THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Mapping the proteome with data-driven methods

A cycle of measurement, modeling, hypothesis generation, and engineering

FILIP BURIC

Department of Biology and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2021

Mapping the proteome with data-driven methods:
A cycle of measurement, modeling, hypothesis generation, and engineering

Filip Buric

ISBN: 978-91-7905-570-7
Löpnummer: 5037
i serien Doktorsavhandlingar vid Chalmers tekniska högskola.
Ny serie (ISSN 0346-718X)

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology

SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Conceptual representation of the data-driven cycle explored in this thesis

Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Mapping the proteome with data-driven methods:
A cycle of measurement, modeling, hypothesis generation, and engineering

Filip Buric
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

The living cell exhibits emergence of complex behavior and its modeling requires a systemic, integrative approach if we are to thoroughly understand and harness it. The work in this thesis has had the more narrow aim of quantitatively characterizing and mapping the proteome using data-driven methods, as proteins perform most functional and structural roles within the cell. Covered are the different parts of the cycle from improving quantification methods, to deriving protein features relying on their primary structure, predicting the protein content solely from sequence data, and, finally, to developing theoretical protein engineering tools, leading back to experiment.

High-throughput mass spectrometry platforms provide detailed snapshots of a cell's protein content, which can be mined towards understanding how the phenotype arises from genotype and the interplay between the various properties of the constituent proteins. However, these large and dense data present an increased analysis challenge and current methods capture only a small fraction of signal. The first part of my work has involved tackling these issues with the implementation of a GPU-accelerated and distributed signal decomposition pipeline, making factorization of large proteomics scans feasible and efficient. The pipeline yields individual analyte signals spanning the majority of acquired signal, enabling high precision quantification and further analytical tasks.

Having such detailed snapshots of the proteome enables a multitude of undertakings. One application has been to use a deep neural network model to learn the amino acid sequence determinants of temperature adaptation, in the form of reusable deep model features. More generally, systemic quantities may be predicted from the information encoded in sequence by evolutionary pressure. Two studies taking inspiration from natural language processing have sought to learn the grammars behind the languages of expression, in one case predicting mRNA levels from DNA sequence, and in the other protein abundance from amino acid sequence. These two models helped build a quantitative understanding of the central dogma and, furthermore, in combination yielded an improved predictor

of protein amount. Finally, a mathematical framework relying on the embedded space of a deep model has been constructed to assist guided mutation of proteins towards optimizing their abundance.

Keywords: data-independent acquisition, deep learning, feature learning, machine learning, mass spectrometry, model interpretation, proteomics, sequence feature engineering, tensor factorization

List of Publications

This thesis is based on the work contained in the following papers and manuscripts:

- **Paper I**:
  Parallel factor analysis enables quantification and identification of highly convolved data-independent-acquired protein spectra
  – **Filip Buric**, Jan Zrimec, Aleksej Zelezniak (2020), *Cell Patterns*

- **Paper II**:
  Learning deep representations of enzyme thermal adaptation
  – Gang Li*, **Filip Buric***, Jan Zrimec*, Sandra Viknander, Jens Nielsen, Aleksej Zelezniak, Martin KM Engqvist (2021), *Manuscript*

- **Paper III**:
  DeepTranslation: The sequence grammar behind protein abundance and its support for protein engineering
  – **Filip Buric**, Jan Zrimec, Aleksej Zelezniak (2021), *Manuscript*

- **Paper IV**:
  Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure
  – Jan Zrimec, Christoph S. Börlin, **Filip Buric**, Azam Sheikh Muhammad, Rhongzen Chen, Verena Siewers, Vilhelm Verendel, Jens Nielsen, Mats Töpel, Aleksej Zelezniak (2020), *Nature Communications*

Additional publications during doctoral research not included in this thesis:

- **Learning the regulatory code of gene expression** (review article)
  – Jan Zrimec, **Filip Buric**, Mariia Kokina, Victor Garcia, Aleksej Zelezniak (2021), *Frontiers in Molecular Biosciences*

- **metaGEM: Reconstruction of genome scale metabolic models directly from metagenomes**
  – Francisco Zorrilla, **Filip Buric**, Kiran R. Patil, Aleksej Zelezniak (2021), *Nucleic Acids Research (accepted)*

* Authors contributed equally to this work

Contribution Summary

- **Paper I**: I co-designed the study, developed the software pipeline, curated data, performed the decomposition trials, analyzed the results, produced visualizations, wrote the manuscript.
- **Paper II**: I co-designed the model interpretation (perturbation) analysis, co-wrote the respective software, performed the perturbation analysis, analyzed the results, produced visualizations, co-wrote the manuscript (the section *Interpreting the sequence determinants of thermostability* in particular).
- **Paper III**: I co-designed the study, wrote the software, curated data, performed the machine learning, developed the mathematical framework, analyzed the results, produced visualizations, wrote the manuscript.
- **Paper IV**: I provided mathematical and programming assistance, wrote small parts of the software, curated data, participated in discussions concerning the study design and execution, reviewed the manuscript.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between June 2017 and December 2021 at the division of Systems and Synthetic Biology under the supervision of Aleksej Zelezniak.

The research was mainly funded by SciLifeLab.

Acknowledgments

Firstly, I must thank my supervisor Aleksej for his trust, guidance, encouragement, and patience. I greatly appreciate him giving me the freedom to explore and experiment, while at the same time providing the critical grounding required for undertaking our type of research.

Many thanks as well to my fellow group members and collaborators: Jan, Martin, Sandra, Gang, Mariia, Xiaozhi, and Christoph, among others. Besides the fruitful projects, I have learned much from them and have gained a renewed appreciation for the diverse skills and perspectives we all contribute.

Thanks also go to all SysBio members for the collegial and friendly atmosphere, the fun activities we've had over the years, as well as the feedback, friendship, and words of wisdom they shared with me. They have significantly contributed to the sense of purpose and fellowship I've felt during these past years. Special thanks as well to the administrative staff, for their continued support.

I am also fortunate to have taken part in creating the MBM workshop, for which I thank my fellow organizers Johannes, Barbara, and Felix for a very fun and rewarding experience.

Lastly, thanks to go to my friends and family for their constant support, and to all those with whom I've shared a pleasant conversation about science, life, literature, entertainment, philosophy, or simply the weather.

Contents

1 | Background

1.1 The challenge and reward of mapping life's machinery

Life is beautiful in its complexity and awe-inspiring in its intrinsic drive towards perpetuating itself in spite of the constant pull asunder from the universe around it. This is exhibited in the simplest cells and viruses, as well as across entire species. Yet this complexity often perplexes when we try to grasp the intricacies of life's behavior and trajectories, hindering at the same time the pure intellectual pursuit, the strive to sustainably live in harmony with our environment, as well as the struggle to heal and preserve. As with science and engineering in general, progress in biology relies on a thorough understanding of the constituents of life's machinery, as well as the rules which govern their intricate behavior. Once understood to a certain level, the possibility of harnessing these processes for technological applications and medical intervention appears.

Within all cells, life is organized around the central dogma of molecular biology, conceptually separating information storage, and its logic of retrieval and propagation, from the assembly and maintenance processes that renew the structure of the cell and drive it towards division [Alb15]. This transfer of information between the involved biopolymers (DNA, RNA, and protein) can be seen as the central part of the self-replication that is a defining characteristic of life. This simplistic framing is augmented by the many levels of regulation taking place constantly to maintain the cell's homeostasis, while adapting it to the environment and preparing it for division. The ensemble of these various processes, whether regulatory or more "straight-forward" information forwarding (transcription and translation), is characterized by non-linearity and thermodynamical non-equilibrium [NP77; Str14]. The biochemical reaction networks from which this behavior emerges are dynamic and complex, featuring numerous forms of feedback and robustness. They are also subject to constant evolution, arising from the continual negotiation between environmental pressure and intrinsic change caused by inevitable mutation or, in some species, extrinsic change due to horizontal gene transfer. Evolution thus adds another dimension to this complexity and flexibility, shaping a varied adaptational landscape, with some reaction pathways more conserved than others, and some essentially universal, such as transcription, translation, and central metabolism, evidence of their critical importance. [Alb15].

Whereas the genome (the set of all DNA sequences) could be seen as the intergenerational carrier of information, and the transcriptome (the set of all RNA transcripts) as a messenger (and regulator) towards shaping the phenotype, proteins can conversely be seen as actuators, signalers, regulators, and construction material, both inside the cell, as well as in its community. While not explored here, it is important to keep in mind this larger context, as proteins are also crucial for intercellular communication, be it quorum sensing in bacteria, mating signals in yeast, or more elaborate regulatory signaling networks in multicellular organisms. Whereas intracellular signal transduction pathways are built from proteins, the extracellular signals themselves are proteins in this latter case, increasing both the complexity as well as specificity of these molecules [Alb15]. A striking example of the importance of proteome dynamics is its role in synaptic plasticity, vital for normal brain function and learning [GA21]. Synaptic activity and the associated protein-protein interaction networks rely on quick changes to proteome composition and allocation, often demanding quick synthesis, transport, and degradation [GA21]. For yeast and Gram-positive bacteria, the extracellular factors are peptides [RB12; MB12]. This signaling behavior can be seen as a social extension of the proteome, albeit with varying relevance and occurrence in a given species. Given its myriad functions and dynamics, a comprehensive picture of the proteome thus forms a considerable part of the overall tapestry of knowledge within cellular biology and mapping its behavior is essential to our understanding of life.

Beyond the pursuit of knowledge as intellectual aim of its own, understanding the composition and function of the proteome enables the treatment of disease. Given their ubiquity, one could of course argue that proteins are involved in virtually all disease. However, there are classes of disorders that have protein malfunction (and consequently, the resulting shift in abundance) as a clear driver of pathosis. One such example are the varieties of Alzheimer's disease. While the cause is still debated, its progression is strongly associated with abnormalities in two proteins. Tau proteins are a promoter of microtubule (cytoskeleton) assembly, their activity being regulated by their degree of phosphorylation. An abnormal hyperphosphorylated state of tau protein aggregates intracellularly, also capturing normal tau and evading ubiquitination (degradation). This inhibits cytoskeleton formation and axonal transport, leading ultimately to cell death [Iqb+05; HF21]. This pathology is also present in other disorders classified as tauopathies [Iqb+05]. An overproduction of the peptide amyloid beta, resulting from an alternative cleaving of amyloid precursor protein, coupled with its reduced degradation, leads to its extracellular accumulation in the form of insoluble plaques and neuroinflammation as

a reaction to these [HF21]. On a different side of the protein homeostasis are lysosomal storage diseases, which affect the capacity of cells to degrade proteins. The lysosome is an organelle present in all eukaryotes, responsible with the breakdown and recycling of a diverse range of large molecules and other organelles, thus part of the overall environment-responsive autophagic pathway [PMB21]. The failure of its normal functioning leads to the accumulation or secretion of undesirable proteins, which would have been processed in the lysosome [Alb15].

With an understanding of the proteome and key interactions, the possibilities of bioengineering emerge. Cells can be made into living factories to produce chemical products for a range of needs, from industrial to medicinal. Obtaining a desired product is achieved in different ways, either by optimizing existing pathways (by e.g. overexpressing an enzyme) [NK16], modifying or designing the structure of proteins to alter their performance or endow them with a desired behavior, either through evolution [YWA19] or rational design (the latter still a considerable challenge, especially for *de novo* designs) [KD20], or by introducing an exogenous process into a platform organism (one that is well-characterized and amenable to the task) [NK16]. Proteins are involved in virtually all applications, either as a means of "rewiring" the cell, or as a desired output, often in the form of enzymes. These have many uses, from detergents [vDH13] to breaking down undesired, yet robust matter such as plant biomass [Kme+20] or plastic [Tou+20; Zri+20b] into reusable material, instead of wastefully burning it. Many proteins are produced in cell factories for medical applications such as tissue engineering, for example spine [MR20] and cartilage [Shi+16] regeneration. Examples of pharmacological protein products are insulin [Vec+18], interferon (signaling proteins that are used to treat various diseases) [vDH13], agkisacutacin (an antithrombotic) [WP18], apidaecin Ia (antimicrobial peptide) [WP18], aliglucerase alfa (used to treat Graucher's disease, which manifests with accumulation of glucocerebroside due to a faulty lysosomal enzyme) [PW16], HIV antibodies [PW16], albumin (transport protein in blood plasma) [PW16], and many more for treating various diseases including metabolic disorders, hematological disorders, and cancers [San+16].

Understanding the progression of disease and finding treatment targets, as well as engineering cell factories to produce desired chemicals in an efficient way, naturally presuppose an accurate image of the proteome, but also comprehensive modeling that captures the intricacies of the various subsystems and interactions involved. This is unfortunately a very difficult undertaking, both from the data acquisition and modeling sides. Given the size, complex structure, heterogeneity, and fragility of

cells, measurements that are comprehensive, precise, and time-resolved to arbitrary scale are difficult technologically (though constantly improving). Moreover, proper environmental conditions or community composition required to grow cells in the lab are known only for a relatively small number of species [Ste12]. Additionally, digital storage and computational capacity for big data sets is a severe limitation even if collection were quickly and substantially improved [Ste+15]. Thus biological models are developed with only partial information, whether due to this difficulty in obtaining measurements, lack of knowledge, or simplifications required for analysis feasibility. On the other hand, life's machinery works on a wide range of spatial and temporal scales, and, as outlined above, the processes involved may be very complicated, especially considered in unison. In some cases, emergence of high-level phenomena may not be possible to capture from elementary components [Str14]. Given all of this, establishing clear causality between the various biological factors and events via an e.g. structural causal model [Pea10] is difficult. On a fundamental level, many computational tasks in biology are known to be intrinsically difficult to perform, formally falling into the NP-complete or even NP-hard computational classes, such as multiple sequence alignment [WJ94], protein folding [HI97], protein-protein interaction network analysis [Kar11], and control of gene regulatory networks modeled as Boolean networks [Aku+07]. This implies the running time of the algorithms scales superpolynomially (and often exponentially) with the size of the input and thus solving for large datasets becomes infeasible (see Fig. 1.1). While for a select few hard problems we will likely see a great benefit from the coming paradigm of (practical) quantum computing, NP-complete and harder problems are believed to not be efficiently solvable on these types of machines [Aar07]. Another way to characterize difficult tasks, especially involving continuous quantities, is through the framework of optimization theory. Here one defines the task in terms of minimizing some cost associated to the problem (e.g. finding the folded state of a protein by searching for its lowest energy). In the case of difficult problems, the "landscape" this cost creates, as a function of the many parameters usually involved, is quite intricate and difficult to "traverse" in search of global minima and often only local minima are obtained [SK06].

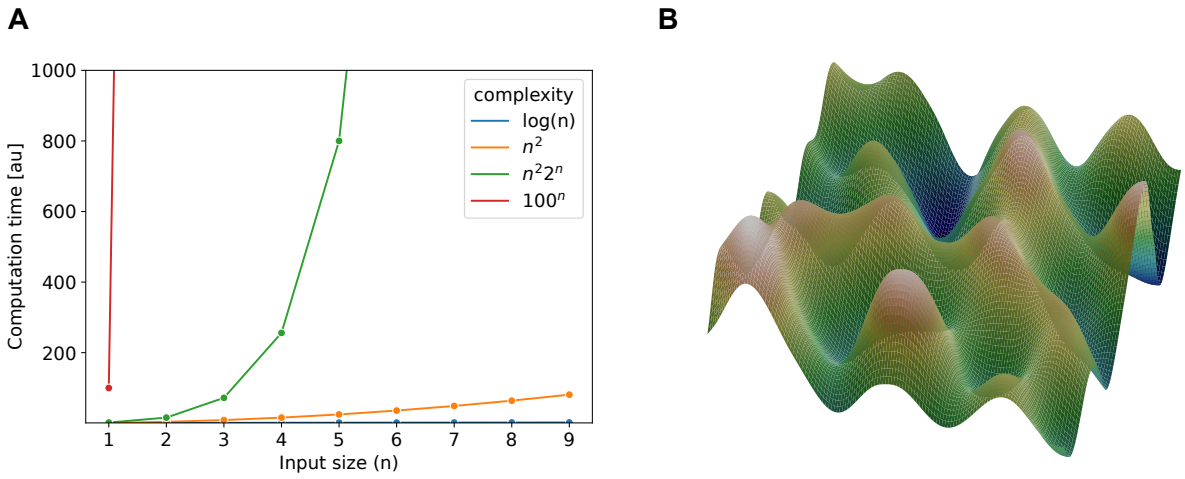


Figure 1.1: The optimum solution for many biological problems is intractable. **A)** The worst-case single CPU time cost (complexity) of solving some problems optimally (time units are arbitrary - au). Searching in a sorted list of length n takes $log(n)$ au, aligning 2 sequences of length n takes n^2 au, protein folding (formulated as an energy-minimization Hamiltonian path problem [HI97; HK62]) takes $n^2 2^n$ au, while aligning n sequences of length 100 takes 100^n au (exhaustively comparing all positions). To illustrate this last cost, if 1 au = 1 ms, the time required to guarantee an optimal alignment (worst case instance) of 100 sequence would take 10^{192} days. **B)** The cost landscape of many biological problems is quite complex and difficult to traverse towards a global optimum, even with stochastic exploration methods. The illustration shows a hypothetical cost landscape of two random variables, whereas typical problems involve many (hundreds) of variables.

