROC) curve. The statistical significance of the results for the top-performing classifier is calculated using Welch's non-parametric t-test with the significance level at 0.1, which is appropriate because it does not assume equal variance of the samples (376) over metrics obtained on each step of two 10-fold cross-validation.

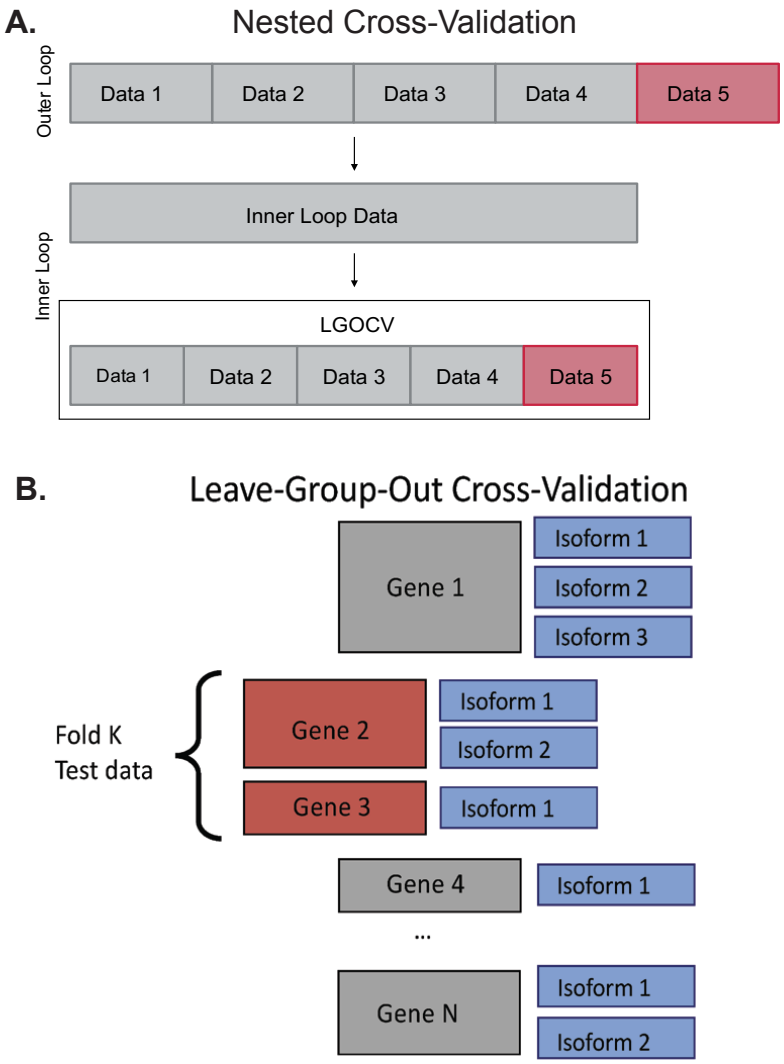

Figure 3.9. Nested 5-fold cross-validation and Leave group out cross-validation (LGOCV) examples. A. Outer loop is a 5-fold cross-validation which splits data into 5 batches. On each step one batch is withheld for testing classifier performance. The other 4 batches are used for another 5-fold cross-validation. This inner cross-validation is used for hyperparameter tuning. LGOCV is a leave-group-out cross validation step. **B.** This type of cross-validation is used when samples are not independent of each other. All samples that share a common origin (produced by the same gene) are put either in train set or in test set. In this way the model generalize substantially better due to no information leakage.

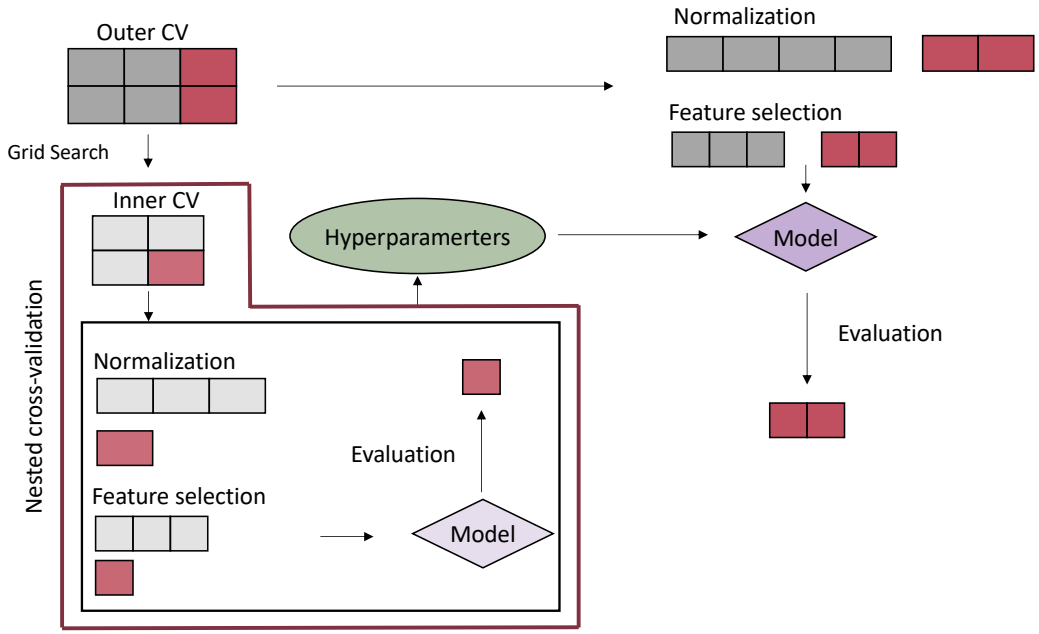

Figure 3.10. Validation protocol. Nested cross validation (see Fig. 3.9) is used for the hyperparameter tuning. For each iteration, inner cross-validation is conducted exclusively on the training data from the outer loop, thus the hyperparameters and normalization procedure derived from this stage are completely agnostic of the outer loop's testing data. This approach prevents information leakage between training and testing phase at each step of cross-validation. In order to account for the information leakage stemming from the sequence similarity among the isoforms, both inner and outer cross-validations are LGOCV (Fig. 3.9) considering isoforms derived from the same gene as a single group.

3.3.4 Results

Method evaluation

First, using D1, we evaluated the prediction accuracy of the three supervised machine learning classifiers: SVM with linear and radial basis function kernels and random forest (Fig. 12F, Tables 1, 2). The results of 10-fold nested cross-validation showed that random forest clearly outperformed the two regular SVM models and SVM+ model that made use of the privileged information, reaching an accuracy of 0.85, F1-score of 0.90, MCC of 0.62, and AUC

of 0.80. Next, to evaluate the importance of protein domain feature information, we assessed the same methods, but with two different feature vector definitions, one that included the protein domain features and another that excluded them. Without protein domains, the performance slightly dropped, with the accuracy values ranging from 0.82 to 0.84, precision from 0.84 to 0.87, recall from 0.90 to 0.93, F1-score from 0.87 to 0.88, MCC from 0.48 to 0.57, and AUC from 0.71 to 0.77. Similarly, to evaluate the importance of using the delta feature information, we assessed the same supervised classifiers with or without these features. For vector representation lacking delta features the performance dropped substantially, with the accuracy values ranging between 0.72 and 0.73, precision ranging between 0.72 and 0.74, and with MCC dropping the most, ranging between 0 and 0.08. Recall was the only metric that improved, ranging from 0.96 to 1.0.

Our second machine learning approach, a semi-supervised learning classifier, incorporated a large number of unknown label data to train the model. As a result, during the cross-validation, the semi-supervised classifier provided the most accurate performance of all other methods. The assessment values included accuracy of 0.87 (improvement of 0.01 over the top supervised learning classifier, $p = 0.65$), precision of 0.93 (improvement of 0.03, $p = 0.09$), recall of 0.89 (lower than the boosting models by 0.09, $p = 0.04$), F1-score of 0.91 (improvement of 0.01, $p = 0.94$), MCC of 0.68 (improvement of 0.06, $p = 0.08$), and AUC of 0.84 (improvement of 0.04, $p = 0.07$).

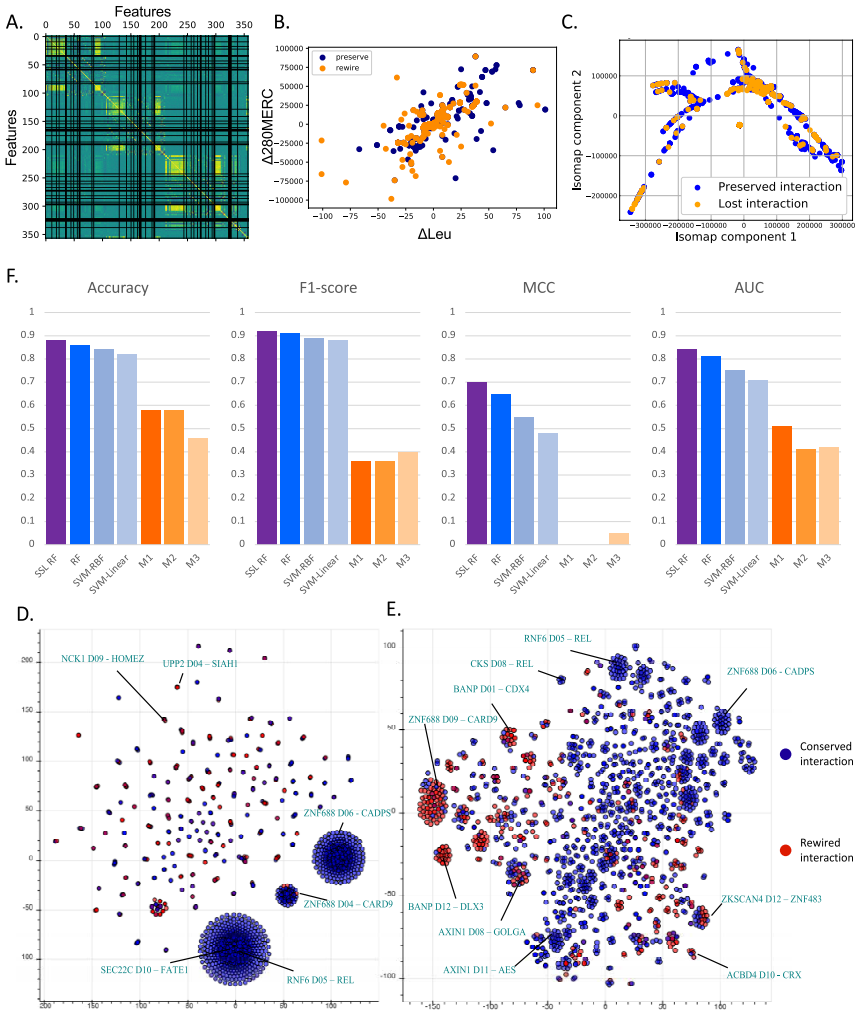

Figure 3.11. Feature analysis and comparison of our machine learning models with general PPI prediction methods across four different metrics (accuracy, F1-score, MCC and AUC). A. A correlation plot between features used for training machine learning models showing three distinct blocks which are associated with biochemical features of reference isoform, biochemical features of interacting protein and delta biochemical features. None of the blocks show high correlations with other blocks. **B**. A scatterplot based on delta frequency of leucine and another delta of 280MERC coefficient is a typical example of how the feature values are distributed between the representatives of two classes, suggesting that the pairwise comparisons cannot separate two classes well. **C.** Isomap visualization of all features through a low-dimensional embedding. In spite of using manifold learning, we are unable to obtain separable classes in 2D space. **D.** t-SNE plot of the entire dataset indicating three large clusters corresponding mainly to the conserved interactions (blue) and small fine-grained clusters of both conserved and perturbed (red) interactions that are linearly separable. **E.** t-SNE plot of the same dataset with columns weighted by feature importance obtained from the random forest model, indicating clearer separation of perturbed and conserved interactions clusters with inner area dominated by the conserved interactions. **F.** Performance of our supervised (blue) and semi-supervised (purple) methods representing mean value of cross-validation runs compared against three current *ab-initio* PPI prediction methods (orange) across four metrics.

The question posed in this work could also be addressed by 1) assuming that the alternative isoform is a new protein, and 2) predicting whether the isoform interacts with the corresponding interaction partner using an existing ab initio PPI prediction method, i.e., without prior knowledge about the interaction between the reference isoform and its interaction partner. Our evaluation of the three state-of-the-art ab initio PPI prediction methods showed that neither of the methods could be reliably used for our task: the values for accuracy ranged between 0.46 and 0.58, recall between 0.29 and 0.5, precision between 0.5 and 0.52, F1-score between 0.36 and 0.4, and MCC between 0 and 0.05 (Fig. 12F).

Algorithm	Feature Selection	Accuracy	Precision	Recall	F1-score	MCC	AUC
Semi-Supervised RF	RFE	**0.87**	**0.93**	0.89	**0.91**	**0.68**	**0.84**
Random Forest	RFE	0.85	0.90	0.89	0.90	0.62	0.80
SVM-RBF	RFE	0.84	0.87	0.93	0.89	0.55	0.75
SVM-Linear	Lasso	0.82	0.86	0.92	0.88	0.44	0.70
LUPI SVM+	None	0.86	0.88	0.93	0.90	0.53	0.78
LUPI SVM+ Boosting	RFE	0.85	0.83	**0.98**	0.90	0.53	0.70
AdaBoost	RFE	0.86	0.87	0.95	**0.91**	0.61	0.78
PPI Prediction Methods							
M1		0.58	0.29	0.50	0.36	0	0.51
M2		0.58	0.29	0.50	0.36	0	0.59
M3		0.46	0.50	0.52	0.40	0.05	0.58
Isoform prediction method							
M4		0.65	0.28	0.74	0.40	0.28	0.73

Table 1. Comparison of AS-specific machine learning models and general *ab initio* PPI prediction methods. Our top performing machine learning model is the semi-supervised random forest. It has the best scores for each metric except recall. PPI prediction methods fairs poorly for our problem. As both F1-score and MCC are also low our conclusion is that M1, M2 and M3 in its current states are unfit for our problem. Low AUC also suggests that we cannot raise other metrics much by simply varying the probability cutoff threshold. The highest scores for each measure are shown in bold.

Unsatisfactory performance of generic PPI prediction methods with comparable results was also previously observed (164). Specifically, the work highlighted a common flaw in the pair input schemes—leaking information on specific protein pairs from a test set to the model. Information leakage effects have been a well-known problem in supervised learning (370-373). In PPI studies, the leakage effects happened because the same interacting partners were present in train and test sets, while the interactions were different because the second interaction partner was different.

Algorithm	Semi-Supervised RF						
Feature Selection	RFE						
Algorithm	**Feature Selection**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**MCC**	**AUC**
Random Forest	RFE	0.31	**$6.6*10^{-2}$**	0.53	0.54	**$8.3*10^{-2}$**	**$7*10^{-2}$**
SVM-RBF	RFE	0.51	**$9*10^{-3}$**	**$5.4*10^{-2}$**	0.22	**$5.5*10^{-3}$**	**$1.7*10^{-3}$**
SVM-Linear	Lasso	**$2.1*10^{-2}$**	**$2.2*10^{-4}$**	**$4.0*10^{-2}$**	0.15	**$3.3*10^{-4}$**	**$1.9*10^{-4}$**
LUPI SVM+	None	0.65	**$9.5*10^{-2}$**	0.55	0.12	**$4.9*10^{-2}$**	**$6.8*10^{-2}$**
LUPI SVM+ Boosting	RFE	0.34	**$2.5*10^{-4}$**	**$3.6*10^{-2}$**	0.94	**$9*10^{-2}$**	**$4.8*10^{-3}$**
AdaBoost	RFE	0.25	**$1.6*10^{-2}$**	**$3.3*10^{-2}$**	0.62	**$2.9*10^{-2}$**	**$7.3*10^{-3}$**

Table 2. Statistical significance of differences in performance measures between the top-performing machine learning model (semi-supervised random forest with RFE feature selection) and other alternative splicing-specific models. Welch's test on two 10-fold cross-validation runs is used to determine statistical significance of the results from Table S2. Statistically significant differences are presented in bold. Semi-supervised random forest (SSRF) completely outperforms SVM-Linear with statistically significant gains for all measures but F1-score. In comparison to SVM-RBF, it loses in Recall but gains in precision, MCC, and AUC. Compared with supervised random forest, SSRF has a statistically significant gain in precision, MCC, and AUC. Factoring in those differences, one can conclude that SSRF with RFE model is able to generalize better than other models, though boosting-based models can be used as an alternative in specific applications as they are able to achieve the highest recall.

To eliminate the aforementioned effects, we used leave-group-out cross-validation (LGOCV) for our method assessment (Fig. S3). After the transition from 10-fold CV to LGOCV, the best performance of our initially evaluated models dropped sharply from 0.86 accuracy and

0.88 F1-score to 0.67 accuracy and 0.59 F1-score, respectively. Eventually, enhancing the classifier using feature generation as well as the introduction of knowledge-based statistical potentials and additional alternative splicing-specific features, enabled the improvement of the accuracy and F1-score to 0.87 and 0.91, respectively (Fig. 3.11).

Lastly, when comparing the performance of our semi-supervised classifier with the two previously published AS-based methods, we found that ALT-IN Tool substantially outperformed both of them. In particular, a previously published AS-based PPI prediction method (317) reported TPR of 0.33 and FPR of 0.2 for the data with full domain-domain interaction annotation and even worse TPR of 0.31 and FPR 0.1 for the data with partial annotation. In comparison, our method had TPR of 0.93 and FPR of 0.20 based on the 10-fold nested cross-validation for the SSRF classification algorithm. Reported results on the experimentally validated dataset for the second AS-based method (318) included accuracy of 0.65, AUC of 0.73, and MCC of 0.28 (Table S2 in Suppl. data) and were also significantly lower than the measures for both supervised and semi-supervised versions of ALT-IN Tool.

Case study

To demonstrate the utility of ALT-IN Tool and the extent to which AS variation can perturb a disease-centered PPI network, we used our method to predict the edgetic effects due to the disease-specific AS occurring in the brain and liver tissues of Western Diet (WD) fed mouse that developed type 2 diabetes (T2D). Because of the similarities between pathological processes in humans and mice bear enough similarities for the latter to be used as model organism, we expect to find biologically relevant highlights of the alternative splicing influence on mouse with environmentally induced T2D based on ALT-IN results.

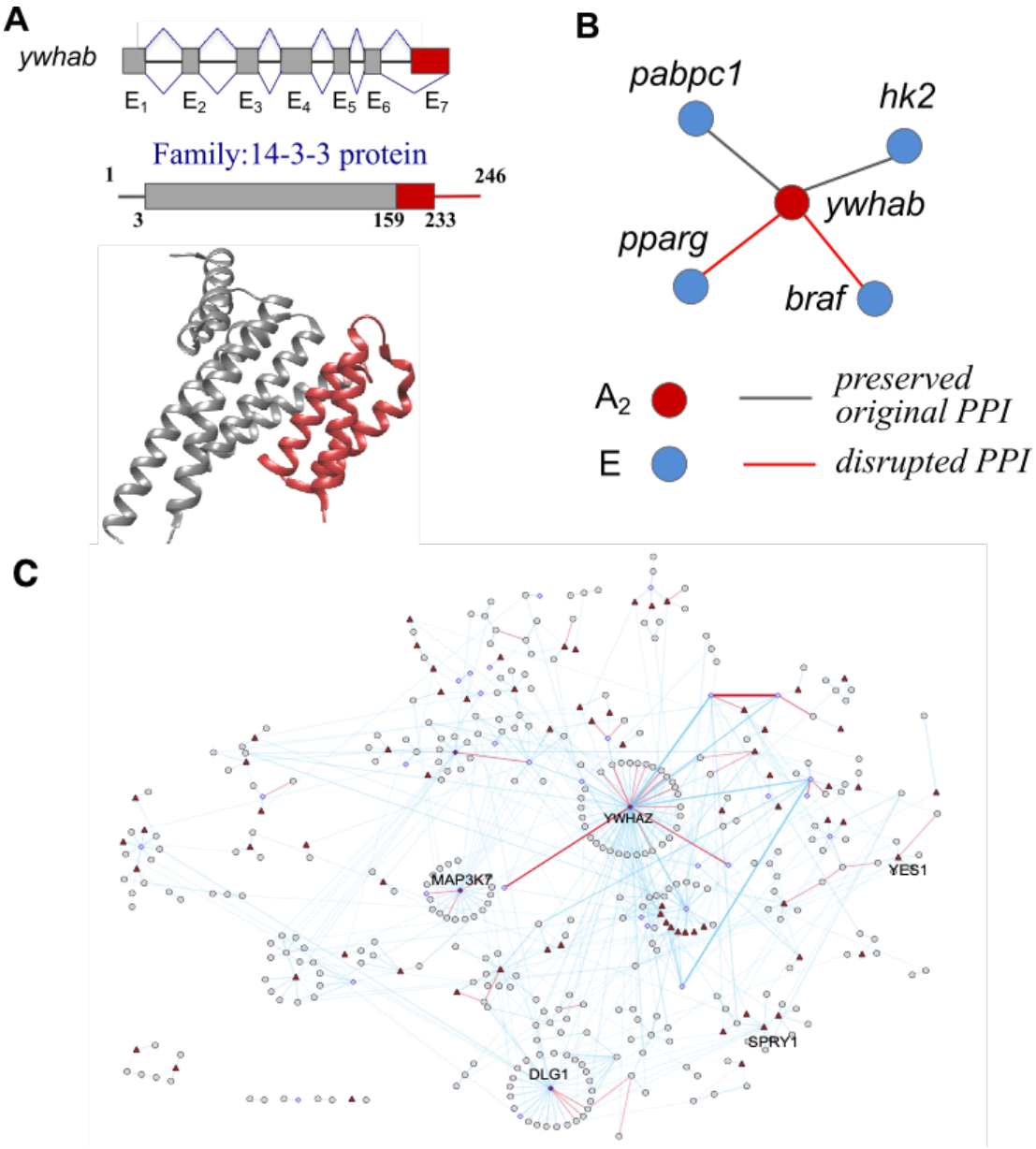

Figure 3.12. A case studies of a gene associated with T2D, which alternatively spliced isoforms were predicted by AS-IN Tool to rewire some of the currently known PPIs and AS-centered protein–protein interaction network perturbation. A. The gene architecture, protein domain architecture, and structure-based characterization of the alternatively spliced isoform of *ywhab* gene. The red part of the protein corresponds to the seventh exon and is spliced out in the alternative isoform A_2. B. As a result, two interactions were predicted to be disrupted by the alternatively spliced isoform A_2 that had been determined to be significantly overexpressed in the tissue samples of WD-fed mouse with T2D disease phenotype. C. Network centered around alternatively spliced isoforms expressed in the liver and brain tissues, which were found drastically different (at least 5 fold of log2 expression values) between the control and T2D mice induced through Western Diet. The effect of the alternative isoforms was predicted as either disrupting the original PPI (red edges) or preserving it (light blue edges). Genes that are associated with T2D are represented as dark red nodes, while their interaction partners are colored gray. Hub nodes (30 interactions or more) associated with T2D are represented as diamond shapes, while the rest of T2D-associated genes are represented as triangles. A few well-studied genes linked to T2D are highlighted: *map3k7*, *yes1*, *spry1*, *dlg1*, and *ywhaz*.

As information on molecular mechanism of action is not directly transferable between species, however similar, we emphasize that the purpose of this case study is solely to investigate biological relevance of the ALT-IN findings and not to provide new insights into the T2D, and ask readers to use method cautiously for cross-species applications. Our deep RNA-sequencing data extracted from the tissue samples of the healthy mouse and Western Diet (WD) fed mouse resulted in 1,899 AS isoforms from 1,608 genes for the brain samples and 5,951 AS isoforms from 3,942 genes for the liver samples that had drastically different expression levels (>5 fold) between diabetic and normal mice samples. In total, 6,745 unique isoforms that were substantially differentially expressed between the normal and diabetic samples were collected for both tissue types.

Only this subset of the differentially expressed isoforms was considered for further analysis. Extracting meaningful information from the direct comparison with the two aforementioned alternative splicing PPI prediction methods turned out to be problematic. The first method based on the domain interactions approach (317) covered only 34% of the proteins which corresponded to >5-fold RNA-Seq change. The second method (318) was not publicly available, however, we conducted a comparison with the LR_PPI (319) predictions which (318) treated as a golden standard for the training purposes.

In the following analysis, we refer to a PPI as "T2D-associated" if at least one interaction partner comes from this subset of 183 genes, otherwise, we call the interaction "normal". After applying our method, ALT-IN Tool predicted 29 interactions as disrupted (4.5% of total interactions), including 23 T2D-associated interactions (79.3% of rewired interactions) (Fig. 3.13). We then cross-tabulated information on the rewiring events among the normal and T2D-associated interactions (Table A3 in Appendix) and applied Fisher's exact test (377, 378). As a

result, we obtained that there was a statistically significant difference between the rewiring frequencies in T2D-associated interactions and normal ones, with $p = 0.012$.

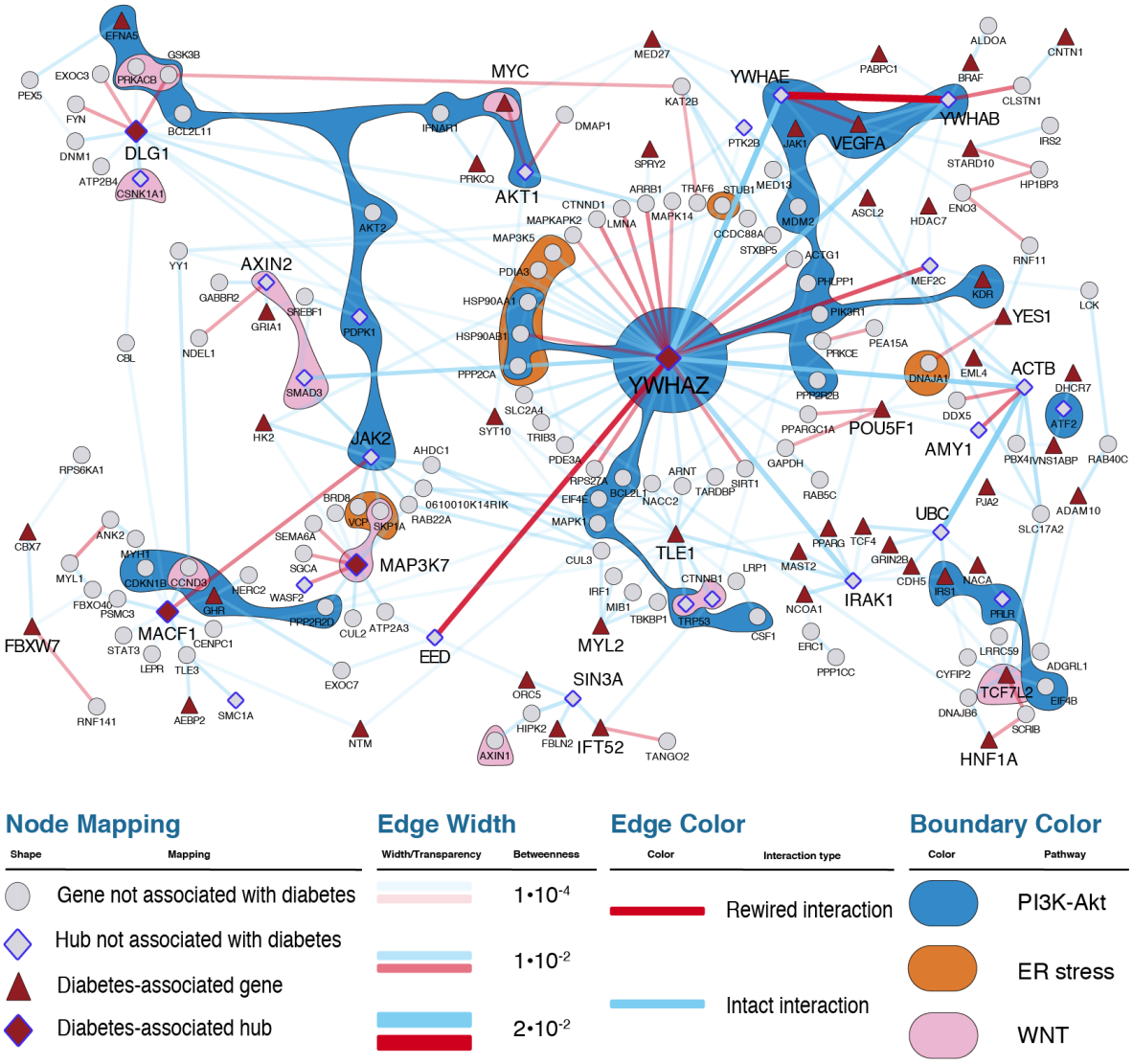

Figure 3.13 Diabetes-centered AS-induced network perturbation. A subnetwork centered around perturbed protein interactions caused by isoforms of genes with five fold of log 2 changes in isoform expression. We highlight three diabetes-related pathways that had significant presence: PI3K-Akt, ER stress, and WNT. Edge width and opacity corresponds to the betweenness centrality, a measure that identifies bottlenecks in graphs, computed for comprehensive PPI network. Hubs were determined as nodes with at least 20 connections in underlying PPI network. It worth noting that there is no clear link between the number of isoforms corresponding to distinct gene and interaction rewiring frequency.

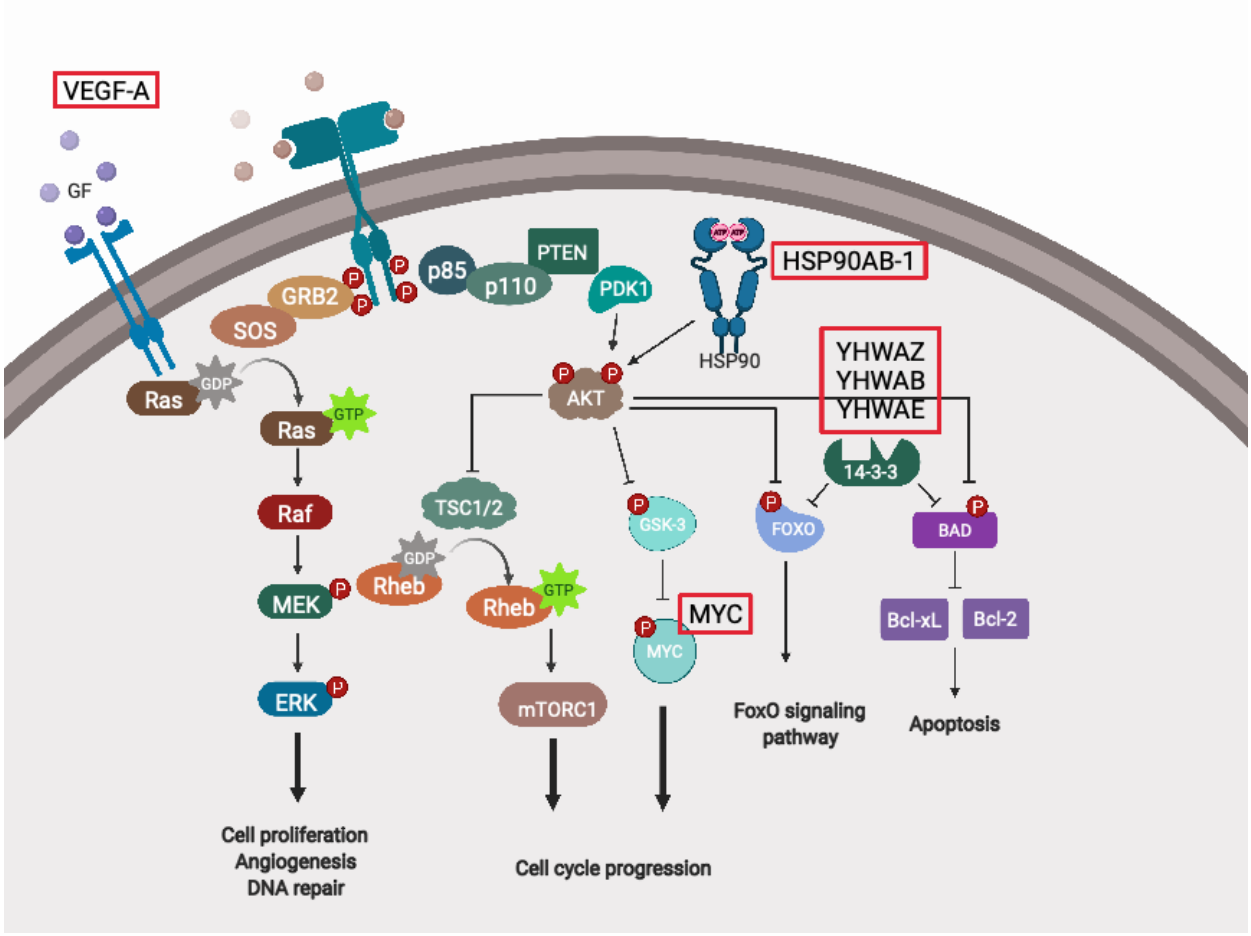

Figure 3.14 The role of alternatively spliced genes perturbing PPIs in signaling PI3K-Akt pathway. Parts of PI3K-Akt pathway with significant rewiring are highlighted red and corresponding genes are listed. HSP90 is a highly conserved protein which take in refolding denatured proteins under stress condition (379). Previous studies linked heightened HSP90 concentration with the high sugar high fat diet (380), and HSP90 inhibitors were demonstrated to reverse hyperglycemia in diabetic mice (332, 381). VEGF is a mitogen protein that takes part in angiogenesis. Increased levels of VEGF-A can improve insulin sensitivity in obese patients (382-384). 14-3-3 is a family of conserved proteins that regulate multitude of phosphoproteins, including those that are deregulated in diabetes, neurological disorders and cancer (385). Created with BioRender.com.

This T2D-associated network was then further curated down to 178 unique genes, which were involved in 244 PPIs, with the focus on rewired interactions for visualization purposes (Fig. 4). We highlighted the network hubs, important interactions based on betweenness centrality

measure (it reflects the number of paths in the underlying human interactome network that come through this edge), and the relevant biological pathways.

Perhaps the most interesting contributor to the AS-induced network perturbation was YWHAZ, a network hub protein implicated in the regulation of several signaling pathways that had recently been linked to a diverse number of diseases including T2D (386-388). Furthermore, three pathways previously implicated in T2D, PI3K-Akt, ER stress, and WNT (389-391), had a significant presence in the alternative splicing network (Figs. 3.13, 3.14).

3.4 Alternative splicing impact factor

3.4.1 Introduction

Alternative splicing is regarded as one of the major regulatory processes responsible for the production of multiple distinct RNA molecules from a single precursor. Even though this phenomenon has been known for 45 years and most of the mammalian genes undergo alternative splicing, there are conflicting opinions on the effects it has on functional diversity in complex organisms, with one camp arguing for it being a major player (392-394) and another pointing out a lack of empiric evidence of its large scale impact (395, 396). These issues impede research on alternative splicing role in diseases and other regulatory mechanisms, and even in many experimental studies, functional changes remain elusive. E.g., the authors of the extensive literature review (395) found conclusive evidence of functional distinctness (either presence or absence) for isoforms in <10% of considered studies. These issues call forth for finding additional insights from large-scale isoform analyses.

There is an argument that for a lot of isoforms abundance of alternatively spliced RNA material is relatively low, and such low-expressed molecules do not have the capacity to

influence cellular mechanisms. This is a very keen insight, as current methods in molecular biology cannot reliably detect the effects of such small-scale actors unless their role is crucial. Another point raised by the community is the observation that a high expression level does not guarantee a similar abundance of corresponding proteins. Though, recent studies point out limitations of applying current protein detection techniques to the alternative isoform studies due to technical reasons (e.g., trypsin cleavage specificity) (397). Considering the aforementioned points, we conclude that the RNA expression level of isoform is one of the viable proxies for its functional impact.

A common criterion for the functional necessity of the molecule is the effects observed in its absence (398, 399), which can be detected in an experimental setting using isoform-specific knockout. Unfortunately, this approach is time- and cost-demanding and currently is not viable on a large scale. Though we can detect one of the critical components for providing function – binding sites – using computational means and derive insights from the structural impact of the alternative splicing.

In spite of the growing number of links between the abnormal alternative splicing and many complex diseases [ref], our mechanistic understanding of this links is still limited because of the following reason. On one hand, scientists have been mounting evidence that many AS variants are functionally modulated through altered macromolecular interactions, including protein-protein interactions (64, 400-404), suggesting that AS variants could operate as phenotypic drivers in physiological and pathological pathways associated with the disease or medical condition. On the other hand, large scale studies of the expression of tissue-specific, cell development-specific, and disease-specific AS isoforms have revealed a wide expression range of alternatively spliced isoforms [refs]. Thus, an alternatively spliced isoform that has the

capacity of changing a protein function, compared to its reference isoform, might have a minimal physiological impact because it is not expressed in the tissue or cell of interest. Similarly, a predominantly expressed alternative isoform might have no physiological impact because it does not alter the original function carried by the reference isoform.

In this study, we propose a novel way of screening the functional impact of alternative isoforms – **A**lternative **S**plicing **I**mpact **F**actor (AS-IF). It is a quantification measure that combines RNA expression level with the modification extent of binding sites. AS-IF provides a high-level overview of the alternative isoforms that are likely to be functionally distinct from the reference isoforms, narrowing the focus.

3.4.2 Impact Factor Concept

The joint impact of these two factors, the functional modulation and expression, has never been addressed due to the lack of a formal mathematical framework that could quantify their contribution. Here, we define a concept of Alternative Splicing Impact Factor (ASIF), a novel quantitative measure of the functional impact calculated at the transcription level. Conceptually, the ASIF measure is first defined for a specific isoform I_X of a gene X expressed at one tissue type, cell type, or condition T (Fig. 3.16). We will refer to it as $ASIF(I_X, T)$. Second, it is generalized to all isoforms for the same gene X at the same tissue type, cell type, or condition T. We will refer to this measure as $ASIF(I_X, T)$. Lastly, it is generalized to all isoforms of a gene across all tissue types, cell types or conditions. We will refer to it as $ASIF(X)$.

The cornerstone of the impact factor measure is $ASIF(I_X, T)$. To formally introduce it, for each gene X, we first define an artificial construct, which we call a canonical isoform, and which is defined as a concatenation of all different exons discovered in all isoforms for that gene. Second, we define N as a number of all protein binding sites mapped onto the canonical isoform.

Third, we determine how an alternately spliced isoform I_X of X affects each of the molecular binding sites of the canonical isoform.

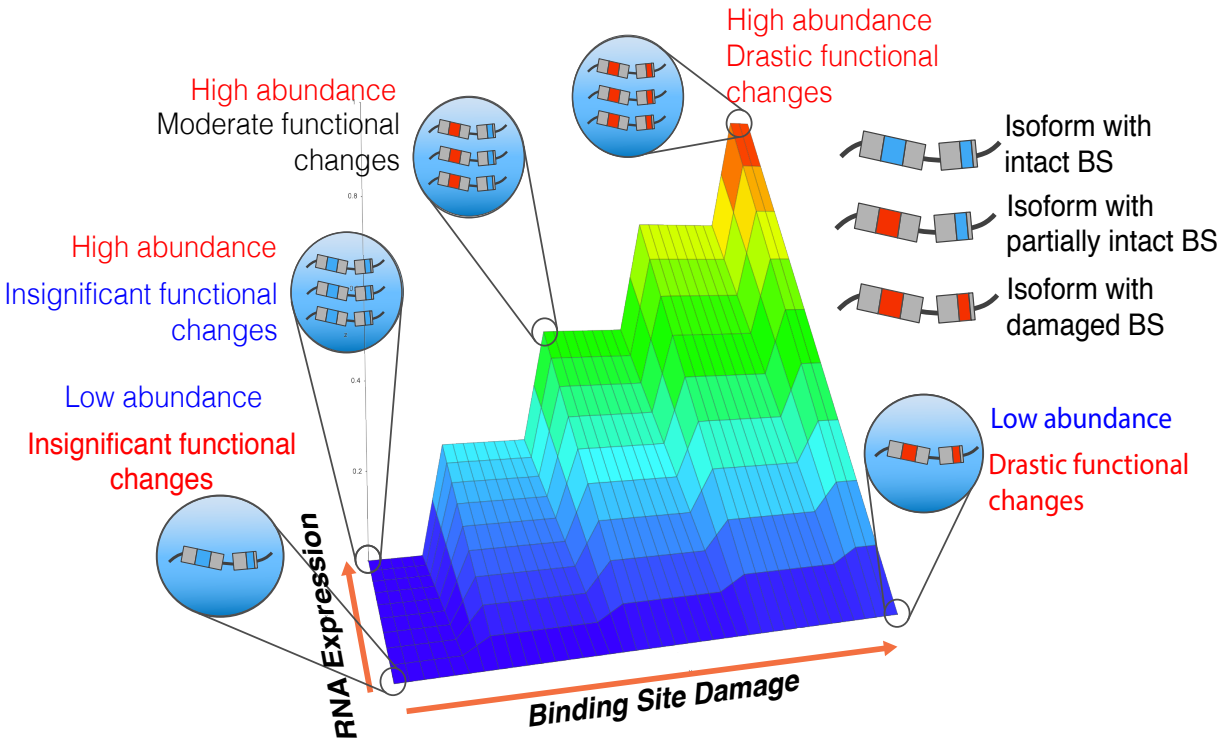

Figure 3.15. Impact factor landscape. The impact factor encompasses RNA-Seq expression and binding sites change of the alternatively spliced transcript in order to quantify the importance of the functional role it plays in the organism. Low scores in either characteristic (functional changes or mRNA expression level) result in small values for the impact factor. The figure demonstrates the hypothetical landscape of the values AS-IF takes for the transcript with four binding sites. It is based on the variability across the two coordinates: binding sites modification and mRNA expression. The expression level changes linearly and translates into the skewed slope from the figure. The conservation rate of the binding sites subsequently changes, meaning the conservation rate of the current binding site is gradually reduced until it hits zero. Only after that next binding site is getting modified. These changes translate into the 'ladder' pattern from the figure.

If $c_i(I_X)$ represents the preserved fraction of the i-th binding site (i.e., $c_i(I_X)$ = 0.0 when the binding site is fully removed and $c_i(I_X)$=1.0 when it is fully preserved), then the measure is defined as:

$$ASIF(I_X, T) = E(I_X, T)\left[1 - \frac{1}{N}\sum_{i=1}^{N} S\big(\alpha(c_i(I_X) - \beta)\big)\right],$$

$$S(x) = \frac{1}{1 + e^{-x}}$$

where $E(I_X, T)$ is the expression value of IX in T, $S(x)$ is the sigmoid function, α is the scaling ratio of the sigmoid function input, and β is the offset of the sigmoid function input. Both α and β are the parameters that ensure mapping of the binding site preserved fraction into a steep sigmoid curve and are determined empirically. Conceptually, the sigmoid function is selected to model a "step-wise" behavior of the isoform's impact w.r.t protein binding in a general case of N binding sites. Specifically, this function, on one hand, is dependent on the damage to the binding site(s) and deteriorates very quickly after a certain portion of the binding site is spliced out, but on the other hand, it is linearly dependent on the transcription level of that isoform for a specific tissue/cell/condition type T.

Each sigmoid estimates binding site function preservation, with value of 1 corresponding to fully functioning binding site and with value of 0 corresponding to the total functional loss. The average value of sigmoids represent cumulative function conservation of the isoform. Consequently, we have to invert this estimate in order to obtain functional change quantification for the transcript tr. Higher functional change directly proportional to the increase of impact factor measure. The values of α and β were set at 63 and 0.3 respectively in order to accommodate the following behavior: the first 10% modifications to the binding site leave it intact, however if the are additional 10% changes then binding site is not considered functional.

Using the above measure, we next define a tissue-type, cell-type, or condition specific ASIF measure for a gene X:

$$ASIF(X, T) = \max_{I_X}[ASIF(I_X, T)].$$

Once defined, for each gene one can calculate its functional impact profile across all types T: if for each T, $ASIF(X,T)$ is 0, then the gene is fully functional, and its functionality does not depend on tissue- or cell-type or on specific condition. Another extreme case is a gene with $ASIF(X,T)$ value of 1 for each T, which means that the gene lost its protein binding function due to predominantly expressed non-functional alternative isoform (Fig. 3.15).

Lastly, the overall functional impact factor of gene X across all types has a purpose of highlighting potential effect that alternative splicing can exhibit via the protein products of the gene:

$$ASIF_{ABS}(X) = \max_{T}\big(ASIF(X,T)\big)$$

We provide separate scores for the pathway analysis.

1. Cumulative

Highlights absolute impact value for a given gene set S. Corresponds to the total value of IF for all genes. Larger gene sets may be favoured.

$$S_{cu} = \sum_{g\in S} ASIF_{ABS}(g),$$

where S is a set of genes belonging to the pathway and $ASIF_{ABS}(g)$ is a system-level impact factor of the gene g.

2. Maximum

Indicates whether gene set S contains individual genes with high IF:

$$S_{max} = \max_{g\in S}\big(ASIF_{ABS}(g)\big),$$

where S is a set of genes belonging to the pathway and $ASIF_{ABS}(g)$ is a system-level impact factor of the gene g.

3. Average

Estimates AS impact in proportion to the gene set size:

$$S_{avg} = \frac{1}{|S|}\sum_{g \in S} ASIF_{ABS}(g),$$

where S is a set of genes belonging to the pathway and $ASIF_{ABS}(g)$ is a system-level impact factor of the gene g.

4. Tissue contrast

Identifies gene sets for which AS regulation significantly differs in small number of tissues. It helps to highlight tissue-specific processes among multiple gene sets.

$$S_{contr} = \max_{t_i \in T}\left(\frac{1}{|S|}\sqrt{\sum_{\substack{g \in S \\ t_j \in T, t_i \neq t_j}} ASIF(g, t_i) - ASIF(g, t_j)}\right),$$

where S is a set of genes belonging to the pathway, $ASIF(g, t_i)$ is a transcript-level impact factor for the gene g and tissue t_i. Bone marrow is excluded from the list of tissues T because of its low variability.

3.4.3 Data

This study focuses on the transcripts that were detected with high level of confidence. In order to achieve this goal we integrate data from six transcript databases – AS-Alps (405), VEGA (406), ASPicDB (274), ASTD (407), and Gencode (408). Because these databases

contain entries that were computationally predicted or manually curated, we select only those entries that have presence in at least four different sources. The result is a compendium of alternatively spliced isoforms (COMP-AS) that we are using further in study.

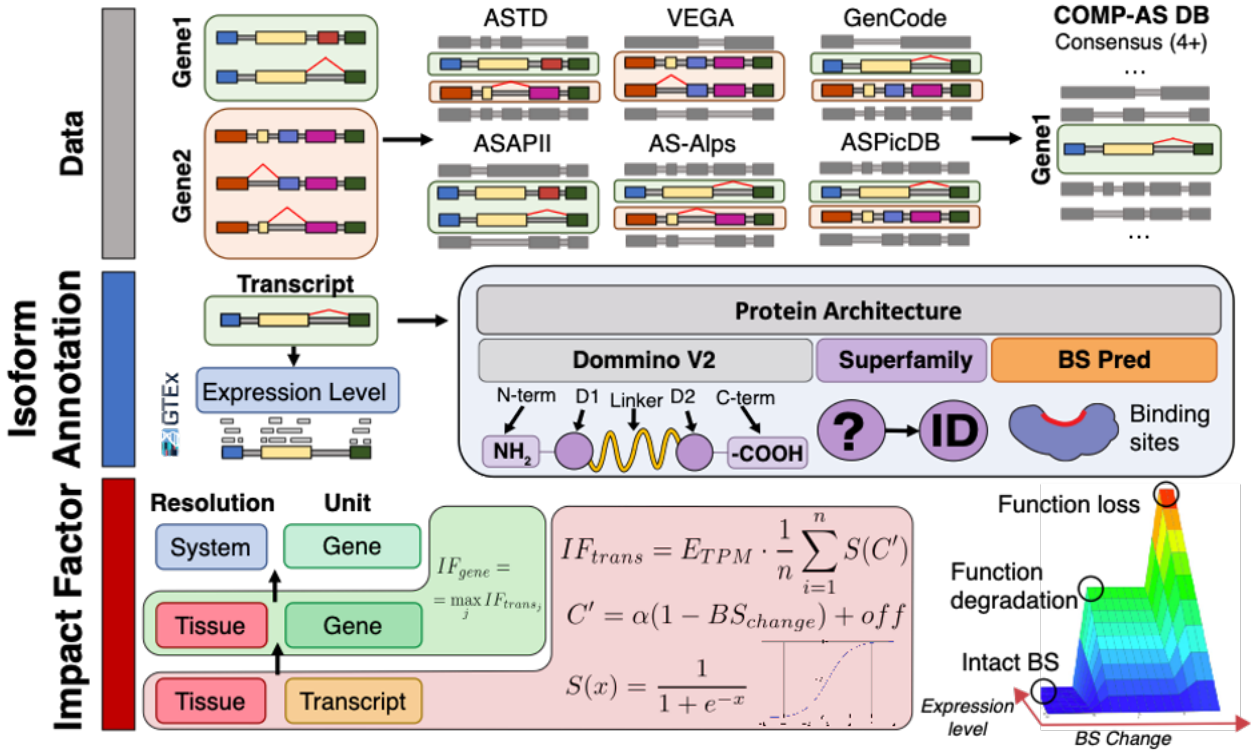

Figure 3.16. Impact factor computational pipeline. *Data.* Transcripts information is collected from six sources – Gencode, ASTD, Vega, ASPicDB, AS-Alps, ASAPII. Then, based on the consensus from at least four databases, high-confidence transcripts are selected into COMP-AS DB. *Isoform annotation.* At this step, each transcript is matched with the averaged mRNA expression level from the GTEx for the healthy individual for each tissue separately (75). Structural components such as N- and C-termini, linkers, and domains are annotated based on the Dommino V2 (247) and SUPERFAMILY (338) databases. Binding sites are predicted using a supervised machine learning algorithm. *Impact Factor.* The mathematical formulation of the impact factor on the tissue/transcript and tissue/gene levels along with the critical points depicted on the landscape figure of the transcript with two binding sites.

Binding sites were detected based on two competing approaches. The first one employs template-based search. It provides high-confidence predictions but may have limited coverage. To mitigate this issue a second approach leverages SCOP protein domain annotation and use

PSI-BLAST to map results to the sequences. It is a secondary source that is used when there are no available templates for the first method.

RNA-Seq data is obtained from the GTEx (75) public repository and contain 1,170 samples based on the material from the healthy humans.

Proteins were annotated based on DOMMINO V2 database (247). The boundaries of structural units such as N- and C-termini, linker regions were mapped onto each protein sequence. Functional domains were annotated via SUPERFAMILY (338) Hidden Markov Model.

3.4.4 Results

Alternative splicing is a complex tissue-specific process that exhibits profound changes in the organism during the developmental stage. Because of this, we provide several location-specific levels of quantification measure, though it should be noted that RNA expression levels in this study correspond to the adult human population, and variations occurring during developmental stages are not addressed in this work.

This work systematically studied alternative splicing and its effect on basic functional units of the proteins – binding sites, C- and N-termini, linker region, and domains. All reported results are based on compendium of alternatively spliced isoforms (COMP-AS), derived from six publicly available transcriptomics databases. The total number of transcripts in the compendium is 218,222 and they cover 16,682 distinct genes. Alternative splicing did not display high level of selectivity – most events partially modified structural units; the large number of clean deletions was observed only for linkers. Overall, 83% of binding sites were affected by alternative

splicing. This regulatory mechanism drastically alters structural composition of the protein, affecting more than 50% of important components.

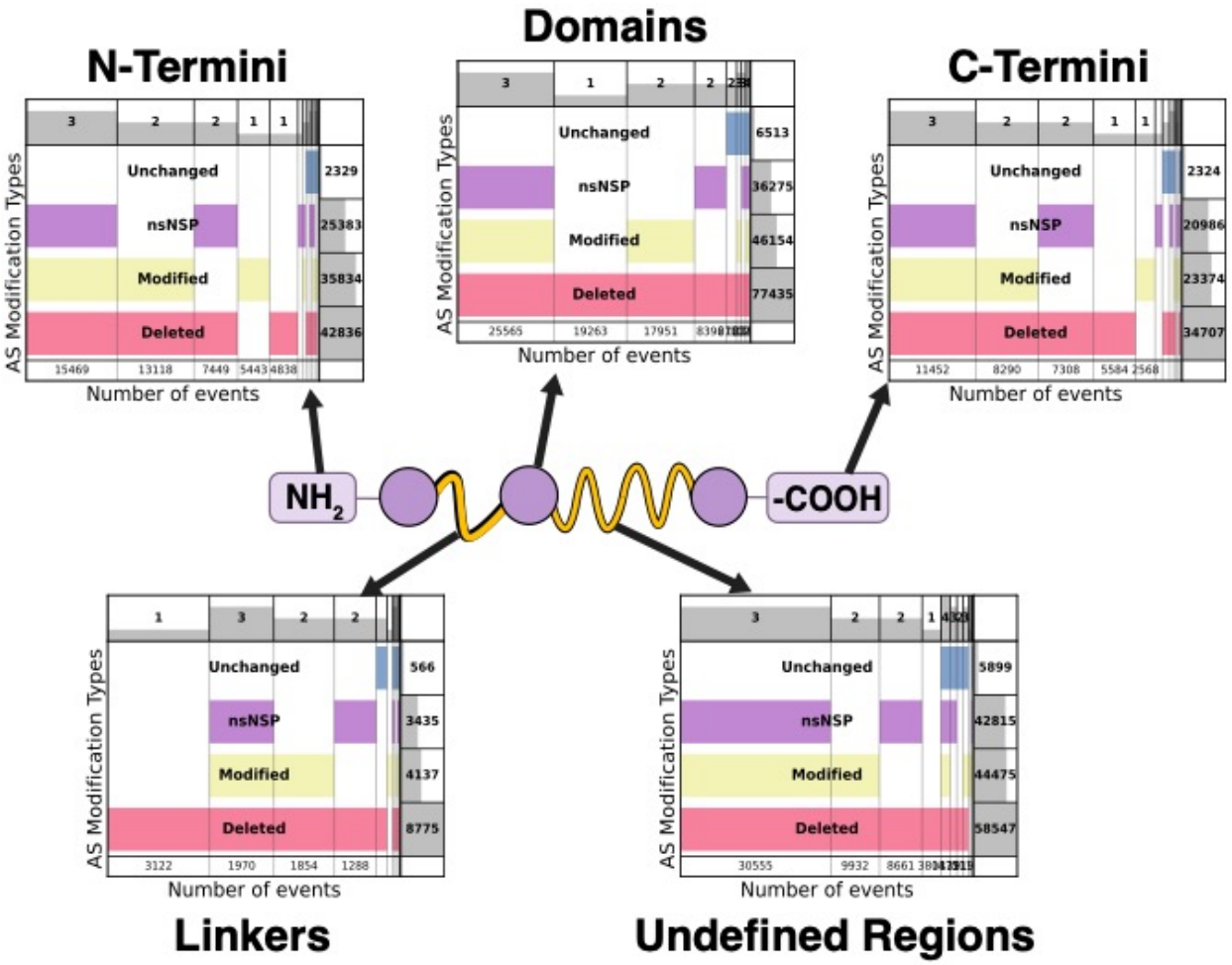

Figure 3.17. Structural units' modifications by the alternative splicing. Incidence of the four possible modification outcomes (modified, deleted, unchanged, non-synonymous single nucleotide polymorphism) across different structural units in transcripts: N- and C-termini, functional domains, linkers, and unidentified regions. It highlights non-specificity of the alternative splicing modifications in regards to those regions. Major part of studied transcripts contain both deletion and modification of nsSNP events. The linkers pose a notable exception, as a significant number of the stand-alone deletion events take place there.

The base variation of alternative splicing impact factor $ASIF(I_X, T)$ was calculated each transcript across 31 tissues. The average score for the transcripts in reach tissues is demonstrated in Fig. 3.18*B*. The most drastic changes are observed across testis, neural system, and fallopian tube. The most conservative tissue is bone marrow.

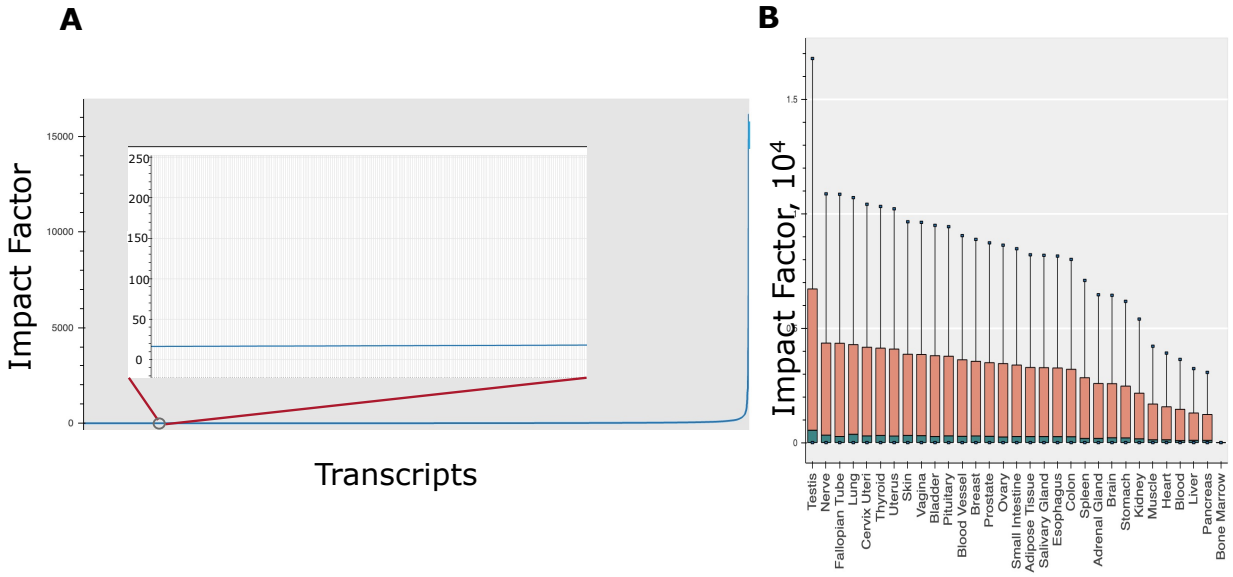

Figure 3.18. Structural units modifications by the alternative splicing. A. Ordered AS-IF values across different transcripts. Except for the small number of gene products (~5,000), for which we see exponential growth, impact factor measure increases linearly. B. Boxplot of the AS-IF scores of the isoform from different tissues. The sex-related tissues belong to the top-3 the most impacted, along with the neural system. Bone marrow was virtually uninfluenced by the alternative splicing.

Another important application of impact factor is the pathway profiling, it is demonstrated via the Fig. 3.19. This approach combines hierarchical clustering with data visualization to help researchers identify potential hotspots of alternative splicing activity and focus on the corresponding gene set.

Impact Factor (IF) Profiles summarize information Gene/Tissue level information for a given gene set S. This concept is similar to gene expression profiles. The same way clustering can be applied to the IF profiles to identify regulatory hotspots. In this work we use unweighted pair group method with arithmetic mean (UPGMA) (409) – a distance-based bottom-up clustering algorithm from scipy package. Due to the extremely high IF scores for a small subset

of genes we used percentile cutoffs (5%, 15%, 25%, 50%, 75%) to compare genes instead of absolute value. Visualizations were produced via bokehheat package.

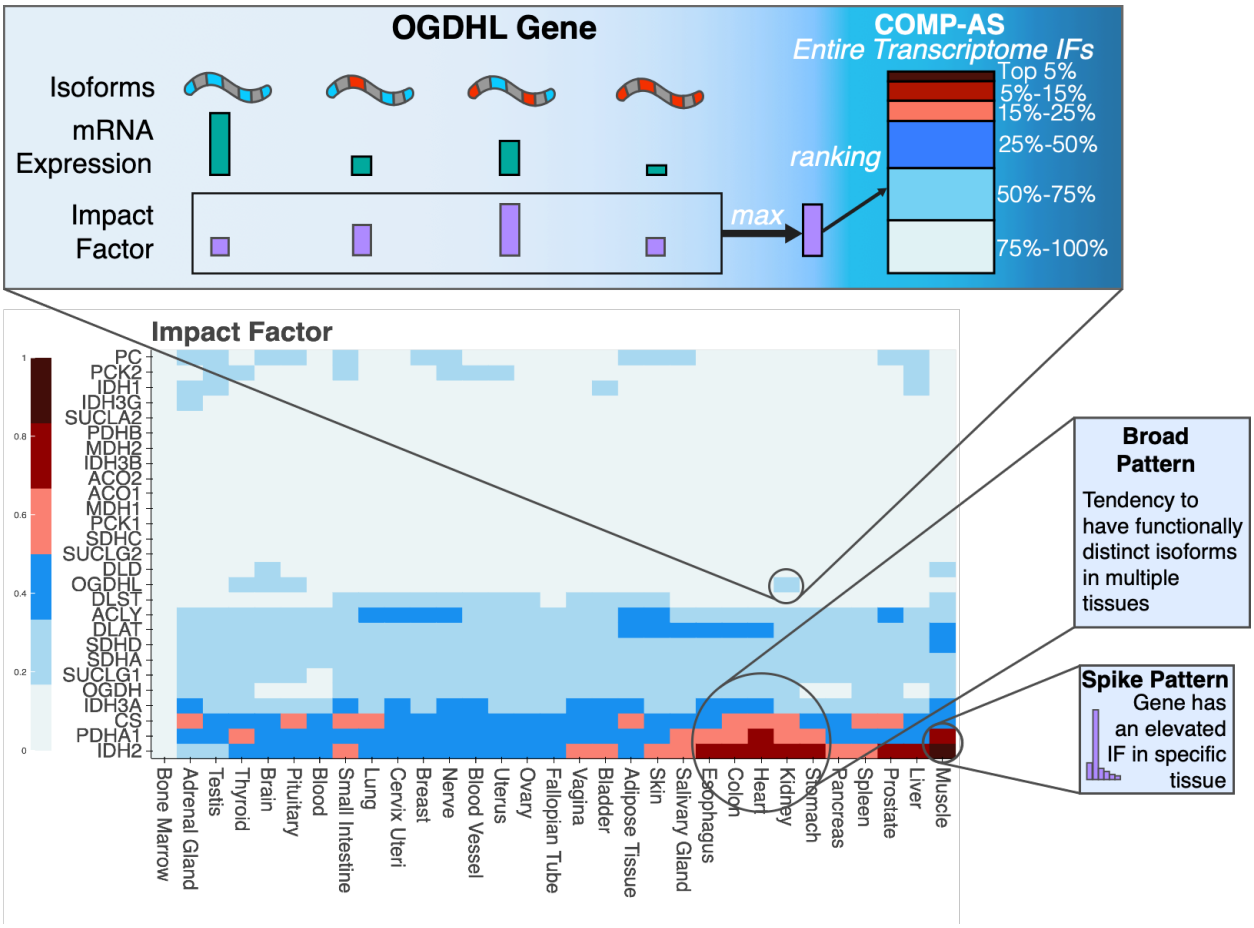

Figure 3.19. AS-IF pathway profile. Visualization of gene/tissue level alternative splicing impact in a set of gene provides opportunity for the comprehensive examination of the effector groups of genes. Hierarchical clustering can assist in identification of such groups. Potential alternative splicing impact can be studied for each tissue separately. Coloring scheme is based on the gene impact factor ranking amid all available data points. Genes from top 5% are depicted via dark red, 5-15% are wine-red, 15-20% are salmon, 25-50% are dark blue, 50-75% are blue, and 75-100% are light blue.

An IF profile collapsed to the Gene/System detalization level. Allows to quickly compare maximum potential impact of AS across multiple gene sets.

Existing methods for differential expression analysis can be adopted for the usage with IF measure, e.g., GSEA (410). In this work we compare control group with a group of breast cancer patients.

3.5 Isoform stability prediction

3.5.1 Introduction

Alternative splicing immensely contributes to the regulatory processes and allows an organism with roughly the same number of genes as its distant relatives from the evolutionary path to achieve new levels of complexity. The potential for the proteome enrichment it offers is immense. Applications previously described in this chapter assume that each alternatively spliced transcript is represented by a corresponding protein. However, there is a certain criticism directed at this point of view. Some researchers go as far as claiming that only ~5% of alternatively spliced isoforms result in functional proteins (396). These estimates are based on the comparison of transcript sequence databases (e.g., Ensembl (327), Gencode (408)) with the available mass-spectrometry (MS) data (Proteomics DB (411)).

MS methods make possible the analysis of the mass-to-charge ratio of ions derived from the peptides (412). This experimental technic evaporates the biological sample and sorts obtained ions according to the aforementioned ratio. Then relative abundance is quantified, producing an isotopic distribution spectrum. This information can be used to identify known peptides by matching them against the library. Deciphering a signal from novel peptides is also possible; however, it requires an additional manual or software-assisted intervention. The main challenge of applying this method to the alternative splicing proteome studies is that the entire protein macromolecule is too large for the analysis in one go, so it has to be broken down into a series of short peptides. This process is called digestion and has a significant chemical bias that produces peptides that are not evenly, or even regularly, distributed across the protein sequence.

Studying highly tuned proteins, proteoforms (413, 414), that have slight variation either due to the alternative splicing or posttranslational modifications is an important task as it allows to directly obtain actionable insights into the diseases and elucidates molecular processes taking place in the organism. It means for the AS research, it is insufficient to either limit the scope to the mRNA resolution or naively translate every sequence into the protein. Isoform stability concerns are completely valid. There is evidence that some isoforms result in 'junk' proteins that are not sustainable (395, 415, 416). However, our ability to precisely detect specific splicing variants is also highly limited by the current MS methods and particular protein digestion strategies (397). This situation necessitates a closer interplay between computational and experimental tools.

3.5.2 Dataset construction

Constructing a negative dataset is a highly non-trivial task, as the current scientific community does not possess protein detection methods with accuracy close to 100%. Current MS methods rely on protein digestion to produce short peptides that can be detected by the spectrometry. The problem lies with the fact that not every peptide can provide relevant information on the alternative splicing event and help us to confirm that we are indeed dealing with the isoform. So, the scope of the peptides that are relevant to our study is greatly reduced.

To mitigate this disadvantage, a DL-based model is used to identify peptides that are theoretically detectable by MS methods under a trypsin-based protein digestion strategy. Proteins without theoretically detectable peptides that contain information on alternative splicing events (e.g., splice junction, a peptide from a switched exon) are discarded. This is the main criteria on which candidates for the dataset are pre-screened. In order to confirm that a sufficient amount of biological material for the selected candidate is present in a sample tissue, an additional long-

read RNA-Seq experiment is conducted. It is necessary to further narrow down the scope of the search to only highly-expressed isoforms; otherwise, the probability of missing translated amino acid sequences would significantly increase. The last step is the confirmation of the isoform presence using MS methods. If necessary alternative splicing-enriched peptides are detected, the isoform is classified as stable. Otherwise, it is considered unstable.

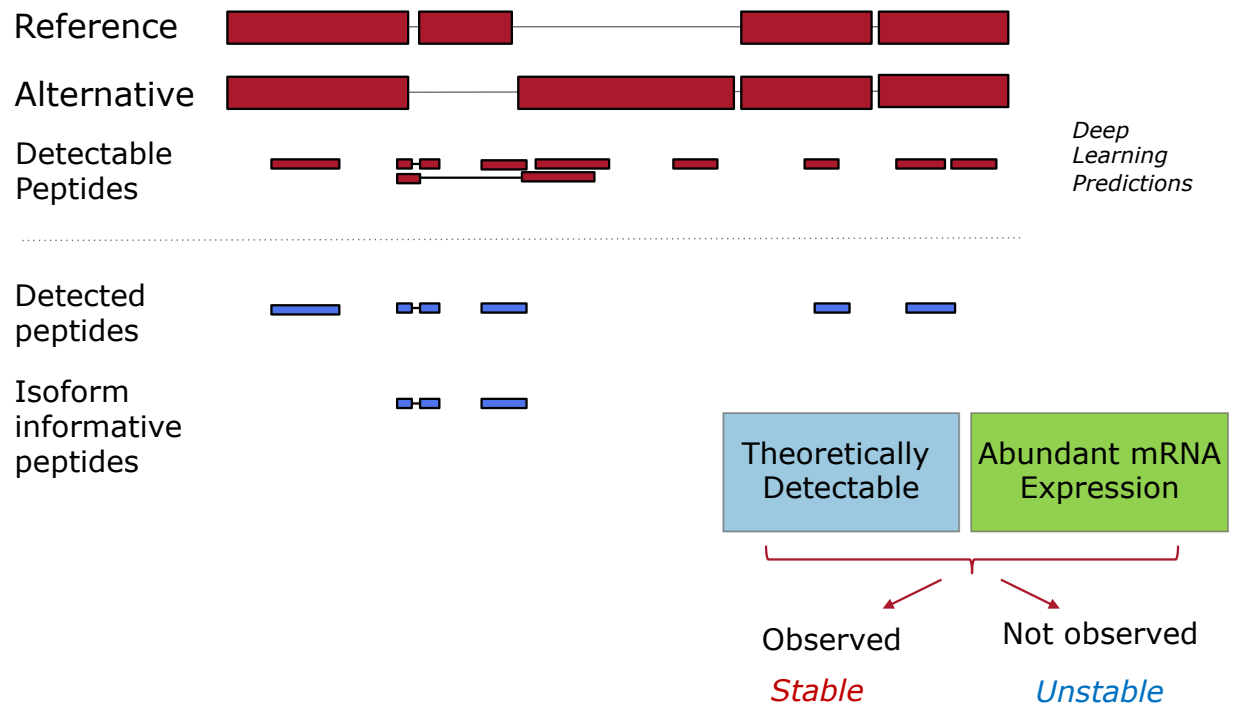

Figure 3.20. Construction of the negative dataset for protein isoform stability. The candidate alternatively spliced proteins should contain theoretically detectable peptides for the MS methods related to the AS events and be highly expressed on the mRNA level.

Overall, experimental confirmation of existence obtained 257 isoforms (stable) and 500 more were considered unstable.

3.5.3 Fine tuning of the DL model

Machine learning methods demonstrate their best performance when extensive amount of training data is available. However, expanding isoform stability dataset is an expensive task that

require usage of two distinct experimental techniques – mass-spectrometry and long-read RNA-Seq. Due to this restriction, we have to look for the other ways to improve model performance. As we are dealing with the protein sequences, one of the viable ways to achieve this issue is to adapt transfer learning.

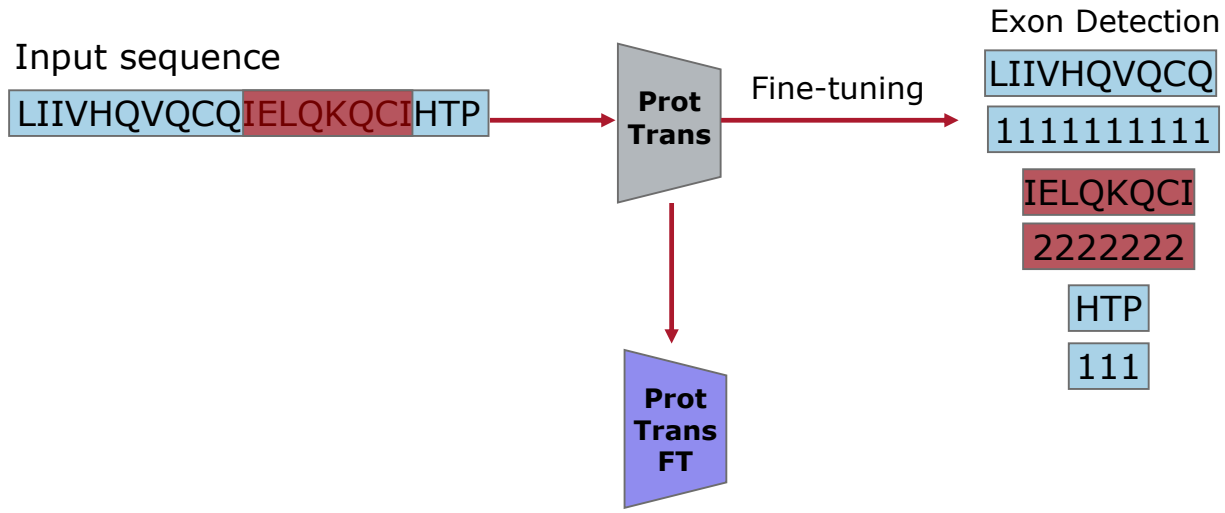

Figure 3.21. Fine tuning ProtTrans model on the exon segmentation task for the protein sequence. Auxiliary task for the fine tuning is to segment original protein sequence into distinct exons. Each exon is encoded by alternating numeric label of '1' or '2'.

One of the most comprehensive recent models that extract features based solely on the protein sequences is ProtTrans (417, 418). This approach employ a transformer-based self-attention model to learn how to predict masked elements during sequence-to-sequence training (419). It allows model to create context for each individual amino acid. ProtTrans model is able to distill a lot of relevant information about protein properties, e.g., biophysical characteristics of the amino acids, secondary structure, conserved motifs, and domain of origin (archaea, bacteria, eukarya, and viruses) (417). This results in a highly informative embedding that compressed protein profile. However, it is increasingly unlikely that this model learned relevant information about splicing without explicitly training on AS-centered task as this is a very complex problem.

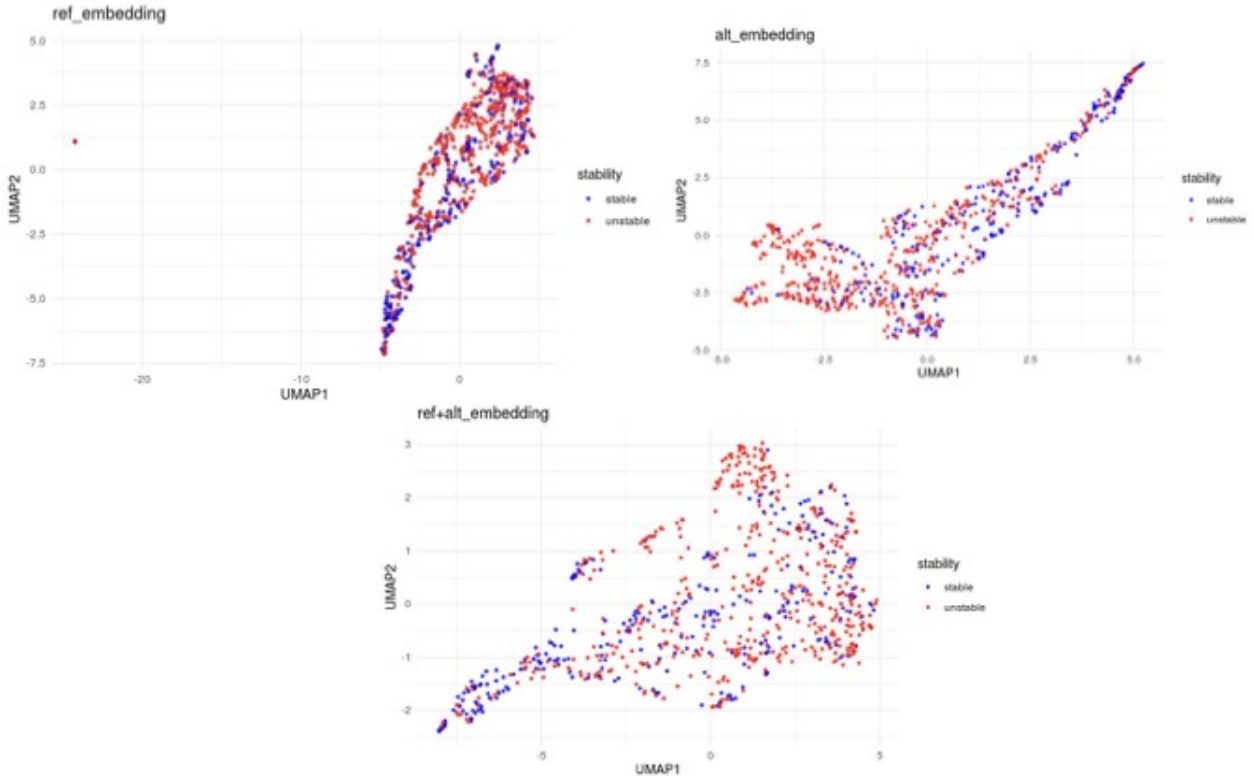

Figure 3.22. ProtTrans embedding of the protein isoforms. UMAP visualization of isoform stability dataset that depicts partial separation of stable (blue) isoforms from unstable (red). Embeddings are derived from reference isoform (top left), alternative isoform (top right). A pairwise concatenation of both embeddings is visualized at the bottom mid.

The problem of insufficient specialization of the machine learning model can be solved using fine-tuning on the auxiliary task. For the alternative splicing exon segmentation problem was selected. In essence, it is similar to the secondary structure prediction problem that assigns each amino acid its label – alpha helix, beta sheet, linker. When applied to the AS, the task becomes a task of identifying exon switching event. In order to accommodate complex proteins that cover a large number of exons we use only two labels that correspond to ‘starting exon’ and ‘different exon’. Each protein can be described using this set of alternating labels.

Training dataset for the auxiliary task was obtained from the Gencode database. Based on chromosomal coordinates of protein-coding regions and exons from annotated GFF3 files with we extracted corresponding DNA sequences from GRCh38 reference human genome (420). The length of translated into amino acid sequence was estimated for each derived exon and the obtained boundaries were mapped onto protein-coding sequences from the Gencode.

3.5.4 Results

Current results cover four distinct input data – main isoform embedding, alternative isoform embedding, selected top-300 features from both embeddings, selected top-300 features from concatenated embedding and ALT-IN Tool features. Performance of the auxiliary learning task has cross-entropy loss equal to 0.61, accuracy is 0.38, precision is 0.47, recall is 0.18, F1 score is 0.26. This indicates insufficient training depth of the machine learning model. One of the ways to mitigate it is to provide an additional biologically relevant information, another is to perform further model tuning using different class weights or an alternative loss function.

Instead of creating an additional dense layer in neural network we perform classification using extreme gradient boosting algorithm (XGBoost) (421) that took protein sequence embeddings as an input features. For the model assessment we used a stratified 10-fold cross-validation. The results are summarized on Fig. 3.23.