For the above reasons concerning the complexity of some computational tasks required to study biological systems, we must therefore rely on heuristics and stochasticity to make these tasks tractable (for instance, by using artificial neural networks), abandoning the guarantee of finding globally optimum solutions, focusing instead on more narrow cases and relying on assumptions about the shape of the cost landscape in order to constrain the problem, while at the same time trying to avoid over-specialization (overfitting) to the datasets we use to derive solutions. However, these types of approaches often require knowledge of either the systems being studied, or at least, abstractly, a characterization of the quantitative variables that describe the systems. As outlined above, the former is a work in progress, while the latter requires experimentation and data acquisition. Within the realm of proteomics, there is great opportunity in the large amounts of data generated by high-throughput experiments from the past two decades [Deu+20]. The quantity and variation provide good grounds for statistical inference and model construction using machine learning.

The perspective of the thesis is this data-driven approach to characterizing the proteome, which relies on large collections of experimental data to support the distillation of empirical models and which serves to

complement hypothesis-driven research. This is because a data-driven approach is beneficial in a stage where much data exists, but mechanistic models are lacking. In terms of complexity and biological knowledge, data-driven techniques have the advantage that one may derive agnostic relations between observed quantities, bypassing the need for a detailed image of the structure of the system under investigation. On the other hand, mining for patterns in data is often itself a difficult task.

After briefly describing the proteome and going into more detail about the data-driven approach, as well as justifying the focus on sequence data for quantitative predictions of protein properties, the subsequent chapters present results from the included papers, covering measurement of protein quantities, learning protein features from their amino acid sequence, then, more generally, predicting expression levels from sequence data only, as well as sketching a meta-modeling approach to bridge the deep models that predict molecular quantities around the central dogma. Finally, a way to use such deep sequence models for protein engineering is presented.

1.2 The life and struggles of proteins

Proteins are arguably the most dynamic class of molecules in the cell, involved in both functional (catalysis, signaling, DNA replication, protein folding, transport) and structural (cytoskeleton) roles [Alb15]. In *Saccharomyces cerevisiae* (baker's yeast) they make up roughly half of the biomass [She02; Ono+17; Jac+19]. To properly understand the dynamics of the cell, one requires a thorough grasp of the proteome composition. However, proteins are complex molecules, both in terms of their structure and regulation, so to understand how the composition of the proteome arises, one needs an overall picture of protein synthesis, degradation, and regulation [VM12; MDR17]. Protein amount reflects their biological process: regulatory proteins (such as those involved in chromatin organization and transcriptional regulation) are rapidly degraded, while housekeeping proteins (such as metabolism) have long half-lives [VM12]. Beyond homeostasis, the significance of balance between these three processes is evident in signaling, where the complex intracellular logic of response to extracellular signals depends not only on the given concentration of a protein, but also on the speeds of degradation and synthesis, i.e. the protein turnover [Alb15]. More so complicated by the fact that the dynamic range of signal sensitivity of the different pathways varies considerably, with some responses depending on only small changes of signaling protein amount, while others requiring large changes [Alb15].

The *synthesis* of proteins is one of the biggest cellular energy

expenditures. In yeast, the fraction of all ATP assigned to it varies between 21% [Lah+17] and 50% for rapidly growing cells [LBA16]. This shows the importance of producing protein for the cell, and the high energy cost implies the optimization of its use through regulation and protein longevity. The information that encodes a protein is held in the coding sequence of its gene. The process of *transcription* consists in copying the information from DNA into messenger RNA (mRNA) strands that serve as the templates of protein synthesis through the process of *translation*, performed by ribosomes, and consisting in constructing a chain of amino acids. Each nucleotide triplet or *codon* in the mRNA sequence corresponds to a certain amino acid. The translation of proteins from mRNA consists of three main stages: initiation, elongation. and termination. Each of these steps is affected by a number of factors, among which the following have the highest known impact: mRNA level [VM12; Lah+17; HBB18], codon usage bias, ribosome density, amino acid composition, and hydrophobicity [Rib+19].

After translation, the amino acid chain folds into its final three-dimensional shape, often in steps, as longer proteins fold along conserved regions called *protein domains*, frequently assisted by helper proteins called *chaperones*. The various forces and chemical bond constraints acting between the amino acids is what determines the final shape, in a process that can be understood in terms of an energy minimization of the conformation of the protein. This shape is made up of structural motifs in the shape of sheets, helices, or coils, a classification referred to as *secondary structure*. It is the overall, *tertiary*, structure of a protein that gives it its function, either by itself or in a complex (see Fig. 1.2). Proteins may be characterized by the distribution of amino acid physicochemical properties, for example hydrophobic amino acids tend to be localized at the interior of proteins and, conversely, hydrophilic ones at the surface. The distributions of these properties, along with the shape and mechanical properties, determine their interaction with other molecules, often in specific fashion due to structural elements such as binding pockets [Alb15].

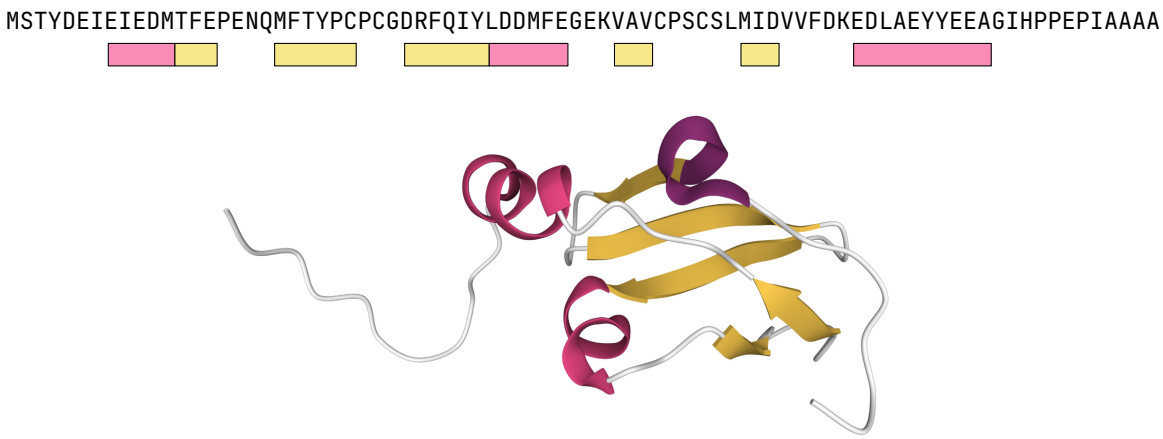


Figure 1.2: Protein structure. The amino acid sequence (primary structure) of the *S. cerevisiae* Diphthamide biosynthesis protein 3 (UniProt ID: Q3E840), with secondary structure annotation and 3D structure, the latter two from PDB entry 1YOP, based on nuclear magnetic resonance spectroscopy [Sun+05]. Pink stretches denote helices and yellow stretches denote sheets. 3D image created using Mol* Viewer [Seh+21]

Degradation is known to be (weakly) influenced by disordered protein regions, the length of β-sheets, and by so-called *degron* short sequence motifs [VM12; MDR17; MV17]. Unfortunately, this process has received far less attention than synthesis [VM12]. The organelle responsible for degrading proteins is the *lysosome* and its behavior is also quite complex. Besides its role in degrading protein, this organelle reacts to energetic conditions in the cell by recruiting various factors to its surface, for example promoting anabolic processes in the presence of high nutrient levels by recruiting mTORC1, a protein complex that promotes translation [PMB21]. Its membrane houses about 60 types of digestive enzymes and some disorders result when certain of these are not present in the lysosome, as a results of defects in their genes, thus the accumulation of would-be target proteins. In other cases, such as inclusion-cell disease, a lack of proper sorting via glycosylation of most of these enzymes (due to a defective phosphotransferase) results in them being secreted rather than transported to and kept in the lysosome [Alb15]. However, there are cases of undegraded molecules that are not associated to defective enzymes, as well as problems with signaling and vesicle trafficking, hinting at more complex relationships in the overall autophagy-lysosomal pathway [PMB21].

Many proteins undergo *post-translational modifications* (PTMs), whether as a normal part of their function or as a result of e.g. oxidative stress. While the latter is one example of modifications linked to pathological states, PTMs are essential for enzyme activation and deactivation, protein localization, degradation, and transmitting signals [Alb15]. Thus, modifications serve proteomic regulation and in order to

properly characterize the state of a cell, one needs to assess these in a precise way.

When considering the molecular flow from transcription to translation, the relation between transcript level and protein abundance is complex, with various factors such as translation rates (as a function of mRNA structure), translation rate modulation (via RNA-binding proteins and noncoding RNAs [Ho+21] or ribosome availability), the modulation of protein half-life via degradation, and protein transport [LBA16] In general, mRNA half-life is an order of magnitude shorter than protein in mammalian cells and two orders in bacteria [VM12]. The importance of these additional translational and post-translational processes depends on the organism, biological system, and conditions, with mechanisms such as the unfolded protein response acting to halt translation and remove misfolded protein under endoplasmatic reticulum stress [LA16]. Indeed, the image put forward by the authors is that relative contribution of mRNA to protein abundance is roughly inversely proportional to the strength of perturbation of the cell, with steady state levels being largely determined by mRNA. The work in this thesis has been concerned with bulk quantities in cell populations, not single cell measurements and behaviors. Thus, transient perturbations or short-lived changes in the proteome due to e.g. cellular division are averaged out.

All these processes and factors offer a diverse quantitative assessment of the proteome and, while some are better characterized than others, there is already a large collection of useful datasets available for analysis with data-driven approaches [Deu+20]. The articles included in this thesis have sought to find connections between the information captured by the data-driven models and the various processes and protein properties outlined above.

1.3 The data-driven approach and machine learning

All too often, one starts the journey of scientific understanding with limited knowledge, a handful of hypotheses, and a large collection of data that are to be scrutinized for patterns and phenomenological laws. These laws serve as the basis for distilling models and generating new hypotheses, leading thus back to experimentation and data collection. Indeed, this sort of development generally appears to be rather cyclical in science in general.

The abundance of biological data (e.g. the capture of almost full proteomes in single mass spectrometric runs [Nav+16; Vow+18; Mes+21]) is fueling a paradigm shift towards data-driven methods. This is a rather

broad category, but it usually involves a philosophy of induction and regression, complemented by validation through recapitulation of known results, then prediction and experimental testing [Dha13; Leo20]. This contrasts (but also complements) hypothesis-driven methods, in which mechanistic models are built from *a priori* knowledge and hypotheses about the structure of the biological systems, either from first principles or more abstract levels (e.g. differential equations, metabolic networks) [FH20]. One purported benefit of the data-driven approach is avoiding preconceptions regarding the nature of the system under study [Leo20], or at least suspending them until patterns in the data can be inspected and cross-checked with existing hypotheses. Conversely, new hypotheses may be generated from such patterns, especially in initial, exploratory studies.

Figure 1.3: The data-driven approach complements hypothesis-driven investigation. Illustration: *Drawing Hands* by M.C. Escher. All M.C. Escher works (C) 2021 The M.C. Escher Company - the Netherlands. All rights reserved. Used by permission. www.mcescher.com

To put it into perspective, this methodology is nothing new, only brought back into focus due to recent jumps in computing power (massive parallelism, graphical processing unit improvements) and the democratization of powerful data processing and machine learning software through open source distribution (e.g. Tensorflow), a trend starting roughly in the early 2010s. To give a sense of scale, our current technological capacity allows for near-routine analysis of data sets on the order of hundreds of terabytes in the order of hours. Historically, sciences like astronomy relied heavily on data accumulation and induction (observation of trajectories), such as Kepler's laws of planetary motion (derived from Tycho Brahe's detailed records) [Thu94], prior to the derivation of mechanistic models. Other examples of phenomenological

models throughout science are Newton's universal law of gravitation, Balmer's equation for the emission spectrum of atomic hydrogen [Bok11], and Monod's bacterial growth rate model as a function of a limiting substrate [El-12].

The act of inferring from observation is carried out routinely today with machine learning, which is a collection of methodologies used to look for statistically significant variable associations and patterns in potentially heterogenous data, and aimed to build models with predictive power [Dha13]. Machine learning concepts and methodologies are treated in depth elsewhere [HTF09; GBC16]. In this work I will merely point out salient aspects of the models under discussion. More formally in statistical learning theory, one speaks of a *learning task* defined on a set of variables or *features* (e.g. temperature, nutrient concentrations) with either the objective of finding a relation between said features and an outcome or *target* variable (e.g. protein abundance) - referred to as *supervised learning*, or discovering clusters or other types of patterns in the features that may hint towards structure within the system under investigation [HTF09] - referred to as *unsupervised learning*. Both features and outcome variables may be quantitative or qualitative (e.g. discrete classes). Finding such patterns and relation models is nontrivial and proper data processing and statistical methods must be employed.

Supervised learning typically consists of starting with an *a priori* decided model class, namely a parametrized function between feature domain and target codomain (e.g. expressing outcome as a polynomial combination of input features). Various types of iterative methods are used to adjust the parameter values (e.g. polynomial coefficients) so that the model predictions best match the example feature data, a process referred to as *fitting* [HTF09]. In the case of unsupervised learning, a task consists of inferring the properties of the joint distribution of input features, i.e. how the observations are structured. Often one performs a clustering of the features or fits a type of algebraic decomposition of the features (e.g. principal component decomposition, factorization). The former seeks to explain the data using combinations of simpler distributions (the clusters), while the latter seeks to identify lower-dimensional manifolds that capture the most feature density and explain the data as combinations of such latent variables. Often there is no clear general measure of the quality of this type of learning as one no longer relies on any "ground truth" in the shape of targets. The evaluation thus becomes heavily domain-specific, but is routinely used to explore data for structure, generate or support hypotheses, and extract novel associations [HTF09].

For both types of learning tasks, the advent of deep (many-layered) neural networks has brought clear progress across a variety of fields, fueled

by the technological advances previously mentioned [GBC16]. Neural networks rely on the composition of a large ensemble of simple nonlinear functions (neurons or *units*) to obtain a more complex mapping between feature and target space (for a supervised task). When these units are organized in layers and the network features "many" such layers (rather arbitrary measure, depending on the precise network architecture, but often tens of layers), the consensus is that layers will learn increasingly complex or high-level associations between features, based on this layered hierarchy of function composition, parametrized by millions of inter-unit connection weights [GBC16].

While there is a large variety of network architectures, the work in this thesis mainly concerns itself with convolutional (CNN) and Transformer-type networks. CNNs were designed for input data with a grid-like structure, such as images or time series, and their deeper incarnations have seen much success especially in image-processing tasks [GBC16], but also in various biological tasks [Ang+16; LCC19; Tra+19; Gai+20; BV20; Zri+21]. In CNNs, small regions of the input are convolved with learned kernels (in the image processing sense of the word, essentially weight matrices, also referred to as *filters*). As each layer feeds into the next, the input "cones" (*receptive fields*) of the convolutions will span the entire input. Additionally, sub-sampling is performed between layers, which enforces translational invariance, meaning features in the input will be recognized regardless of their position. See Fig. 1.4 for a simplified diagram of a CNN.

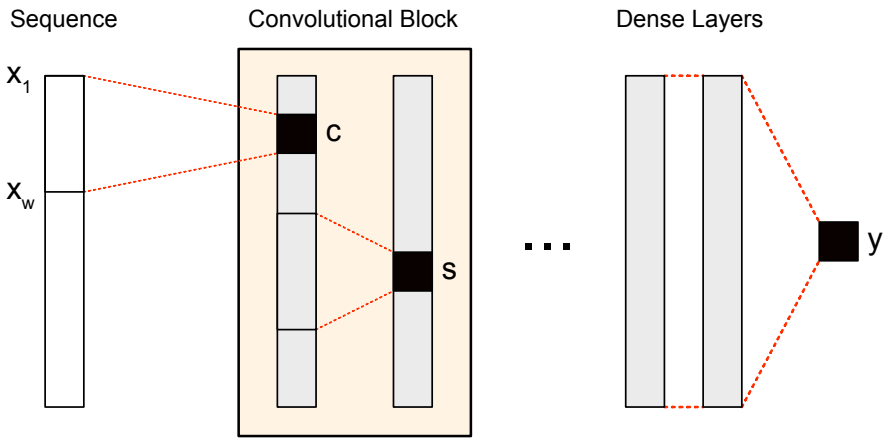


Figure 1.4: Simplified diagram of a typical convolutional neural network (CNN). Convolutional blocks consist of alternating convolutional layers and downsampling layers. The input sequence is x and the output value is y. Each output c of units in the convolutional layer results from the convolution of an input region of width w with a learned kernel (weight matrix) k, i.e. $\phi\left(\sum_{i=1}^{w} x_i k_i\right)$, where ϕ is a nonlinear activation (gating) function. The nonlinear downsampling s is used as a strong prior to enforce translational invariance. The final dense layers are fully connected, that is the input of each unit is the entire previous layer's width.

Transformer networks differ from CNNs in a number of ways, but their distinguishing feature is the *attention* mechanism they employ in the majority of layers. Attention values give the relative importance of pairs of variables in the input, e.g. pairs of words in a sentence. Whereas the convolutions in CNNs span small regions of the input, attention nodes receive information from the entire input and learn association strengths between pairs of words in a sentence. Thus, these types of networks are explicitly designed to capture long-distance relationships between sequence elements. The attention mechanism aims to have the network learn a representation space or *word embedding* (or rather, a transformation of the input into this space) in which words of similar meaning are closer together and functionally connected pairs of words have high attention values [Vas+17; Vig19]. "Meaning" here is application-dependent and may reflect how words across languages are similar in function when performing translation, or how different words or structures convey the same meaning in emotion detection tasks [RKR20]. See Fig. 1.5 for a simplified diagram of a Transformer network. For more information on this quite intricate architecture, see the original article [Vas+17] for more details. In **Paper III**, where such a network was used to predict protein abundance from amino acid sequence, my assumption was that the structure of this embedded representation space reflects the ordering of the predicted value and the "meaning" (or semantics) is the protein abundance. Based on this, the space was probed and the structure exploited to perform

guided mutation aiming at increasing predicted protein abundance.

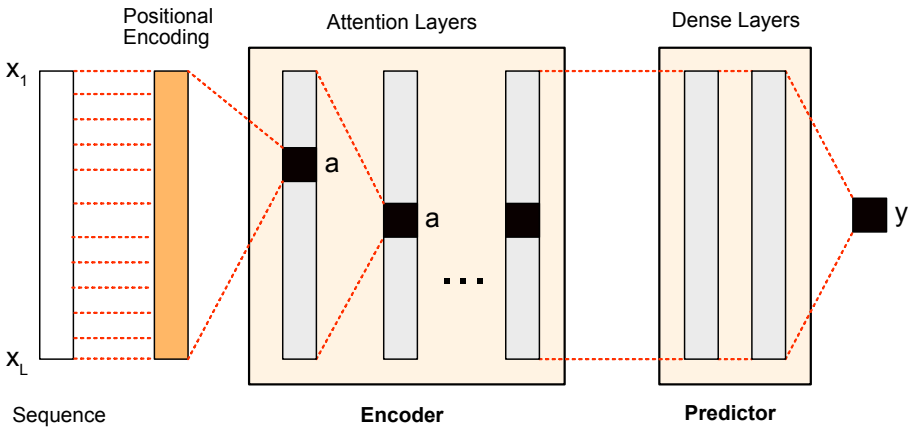


Figure 1.5: Simplified diagram of a Transformer-type network. This in particular is the BERT sequence-to-value network used in **Paper III**, while typical sequence-to-sequence Transformers have a symmetrical Decoder instead of the Predictor. The initial positional encoding captures information about the order of letters in the input sequence x. The attention layers learn directional associations between pairs of letters, each unit a having input spanning the entire sequence. Finally, a few dense (fully connected) layers produce the final single-value output y.

The motivation and one reason for the proven versatility of deep neural network is the universal approximation theorem, which states that even rather shallow networks may approximate any continuous (Borel-measurable) function from features to targets, provided the layers are "wide" enough and training proceeds successfully. While the currently known worst-case bounds for the necessary amount of units is generally exponential in the number of features, for constrained classes of functions and input domains, it is expected that the number of units required for good results is considerably smaller [GBC16].

There are a few important things to keep in mind when considering best practices of machine learning. Learned models must *generalize* as much as possible beyond the example feature data they were exposed to. *Overfitting* to the example set is increasingly likely with more complex models, as the high number of degrees of freedom allows for very close fitting to the provided examples, but with a drastic loss of generality. Generally speaking, in order to reduce generalization error, models of appropriate complexity must be considered (e.g. perhaps more than linear, yet constrained by the principle of parsimony) and large enough sample sizes should be acquired, to reduce estimate bias. The overfitting behavior may also be expressed as a tradeoff between the bias and variance of the model (trained predictor function): with higher complexity of this function, the lower the bias but higher the variance [HTF09]. See Fig. 1.6 for an

illustration of how these concepts relate. Assuming the available sample data is a good representation of a larger population, besides constraining the complexity of the model, a common technique to test generality is to use cross-validation, consisting in putting aside a small portion of the dataset on which to test the prediction performance of the model after a certain number of training iterations.

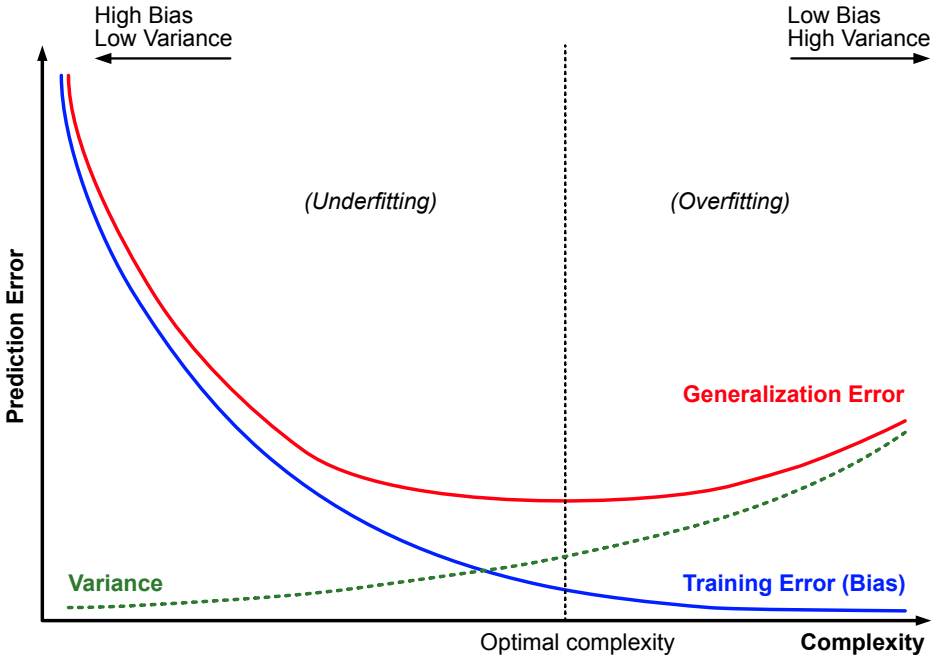


Figure 1.6: Bias-variance tradeoff, illustrated in an idealized model selection scenario. When considering model complexity using a cross-validation assessment of training, there is a theoretical optimum that balances bias and variance of the model's prediction. With an increase in complexity, bias (training error) and generalization error both drop, but past an optimum point, generalization error and variance increase significantly, as one obtains overfitted models, tailored to the training set but unable to generalize.

Given the complexity of deep models, a large amount of data is thus desirable for good performance. However, another difficulty arises from using many feature variables, stemming from the exponential increase in volume of the space spanned by these variables. Referred to as the "curse of dimensionality", there are two major consequences of dealing with these high-dimensional spaces. One is the fact that the sample date becomes quite sparse in relation to the volume of space, an exhaustive sampling requiring an exponential increase in the combinations of feature values. The other is that in high-dimensional space, most commonly used distances between points tend to be effectively equal, which poses obvious issues for clustering procedures [HTF09]. The typical approaches to reduce this issue involve pruning redundant features, if there are any, and reducing the number of dimensions, either linearly (e.g. with principal component

analysis, by choosing components with most variance) or nonlinearly (e.g. with t-SNE [MH08] or UMAP [MHM18], which seek to recover local topology between points, at the cost of global distortion). Part of the success of deep learning is that it was developed to better deal with high-dimensional data and in many cases feature selection is not required [GBC16].

Finally, it is worth keeping in mind the human factor, as we are prone to many cognitive biases, including a preference for simplistic explanations. Moreover, a researcher or community may have certain preconceptions about the relations of the variables under study. The data-first perspective cannot completely prevent this, but the methodology does not require *a priori* assumptions on the mathematical structure of the data and, moreover, may yield surprising relationships [Dha13]. Needless to say, there will always be some form of "bias": to understand a thing, it must be placed within an existing ontological framework (whether this is a formal endeavor or implicit thinking), thus subject to relativism. Indeed, formulating theories and testing them requires an *a priori* framework and paradigm shifts may occur when contradictions are identified within a given framework, causing a revisit of axioms, assumptions, and so on.

1.4 The case for relying on sequence in building models

Sequence encodes the necessary information to create and maintain a cell and may be obtained relatively easily. With the reduced costs of present day sequencing (as low as US cents per million base pairs [Kul16]), sequence is one of the most abundant types of biological data. For example, genomic data is projected by 2025 to require between 2 and 40 million terabytes storage per year and in 2015 about 32000 microbial, 5000 plant, and 250000 individual human genomes were stored by the US NIH/NCBI-maintained Sequence Read Archive alone [Ste+15]. In terms of phenotypic data, many fields have transitioned to high-throughput methodologies, as is the case for proteomics, with single experiments generating on the order of hundreds of gigabytes of data [Zel+18] and the total size of ProteomeXchange consortium datasets exceeding 1 million gigabytes as of 2019, spanning over 14000 datasets [Deu+20].

The moment-to-moment state of the cell is determined by the interaction between the functional ensemble of components that make up the cell, and the environment (potentially including other cells), as well as some inevitable stochasticity [Alb15]. Concerning the cell alone, the various components and their interactions ultimately derive

from sequence information, albeit in an emergent way, confounding simple or reductionistic characterizations in terms of sequence elements (e.g. composition of nucleotides or amino acids, motifs, regulatory regions) [Nie17].

Sequence information thus appears as an abundant "primary" source of understanding and prediction. Coupled with the aforementioned recent technological leaps, data-driven characterization of phenotype from sequence information has seen many breakthrough results, including learning regulatory interactions at the nucleotide level [Zri+20a; Zri+21]. When learning occurs from sequence data, salient features consisting of patterns within the input sequences are distilled across the layers (and implicitly, the high-dimensional spaces) that make up the network model [GBC16]. Importantly, assumptions about the physical process are not required, though they may serve as validation. Since deep learning alleviates the need to reduce the number of features, sequences need not be summarized by quantitative variables such as abundance of amino acids or k-mers, physicochemical values, or adaptation indices. Besides the extra amount of effort, feature selection in classical machine learning methods may lose information by removing features that are not fully redundant (especially by relying solely on linear measures such as correlation).

With respect to the proteome, there are many studies linking various regions of a gene to molecular quantities using data-driven machine learning. Most have been classic ("shallow") models that numerically summarize sequences or used derived features such as physicochemical properties. Relatively few have utilized sequences as-is, either due to the limitation of the classic machine learning models used or due to technical limitations (primarily memory) of the hardware being used at that time.

Using the sequence as input directly (effectively assigning each letter as a separate variable) has become feasible with deep models, however. The model in **Paper IV** for example considers in total 2150 base pairs for regulatory regions. If the model were extended to include full coding sequences of say 6000 base pairs, covering the majority of yeast genes, this would result in 8150 input variables. While these types of network models would be quite large, depending on the actual architecture, the required memory could however be accommodated if not on a typical consumer-grade GPUs (32 GB of RAM for two paired GPUs at the time of writing), then certainly on the more powerful configurations of datacenter-grade GPUs increasingly available to the research community, such as the NVIDIA A100 with up to 80 GB of RAM per each card. The information content of the sequence may thus be kept intact and various types of neural network architectures may be used to learn relevant associations rules between these letters. We have referred to these learned

association rules as "grammars" [Zri+20a; Zri+21], both for the intuitive explanation but also because much of this work borrows heavily from natural language processing.

Protein abundance of a single yeast gene, for instance, has been predicted directly from 5' UTR sequences with a coefficient of determination R^2 of 62% [Cup+17]. (Unless otherwise stated, all R^2 values in this thesis are reported over held-out test sets.) The authors trained a CNN model from experimental measurements using about 0.5 million 50-base-pair 5' UTR sequences for the gene HIS3. While this produced good results and the authors argue for generality across genes, in **Paper IV** we showed evidence that the different gene regions are a part of a co-evolved unit and all contribute towards transcription, thus underlining the need for comprehensive, gene-specific data, though this a difficult experimental undertaking. Regardless, taking advantage of the sequence is also underlined by the authors [Cup+17], which note that more complex shallow models without positional information perform more poorly than simpler ones with this information. By probing the trained network filters, they extracted significant motifs informing abundance.

Another study using a random forest (i.e. "shallow") model to predict protein abundance from 5' UTR, for a single *E. coli* gene, achieved an R^2 of 82% [Bon+16]. The authors experimentally measured the effect in abundance difference for about 3000 Shine–Dalgarno sequences (ribosomal binding sites in bacteria, contained in the 5' UTR). The sequences were numerically summarized with hybridization energy between these and 16S rRNA. In contrast, a hypothesis-based thermodynamical model using ribosome-mRNA binding energy as input achieved an R^2 of 54% on *E. coli* genes [SMV09].

While various quantities have been predicted for transcription and translation from regulatory region sequence, with both classic and deep models, relatively few studies considered interaction between all regions using sequence [Zri+21], which served as motivation for **Paper IV** [Zri+20a].

The coding sequence has been used to predict protein abundance by a few studies using classical machine learning models. [Fer+21] used codon usage to differentiate between the highest and lowest 10% yeast protein by abundance in terms of codon frequency using principal component analysis. They trained an AdaBoost model on codon metrics (without positional information) to predict protein abundance and achieved a Spearman rank correlation between prediction and experimental values of $\rho = 0.74$. The same model trained on *S. cerevisiae*. was used to predict abundance for *E. coli*, *Schizosaccharomyces pombe*, and *Kluyveromyces marxianus* genes, with correlations to experimental values of 0.5, 0.7, and 0.62, respectively.

The amino acid sequence has been used in a few studies to predict several protein aspects, using deep models developed for natural language processing. For example, in terms of structural models, secondary protein structure was predicted with an accuracy of 0.75 and contact maps with a precision of 0.4 (measured with a metric used in the Critical Assessment of Methods of Protein Structure prediction competition, with 1 as maximum) [Rao+19]. Higher results were obtained by including alignment information (0.8 and 0.64, respectively), whereas the former were from sequence alone. In terms of quantitative associations, fluorescence intensity was predicted with a Spearman ρ of 0.68, and protein stability with a ρ of 0.73 [Rao+19]. **Papers II and III** in this thesis are concerned with this type of association, relating amino acid sequence to optimal catalytic temperature and protein abundance, respectively.

While in some cases the deep sequence-based models have lower performance than their shallow counterparts, this can certainly be improved with better architectures, allowing the networks to learn the numeric features that were given as input to the shallow models (recapitulating this information). While this might seem like overhead for such cases, this approach has two major advantages. From a practical point of view, it accelerates studies by removing the need to select features (and to iterate evaluation over combinations of such features). But more importantly, by learning rules in the sequences themselves, one may discover determinants of the target variables. For example, what sort of amino acid combinations are associated with high-abundance proteins? Even if there are cases where the performance is not greater than shallow models using proxy feature variables, this benefit alone is worth the training of sequence-based models. And as I outline in the next section on model interpretability, there is great effort being put into developing models and probing procedures to extract meaningful rules out of deep models. This, coupled with the improvements in computing power mentioned above, makes "direct-from-sequence" models quite attractive.

Nonetheless, the complex nature of the processes involved may require infeasible deep models (in terms of size and/or training time) should one aim for highly detailed characterizations of phenotype from DNA sequence, especially given that phenotype arises additionally from interaction with the environment, various emergent behaviors, as well as from the compositional inheritance of the cell from its mother (i.e. no cell is created *ex nihilo* from sequence, there is always a "bootstrapped" intracellular context). Given this barrier, an obvious approach is to rely on integrations of multiple models that capture different processes, systems, or abstract properties. I outline a simple example of this approach in Chapter 4.