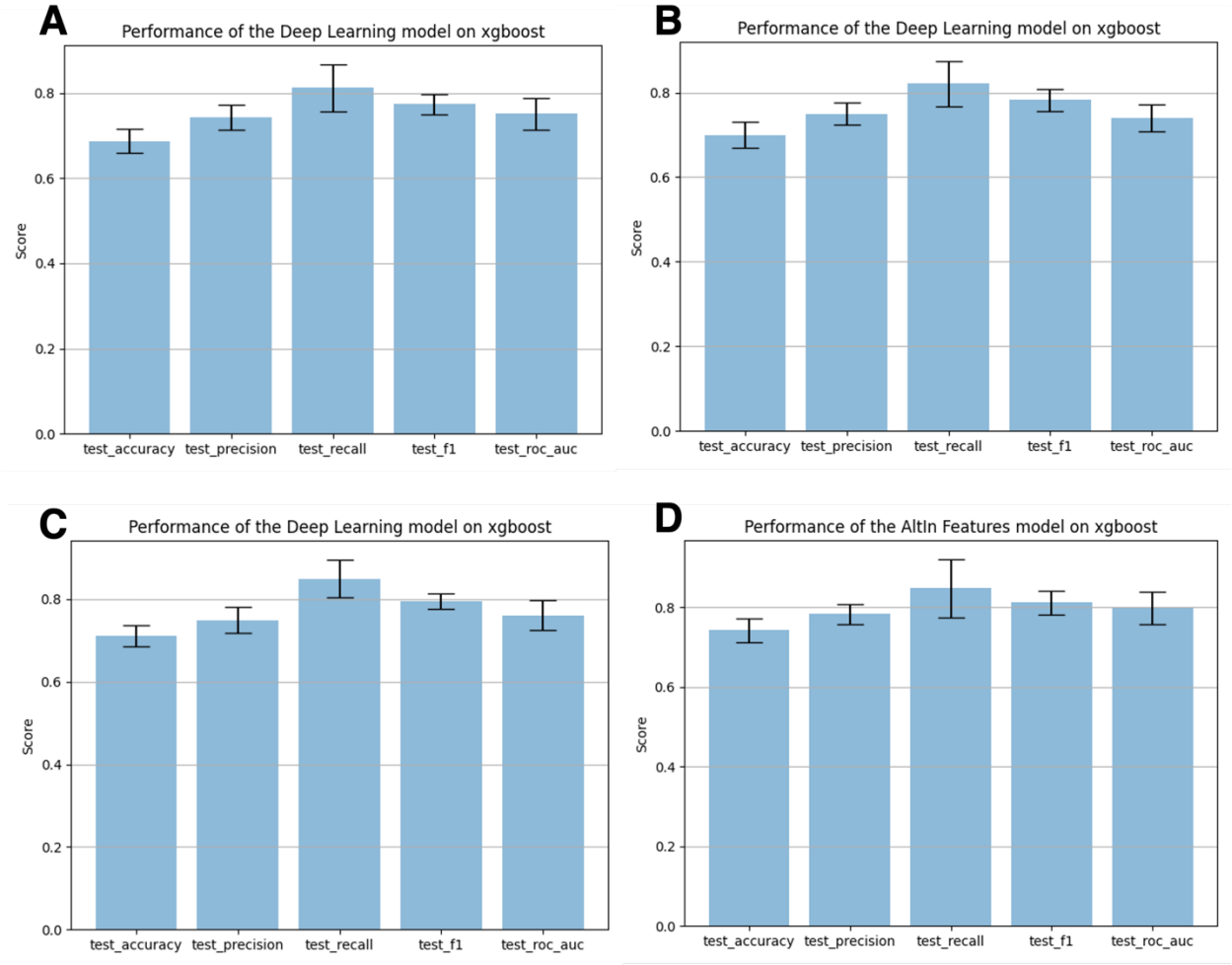

Figure 3.23. Isoform stability prediction results. Bars denote an average score and whiskers indicate standard deviation of each cross-validation run. A. Performance of the XGBoost model on the reference isoform embeddings. B. Performance of the XGBoost model on the alternative isoform embeddings. C. Performance of the XGBoost model on the 300 selected features from both reference and alternative isoform embedding. D. Performance of the XGBoost model on the 300 selected features from alternative isoform embeddings and ALT-IN features.

The current best performing model was a combination of deep learning and manually engineered featured (Fig. 3.23 D) based on five distinct metrics – accuracy, precision, recall, F1-score, and AUC. It indicates that fine-tuning procedure does not exhaustively extract information related to the alternative splicing and have to be further improved.

3.6 Proteomics-based pain studies

3.6.1 Introduction

Gulf War Syndrom, also known as Gulf War Illness (GWI), is a condition that affects around 30% of veterans of the Operation Desert Storm/Desert Shield – the Gulf War (1990-1991) and encompasses several varying and seemingly unrelated symptoms (422). The adverse effects that occur in the affected population include musculoskeletal pain (especially in the lower back region), fatigue, cognitive problems, skin rashes, respiratory complaints, brain imaging abnormalities, and diarrhea.

The etiology of GWI is not entirely understood, as multiple factors could contribute to the disease. They encompass multiple categories: environmental factors (oil-well fires' smoke, dust storms, insects, heat), physiological conditions (epigenetic causes, psychological stress due to the deployment and combat activities), chemical pollutants (organophosphates, carbamates, pyrethroids, insect repellents, organochlorine), and chemical warfare related substances (sarin, cyclosarin, mustard gas, pyridostigmine bromide tablets for prophylaxis against nerve gas agents) (423). Multiple studies investigated the potential contribution of each of those factors. However, due to the heterogeneous nature of symptoms, no single pre-clinical model was able to comprehensively cover all adverse conditions experienced by the veterans (423-426). Though, these models still played a significant role in devising intervention strategies.

The interplay between a large number of chemicals, environmental factors, and age-related ailments produces one of the most complex acquired syndromes. It significantly increases the difficulty of finding an efficient treatment strategy for the patients. The therapeutical interventions for the GWI include nutrition supplements, medication, exercise regimen,

mindfulness-based stress reduction, continuous positive airway pressure (CPAP), and acupuncture (427-432).

Pain is one of the most prevalent GWI symptoms that significantly reduces patients' quality of life and creates a burden on personal and social levels. It interferes with daily activities and creates constant distractions, reducing the individual's ability to stay engaged, interfering with sleep patterns. Among previously discussed intervention strategies, detox regimen, mindfulness, CPAP, and acupuncture demonstrated an ability to successfully reduce the pain level experienced by the patient (423). However, among presented treatment methods, acupuncture, though effective (431, 432), presents a certain level of risk of exposing the individual to adverse effects, a small number of which leads to major complications (433, 434). Due to these issues, medical providers, in order to make the best possible decision for a given patient, would need a pre-screen test that is able to estimate a response degree.

One of the viable approaches to this problem is an identification of the response biomarkers – a set of characteristics that can predict the outcome of a particular therapeutical approach (435-437). They may include various characteristics – genetic testing, brain imaging, metabolic products. One of the most convenient and ubiquitous tests in medical practice is the protein level measurement in blood. It requires only a liquid biopsy; it is a low-intrusion technique that does not significantly inconvenience patients. This study identified response biomarkers based on protein expression levels obtained via the SomaScan assay.

3.6.2 Methods

Clinical trial design

Clinical trials included thirty veterans that underwent acupuncture therapy for six months; 27 completed all questionaries necessary for the studies. To estimate a highly subjective characteristic, such as pain, two types of commonly used pain scales were utilized – SF36 and McGill (438, 439). For each scale, a delta pain measure was calculated – a percentage of change in questionary score from the base measurement that took place before the start of the acupuncture treatment. A liquid biopsy sample was collected from each patient no later than two months into therapy and then processed via SomaScan manual assay, providing an expression level of 1317 distinct proteins.

Pain measurement

The Medical Outcomes Study short-form general health survey (SF36) is a short questionary consisting of 36 distinct queries. It is ubiquitously used in health-related quality of life studies and is able to reliably demonstrate differences between a healthy population and a cohort affected by chronic diseases (440). SF-36 provides scores for the eight distinct profiles that form the physical and mental composite scales. The physical composite scale includes vitality, physical functioning, physical problems, and pain. The composite mental scale covers the following profiles: general health perception, role limitation due to the emotional problem, social functioning, and mental health. In this study, we focus on the pain scale.

McGill pain questionary is designed to quantify different aspects of the subjective pain experience (441). The participant has to assign a number to the 5-point intensity scale that reflects his experience with 78 distinct words describing affective, evaluative, and sensory aspects. This questionary is widely used in clinical pain studies.

Measuring protein levels can provide a snapshot of the current processes taking place in the organism and can be indicative of various health issues: blood pressure, inflammation, coronary artery disease (442, 443). Among the various tools in the precision medicine arsenal, proteomics is one of the most practical instruments as it opens doors to the design of inexpensive liquid biopsy-based tests (444). Recently, many efforts have been dedicated to the development and improvement of high-throughput proteomics assays from the companies such as Olink and SomaLogic. SOMAscan is one of the high-throughput assays provided by SomaLogic that is based on short oligonucleotides with highly selective binding affinity – aptamers.

High-throughput proteomics

The aptamer-based technology has the following workflow. Slow Offrate Modified aptamers (SOMAmers) reagents are synthesized along with supporting entities – fluorophore, photocleavable linker, and biotin, which is used to place the entire complex on streptavidin bead. Fixed in place reagents capture proteins with the corresponding binding affinity from biological material. Excess proteins that were not captured are washed away, and molecules on top of the bead are labeled with biotin. Then UV light is used to remove the photocleavable linker and allows formed protein complexes to drift freely in the solution. When non-specific complexes are subjected to the motion, they dissociate, and polyanionic competitors are added to prevent rebinding. In the next step, biotin labels are used to reattach protein-aptamer complexes back to the streptavidin bead, which is afterward isolated and added to the fluorescent array. The final optical intensity is measured by a fluorescent microscope, similar to the gene expression arrays. Intensity is quantified in relative fluorescence units (RFU). To keep RFU measurements roughly on the same scale, specific protein groups undergo a different degree of dilution. It is necessary to avoid a scenario of being unable to measure differential expression for the ubiquitous protein

because, during each experiment, it saturated streptavidin bead. Each biological sample is processed on a specific plate, where it is placed in one of the subarrays that hold a total of eight portions of biological material, including technical entries such as calibrator and buffer.

The presence of technical variations introduces noise into laboratory measurements. It makes necessary the data normalization procedures. In this work, I explore the effect of adopting various normalization strategies at the intraplate and interpolate step.

Normalization methods

Normalization methods holds a purpose of reducing technical variability of the data samples obtained via particular experimental technology, in this case – SOMAScan proteomic assay. SomaLogic normalization procedure is split into two parts: intraplate and interplate. Intraplate procedures include following steps: hybridization normalization and median normalization. In this study I explore additional alternatives that include computational methods based on the background correction (EK), quantile normalization, and elastic net linear regression. Intraplate methods that were studied include calibrator normalization procedure (standart SomaLogic approach) and ComBat, a wide-spread statistical approach for batch effect correction for the genomics data.

Hybridization normalization

A part of the standard SomaLogic workflow, hybridization normalization is making use of 12 artificial proteins that are added to each sample – hybridization probes. These probes highlight the differences between the proteins that fall on different parts of the RFU scale among samples from the same plate. It helps to reduce technical bias, as the amount of each hybridization probe in every sample is the same.

Step 1. For each hybridization probe i identify median among calibrators M_i.

Step 2. Calculate scaling factor $SF_i = \frac{1}{M}$ for each hybridization probe i.

Step 3. For each sample, multiply each protein's expression by the SF_i corresponding to the closest probe i.

Hybridization normalization step uses specific technical information added to SOMAScan assays explicitly for this purpose and is indispensable. This normalization step is always applied before any of the subsequent intraplate methods.

Median normalization

This normalization step is a part of the SomaLogic workflow. It is centered around factoring technical aspect of the assay – the usage of varying dilution groups and distinct sample categories. It is necessary to account for both technical and biological variations.

The normalization steps are the following. For a given sample type category (e.g., calibrator, technical replicate, biological sample) and a protein group with the same dilution (e.g., 40, 0.05, etc.):

Step 1. For each protein and each sample, calculate ratio r:

$$r = \frac{RFU}{RFU_{median}}$$

Step 2. Calculate scaling factor SF for the sample as

$$SF = \frac{1}{r_{median}}$$

Step 3. Median SF' is calculated across all proteins in the dilution group.

Step 4. Multiply all proteins in the same dilution group by SF'

Median norm for calibrators

The same as the Median normalization but is applied only to the calibrator samples (445). It was hypothesized that removing sample-to-sample differences from the inherently heterogeneous biological samples increases plate bias and have adverse effects during interpolate normalization step.

ElasticNet normalization

This normalization is closely related to the median normalization procedure, however, instead of the directly scaling by the RFU_{median} we solve a problem based on the elastic net – regularized linear regression (446):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta}[L(\lambda_1, \lambda_2, \beta)]$$

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 \|\beta\|_2 + \lambda_1 \|\beta\|_1$$

Step 1. For each SOMAmer find median value for calibrators samples.

Step 2. For each SOMAmer calculate scaling factor from calibrators RFUs based on the elastic net without intercept that fits median for this SOMAmer from Step 1.

Step 3. Apply corresponding scaling factor to each SOMAmer.

Quantile normalization

Quantile normalization is a popular normalization method in the genomics area that map samples into same distribution (447). It was developed for the gene expression microarrays but

was later adopted for the use with other high-dimensional omics data types, e.g., RNA-sequencing.

Steps of the quantile regression:

1. Order values within each sample (column)
2. Average across rows and substitute values with the average
3. Reorder averaged values in the original order

EK Normalization

Inspired by the similarities between gene expression array technology and SomaScan fluorescent microarrays, EK normalization incorporates plate-level background correction for the total fluorescence. Because specific positions of the probes are not disclosed, we normalize expression values based on the total signal intensity of the plate. This method assumes proportional distribution of the biological samples across different plates and designed to mitigate systematic effects introduced during each experimental batch.

Relative intensity can be calculated using the following steps:

1. Calculate total plate fluorescence level RFU_{sum} by summing all expression values.
2. Divide RFUs corresponding to each individual aptamer by RFU_{sum}.

EK-based methods

Previous normalization strategies can also be applied on top of total fluorescence normalization such as EK method. ElasticNet EK normalization is an application of RFU_{sum} adjustment on top of ElasticNet adjusted data. Median EK normalization is constructed in the same way for the Median-normalized input.

ComBat

ComBat normalization methods amends insufficient strength of linear and qspline normalization methods by leveraging Bayesian framework to adjust correction for both scale and probe location (448). Comparative study demonstrated that ComBat performance is more precise than selection of five other popular batch effect correction algorithms (449).

ComBat has two distinct steps – data normalization and batch effect adjustment. Normalization step is based on the ordinary least squares (OLS) estimates for the expression level mean and standard deviation of $k-th$ probe:

$$Z_{ijk} = \frac{Y_{ij} - \hat{m}_k - X\hat{\beta}_k}{\hat{\sigma}_k},$$

where $\hat{m}_k$ is the mean, $\hat{\sigma}_k$ is the standard deviation, Y_{ij} is the unadjusted expression level, $X\hat{\beta}_k$ are the biological covariates, Z_{ijk} is the normalized expression that follow normal distribution $Z_{ijk} \sim N(m_{normal}, \sigma_{normal})$. Final expression level is computed using the following formula:

$$Y^*_{ijk} = \frac{\hat{\sigma}_k}{\sigma_{normal}}\left(Z_{ijk} - m_{normal}\right) + \hat{m}_k + X\hat{\beta}_k$$

Linear regression

Linear regression describes a linear relationship between the response variable and explanatory variables (factors)(450). It can be mathematically formulated in the following way:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where y is the response variable, k is the number of factors, x_k is the explanatory variable, and is the residual noise. One of the most widely used implementations of linear regression is the

ordinary least squares (OLS) model. It works by minimizing the squared distance from training samples to the predicted values.

Multicollinearity in the data

There are two main conclusions. First, the variability in the model that was considered high by the statistical library we used was mostly a numerical issue; a simple Z-normalization completely removed those artifacts. This type of normalization performs linear transformation over the data, which should not affect linear regression. To emphasize this point, we made sure that the regression results stayed the same event if we computed mean and standard deviation over as low as three random samples. Still, to make sure that each biomarker contributes significant information, we performed the following procedure. We added new features to the regression model one by one, starting from the most significant (according to the feature selection method) while calculating the variance inflation factor (VIF) of each new candidate feature (451). If VIF was larger than a cutoff (set to 5 based on the general standards), then we did not include the corresponding feature in the final model.

Adjusted R^2

Similar to the regular coefficient of determination R^2, adjusted R^2 characterizes data goodness of fit but takes into account a number of factors in the model. It penalizes the addition of variables that do not substantially improve the model. Adjusted R^2 can be calculated using the following formula:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1},$$

where R^2 is the coefficient of determination, n is the number of samples, and k is the number of factors.

Spectral co-clustering

Detecting general groups of highly correlated protein expressions proved to be a more challenging task. We used spectral co-clustering (452, 453) to construct block matrices from the correlation matrix of the 20 most significant features. The downside of this algorithm is the inability to determine the true number of clusters; it needs to be specified manually. Because of this, we include several values in the analysis. There are images of reordered correlation matrix where each gene is assigned to the cluster. We found that clustering for 7-8 groups looks the most reasonable for detecting highly correlated groups of features.

Coefficient of variation

Coefficient of variation (CV) is a statistical measure that assesses variability of series independently from the units of measurement (454, 455). It has an advantage of creating a unified score for the series belonging to different scales, which is impossible to achieve with standard deviation. CV can be calculated directly by dividing standard deviation by mean:

$$CV = \frac{\sigma}{M}$$

gPCA

Guided principal component analysis (gPCA) is a statistic designed to quantify batch effect presence in the genomic data (456). It is based on incorporating information about batches into standard principal component analysis (PCA) algorithm. In the core of PCA algorithm lies singular value decomposition (SVD).

SVD is rooted in the mathematical fact that you can present a matrix of real numbers as a product of three distinct matrices. For a n by p matrix X the following entities exist: $n \times n$ orthogonal matrix U, $n \times p$ diagonal matrix D that contains singular values, and $p \times p$ orthogonal matrix V, such as

$$X = UDV^T$$

Based on the matrix of right singular vectors V we then can calculate principal components from the product

$$P = XV$$

and are encoded by the column vectors $(P_1, P_2, \dots, P_p)$.

In contrast to the regular PCA, gPCA incorporates an additional information on batches with the help of $n \times b$ indicator matrix Y, where b is the number of samples. An entry y_{ij} is 1 if sample i belong to the batch j and is 0 for all other positions. In gPCA, an SVD decomposition is performed on the product $Y^T X$. In such a setup orthogonal matrix U contains information on batch loadings, while matrix V continue to hold data on feature loadings.

The final statistic is calculated according to the following formula:

$$gPCA = \frac{var(XV_{g1})}{\sum_{k=1}^{b} var(XV_{gk})}$$

The larger gPCA value, the more prominent is the batch effect presence.

3.6.3 Data

27 sample liquid biopsy samples from the veterans affected by GWI were collected prior or at the beginning (no later than two months) of six-months long acupuncture therapy course.

Each sample was processed on the 1.3k SOMAScan manual assay in plasma matrix. The samples were located on three distinct plates, containing two paired bridge samples.

The dataset for normalization methods assessment was produced by the Trans-NIH Center for Human Immunology, Autoimmunity, and Inflammation (CHI), National Institute of Health (445). The samples were processed via two distinct assays – one portion of dataset is run manually on 1.1k Assay with 32-well plates, and another were processed on semi-automated 96-well plates 1.3k Assay. For this study we focused on the plasma matrix and leaving out serum-based samples. The 1.3k Assay data contains artificial bridge samples QC_CHI that were constructed from pooled plasma from 21 individuals (10 male, 11 female) with the median age of 57.

3.6.4 Results

We conducted a comprehensive comparison of the normalization methods for both intraplate and interplate stages with the goal of mitigating batch effect observed in the clinical samples. This part of the study is based on comprehensive dataset for plasma collected by CHI.

Initial assessment of intraplate normalization step was conducted based on variability across technical calibrator samples. As you can see in Fig.3.24, results for different strategies are comparable, except for Median EK Normalization, which drastically increases variability.

The primary goal of normalization methods analysis is the identification of optimal strategy to reduce batch effect for biological samples. The best way to identify such a strategy is the study of technical replicates – biological samples that contain near identical material. However, this type of data is present only for automated 1.3k assay (Fig.3.26 *B*). Because of this, we are using calibrator values across different plates that contain artificial and biologically

irrelevant information but is still possible to use as a proxy to distinguish relative changes in median CV value for the entire set of aptamer RFU values, like we did for 1.1k manual assay (Fig.3.25 *A*). A correlation between calibrators' and technical replicates median CV values can be observed on Fig.3.26 *A,B*.

Intraplate calibrator strategy demonstrates slight reduction for the CV in comparison to the raw interpolate data. No adverse effects on CV are observed across all base methods, with most of them experiencing a significant decrease, except for median normalization. ComBat method also demonstrates its ability to reduce median CV across multiple batches.

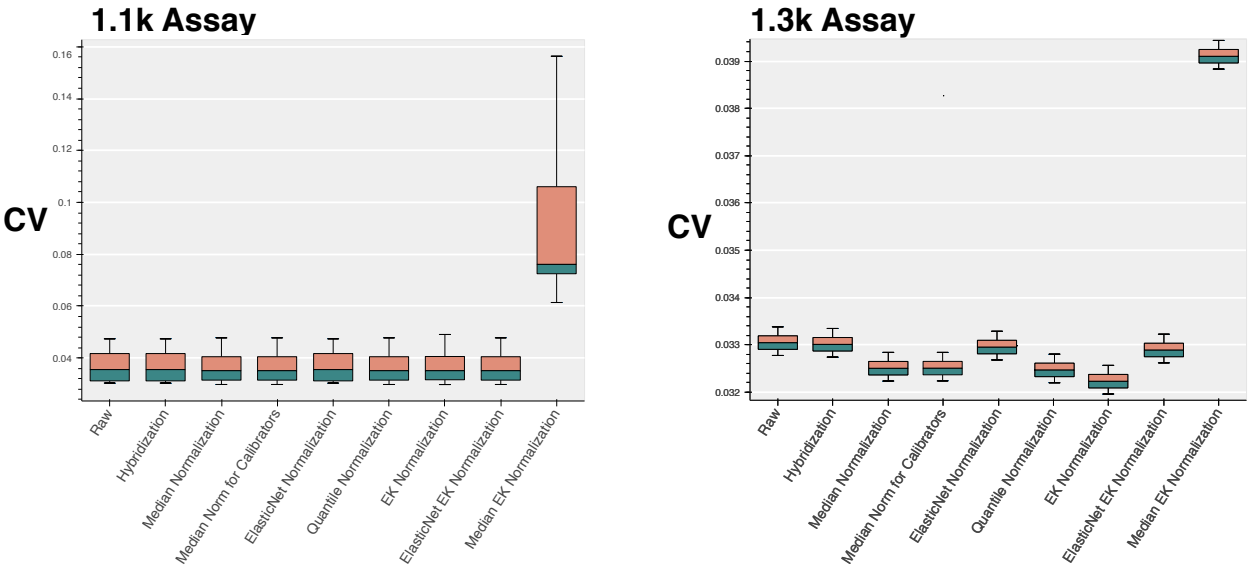

Figure 3.24. Intraplate normalization. CV estimates for the calibrator probes for 1.1k and 1.3k assays after applying various normalization strategies. "Raw" category corresponds to not applying an interpolate normalization methods. Most of normalization methods provide comparable results, except for the Median EK Normalization, which drastically increased CV values.

We also note that methods based on adjustment for total fluorescence demonstrate a significant decrease CV values for 1.1k manual assay (Fig 3.25 *A*), but effect of this correction

strategy is limited for automated 1.3k assay (3.26 *A,B*). It allows us to draw a conclusion that application of EK-based methods should be limited to the scenarios where technical variability for the individual plate is prominent, i.e., a manual work was conducted.

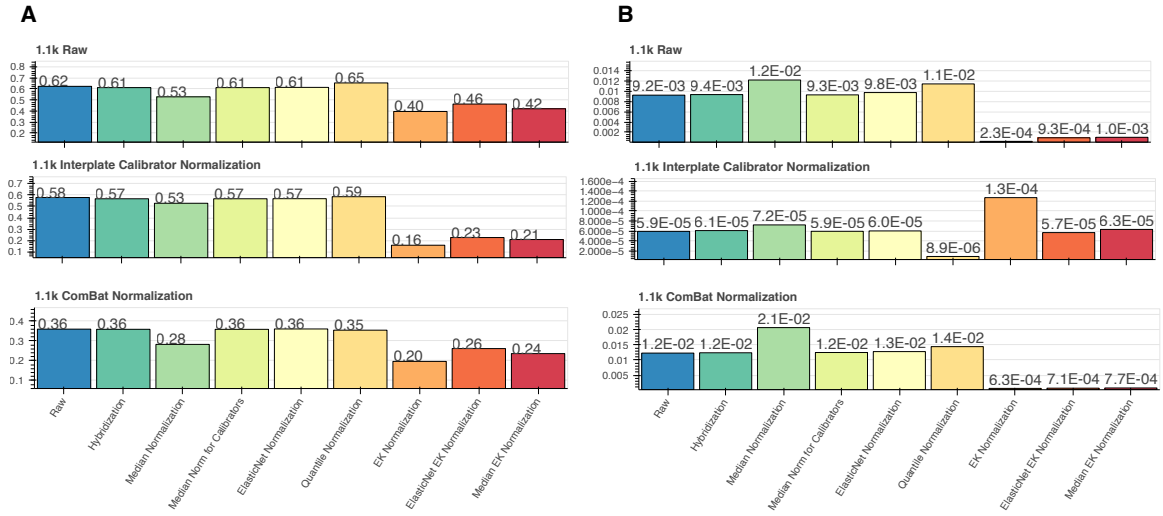

Figure 3.25 Interplate normalization of the 1.1k manual assay. A. Median CV expression across calibrator probes for three interplate normalization strategies: no normalization, calibrator normalization, ComBat. These expressions applied to nine intraplate normalization strategies. B. gPCA values across calibrators for interplate normalization strategies. ComBat demonstrates

Another potential contributing factor to the observed differences is more than 200 additional probes present in 1.3k assay. Higher CV values for the raw data may indicate. However, interplate calibrator normalization displays consistently lower median CV scores for the 1.3k array. It means that the most of variability due to the difference in probes was mitigated by this procedure.

The batch effect measure used in this study is gPCA. The results for the 1.1k manual and for the 1.3k automatic assays are listed in Fig. 3.25 B and Fig. 3.26 C correspondingly. Interplate calibration consistently showed better results than ComBat normalization procedure. It may be due to the usage of technical information from the SOMAScan platform. As ComBat

demonstrates higher gPCA values along with consistently lower median CV, it is likely that reduction in variability is due to the loss of biologically relevant information.

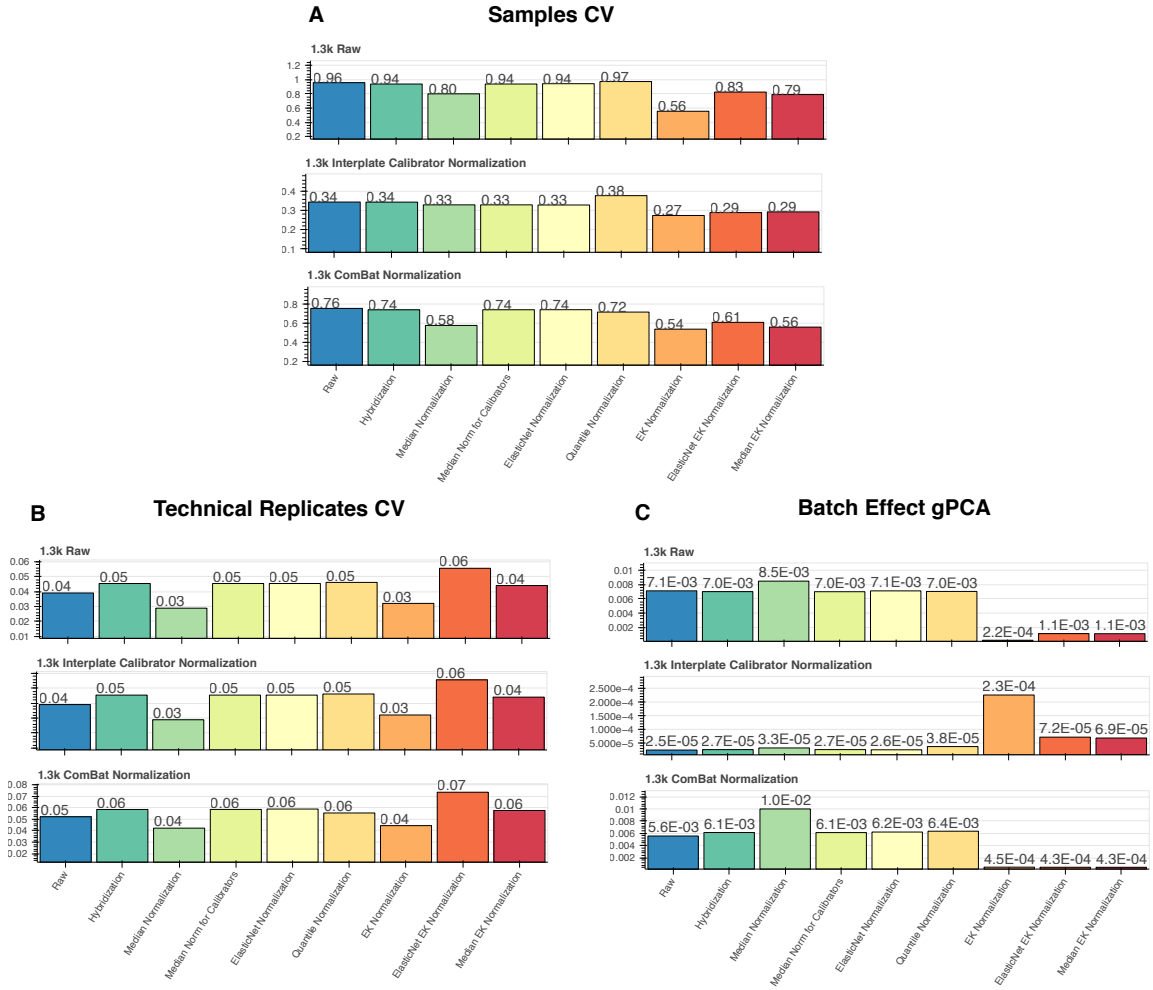

Figure 3.26 Interplate normalization of the 1.3k automated assay. A. Median CV expression across calibrator probes for three interplate normalization strategies: no normalization, calibrator normalization, ComBat. These expressions applied to nine intraplate normalization strategies. B. Median CV across technical replicates (pooled plasma samples1). C. gPCA values across calibrators for interplate normalization strategies.

Overall, we based the final selection of the normalization strategy on the 1.1k manual assay result because manual component exhibited much more prominent effect on the results than an addition of aptamers to fluorescent array. We decided to apply ElasticNet EK

normalization, as it displays a significant reduction of the median CV values across calibrators, along with slight reduction of gPCA (Fig. 3.25).

Based on normalized data, we built a linear regression model for SF-36 and McGill delta pain response based on the individual proteomic makeup of the veterans that was assessed before the acupuncture therapy started (Fig. 3.27, Fig 3.28).

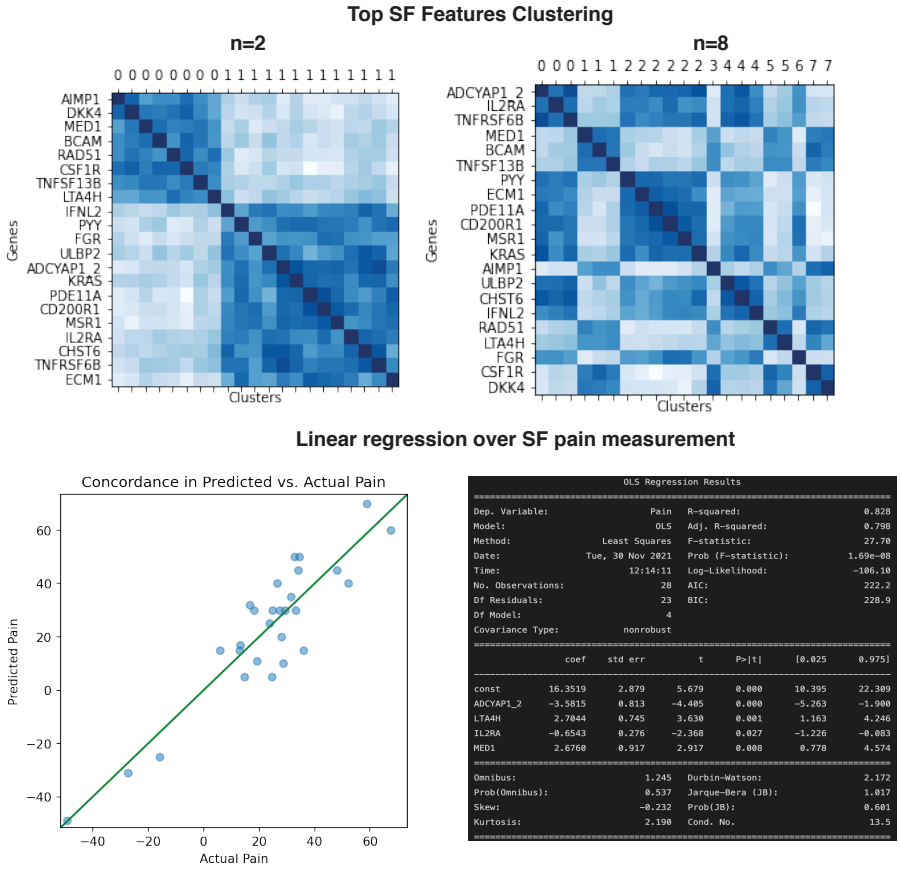

Figure 3.27. SF scale delta pain response prediction. Top. Spectral co-clustering results for the top-21 features from the largest linear regression model into two (left) and eight (right) distinct groups. Bottom. Trimmed model based on the four factors. Concordance plot is on the left side, statistics on the OLS model performance is on the right.

We applied univariate feature selection approach – F-test for regression (457) from scikit-learn package (356). At this step we selected top-50 features. We selected an optimal model by subsequently adding new factor variables into the model and measuring adjusted R^2 score for 10-

fold cross-validation. New variables that had VIF larger than 3 when compared to the current model were skipped. The best-performing models were then trimmed based on the p-value significance. In the final models only protein product based of the IDUA gene displayed $p = 0.057$, while all other features are classified as significant with $p < 0.05$.

Performance of the response level prediction models have high correlation with true values. For the SF-36 model the $R^2_{adj} = 0.8$ and for the McGill pain response it is $R^2_{adj} = 0.79$. Both models contain four variable that do not overlap. In order to provide an overview of the most significant parts we map protein product detected by SOMAScan protein assay back to the gene of origin.

SF-36 delta pain model is based on proteins derived from four distinct genes: ADCYAP1, LTA4H, IL2RA, and MED1. Downregulated ADCYAP1 is one of the indicators that patients may benefit from acupuncture therapy. This gene regulates neuropeptide hormone activity (458) and is known to affect hypersensitivity and related to pain sensations (459, 460). LTA4H takes part in the synthesis of the proinflammatory mediator (461, 462) and its upregulation seems to be a good indicator of the therapy response. Though IL2RA also has proinflammatory properties (463), it seems mostly unaffected by acupuncture therapy, and patients that suffer chronic pain due to this reason may not be the best candidates for this treatment. MED1 contributes to pain and depression (464), and upregulated expression of this gene product indicates a more favorable therapeutic outcome.

McGill delta pain model works with four features based on gene products of PCNA, TNSF9, IFNG, and IDUA. PCNA is the cell proliferation and repair marker that can contribute to neuropathic pain resistance (465). TNSF9 encodes a ligand from the tumor necrosis factor family that can bind to the TNFRSF9, a receptor molecule in T lymphocytes (466). IFNg is an

inflammatory cytokine (467) and its production may be related to the increased stress level (468). IDUA is responsible for breaking down glycosaminoglycans – large sugar molecules, and its misexpression may affect metabolic functions (469).

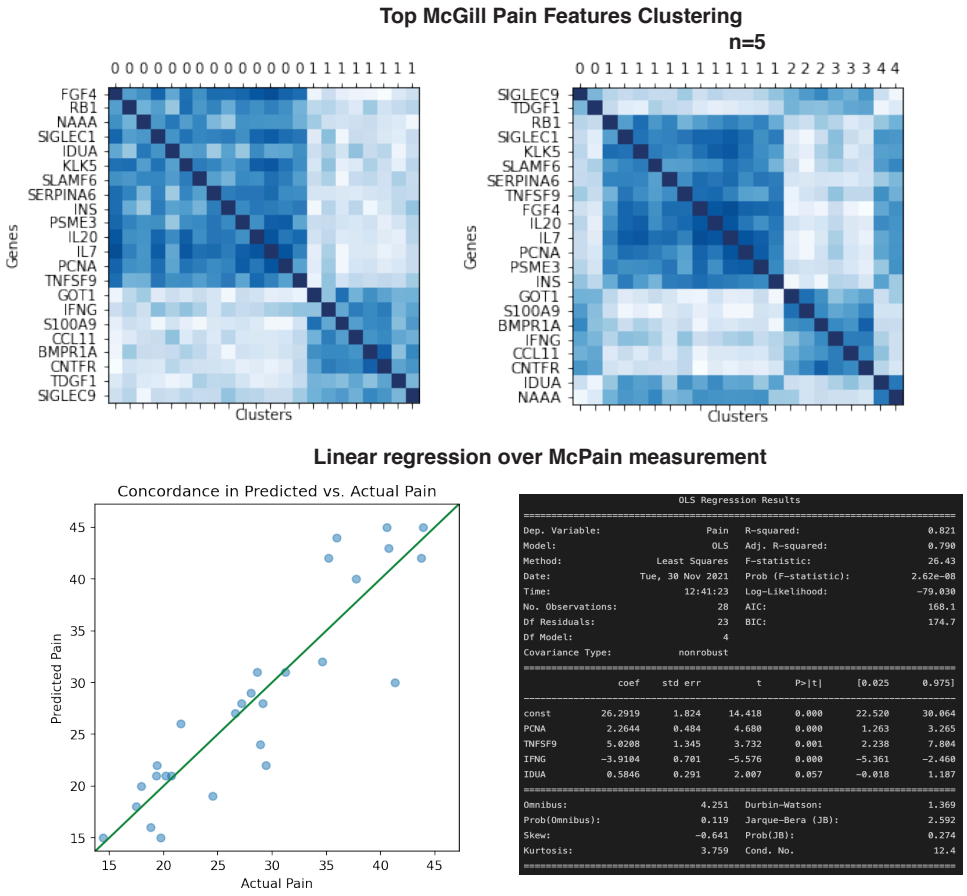

Figure 3.28. McGill scale delta pain response prediction. Top. Spectral co-clustering results for the top-21 features from the largest linear regression model into two (left) and eight (right) distinct groups. Bottom. Trimmed model based on the four factors. Concordance plot is on the left side, statistics on the OLS model performance is on the right.

3.7 Conclusions

Our computational approach, the ALT-IN Tool, aims to characterize functional effects of alternatively spliced isoforms by determining if they perturb a protein-protein interaction. In addition to those naturally occurring in different tissues, cells, and under different cellular conditions, a growing number of alternative splicing events have been associated with genetic

disorders, including cancer, neurodevelopmental, and heart diseases (67-69). Thus, understanding how differentially expressed alternative isoforms perturb PPIs may provide new molecular insights into the functioning of healthy as well as disease cells and tissues.

Given the data scarcity for the problem of AS-induced perturbation of PPIs, there are a number of advantages that random forest-driven supervised and semi-supervised learning methods provide. The important questions that can be affected by the data scarcity include efficient training of the machine learning model and its subsequent evaluation. Random forest algorithms have previously been successful in providing a rapid learning rate and robustness while being among the most accurate algorithms, especially on small data sets (133, 213, 470). In addition, previous work has argued that deep learning models may not be the best choice for small datasets (471, 472). Another viable option for the overcoming limitations imposed by the small sample size is the usage of transfer learning, especially when the learning task can be linked to the high-level entity that remains invariant, e.g., gene and protein sequences, images, texts. Our transfer learning approach for the isoform prediction problem shows promise but has to be further refined.