1.5 Model interpretation and protein properties

Once a phenomenological association between observables has been achieved with a data-driven method, the next phase of inquiry is explaining how the variables in question contribute to yield target values (for supervised tasks) or what distinguishing features characterize clusters of data (in unsupervised tasks). Limiting the discussion to supervised models, one usually wishes to find the main determinants (groups of variables) of the predicted values, as well as how these associate. This implies a simplification of the trained model into a set of human-understandable or at least more intuitive and wieldy rules or equations, as the explicit mathematical formulation of the e.g. deep neural network is in fact available but defies simple description.

For the current generation of deep models, a major drawback is that they are essentially black boxes that offer little transparency and even less human-understandable insights into the meaning and behavior of the patterns they have learned. At best one can describe neural networks in very abstract terms concerning how the groups of functions being composed process their inputs, but the predictive power of this approach is very small and usually serves as a guide for network design [GBC16]. Still, attempts are being made in the broader machine learning community to develop interpretable models. Here, there are subtle differences in aims and it is helpful to distinguish between *transparency* and *interpretability* [Lip17; WP19]. The former concerns itself with the possibility of inspecting the "inner workings" of models and assigning familiar concepts and mathematical constructs to the various components of the models. The latter refers to offering human-understandable explanations for the way in which the models function. Some authors also distinguish *explainability* [WP19] as the capacity for a model to offer its decision-making process, which would be especially valuable for medical applications.

Within this framing, *transparency* of deep models is an easier task and achieved to some extent in this work, especially in **Paper III**, with post-hoc analyses in the form of parameter inspection and probing of the high-dimensional space the model uses internally to represent transformations of the input data. The latter goal of *interpretation* is much harder, as one is faced on the one hand with the challenge of framing exceedingly complex functionality within familiar human metaphors, and on the other hand, with the emergent behavior of these models. This second issue is important to keep in mind, as it can to some extent undermine the goals of transparency and explainability (when "the whole is greater

than the sum of its parts") and, more importantly, it raises the bar of understanding beyond reductionistic explanations. Thus, ideally a systems type of thinking and even perhaps new theoretical frameworks would be in order.

While aiming for at least transparency, a series of probing techniques are normally used to determine the behavior of deep networks. *In silico* perturbation experiments may be performed with virtually any type of model to probe the relevance of each variable. For images and sequences, this typically involves systematically covering or "occluding" regions of an image or sequence to find important contributions to the prediction [ZF14; Zri+20a]. In convolutional neural networks, one may inspect layer activation (i.e. which units "fire") in response to input, though this tends to be very sparse. One can also inspect the weights of the convolutional kernels to see what types of features they have learned to recognize (for images, first layers usually learn simpler shapes such as lines, while deeper layers learn more complex shapes) [KSH12; ZF14; GBC16]. One major issue with such approaches is that their extraction of salient features in the input data is easy to verify when one is working with images. Thus one can distinguish a probing technique that is able to capture useful information for interpretation, whereas with e.g. protein sequences one essentially does not know what they are looking for. At best, correlations can be made to various protein properties.

For attention-based neural networks such as the one used in **Paper III**, the weight assigned to each variable (the attention) may be directly inspected. The attention mechanism was created to capture salient associations between words in a sentence [Vas+17] and appears as a better way to gauge both the importance of amino acids and the strength of interactions towards producing the predicted target value [Vig+20].

Other probing approaches seek to capture the structure of the high-dimensional embedded (or "latent") space in the interior part of the network, to recover associations between words in a sentence, in the shape of syntactical (sentence parsing) or semantical (word-association) structures. These probes however can be quite complex (even neural networks themselves) and a major concern is that the patterns they recover may be more specific to the probes than to the neural network under investigation [RKR20]. Depending on the aim, points from this space can be nonlinearly reduced to a few dimensions using t-SNE or UMAP in the hope that clusters may be recovered (e.g. sequences with low or high predicted value cluster together). In **Paper III**, I have built such a nonlinear reduction to support guided protein mutation. In natural language processing models, parse trees describing the structure of the sentence in terms of functional roles of words have been built from the

high-dimensional embedded space [RKR20].

Within existing work and typical practice today, there is often the tacit assumption that only a handful of feature variables provided in the input are highly impactful for the end result. Given the complexity of the model, this is also likely the simplest part in terms of interpretation, as perturbation experiments often identify these variables. Such an investigation was for example used in **Paper II**, where parts of the sequence were occluded to determine the impact of those amino acids towards the prediction. Then the natural question is how these salient features jointly determine the prediction. At this stage, measurements of variable relevance such as perturbation profiles, activation maps, or convolutional filters are correlated with known quantities. In the case of sequence models, these profiles may be correlated with physicochemical properties of the polypeptide chain, for instance. Moreover, mechanistic models may be tested against these determinant variables, besides just deep model predictions.

In terms of finding simpler, human-understandable models that still retain a sufficiently close predictive power to the deep model, some success has been obtained recently in deriving symbolic expressions of the relations between variables that best explain the trained model [Cra+20]. In this study, known force laws (such as Newton's law of gravitation) were recovered from simulations based on these laws. While this area of symbolic regression appears quite promising, it still requires network design informed by the nature of the process being modeled and, moreover, its demonstrated performance (and role for interpretability) has been on low-dimensional inputs, raising the question of how to further interpret equations of thousands of variables for e.g. sequence models (i.e. a transparent, yet likely not interpretable result).

For the scope of this thesis, the type of protein properties used in attempting to interpret deep models are quickly summarized below. Needless to say, determining such properties relies heavily on experimental work and, coupled with any hypotheses that may be formulated at this point, brings us to closing the loop back to experiment.

Physicochemical properties of the protein, such as hydrophobicity, emerge from the combination, relative position, and interactions of individual amino acids that form the polypeptide chain. Thus, each protein can be seen as a landscape of these properties, with various levels of detail informing various functional roles of the protein or its substructures (e.g. a binding pocket). It is understood that these landscapes are strongly determined through evolution to satisfy the protein's function within the overall cellular physiology [Alb15].

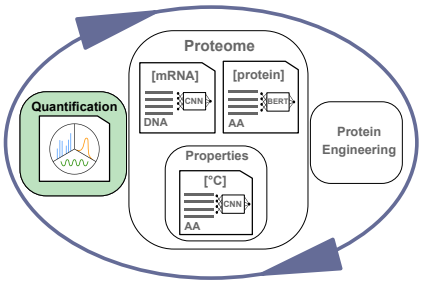
Given that proteins fulfill both a functional and structural role, their

three-dimensional shape and constituent motifs (sheets, helices, etc.), along with *mechanical properties* provide important information about the functioning of polypeptides and where one might try to intervene to alter the operation of these biological machines. The recent breakthrough in tertiary structure prediction from amino acid sequence [Jum+21] opens up many exciting possibilities of constructing accurate *in silico* representations of proteins and their physicochemical properties.

Sequence properties then refer to numerical characterizations of the amino acid sequence, be it residue frequencies, positional patterns, sequence complexity, residue association rules, or statistically significant motifs. Additionally, such measures may be linked to protein domains.

Within this area of research there are many largely reductionistic studies that attribute protein abundance to specific such properties, with rather weak associations based usually on correlation (i.e. linear relations) [Rib+19; Ver+19; DML19; Cha+20; Web+20]. While each feature contributes to the overall function of a protein, one would wish an integrative approach towards understanding the overall physiology of the cell, hence the expected benefit of deep models and the data-driven approach.

2 | Measurement (Paper I)



It is essential to have precise and in-depth quantification when mapping out a biological system, more so if one is aiming at a time-resolved picture. For instance, to assess cellular regulation under stress, frequent proteomic snapshots are desirable, especially since the protein abundance plays a greater role than that of transcript in such perturbed conditions [LA16; LBA16].

Mass spectrometry, often coupled with a physical separation step such as liquid chromatography, is the workhorse of proteomics and innovations in the past ten years have allowed relatively rapid and cheap quantification through high-throughput platforms. This has proven extremely valuable, especially in crisis situations like the current viral pandemic, where blood, plasma, serum, and immune cell samples from large cohorts have been used to identify biomarkers for diagnostic and prognostic purposes, to understand cellular mechanisms, and to support treatment development [Spe+20; Dem+21; Mes+21]. Towards these aims, especially in initial stages, data-driven approaches have been employed [Spe+20], for example to construct a time-resolved SARS-CoV-2 phenotype evolution map, from which prognostic signatures were identified with machine learning [Dem+21].

The field of mass spectrometry is quite intricate and a detailed treatment requires understanding of the instruments, the analytical techniques, as well as the statistics involved. In this chapter, I will introduce technical details as relevant to the discussion. For an introduction to the field, the tutorial by [Smi+14] gives a good overview.

In terms of proteome quantification, the high-throughput experimental platforms have shifted the burden to downstream analytics, as the data produced is quite large and, more problematically, highly dense and consisting of many overlapped signals spanning a wide variety of dynamic ranges, making the recovery of analyte quantities very difficult. [PMW19; Dem+19]. Proteins with low copy numbers can be especially problematic, as they are close to baseline signal levels [Nav+16; BS20], yet important for mapping heterogeneous systems such as the nervous system and in single-cell contexts [GA21]. Thus new techniques that leverage the entirety

of the data in an efficient way are required [Lik09; Zha+; Deu+18].

My work in **Paper I** has focused on processing the dense scans from data-independent acquisition (DIA) runs, using an unsupervised decomposition method that yields separate molecular signal fingerprints. The methodology itself is elaborated in the section below. The work consisted in building a parallelized, GPU-accelerated pipeline called CANDIA that can process large scan series in the order of hundreds of GB in a matter of hours. While the decomposition method itself has been used before, up to my implementation it was not feasible on DIA data, as existing use cases were using single-machine, CPU-bound software implementations on small datasets (in the order of MB) consisting in a small number (tens or a few hundreds) of analytes, whereas in high-throughput DIA scans one expects tens of thousands of peptides and a single scan is around 10 GB. Moreover, they required a human specialist to validate parameter choices by visual inspection, which is not tenable for the large DIA runs, whereas CANDIA is fully automated.

DIA platforms [Ven+04] strike a compromise between the high precision but low throughput of the gold standard selected reaction monitoring (SRM) technique, and the high throughput but low precision of shotgun techniques [ZKZ16]. The "data-independent" qualifier refer to the fact that the instrument performs an exhaustive scan of its mass range, regardless of where the highest peaks occur in survey scans, as is the case with shotgun proteomics. The SWATH-MS platform [Lud+18], introduced about ten years ago [Gil+12] has in particular proven quite performant, achieving good precision and reproducibility, as well as high throughput [Col+17; Ros+17; Vow+18]. Due to the high overlap of signal, the standard approach to identification and quantification relies on an *a priori* spectral library consisting of fingerprint signals (or spectra) of expected protein fragments, which are checked against the DIA scans. This however only recovers a small percentage of total signal, just over 2% on the benchmark dataset used in **Paper I** (Fig 3A in the article), the large remaining fraction referred to as proteomic "dark matter". Moreover, constructing libraries incurs experimental cost, as they require additional shotgun mass spectrometric runs, the process being specific to the instrument and conditions [Sch+15].

The CANDIA pipeline functions as a preprocessing and data extraction step for DIA scans series, upstream of conventional database search engines used in shotgun proteomics such as Crux [McI+14], TPP [Deu+15], and MS-GF+ [KP14]. Briefly, such tools use a database of protein sequences to match spectra in the shotgun scans with peptides (protein fragments) from the database. Additionally, CANDIA performs a summarization of the entire dataset, outputting a single file containing the deconvolved

spectra present across all scans. As I will expand upon in the technical section at the end of this chapter, one may obtain from CANDIA the contribution of a deconvolved spectrum to each scan, thus distinguishing between potentially varying quantities of the corresponding analyte or even its absence across scans. This way of extracting a "trace" of an analyte's amount in each scan alleviates the issue of missing identifications, which may otherwise even happen within technical replicates, as demonstrated in the paper (Fig. 2A). While this trace information (the sample mode, as explained below) proved to be too imprecise to directly obtain high quality quantification from it, this is theoretically possible and a point of future improvement. Instead, bypassing the need for additional mass spectrometric runs, CANDIA builds a spectral library directly from its output using DIA-NN [Dem+19], a tool that uses machine learning to distinguish between true and false positive matches, and can also generate and export a spectral library from a sequence database and a spectral input as provided by CANDIA. It should be noted that software such as DIA-NN, PECAN [Tin+17], and Spectronaut [Bru+15] that have (peptide-centric) "library-free" modes generate their own libraries *in silico* as part of their execution (based on e.g. simulated or machine-learned predictions of spectra for a given peptide). They still however rely on the collection of spectra within this generated library to match against the DIA scans, whereas the decomposition method in CANDIA extracts analyte signals in an untargeted fashion. In terms of procedure compatibility, the CANDIA / DIA-NN library may be used in standard DIA quantification software such as Spectronaut, Skyline [Pin+20], OpenSWATH [Rös+14], and DIA-NN itself.

Results produced using CANDIA have high precision and lower false positive rates compared to conventional alternatives. While the number of identified proteins is comparable with established software, more unique proteins to the CANDIA workflow were identified both in our yeast lysate dataset, as well as in the more complex (mixed-organism) LFQbench [Nav+16] benchmark dataset (Fig. 2A, C in **Paper I**). Quantification using a CANDIA / DIA-NN library is precise (coefficient of variation CV = 9.3% on our yeast technical replicate dataset). Twice more post-translational modifications were identified by MS-GF+ running on CANDIA output, and at higher prevalence across technical replicates, compared to the alternative approach of running MS-GF+ on output from DIA-Umpire [Tso+15]. This latter established tool extracts pseudo-spectra from DIA data by detecting covarying molecular precursor-fragment signal groups, which may be used by shotgun search engines. The main difference is that DIA-Umpire does not perform true decomposition and generates an output file per each scan, whereas the PARAFAC method in CANDIA uses the variation across all

scans to generate its decomposed output.

Additionally, *de novo* sequencing results were improved by running Novor [Ma15] and DeepNovo [Tra+17] on CANDIA output, the former showing a 24-fold increase in high-confidence sequences, compared to running on DIA-Umpire output. *De novo* sequencing is the task of inferring a peptide's sequence exclusively from the mass spectra, a much more difficult combinatorial task than the current standard approach of searching for *a priori* known spectra, and one that benefits from deconvolved signals [MR18]. Library-free approaches are advantageous to less studied organisms, especially given the complementary advances in high-throughput sequencing [HN20]. The possibility of quickly characterizing a new organism has far-reaching consequences in the broader life sciences and biotechnology. Towards this end, *de novo* protein sequencing appears as a very attractive tool, despite its current reduced competitiveness [MR18]. One of the aims of the CANDIA paper was to help improve the existing status quo by offering a simplified problem for *de novo* sequencing algorithms to solve.

While the total signal (ion count) of identified CANDIA spectra was clearly higher than that of library-matched spectra in our datasets, a large amount of the CANDIA output was not matched by the search engines. The unidentified deconvolved spectra were non-redundant with those that were identified and had a good overall decomposition metric (i.e. unimodality, described in the section below), showing that these represent useful data that could be leveraged by machine learning methods.

Besides the work on CANDIA, in **Paper I** we investigated some methodological issues in the current software ecosystem. Most current identification methods rely on a target-decoy strategy to control the number of false positives. In brief, besides the search with a database consisting of the target proteins, a control search is typically performed on a *decoy* database consisting of shuffled or mirrored versions of the target sequences [Käl+08]. We however expressed some doubt regarding how appropriate the shuffled or mirrored varieties of decoys are, as such unnatural sequences are highly unlikely to appear in samples (i.e these amino acid sequences are not sufficiently "peptide-like"). A striking result was obtained when we assessed this fact in standard tools by using shuffled versions of peptide sequences. Namely, we constructed a spectral library from a database with purely spurious (30% shuffled) peptides, alongside further shuffled versions of these as decoys. Two out of three tools (DIA-NN and Skyline) still yielded a high number (hundreds and thousands, respectively) of confident protein identifications (on average 185 times higher than the expected number of false discoveries at 1% FDR) when given this library with which to search the scans. This

shuffling strategy to generate decoys may therefore not do a good job of implementing an accurate model of the null hypothesis as a match against decoys, leading to all significance results relying on it to be inaccurate [Käl+08]. Besides the spurious results, this is further evidenced by the scores of these peptide sets, which showed distributions with heavy right tails (i.e. scores) for both targets and decoys (Fig. S2 in **Paper I**). One way then to interpret the high number of false positives on spurious input is that, since the decoys are not natural-peptide-like, they don't provide a strong enough "attraction" for the matching procedures. So there is no essential distinction between targets and decoys, both sets representing sequences that would not be observable in samples. If the converse were true (proper decoys), then the matching procedures ought to have ignored the spurious targets and focused on the peptide-like decoys. Using a CANDIA library helped reduce this effect, with no spurious such identifications made by Skyline and a reduced number of 22 times higher than what was expected at 1% FDR for DIA-NN (Fig. S1 in the paper), the hypothesis being that the simplified data provided by the pipeline in the form of deconvolved spectra helped against spurious (random) matches.

The work in **Paper I** thus demonstrates a data-driven approach to analyzing large and complex data, leveraging variability in the data itself, in combination with domain-specific assumptions and hypotheses. The approach also shows the significant benefits of the large computing power available today, allowing for a considerable amount of brute-force solution searching when no efficient procedure exists for selecting models *a priori*, as exemplified by the parallelized decomposition of many candidates models. Additionally, the implementation of CANDIA demonstrates the great benefit of open source software to innovation, as the code of the underlying libraries like TensorLy [Kos+19] were readily adapted to the needs of the pipeline. Finally, the PARAFAC method and core functionality of CANDIA are generic and may be employed for different (multilinear) factor separation tasks.

In the following technical section, I will briefly overview the methodology at the core of the CANDIA pipeline. Additional detail may be found in the *Supplemental Information* of **Paper I**.

2.1 Parallel factor analysis for proteomics

The task solved by CANDIA is that of chemometric measurement of individual analytes in a mixed ("convolved") signal. The method used is a type of factorization of the data, called parallel factor analysis (PARAFAC) or canonical decomposition (CANDECOMP) [KB09].

It can be considered an unsupervised method from a machine learning perspective, in the sense that no expected target values are used to train or derive a solution. This is in contrast to the related independent component analysis (ICA) approach taken by the Specter software [Pec+18], which factorizes each mass spectrometric scan separately using an *a priori* constructed spectral library. Rather, PARAFAC considers a whole scan series as a single input and decomposes it by exploiting the variation across samples. The formulation of this data model can be thought of in terms of the bottom-up construction of the observed signal, namely how has the overall intensity arisen from individual contributions of analyte values?

The shape of observed signal data may be described as a three-dimensional or "three-way" tensor, formed as follows. Each scan is a two-dimensional map with axes *mass/charge* (mass, for simplicity) and *elution time* (or *retention time*), the latter tracking the time at which the analyte particle "clouds" have appeared in the scanner, while the former records the fragmentation patterns of these particles. The time dimension (via e.g. liquid chromatography) is introduced as a way to achieve higher specificity, by physically separating peptides according to their physical properties, leading to less overlap on the mass dimension. These maps are stacked as separate observations, thus organizing the third, *sample* dimension of the data. Given perfect elution, the signal from a single analyte will appear at the same mass - elution time coordinates across all samples, but with different values, depending on actual sample concentrations and noise. It is precisely this sample variation that is exploited to regress out the contribution of each peptide within each sample [SBG04]. This data tensor or cube is thus the input to the PARAFAC decomposition.

The central assumption behind PARAFAC is that observed values in the tensor arise out of a trilinear combination of the three mass, time, and sample *factors* or *modes* (see Fig. 2.1). Additionally, a non-negativity constraint is imposed on all modes, a natural assumption as these capture particle counts. The output of the decomposition consists of three matrices or *modes* corresponding to each dimension of the input tensor, each with F columns, where F is the number of components for which the decomposition is made. More explicitly, one can write:

$$\underline{\mathbf{D}} = \sum_{r=1}^{F} s_r \otimes t_r \otimes m_r + \underline{\mathbf{E}} = [\![S, T, M]\!] + \underline{\mathbf{E}} \tag{2.1}$$

where $\underline{\mathbf{D}}$ is the input data tensor, $\otimes$ denotes outer product between the three column vectors of each component r, the matrices S, T, and M are the sample, time, and mass modes, respectively, and $\underline{\mathbf{E}}$ is the residual

(error) tensor. The second form of the inequality uses the Kruskal tensor product operator [KB09]. Graphically:

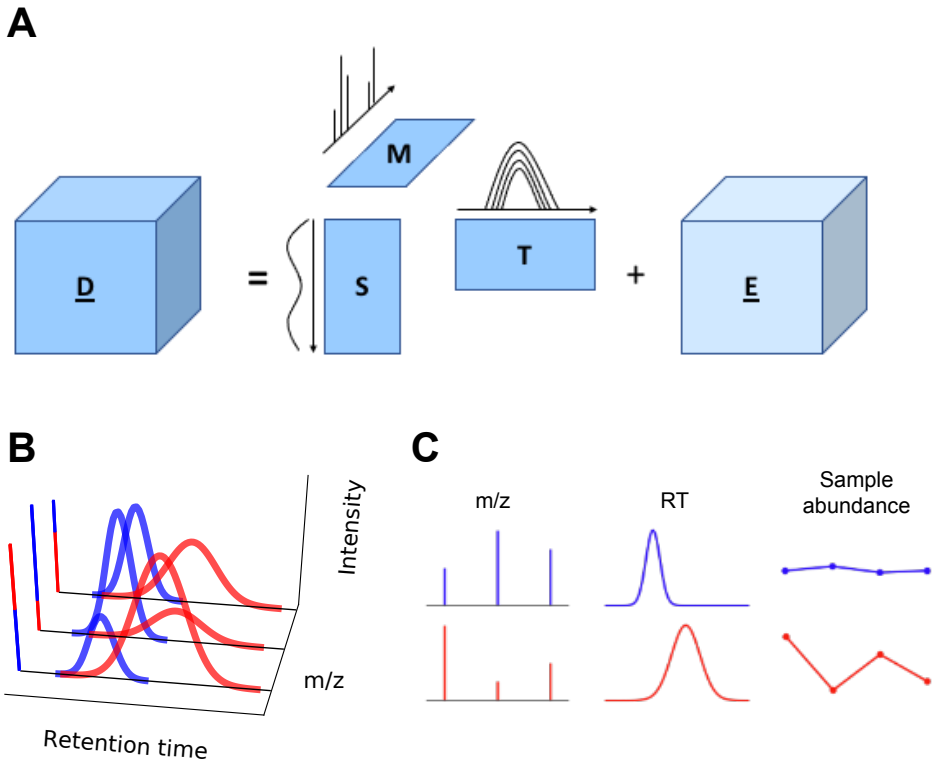


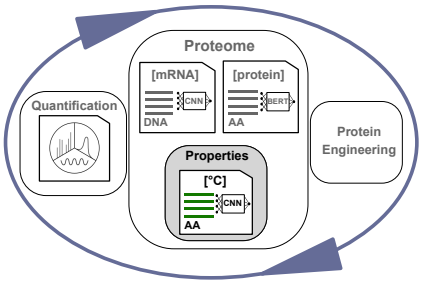
Figure 2.1: PARAFAC for proteomics (reproduced from [BZZ20]). **A)** The tensor of observed intensity values $\underline{\mathbf{D}}$ consists of stacked liquid chromatography-mass spectrometric scan maps. The model expresses these values as a trilinear combination of three matrices (*modes* or *factors*), namely the mass M, elution time T, and sample variation S, plus residual errors $\underline{\mathbf{E}}$. The width of each mode matrix is the number of components (analytes) for which to decompose. **B)** Illustration of convolved signal of two analytes (blue and red) in a single scan. Each analyte has a single elution curve and a specific mass profile, with considerable overlap along these two dimensions. **C)** Given several scans as the example in B), a PARAFAC decomposition for 2 components (each assumed to capture an analyte) yields mass (m/z), retention time (RT), and sample modes for each component.

The number of components F must be decided *a priori*, based on knowledge of the problem, or by examining the decomposition quality of several models over a range [SBG04]. The assumption here is that each separable peptide corresponds to such a component and is described by an individual mass fingerprint, elution curve, and sample variation trace. A complicating factor in the choice of value is that models do not "nest", that is $F + 1$ is not a model for F with an additional component [SBG04]. On the other hand, the solution for a given F is unique (modulo permutation and scaling of the component matrices). Another advantage is that the three decomposed modes have the natural interpretation of the dimensions they capture, and are not some type of abstract component.

While the oldest application of factor analysis to gas-chromatography-coupled mass spectrometry that I could find dates back to 1976 [Rit+76] and in spite of the linear algebra behind the decomposition being relatively straightforward and well characterized [SBG04], only with the recent advances in powerful graphical cards (GPUs) and distributed computing software could PARAFAC be made feasible for DIA proteomics datasets, as demonstrated in **Paper I**. Given the sheer size of the scans (a series easily reaching 500 GB) and the uncertainty around the correct number of analytes to decompose for (requiring solving multiple candidate models), one quickly reaches the limits of CPU-bound processing and memory (64-128 GB of RAM) for typical a workstation, as the tensor must be loaded into memory to be processed. Thus a partitioning scheme is necessary and this relies on another advantage of the SWATH-MS DIA scanning regimes. My insight here was that, as the mass scanning windows in the SWATH platforms are expected to be independent in term of precursor-fragment spectra [Lud+18], it would be fairly "safe" to partition the data tensor along them, besides partitioning along retention time windows. As for the time windows, their width was decided based on balancing the expected number of elutants (and consequentially, range of models) with the resulting number of such windows, taking the full length of chromatography into consideration. Each resulting partition or *slice* is an independent small tensor (on the order of MBs) spanning a narrow range of the mass and time dimensions (see Fig. 1A in **Paper I**), making it more more manageable, even for a typical workstation, while benefiting most from distributed processing across a computing cluster, CANDIA supporting both types of running modes seamlessly. More details on data management and preprocessing are further described in the *Supplemental Information* of **Paper I**.

The remaining issue of choosing the number of components F was based on a more straightforward iteration of all values within a reasonable range (i.e. how many peptides would one expect in a given time window), leaving model selection as a post-hoc problem. Among the criteria I considered for this selection, the most effective was that elution curves should be single peaks (or unimodal), given that peptides are expected to elute only once. Optimally deriving the number of components from a tensor is a hard (NP-complete) problem [Hås90] and while there are diagnostic routines one could run [Joh+14], the simpler approach based on the unimodality criterion proved very efficient computationally, as all decompositions could be run in parallel and the overall procedure performed well in terms of results quality. See *Algorithm 1* in the *Supplemental Information* of **Paper I** below for a detailed description.

3 | Learning protein features from sequence (Paper II)



It stands to reason that the amino acid sequence of a protein could be used to infer its 3D structure, and thus a good deal about its function. Indeed, we have seen just this year a significant leap in solving the folding problem [Jum+21]. Thus, as structural information is present in sequence, various physicochemical properties may be predicted from it and deep models have been used to capture a diverse array of properties, such as stability [CFC05], fitness landscapes [RKA13], and function prediction [Sur+19].

Moreover, these models learn various abstract features of the input sequence space that may be reused towards predicting other properties than what the original model was trained for. This is referred to as *transfer learning* and has had many successful classification, regression, and clustering applications, in areas such as natural language processing, image classification, genomics, medicine, and climate science [GBC16; WKW16; PY10], with many of the models quite generic. One of the major benefits of this approach is that it may bypass the limitation of data unavailability [PY10; GBC16]. To be more precise, one speaks about a source domain (variables with an associated marginal probability distribution over them) and a learning task on this domain (a function to be learned, that maps the domain to e.g. labels or output values). Transfer learning consists in using features learned as part of this source task to enable or improve a similar task on another target domain (different variables or distributions) [WKW16]. Concretely, in the case of neural networks, this typically consists of "transplanting" parts of one network into a target one, or partially freezing or reducing the change of certain network weights while training the source network onto the target task, a process often referred to as *fine tuning*. The transfer assumes that there are common (low-complexity) features shared between the two domains that capture the variance in both of these [GBC16]. Another assumption and often the motivation is that the source domain possesses significantly more data, such that it would be easier to generalize to a target domain containing relatively few data.

This technique presents the clear benefit of using proxy variables where

measurements are missing for a desired variable, as a model may in principle be trained on a different, derived or highly correlated proxy variable, for which there is abundant data. The correlation or dependence between this proxy variable and the desired one ought to be high however, as it was shown that transfer learning will not be optimal when the marginal probability distributions of the source and target domains are different [WKW16; Shi00]. Moreover, care must be taken as correlation is not generally a transitive relation, unless the correlation is arbitrarily close to 1 [14].

This transfer learning technique was used to good effect in **Paper II**, wherein the DeepET model was first trained to predict optimal cell growth temperature (OGT) on a large set of enzyme sequences (3 million), then subsequently retrained with only minimal tuning to predict optimal catalytic temperature (T_{opt}) on a much smaller set of enzyme sequences (N = 1902), motivated by the assumption that proteins should be functional at the organism's OGT. While similar in performance (R^2 = 57%) to a previous random forest model relying on amino acid compositions and OGT, the DeepET model has the major advantage of relying solely on sequence, as OGT data may not be available for a given organism. Moreover, the performance provides evidence that the deep model has learned repurposable sequence features, especially since the best performance was obtained by only fine tuning the last two (dense) layers of the network, keeping the convolutional layers and the residual block frozen (see Fig. 1 in **Paper II**).

A natural aim then was to probe these features in the context of T_{opt} prediction and relate them to known physical protein factors. Due to the black box nature of the deep convolutional-residual network, a perturbation approach was chosen. Namely, each protein sequence was covered (*occluded*) with a sliding window and, for each position of the window, fed to the network. The percentual deviation in prediction from the unoccluded sequence was considered as the *relevance* of the given position toward prediction [ZF14]. The resulting sequence relevance profiles were then smoothed with a moving average. Finally, only significant deviations (over ±2 standard deviations) were considered for matching against protein properties.

To check against properties that are known to affect thermal stability, I matched the significant relevance profiles with amino acid composition, secondary structure annotation, and protein domains. The first was a simple assessment of how the presence of certain residues influences the model, while the latter two were more specific to the sequence information, taking advantage of the position-dependent relevance profiles. This was done separately for enzymes from mesophilic organisms (with OGT 20-45

°C) and those from thermophilic organisms (OGT > 45 °C), in an attempt to contrast the most relevant matched elements between these two adaptation classes. See Fig. 3 in **Paper II**.

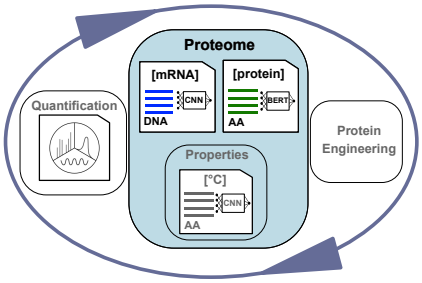
The amino acids that were enriched (overrepresented) at significantly relevant positions, when compared to the background amino acid count in the proteome, appear overall in line with known associations. In terms of composition, the majority of enriched amino acids were hydrophobic, which is a known (weak) determinant of thermostability, the hydrophobic content increasing with temperature [MMS16]. The set of such amino acids differed between thermal adaptation classes, with only Leu, Met, Phe common to both. Matching relevant positions with secondary structure annotation showed that for the prediction of T_{opt}, there is more distinction between the two thermal classes. For mesophilic enzymes, a larger number of structure types are relevant, while for thermophilic enzymes, only helices and turns are determinative. Helical content is known to increase with temperature, due to its stabilizing role, and the enrichment of Arg in the relevance profiles of thermophiles aligns with this, given that this amino acid is favored for helical formation [KTN00]. When measuring the coverage of InterPro protein domains [Blu+21] by relevance profiles across both thermal classes, only about 3% of searched domains were significantly overlapped (30% of their length), compared to control. This would seem to indicate a narrow range of functionality and to assess this, I derived GO slim terms from the GO annotations of the InterPro domains. While the biological processes associated with mesophilic enzymes is rather broad, thermophile terms were limited to metabolic processes and response to stress, the hypothesis being that these domains become more determinative for prediction of T_{opt} at higher thermal adaptation.

Thus, the DeepET network, trained on a very large set of enzymes to predict optimal growth temperature, learned to represent various protein properties as model features. These sequence features capture physicochemical properties, secondary structure, and a narrow set of protein domains. With only slight fine tuning, they were shown to be determinative in predicting optimal catalytic temperature. The first two types of properties were also seen to be captured by UniRep features, an unsupervised deep network model [All+19]. However, for the prediction tasks in **Paper II**, models trained on features exported by UniRep yielded the lowest performance, specifically $R^2 = 35\%$ for T_{opt}. Lastly, as the combination and strength of the determinants differ between enzyme families [MMS16], future inquiry ought to include a clustered perspective across the enzyme set.