Developing a robust evaluation protocol for a small dataset is critical because of a risk of information leakage between the training and testing set, resulting in overoptimistic scores, even though obtained model lacks generalization capacity (164, 370-373). We can see very similar situation with the

When comparing our classifiers with the start-of-the-art sequence-based PPI prediction tools and currently available AS-based prediction methods, the accuracy of both supervised and semi-supervised ALT-IN Tool classifiers dominated all current methods, with the semi-supervised classifier being the best method overall. Several reasons may account for such

performance. First, the task considered in this work is different from a standard *ab initio* PPI prediction task. We formulate our problem by taking advantage of a known PPI and then characterizing the difference between the reference and alternative isoforms. Thus, if this information is properly used, the new AS-based problem could be easier to solve. Second, the inclusion of the knowledge-based statistical potential also improves the prediction accuracy because this potential is based on the previously obtained functional knowledge shared between the evolutionary and structurally related proteins and protein domains. Lastly, the usage of a large unlabeled dataset engineered through the integration of human interactome and spliceome data for the semi-supervised learning task also improves the prediction accuracy, albeit not substantially, which can be explained by the already high accuracy of the supervised approach.

While the alternative splicing events in the unsupervised dataset D2 do not constitute an exhaustive list of currently known AS events because it is currently limited to the Ensemble database (327), D2 provides sufficient variety in the data for the semi-supervised algorithm, as was demonstrated by our results. It is possible to improve the semi-supervised learning method further either by adding more AS events from other recent databases, such as VastDB (473) and including the information on the novel exons, or by improving the approach itself. Based on our analysis, we expect a moderate extension of the unlabeled dataset (a fraction of the current dataset) to provide minimal, if any, improvement in the accuracy. Improving the semi-supervised learning approach seems to be a more promising direction and will be one of the immediate future steps for this work.

Overall, with the accuracy, precision, and recall surpassing 90% when tested on the most stringent evaluation protocol, ALT-IN Tool becomes a great alternative to the experimental approaches and the only accurate computational approach for this task available to date. In

particular, ALT-IN Tool can be used for understanding system-wide alternative splicing-driven variation in molecular mechanisms implicated in complex diseases, which is highlighted in this work by applying the tool to Western Diet fed mouse transcriptomics data. Our analysis has revealed widespread PPI network perturbation in both liver and brain tissues that affect T2D-associated pathways and is centered around genes previously associated with the disease. We hope that our method could streamline the expensive and time-consuming high-throughput interactomics approach by first identifying a pool of candidate genes for the primer libraries and then pinpointing the isoforms of the utmost interest.

Alternative splicing impact factor provides a context frame for the ALT-IN Tool, allowing researchers to focus on particular genes and transcripts of interest instead of conducting an exhaustive search across the entire network. Further narrowing down scope of the search by selecting only stable isoforms allows us to increase confidence of findings and come up with more specific drug targets. Finally, merging this information with the high-throughput proteomics studies allows to closely follow molecular processes currently taking a place in the organism and hypothesize on the molecular basis of the disease, which would allow us to come up with actionable insights.

Chapter 4. Structural Modeling, Molecular Dynamics, and High-Performance Computing

4.1 Background

It has been two years since the release of the first SARS-CoV-2 genome (474), which provided scientists with critical knowledge about its proteins. Thanks to the unprecedented experimental efforts by scientists worldwide, we have now obtained structural knowledge about most SARS-CoV-2 proteins, determining their three-dimensional (3D) shapes. Perhaps even more critical is the structural knowledge of the protein complexes that underlie the basics of viral functioning. Months before the experimental protein structures were solved, computational efforts by several groups provided researchers with accurate 3D models of the viral proteins and their physical interactions with each other and with host proteins.

3D molecular information is instrumental in basic research, to understand mechanisms behind viral entry and replication, as well as in structure-based drug design, to determine new antiviral targets, or in vaccine development, to study the effects of novel mutations on antigen-antibody binding. Given that it is not ‘if’ but ‘when’ a new viral pandemic will emerge (475), it is crucial to know whether computational modeling methods can facilitate the structural characterization of viral proteins and their essential complexes. After one year of intensive research by the structural biology community, we have accumulated enough data to evaluate the impact of computational modeling efforts toward understanding the structural nature of the virus.

4.2 Structural Modeling of the SARS-CoV-2 Proteins

4.2.1 Introduction

The novel coronavirus, SARS-CoV-2, recently emerged in 2019 and has brought devastating effects to the global worldwide community, infecting more than 480 million people and has taken more than six million lives. However, the swiftly spreading virus also caused an unprecedentedly rapid response from the research community facing the unknown health challenge of potentially enormous proportions. Unfortunately, the experimental research to understand the molecular mechanisms behind the viral infection and to design a vaccine or antivirals is costly and takes months to develop. To expedite the advancement of our knowledge, we leveraged data about the related coronaviruses that is readily available in public databases and integrated these data into a single computational pipeline.

Within two and a half months since its initial discovery, the novel deadly coronavirus SARS-CoV-2 had infected more than 100,000 people, with the death toll already surpassing that of the 2003 severe acute respiratory syndrome coronavirus (SARS-CoV) and 2012 Middle East respiratory syndrome coronavirus (MERS-CoV) outbreaks combined (476-478). In spite of the instantaneous reaction by the scientific community and extensive worldwide efforts to address this health crisis, it took significant time for vaccines to finish the clinical trial and obtain emergency authorization (479, 480). If not for urgency, a regular timeline may take years. For instance, the Phase I trial for a vaccine that treats SARS was announced in December 2004, two years after the disease outbreak (481). Additionally, a vaccine against MERS, another infectious outbreak of the related coronavirus that emerged in 2012, was patented in 2019, with Phase I trials introduced in the same year (482, 483). Nevertheless, in the past two decades, a massive amount of work has been done to understand the molecular basis of the coronavirus evolution

and infection, develop an effective treatment in the forms of both vaccines and antiviral drugs, and propose efficient measures for viral detection and prevention (484-488). Structures of many individual proteins of SARS-CoV, MERS-CoV, and related coronaviruses, as well as their biological interactions with other viral and host proteins, have been explored along with the experimental testing of the antiviral properties of small-molecule inhibitors (489-494). This rich bank of knowledge allows us to rapidly unravel key point differences of the newly encountered virus via *in silico* modeling.

4.2.2 Protein Sequence Data Collection

Available sequences for protein candidates wS, wORF3a, wE, wM, wORF6, wORF7a, wORF7b, wORF8, wN, and wORF10 were extracted from the NCBI Virus repository (495) (collected on 29 January 2019) and then used in the sequence analysis and structural modeling. The Uniprot BLAST-based search was performed for each of the proteins using default parameters. From the results of each search, the final selection was made based on the pairwise sequence identity (>60%) as well as the evolutionary relationship (*Coronaviridae* family). Each of SARS-CoV-2 proteins was then aligned with the corresponding coronavirus proteins using a multiple sequence alignment method Clustal Omega (EMBL-EBI, Cambridge, UK) (496).

4.2.3 Template-Based Structural Characterization of Protein and Protein Complexes

The structure of each protein was determined using single-template comparative modeling protocols with the MODELLER software (UCSF, CA, USA) (497). First, the template for each protein sequence was identified using a PSI-BLAST search in the Protein Data Bank (PDB) (Research Collaboratory for Structural Bioinformatics) (498). In general, a structural template with the highest sequence identity was selected out of those that covered at least 50

residues of the target sequence with at least 30% sequence identity. The polyprotein wORF1ab was first split into 16 putative proteins based on its alignment with the human SARS-CoV polyprotein, with each protein then independently searched against PDB. In total, structural templates for 17 proteins were chosen (Fig. 4.1). In some cases, several independent templates, each covering an individual protein domain of a target SARS-CoV-2 protein, were selected. The obtained template was used in the comparative modeling protocol, generating five models. Each model was assessed using the DOPE statistical potential (499); the best-scoring model was selected as a final prediction.

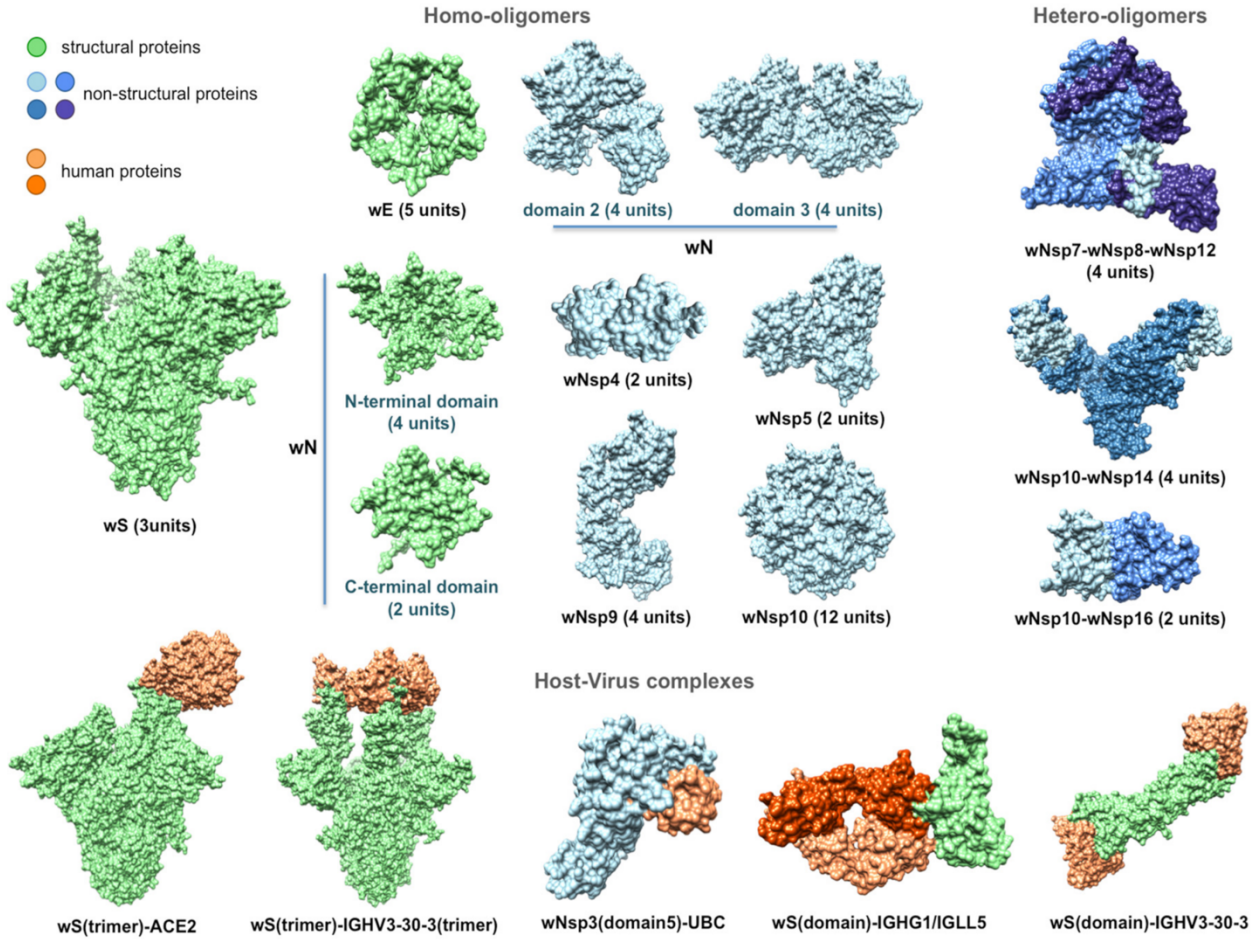

Figure 4.1. Structurally characterized intra-viral and host–viral protein–protein interaction complexes of SARS-CoV-2. Human proteins (colored in orange) are identified through their gene names. For each intra-viral structure, the number of subunits involved in the interaction is specified.

Using comparative modeling, we structurally characterized protein interaction complexes for both intra-viral interactions (homo- and hetero-oligomers) and host–viral interactions, where the host proteins were exclusively human. In total, we obtained structural models for 16 homo-oligomeric complexes, three hetero-oligomeric complexes, and eight human–virus interaction complexes including multiple conformations (Fig. 4.1). The intra-viral hetero-oligomeric complexes included exclusively the interactions between the non-structural proteins (wNsp7, wNsp8, wNsp10, wNsp12, wNsp14, and wNsp16). The modeled host–viral interaction complexes included three types of interactions: non-structural protein wNsp3 (papain-like protease, PLpro, domain) interacting with human ubiquitin-aldehyde, surface protein wS (in its trimeric form) interacting with human receptor angiotensin-converting enzyme 2 (ACE2) in different conformations, as well as the same protein wS interacting with several neutralizing antibodies. Based on the obtained models, the protein interaction binding sites were extracted and analyzed with respect to their evolutionary conservation.

4.2.4 De-novo modeling – (M)embrane protein

Unlike for proteins S and E, no experimental structures have been solved for SARS-CoV-2 M protein or any of its homologs, neither as a monomer nor as a dimer, which is its physiological conformation in the envelope. Thus, a comparative modeling approach cannot be applied, and a novel integrative approach was introduced that utilized geometric constraints of the M dimer in the envelope derived from the low-resolution CryoEM images of the envelopes of the closely related coronaviruses, SARS-CoV and MHV, as well as a high-resolution CryoEM structure of a dimer for another SARS-CoV-2 protein sharing substantial structural similarity (see Figs. B1, B2).

The approach included six steps. First, an ensemble of models of M monomer was obtained using *de novo* modeling methods AlphaFold and I-TASSER (500-502). The top 5 models from each method were then selected, and 200 models of M homodimers overall were obtained using two symmetric docking approaches, SymDock and Galaxy (503-505) (10 docking models for each of the 10 monomers for each docking approach). Next, a set of geometric constraints was applied to an ensemble of the 100 top-scoring homodimer models (50 for each symmetric docking). The geometric constrains include (1) the dimer axial dimensions and the shape of the part of the packaged dimer located on the envelope surface, (2) the orientation of the monomers in the membrane, and (3) the approximate dimensions of the transmembrane domain (TMD) of a single packaged M dimer defined by the envelope membrane's thickness. Specifically, from the previous analysis of SARS-CoV envelope (506, 507), it follows that TM domains of an average M dimer form a parallelogram, with rough dimensions between the two centers of adjacent dimer parallelograms measured to be 6.0 nm and 7.5 nm (Fig. 1). As a result, we filtered out those M dimer models whose TM domains would not fit into a parallelogram-shaped grid of these dimensions. We note that we did not require the TM domains to form a parallelogram-like shape, rather these constraints primarily affected the length of the modeled helices, filtering out models with significantly elongated one or two helices. Furthermore, the dimers were required to have N-termini of both monomers located on the exterior surface of the virion's envelope, and the thickness of the envelope's membrane was set to be equal to 4 nm (508, 509).

The analysis of the three best-scoring M dimers that satisfy all above geometric constraints provided us with an interesting finding: all three dimer models share striking structural similarity with the ORF3 protein of SARS-CoV-2 (Figs. 4.2, B1, B2), whose CryoEM

structure was recently solved (510). In particular, we found that the *de novo* modeled structure of the endodomain of M and the experimentally obtained structure of C-terminal domain of ORF3 were structurally similar, while the transmembrane domain structure of M and N-terminal domain of ORF3 shared the same secondary structure elements (four helices) of the same lengths while the mutual arrangement of the helices was somewhat different. Therefore, we hypothesized that M dimers and ORF3a were structural homologs, and the model of the M dimer could be further refined using the structural information from ORF3a dimeric structure (Fig. 4.2).

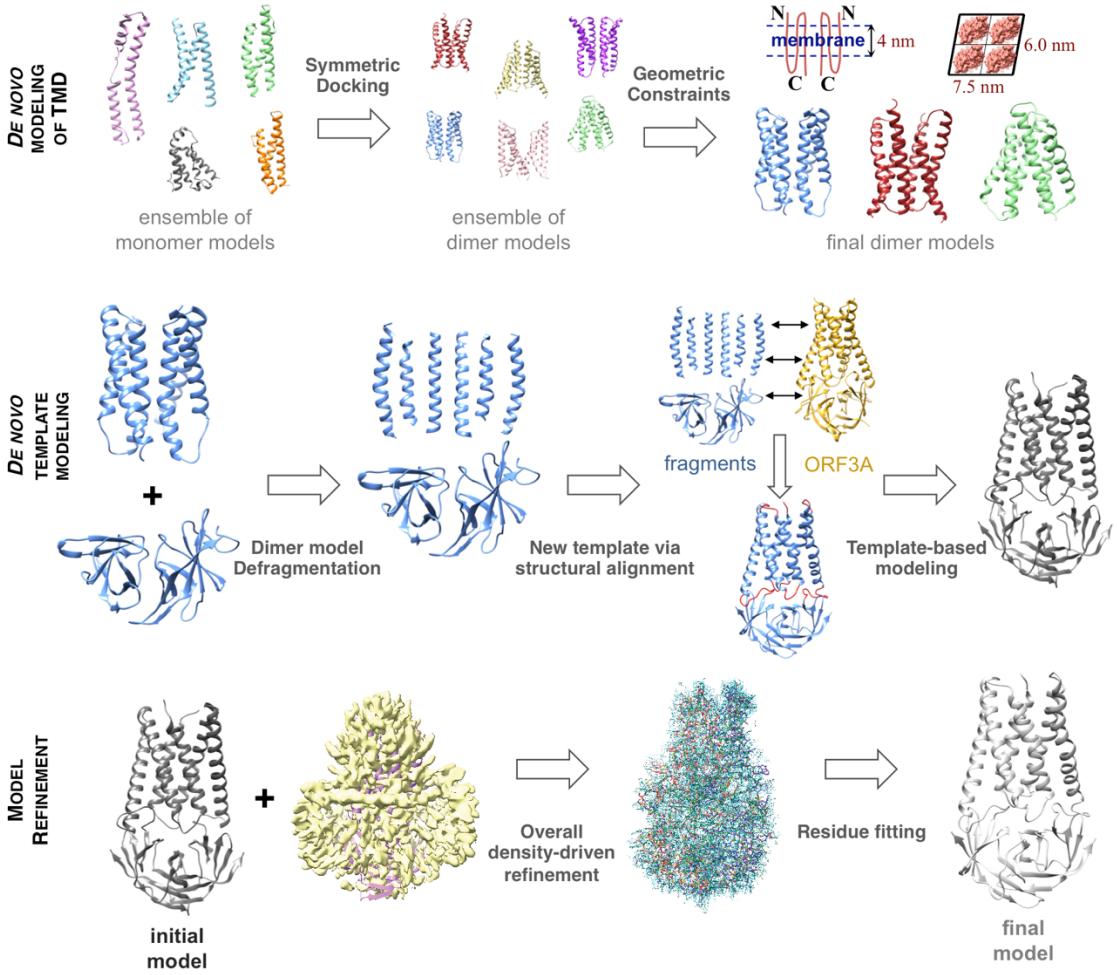

Figure 4.2. Basic stages of structural characterization of M protein's dimeric complex using integrative modeling.

To use the structural information from ORF3a, we first created a "fragmented' structural template by individually structurally aligning the four helices and endodomain of M dimer with the corresponding helices and endodomain in the ORF3a template structure (PDB ID: 6XDC). Then, we obtained a preliminary comparative model using the newly created fragmented structural template of M dimer. The obtained M dimer model was refined using a protocol similar to the one used to obtain a full-length model S trimer (Fig. B2) (511). First, the linker regions of the two TMDs in the obtained comparative model of M dimer were refined by energy minimization, followed by refinement of the whole TMDs using Molecular Modeling Toolkit (512). Next, the overall M dimer structure was minimized using the CHARMM36 force field in GROMACS (513). We then placed the M dimer model into the experimental EM density map of the ORF3a dimer (EMD-22139 (510)) using Phenix (514) and relied on the EM map to further refine the structure in ISOLDE (515), a package for UCSF Chimera X molecular visualization program (516). ISOLDE uses OpenMM-based interactive molecular dynamics flexible fitting (517) using AMBER force field (518) and allows for the real-time assessment and validation of the geometric clashing problems. Each residue of the M dimer model (1-946) was then inspected and remodeled to maximize its fit into the density map. We considered both a deposited electron-microscopy map and a smoothed version with a B-factor of 100 $Å^2$ as proposed in (511), but in our case there was no significant difference between these two versions of the refinement protocol.

4.2.5 Computational Protein Modeling – Race Against Time

Structural genomics efforts to characterize the protein repertoire of a virus are usually carried out by comparative—or template-based—modeling (519). A newer technique described

in the previous section, de novo protein modeling (520), does not require a template structure and may complement existing methods. However, it takes significant efforts to validated proposed finding. Template-based models are often more accurate than de novo ones; however, the former technique is dependent on previously solved structures of homologous proteins or protein complexes while the latter can be applied to novel proteins. The latest success in protein modeling has been primarily due to recent technological innovations in the developme nt of novel protein structure prediction algorithms, which use deep learning and are empowered by advances in graphical processing unit (GPU)-accelerated computing.

We surveyed accurate template-based and de novo models of SARS-COV-2 proteins and protein complexes that were also experimentally solved to determine (i) model accuracy when compared with the experimental structure and (ii) how far ahead of the experimental structures they were obtained (Fig. 4.3). We considered comparative models generated by our group (521) and de novo models reported by AlphaFold (149) and C-I-TASSER (522), which have also contributed to structural characterization of SARS-COV-2 proteins (Fig. 4.3 and Table B1). Of the 29 putative proteins, 17 were at least partially experimentally and computationally resolved, while 5, including key structural protein M, were characterized only computationally. Six putative proteins have not been structurally characterized at all. The computational methods were fairly accurate, producing an average root mean squared deviation (r.m.s.d.) error of 4.1 Å for all 17 proteins (Table B1). On average, computational models covered roughly 80% of the viral protein sequence, while experimental structures covered 82%. Most importantly, 3D models of viral proteins were released on average 86 days earlier than the corresponding experimental structures.

Even if we had structural knowledge of all SARS-COV-2 proteins, our understanding of the virus's functional units would be far from complete: most, if not all, viral proteins carry out their functions by forming macromolecular complexes. Recent efforts to map all protein complexes formed by SARS-CoV-2 proteins have identified hundreds of putative interactions[8]. Unfortunately, only a small fraction of these complexes have been structurally characterized (Fig. 4.3B and Table B2 in Appendix): 18 protein complexes have been characterized experimentally and 16 computationally. Overall, for 13 protein complexes, the structure was both modeled and resolved experimentally. For 5 of these, an incorrect oligomer conformation was derived from homologous complexes; for the remaining 8, the computational models yielded accurate protein complexes in correct conformations, with an average r.m.s.d. of 2.6 Å over the entire multimeric structure (Table B1). The models were available on average 53 days earlier than experimental structures, covering on average 77% of all protein sequences involved in the complex. Lastly, for 4 modeled complexes, no experimental structures have yet been obtained.

Comparison between experimental and computational structures

The experimental structures were obtained from RCSB data bank using BLAST search of each SARS-CoV-2 protein. Overall, 19 PDB structures of the individual proteins and 116 PDB structure of the protein complexes were retrieved and analyzed. Some of the structures were independently solved by multiple groups and in the presence of different mutations or substrates. In such case, only the structure with the earliest release was kept.

Three measures were computed when comparing the computational model of a SARS-CoV-2 protein and its experimentally resolved structure: sequence coverage (of the model and the experimental structure), model accuracy, and the difference between the release dates. Sequence coverage was computed as a fraction C/S, where C is the number of amino acid

residues covered by either a model or experimental structure, and S is the total length of residues sequenced for the corresponding SARS-CoV-2 protein. The model's accuracy is calculated as all-pair root-mean-square deviation (RMSD) defined as:

$$RMSD(M, M') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2)},$$

where (x_i, y_i, z_i) are coordinates of the i-th atom of the model M and (x'_i, y'_i, z'_i) are the coordinates of the corresponding atom from the experimental structure M'; the one-to-one atomic correspondence was obtained through structural superposition of the computational model and experimental structure using least-squares fitting. The difference, $\Delta T_{\text{Release}}$, is defined between the dates of the public release of the corresponding experimental and computational structures.

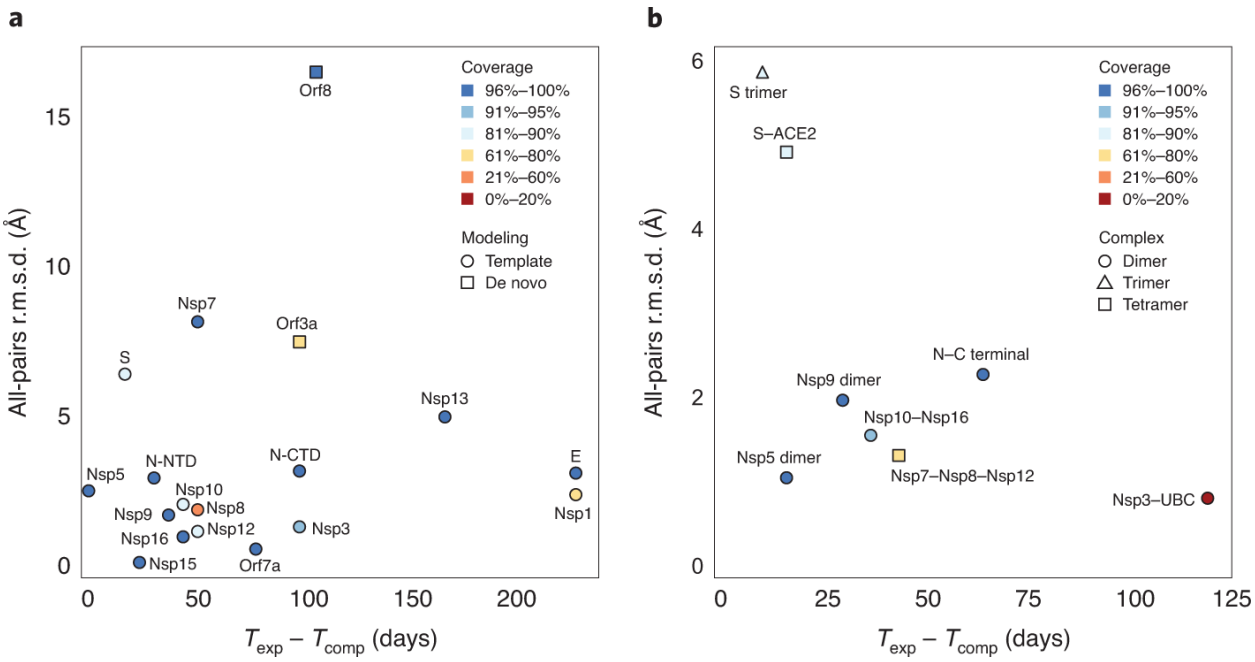

Figure 4.3. Accuracy vs timeline of appearance of protein structures. a. Analysis of 17 individual proteins that were both experimentally characterized and computationally modeled, using comparative (circles) and de novo (squares) methods. **b.** Analysis of 8 protein complexes; each complex consists of two (circle), three (triangle) or four (square) protein subunits. For each modeled protein or protein complex, its r.m.s.d. error between the model and experimental structure, the number of days between the releases of experimental and computational structures, and the model's coverage of the protein sequence (color) are calculated.

Protein complex models.

Similar measures were calculated for the protein complexes. Sequence coverage of a protein complex was computed as a fraction $\frac{\sum_{i=1}^{n} C_i}{\sum_{j=1}^{n} S_j}$ where C_i is a number of amino acid residues covered by the model of each protein i that is a part of the protein complex and S_i is a total length of residues sequence for corresponding $experimental$ structure in protein complex. To calculate the modeling error of a protein complex, each model was superposed against the corresponding experimental structure using the least-squares fitting, and all-pair root-mean-square deviation (RMSD) was calculated using all atoms in the model. Correspondence between the individual protein chains from the experimental and computational structures was manually curated to ensure a superposition between 3D structures that produces the lowest RMSD score. The difference, $\Delta T_{\text{Release}}$, was defined in the same way as for the individual protein models.

Structural characterization of individual SARS-CoV-2 proteins

Five experimentally uncharacterized proteins include one of the main structural proteins, M, and four non-structural proteins, Nsp2, Nsp4, Nsp6, and Nsp14. Furthermore, one protein, ORF9b, was characterized only experimentally, and 6 putative proteins, Nsp11, ORF3b, ORF6, ORF7b, ORF9c, and ORF10, have not been structurally characterized at all, either computationally or experimentally. For the 15 models that were obtained exclusively with comparative modeling the RMSD was substantially smaller, 3.1 Å.

We considered protein complexes of two types: either homo-oligomers (*e.g.* a functional form of a key protein spike, S, is a homo-trimer) or hetero-oligomers (*e.g.* a virus-host protein interaction complex between S and human ACE2 receptor). Overall, 18 protein complexes were computationally characterized, including oligomeric structures of 4 individual domains of

nucleocapsid (N) protein and multiple interactions between S and the host antibodies. The modeled complexes included 11 homooligomer and 7 heterooligomer structures. The number of computationally resolved protein complexes was smaller due to the lack of homologous complexes previously solved for the related viruses. The incorrect oligomer confirmation, or binding mode, for 7 protein complexes, (Nsp3, Nsp10, Nsp13, N-Nterminal, E and antibodies), was predicted because of the difference between the native conformations of the homologous complexes that served as templates and the corresponding complexes involving SARS-CoV-2 proteins. It is worth noting that experimental structure that can be considered Nsp10 dimerization is a part of Nsp10-Nsp16 complex and is not a dedicated Nsp10 complex study, which was not conducted up to date.

4.3 Molecular Dynamics of the SARS-CoV-2 Envelope and High-Performance Computing

4.3.1 Introduction

The year 2020 brought forth the largest pandemic of the past century, caused by the SARS-CoV-2 virus. Scientific community responded with unprecedented collaborative efforts, unraveling genomic, structural and interactome information about this virus with the goal of finding a cure.

Currently, even though we have a range of highly efficient vaccines that successfully combat the spread of the virus, it is unlikely that COVID would follow footsteps of smallpox - the only eradicated infectious disease. This means that our fight is not over, we already see a wide range of variants arising in different locations on Earth; we have to find means to curb the disease in people who already got sick, akin to therapeutic drugs for the influenza.

For this purpose, besides well-studied (S)pike protein, main building block of the viral shell - (M)embrane protein – is another lucrative target, though working with it presents unique challenges. Even though a large part of its structural proteins was experimentally resolved, we have a limited understanding on protein complexes (up to date there is no experimental M dimer structure), and even less is known about membrane surface, which consists from lipids, several (~20) spikes, a few (2-3) (E)nvelope proteins and a large amount (~1000) of M proteins that constantly interact with each other.

Because of this targeting M protein without knowledge of dynamic processes surrounding it is an error-prone approach, and designing efficient antiviral drugs require better understanding of the SARS-CoV-2 surface.

We present a physically tractable mesoscale system of viral envelope that amalgamates the most recent information on viral envelope (protein structures, stoichiometry, and geometry) and corresponding molecular dynamics simulations of its behavior in solvent.

4.3.2 Molecular Dynamics of the M-dimer

To enable coarse-grained (CG) simulations of the entire virion envelope, high resolution structures of the constituting S, E, and M proteins are required. The CG representations of the homooligomers of the structural proteins were obtained from available atomic models (S and E) and from *de novo* modeling (M), which were coarse-grained and then refined in the presence of a lipid bilayer using the Martini 3 force field (Methods). The initial model of the full-length S trimer was obtained previously using an integrative modeling approach, and a model of the E pentamer was obtained previously using homology modeling. In contrast to the S and E proteins, a structure of M or its homodimer supported by experimental observations or evolutionary inference did not exist. Therefore, we first modeled the structure of M dimer using an integrative

approach (Supp. Figs. S3-S6). The procedure started with the *de novo* modeling of the monomeric structure of M, followed by constraint symmetric docking to create a homodimer that satisfies the geometric constraints obtained based on 1) the envelope's membrane thickness, 2) mutual orientation of the monomers, and 3) the approximate local geometric boundaries of a single M dimer complex, previously obtained from microscopy data of the SARS-CoV envelope. However, preliminary CG MD simulation of the envelope using the obtained top-scoring *de novo* model of the M dimer revealed the structural instability of the dimeric complex prompting us to the further refinement of the model.

We found that the *de novo* homodimer models of M that satisfy all the above constraints appeared to share striking structural similarity with another recently resolved homodimer of SARS-CoV-2, the ORF3a protein. The similarity included 1) the same two-domain fold composition and 2) the same combination of secondary structure elements as in our *de novo* model, but with a slightly different arrangement of the secondary structure elements in the transmembrane domain (Supp. Figs. S4, S5). We thus further refined the M dimer model by constructing a new structural template as a scaffold of the same secondary structure elements as in the original *de novo* model, each of which was structurally superposed against the ORF3a dimer. We then applied a novel integrative template-modeling protocol using the newly designed template and followed by a refinement protocol guided by the electron density of ORF3a (Methods, Supp. Fig. S6). The rationale for this approach was that the newly designed template, based on the ORF3A secondary structure topology and including the original secondary structure fragments of M dimer extracted from the top-scoring *de novo* model, would improve the arrangement of the secondary structure elements, making the model more stable, while maintaining the structural similarity with the original *de novo* model. The resulting model not

only provided a tighter, more stable packing of M monomers in the dimer; but the shape complementarity of M dimers with each other allowed for a natural tiling of multiple dimers into the "filament" structures, supporting the previously proposed model of M dimer lattices based on the microscopy study of SARS-CoV envelope (507, 523). Importantly, the *de novo* M dimer model proved stable in subsequent simulations of the full envelope. Furthermore, a 200 ns all-atom simulation of the two TM domains embedded in a lipid bilayer also resulted in a stable complex, while a 200ns simulation of the TM dimer of the top-scoring *de novo* model appeared to be unstable.

In the 4μs simulations, we consistently observed changes in the viral shape (Fig. 2D, Supp. Figs. S9, S10), with the initial diameters of the ellipsoid of a model in composition C1 changing from $d_1 \approx 109.9$nm, $d_2 \approx 97.8$nm, and $d_3 \approx 76.2$ at t = 0μs to $d_1 \approx 103.1$ nm, $d_2 \approx 97.8$ nm, and $d_3 \approx 81.3$ nm at the end of the simulation, t = 4μs. The obtained diameters were close to the range observed for the particles from the Cryo-ET images of SARS-CoV-2 (509). We also note, that while the experimentally observed structures of the virions had variable shapes, from a nearly spherical shape to a significantly elongated ellipsoid, the average shape of SARS-CoV-2 according to the Cryo-ET study is an elongated ellipsoid, not a sphere, hence the rationale for our initial model dimensions. The elongated shape of a virion particle was also reported in previous studies of SARS-CoV and the related betacoronaviruses (507). The calculated d_{MAX}/d_{MIN} ratio of 1.27 for our final model falls within the range of average ratios observed in CryoET of the SARS-CoV-2 virion (509). Interestingly, the changes in the shape did not have significant effects on the surface area of the envelope (1.3% reduction) or its volume (0.6%) (Supp. Figs. S9, S10). Along with the diameters of the envelope shape, the principal radii of gyration were also converging, reflecting shape stabilization (Fig. 2D). Furthermore, analysis of the temporal

changes of the viral dimensions together with the connectivity patterns of the envelope proteins suggested the presence of two distinct concurrent relaxation processes, separately affecting the two smallest and the two largest diameters. Specifically, we observed a faster process (0−1μs), followed by a slower process (0.5−4μs). During the faster process, the minor circumference (principal radii r_2 and r_3 corresponding to diameters d_2 and d_3) became more circular while during the slower process, the major circumference (principal radii r_1 and r_2 corresponding to diameters d_1 and d_2) also became more circular, thus making the minor circumference to become more elliptical again (Fig. 2D).

4.3.3 SARS-CoV-2 Envelope Construction

The modeling of the entire envelope started with the generation of a mesoscale model using dynamic triangulated surface (DTS) simulation (524, 525) on a triangulated mesh, matching the dimensions of the virion envelope. The mesh included a set of vectors each representing one protein and its orientation in the envelope surface. To set up the initial positions of the structural proteins in the envelope structures, available EM data was used only to obtain information on local geometry (Methods, Supp. Fig. S1). The global geometric patterns observed in the EM studies were not used during modeling, but only to evaluate our model (*vide infra*).

The DTS simulation provided us with an initial guess of the protein organization and orientation on the fixed geometry of the envelope. This model was subsequently backmapped using TS2CG to near-atomic resolution (526), based on the CG Martini 3 models of the proteins and lipids (527) with specified stoichiometries. This resulted in an initial arrangement of the oligomeric protein structures embedded in a lipid bilayer comprising up to six types of lipid molecules (palmitoyl-oleoyl-phosphatidylcholine, POPC; PO-phosphatidylethanloamine POPE; PO-phosphatidylserine, POPS; PO-phosphoinositol, POPI; cardiolipin, CDL2; and cholesterol,

CHL), with a composition reflecting that of the endoplasmic reticulum (ER), but also considering enrichment of specific lipids due to interactions with the proteins. Overall, lipid-to-protein ratio was a crucial factor in constructing a stable model. When number of lipids was insufficient, proteins tended to agglomerate together, leaving no space for lipid molecules in between. After pressure was applied to such model it tended to loose stability and get punctured.

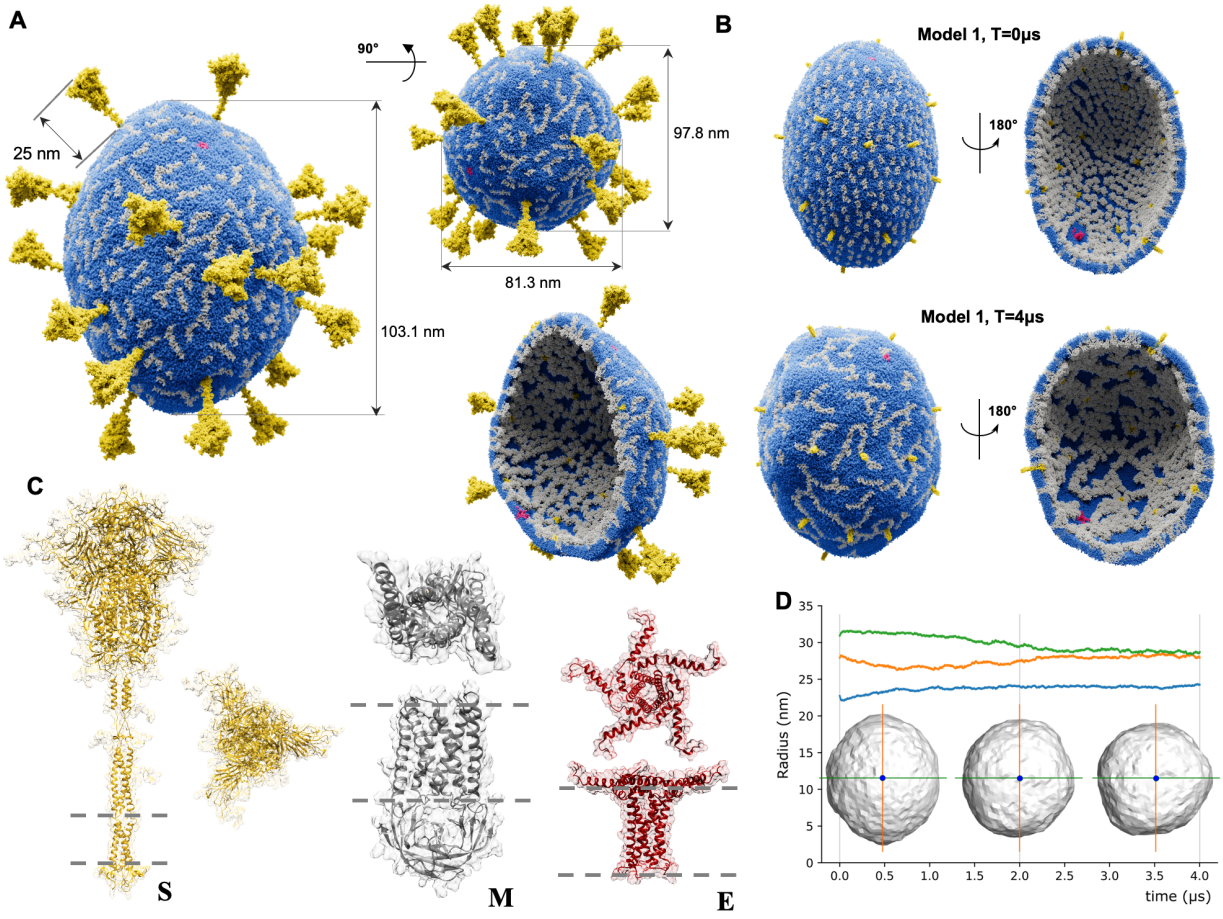

Figure 4.4. Structural characterization of SARS-CoV-2 viral envelope and its components. **A.** An envelope model (M2) obtained from molecular composition C1 (2 E pentamers, 25 S trimers, 1003 M dimers) and including full-length structures of S trimers after 1μs simulation run. Lipid molecules are depicted in sapphire blue, E pentamers in ruby red, M dimers in silver, and S trimers in gold. Principal diameters have values of 81.3 nm, 97.8 nm, and 103.1 nm. Height of the outer part of S protein is 25 nm. Surface of the envelope displays "filament" patterns formed by transmembrane domains of M dimers, while the internal part of the envelope shows tight packing of M dimers' endodomains assemblies; **B**. Envelope model M1 from molecular composition C1 using truncated S trimer structures at the start of the simulations (top) and after 4μs (bottom); **C**. Structural proteins S, M, and E representing the main structural building blocks of SARS-CoV-2 envelop in their physiological oligomeric states, in side and top views: S trimer, M dimer, and E pentamer. The grey dashed lines correspond to the membrane boundaries. The structures are shown in different scales. **D**. Change of the viral shape during the

simulation defined through the principal gyration radii. The two largest principal radii converge to the value ~28 nm while the third one converges to ~24 nm. The actual diameters of the model after 4μs simulation were 103.1 nm, 97.8 nm, and 81.3 nm, respectively.