Due to the black box nature of the model, there are downsides to this sort of perturbation analysis, however. Primarily, the biggest issue

is the arbitrariness of the perturbation procedure, namely the positioning and size of occlusions - whether these should be contiguous windows or rather scattered points, perhaps relying on external information, such as 3D contact maps. Moreover, occlusion assesses the impact of information *removal* from the sequence. While a subtle point, it begs the question if this measure of relevance indeed fully captures the effect of the region being present and *interacting* with the rest of the sequence. For contrast, in image processing, where the technique originates, often the subjects of identification are (essentially) independent groups of pixels (e.g. a dog on a green field). Thus excision of an object in a scene is adequately informative with respect to its detection by the network, whereas the situation is less clear with protein sequences. Given these aspects, results stemming from window occlusions may tend towards capturing local subsequence relevance and miss long-range interactions. From this interpretability perspective, the BERT architecture described in the next chapter serves as an attractive alternative, as it exposes its learned weights in a more transparent fashion, intrinsically highlighting pairwise amino acid association strengths.

4 | Learning expression levels from sequence (Papers III and IV)



In the previous chapter, I explored a case of using amino acid sequence to capture physical properties of proteins, as well as using this data-driven approach to infer the determinative factors of these properties. Here I will cover a broader goal, that of determining the composition of the proteome from sequence data. The work can be seen as part of a larger undertaking to quantify the central dogma and get comprehensive phenotypic predictions from genomic and proteomic sequence. Additionally, "grammars" of expression were sought, i.e. rules of nucleotide and amino acid associations determining expression levels, as expanded on in the first section below.

As briefly overviewed in the introduction chapter, proteome composition arises out of the balance of different processes and cellular needs. Much work so far regarding protein abundance has been done either from genomic or transcriptomic information [Zri+21], though in most cases sequence information was summarized as e.g. amino acid frequencies or codon usage bias and the majority of models considered were very simplistic, often linear, and with rather poor explanatory power [VM12; CR18; Rib+19]. These simplifications, while providing valuable indication of the relations between the various factors, lead to many of the more complicated interactions between said factors to not be captured, whereas we know that between the "levels" of central dogma there is much regulation and the molecular abundances span different dynamic ranges (e.g. as is the case of mRNA and protein abundance) [VM12; BS20].

The results in **Paper III** are derived from *S. cerevisiae* median protein abundances (N = 5202, median over 21 experiments), collected predominantly from mid-exponential growth phase. They show that much information about abundance is encoded in a protein's amino acid sequence alone (the best model had an $R^2 = 41.6\%$), which is perhaps not surprising given evolution's imprint and the fact that function is given by structure, thus indirectly the sequence. There is also the fact that protein levels have relatively low variance across different conditions (i.e. within 1 or 2 orders of magnitude), relative to the wide range of levels observed across all proteins (i.e. 5 orders of magnitude) [HBB18] (see Fig. 1A in **Paper III**), hinting a rather tight coupling between protein structure and proteome composition.

Of note that on the data from [HBB18] (which were used in this paper), there is no correlation between amino acid sequence length and protein abundance (Pearson $r = -0.1$, p-value = 1.26e-10) in spite of its reported determination of mRNA levels [VM12; Rib+19], even when restricted to the set of median abundance with at most 1 standard deviation across experiments).

The deep neural network model chosen to learn this sequence-to-abundance relationship was BERT, a Transformer-type model using the attention mechanism described in the introduction chapter [Dev+19]. The choice of network architecture was encouraged by the expectation that attention values would provide an intrinsic way to inspect amino acid association rules, and by previous successes in capturing contact maps, binding sites, and substitution likelihoods through the attention mechanism (either directly or through a simple probe) [Vig+20]. The TAPE implementation of this model [Rao+19], specialized in protein sequence, served as starting point for my code base.

And indeed, attention profiles were seen to correlate well with physicochemical properties of the polypeptide chain, as well as to preferentially cover protein domains and some homorepeats (stretches of repeated occurrence of the same amino acid). The GO terms associated with these were rather diverse, including translation, protein folding, post-translational modification, carbohydrate and ion transport, stress response, organelle fission, cell cycle, sporulation, and cell division.

The study performed previously in **Paper IV** was one "level" upstream in the central dogma, assessing how much information about transcript abundance is encoded in the entire gene. One main finding of this study was that regulatory regions (as sequence input) jointly explained 49% of transcript level, while augmenting by including codon frequencies and stability values (as numerical variables) in the input, 82% of mRNA levels could be explained. The breakdown of these different predictions is explored more in the second section of this chapter. Both learning tasks used convolutional neural network models, which conferred a performance increase compared to simpler shallow models.

A second result of this investigation was that a gene's mRNA level arises out of the interplay of its coding region and the full cis-regulatory structure (**Paper IV**). Moreover, codon frequencies could be 58% explained (predicted) from regulatory regions (Fig. 2c, d in **Paper IV**), showing overlap in information and hinting that the ensemble of regions is a co-evolving unit, backed by evidence that in eukaryotes non-coding and coding regions are under weakly coupled selective pressure in orthologs [CHA04; Che+10], as well as mutation rate correlation computed between regulatory and coding regions (Pearson $r = 0.42$ and 0.47 for promoters

and terminators, respectively) in multiple orthologous yeast genes (Fig. 2e, f in **Paper IV**). In line with protein abundance, mRNA variability in yeast across conditions is smaller than between genes (within 1 relative standard deviation for 85% of genes), another hint that much of the information on transcript homeostasis has been imprinted in DNA.

In order to inspect how different regions and relative positions influence the prediction, a perturbation study was performed using the model. Region boundaries were found to be most impactful for the prediction (Fig. 3a in **Paper IV**) and the relevance profiles of promoters were found to anticorrelate (Pearson $r = -0.7$) with nucleosome occupancy scores (the frequency of histone octamer occupation of a given DNA region across a cell population [SS13]). These relevance profiles were clustered and two of the resulting clusters matched low- and high-expression genes, enriched in cell cycle regulation and DNA repair, and metabolic processes, respectively.

This approach was thus based on extrinsic probing, measuring the impact (i.e. relevance) of systematic occlusion of sequences. This contrasts with the strategy taken in the later **Paper III** study, using a BERT model to predicting protein abundance, as that relied on an intrinsic measure of position relevance, namely the attention mechanism. These aspects are expanded upon in the first section below.

It is known that mRNA is a major determinant of protein abundance [LBA16; Lah+17], however the coupling between these two quantities is not tight in yeast. Using the datasets in **Papers III and IV**, Pearson $r = 0.74$ on the low-variability subset of genes (N = 3399) and $r = 0.69$ on all (N = 4859) genes. Besides the missing explanatory fraction, correlation is a linear assessment between the variables, implying that the missing fraction might "hide" a more complex relation between the two quantities (see Fig. 4.1). Indeed, as outlined in the introduction, we know the post-transcriptional translation, regulation, and degradation processes further determine protein levels, beyond the availability of mRNA amount [VM12; LBA16; BS20; Ho+21], even if to a lesser extent during steady state [LA16; LBA16].

The second section below bridges these two studies, towards an improved prediction of protein abundance.

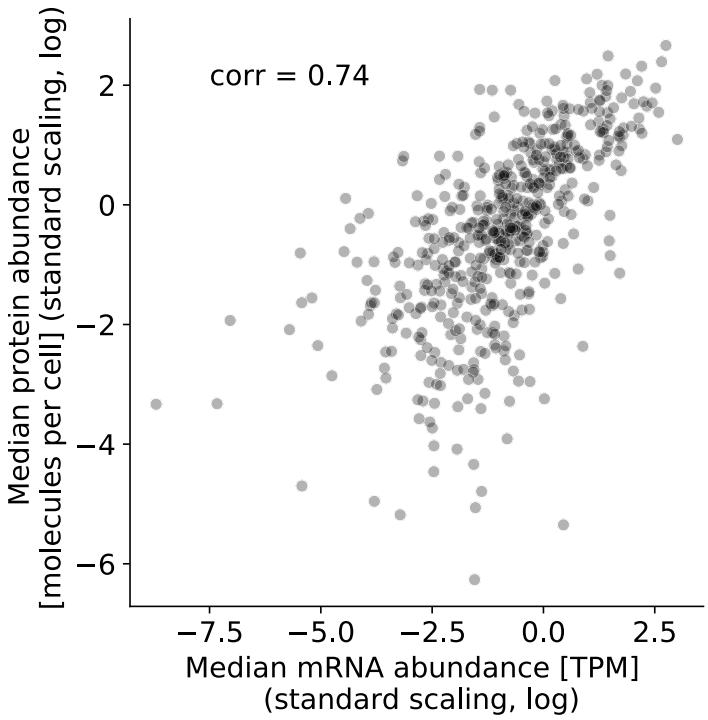


Figure 4.1: Correlation between mRNA and protein abundance, using data from Papers III and IV. Plotted are genes (N = 3399) with coefficient of variation at most 100% across experiments. Values were standard-scaled and log-transformed. The Pearson correlation on all (N = 4859) genes was $r = 0.69$ (p-value below double floating precision).

As was mentioned in the introduction, it is a big step to learn expression from sequence alone, accounting for different conditions, intercellular interactions, and abstracting over complex pathways. But as evidenced by these two studies, integrative or holistic sequence-driven modeling can offer much insight into the information evolution has imprinted in a genome. While one could in principle train a protein model on DNA sequence directly, in spite of the complexity involved, the modeling in **Paper III** focuses the analysis on the primary structure of proteins and post-translational interactions. It also provides a simpler input than DNA sequences, allowing for smaller models, and also simplifies the already complicated task of interpreting deep models. Moreover, different models could be integrated in a number of ways. In the second section of this chapter I describe a simple meta-modeling approach that pools predictions from the models in both papers for an improved protein abundance prediction.

4.1 The grammars of expression

In terms of sequence informing expression product, **Paper IV** is roughly analogous to **Paper III**, albeit methodologically different and with different focus, the two being placed on different "levels" of the central dogma. As was centrally highlighted in **Paper IV**, it is the interactions between the different regions that are most informative for the product levels (in this case, transcript). This parallels the importance of amino acid interactions towards predicting abundance in **Paper III**. Both studies aimed at recovering a set of rules that may explain how association between the respective elements inform their product quantities. These sequence "grammars" are assumed to be learned by the deep neural networks and are thus implicit to the predictor. By using different probing techniques, associations of elements (regions, motifs, and amino acids) were uncovered and used for protein engineering.

The input data for **Paper IV** is heterogenous (different regions with different functions), raising the intriguing question of how the interplay of regulatory and coding regions gives rise to mRNA levels. The probing technique used in the study is perturbation by occlusion of windows along the sequence length, in order to test which positions are the most relevant for the prediction (see also Fig. S14 in the article), the approach also taken in **Paper II**. This yields a prediction relevance profile for all genes, which may be further mined for connections to known important sequence patterns.

As first quantitative evidence of the importance of position (and hence, sequence structure), the relevance profiles featured the largest perturbation in promoters and terminators, irrespective of nucleotide composition across all regions, and that region boundaries were most impactful. To obtain structural sequence patterns, 2200 regulatory motifs (i.e. contiguous short sequences) were extracted from the relevance profiles of all four regulatory regions, by picking relevant DNA sequences (with absolute values above 2 standard deviations), then clustering and aligning them. Validation was performed against known motifs in databases. For more details, see Methods in **Paper IV**. The majority of motifs are specific to each region and it is their co-occurrence that is mostly predictive of transcript levels. The co-occurrence of motifs was assessed with *association rule learning*, a type of unsupervised task that identifies significant combinations of items (motifs) from a large collection of observations [HTF09]. Association rules are expressed in terms of implications such as {motif 1, motif 2} $\Rightarrow$ {motif 3}, meaning that, across the data, should motif 1 and 2 appear in the same observation, this implies (with a certain confidence) the presence of motif 3 as well. Almost 10000 significant

rules were identified. These were often comprised of a few motifs (2 to 5 typically) clustered within small sets ($\leq$ 10) of genes, and were discriminative between low and high transcript values. Rules frequently (88%) spanned all regulatory regions and covered a wider range of transcript values than single motifs, as well as showing lower variance, hinting at the important role of the rule presence in these genes. By swapping an increasing number of motifs within a given gene, the predicted expression level could be changed considerably, up to 3 orders of magnitude when swapping 3 out 4 motifs. The most common motifs across all genes were not enriched in any specific cellular function, comprising thus a common vocabulary across the genome.

The motif grammar was validated *in vivo* (yeast) in a simplified scenario, using a model trained on the most relevant segments of the various regions (due to experiment limitations on region length). The experiment used the native promoters of 6 constitutive genes for the green fluorescent reporter protein (GFP), along with all combinations of native, weak, and strong terminators (18 combinations in total). The changes in expression level correlated quite well with prediction (Pearson $r = 0.65$), showing the potential of the motif grammar in screening promising promoters and terminators from the very large space (millions) of possible combinations.

In terms of model interpretation, aspects of the transcription regulation grammar learned by the deep network may therefore be represented as combinatorial associations of motifs. Due to limited transparency of convolutional neural network model class used in this study, however, probing it to find the main determinants and their interactions toward prediction requires the extrinsic perturbation approach. This comes with some arbitrariness and limitations, namely occlusion is performed on contiguous regions, capturing local interactions. While pertinent for this application, such an approach was perceived to be limiting for modeling sequences of amino acids, which are expected to show long-range interactions (i.e. via folding). Moreover, the perturbation analysis is further complicated by the chaining of downstream analytic tasks, such as clustering, alignment, and rule mining. Lastly, the motif rules and their matching to expression levels were mined post hoc from the set of perturbation profiles. Thus, while the model indeed has good performance, has given insight into important regions, and yielded predictive motif co-occurrence rules, one would ideally like a more intrinsic weighting of the associations of various sequence sections, coming directly from the model itself. The attention mechanism used in the BERT model in **Paper III** is such an alternative.

While the "grammar" analogy to language was used rather intuitively in **Paper IV**, my approach in **Paper III** was to explicitly rely on frameworks

from natural and formal language processing, distinguishing between syntactical (structural) and semantical (quantitative) characterization of amino acid interactions learned by the BERT model to produce protein abundance predictions, partially motivated by the use of this type of deep model to study language in this way [Vig+20]. As a simple illustration of how these two notions frame the grammatical understanding, consider the two phrases "I eat the cake" and "The cake eats me". From a syntactical point of view, both of these are structurally correct English sentences and the rule of their formation obeys the pattern subject-verb-object. However, the meaning or semantics of the two sentences are quite different. Indeed, the same distinction is made in formal language processing, when a computer program is first analyzed syntactically, then is ultimately converted to purely numeric values. As an analogy closer to the protein application, my intent was that given a sentence such as "one plus two", a hypothetical BERT model would learn the functional role of each word (i.e. number and operator words), the correct value to assign to each number word, as well as the correct operation to perform on these values (given by the operator word), finally outputting the prediction "3". The more ambitious goal was to also extract such operational understanding.

The syntax-semantics analytical paradigm was motivated by the fact that attention-based models appear amenable to this type of analysis, the common understanding being that attention layers learn syntax, especially in lower layers, while the topmost (deepest) layers, as well as the embedded space of the network, capture more semantical aspects, though the exact cutoff is debated and may very well vary depending on task [RKR20].

In terms of syntax, to characterize the attention-based sequence association rules, I represented attention matrices for a given sequence as *dependency trees*, where each residue is a node and a connection is given by an attention value (for more details on the construction, see **Paper III** Methods). In natural language processing (NLP), these trees are a way to describe the structure of a sentence, by connecting words that depend on each other. The strength of dependency between words is determined in different ways, often by computing word pair association weights over a large corpus. The relation is often not symmetric, and one word is considered to depend on another (Fig. 4.2A) [JM09].