Several envelope models were thus built and subjected to MD simulations to test the stability of the envelope structure. The overall models prepared for the simulation consisted of 20-30 million CG particles, representing about 100 million heavy atoms. The models varied in several key parameters: (i) different protein-to-lipid ratios, (ii) different stoichiometries of the structural proteins, (iii) full or partial ectodomains of the spike trimer included in the envelope structure, and (iv) different lipid compositions. In total, three independent simulations turned out to be stable. For the unstable models, the integrity of the envelope surface became compromised (an example of unstable structure simulation is shown in Suppl. Movie S2). The selected stable models (M1-3), simulated for 1-4μs, included ~1.0-1.2M protein particles, ~0.5-0.8M lipid particles, and ~13M-35M solvent particles (Fig. 4.4A,B, Table 3,B3 in Appendix). A 4μs simulation took ~1,560,000 CPU hours to compute on the TACC Frontera supercomputer.

We found that each of the key parameters played a role in the simulation. First, when selecting between two different protein-to-lipid particle ratios, the higher ratio value of 2.36 resulted in unstable structures, while the ratio of 1.44 resulted in a structural model that remained stable (models M1, M2). Given that the molecular composition was the same in both models (1,003 M dimers, 25 S trimers, and 2 E pentamers; we refer to this molecular composition as C1), the different ratios were due to the different lipid numbers (36,645 and 60,141 molecules, respectively), suggesting that the lipid concentration plays a role in the envelope stability, a finding supported by recent CG simulations of cell-scale envelopes (528). Another factor that affected the model stability was a higher number of solvent particles, compared to the stable models, leading to pore formation and subsequent membrane rupture.

Second, we found that varying the stoichiometries of E pentamer, S trimer, and M dimer under the same conditions does not affect the envelope stability. For instance, when we significantly increased the proportion of S trimers (truncated form) creating a model in an oligomeric composition C2 that included 3E pentamers, 71 S trimer, and 1080 M dimers (Supp. Table S3), while maintaining the same protein-to-lipid ratio and the number of solvent particles as in models M1 and M2, we found that the new model, M3, was also stable after 4μs simulation (Supp. Fig. S8). The stability of viral envelope with different stoichiometries is in line with the experimental evidence suggesting a range of different stoichiometries to be found *in vivo*. The behavior of the envelope model M2 that included the full-length S trimers and composition C1 was similar to the ones of the truncated model M1 (Fig. 4.4A,B): a 1μs trajectory of the former was comparable to the first 1μs of the 4μs trajectory for the latter. Finally, variations in lipid composition did not appear to impact the stability of the envelope.

4.3.4 Molecular Dynamics Simulation of the Envelope

Our envelope models and the time scales of the simulations allowed a detailed assessment of the interactions between the different constituents, in particular those involving M dimers. To characterize these interactions, we focused on the preferential relative orientations of protein neighbors and second neighbors (Fig. 4.5A-C), which were determined using a method for orientation analysis (529). The results showed that the M dimer transmembrane domains (TMDs) preferentially formed filament-like assemblies, without contacts between adjacent filaments (Figs. 4.5A,B). In contrast, the endodomains (EDs) appear tightly packed, binding neighbors in two directions (Figs. 4.5A,B), thus showing a tendency for the formation of a well-ordered lattice. Intriguingly, the lattice vectors of the TMD domains extracted from M dimers in our model were identical to the ones extracted from the previously reported EM data (507).

Combining the (averaged) relative orientations of the TMDs and the EDs with the projected densities of the proteins revealed the characteristic patterns of densities. Specifically, the averaged orientations of TMDs fit almost exactly with the 'lattice' patterns previously observed from the averaged Cryo-EM virion structures of SARS-CoV and related betacoronaviruses (Fig. 3A,B) (507), despite the fact that this information was not used during construction of the models. Even the characteristic lack of density in the unit cells' corners previously observed in MHV and SARS-CoV was clearly noticeable in our data (Fig. 4.5B). These patterns were consistent in the models with both compositions, C1 and C2. The orientations of EDs revealed a different kind of pattern, compared to the one of TMs. Specifically, our model showed that the ED dimers formed triangulated structures, a pattern common in engineering rigid frame structures (530). Unlike TMs, the formation of EDs have not been experimentally characterized by microscopy before, because it can only be observed from the virion's interior. To ensure that the observed property of M dimers forming the filament-like assemblies was not a consequence of the initial setup, we have performed additional simulations of a subset of randomly placed and oriented 41 M dimers in a flat bilayer system with the same lipid composition as the full envelope simulations (Methods). After a 13μs simulation, we observed that the M dimers formed filaments that were reminiscent of the ones we observed in the full envelope simulations (Fig. 4.5A). In contrast to the strong preferential orientation of M dimers, the orientation of M dimers around S trimers did not show clear preference in attachment (Fig. 4.5C), which could be due to the stronger orienting effects of M dimer interactions as well as to the worse statistics for the interactions between S and M.

To further characterize formation of the higher-order assemblies in the envelope through interactions between M dimers, a temporal analysis was performed of the domain-level physical

interaction networks between the TMDs and EDs of M dimers and the transmembrane domains of S trimers. This analysis further illustrated the strikingly different nature of the M dimer's key domains (Fig. 4.5D-G), supporting our previous findings. Throughout the simulations, the domain interaction networks appeared to undergo drastic rearrangement via two distinct phases, with the number of connected components of three and more dimers first rapidly rising during 0-200ns up to ~160 components and then slowly saturating to ~30 components (Fig. 4.5D). The total number of connected components closely follows a bi-exponential law, suggesting two processes running concurrently: a faster local rearrangement and a slower filament assembly. The two-process formation of the connected components was also evident from the clustering analysis (Fig. 4.5F), indicating the initial formation of many small clusters, followed by the preferential growth of the largest connected components. It was also interesting to see that the second, slower process of growing connected components started before the first, faster stage of forming initial small assemblies of M dimers was over. The analysis of the interaction network of TMDs also supported the formation of filaments observed on the surface of the envelope (Fig. 4.5E); the network revealed that these filaments were occasionally connected even further into larger assemblies. In contrast to the filament-like network topology, the EDs network consisted of connected triangulated components, which may contribute to the structural rigidity of the envelope. The difference between the TMD and ED network topologies also followed from the temporal node degree distributions for each network: the average node degrees converge to ~2 in TMD network and ~3 in ED network (Fig. 4.5G). Lastly, all other major network parameters appeared to converge to the stable values, thus further drastic changes in network topologies were not observed.

4.3.5 Network Analysis of Macromolecular Spatial Organization

To get further insights into the dynamic reordering of structural proteins that took place in the envelope, we conducted the connectivity analysis, constructing domain-level protein-protein interaction networks. This approach was a scaled-up version of the protein structure network (PSN) analysis, which was previously employed in the structural characterization of individual proteins. Here, we defined a network node not as an individual amino acid residue, but as an entire protein domain. In this analysis, we differentiated only between the proteins' transmembrane domains (TMD) and endodomains (ED), which resulted in two separate physical contact networks: one within the lipid membrane and one on the inside of the virion.

We processed the reduced trajectory files (see Section 6 in Materials and Methods) that preserve the center of mass and orientation for each protein, substituting each entry with the canonical pdb model used in this study and conducting contact analysis with a cutoff distance of 0.6nm, a frequently used threshold for the coarse-grained structures (531-533). This procedure is performed separately on TMDs and EDs of the structural proteins. As a result, two sets of temporal dynamic networks for each of the truncated spike models, M1 and M3, were obtained. Each temporal dynamic network was a series of domain-domain interaction network snapshots, taken every 1ns, from 0μs to 4μs. Several key network properties were calculated for each network snapshot: average degree, number of connected components, average diameter, transitivity, and average degree connectivity for nodes with degrees 1, 2, and 3.

The degree d_i of a node i is the number of edges incident to it. The average degree in a network of N nodes is calculated according to the following formula: $d_A = \frac{1}{N}\sum_{i=1}^{N} d_i$. The number of connected components is the minimal number of subgraphs that have a path for any pair of vertices. The diameter of a graph G is the maximum length of the shortest paths for each

pair of vertices (534): $\max_{i,j \in G} \delta(i,j)$, where $\delta(i,j)$ is a shortest path between vertices i and j. Because our network was composed of multiple connected components, we calculated the diameter for each connected subgraph, and defined an average of the diameters: $D_A = \frac{1}{n}\sum_{s \in S} \max_{i,j \in s} \delta(i,j)$, where S is a set of connected components and n is its cardinality. The transitivity (also called the clustering coefficient) is the relative number of triangles ($\#triangles$) present in a given network compared to the number of all possible triangles ($\#triads$): $T = 3\ (\#triangles)/(\#triads)$ (535). Finally, the average k-degree connectivity is the average degree of the nearest neighbors for the nodes with the degree k. In this study, we computed the average degree connectivity for k=1, 2, and 3.

The number of connected components over time was fit with exponential and biexponential models:

$$Exp(x) = c_2 e^{c_1 x} + c_0$$

$$BiExp(x) = c_4 e^{c_3 x} + c_2 e^{c_1 x} + c_0$$

The resulting models were compared using Akaike's information criterion (AIC) and Bayesian information criterion (BIC). Results were plotted using Python 3.9 and Bokeh library. A Sankey diagram of connected components was plotted in MATLAB using visualization module of PisCES algorithm.

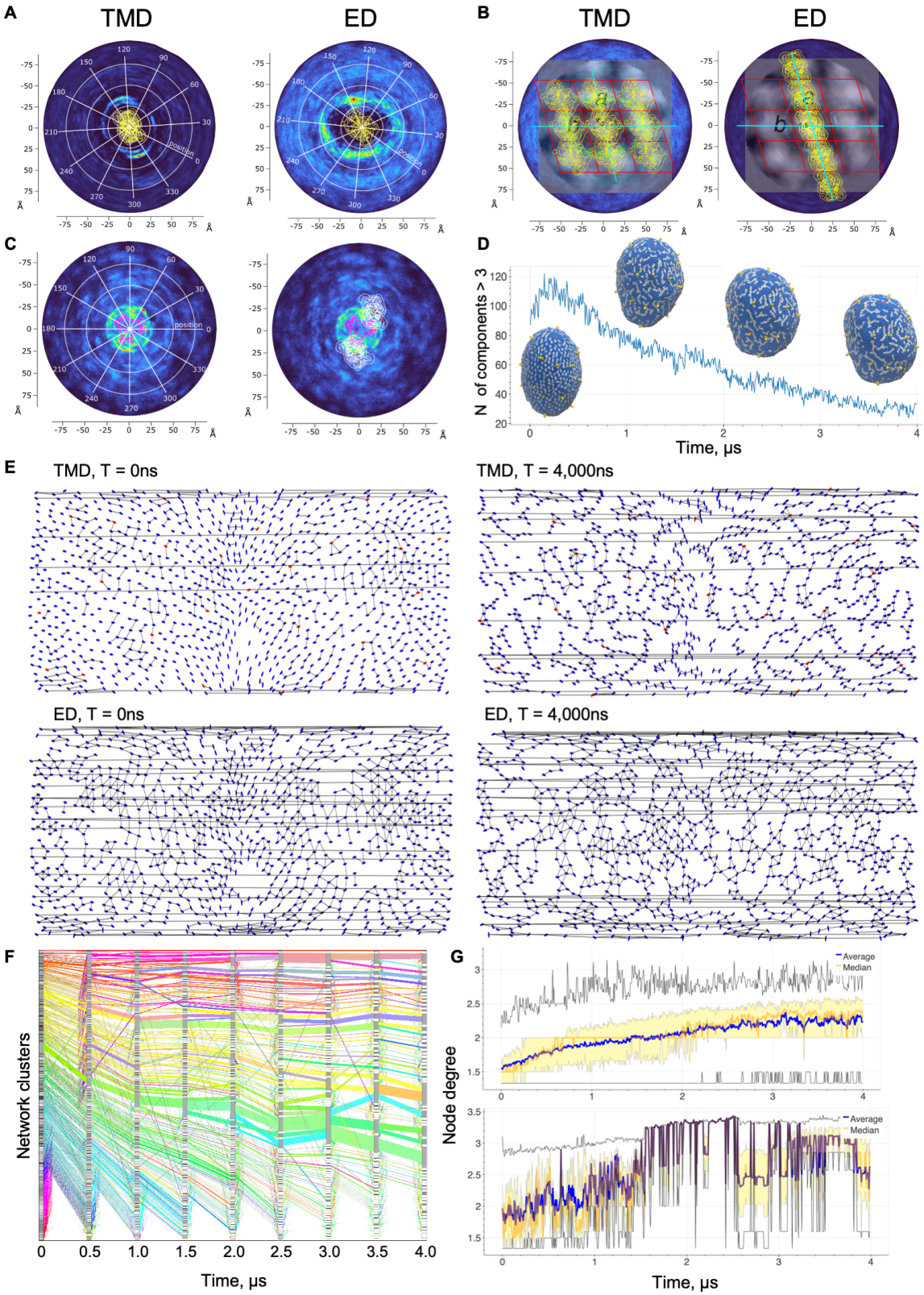

Figure 4.5 Structural and network analysis of the envelope assembly. A. Orientation preference of the transmembrane domains (TMD) and endodomains (ED), two main components of the M proteins in their dimeric state. TMD components display two preferred locations at 135° and 315°, ED

components display four preferable locations at 135°, 315°, 225° and 75°; **B**. Super position of averaged Cryo-EM images previously obtained from SARS-CoV envelope with M dimer models obtained separately for TMDs and EDs and arranged according to the preferred interaction positions from panel A, demonstrating a near-perfect correspondence between our model and the Cryo-EM images. **C**. Orientation analysis of the contacts between S trimers and M dimers. Shown are positions of the S trimer (pink), the density of positions of M dimers (left panel) and the same density together with two M dimers (white) positioned around two distinct high-density locations at 60° and 260° (right panel). **D**. The number of connected components that have at least three nodes dynamically change during the simulation; included into the same plot are envelope models at t=0μs, 1.5μs, 2.5μs, and 4μs in molecular composition C1. **E**. Mercator projection of physical domain-domain interaction network established for molecular interactions of M dimers (blue ellipses) with each other and with S trimer (red ellipses) for the models in molecular composition C1. M dimer's TMDs and EDs are arranged into separate domain-domain interaction networks. The orientation of ellipses corresponds to the orientation of the corresponding oligomers. An ellipse with the major axis positioned horizontally corresponding to the canonical orientations of the oligomers, as defined in panels A and C, for M dimer and S trimer, respectively. Connectivity increases after 4μs of simulations for both TMD and ED components; **F**. Sankey diagram showing clustering dynamics of the TMD components over the time for the model in molecular composition C1. Each of the displayed "flows" contains at least 25 network nodes. One can see a drastic increase in the cluster size for a small number of connected components as the simulation progresses; **G**. Average node degree dynamics during the simulation for TMDs (top) and EDs (bottom). Yellow bands indicate values for the 2nd and 3rd quantiles, black lines denote the minimum and maximum values. There is a clear trend for the increase of the node degree value for TMD components; the value plateaus during the last microsecond (3-4μs) around the value of 2.3. EDs' node degrees have a much smaller spread and tend to converge to the value of 3.2.

Solvent molecules were excluded from the visualization. To capture the whole dynamic simulation process, one frame per one nanosecond was extracted from the original trajectory using *gmx trjconv* (536); the extracted XTC file containing the frames was then loaded into VMD. Lastly, a *tcl* script was used to visualize each frame, followed by rendering the frames into TGA images using tachyon.

Visualization of the simulation trajectory of M1 was made by rendering each frame in the 4μs trajectory of the envelope model M1 and rotating 2,000 steps around y-axis clockwise and 2,000 steps counterclockwise; each step was 0.9 degrees. Another visualization of the simulation trajectory of M1 (Suppl. Movie S4 was made with imagemagick montage by showing two identical 4μs trajectories, first one revealing the outside and the second one revealing the inside of the envelope's structural model M1. For Visualization of 1μs trajectory of model M2 (Suppl. Movie S5), we used FFmpeg to convert images into a MP4 movie with 60 fps. Rendering a

movie with 4,000 frames corresponding to 4μs of simulation took ~670 CPU-hours on TACC Frontera supercomputer.

To visualize the protein network (Suppl. Movies S7 and S8), we first convert each node's 3D position into its 2D projection. We used the Mercator projection, which is a cylindrical projection that preserves local directions and shapes. Specifically, for each node, its 2D projection (a,b) is calculated by:

$$r = \sqrt{x^2 + y^2 + z^2}$$

$$b = y/r$$

$$\mathrm{a} = \tan^{-1} x/z$$

where (x,y,z) is its normalized 3D position. The orientation of each protein domain is also converted in a similar manner. Then, we plot each node as a ellipse, where the its long axis follows the orientation and an orange tip suggests the direction, and colored spike proteins as red nodes and membrane proteins as blue nodes. The edges are represented as black lines where the new edges from the previous frame are highlighted as orange. We generated 1 frame per 10 ns, which resulted in 400 frames corresponding to 4μs of simulation.

4.3.6 High-Performance Computing for Molecular Dynamics Simulations

Production runs for the systems were performed on the TACC Frontera supercomputer on nodes equipped with Intel Xeon Platinum 8280 with 56 AVX_512 logical cores per node. For the full-spike model, the 1μs simulation took ~208,000 CPU hours; the truncated spike systems, running for 4μs, took ~925,000 CPU hours on average.

Composition		C1, model M1 truncated S, 4μs	C1, model M2, full S, 1μs
Proteins	S	25	25
	E	2	2
	M	1,003	1,003
	Total particles	**968,373**	**1,237,398**
Lipids	POPC	34,923	34,923
	POPE	11,837	11,837
	POPI	5,918	5,918
	POPS	1,183	1,183
	CHOL	2,663	2,663
	CDL2	2,663	2,663
	Total particles	**751,373**	**751,373**
Solvent	Na	175,565	409,000
	Cl	184,999	517,855
	Water	13,045,399	34,236,835
	Total particles	**13,220,964**	**35,163,690**

Table 3. Overview of system compositions. Shown are the compositional details of the two stable envelope models in molecular composition C1 (2 E pentamers, 25 S trimers, 1003 M dimers) with truncated (model M1) and full (model M2) S trimers. A CG water particle corresponds to 4 real water molecules.

4.4 Conclusions

This work provides an initial large-scale structural genomics and interactomics effort towards understanding the structure, function, and molecular dynamics of the SARS-CoV-2 virus with the aim to facilitate the process of structure-guided research where accurate structural models of proteins and their interaction complexes already exist. It also analyzes current capacities for the computational modeling of 3D structure of viral proteins. The importance of rapidly uncovering functional complexes and coming up with therapeutic targets is highlighted by the ongoing COVID-19 pandemic that demonstrated how vulnerable modern globalized

society is to the emergence of novel highly transmissible disease. It means we have to advance computational modeling tools and streamline pipelines in order to be ready to fight next biological threat to the human lives. One of such initiatives is the development of mesoscale modeling approach for the SARS-CoV-2 viral envelope.

Despite remarkable similarity of the virion structures shared between betacoronaviruses, such as SARS-CoV, MHV, and SARS-CoV-2, and the efforts to elucidate the structure of these viruses using imaging and computational methods (509, 523, 537-540), no high-resolution structure of the envelope currently exists. Our integrative approach allows combining experimental information at different resolutions into a consistent model, providing structural and functional insights beyond what can be obtained by a single experimental method. The obtained model is an important step towards our understanding of the underlying molecular architecture of the entire virus and successfully bridges the gap between molecular simulations and electron microscopy of virions, reproducing the experimentally observed density profiles of the local envelope structure. Furthermore, the developed computational protocol can be applied to study the envelopes of other coronaviruses, once the models or experimental structures and stoichiometries of the structural proteins comprising the envelope are obtained. The model will join in other efforts to structurally characterize virion particles with molecular dynamics such as influenza A, human immunodeficiency virus (HIV), and hepatitis B virus (HBV) (541-545).

The model can serve as a structural scaffold for understanding the interplay between mutual orientations of neighbor spike trimers and the role of this orientation in the viral interaction with the host receptors, as well as for studying the interactions between the proposed elongated form of M dimers and N-proteins(507) to discern the mechanistic determinants of the virion's stability. The structures of M dimers and the complex inter-dimeric assemblies they

form can provide the structural basis for understanding the molecular mechanisms behind the viral assembly. The structural knowledge of complexes formed by M proteins can also be helpful when designing new antiviral compounds targeting the interaction interface of the M protein and thus preventing formation of the envelope, an approach recently suggested for other viruses (546, 547). Targeting of the viral envelope with antiviral drugs is directly accessible within the Martini model framework (548). Finally, the structural model of the viral envelope will facilitate the development of viral-like nanoparticles for novel vaccines (549).

Chapter 5. Discussion and Future Directions

5.1 Discussion

The field of precision medicine currently experience unprecedented advancement. Fueled by the decrease in experimental cost, rapid improvement of the machine learning algorithms, and high-profile initiatives, it fosters collaborative environment for the scientists with various backgrounds – physics, biology, chemistry, and computer science. The latter can discover a fertile field of opportunities in advancing computational methods that directly affect an experimental outcomes and findings or adopting existing or devising novel cutting-edge algorithms for secondary analysis. Biological data is not easy to work with – the noise and technical variation are inevitable, confounding variables are plenty, its description is often much less structured than that of counterparts from business and tech worlds. However, these exact difficulties can stimulate advances in AI, pushing boundaries in the domains of knowledge discovery and advocating for the increase in reproducibility and more robust model evaluation and validation strategies.

The most common challenges in working with biological data are non-interpretability of the data from the human perspective, limited sample size, and technical variations. The first problem is the most global one, as it circumvents our ability to access analysis results and machine learning models. The ML applications are among the approached that experience the most scrutiny from the clinical community, and this attention in not an unwarranted one. It is often quite easy to obtain high scores for common machine learning metrics, however, the performance can be artificially boosted because of the information leakage problems, or it may

be not generalizable outside of the limited dataset it was trained on. Due to this, for biological and medical applications it is not sufficient just to assess performance on validation set, it is also important to peer into results obtained with its help and check their soundness from multiple points of view. E.g., in this work I dive deep into the biological validation by conducting case studies and assessing relevance of the created models. The second challenge, small datasets, can be potentially overcome with the usage of transfer learning by the reusing high-level general-purpose information extracted from gene or protein sequences. As an alternative, domain adaptation methods could be used in tandem with the batch effect removal approaches to expand scarce data by integrating them with the information available in public domain. Ubiquitously present technical noise calls for more sophisticated normalization approaches, however, blindly using statistical methods to correct for this variation may have adverse effects, e.g., reducing level of the true biological signal, changing a measurement scale and making adjusted samples non-comparable with other samples obtained via the same technology.

Computational methods keep increasing their impact on another crucial aspect of biological research – structural modeling. Novel *de novo* algorithms produce very promising results by demonstrating significant advances in international competitions. Though, more traditional homology-based approaches still demonstrate their usefulness by allowing scientist to pour in their biological expertise and integrate information from multiple sources into feasible constraints. Produced models with the resolved 3D structure are further studied using molecular dynamics approaches. These family of methods is especially demanding on computational resources and international teams of scientists constantly study ways to accelerate physics-based simulations. Besides adding raw computational power, some groups develop ML-aided force

fields that describe the nature of interaction between individual atoms and allows to approximate simulation results when large number of steps is skipped.

5.2 Schematics of algorithms improvement

5.2.1 Predicting alternative splicing isoforms from scRNA-Seq short-read data

Previous cell-pool studies on scRNA-Seq **(301)** demonstrated that small groups of cells increase data analysis robustness and experiment reproducibility. Inspired by those successes, I propose to transfer cell-pool analysis to the *in-silico* stage. The utility of a similar pseudo-bulk approach, MetaCell **(550)**, was previously demonstrated by the wide range of applications to cell profiling **(551-554)**. The idea is to use a combination of data cleaning methods **(302, 304)** and high accuracy clustering like recent deep-learning-based DUSC **(555)** to identify subpopulations in scRNA-Seq data and combine them into "metacell." This algorithm previously demonstrated separating scRNA-Seq data from the mouse cortex into well-defined clusters (Fig.12).

The "metacell" data would allow us to extend read depth and, with high probability, coverage. It would enable estimating splicing events in specialized cells and potentially use bulk RNA-Seq quantification methods. Another possible study direction is the detection of subpopulation snapshots that correspond to distinct cell cycle phases.

The study on the subpopulation-based methods for alternative splicing detection can be conducted according to the following methodology. First, a golden-standard dataset for alternative splicing detection can be constructed from publicly available single-cell and bulk paired samples, *e.g* in **(556)**, **(303)**, **(557)**. This would allow us to compare several isoforms detected via high coverage bulk RNA-Seq data and a total number of isoforms across individual cells identified based on single-cell data. Second, I would construct a simulated dataset

according to the protocol from **(299)** to provide a controlled environment for accessing technical noise and study the effect of 'metacell' size on AS events detection to get insights into the nature of a trade-off between the size of sub-population and the data coverage.

The final algorithm follows the next steps:

1. Refining splicing boundaries annotation using SpliceAI in tandem with DeepSplice, keep only novel isoform identified by both algorithms

2. Aggregate cells with specified granularity via DUSC algorithm into meta-cells

3. Merge sequence reads into pseudo-bulks in order to increase read depth

4. Quantify isoforms abundance according to the Salmon statistical model based on the alignment to the references that were updated at the Step 1.

5. Incorporate machine learning model estimating sufficient sequencing depth based on the meta-cell granularity and alternative splicing patterns across different cell types.

Benchmarking

The hypothesis is that combining cells from the same subpopulation would essentially increase the alternative splicing signal. In order to test this hypothesis, I will leverage three types of AS prediction methods. First are the traditional quantification methods used in bulk RNA-Seq setting – DEXSeq and Census. I expect that cell subpopulation would behave akin to the tissue type, just more homogeneous. Second is the single cell-specific methods because the collection could be considered as enriched single-cell data. The third is our targeted middle ground – subpopulation-based methods which compensate for low coverage of Smart-based and,

especially, UMI-based methods while maintaining enough specificity for each cell type aligned across pseudotime.

Approaches such as RSEM and Splatter heavily rely on provided examples of scRNA-Seq experiments for simulations. A strong point of such an approach is that synthetic dataset closely resembles real-world biological experiments. Still, the downside is our "ground truth" isoform quantification data is biased and does not represent an accurate biological picture. Nevertheless, this is a widely used approach for benchmarking. It would be interesting if serious discrepancies would arise between evaluations on the simulated dataset and paired single-cell and bulked datasets.

For evaluation of isoform expression on synthetic single-cell ground truth data $\boldsymbol{O}$ I would use normalized root mean square error (NRMSE):

$$\boldsymbol{NRMSE = 100\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2}}{\sigma(O)}},$$

where $\boldsymbol{S}$ is an estimation of the isoform expression, $\boldsymbol{N}$ is the total number of isoforms, and $\boldsymbol{\sigma(O)}$ is the standard deviation of the estimated expressions on the ground truth.

For the comparison with paired samples, I employ a formula for the RMSE:

$$\boldsymbol{RMSE = 100\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j}^{M} S_{ij} - O_i\right)^2}},$$

where ground truth data $\boldsymbol{O}$ comes from the bulk RNA-Seq experiment, hence, it has only a single value, $\boldsymbol{S}$ is an estimation of the isoform expression, $\boldsymbol{N}$ is the total number of isoforms, and $\boldsymbol{M}$ is the number of cells.

5.2.2 Domain Adaptation for ALT-IN Tool

Recently published method, ALT-IN **(236)**, leverages traditional semi-supervised approach in order to improve generalization ability of the machine learning model. However, this approach demonstrated its limitation in improving generalization capability of the model. Clustering of the original features reveal protein family-dependent groupings of interactions. Large amount of heterogeneity between features derived from distinct protein families makes it difficult to achieve generalization across all data space.

To account for this problem, I introduce transfer learning-based approach for predicting alternative splicing functional effects on protein-protein interaction networks that improves generalization by accounting for different distinct domain arising in data space. The information on PPIs can be bolstered by the large number of available transcript sequences **(327)** and interactomics data **(195)**.

The initial domain identification would be conducted via SCOP protein annotation **(333)**. SCOP is predominantly manual structural classification of the proteins. It four hierarchical levels: class, fold, superfamily and family. We have to experiment with granularity of the selected domain and carefully balance it based on validation results, as each distinct domain would decrease amount labeled samples available for each adaptation subtask.

In order to refine decision boundaries inside the domains I propose to add structural features during to the training phase, following Learning Under Privileged Information paradigm **(366)**. I describe my approach for incorporating LUPI into Random Forest algorithm

In addition to introducing a transfer learning method for predicting functional effects of alternative splicing, we are aiming for improving model explainability and bolstering its potential for knowledge discovery applications. Therefore, we would like to preserve as much interpretability for our feature set as possible, opting for the set of biochemistry- and alternative splicing-based features introduced in **(236)**. For the aforementioned purposes, inference-based methods are our best choice. In particular, the work **(215)** discusses the usage of Bayesian inference for weighing features for age prediction from DNA methylation. Due to the diverse nature of our data, in order to use this approach, we have to extend it for a multisource domain setting. The theoretical results **(227, 558)** indicate that such a problem is tractable, and **(559, 560)** demonstrate that a weighted ensemble of classifiers over source domains is a reasonable and widely adopted approach. This methodology is highly suitable for the proposed semi-supervised approach because it strives to create a hierarchical Gaussian mixture model over available data and individual trees can be used to derive underlying source domains.

Feature Confidence Scores in WENDA Method

The WENDA method **(215)** for multiple target domains adaptation that we are taking for a basis has the following assumptions and definitions:

1. $\boldsymbol{n}$ labeled examples $(\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2}), \dots, (\boldsymbol{x_n}, \boldsymbol{y_n})$ for the source domain, where $\boldsymbol{x_i}$ is the $\boldsymbol{p}$-dimensional vector and $\boldsymbol{y_i}$ is a scalar.
2. $\boldsymbol{m}$ labeled examples for the target domain $(\widetilde{\boldsymbol{x}}_{\boldsymbol{1}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{1}}), (\widetilde{\boldsymbol{x}}_{\boldsymbol{2}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{2}}), \dots, (\widetilde{\boldsymbol{x}}_{\boldsymbol{m}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{m}})$, where $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}$ is the $\boldsymbol{p}$-dimensional vector and $\widetilde{\boldsymbol{y}}_{\boldsymbol{i}}$ is a scalar.
3. The data comes from the two different distributions: $\boldsymbol{P_S(X,Y) = P_S(Y|X) \cdot P_S(X)}$ and $\boldsymbol{P_T(X,Y) = P_T(Y|X) \cdot P_T(X)}$ for the (S)ource and (T)arget domains correspondingly.
4. Weakened covariance shift assumption: $\boldsymbol{P_S(X) \neq P_T(X)}$, $\exists \boldsymbol{M} \subseteq \{\boldsymbol{f_1}, \dots, \boldsymbol{f_p}\}$:

$$P_S(X_f|X_{\neg f}) \approx P_T(X_f|X_{\neg f}) \Rightarrow P_S(Y|X_M) \approx P_T(Y|X_M), \qquad \forall f \in M$$

It means that for our data there exists a feature subset M of that has the same influence on source and target domain and features outside this subset are allowed to influence domain distributions differently, unlike for the strict covariance shift assumption.

These assumptions are used as a basis for a Gaussian process model that takes for the inputs a p-dimensional vector $x_{\cdot,f} = (x_{1,f}, x_{2,f}, \dots, x_{n,f})$ that contains all values for the feature f and $n \times (p-1)$ matrix $X_{\cdot,\neg f} = (x_{1,\neg f}, x_{2,\neg f}, \dots, x_{n,\neg f})$ that contains the rest of the features and maximizes log-likelihood

$$\log p(x_{\cdot,f}|X_{\cdot,\neg f}) = -\frac{1}{2}x_{\cdot,f}^T(K + \sigma_n^2 I_n)^{-1}x_{\cdot,f} - \frac{1}{2}\log|K + \sigma_n^2 I_n| - \frac{n}{2}\log 2\pi$$

where linear kernel matrix $K = \sigma_p^2 X_{\cdot,\neg f} X_{\cdot,\neg f}^T$, σ_p^2 is the variance of the prior on coefficients, σ_n^2 is the variance of the noise and $I_n$is the identity matrix of the dimensionality n. The posterior distribution of the coefficients ω has the following closed-form solution, according to **(215)**:

$$p(\omega|x_{\cdot,f}, X_{\cdot,\neg f}) \sim \mathcal{N}(\sigma_n^{-2} A^{-1} X_{\cdot,\neg f}^T; x_{\cdot,f}^T, x_{\cdot,f} A^{-1}),$$

where $A = \sigma_n^{-2}; X_{\cdot,\neg f}^T; x_{\cdot,f}^T + \sigma_p^{-2} I_{p-1}$. According to these probabilities, the goodness of fit for the observable value $\tilde{x}_{i,f}$ into the feature model g_f is calculated. Based on the Gaussian process g_f the posterior distribution for the $\tilde{x}_{i,f}$ value based on $\tilde{x}_{i,\neg f}$ is defined as $\mathcal{N}(\mu_{g_f}(\tilde{x}_{i,\neg f}), \sigma_{g_f}(\tilde{x}_{i,\neg f}))$. The goodness of fit for the $\tilde{x}_{i,f}$ is quantified in confidence scores proposed in the **(561)**:

$$c_f(\tilde{x}_i) = 2\Phi\left(-\left|\frac{\tilde{x}_{i,f} - \mu_{g_f}(\tilde{x}_{i,\neg f})}{\sigma_{g_f}(\tilde{x}_{i,\neg f})}\right|\right),$$

where Φ corresponds to the CDF of standard normal distribution, and **(215)** define an overall confidence score of the feature f as

$$\boldsymbol{c_f = \frac{1}{m}\sum_{i=1}^{m} c_f(\tilde{x}_i)}$$

Protein-protein interaction prediction as a domain adaptation problem

Large amounts of data can be derived from protein sequences. Still, the resulting datasets are not homogeneous and additional steps are needed in order to ensure adequate generalization ability of the ML models trained on them **(562, 563)**. For example, **(562)** uses protein superfamily split in protein homology detection application. Here, we are expanding upon our previous work on the alternatively spliced isoforms' effects on PPIs, ALT-IN **(564)**.

My idea is to treat each protein family or superfamily (the most appropriate way to designate split for our application is also a subject for the research, as currently there is no community standard) as separate domains. PPI prediction generally falls under a pair-input machine learning scheme **(164)**, which means that features for each labeled data point come from two different sources, in our case, proteins. In the general case this adds an extra layer of complexity because we would not have a single source domain of the classical DA problem, but a composite one $\boldsymbol{P_S(X_i, X_j, Y)}$, where $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ correspond to the proteins belonging to families $\boldsymbol{i}$ and $\boldsymbol{j}$ correspondingly. Our ALT-IN method deals only with isoforms derived from the same gene that belongs to a single protein family, but the interactor partner does not have such restrictions, thus bears a similarity to the generic PPI prediction methods in this regard. We are taking a straightforward approach to this problem by performing domain adaptation for the $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ independently.

Among the available DA methodologies, we chose to follow inference-based one which strives to find a robust subset of features. Thus, our problem can be formulated as

$$\min_{c_f, F} \mathcal{L}\left(c_f, F\right),$$

where c_f is the vector of feature weights and F is a classifier.

LUPI in Random Forest

One of the crucial steps in Random Forest is node split. Here, we would consider K -best split approach, where on each step of tree building we select the best performing split on one of the K selected features. One of the most ubiquitous criteria for selecting this split is Gini impurity **(565)**:

$$G(m) = 1 - \sum_{j=1}^{m} \left(p(c_j)\right)^2,$$

where m is the number of elements in node, $p(c_j)$ is the probability of assigning class c_j to the element if it would be done on the random. Effectively, $p(c_j)$ equals a fraction of elements belonging to the class c_j in the node. Another widely used approach is an Information Gain criteria **(91, 565)**. For the sake of generality, let us denote the impurity index employed by Random Forest as $I(m)$, where m is a number of classes.

We are proposing to improve the node splitting procedure by incorporating the LUPI paradigm into this step. Though LUPI is a generic paradigm and is not specific to selected classifiers, current general-purpose implementations are limited to the SVM+ implementations **(339, 340, 343, 566)** and neural networks **(344, 345)**.

Under the LUPI paradigm let us consider X to be an entire set of features, vector x of size M, where M is a number of samples on each split feature and complimentary privileged feature matrices $X_1^*, X_2^*, \ldots, X_L^*$ with the dimensions $m_i \times n_i$, $2 < m_i < M$ and $n_i > 0$. Each matrix X_i^*

corresponds to a separate privileged feature set which may include different data samples from $\boldsymbol{X}$.

Our idea is to add a regularization term for the split reflecting the ability to separate data in supplementary privileged data space $\boldsymbol{X_i^*}$ to the impurity index $\boldsymbol{I(m)}$. One of the most natural ways to achieve this is to construct a split using linear SVM+ and calculate corresponding margin size $\boldsymbol{S = \frac{2}{|w|}}$, where $\boldsymbol{w}$ are the feature weights of linear SVM. The resulting expression would be:

$$\boldsymbol{I^*(m) = I(m) + \sum_{i=1}^{L} \lambda_i \frac{1}{S_i}},$$

where $\boldsymbol{L}$ is the number of privileged datasets, $\boldsymbol{\lambda_i}$ is the regularization parameter and $\boldsymbol{S_i}$ is the margin size of the linear SVM+ trained on $\boldsymbol{i}$-th privileged dataset.

Semi-supervised learning and low-density separation

Node split procedure also became a focus in Sherwood **(567)** modification of random forest for semi-supervised learning. There an additional information gain regularization term is based not on the class impurity in the node, but on the purity of data distribution in child nodes $\boldsymbol{L}$ and $\boldsymbol{R}$:

$$\boldsymbol{IG = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i)},$$

where $\boldsymbol{H(S)}$ is the Shannon entropy and $\boldsymbol{S}$ is a data subset. For the continuous distribution, entropy is defined as

$$\boldsymbol{H(S) = -\int p(y) \log (p(y)) dy}$$

Sherwood operates takes Gaussian distribution as a basis for the low-density separation because of the computational efficiency, as for the $\boldsymbol{d}$-variate Gaussian entropy can be written in a closed form:

$$\boldsymbol{H(S) = \frac{1}{2}\log((2\pi e)^d|\Lambda(S)|)}$$

Then it is used as a normalization term in the final impurity criteria analogous to the previous section:

$$\boldsymbol{I^*(m) = I(m) + \lambda \cdot IG(U),}$$

where $\boldsymbol{U}$ corresponds to the unlabeled data and $\boldsymbol{\lambda}$ is a regularization term.

Domain Adaptation in ALT-IN

ALT-IN method requires the input dataset to be a structured input of triplets $\bar{\mathbf{X}} = (\boldsymbol{X, D, R})$ with dimensionality $\boldsymbol{N \times M}$, where $\boldsymbol{X}$ is the main isoform features, $\boldsymbol{D}$ correspond to delta features (difference between main isoform feature vector $\boldsymbol{X}$ and the feature vector derived from the alternative isoform), and $\boldsymbol{R}$ corresponds to interaction partner's features and $\bar{\mathbf{X}}$ is the entire dataset. $\boldsymbol{X}$ and $\boldsymbol{I}$ can belong to the different protein families, and delta features may form their own alternative splicing-related dependencies, i.e., all three feature groups can be considered as coming from the different statistical distributions. Ideally, we would like to capture nuanced interplays arising from conditional probabilities for the distributions of those domains. Still, our limited dataset does not provide enough information to make such a study viable. Because of this, we are making several simplifications.

We would study the effects of $\boldsymbol{X, D}$ and $\boldsymbol{R}$ coming from different protein families separately. Thus, for each protein family $\boldsymbol{\mathcal{F}_i}$ we are calculating confidence scores $\boldsymbol{c_f}$ in a disjoint

manner. Let's denote the confidence score calculation procedure from the previous section that produces confidence $\boldsymbol{c_f}$ for every feature from the matrix $\boldsymbol{X}$ as $\boldsymbol{C(X)}$. Then we will generalize this expression for subsets of $\boldsymbol{X}$: $\boldsymbol{C_S(X_S, X_A)}$ such that concatenation $(\boldsymbol{x_S, x_A}) \in \boldsymbol{X}, \forall \boldsymbol{x_S} \in \boldsymbol{X_S}, \boldsymbol{x_A} \in \boldsymbol{X_A}$ belongs to the original set $\boldsymbol{X}$, where $\boldsymbol{S}$ is a fixed subset and $\boldsymbol{A}$ corresponds to the lack of any restrictions.

This results in the following scheme: to perform domain adaptation we have to consider two potentially different protein families $\boldsymbol{\mathcal{F}_i, \mathcal{F}_j}$ and corresponding datasets $\left(\boldsymbol{X_{\mathcal{F}_i}, D_{\mathcal{F}_i}, R_{\mathcal{F}_j}}\right)$, calculate an $\boldsymbol{M}$-dimensional vector of confidence scores

$$\boldsymbol{C} = \left(\boldsymbol{C_{\mathcal{F}_i}}\left(\boldsymbol{X_{\mathcal{F}_i}, D_A, R_A}\right), \boldsymbol{C_{\mathcal{F}_i}}\left(\boldsymbol{X_A, D_{\mathcal{F}_i}, R_A}\right), \boldsymbol{C_{\mathcal{F}_j}}\left(\boldsymbol{X_A, D_A, R_{\mathcal{F}_j}}\right)\right)$$

Then we modify the procedure of the random subspace selection and corresponding vector $\boldsymbol{\Theta_k}$ for the Random Forest algorithm. We apply stratified sampling to the feature selection procedure. It was previously shown that stratified sampling of the features plays a positive role in the classifier's performance **(568-570)**. More specifically, while selecting feature subspace of size $\boldsymbol{k}$ during node split we are visiting individual features in random order and sample with replacement feature $\boldsymbol{i}$ with probability $\boldsymbol{p_i} = \frac{\boldsymbol{c_{f_i}}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} \boldsymbol{c_{f_j}}}$.

By adopting this approach, we ensure that the main information flow is confined to the subset of features, robust for the specific protein family while not losing flexibility provided by the entire set.

Appendix A

Supplementary Figures

1: X_{train} = train set data samples
2: Y_{train} = train set labels
3: U = unlabeled data samples
4: N = number of elements to add to train set
5: T = *False*
6: ε = threshold value
7: $X_{best} = X_{train}$
8: RF_{model} ←run Random Forest classifier on train set (X_{train}, Y_{train})
9: $Fscore_{best}$ ← F-score of RF_{model} based on 10-fold CV on (X_{train}, Y_{train})
10: $RF_{bestmodel} \leftarrow RF_{model}$
11: **while not** T **do:**
12: Y_U ← Classification result of RF_{model}on U
13: $\left(U_{ordered}, Y_{U_{ordered}}\right)$ ←order U along with corresponding Y_U according to the probability of Y_{U_i} to be correct label for sample U_i in descending order
14: (K, Y_K) ← first N elements from set $\left(U_{ordered}, Y_{U_{ordered}}\right)$
15: (X_{new}, Y_{new}) ← merge (X_{train}, Y_{train}) and (K, Y_K)
16: $RF_{newmodel}$ ← run Random Forest classifier on train set (X_{new}, Y_{new})
17: $Fscore_{new}$ ← F-score of $RF_{newmodel}$ based on 10-fold CV on (X_{train}, Y_{train})
18: **if** $Fscore_{new} > Fscore_{best}$:
19: $Fscore_{best} \leftarrow Fscore_{new}$
20: $RF_{bestmodel} \leftarrow RF_{newmodel}$
21: $(X_{train}, Y_{train} \leftarrow (X_{new}, Y_{new})$
22: **remove** K from U
23: **if** $|Fscore_{old} - Fscore_{new}| < \varepsilon$**:**
24: T ←True

Figure A1. A pseudocode of iterative self-learning random forest algorithm used in AS-IN Tool.

Supplementary Tables

	Normal	T2D	Total
Conserved	279	332	611
Rewired	6	23	29
Total	285	355	640

Table A1. Contingency table for the rewiring for normal and T2D-related interactions. This table cross tabulate information on rewiring of normal and T2D-related interactions.

Feature Group		Related Proteins	Feature List
Biochemical		**A_1** **B**	Molecular weight Number of residues Average residue weight Charge Isoelectric point A280 molecular extinction coefficient for reduced and cystine bridges Frequency, Molarity, DayhoffStat for each residue and residue property (Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Basic, and Acidic)
Statistical Potentials		**(A_1,B)**	3 largest statistical potentials among all combinations of domain from protein and protein
		A_1	2 largest statistical potentials for individual domains of protein
Delta Features	Biochemical	**A_1-A_2**	*See feature list for Biochemical*
	Statistical potentials	**(A_1,B)-(A_2,B)**	*See feature list for Statistical Potentials*
		A_1-A_2	
	Sequence Alignment Based	**(A_1,A_2)**	Length change (ratio) Length change (absolute) N-termini C-termini Maximum alignment gap size Mean alignment gap size Number of alignment gaps Number of large gaps (>=10 bases) Number of small gaps (<10 bases)
	Domain Based	**(A_1,A_2)**	Domains lost Domains changed Domain or linker

Table A2. List of features used for machine learning classifiers. The list includes three main groups – biochemical features, statistical potentials, and AS-related "delta" features. Related Proteins column indicates which proteins among the reference isoform **A_1**, its interacting partner **B,** and alternatively spliced isoform **A_2** are used as the feature source. Two stand-alone proteins indicate that features described in the corresponding group were obtained for each protein independently. **(X,Y)** grouping of proteins indicate that both proteins, **X** and **Y,** are required to obtain the corresponding features. **X-Y** indicates that the corresponding features reflect the difference between the individual features of proteins **X** and **Y**.

Tool	Structural features list
FoldX	Total interaction energy
	Backbone Hbond
	Sidechain Hbond
	Van der Waals
	Electrostatics
	Solvation Polar
	Solvation Hydrophobic
	Van der Waals clashes
	Entropy sidechain
	Entropy mainchain
	sloop_entropy
	mloop_entropy
	cis_bond
	Torsional clash
	Backbone clash
	Helix dipole
	Water bridge
	Disulfide
	Electrostatic contribution
	Partial covalent bonds
	Energy Ionisation
	Entropy Complex
OPUS-PSP	Energy score
GOAP	Energy potential
Naccess2	Accessible surface area
Geometric	Energy potential
Dfire2	Energy score
InterproScan	Removed binding sites number
	Changed binding sites number
	Affected binding sites number
	Removed binding sites percentage
	Changed binding sites percentage
	Affected binding sites percentage
	Total binding sites number

Table A3. List of the privileged features for the SVM+ and SVM+ Boosting algorithms. This list includes structure-based characterization of the individual proteins and protein interactions.

Critical commercial assays		
Qiagen's RNeasy Mini Kit	N/A	https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/rna-purification/total-rna/rneasy-mini-kit/
TruSeq RNA v2	(571)	https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-rna-v2.html
Agilent 2500 BioAnalyzer	N/A	
Illumina HiSeq 2000	N/A	
Deposited Data		
Experimentally validated isoform interaction dataset	(64)	http://www.interactome-atlas.org/data/Yang-16.tsv
Interactome data	(64, 246, 252, 320, 321)	http://www.interactome-atlas.org/download
Processed mRNA	This paper (572)	http://dx.doi.org/10.17632/wc32wwvdwk.1 Link to preview: https://data.mendeley.com/v1/datasets/wc32wwvdwk/draft?preview=1
Type 2 Diabetes Knowledge Portal	(573)	http://www.type2diabetesgenetics.org/
STRING database	(574)	https://string-db.org/
ENSEMBL	(327)	https://useast.ensembl.org/
GRCm38.p5	(575)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.25/
Type 2 Diabetes GWAS studies	(576) (577) (578)	http://diagram-consortium.org/downloads.html https://www.ebi.ac.uk/ega/studies/EGAS00001001459 https://www.ebi.ac.uk/ega/studies/EGAS00001001460
DOMMINO	(247)	http://korkinlab.org/dommino
Software and algorithms		
ALT-IN Tool	This paper (579)	https://doi.org/10.5281/zenodo.5234256 https://github.com/KorkinLab/alt-in-tool https://hub.docker.com/r/narykov/alt-in
Trimmomatic	(580)	http://www.usadellab.org/cms/?page=trimmomatic
Tophat v2	(581)	https://ccb.jhu.edu/software/tophat/index.shtml
Cufflinks v2	(582)	http://cole-trapnell-lab.github.io/cufflinks/

SUPERFAMILY	(338)	https://supfam.org/
SVM+	(340)	Requested from authors
FoldX	(347)	http://foldxsuite.crg.eu/
OPUS-PSP	(349)	http://ma-lab.rice.edu/MaLab/soft.php
GOAP	(350)	https://sites.gatech.edu/cssb/goap/
NACCESS2	(351)	http://www.bioinf.manchester.ac.uk/naccess/
Geometric	(352)	Requested from authors
DFire2	(353)	https://sparks-lab.org/downloads/
InterProScan	(583)	https://www.ebi.ac.uk/interpro/
EMBOSS	(584, 585)	https://www.ebi.ac.uk/Tools/emboss/
LR_PPI	(319)	http://www.csbio.sjtu.edu.cn/bioinf/LR_PPI/
TRI_Tool	(336)	http://www.vin.bg.ac.rs/180/tools/tfpred.php
DIIP	(317)	https://github.com/MohamedGhadie/isoform_interactome_prediction
scikit-learn	(356)	https://scikit-learn.org/
pyMannKendall	(323)	https://pypi.org/project/pymannkendall/
Cytoscape	(586)	https://cytoscape.org/
Adobe Illustrator		https://www.adobe.com/products/illustrator.html
BioRender		https://biorender.com/

Table A4. List of key resources used during creation or evaluation of the ALT-IN machine learning model.

Appendix B

Supplementary Figures

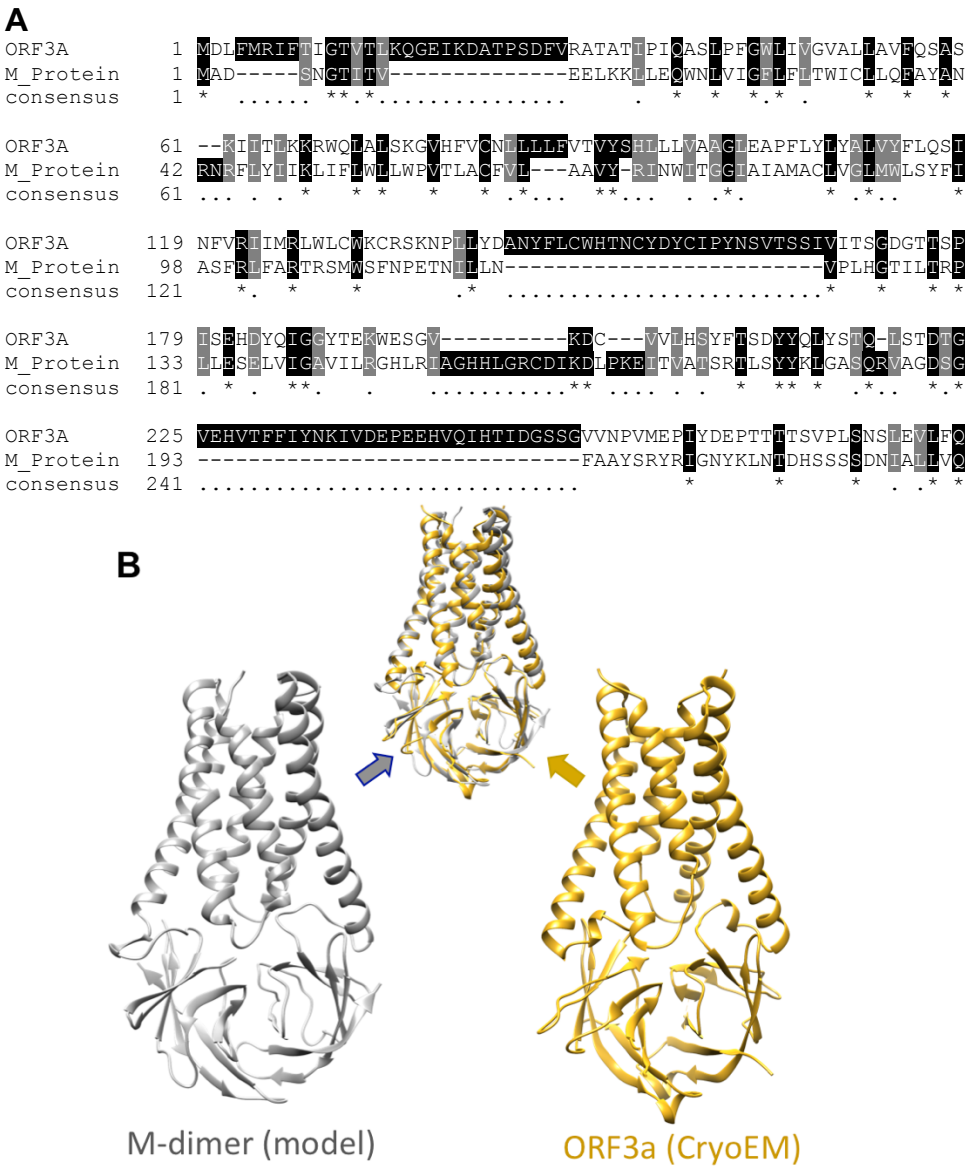

Figure B1. In spite of substantial difference in protein sequences M model resembles closely a structurally resolved ORF3a dimer. A. Protein sequence alignment between M and ORF3A proteins of SARS-CoV-2 using TCoffee. Sequence identity is 16%. B. Structural alignment of M dimer model and ORF3A structure (PDB ID: 6XDC) using TM-align. All-residue RMSD is 1.64A.

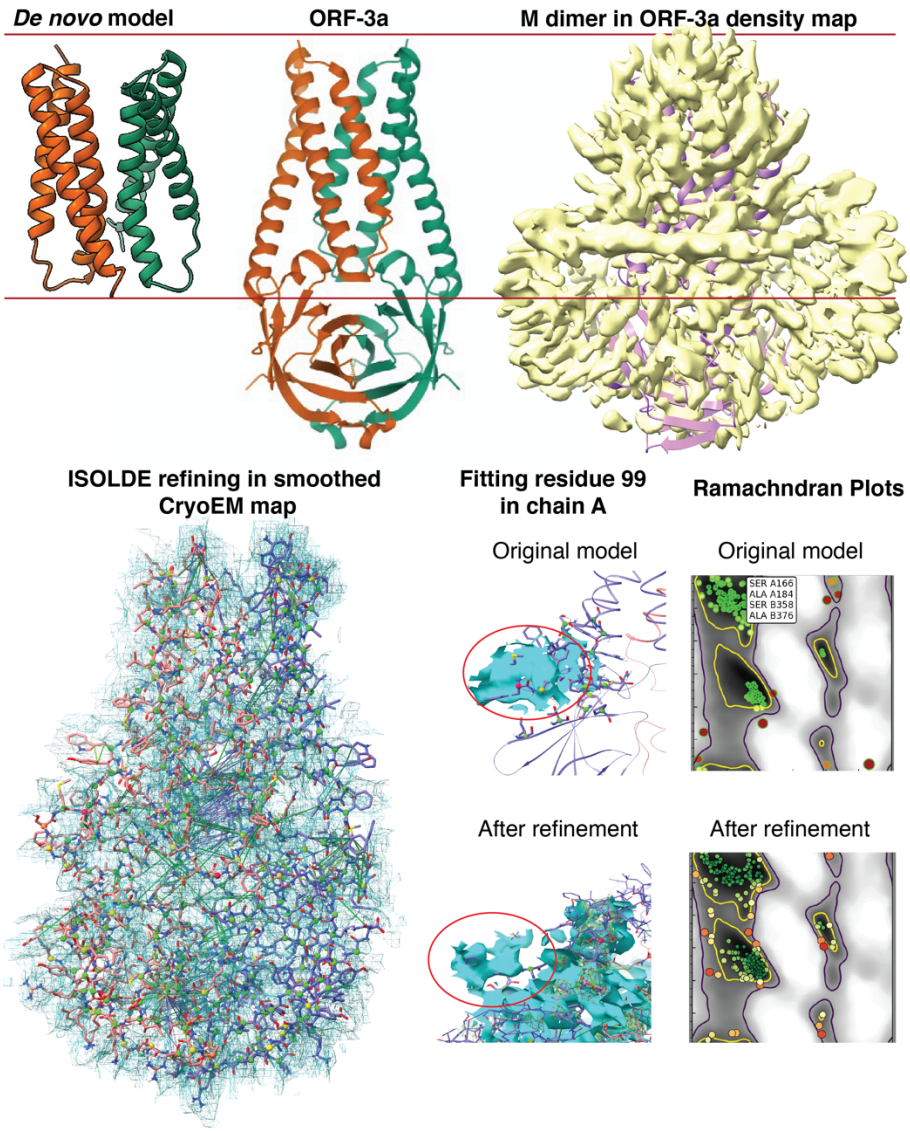

Figure. B2. Structural refinement of M dimer. Top left: *de novo* model of the M dimer without endodomain exhibit similarity to the ORF-3a protein from SARS-CoV-2. Top right: M dimer model with endodomain in Cryo-EM density map of ORF-3a in lipid nanodisc (EMD-22139). Bottom left: M dimer model during ISOLDE simulation run in ChimeraX molecular visualization suite; blue mesh corresponds to the CryoEM map with Gaussian smoothing with B-factor equal to 100, green lines correspond to the local restraints at the distances of 1.5Å or more, and the colors of residue's atoms correspond to atoms' goodness of fit. Bottom middle: an example of fitting a specific residue 99 in chain A; in the original model this residue sticks out of the density map (blue surface), and during the process of refinement it acquires a better fit. Bottom right: Ramachandran plots of the original and refined models; one can see that the original model.

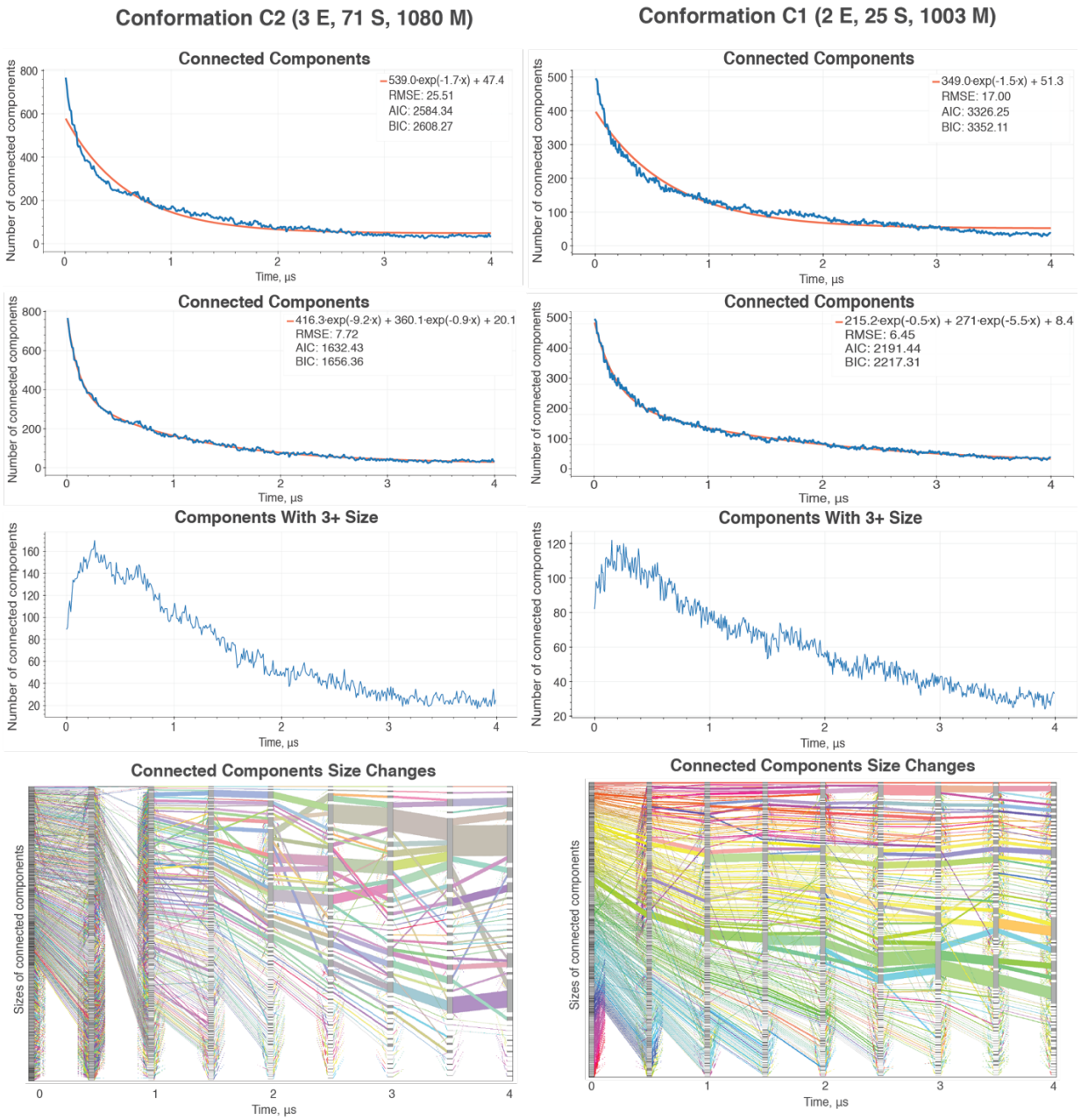

Figure B3. Change in the number of connected components in the TM domain-domain interaction networks throughout the simulation for models M2 (conformation C2) and M1 (conformation C1). Between two possible fitted lines, exponential model (first row) and biexponential model (second row), one can see that the biexponential model provides better RMSE score along with smaller values for Akaike's Information Criterion (AIK) and Bayesian Information Criterion (BIC), suggesting that the biexponential model better represents the underlying process and supporting our hypothesis about two separate processes happening during the simulation. Specifically, for both M1 and M2 models, there are two distinct exponential processes, fast and slow, differing in speed ~10x times. Shown in the third row is the number of components with at least 3 proteins. Again, there is an increase in the number of components at the beginning of simulation, but later the components tend to merge, bringing a total number of components under 40 for both M1 and M2. The last row represents Sankey diagram of connected components, demonstrating the tendency of the envelope structure to form larger clusters of TMDs for M dimers. Each flow depicts rearrangement of 25 or more proteins.

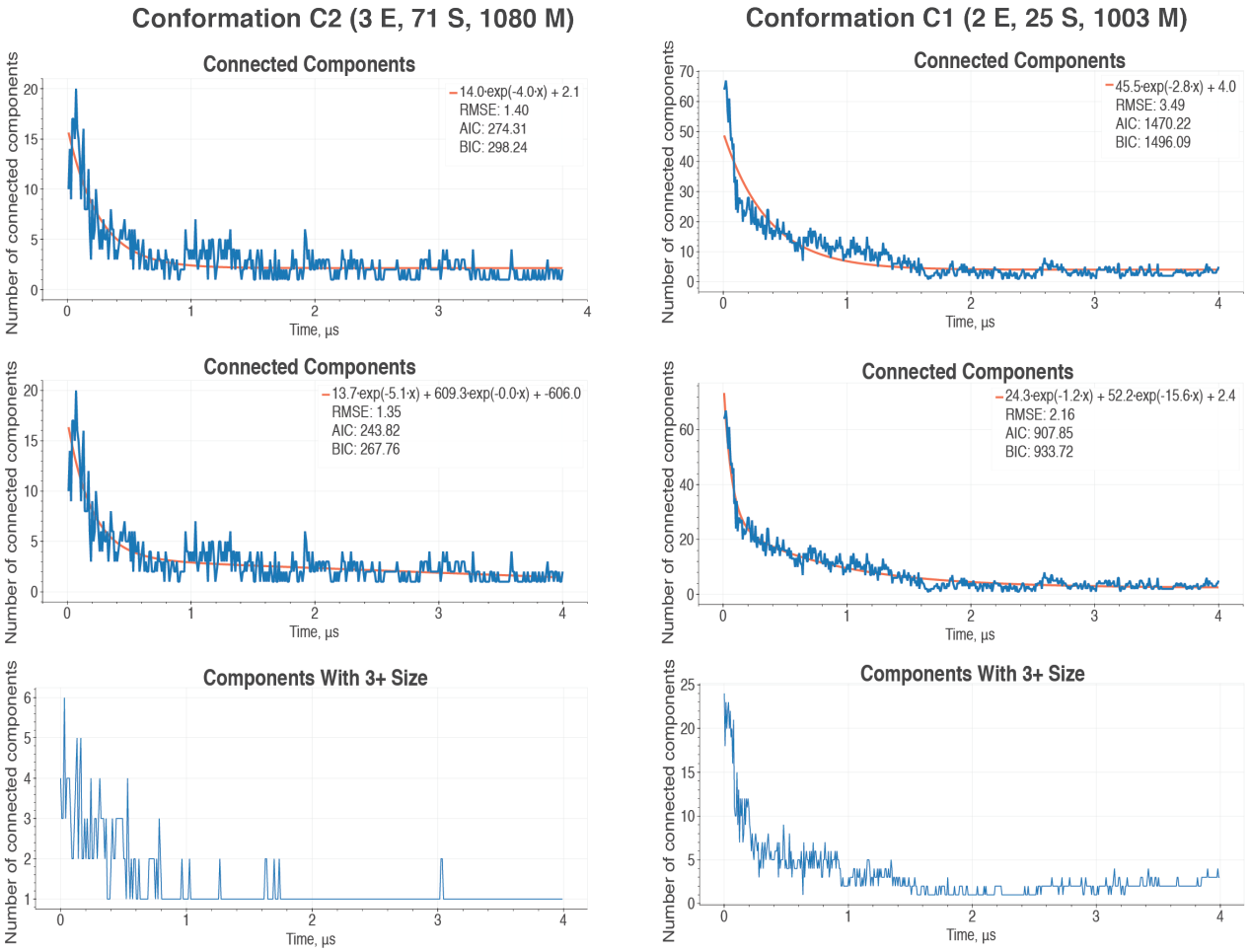

Figure B4. Change in the number of connected components in the ED domain-domain interaction networks throughout the simulation for models M2 (conformation C2) and M1 (conformation C1). Similar to Suppl. Fig. S8, one can see that ED rearrangement can be described as a biexponential process, although for M1 the difference between biexponential and exponential models is minimal because of extremely tight packing. Also, for M2 its EDs tend to conglomerate to such an extent that they have only one, giant, connected component of the size of three or more; for M1 this number varies over time but stays under five connected components.

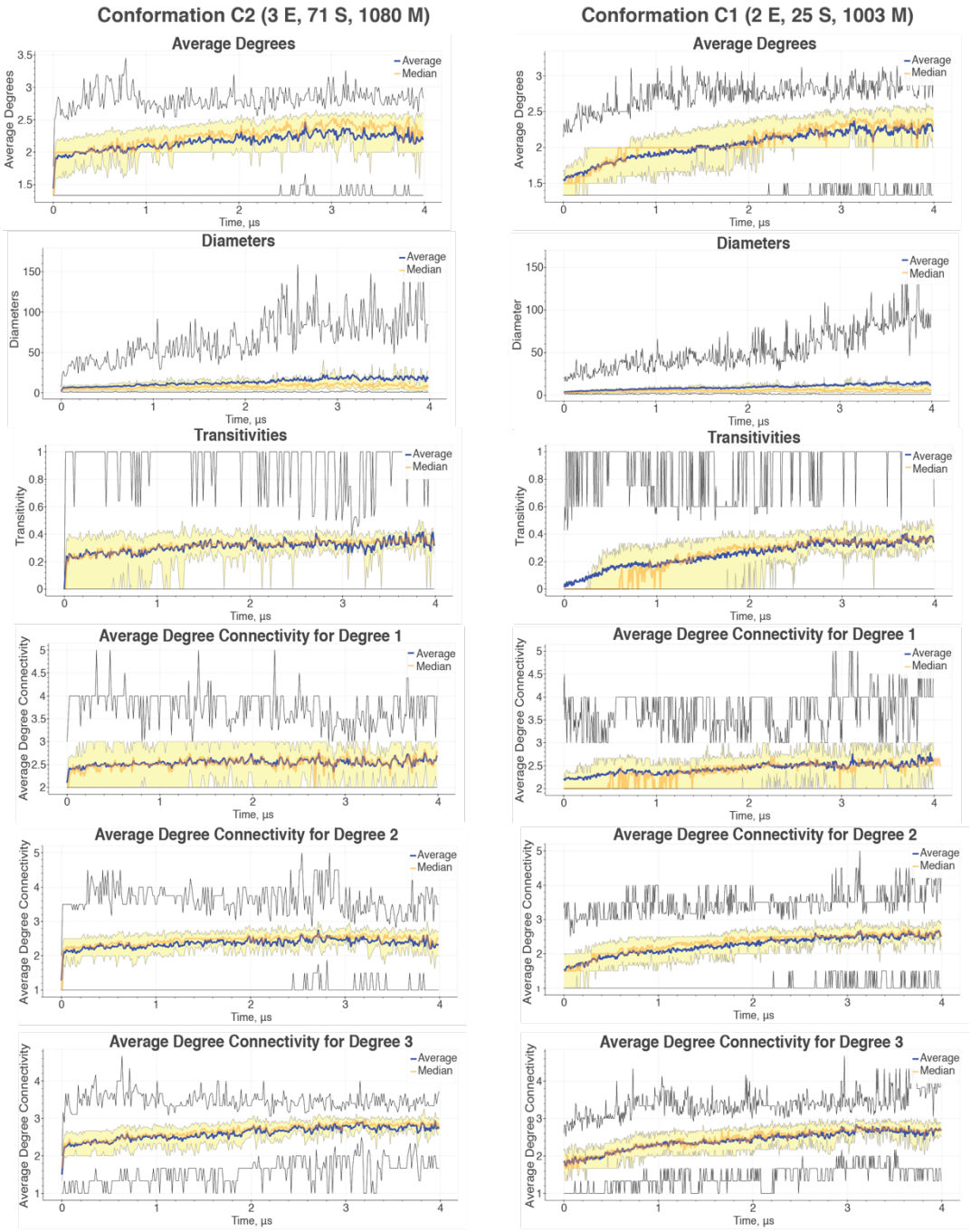

Figure B5. Dynamics of the basic network parameters for TMD domain-domain interaction networks during the simulation of models M2 (conformation C2) and M1 (conformation C1). Yellow bands indicate values for the 2nd and 3rd quantiles, black lines denote the minimum and maximum values. There is a clear trend for the increase in node degrees for TMD components, plateauing during 3-4 μs around the value of 2.3. The diameter sizes tend to grow over time. The vast difference between the maximum and average or median values suggests that small number of connected components cover most of the network. The transitivity measure depicts a ratio of fully connected components in the network with respect to the maximum number of possible components for the given network. The measure stabilizes over time around the value 0.35 with very limited spread for the 2nd and 3rd quantiles, suggesting that this characteristic is similar for the most connected components. Average degree connectivity characterizes the nearest neighbors of the nodes of degree k. The values for the nodes of

degrees 1, 2, and 3 all converge to comparable average degree connectivities (around 2.5), suggesting that those nodes are evenly distributed in the protein connectivity network.

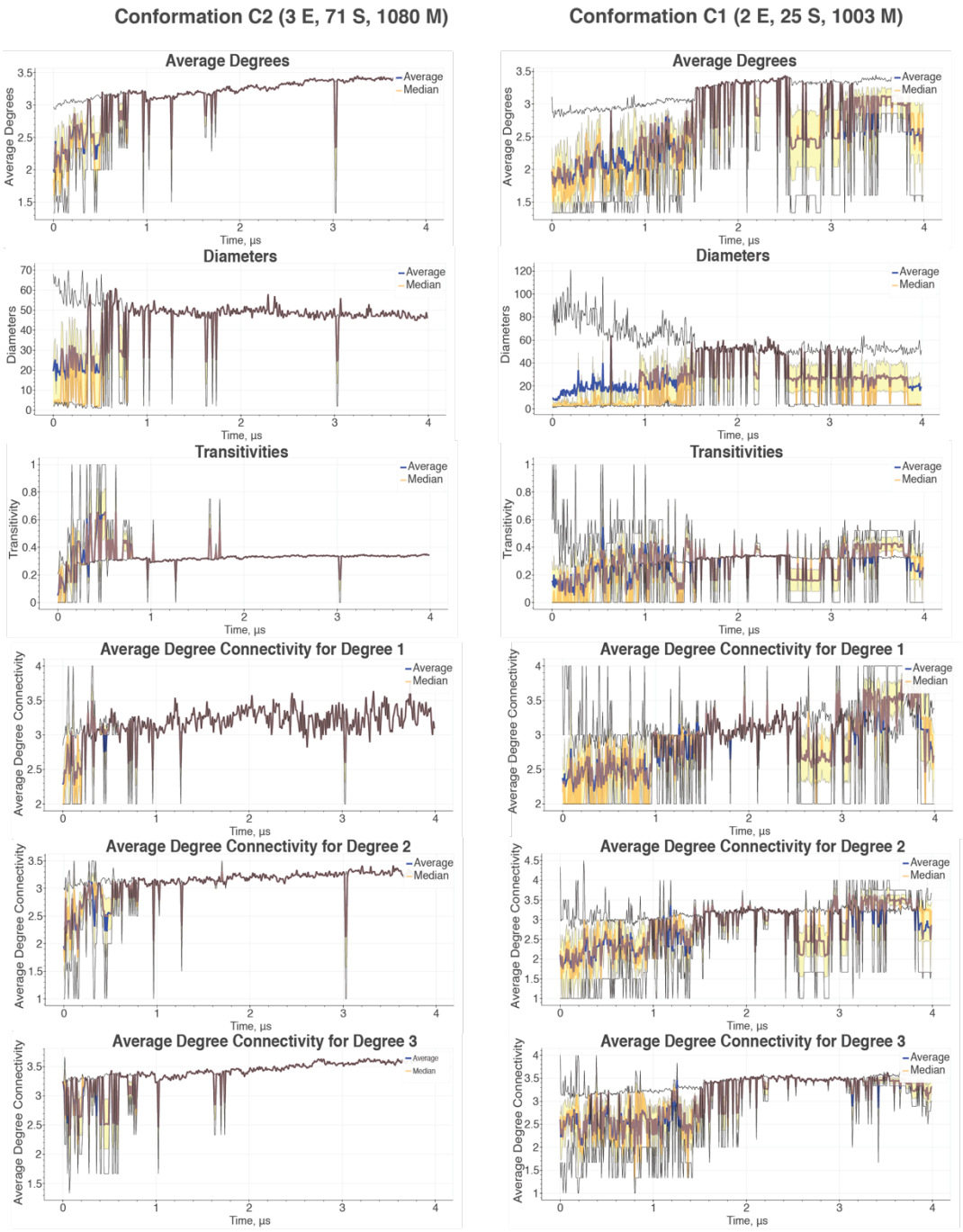

Figure B6. Dynamics of the basic network parameters for ED domain-domain interaction networks during the simulation of models M2 (conformation C2) and M1 (conformation C1). Yellow bands indicate values for the 2nd and 3rd quantiles, black lines denote the minimum and maximum values. ED domain-domain network is almost fully connected, and therefore has a much smaller number of connected components. The node degrees converge to the values of 2.5 for M2 and slightly above 3 for M1. Diameters for the connected components of M2 stays around 50, while for M2 it fluctuates around the value of 30. Transitivity value approaches 0.4, which seems to be a limit for the triangulated mesh.

Average degree connectivity suggests a much tighter coupling between ED domain-domain network compared with that one of TMD network.

Supplementary Tables

Proteins	Experimental				Computational			Comparison	
	PDB ID	Method	Release Date	Coverage (%)	Type	Release Date	Coverage (%)	All-Pair RMSD, Å	$\Delta T_{Release}$, days
Nsp1	7K3N	EM	9/30/20	100	Comparative	2/5/20	63	2.79	238
Nsp2	—	—	—	—	De novo	3/4/20	85	—	—
Nsp3	6WUU	X-Ray	5/20/20	16	Comparative	2/5/20	16	1.71	105
Nsp4	—	—	—	—	De novo	3/4/20	97	—	—
Nsp5	6LU7	X-Ray	2/5/20	100	Comparative	2/5/20	100	2.91	0
Nsp6	—	—	—	—	De novo	3/4/20	95	—	—
Nsp7	6M71	EM	4/1/20	100	Comparative	2/5/20	100	8.57	56
Nsp8	6M71	EM	4/1/20	100	Comparative	2/5/20	57	2.28	56
Nsp9	6W4B	X-Ray	3/18/20	100	Comparative	2/5/20	97	2.10	42
Nsp10	6W61	X-Ray	3/25/20	100	Comparative	2/5/20	87	2.46	49
Nsp12	6M71	EM	4/1/20	100	Comparative	2/5/20	82	1.55	56
Nsp13	6XEZ	EM	7/29/20	100	Comparative	2/5/20	99	5.39	175
Nsp14	—	—	—	—	Comparative	2/5/20	99	—	—
Nsp15	6VWW	X-Ray	3/4/20	100	Comparative	2/5/20	99	0.52	28
Nsp16	6W61	X-Ray	3/25/20	100	Comparative	2/5/20	96	1.37	49
E	7K3G	Solid NMR	9/30/20	41	Comparative	2/5/20	77	3.51	238
M	—	—	—	—	De novo	3/4/20	86	—	—
N-NTD	6VYO	X-Ray	3/11/20	30	Comparative	2/5/20	37	3.34	35
N-CTD	6YUN	X-Ray	5/20/20	28	Comparative	2/5/20	28	3.58	105
S	6VSB	EM	2/26/20	95	Comparative	2/5/20	87	6.81	21
Orf3a	6XDC	EM	6/17/20	100	De novo	3/4/20	71	7.89	105
Orf3b	—	—	—	—	—	—	—	—	—
Orf6	—	—	—	—	—	—	—	—	—
Orf7a	6W37	X-Ray	4/29/20	55	Comparative	2/5/20	78	0.97	84
Orf7b	—	—	—	—	—	—	—	—	—
Orf8	7JTL	X-Ray	8/26/20	85	De novo	1/31/20	100	16.92	113
Orf9b	7KDT	EM	10/21/20	100	—	—	—	—	—
Average				81.57			79.82	4.15	86.39

Table B1. Comparison between experimental and computational structures of individual SARS-CoV-2 proteins.