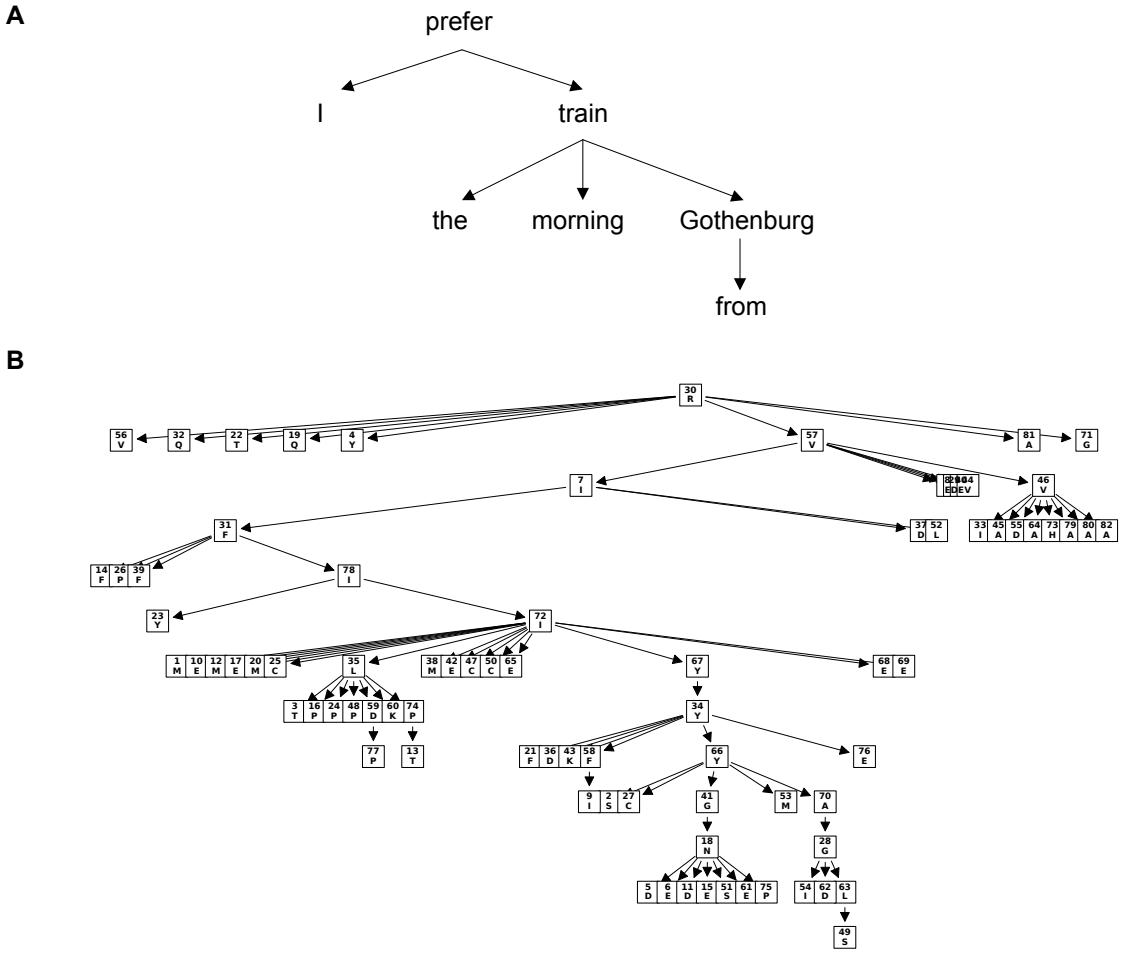


Figure 4.2: Dependency tree examples. These trees are constructed from the pairwise association weights between the tokens of a sequence (words or symbols), typically by deriving the maximum spanning arborescence graph from the aforementioned weight matrix. **A)** Dependency tree for the sentence "I prefer the morning train from Gothenburg". **B)** Amino acid sequence dependency tree for the protein Diphthamide biosynthesis protein 3 (UniProt ID: Q3E840), constructed from the attention matrix of a BERT head when given this sequence as input. The nodes are marked with the amino acid and position in the sequence.

Accordingly, the attention matrix produce by every BERT head for a sequence expresses how much a given residue in the sequence *attends* every other, normalized as a percentage across all pairs. As was observed in **Paper III**, for protein sequences this relation is asymmetric. Similar types of attention-derived trees have been constructed for Transformer-type models in NLP, with overall quite good success in matching curated trees [RT18; Htu+19]. The directional relation captured by attention between pairs of residues thus serves as the basis for constructing sequence dependency trees that show a hierarchical, position-dependent description of the importance of each residue towards predicting abundance (Fig. 4.2B). Another motivation for such a hierarchical representation is that residue interactions are expected to be complex and span the entire sequence (when thinking of e.g. protein folding).

While it is still unclear how this or similar representations can be connected to the semantics posited to be captured in the network's embedded space, the dependency tree does highlight some interesting properties. Within each tree, there are only a few "hub" residues on which many others residues spread across the entire sequence depend. Additionally, the majority of hubs are "occupied" by just three amino acids, Ser, Lys, Ala, in order of increasing depth (Fig. 3D in **Paper III**). Given the hierarchical organization, the depth in the tree could be interpreted as importance ranking, with most important hubs at the top (however "importance" may be quantified towards prediction). Interestingly, amino acids do not seem to be preferred by attention at any given position along the length of sequences. Moreover, there is little variation in the above patterns across protein abundance bins. These last two observations hint at the BERT attention-based prediction arising out of complex interactions of residues across the entire sequence, rather than special prevalent regions or amino acid motifs.

As mentioned above, overlapping semantical information onto these trees was less fruitful, although it is possible to implicitly harness the embedded space of the model to effect a significant change in prediction, as I expand upon in the next section. How these two grammatical dimensions may be seamlessly integrated such that one may observe how semantical operations "endow" syntax with value, is an avenue for further research. This problem, recognized across other areas such as image recognition, is referred to as the "semantic gap" [DK21] and could be summarized as the difficulty in assigning human meaning to features learned by a deep model, part of the overall challenge of interpretability.

4.2 A surrogate model spanning the central dogma

To assess the composition of the proteome, ideally one would like to construct an overarching transparent model predicting protein abundance directly from DNA sequence, to the full extent at which this information is encoded in sequence. However, such models may not yet be feasible due to computational limitations and perhaps model complexity (at least, for the current classes of existing deep models). Moreover, transparency or even interpretation might be difficult to achieve through such models. In spite of this, given the strong connections between the subprocesses involved, one may use the partial modeling performed on these subprocesses. The models in **Papers III** and **IV** helped construct a piecewise quantitative characterization of the central dogma. The variables in question correlate

and jointly contribute to the overall "information flow".

All of this raises the possibility of somehow building on these piecewise predictions. For the purpose of predicting protein abundance, one simple way to do this is to construct a *surrogate* (meta-)model [DCA19] using the mRNA and protein estimators obtained with the deep models (i.e. the predicted values alone as surrogates for the models themselves). To this end, estimator variables (i.e. predicted values for all genes) were produced using the various models. In order to guard against overfitting, the distinction between the training and test sets used in **Paper IV** was kept for all models, including the CNN model in **Paper III**. This model was retrained on a repartitioning of the protein abundance data, respecting the training-test split of the mRNA data. A new hyperparameter search was performed (see the thesis *Appendix* for the best hyperparameter values) and the best CNN model performance was 41.6%, the same as the original model described in the paper. Thus, considering the intersection in genes present in the mRNA and protein abundance datasets, the same partitioning of 2708 training and 303 test genes was used across all models here.

The models considered were:

- M_{reg} - the CNN model trained on gene regulatory regions from **Paper IV**
- M_{codons} - a random forest model I trained here on codon frequencies
- M_{gene} - the full CNN model from **Paper IV**, trained on regulatory regions, codon frequencies, and mRNA stability variables
- M_{aa} - the CNN re-trained here on amino acid sequences to predict protein abundance.

The predictions from these models are considered as estimators of the mRNA ($\hat{R}_{\square}$) and protein ($\hat{P}_{aa}$) abundance. As was shown in **Paper IV**, there is a significant amount of interdependence between the different variables the models capture, and it is also known that there is a high correlation between mRNA and protein abundance [Lah+17], which I illustrated in the first part of this chapter (see Fig. 4.1). Because the interdependencies of the variables (and, in fact, processes) in question are however not tight (the variables are not perfectly correlated), the expectation is that combinations of these different piecewise estimators would yield a better prediction of protein abundance (Fig. 4.3).

Indeed, even a simple linear model fitted to $\hat{R}_{regions}$ and $\hat{P}_{aa}$ to predict protein abundance gave an R^2 of 50.28% (on the hold-out test set), a

clear improvement on the performance of the CNN model in **Paper III**. To get a full overview, I considered here all combinations of at least two estimators as input for linear and random forest models, as shown in Fig. 4.3. The random forest class of model was chosen as it is nonlinear, computationally inexpensive, yet readily exposes the relative importance of its input variables.

Random forests consist of an ensemble of many de-correlated regression trees, all trained on subsamples (with replacement) of the same input data. The output of all trees in the ensemble is averaged to give the final prediction. Through the splits at each level, these trees partition the domain of input variables into smaller regions, for which simple constraints hold. These constraints are then hierarchically organized such that they model a decision process for each datum. Using the average across multiple tress serves to the reduce variance of the final estimator (generally leaving bias of the forest the same as any of the trees) and increases training stability (i.e. the dependence of the model structure on the training set), both of which are characteristic problems of single decision trees. This also helps with overfitting, which can be further restrained by limiting the depth of the trees in the ensemble [HTF09].

The random forest models considered shared the same hyperparameters across all estimator combinations (see the *Appendix*). Both mRNA and protein quantities have been Box-Cox-transformed as input to the original deep models ($\lambda = 0.22$ and -0.05155, respectively, same as in the two papers), and the predictions of these models (the estimators) were normalized with standard score.

Of all the combinations that do not include $\hat{R}_{gene}$, the highest performance ($R^2 = 63.51\%$) was obtained by the random forest trained on all three estimators derived from sequence (regulatory regions, codon frequencies, and amino acid sequence), once again illustrating how these sources of information jointly predict the quantity of the end translation product (Fig. 4.3B). There is redundancy between the mRNA models, as pointed out in **Paper IV**, and $\hat{R}_{gene}$ is additionally trained on mRNA stability variables. Moreover, the codon frequency and amino acid sequence data are clearly overlapping in terms of information as well. The redundancies in these four estimators is evident in the performance of the combinations listed here, with very close performance obtained from the various combinations of three predictors, and the highest variance explained obtained from using all four.

Considering only the sequence-based random forest model (Fig. 4.3C), each estimator contributes roughly a third in terms of importance to the model, in decreasing order of codons, amino acid sequence, and regions. It is known [HTF09] that the random forest variable importance

measure tends to be rather uniform, with the ranking as the most salient distinction. It should be stressed however that these percentages reflect their importance to the random forest *model*, based on the respective partial deep models used to obtain these estimators (the performance of said models incidentally decreasing in the same order), and should not be interpreted as a direct reflection of the ranking of the associated molecular quantities and physical processes in their roles toward the resulting protein abundance.

This simpler surrogate model outlined here, based on regulatory regions, codon frequencies, and amino acid sequence shows that piecewise modeling along the processes in the central dogma can yield good predictions from sequence alone, leveraging the good performance of each partial model. Moreover, such an approach alleviates computational cost. In the example presented here, we are dealing with few estimators and partial models on a relatively simple (essentially linear) network model. Regardless, as argued above, bridging the work in **Papers III** and **IV** by training a full DNA-to-protein-abundance model (especially of the BERT variety) is severely limited by the size of the input data (full genes and upstream and downstream regions) and the model complexity required to learn these data. However, even when such a technical challenge will be overcome for this particular case, the approach presented in this section could conceivably be taken with much larger process networks and perhaps expanded to consider graphical models or similar formalisms that account for the topology of the network (and associated inferences), as well as the uncertainty around each partial prediction. Conversely, training deep neural network models may not be feasible for large models of this sort due to both the size of the data as well as the way the model complexity (and, implicitly, training time) might scale with the data.

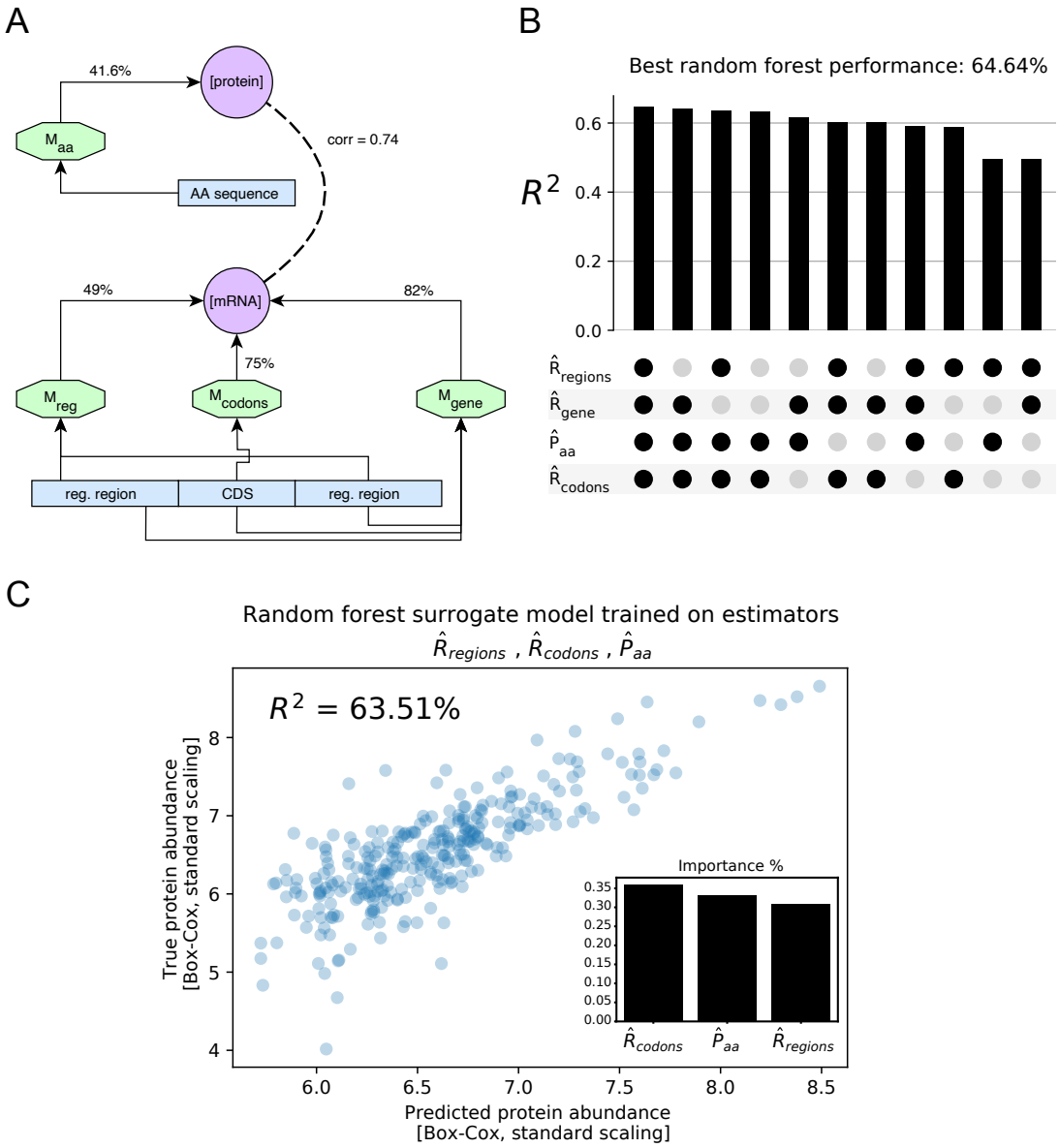
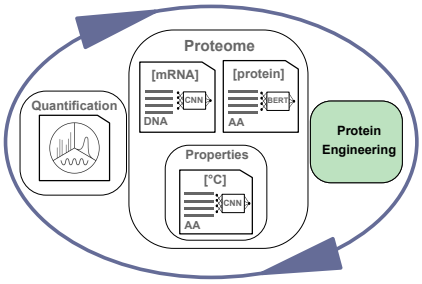


Figure 4.3: A surrogate model from piecewise estimators improves the prediction of protein abundance. A) Conceptual diagram with explanatory variables and models around the central dogma. Purple circles depict molecular quantities and light blue boxes depict DNA and amino acid sequences. Models trained to predict these molecular quantities from sequence are shown as green octagons, with arrows connecting the sequences as sources of information to the predicted quantities. The variance explained (R^2) by these models are shown as percentages on the arrows. The CNN models trained on gene regulatory regions (M_{reg}) and the full gene, including mRNA stability values (M_{gene}), have been characterized in **Paper IV**, while the model predicting protein abundance from amino acid sequence (M_{aa}) is a CNN model retrained on the same train-test dataset split as the models in **Paper IV**. M_{codons} is a random forest model trained here from the codon frequency dataset used in **Paper IV**. **B)** Performance (as variance explained) of random forest models constructed from all combinations of at least two estimators (and evaluated on the same test set). All models share the same hyperparameters. **C)** The performance as variance explained of a random forest surrogate model trained on the mRNA and protein abundance estimators derived from sequence information only (i.e. regulatory regions, codon frequencies, and amino acid sequence). In terms of model importance, the three estimators contribute roughly a third each.