Protein complex	Experimental					Computational				Comparison	
	PDB ID	# of units	Method	Release Date	Coverage (%)	# of units	Type	Release Date	Coverage (%)	All-Pair RMSD, Å	$\Delta T_{Release}$, days

S trimer	6VSB	3	EM	02/26/20	95	3	Comparative	2/14/20	87.11	6.01	12
S-ACE2 tetramer	6VW1	4	X-Ray	03/04/20	18.35	4	Comparative	2/14/20	88.79	5.06	18
S-IGHG1/IGLL5 tetramer***	7BZ5	3	X-Ray	05/13/20	39.31	3	Comparative	2/14/20	34.29	—	—
S-IGHV3-30-3 dimer***	6YZ5	2	X-Ray	06/03/20	37.24	4	Comparative	2/14/20	28.41	—	—
Nsp3 domain2* dimer	6VXS	2	X-Ray	03/04/20	8.74	4	Comparative	2/14/20	8.63	15.22	18
Nnsp3 domain3 tetramer	—	—	—	—	—		Comparative	2/14/20	13.62	—	—
Nsp3-UBC dimer	6XAA	2	X-Ray	06/17/20	19.49	2	Comparative	2/14/20	19.29	0.98	111
Nsp4 dimer	—	—	—	—	—		Comparative	2/14/20	18.23	—	—
Nsp5 dimer	6Y2G	2	X-Ray	03/04/20	100	2	Comparative	2/14/20	100	1.19	18
Nsp7-Nsp8-Nsp12 tetramer	6M71	4	EM	04/01/20	100	4	Comparative	2/14/20	77.53	1.46	46
Nsp9 dimer	6W4B	2	X-ray	03/18/20	100	4	Comparative	2/14/20	97	2.11	32
Nsp10 dodecamer*	6W75	4	X-ray	03/25/20	100	12	Comparative	2/14/20	95.71	—	—
Nsp10-Nsp14 dimer	—	—	—	—	—	2	Comparative	2/14/20	99.09	—	—
Nsp10-Nsp16 dimer	6W61	2	X-ray	03/25/20	100	2	Comparative	2/14/20	94.05	1.69	39
Nsp13 dimer*	5RM6	2	X-ray	07/29/20	100	2	Comparative	2/14/20	99.16	40.20	165
N-Nterminal pentamer*	6VYO	4	X-Ray	03/11/20	100	5	Comparative	2/14/20	100	—	—
N-Cterminal dimer	6WJI	5	X-Ray	04/22/20	100	2	Comparative	2/14/20	100	2.42	67
E pentamer*	7K3G	5	Solid NMR	09/11/20	40	5	Comparative	2/14/20	77.31	—	—
Average					70.54 (74.16)				68.79 (72.24)	7.63	52.6
Average**					80.36				75.87	2.61	42.9

* - Protein complexes in different conformations or with the different number of base monomers

** - Statistics computed without complexes in different conformations and missing experimental structures

*** - Complexes of S protein with antibodies. Corresponding experimental structures were looked up based on antibody protein sequence from the computational model. As matches were not exact they are excluded from the analysis

Table B2. Comparison between experimental and computational structures of protein complexes that involve SARS CoV 2 proteins.

Element	N of Martini3 particles
Protein homo-oligomers	
S-TM trimer	459
S-truncated trimer	711
S-full trimer	3,822
E pentamer	880
M dimer	946
Lipids	
POPC	12
POPE	12
POPI	14
POPS	12
CHOL	8
CDL2	27

Table B3. Number of Martini3 particles per element. Lipid molecules include 1-palmitoyl-2-oleoylphosphatidylcholine (POPC), 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-ethanolamine (POPE), 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPI), 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (POPS), cholesterol (CHOL), and cardiolipin (CDL2). S-TM corresponds to the transmembrane domains of S trimer, S-truncated corresponds to the truncated model of S timer without its endodomains, S-full corresponds to a model of entire S trimer.

References

1. R. R. Schaller, Moore's law: past, present and future. *IEEE spectrum* **34**, 52-59 (1997).
2. G. O'Regan, "Integrated Circuit and Silicon Valley" in A Brief History of Computing. (Springer, 2021), pp. 89-96.
3. T. S. Perry, Move over, Moore's law. Make way for Huang's law [Spectral Lines]. *IEEE Spectrum* **55**, 7-7 (2018).
4. M. Barba, H. Czosnek, A. Hadidi, Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* **6**, 106-136 (2014).
5. K. K. Jain, Personalized medicine. *Current opinion in molecular therapeutics* **4**, 548-558 (2002).
6. L. P. Garrison Jr, M. F. Austin, Linking pharmacogenetics-based diagnostics and drugs for personalized medicine. *Health Affairs* **25**, 1281-1290 (2006).
7. H.-G. Xie, F. W. Frueh, Pharmacogenomics steps toward personalized medicine. (2005).
8. K. A. Wetterstrand (2021) DNA Sequencing Costs: Data.
9. N. R. Council, Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. (2011).
10. G. S. Ginsburg, K. A. Phillips, Precision medicine: from science to value. *Health Affairs* **37**, 694-701 (2018).
11. E. A. Ashley, The precision medicine initiative: a new national effort. *Jama* **313**, 2119-2120 (2015).
12. D. S. Singer, T. Jacks, E. Jaffee, A US "Cancer Moonshot" to accelerate cancer research. *Science* **353**, 1105-1106 (2016).
13. N. E. Sharpless, D. S. Singer, Progress and potential: the Cancer Moonshot. *Cancer cell* **39**, 889-894 (2021).
14. R. Lowe, N. Shirley, M. Bleackley, S. Dolan, T. Shafee, Transcriptomics technologies. *PLoS computational biology* **13**, e1005457 (2017).
15. A. Rao, D. Barkley, G. S. França, I. Yanai, Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211-220 (2021).
16. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**, 57-63 (2009).
17. W. C. Cho, Proteomics technologies and challenges. *Genomics, proteomics & bioinformatics* **5**, 77-85 (2007).
18. S. Hanash, Disease proteomics. *Nature* **422**, 226-232 (2003).
19. B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, M. H. Rasool, Proteomics: technologies and their applications. *Journal of chromatographic science* **55**, 182-196 (2017).
20. K. Luck, G. M. Sheynkman, I. Zhang, M. Vidal, Proteome-scale human interactomics. *Trends in biochemical sciences* **42**, 342-354 (2017).
21. M. Vidal, Interactome modeling. *FEBS letters* **579**, 1834-1838 (2005).
22. S. G. Choi, A. Richardson, L. Lambourne, D. E. Hill, M. Vidal, "Protein interactomics by two-hybrid methods" in Two-Hybrid Systems. (Springer, 2018), pp. 1-14.
23. Q. Zhong *et al.*, Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321 (2009).

24. B. Charloteaux *et al.*, "Protein–protein interactions and networks: forward and reverse edgetics" in Yeast Systems Biology. (Springer, 2011), pp. 197-213.
25. J. R. Idle, F. J. Gonzalez, Metabolomics. *Cell metabolism* **6**, 348-351 (2007).
26. A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Modern analytical techniques in metabolomics analysis. *Analyst* **137**, 293-300 (2012).
27. D. R. Schmidt *et al.*, Metabolomics in cancer research and emerging applications in clinical oncology. *CA: a cancer journal for clinicians* **71**, 333-358 (2021).
28. A. L. Non, Social epigenomics: are we at an impasse? *Epigenomics* **13**, 1747-1759 (2021).
29. P. A. Callinan, A. P. Feinberg, The emerging science of epigenomics. *Human molecular genetics* **15**, R95-R101 (2006).
30. D.-X. Zhou, Y. Hu, Y. Zhao, "Epigenomics" in Genetics and Genomics of Rice. (Springer, 2013), pp. 129-143.
31. M. Egert, A. A. De Graaf, H. Smidt, W. M. De Vos, K. Venema, Beyond diversity: functional microbiomics of the human colon. *Trends in microbiology* **14**, 86-91 (2006).
32. S. K. Shukla, N. S. Murali, M. H. Brilliant, Personalized medicine going precise: from genomics to microbiomics. *Trends in molecular medicine* **21**, 461 (2015).
33. B. H. Mullish *et al.*, Functional microbiomics: evaluation of gut microbiota-bile acid metabolism interactions in health and disease. *Methods* **149**, 49-58 (2018).
34. H. L. Gallagher, C. D. Frith, Functional imaging of 'theory of mind'. *Trends in cognitive sciences* **7**, 77-83 (2003).
35. M. Ingvar, Pain and functional imaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **354**, 1347-1358 (1999).
36. D. Lee, J. Gwak, T. Badloe, S. Palomba, J. Rho, Metasurfaces-based imaging and applications: from miniaturized optical components to functional imaging platforms. *Nanoscale Advances* **2**, 605-625 (2020).
37. V. S. Parekh, M. A. Jacobs, Deep learning and radiomics in precision medicine. *Expert review of precision medicine and drug development* **4**, 59-72 (2019).
38. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, H. J. Aerts, Artificial intelligence in radiology. *Nature Reviews Cancer* **18**, 500-510 (2018).
39. W. E. Brant, C. A. Helms, Fundamentals of diagnostic radiology. (2012).
40. W. S. Bush, J. H. Moore, Chapter 11: Genome-wide association studies. *PLoS computational biology* **8**, e1002822 (2012).
41. E. W. Demerath *et al.*, Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human molecular genetics* **24**, 4464-4479 (2015).
42. V. K. Rakyan, T. A. Down, D. J. Balding, S. Beck, Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* **12**, 529-541 (2011).
43. T. Hulsen *et al.*, From big data to precision medicine. *Frontiers in medicine*, 34 (2019).
44. J. Anuradha, A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science* **48**, 319-324 (2015).
45. I. The, T. P.-C. A. of Whole, G. Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82 (2020).
46. M. Gerstung *et al.*, The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128 (2020).

47. E. Rheinbay *et al.*, Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111 (2020).
48. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
49. Y. Li *et al.*, Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121 (2020).
50. C. Calabrese *et al.*, Genomic basis for RNA alterations in cancer. *Nature* **578**, 129-136 (2020).
51. S. Roy *et al.*, Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. *Archives of pathology & laboratory medicine* **140**, 958-975 (2016).
52. S. Roy, Molecular pathology informatics. *Surgical Pathology Clinics* **8**, 187-194 (2015).
53. I. W. P. Consortium, Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine* **360**, 753-764 (2009).
54. J. H. You, Pharmacogenetic-guided selection of warfarin versus novel oral anticoagulants for stroke prevention in patients with atrial fibrillation: a cost-effectiveness analysis. *Pharmacogenetics and genomics* **24**, 6-14 (2014).
55. S. A. Dugger, A. Platt, D. B. Goldstein, Drug development in the era of precision medicine. *Nature reviews Drug discovery* **17**, 183-196 (2018).
56. F. Stegmeier, M. Warmuth, W. Sellers, M. Dorsch, Targeted cancer therapies in the twenty-first century: lessons from imatinib. *Clinical Pharmacology & Therapeutics* **87**, 543-552 (2010).
57. D. J. Slamon *et al.*, Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England journal of medicine* **344**, 783-792 (2001).
58. D. Ho *et al.*, Enabling technologies for personalized and precision medicine. *Trends in biotechnology* **38**, 497-518 (2020).
59. X. Tian *et al.*, Wireless body sensor networks based on metamaterial textiles. *Nature Electronics* **2**, 243-251 (2019).
60. T. Djenizian, B. C. Tee, M. Ramuz, L. Fang (2019) Advances in flexible and soft electronics. (AIP Publishing LLC), p 031201.
61. Y. Zhao, A. Kim, G. Wan, B. C. Tee, Design and applications of stretchable and self-healable conductors for soft electronics. *Nano convergence* **6**, 1-22 (2019).
62. J. Kim *et al.*, Miniaturized battery-free wireless systems for wearable pulse oximetry. *Advanced functional materials* **27**, 1604373 (2017).
63. J. A. Rogers, Nanomesh on-skin electronics. *Nature nanotechnology* **12**, 839-840 (2017).
64. X. Yang *et al.*, Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805-817 (2016).
65. H. Keren, G. Lev-Maor, G. Ast, Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* **11**, 345 (2010).
66. J. Merkin, C. Russell, P. Chen, C. B. Burge, Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-1599 (2012).
67. H. Cui, A. Dhroso, N. Johnson, D. Korkin, The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods* **79**, 18-31 (2015).

68. R. Corominas *et al.*, Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature communications* **5**, 3650 (2014).
69. E. Lara-Pezzi, J. Gómez-Salinero, A. Gatto, P. García-Pavía, The alternative heart: impact of alternative splicing in heart disease. *Journal of cardiovascular translational research* **6**, 945-955 (2013).
70. O. Kelemen *et al.*, Function of alternative splicing. *Gene* **514**, 1-30 (2013).
71. S. Buonamici *et al.* (2016) H3B-8800, an orally bioavailable modulator of the SF3b complex, shows efficacy in spliceosome-mutant myeloid malignancies. (American Society of Hematology Washington, DC).
72. R. Sciarrillo *et al.*, Splicing modulation as novel therapeutic strategy against diffuse malignant peritoneal mesothelioma. *EBioMedicine* **39**, 215-225 (2019).
73. S. X. Lu *et al.*, Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. *Cell* **184**, 4032-4047. e4031 (2021).
74. K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).
75. J. Lonsdale *et al.*, The genotype-tissue expression (GTEx) project. *Nature genetics* **45**, 580-585 (2013).
76. M. Lathrop *et al.*, International network of cancer genome projects (The International Cancer Genome Consortium). *Nature Digest* **464**, 993-998 (2010).
77. F. Jiang *et al.*, Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* **2** (2017).
78. A. M. Turing, Mind. *Mind* **59**, 433-460 (1950).
79. C. S. Strachey (1952) Logical or non-mathematical programmes. in *Proceedings of the 1952 ACM national meeting (Toronto)*, pp 46-49.
80. F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65**, 386 (1958).
81. J. G. C. Devol (1961) Programmed article transfer. (Google Patents).
82. E. S. Brunette, R. C. Flemmer, C. L. Flemmer (2009) A review of artificial intelligence. in *2009 4th International Conference on Autonomous Robots and Agents* (Ieee), pp 385-392.
83. M. Oravec, P. Podhradsky, Medical image compression by backpropagation neural network and discrete orthogonal transforms. *WIT Transactions on Biomedicine and Health* **4** (1970).
84. A. T. Greenhill, B. R. Edmunds, A primer of artificial intelligence in medicine. *Techniques and Innovations in Gastrointestinal Endoscopy* **22**, 85-89 (2020).
85. D. Cohn, G. Tesauro, How tight are the Vapnik-Chervonenkis bounds? *Neural Computation* **4**, 249-269 (1992).
86. S. R. Sain (1996) The nature of statistical learning theory. (Taylor & Francis).
87. V. Vapnik, *The nature of statistical learning theory* (Springer science & business media, 1999).
88. T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009), vol. 2.
89. V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks* **17**, 113-126 (2004).

90. Y. Cao *et al.* (2006) Adapting ranking SVM to document retrieval. in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 186-193.
91. L. Breiman, Random forests. *Machine learning* **45**, 5-32 (2001).
92. S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **6**, 107-116 (1998).
93. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).
94. J. M. McGinnis, P. Williams-Russo, J. R. Knickman, The case for more active policy attention to health promotion. *Health affairs* **21**, 78-93 (2002).
95. P. J. Park, ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics* **10**, 669-680 (2009).
96. S. Ma, Y. Zhang, Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Molecular Biomedicine* **1**, 1-13 (2020).
97. J. D. Buenrostro, B. Wu, H. Y. Chang, W. J. Greenleaf, ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology* **109**, 21.29. 21-21.29. 29 (2015).
98. J.-M. Belton *et al.*, Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276 (2012).
99. V. K. Tiwari, *Carbohydrates in Drug Discovery and Development: Synthesis and Application* (Elsevier, 2020).
100. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Structure, Function, and Bioinformatics* **82**, 1-6 (2014).
101. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics* **86**, 7-15 (2018).
102. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* **87**, 1011-1020 (2019).
103. B. Chakrabarty, D. Das, G. Bulusu, A. Roy, Network-based analysis of fatal comorbidities of COVID-19 and potential therapeutics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021).
104. M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, J. R. Green, A review of network-based approaches to drug repositioning. *Briefings in bioinformatics* **19**, 878-892 (2018).
105. K. Glass, C. Huttenhower, J. Quackenbush, G.-C. Yuan, Passing messages between biological networks to refine predicted interactions. *PloS one* **8**, e64832 (2013).
106. D. Schlauch, K. Glass, C. P. Hersh, E. K. Silverman, J. Quackenbush, Estimating drivers of cell state transitions using gene regulatory network models. *BMC systems biology* **11**, 1-10 (2017).

107. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, 1097-1105 (2012).
108. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
109. L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681-694 (2020).
110. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *nature* **521**, 436-444 (2015).
111. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35**, 1798-1828 (2013).
112. A. Kolesnikov, X. Zhai, L. Beyer (2019) Revisiting self-supervised visual representation learning. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1920-1929.
113. G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *science* **313**, 504-507 (2006).
114. J. Ngiam *et al.* (2011) Multimodal deep learning. in *ICML*.
115. D. Bhaskar, J. D. Grady, M. A. Perlmutter, S. Krishnaswamy, Molecular Graph Generation via Geometric Scattering. *arXiv preprint arXiv:2110.06241* (2021).
116. A. Vaswani *et al.* (2017) Attention is all you need. in *Advances in neural information processing systems*, pp 5998-6008.
117. M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile (2019) Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. in *6th Italian Conference on Computational Linguistics, CLiC-it 2019* (CEUR), pp 1-6.
118. Z. Lan *et al.*, Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
119. J.-H. Oh, R. Iida, J. Kloetzer, K. Torisawa (2021) BERTAC: Enhancing Transformer-based Language Models with Adversarially Pretrained Convolutional Neural Networks. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp 2103-2115.
120. J. Lee *et al.*, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234-1240 (2020).
121. M. Zhang, Z. Cui, M. Neumann, Y. Chen (2018) An end-to-end deep learning architecture for graph classification. in *Thirty-Second AAAI Conference on Artificial Intelligence*.
122. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nature reviews genetics* **12**, 56-68 (2011).
123. B. Zhang, C. Quan, F. Ren (2016) Study on CNN in the recognition of emotion in audio and images. in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (IEEE), pp 1-5.
124. S. A. Siddiqui *et al.*, Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science* **75**, 374-389 (2018).

125. P. Sermanet, S. Chintala, Y. LeCun (2012) Convolutional neural networks applied to house numbers digit classification. in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (IEEE), pp 3288-3291.
126. Y. Wang *et al.*, A Depthwise Separable Fully Convolutional ResNet With ConvCRF for Semisupervised Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 4621-4632 (2021).
127. M. Pak, S. Kim (2017) A review of deep learning in image recognition. in *2017 4th international conference on computer applications and information processing technology (CAIPT)* (IEEE), pp 1-3.
128. A. Shrestha, A. Mahmood, Review of deep learning algorithms and architectures. *IEEE Access* **7**, 53040-53065 (2019).
129. D. Xu *et al.*, Infrared and visible image fusion with a generative adversarial network and a residual network. *Applied Sciences* **10**, 554 (2020).
130. A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, F. Makedon, A survey on contrastive self-supervised learning. *Technologies* **9**, 2 (2021).
131. P. Baldi (2012) Autoencoders, unsupervised learning, and deep architectures. in *Proceedings of ICML workshop on unsupervised and transfer learning* (JMLR Workshop and Conference Proceedings), pp 37-49.
132. C. Doersch, Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
133. T. Han, D. Jiang, Q. Zhao, L. Wang, K. Yin, Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control* **40**, 2681-2693 (2018).
134. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators. *Neural networks* **2**, 359-366 (1989).
135. I. Sutskever, O. Vinyals, Q. V. Le (2014) Sequence to sequence learning with neural networks. in *Advances in neural information processing systems*, pp 3104-3112.
136. N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, K. M. Borgwardt, Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research* **12** (2011).
137. J. Godwin *et al.*, Very Deep Graph Neural Networks Via Noise Regularisation. *arXiv preprint arXiv:2106.07971* (2021).
138. L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017).
139. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning. *Journal of Big Data* **6**, 1-48 (2019).
140. A. Mikołajczyk, M. Grochowski (2018) Data augmentation for improving deep learning in image classification problem. in *2018 international interdisciplinary PhD workshop (IIPhDW)* (IEEE), pp 117-122.
141. S. Ioffe, C. Szegedy (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *International conference on machine learning* (PMLR), pp 448-456.
142. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345-1359 (2009).

143. H. Malik *et al.*, A Comparison of Transfer Learning Performance Versus Health Experts in Disease Diagnosis From Medical Imaging. *IEEE Access* **8**, 139367-139386 (2020).
144. M. M. Palatucci, D. A. Pomerleau, G. E. Hinton, T. Mitchell, Zero-shot learning with semantic output codes. (2009).
145. C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, S. Jegelka, Debiased contrastive learning. *arXiv preprint arXiv:2007.00224* (2020).
146. T. Chen, S. Kornblith, M. Norouzi, G. Hinton (2020) A simple framework for contrastive learning of visual representations. in *International conference on machine learning* (PMLR), pp 1597-1607.
147. I. R. Kondor, *Group theoretical methods in machine learning* (Columbia University, 2008).
148. G.-W. Wei, Protein structure prediction beyond AlphaFold. *Nature Machine Intelligence* **1**, 336-337 (2019).
149. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
150. M. AlQuraishi, AlphaFold at CASP13. *Bioinformatics* **35**, 4862-4865 (2019).
151. A. Perrakis, T. K. Sixma, AI revolutions in biology: The joys and perils of AlphaFold. *EMBO reports* **22**, e54046 (2021).
152. J. Tang, F. Wang, F. Cheng (2021) Artificial Intelligence for Drug Discovery. in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp 4074-4075.
153. C. Shi *et al.*, Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382* (2020).
154. S. Tang, B. Li, H. Yu, Chebnet: Efficient and stable constructions of deep neural networks with rectified power units using Chebyshev approximations. *arXiv preprint arXiv:1911.05467* (2019).
155. F.-Y. Sun, J. Hoffmann, V. Verma, J. Tang, Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000* (2019).
156. S. Mo *et al.*, Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. *arXiv preprint arXiv:2110.05231* (2021).
157. G. Carleo *et al.*, Machine learning and the physical sciences. *Reviews of Modern Physics* **91**, 045002 (2019).
158. A. Radovic *et al.*, Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560**, 41-48 (2018).
159. O. A. von Lilienfeld, K. Burke, Retrospective on a decade of machine learning for chemical discovery. *Nature communications* **11**, 1-4 (2020).
160. T. J. Keyes, P. Domizi, Y. C. Lo, G. P. Nolan, K. L. Davis, A cancer biologist's primer on machine learning applications in high-dimensional cytometry. *Cytometry Part A* **97**, 782-799 (2020).
161. M. L. De Prado, *Advances in financial machine learning* (John Wiley & Sons, 2018).
162. K. C. A. Khanzode, R. D. Sarode, Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review. *International Journal of Library & Information Science (IJLIS)* **9**, 3 (2020).

163. T. Wuest, D. Weimer, C. Irgens, K.-D. Thoben, Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research* **4**, 23-45 (2016).
164. Y. Park, E. M. J. N. m. Marcotte, Flaws in evaluation schemes for pair-input computational predictions. **9**, 1134 (2012).
165. T. V. Riepe, M. Khan, S. Roosing, F. P. Cremers, P. A. 't Hoen, Benchmarking deep learning splice prediction tools using functional splice assays. *Human mutation* **42**, 799-810 (2021).
166. N. Kilbertus, Beyond traditional assumptions in fair machine learning. *arXiv preprint arXiv:2101.12476* (2021).
167. R. Balestriero, J. Pesenti, Y. LeCun, Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485* (2021).
168. F. G. Cozman, I. Cohen, M. C. Cirelo (2003) Semi-supervised learning of mixture models. in *ICML*, p 24.
169. L. Bruzzone, M. Chi, M. Marconcini, A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **44**, 3363-3373 (2006).
170. J. E. Van Engelen, H. H. Hoos, A survey on semi-supervised learning. *Machine Learning* **109**, 373-440 (2020).
171. V. M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine* **32**, 53-69 (2015).
172. S. Sun, H. Shi, Y. Wu, A survey of multi-source domain adaptation. *Information Fusion* **24**, 84-92 (2015).
173. W. M. Kouw, M. Loog, A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence* (2019).
174. B. Mieth *et al.*, Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Scientific reports* **9**, 1-14 (2019).
175. J. Hu *et al.*, Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nature machine intelligence* **2**, 607-618 (2020).
176. Y. Lieberman, L. Rokach, T. Shay, CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PloS one* **13**, e0205499 (2018).
177. X. Zhou, H. Chai, H. Zhao, C.-H. Luo, Y. Yang, Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network. *GigaScience* **9**, giaa076 (2020).
178. B. Kang *et al.*, "Online Single-cell RNA-seq Data Denoising with Transfer Learning" in Practice and Experience in Advanced Research Computing. (2020), pp. 469-472.
179. T. Wang *et al.*, BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome biology* **20**, 1-15 (2019).
180. J. J. Bird, J. Kobylarz, D. R. Faria, A. Ekárt, E. P. Ribeiro, Cross-domain mlp and cnn transfer learning for biological signal processing: eeg and emg. *IEEE Access* **8**, 54789-54801 (2020).

181. Y. Lin *et al.*, scJoint: transfer learning for data integration of single-cell RNA-seq and ATAC-seq. *bioRxiv*, 2020.2012. 2031.424916 (2021).
182. G. A. Tadesse *et al.* (2019) Cardiovascular disease diagnosis using cross-domain transfer learning. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE), pp 4262-4265.
183. Q. Wang, H. Wang, L. Wang, F. Yu, Diagnosis of chronic obstructive pulmonary disease based on transfer learning. *IEEE Access* **8**, 47370-47383 (2020).
184. B. Cheng, M. Liu, D. Shen, Z. Li, D. Zhang, Multi-domain transfer learning for early diagnosis of Alzheimer's disease. *Neuroinformatics* **15**, 115-132 (2017).
185. A. Mehmood *et al.*, A transfer learning approach for early diagnosis of alzheimer's disease on MRI images. *Neuroscience* **460**, 43-52 (2021).
186. R. Colbaugh, K. Glass, G. Gallegos (2017) Ensemble transfer learning for Alzheimer's disease diagnosis. in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE), pp 3102-3105.
187. G. López-García, J. M. Jerez, L. Franco, F. J. Veredas, Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PloS one* **15**, e0230536 (2020).
188. J. Chen *et al.*, Deep Transfer Learning of Drug Sensitivity by Integrating Bulk and Single-cell RNA-seq data. *bioRxiv* (2021).
189. B. Schmauch *et al.*, A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature communications* **11**, 1-15 (2020).
190. G. L. Stein-O'Brien *et al.*, Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell systems* **8**, 395-411. e398 (2019).
191. P. Wang, B. Yan, J.-t. Guo, C. Hicks, Y. Xu, Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proceedings of the National Academy of Sciences* **102**, 18920-18925 (2005).
192. M. Buljan *et al.*, Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell* **46**, 871-883 (2012).
193. H. G. Sequencing, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).
194. W. Jiang, L. Chen, Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal* **19**, 183-195 (2021).
195. K. Luck *et al.*, A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
196. M. Li, Y. Guo, Y.-M. Feng, N. Zhang, Identification of triple-negative breast cancer genes and a novel high-risk breast cancer prediction model development based on PPI data and support vector machines. *Frontiers in genetics* **10**, 180 (2019).
197. Z. Li *et al.*, The OncoPPi network of cancer-focused protein–protein interactions to inform biological insights and therapeutic strategies. *Nature communications* **8**, 1-14 (2017).
198. M. Farahani, M. Rezaei-Tavirani, A. Zali, M. Zamanian-Azodi, Systematic Analysis of Protein–Protein and Gene–Environment Interactions to Decipher the Cognitive

Mechanisms of Autism Spectrum Disorder. *Cellular and Molecular Neurobiology*, 1-13 (2020).
199. A. E. Apostolakou, X. K. Sula, K. C. Nastou, G. I. Nasi, V. A. Iconomidou, Exploring the conservation of Alzheimer-related pathways between H. sapiens and C. elegans: a network alignment approach. *Scientific reports* **11**, 1-11 (2021).
200. J. Peng, Y. Zhou, K. Wang, Multiplex gene and phenotype network to characterize shared genetic pathways of epilepsy and autism. *Scientific reports* **11**, 1-16 (2021).
201. M. A. Skinnider *et al.*, An atlas of protein-protein interactions across mouse tissues. *Cell* **184**, 4073-4089. e4017 (2021).
202. A. Bergamo, L. Torresani, Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *Advances in neural information processing systems* **23**, 181-189 (2010).
203. H. Zhao *et al.*, Adversarial multiple source domain adaptation. *Advances in neural information processing systems* **31**, 8559-8570 (2018).
204. L. T. Triess, M. Dreissig, C. B. Rist, J. M. Zöllner, A Survey on Deep Domain Adaptation for LiDAR Perception. *arXiv preprint arXiv:2106.02377* (2021).
205. M. Toldo, A. Maracani, U. Michieli, P. Zanuttigh, Unsupervised domain adaptation in semantic segmentation: a review. *Technologies* **8**, 35 (2020).
206. A. Storkey, When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning* **30**, 3-28 (2009).
207. W. Fan, I. Davidson (2007) On sample selection bias and its efficient correction via model averaging and unlabeled examples. in *Proceedings of the 2007 SIAM International Conference on Data Mining* (SIAM), pp 320-331.
208. A. Axelrod, X. He, J. Gao (2011) Domain adaptation via pseudo in-domain data selection. in *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp 355-362.
209. S. Ruder, B. Plank, Learning to select data for transfer learning with bayesian optimization. *arXiv preprint arXiv:1707.05246* (2017).
210. D. Guo *et al.*, Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366* (2020).
211. A. Ramponi, B. Plank, Neural Unsupervised Domain Adaptation in NLP---A Survey. *arXiv preprint arXiv:2006.00632* (2020).
212. L. Luo, L. Chen, S. Hu, Y. Lu, X. Wang, Discriminative and geometry-aware unsupervised domain adaptation. *IEEE transactions on cybernetics* **50**, 3914-3927 (2020).
213. Y. Zhang *et al.*, Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomedical engineering online* **16**, 125 (2017).
214. S. Mourragui, M. Loog, M. A. Van De Wiel, M. J. Reinders, L. F. Wessels, PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510-i519 (2019).
215. L. Handl, A. Jalali, M. Scherer, R. Eggeling, N. Pfeifer, Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics* **35**, i154-i163 (2019).

216. J. Howard, S. Ruder, Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).
217. J. Blitzer, R. McDonald, F. Pereira (2006) Domain adaptation with structural correspondence learning. in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp 120-128.
218. S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, Z. Chen (2010) Cross-domain sentiment classification via spectral feature alignment. in *Proceedings of the 19th international conference on World wide web*, pp 751-760.
219. R. Gopalan, R. Li, R. Chellappa (2011) Domain adaptation for object recognition: An unsupervised approach. in *2011 international conference on computer vision* (IEEE), pp 999-1006.
220. B. Gong, Y. Shi, F. Sha, K. Grauman (2012) Geodesic flow kernel for unsupervised domain adaptation. in *2012 IEEE conference on computer vision and pattern recognition* (IEEE), pp 2066-2073.
221. K. A. Gallivan, A. Srivastava, X. Liu, P. Van Dooren (2003) Efficient algorithms for inferences on grassmann manifolds. in *IEEE Workshop on Statistical Signal Processing, 2003* (IEEE), pp 315-318.
222. A. M. Martinez, A. C. Kak, Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence* **23**, 228-233 (2001).
223. J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction. *IEEE transactions on image processing* **27**, 2420-2432 (2018).
224. A. Gretton *et al.*, Covariate shift by kernel mean matching. *Dataset shift in machine learning* **3**, 5 (2009).
225. A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, A. J. Smola (2007) A kernel approach to comparing distributions. in *Proceedings of the National Conference on Artificial Intelligence* (Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999), p 1637.
226. H. Callaert, N. Veraverbeke, The order of the normal approximation for a studentized U-statistic. *The Annals of Statistics*, 194-200 (1981).
227. S. Ben-David *et al.*, A theory of learning from different domains. *Machine learning* **79**, 151-175 (2010).
228. Y. Ganin, V. Lempitsky (2015) Unsupervised domain adaptation by backpropagation. in *International conference on machine learning* (PMLR), pp 1180-1189.
229. I. Goodfellow *et al.*, Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).
230. H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* **90**, 227-244 (2000).
231. Y. Ganin *et al.*, Domain-adversarial training of neural networks. *The journal of machine learning research* **17**, 2096-2030 (2016).
232. J. Shen, Y. Qu, W. Zhang, Y. Yu (2018) Wasserstein distance guided representation learning for domain adaptation. in *Thirty-Second AAAI Conference on Artificial Intelligence*.

233. D. McClosky, E. Charniak, M. Johnson (2006) Effective self-training for parsing. in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp 152-159.
234. Y. Zou, Z. Yu, X. Liu, B. Kumar, J. Wang (2019) Confidence regularized self-training. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp 5982-5991.
235. M. Chen, K. Q. Weinberger, J. Blitzer (2011) Co-Training for Domain Adaptation. in *Nips* (Citeseer), pp 2456-2464.
236. O. Narykov, N. T. Johnson, D. Korkin, Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. *Cell reports* **37**, 110045 (2021).
237. S. Lu, Z. Lu, Y.-D. Zhang, Pathological brain detection based on AlexNet and transfer learning. *Journal of computational science* **30**, 41-47 (2019).
238. H. Chen *et al.* (2021) Pre-trained image processing transformer. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 12299-12310.
239. A. Farahani, S. Voghoei, K. Rasheed, H. R. Arabnia, A brief review of domain adaptation. *Advances in Data Science and Information Engineering*, 877-894 (2021).
240. A. Søgaard, Y. Goldberg (2016) Deep multi-task learning with low level tasks supervised at lower layers. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp 231-235.
241. J. Zhang, W. Li, P. Ogunbona, Unsupervised domain adaptation: A multi-task learning-based method. *Knowledge-Based Systems* **186**, 104975 (2019).
242. F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* **77**, 443-477 (2008).
243. X. Wang *et al.*, Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Molecular & Cellular Proteomics*, mcp. RA117. 000155 (2017).
244. Z. Hu *et al.*, Revealing missing human protein isoforms based on ab initio prediction, RNA-seq and proteomics. *Scientific reports* **5**, 10940 (2015).
245. A.-C. Gavin *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141 (2002).
246. J.-F. Rual *et al.*, Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173 (2005).
247. X. Kuang, A. Dhroso, J. G. Han, C.-R. Shyu, D. Korkin, DOMMINO 2.0: integrating structurally resolved protein-, RNA-, and DNA-mediated macromolecular interactions. *Database* **2016** (2016).
248. H. M. Berman *et al.*, "The protein data bank, 1999–" in International Tables for Crystallography Volume F: Crystallography of biological macromolecules. (Springer, 2006), pp. 675-684.
249. A. Stein, R. B. Russell, P. Aloy, 3did: interacting protein domains of known three-dimensional structure. *Nucleic acids research* **33**, D413-D417 (2005).
250. N. Zhao, B. Pang, C. R. Shyu, D. Korkin, Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. *Protein Science* **20**, 1275-1284 (2011).

251. J. Andreani, G. Faure, R. Guerois, Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS computational biology* **8**, e1002677 (2012).
252. T. Rolland *et al.*, A proteome-scale map of the human interactome network. *Cell* **159**, 1212-1226 (2014).
253. A.-L. Barabasi, Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**, 101 (2004).
254. K. Mitra, A.-R. Carvunis, S. K. Ramesh, T. Ideker, Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics* **14**, 719 (2013).
255. M. Vidal (2016) How much of the human protein interactome remains to be mapped? (American Association for the Advancement of Science).
256. M. P. Stumpf *et al.*, Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences* **105**, 6959-6964 (2008).
257. N. Zhao, J. G. Han, C.-R. Shyu, D. Korkin, Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS computational biology* **10**, e1003592 (2014).
258. A. Singh *et al.*, MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic acids research* **36**, D815-D819 (2007).
259. K. Kessenbrock, Q. Nguyen, N. Pervolarakis, K. Nee, Experimental Considerations for Single Cell RNA Sequencing Approaches. *Frontiers in cell and developmental biology* **6**, 108 (2018).
260. A. Haque, J. Engel, S. A. Teichmann, T. Lönnberg, A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9**, 75 (2017).
261. C. Ziegenhain *et al.*, Comparative analysis of single-cell RNA sequencing methods. *Molecular cell* **65**, 631-643. e634 (2017).
262. A. Byrne *et al.*, Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature communications* **8**, 1-11 (2017).
263. K. Karlsson, S. Linnarsson, Single-cell mRNA isoform diversity in the mouse brain. *BMC genomics* **18**, 1-11 (2017).
264. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
265. E. Z. Macosko *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
266. Á. Arzalluz-Luque, A. Conesa, Single-cell RNAseq for the study of isoforms—how is that possible? *Genome biology* **19**, 110 (2018).
267. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360 (2015).
268. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
269. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

270. P. E. Compeau, P. A. Pevzner, G. Tesler, How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* **29**, 987-991 (2011).
271. Y. Zhang, X. Liu, J. N. MacLeod, J. Liu (2016) DeepSplice: Deep classification of novel splice junctions revealed by RNA-seq. in *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)* (IEEE), pp 330-333.
272. M. F. Rogers, J. Thomas, A. S. Reddy, A. Ben-Hur, SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome biology* **13**, 1-17 (2012).
273. P. Bonizzoni, R. Rizzi, G. Pesole, ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. *BMC bioinformatics* **6**, 1-16 (2005).
274. P. L. Martelli *et al.*, ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic acids research* **39**, D80-D85 (2010).
275. A. Moles-Fernández *et al.*, Computational tools for splicing defect prediction in breast/ovarian cancer genes: how efficient are they at predicting RNA alterations? *Frontiers in genetics* **9**, 366 (2018).
276. S. Degroeve, Y. Saeys, B. De Baets, P. Rouzé, Y. Van de Peer, SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332-1338 (2005).
277. D. Wei, H. Zhang, Y. Wei, Q. Jiang, A novel splice site prediction method using support vector machine. *Journal of Computational Information Systems* **9**, 8053-8060 (2013).
278. A. Stanescu, D. Caragea (2014) Ensemble-based semi-supervised learning approaches for imbalanced splice site datasets. in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE), pp 432-437.
279. X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human mutation* **37**, 235-241 (2016).
280. M. Stanke *et al.*, AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435-W439 (2006).
281. X. Jian, E. Boerwinkle, X. Liu, In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genetics in Medicine* **16**, 497-503 (2014).
282. K. Jaganathan *et al.*, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535-548. e524 (2019).
283. A. Dutta, T. Dubey, K. K. Singh, A. Anand, SpliceVec: distributed feature representations for splice junction prediction. *Computational biology and chemistry* **74**, 434-441 (2018).
284. R. Liu, A. E. Loraine, J. A. Dickerson, Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC bioinformatics* **15**, 1-16 (2014).
285. Y. Li, X. Rao, W. W. Mattox, C. I. Amos, B. Liu, RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. *PloS one* **10**, e0136653 (2015).
286. S. Shen *et al.*, MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic acids research* **40**, e61-e61 (2012).
287. W. Wang, Z. Qin, Z. Feng, X. Wang, X. Zhang, Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* **518**, 164-170 (2013).

288. Z. Zhang *et al.*, Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature methods* **16**, 307-310 (2019).
289. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).
290. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 1-16 (2011).
291. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).
292. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14**, 417-419 (2017).
293. J. D. Welch, Y. Hu, J. F. Prins, Robust detection of alternative splicing in a population of single cells. *Nucleic acids research* **44**, e73-e73 (2016).
294. Y. Song *et al.*, Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular cell* **67**, 148-161. e145 (2017).
295. Y. Huang, G. Sanguinetti, BRIE: transcriptome-wide splicing quantification in single cells. *Genome biology* **18**, 123 (2017).
296. J. Gilis, K. Vitting-Seerup, K. Van den Berge, L. Clement, satuRn: Scalable Analysis of differential Transcript Usage for bulk and single-cell RNA-sequencing applications. *bioRxiv* (2021).
297. Y. Hu, K. Wang, M. Li, Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLOS Computational Biology* **16**, e1007925 (2020).
298. J. Westoby, M. S. Herrera, A. C. Ferguson-Smith, M. Hemberg, Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome biology* **19**, 1-14 (2018).
299. J. Westoby, P. Artemov, M. Hemberg, A. Ferguson-Smith, Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome biology* **21**, 1-19 (2020).
300. C. F. B. A. Najar, N. Yosef, L. F. Lareau, Coverage-dependent bias creates the appearance of binary splicing in single cells. *Elife* **9**, e54603 (2020).
301. G. K. Marinov *et al.*, From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome research* **24**, 496-510 (2014).
302. F. Buettner *et al.*, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* **33**, 155 (2015).
303. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381 (2014).
304. L. Haghverdi, A. T. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421 (2018).
305. J. Westoby (2020) Alternative splicing and single-cell RNA-sequencing: a feasibility assessment. (University of Cambridge).
306. M. Zhao *et al.*, DNA methylation and mRNA and microRNA expression of SLE CD4+ T cells correlate with disease phenotype. *Journal of autoimmunity* **54**, 127-136 (2014).

307. J. M. Taliaferro *et al.*, Distal alternative last exons localize mRNAs to neural projections. *Molecular cell* **61**, 821-833 (2016).
308. L. Li *et al.*, Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell stem cell* **20**, 858-873. e854 (2017).
309. R. K. Bradley, J. Merkin, N. J. Lambert, C. B. Burge, Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS biology* **10** (2012).
310. B. Taneri, B. Snyder, A. Novoradovsky, T. Gaasterland, Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biology* **5**, R75 (2004).
311. D. Lipscombe, E. J. L. Soto, Alternative splicing of neuronal genes: new mechanisms and new therapies. *Current opinion in neurobiology* **57**, 26-31 (2019).
312. I. Meininger *et al.*, Alternative splicing of MALT1 controls signalling and activation of CD4+ T cells. *Nature communications* **7**, 1-15 (2016).
313. J. Fang, A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings in bioinformatics* (2019).
314. M. Koohi-Moghadam *et al.*, Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nature Machine Intelligence* **1**, 561-567 (2019).
315. A. Kulandaisamy, J. Zaucha, R. Sakthivel, D. Frishman, M. Michael Gromiha, Pred-MutHTP: Prediction of disease-causing and neutral mutations in human transmembrane proteins. *Human Mutation* (2019).
316. O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks* **20**, 542-542 (2009).
317. M. A. Ghadie, L. Lambourne, M. Vidal, Y. Xia, Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS computational biology* **13**, e1005717 (2017).
318. J. Zeng, G. Yu, J. Wang, M. Guo, X. Zhang (2019) DMIL-III: Isoform-isoform interaction prediction using deep multi-instance learning method. in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE), pp 171-176.
319. X.-Y. Pan, Y.-N. Zhang, H.-B. Shen, Large-Scale prediction of human protein− protein interactions from amino acid sequence based on latent topic features. *Journal of Proteome Research* **9**, 4992-5001 (2010).
320. K. Venkatesan *et al.*, An empirical framework for binary interactome mapping. *Nature methods* **6**, 83 (2009).
321. H. Yu *et al.*, Next-generation sequencing to generate interactome datasets. *Nature methods* **8**, 478 (2011).
322. Q. Zhong *et al.*, An inter-species protein–protein interaction network across vast evolutionary distance. *Molecular systems biology* **12**, 865 (2016).
323. M. M. Hussain, I. Mahmud, pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software* **4**, 1556 (2019).
324. H. B. Mann, Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, 245-259 (1945).
325. M. G. Kendall, Rank correlation methods. (1948).

326. L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
327. D. R. Zerbino *et al.*, Ensembl 2018. *Nucleic Acids Research* **46**, D754-D761 (2018).
328. J. De Fauw *et al.*, Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* **24**, 1342-1350 (2018).
329. J. Wang, L. Perez, The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11 (2017).
330. J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* **24**, 279-283 (2017).
331. H. C. Jubb *et al.*, Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology* **128**, 3-13 (2017).
332. X. Yang *et al.*, Potential role of Hsp90 in rat islet function under the condition of high glucose. **53**, 621-628 (2016).
333. N. K. Fox, S. E. Brenner, J.-M. Chandonia, SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research* **42**, D304-D309 (2014).
334. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
335. S. Subhash, C. Kanduri, GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. *BMC bioinformatics* **17**, 1-10 (2016).
336. V. Perovic *et al.*, TRI_tool: a web-tool for prediction of protein–protein interactions in human transcriptional regulation. *Bioinformatics* **33**, 289-291 (2017).
337. X. Kuang *et al.*, DOMMINO: a database of macromolecular interactions. *Nucleic acids research* **40**, D501-D506 (2011).
338. D. Wilson, M. Madera, C. Vogel, C. Chothia, J. Gough, The SUPERFAMILY database in 2007: families and functions. *Nucleic acids research* **35**, D308-D313 (2006).
339. V. Vapnik, R. Izmailov, Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.* **16**, 2023-2049 (2015).
340. M. Lapin, M. Hein, B. Schiele, Learning using privileged information: SVM+ and weighted SVM. *Neural Networks* **53**, 95-108 (2014).
341. N. Gauraha, L. Carlsson, O. Spjuth (2018) Conformal prediction in learning under privileged information paradigm with applications in drug discovery. in *Conformal and Probabilistic Prediction and Applications* (PMLR), pp 147-156.
342. W. Li, D. Dai, M. Tan, D. Xu, L. Van Gool (2016) Fast algorithms for linear and kernel svm+. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 2258-2266.
343. W. A. Abbasi, A. Asif, A. Ben-Hur, Learning protein binding affinity using privileged information. *BMC bioinformatics* **19**, 1-12 (2018).
344. J. Lambert, O. Sener, S. Savarese (2018) Deep learning under privileged information using heteroscedastic dropout. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 8886-8895.
345. Y. Chen, X. Jin, J. Feng, S. Yan, Training group orthogonal neural networks with privileged information. *arXiv preprint arXiv:1701.06772* (2017).

346. Z. Gao *et al.*, Learning the implicit strain reconstruction in ultrasound elastography using privileged information. *Medical image analysis* **58**, 101534 (2019).
347. J. Delgado, L. G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).
348. E. Verschueren, P. Vanhee, F. Rousseau, J. Schymkowitz, L. Serrano, Protein-peptide complex prediction through fragment interaction patterns. *Structure* **21**, 789-797 (2013).
349. M. Lu, A. D. Dousis, J. Ma, OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of molecular biology* **376**, 288-301 (2008).
350. H. Zhou, J. Skolnick, GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal* **101**, 2043-2052 (2011).
351. S. Hubber, J. Thornton, NACCESS computer program. *Department of Biochemistry and Molecular Biology, University College, London* (1993).
352. X. Li, J. Liang, Geometric packing potential function for model selection in protein structure and protein-protein binding predictions. (2012).
353. Y. Yang, Y. Zhou, Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein science* **17**, 1212-1219 (2008).
354. A. Amos-Binks *et al.*, Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC bioinformatics* **12**, 1-13 (2011).
355. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2**, 27 (2011).
356. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).
357. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **55**, 119-139 (1997).
358. Y. Freund, R. Schapire, N. Abe, A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* **14**, 1612 (1999).
359. X. Xie, S. Wu, K.-M. Lam, H. Yan, PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. *Bioinformatics* **22**, 2722-2728 (2006).
360. L. Zhong *et al.*, Effective classification of microRNA precursors using feature mining and AdaBoost algorithms. *Omics: a journal of integrative biology* **17**, 486-493 (2013).
361. B. Niu, Y.-D. Cai, W.-C. Lu, G.-Z. Li, K.-C. Chou, Predicting protein structural class with AdaBoost learner. *Protein and peptide letters* **13**, 489-492 (2006).
362. R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, J. H. Moore (2018) Data-driven advice for applying machine learning to bioinformatics problems. in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* (World Scientific), pp 192-203.
363. A. Verma, S. Mehta (2017) A comparative study of ensemble learning methods for classification in bioinformatics. in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence* (IEEE), pp 155-158.

364. P. Yang, Y. Hwa Yang, B. B Zhou, A. Y Zomaya, A review of ensemble methods in bioinformatics. *Current Bioinformatics* **5**, 296-308 (2010).
365. X. Li, L. Wang, E. Sung, AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence* **21**, 785-795 (2008).
366. D. Pechyony, V. Vapnik (2010) On the theory of learnining with privileged information. in *Advances in neural information processing systems*, pp 1894-1902.
367. D. Pechyony, R. Izmailov, A. Vashist, V. Vapnik (2010) SMO-Style Algorithms for Learning Using Privileged Information. in *Dmin* (Citeseer), pp 235-241.
368. A. Criminisi, J. Shotton, E. Konukoglu, Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. (2012).
369. Z. M. Hira, D. F. Gillies, A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics* **2015** (2015).
370. P. Smialowski, D. Frishman, S. J. B. Kramer, Pitfalls of supervised feature selection. **26**, 440-443 (2009).
371. A. Barla *et al.*, Machine learning methods for predictive proteomics. **9**, 119-128 (2008).
372. A. Dupuy, R. M. J. J. o. t. N. C. I. Simon, Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. **99**, 147-157 (2007).
373. Y. Drier, E. J. P. o. Domany, Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? **6**, e17795 (2011).
374. R. V. Rossel, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **158**, 46-54 (2010).
375. D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics* **6**, 10 (2014).
376. C. Lohrmann, P. Luukka, A novel similarity classifier with multiple ideal vectors based on k-means clustering. *Decision Support Systems* **111**, 27-37 (2018).
377. R. A. Fisher, On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87-94 (1922).
378. A. Agresti, A survey of exact inference for contingency tables. *Statistical science* **7**, 131-153 (1992).
379. S. Sato, N. Fujita, T. J. P. o. t. N. A. o. S. Tsuruo, Modulation of Akt kinase activity by binding to Hsp90. **97**, 10832-10837 (2000).
380. D. Stygar *et al.*, The influence of high-fat, high-sugar diet and bariatric surgery on HSP70 and HSP90 plasma and liver concentrations in diet-induced obese rats. **24**, 427-439 (2019).
381. J.-H. Lee *et al.*, Heat shock protein 90 (HSP90) inhibitors activate the heat shock factor 1 (HSF1) stress response pathway and improve glucose regulation in diabetic mice. **430**, 1109-1113 (2013).
382. S. Corvera, O. J. B. e. B. A.-M. B. o. D. Gealekman, Adipose tissue angiogenesis: impact on obesity and type-2 diabetes. **1842**, 463-472 (2014).
383. O. Gealekman *et al.*, Depot-specific differences and insufficient subcutaneous adipose tissue angiogenesis in human obesity. **123**, 186-194 (2011).

384. F. J. Tinahones *et al.*, Obesity-associated insulin resistance is correlated to adipose tissue vascular endothelial growth factors and metalloproteinase levels. **12**, 4 (2012).
385. S. Chen, S. Synowsky, M. Tinti, C. J. T. i. E. MacKintosh, Metabolism, The capture of phosphoproteins by 14-3-3 proteins mediates actions of insulin. **22**, 429-436 (2011).
386. Y. Liu *et al.*, Exploring the pathogenetic association between schizophrenia and type 2 diabetes mellitus diseases based on pathway analysis. *BMC medical genomics* **6**, S17 (2013).
387. Y. Nishimura *et al.*, Overexpression of YWHAZ relates to tumor cell proliferation and malignant outcome of gastric carcinoma. *British journal of cancer* **108**, 1324-1331 (2013).
388. T. Kang *et al.*, Characterization of signaling pathways associated with pancreatic β-cell adaptive flexibility in compensation of obesity-linked diabetes in db/db mice. *Molecular & Cellular Proteomics* (2020).
389. T. Vatseba, Influence of pathogenetic factors of type 2 diabetes on activation of PI3K/AkT/mTOR pathway and on the development of endometrial and breast cancer. *Regulatory Mechanisms in Biosystems* **10**, 295-299 (2019).
390. S. H. Back, R. J. Kaufman, Endoplasmic reticulum stress and type 2 diabetes. *Annual review of biochemistry* **81**, 767-793 (2012).
391. W. Ip, Y.-t. A. Chiang, T. Jin, The involvement of the wnt signaling pathway and TCF7L2 in diabetes mellitus: The current understanding, dispute, and perspective. *Cell & bioscience* **2**, 28 (2012).
392. D. L. Black, Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367-370 (2000).
393. E. V. Kriventseva *et al.*, Increase of functional diversity by alternative splicing. *Trends in Genetics* **19**, 124-128 (2003).
394. S. Chaudhary *et al.*, Alternative splicing and protein diversity: plants versus animals. *Frontiers in plant science* **10**, 708 (2019).
395. S. A. Bhuiyan *et al.*, Systematic evaluation of isoform function in literature reports of alternative splicing. *BMC genomics* **19**, 1-12 (2018).
396. M. L. Tress, F. Abascal, A. Valencia, Alternative splicing may not be the key to proteome complexity. *Trends in biochemical sciences* **42**, 98-110 (2017).
397. X. Wang *et al.*, Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Molecular & Cellular Proteomics* **17**, 422-430 (2018).
398. F. Gifford, Genetic traits. *Biology and philosophy* **5**, 327-347 (1990).
399. L. Gannett, What's in a cause?: The pragmatic dimensions of genetic explanations. *Biology and Philosophy* **14**, 349-373 (1999).
400. S. Stamm *et al.*, Function of alternative splicing. *Gene* **344**, 1-20 (2005).
401. B. J. Blencowe, Alternative splicing: new insights from global analyses. *Cell* **126**, 37-47 (2006).
402. H. R. Christofk *et al.*, The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230-233 (2008).

403. R. L. Brown *et al.*, CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *The Journal of clinical investigation* **121**, 1064-1074 (2011).
404. M.-É. Huot *et al.*, The Sam68 STAR RNA-binding protein regulates mTOR alternative splicing during adipogenesis. *Molecular cell* **46**, 187-199 (2012).
405. M. Shionyu, A. Yamaguchi, K. Shinoda, K.-i. Takahashi, M. Go, AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic acids research* **37**, D305-D309 (2009).
406. L. G. Wilming *et al.*, The vertebrate genome annotation (Vega) database. *Nucleic acids research* **36**, D753-D760 (2007).
407. G. Koscielny *et al.*, ASTD: the alternative splicing and transcript diversity database. *Genomics* **93**, 213-220 (2009).
408. A. Frankish *et al.*, GENCODE 2021. *Nucleic acids research* **49**, D916-D923 (2021).
409. I. Gronau, S. Moran, Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters* **104**, 205-210 (2007).
410. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).
411. M. Wilhelm *et al.*, Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-587 (2014).
412. B. Domon, R. Aebersold, Mass spectrometry and protein analysis. *science* **312**, 212-217 (2006).
413. L. M. Smith, N. L. Kelleher, Proteoforms as the next proteomics currency. *Science* **359**, 1106-1107 (2018).
414. R. Aebersold *et al.*, How many human proteoforms are there? *Nature chemical biology* **14**, 206-214 (2018).
415. S.-N. Hsu, K. J. Hertel, Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA biology* **6**, 526-530 (2009).
416. E. Melamud, J. Moult, Stochastic noise in splicing machinery. *Nucleic acids research* **37**, 4873-4886 (2009).
417. A. Elnaggar *et al.*, ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* (2020).
418. A. Elnaggar *et al.*, ProtTrans: towards cracking the language of life's code through self-supervised learning. *bioRxiv*, 2020.2007. 2012.199554 (2021).
419. K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450* (2019).
420. V. A. Schneider *et al.*, Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research* **27**, 849-864 (2017).
421. T. Chen *et al.*, Xgboost: extreme gradient boosting. *R package version 0.4-2* **1**, 1-4 (2015).

422. R. F. White *et al.*, Recent research on Gulf War illness and other health problems in veterans of the 1991 Gulf War: Effects of toxicant exposures during deployment. *Cortex* **74**, 449-475 (2016).
423. A. C. Ribeiro, L. S. Deshpande, A review of pre-clinical models for Gulf War Illness. *Pharmacology & therapeutics* **228**, 107936 (2021).
424. R. W. Haley, T. L. Kurt, Self-reported exposure to neurotoxic chemical combinations in the Gulf War: a cross-sectional epidemiologic study. *Jama* **277**, 231-237 (1997).
425. L. Steele, A. Sastre, M. M. Gerkovich, M. R. Cook, Complex factors in the etiology of Gulf War illness: wartime exposures and risk factors in veteran subgroups. *Environmental health perspectives* **120**, 112-118 (2012).
426. L. L. Chao, J. C. Rothlind, V. A. Cardenas, D. J. Meyerhoff, M. W. Weiner, Effects of low-level exposure to sarin and cyclosarin during the 1991 Gulf War on brain function and brain structure in US veterans. *Neurotoxicology* **31**, 493-501 (2010).
427. D. J. Kearney *et al.*, Mindfulness-based stress reduction in addition to usual care is associated with improvements in pain, fatigue, and cognitive failures among veterans with gulf war illness. *The American journal of medicine* **129**, 204-214 (2016).
428. K. Kerr *et al.*, A detoxification intervention for Gulf War illness: a pilot randomized controlled trial. *International journal of environmental research and public health* **16**, 4143 (2019).
429. J. N. Baraniuk, S. El-Amin, R. Corey, R. U. Rayhan, C. R. Timbol, Carnosine treatment for gulf war illness: a randomized controlled trial. *Global journal of health science* **5**, 69 (2013).
430. S. T. Donta *et al.*, Benefits and harms of doxycycline treatment for Gulf War veterans' illnesses: a randomized, double-blind, placebo-controlled trial. *Annals of internal medicine* **141**, 85-94 (2004).
431. L. Conboy, M. St John, R. Schnyer, The effectiveness of acupuncture in the treatment of Gulf War Illness. *Contemporary clinical trials* **33**, 557-562 (2012).
432. L. Conboy *et al.*, The effectiveness of individualized acupuncture protocols in the treatment of Gulf War illness: a pragmatic randomized clinical trial. *PLoS One* **11**, e0149161 (2016).
433. G. Ernst, H. Strzyz, H. Hagmeister, Incidence of adverse effects during acupuncture therapy—a multicentre survey. *Complementary therapies in medicine* **11**, 93-97 (2003).
434. J. Wu *et al.*, Systematic review of adverse effects: a further step towards modernization of acupuncture in China. *Evidence-Based Complementary and Alternative Medicine* **2015** (2015).
435. V. N. Harry, Novel imaging techniques as response biomarkers in cervical cancer. *Gynecologic oncology* **116**, 253-261 (2010).
436. M. Parlato, J.-M. Cavaillon, Host response biomarkers in the diagnosis of sepsis: a general overview. *Sepsis*, 149-211 (2015).
437. S. Amin, O. F. Bathe, Response biomarkers: re-envisioning the approach to tailoring drug therapy for cancer. *BMC cancer* **16**, 1-11 (2016).
438. R. Melzack, The short-form McGill pain questionnaire. *Pain* **30**, 191-197 (1987).
439. J. E. Ware Jr, SF-36 health survey update. *Spine* **25**, 3130-3139 (2000).

440. J. E. Ware Jr, C. D. Sherbourne, The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical care*, 473-483 (1992).
441. C. S. Burckhardt, K. D. Jones, Adult Measures of Pain: The McGill Pain Questionnaire (MPQ). *Arthritis & Rheumatism: Arthritis Care & Research* (2003).
442. B. Li *et al.*, TGF-β1 DNA polymorphisms, protein levels, and blood pressure. *Hypertension* **33**, 271-275 (1999).
443. S. Fichtlscherer *et al.*, Elevated C-reactive protein levels and impaired endothelial vasoreactivity in patients with coronary artery disease. *Circulation* **102**, 1000-1006 (2000).
444. Y. Kim *et al.*, Targeted proteomics identifies liquid-biopsy signatures for extracapsular prostate cancer. *Nature communications* **7**, 1-10 (2016).
445. J. Candia *et al.*, Assessment of variability in the SOMAscan assay. *Scientific reports* **7**, 1-13 (2017).
446. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301-320 (2005).
447. Y. Zhao, L. Wong, W. W. B. Goh, How to do quantile normalization correctly for gene expression data analyses. *Scientific reports* **10**, 1-11 (2020).
448. C. Müller *et al.*, Removing batch effects from longitudinal gene expression-quantile normalization plus ComBat as best approach for microarray transcriptome data. *PloS one* **11**, e0156594 (2016).
449. C. Chen *et al.*, Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* **6**, e17238 (2011).
450. J. Gross, J. Groß, *Linear regression* (Springer Science & Business Media, 2003), vol. 175.
451. T. A. Craney, J. G. Surles, Model-dependent variance inflation factor cutoff values. *Quality engineering* **14**, 391-403 (2002).
452. I. S. Dhillon (2001) Co-clustering documents and words using bipartite spectral graph partitioning. in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 269-274.
453. Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research* **13**, 703-716 (2003).
454. C. E. Brown, "Coefficient of variation" in Applied multivariate statistics in geohydrology and related sciences. (Springer, 1998), pp. 155-157.
455. H. Abdi, Coefficient of variation. *Encyclopedia of research design* **1**, 169-171 (2010).
456. S. E. Reese *et al.*, A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* **29**, 2877-2883 (2013).
457. S. Wang, H. Cui, Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis* **117**, 134-149 (2013).
458. D. V. Pinkhasova *et al.*, Regulatory status of pesticide residues in cannabis: Implications to medical use in neurological diseases. *Current research in toxicology* **2**, 140-148 (2021).
459. R. K. Singla, A. Sultana, M. Alam, B. Shen, Regulation of Pain Genes—Capsaicin vs Resiniferatoxin: Reassessment of Transcriptomic Data. *Frontiers in pharmacology* **11**, 1565 (2020).

460. S. Yamaoka *et al.*, Altered gene expression of RNF34 and PACAP possibly involved in mechanism of exercise-induced analgesia for neuropathic pain in rats. *International journal of molecular sciences* **18**, 1962 (2017).
461. R. J. Snelgrove *et al.*, A critical role for LTA4H in limiting chronic pulmonary neutrophilic inflammation. *Science* **330**, 90-94 (2010).
462. A. M. Fourie, Modulation of inflammatory disease by inhibitors of leukotriene A4 hydrolase. *Current opinion in investigational drugs (London, England: 2000)* **10**, 1173-1182 (2009).
463. D. Bouzid *et al.*, Polymorphisms in the IL2RA and IL2RB genes in inflammatory bowel disease risk. *Genetic testing and molecular biomarkers* **17**, 833-839 (2013).
464. R. Infantino *et al.*, MED1/BDNF/TrkB pathway is involved in thalamic hemorrhage-induced pain and depression by regulating microglia. *Neurobiology of disease*, 105611 (2022).
465. A. Usta *et al.*, The usefulness of CD34, PCNA Immunoreactivity, and Histopathological findings for prediction of pain persistence after the removal of Endometrioma. *Reproductive Sciences* **26**, 269-277 (2019).
466. M. Gauthier, C. Laroye, D. Bensoussan, C. Boura, V. Decot, Natural Killer cells and monoclonal antibodies: Two partners for successful antibody dependent cytotoxicity against tumor cells. *Critical Reviews in Oncology/Hematology* **160**, 103261 (2021).
467. D. Padua *et al.*, A long noncoding RNA signature for ulcerative colitis identifies IFNG-AS1 as an enhancer of inflammation. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **311**, G446-G457 (2016).
468. M. Misale, L. Janusek, H. Mathews, Epigenetic basis for psychological stress-induced IFNg production in women diagnosed with breast cancer. *Brain, Behavior, and Immunity* **49**, e11-e12 (2015).
469. S. D. Hartung *et al.*, Correction of metabolic, craniofacial, and neurologic abnormalities in MPS I mice treated at birth with adeno-associated virus vector transducing the human α-L-iduronidase gene. *Molecular Therapy* **9**, 866-875 (2004).
470. D. S. Siroky, Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys* **3**, 147-163 (2009).
471. K. Pasupa, W. Sunhem (2016) A comparison between shallow and deep architecture classifiers on small dataset. in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)* (IEEE), pp 1-6.
472. A. Schindler, T. Lidy, A. Rauber (2016) Comparing Shallow versus Deep Neural Network Architectures for Automatic Music Genre Classification. in *FMT*, pp 17-21.
473. J. Tapial *et al.*, An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome research* **27**, 1759-1768 (2017).
474. P. Zhou *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature* **579**, 270-273 (2020).
475. D. R. Burton, E. J. Topol (2021) Variant-proof vaccines—invest now for the next pandemic. (Nature Publishing Group).
476. C. Huang *et al.*, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* **395**, 497-506 (2020).

477. D. S. Hui *et al.*, The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *International journal of infectious diseases* **91**, 264-266 (2020).
478. N. Zhu *et al.*, A novel coronavirus from patients with pneumonia in China, 2019. *New England journal of medicine* (2020).
479. G. J. Nabel, Designing tomorrow's vaccines. *New England Journal of Medicine* **368**, 551-560 (2013).
480. S. A. Plotkin, S. L. Plotkin, The development of vaccines: how the past led to the future. *Nature Reviews Microbiology* **9**, 889-893 (2011).
481. J. E. Martin *et al.*, A SARS DNA vaccine induces neutralizing antibody and cellular immune responses in healthy adults in a Phase I clinical trial. *Vaccine* **26**, 6338-6343 (2008).
482. A. Zumla, J. F. Chan, E. I. Azhar, D. S. Hui, K.-Y. Yuen, Coronaviruses—drug discovery and therapeutic options. *Nature reviews Drug discovery* **15**, 327-347 (2016).
483. A. Zumla, Z. A. Memish, D. S. Hui, S. Perlman, Vaccine against Middle East respiratory syndrome coronavirus. *The Lancet Infectious Diseases* **19**, 1054-1055 (2019).
484. M. Cotten *et al.*, Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet* **382**, 1993-2002 (2013).
485. E. De Wit, N. Van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: recent insights into emerging coronaviruses. *Nature Reviews Microbiology* **14**, 523-534 (2016).
486. G. Lu *et al.*, Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* **500**, 227-231 (2013).
487. J. K. Millet, G. R. Whittaker, Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* **517**, 3-8 (2018).
488. Y. Yuan *et al.*, Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature communications* **8**, 1-9 (2017).
489. W. Song, M. Gui, X. Wang, Y. Xiang, Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS pathogens* **14**, e1007236 (2018).
490. C.-C. Cho, M.-H. Lin, C.-Y. Chuang, C.-H. Hsu, Macro domain from Middle East respiratory syndrome coronavirus (MERS-CoV) is an efficient ADP-ribose binding module: crystal structure and biochemical studies. *Journal of Biological Chemistry* **291**, 4894-4902 (2016).
491. M. Gui *et al.*, Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell research* **27**, 119-129 (2017).
492. A. C. G. Kankanamalage *et al.*, Structure-guided design of potent and permeable inhibitors of MERS coronavirus 3CL protease that utilize a piperidine moiety as a novel design element. *European journal of medicinal chemistry* **150**, 334-346 (2018).
493. K. Ratia, A. Kilianski, Y. M. Baez-Santos, S. C. Baker, A. Mesecar, Structural basis for the ubiquitin-linkage specificity and deISGylating activity of SARS-CoV papain-like protease. *PLoS pathogens* **10**, e1004113 (2014).

494. N. Wang *et al.*, Structural definition of a neutralization-sensitive epitope on the MERS-CoV S1-NTD. *Cell reports* **28**, 3395-3405. e3396 (2019).
495. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic acids research* **43**, D571-D577 (2015).
496. F. Sievers, D. G. Higgins, Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27**, 135-145 (2018).
497. N. Eswar *et al.*, Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **15**, 5.6. 1-5.6. 30 (2006).
498. S. R. Lehrman, Protein structure. *Fundamentals of protein biotechnology*, 9-38 (2017).
499. M. y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein science* **15**, 2507-2524 (2006).
500. A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).
501. J. Yang *et al.*, The I-TASSER Suite: protein structure and function prediction. *Nature methods* **12**, 7-8 (2015).
502. Y. Zhang, I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9**, 1-8 (2008).
503. I. André, P. Bradley, C. Wang, D. Baker, Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences* **104**, 17656-17661 (2007).
504. W.-H. Shin, G. R. Lee, L. Heo, H. Lee, C. Seok, Prediction of protein structure and interaction by GALAXY protein modeling programs. *Bio Design* **2**, 1-11 (2014).
505. J. Ko, H. Park, L. Heo, C. Seok, GalaxyWEB server for protein structure prediction and refinement. *Nucleic acids research* **40**, W294-W297 (2012).
506. B. W. Neuman *et al.*, "Ultrastructure of SARS-CoV, FIPV, and MHV revealed by electron cryomicroscopy" in The Nidoviruses. (Springer, 2006), pp. 181-185.
507. B. W. Neuman *et al.*, A structural analysis of M protein in coronavirus assembly and morphology. *Journal of structural biology* **174**, 11-22 (2011).
508. H. Fujiwara, M. Fujihara, T. Koyama, T. Ishiwata (2004) New aspect of the spontaneous formation of a bilayer lipid membrane. in *AIP Conference Proceedings* (American Institute of Physics), pp 724-726.
509. S. Klein *et al.*, SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nature communications* **11**, 1-10 (2020).
510. D. M. Kern *et al.*, Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nature Structural & Molecular Biology*, 1-10 (2021).
511. H. Woo *et al.*, Developing a fully glycosylated full-length SARS-CoV-2 spike protein model in a viral membrane. *The Journal of Physical Chemistry B* **124**, 7128-7137 (2020).
512. K. Hinsen, The molecular modeling toolkit: a new approach to molecular simulations. *Journal of Computational Chemistry* **21**, 79-85 (2000).
513. J. Huang, A. D. MacKerell Jr, CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **34**, 2135-2145 (2013).

514. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology* **75**, 861-877 (2019).
515. T. I. Croll, ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology* **74**, 519-530 (2018).
516. E. F. Pettersen *et al.*, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70-82 (2021).
517. P. Eastman *et al.*, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).
518. D. A. Case *et al.*, Amber 2020. (2020).
519. M. A. Martí-Renom *et al.*, Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **29**, 291-325 (2000).
520. B. Kuhlman, P. Bradley, Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **20**, 681-697 (2019).
521. S. Srinivasan *et al.*, Structural genomics of SARS-CoV-2 indicates evolutionary conserved functional regions of viral proteins. *Viruses* **12**, 360 (2020).
522. W. Zheng *et al.*, Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* **87**, 1149-1164 (2019).
523. B. W. Neuman *et al.*, Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. *Journal of virology* **80**, 7918-7928 (2006).
524. W. Pezeshkian, M. König, S. J. Marrink, J. H. Ipsen, A multi-scale approach to membrane remodeling processes. *Frontiers in molecular biosciences* **6**, 59 (2019).
525. W. Pezeshkian, J. H. Ipsen, Fluctuations and conformational stability of a membrane patch with curvature inducing inclusions. *Soft matter* **15**, 9974-9981 (2019).
526. W. Pezeshkian, M. König, T. A. Wassenaar, S. J. Marrink, Backmapping triangulated surfaces to coarse-grained membrane models. *Nature communications* **11**, 1-9 (2020).
527. P. C. Souza *et al.*, Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature methods* **18**, 382-388 (2021).
528. J. V. Vermaas, C. G. Mayne, E. Shinn, E. Tajkhorshid, Assembly and Analysis of Cell-Scale Membrane Envelopes. *Journal of Chemical Information and Modeling* (2021).
529. T. A. Wassenaar *et al.*, High-throughput simulations of dimer and trimer assembly of membrane proteins. The DAFT approach. *Journal of chemical theory and computation* **11**, 2278-2291 (2015).
530. M. Horne, W. Merchant, *The Stability of Frames: The Commonwealth and International Library: Structures and Solid Body Mechanics Division*. W. M. M.R. HORNE, Ed. (Elsevier, 2014).
531. H.-Y. Yen *et al.*, PtdIns (4, 5) P 2 stabilizes active states of GPCRs and enhances selectivity of G-protein coupling. *Nature* **559**, 423-427 (2018).
532. G. Hedger *et al.*, Cholesterol interaction sites on the transmembrane domain of the hedgehog signal transducer and class FG protein-coupled receptor smoothened. *Structure* **27**, 549-559. e542 (2019).

533. P.-C. Hsu, D. Jefferies, S. Khalid, Molecular dynamics simulations predict the pathways via which pristine fullerenes penetrate bacterial membranes. *The journal of physical chemistry B* **120**, 11170-11179 (2016).
534. J. M. Hendrickx, R. M. Jungers, A. Olshevsky, G. Vankeerberghen, Graph diameter, eigenvalues, and minimum-time consensus. *Automatica* **50**, 635-640 (2014).
535. K. Orman, V. Labatut, H. Cherifi, "An empirical study of the relation between community structure and transitivity" in Complex Networks. (Springer, 2013), pp. 99-110.
536. E. Lindahl, B. Hess, D. Van Der Spoel, GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual* **7**, 306-317 (2001).
537. M. Bárcena *et al.*, Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. *Proceedings of the National Academy of Sciences* **106**, 582-587 (2009).
538. C. Liu *et al.*, Viral architecture of SARS-CoV-2 with post-fusion spike revealed by Cryo-EM. *BioRxiv* (2020).
539. Z. Jin, H. Wang, Y. Duan, H. Yang, The main protease and RNA-dependent RNA polymerase are two prime targets for SARS-CoV-2. *Biochemical and Biophysical Research Communications* **538**, 63-71 (2021).
540. L. Casalino *et al.*, AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *The International Journal of High Performance Computing Applications*, 10943420211006452 (2021).
541. T. Reddy *et al.*, Nothing to sneeze at: a dynamic and integrative computational model of an influenza A virion. *Structure* **23**, 584-597 (2015).
542. J. D. Durrant *et al.*, Mesoscale all-atom influenza virus simulations suggest new substrate binding mechanism. *ACS central science* **6**, 189-196 (2020).
543. J. R. Perilla, K. Schulten, Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nature communications* **8**, 1-10 (2017).
544. G. S. Ayton, G. A. Voth, Multiscale computer simulation of the immature HIV-1 virion. *Biophysical journal* **99**, 2757-2765 (2010).
545. J. A. Hadden *et al.*, All-atom molecular dynamics of the HBV capsid reveals insights into biological function and cryo-EM resolution limits. *Elife* **7**, e32478 (2018).
546. I. Galindo *et al.*, Antiviral drugs targeting endosomal membrane proteins inhibit distant animal and human pathogenic viruses. *Antiviral research* **186**, 104990 (2021).
547. R. B. Dodd, T. Wilkinson, D. J. Schofield, Therapeutic monoclonal antibodies to complex membrane protein targets: Antigen generation and antibody discovery strategies. *BioDrugs* **32**, 339-355 (2018).
548. P. C. Souza *et al.*, Protein–ligand binding with the coarse-grained Martini model. *Nature communications* **11**, 1-11 (2020).
549. R. Veneziano *et al.*, Role of nanoscale antigen organization on B-cell activation probed using DNA origami. *Nature nanotechnology* **15**, 716-723 (2020).
550. Y. Baran *et al.*, MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome biology* **20**, 1-19 (2019).
551. H. Li *et al.*, Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell* **176**, 775-789. e718 (2019).

552. M. Cohen *et al.*, Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* **175**, 1031-1044. e1018 (2018).
553. A. Giladi *et al.*, Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nature cell biology* **20**, 836-846 (2018).
554. A. Sebé-Pedrós *et al.*, Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq. *Cell* **173**, 1520-1534. e1520 (2018).
555. S. Srinivasan, N. T. Johnson, D. Korkin, A Hybrid Deep Clustering Approach for Robust Cell Type Profiling Using Single-cell RNA-seq Data. *bioRxiv* 10.1101/511626, 511626 (2019).
556. A. K. Shalek *et al.*, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240 (2013).
557. A. P. Patel *et al.*, Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
558. J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. Wortman, Learning bounds for domain adaptation. (2008).
559. Y. Mansour, M. Mohri, A. Rostamizadeh, Domain adaptation with multiple sources. *Advances in neural information processing systems* **21**, 1041-1048 (2008).
560. S. Zhao, B. Li, P. Xu, K. Keutzer, Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169* (2020).
561. A. Jalali, N. Pfeifer, Interpretable per case weighted ensemble method for cancer associations. *BMC genomics* **17**, 1-10 (2016).
562. R. Rao *et al.*, Evaluating protein transfer learning with tape. *Advances in Neural Information Processing Systems* **32**, 9689 (2019).
563. W. Jin, R. Barzilay, T. Jaakkola, Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908* (2020).
564. O. Narykov, N. Johnson, D. Korkin, Determining rewiring effects of alternatively spliced isoforms on protein-protein interactions using a computational approach. *bioRxiv*, 256834 (2018).
565. L. E. Raileanu, K. Stoffel, Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* **41**, 77-93 (2004).
566. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525 (2016).
567. D. Roberston, J. Shotton, T. Sharp, "The sherwood software library" in Decision Forests for Computer Vision and Medical Image Analysis. (Springer, 2013), pp. 333-342.
568. D. Amaratunga, J. Cabrera, Y.-S. Lee, Enriched random forests. *Bioinformatics* **24**, 2010-2014 (2008).
569. Y. Ye, Q. Wu, J. Z. Huang, M. K. Ng, X. Li, Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition* **46**, 769-787 (2013).
570. Q. Wang, T.-T. Nguyen, J. Z. Huang, T. T. Nguyen, An efficient random forests algorithm for high dimensional data classification. *Advances in Data Analysis and Classification* **12**, 953-972 (2018).

571. M. Sultan *et al.*, A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochemical and biophysical research communications* **422**, 643-646 (2012).
572. N. T. Johnson, RNA-Seq of Mouse Models - Environmentally-Induced T2D and Control.
573. Anonymous (Type 2 Diabetes Knowledge Portal.
574. C. Von Mering *et al.*, STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research* **33**, D433-D437 (2005).
575. D. M. Church *et al.*, Modernizing reference genome assemblies. *PLoS Biol* **9**, e1001091 (2011).
576. F. Jason *et al.*, Sequence data and association statistics from 12,940 type 2 diabetes cases and controls. *Scientific data* **4**, 170179 (2017).
577. K. J. Gaulton *et al.*, Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics* **47**, 1415 (2015).
578. J. M. Mercader *et al.*, A loss-of-function splice acceptor variant in IGF2 is protective for type 2 diabetes. *Diabetes*, db170187 (2017).
579. O. Narykov (2021) AlexandrNP/altintool: ALT-IN with Docker interface. (Zenodo).
580. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
581. D. Kim *et al.*, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36 (2013).
582. C. Trapnell *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562 (2012).
583. P. Jones *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
584. P. M. Rice, P. M. Rice, A. J. Bleasby, J. C. Ison, *EMBOSS User's Guide: Practical Bioinformatics with EMBOSS* (Cambridge University Press, 2020).
585. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite. *Trends in genetics* **16**, 276-277 (2000).
586. P. Shannon *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).