5 | Using deep representations to guide protein engineering (Paper III)



As previously stated in the introductory chapter, once a certain level of information has been captured by a model or framework, a way is opened towards engineering desired products. Such is the case with the BERT sequence-to-abundance model that was investigated in **Paper III**, as its structure is conductive to exploring the effect of variation in sequence space towards the final quantitative prediction. In this chapter, I will describe in more detail the guided mutation framework developed for protein engineering applications, thus closing the cycle back to experiment.

Classically, to obtain a protein with desired properties or influence the composition of the proteome in a certain way, researchers rely on directed evolution to yield a mutant with the desiderata, thus effectively conducting a local search in the space of protein sequences [YWA19]. Besides the experimental costs involved, a random search through sequence space however is a daunting undertaking, due to the astronomical size arising from combinatorial explosion. Mutating just 10 residues results in a search space of $19^{10} \approx 6$ trillion combinations. Moreover, it is assumed that only a relatively small number of "islands" in this vast space correspond to functional proteins [YWA19]. Current methods thus perform a sparse sampling of this space with various mutagenesis methods, then try to move toward optimal regions (i.e. combinations of amino acids maximizing the desired effect) by a screening step [PL15; YWA19].

Machine learning can be beneficial for directed evolution pipelines by providing candidate mutants based on previous rounds of screening, thus acting like a "shortcut" and alleviating the costs associated with screening. The data-driven approach is a way to sidestep the computationally hard problem of sequence optimization and the reliance on detailed understanding of the physics or biochemical networks involved [YWA19]. There is also the question if some of the aforementioned islands are indeed accessible through evolution [PL15] or may require an artificially designed "jump" over regions corresponding to inviable proteins.

For the study in **Paper III**, I developed a proof-of-concept strategy to mutate the amino acid sequence of a protein (by substitution only) in order

to change the predicted output of the deep model. The method allows one to set how many residues to mutate and the algorithm deterministically chooses the substitute residues in order to increase the predicted abundance as much as possible, by taking advantage of the structure of the embedded space that the BERT network learns. Incidentally, while increase was the aim of the study, the method can just as well be used to decrease abundance, should this be desirable. The approach does not require an iterative experiment-machine-learning pipeline [YWA19] and produces a mutant sequence for each starting wild type protein.

Similar work includes the support vector machine model developed in [vdBer+14]. Here, the model was trained on many example amino acid sequences from *Aspergillus niger* to discriminate between low and high production proteins. The predicted classification was used as a criterion in an iterative sequence mutation procedure relying on a genetic algorithm. The common aspect with my method is the reliance on the representation space used by the model to "sort" sequences from low to high abundance. However, the BERT network I used is a continuous map from sequence to abundance. Moreover, as I outline below, no iterative sequence optimization is required. A sequence deemed to be optimized is returned directly. Other current approaches to learning abstract representations of protein sequences focus on unsupervised models (i.e. without target values), which aim to cluster a large number of sequences across many organisms, in order to improve performance of downstream machine learning tasks or at least offer a lower dimensional feature space to model [Yan+18; All+19]. Additionally, these clusters may capture some physicochemical properties or secondary structure features [All+19]. In contrast, the BERT network has been specifically targeted towards predicting abundance in *S. cerevisiae* and it is expected its internal representation reflects this specifically. Connections between the two approaches are an intriguing future direction.

In the second part of this chapter, I will intuitively describe the guided mutation framework in **Paper III** and the motivations behind it.

The core assumption is that the embedded space into which the BERT encoder maps sequences is structured in such a way that point clouds rather "closely" follow the order of abundance values predicted from them. The assumption is motivated by the thin predictor stack further mapping into protein abundance (consisting of only two dense layers, see Fig. 1B in **Paper III**), which would imply that the way the sequences are represented in the embedded space ought to reflect their distribution in the target abundance space. This would be an example of a "manifold hypothesis" [GBC16], stating that representations of input data lie on a lower-dimensional manifold. See Fig. 4A in **Paper III** for

a conceptual illustration. Additionally, given that the target space ($\mathbb{R}$) is a totally-ordered set, my intuition was that this strict structure would further constrain the topology of the embedded space, such that the point clouds corresponding to each sequence would be ordered if not totally, then partially but "close" to a total order. The point clouds are thus expected to lie on a geodesic in E that reflects the total ordering of the real values each cloud is mapped to.

Slightly more formally, let $E \subset \mathcal{P}\left(\mathbb{R}^{1024}\right)$ be the embedded vector space into which the BERT encoder e maps sequences. E consists of sets (or *point clouds*) of 1024-dimensional vectors (or *points*), each vector representing a residue in the amino acid sequence. So $E = \{e(s) \mid \forall \text{ sequence } s \in \text{ sequence space } S\}$. Note that BERT uses a positional encoding of the protein sequence, meaning an amino acid ("letter") will generally have different embedded values (1024-dimensional points) depending on where in the sequence it occurs. Each point cloud c is then mapped by the predictor layer p into $\mathbb{R}$. Let us furthermore assume that points clouds are "clusters", i.e. points in a cloud are generally closer to each other than to points in another cloud. This appears to be the case when tested with the UMAP projection described below. See Fig. 5.1 for an illustration of these structures.

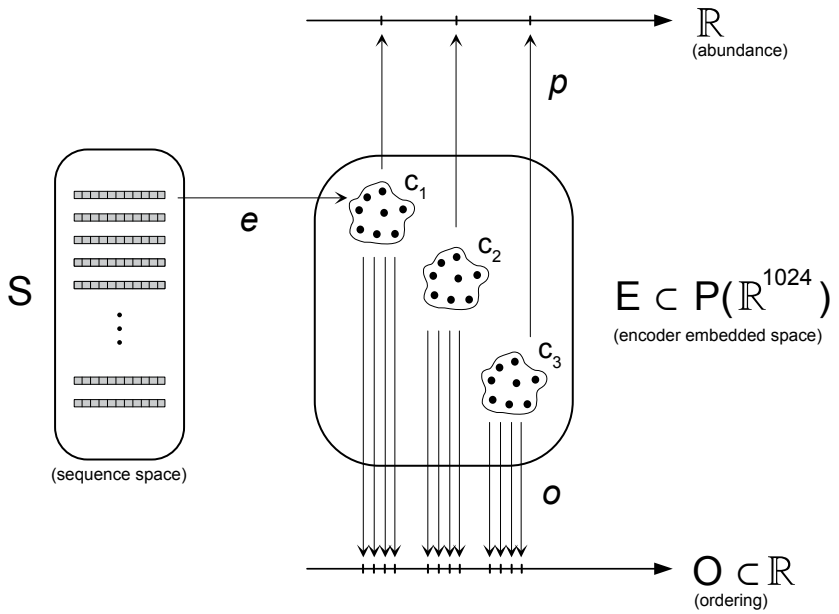


Figure 5.1: Idealized description of the BERT model and the embedded ordering construction used for mutation. The BERT network consists of two composed maps e and p (encoding and prediction, respectively). The first takes a sequence and assigns it a set of points (or point cloud) in the 1024-dimensional embedded space E learned by the model (each residue in the sequence being assigned a point). The map p takes the point cloud of an e-mapped sequence (e.g. c_1) and outputs a real number ($p(c_1)$), the predicted protein abundance. The UMAP projection o takes the point cloud of a sequence and assigns each 1024-dimensional point in the cloud to a real number. The codomain of o is referred to as O to distinguish it form the codomain of p.

Now, to describe the structure of the spaces in question, we have the target abundance space $\mathbb{R}$ with the total ordering $\leq$. Then we posit the order $\prec$ between elements of E satisfying

$$c_1 \prec c_2 \Rightarrow p(c_1) \leq p(c_2) \tag{5.1}$$

where c_1 and c_2 are two point clouds in E. This is perhaps an optimistic ansatz, but is expected to cover most sequences given as input to the encoder and serves to fix the intuition of structure preservation through p. Note that p is not injective, as there are proteins with the same median abundance in the dataset.

It is however unclear how the $\prec$ order arises from the contribution of each point in a cloud. Still, given the previous assumption that point clouds are clusters in this space, conceivably, by virtue of the left-right-to implication, when shifting subsets of points in a cloud c toward regions that are known to be higher-valued in terms of p, the shifted cloud c' is expected to obey $p(c) \leq p(c')$.

But how do we shift a point in $\mathbb{R}^{1024}$ towards a region with high predicted abundance? The geodesic on which this movement would occur is

the one that respects $\prec$ but we do not know how this is expressed in terms of individual points and the very high dimensionality confounds approaches based on distances within E. The approach I took was to perform a non-linear dimensionality reduction using Parametric UMAP [SMG20] from $\mathbb{R}^{1024}$ down to $\mathbb{R}$, as UMAP was designed aiming to preserve local topology [MHM18], in our case the point clouds. The intent was to have a proxy space ($\mathbb{R}$) with understandable movement of points corresponding to residues. To simplify discussion, let the UMAP projection be labeled as $o : \mathbb{R}^{1024} \to O \subset \mathbb{R}$, and, forcing notation, let this designate both the projection of a point or a point cloud, i.e. $o(c) = \{o(r) \mid r \in c\}$, where r is the point corresponding to a residue in the sequence. This construction I call "embedded ordering" in the paper. The "parametric" variety of UMAP was chosen because it returns the map o itself, not just the set of projected points, which allows us to project any new arbitrary points from E. The neural network underlying this type of UMAP was trained on the points corresponding to start tokens of each sequence, as BERT forward "routes" information from the entire sequence through these nodes (i.e. they appeared as good representative points for each cloud).

Because the projection is done down to 1D, this induces a total order $\leq$ between the projected points. Note that for our purpose, the relative distance between point clouds (now on the line O) is less relevant than their order. The only expected problem is that some point clouds will be in swapped order on O compared to $\prec$ in E, given UMAP does not guarantee preserving global topology. To assess this, I computed the centroids in $\mathbb{R}^{1024}$ of the point clouds of all sequences, projected them with o, then rank-correlated them with their BERT abundance predictions, obtaining Spearman $\rho = 0.85$ (Fig. 5.2A). As the Spearman rank correlation can be seen as gauging the monotonicity of the function mapping one of its argument variables to the other, the UMAP projection o thus appears approximately order-preserving (via composition with the average).

In addition, I computed the centroids (in $\mathbb{R}$) of o-projected point clouds of all sequences, and again correlated these centroids with predicted abundance values. This gave a $\rho = 0.83$ (Fig. 5.2B), which is more evidence that o preserves structure with respect to the order $\prec$ on E. So, empirically for our dataset, in most cases we have:

$$\langle o(c_1) \rangle \leq \langle o(c_2) \rangle \overset{\rho=0.85}{\Longrightarrow} p(c_1) \leq p(c_2) \tag{5.2}$$

$$o\left(\langle c_1 \rangle\right) \leq o\left(\langle c_2 \rangle\right) \overset{\rho=0.83}{\Longrightarrow} p(c_1) \leq p(c_2) \tag{5.3}$$

where $\langle \square \rangle$ denotes average, i.e. the centroid of a set of points.

Interestingly, while $\langle\square\rangle$ and o do not commute, in 97% of cases we have

$$\langle o(c)\rangle \leq o(\langle c\rangle) \tag{5.4}$$

(and their rank correlation $\rho = 0.9$). While unclear what this says about o or E, it seems to show a systematic "bias" rather than disorder.

Given the above evidence, while considering the centroids as representatives of point clouds (both in E and O), I deemed the movement along the O axis to be a good enough proxy for the movement of points on the geodesic in E, and, consequently, as a way to tweak the embedded values of a sequence to increase its abundance. An intuitive way to think about the value on O of a single residue is as its global ranking or sorting (i.e. across all sequences) of its contribution to the predicted abundance value of its sequence.

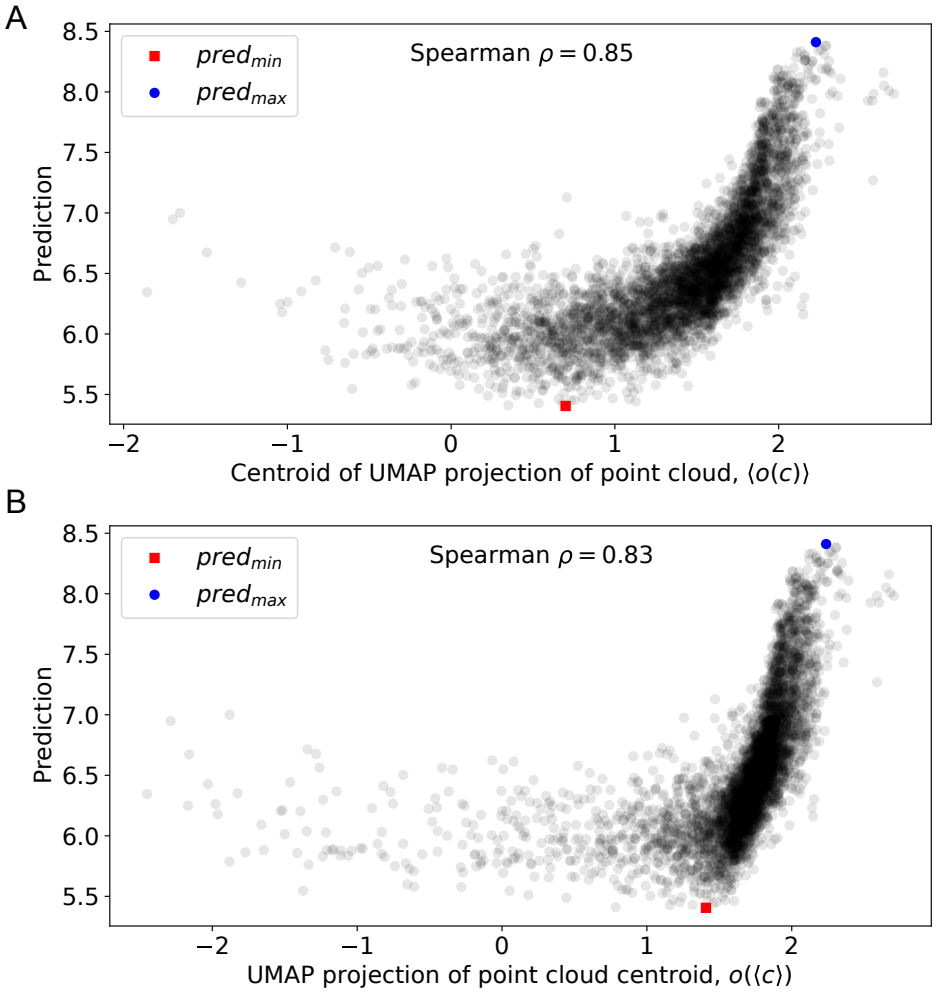


Figure 5.2: The UMAP projection approximately preserves the structure of the embedded space E. Both plots have the points corresponding to the lowest and highest predicted sequence value highlighted by a red square and a blue circle, respectively. **A)** Correlation between centroids on O of UMAP-projected point clouds from the embedded space E with predicted (Box-Cox-transformed) protein abundance. **B)** Correlation between UMAP-projected centroids of point clouds in E with predicted (Box-Cox-transformed) protein abundance.

In order to substitute a residue to increase predicted abundance, the O value of its embedded 1024-point is increased by a large amount, thus shifting it into a region corresponding to high-abundance clouds. This then requires a way to get the amino acid that corresponds to this increased O value. However, we do not have a backward mapping from this space to amino acids. As a workaround solution, I opted for a guided approach, taking the point clouds of the 5 highest abundance sequences as substitute candidates. (The number is arbitrary and more than one sequence was taken to allow for some diversity.) The algorithm chooses the closest point on the O axis in any guide cloud to the increased O value (Figure 4C in the paper). The amino acid corresponding to this guide point is used as substitute for the wild type residue chosen for mutation. The intuition

behind this was that however that guide point contributes to the high abundance of its guide sequence, it should fulfill a similar role for the wild type sequence, as it is closest to the increased O value of the wild type residues under mutation.

To reiterate, the *mutation algorithm* for a sequence s_{wt} using the embedded ordering proceeds as follows:

Algorithm 5.1 Guided Mutation Algorithm

1: Set the number of residues k to mutate
2: Project the embedded w.t. residue points $e(s_{wt})$ with UMAP down to O
3: Choose the top 5 highest abundance sequences as guides and project their point clouds down to O as well
4: Choose the k residues with the lowest O values (lowest "ranking" or "contribution") for shifting
5: Increase the O values of these k points with a large amount. In the paper, a fixed value of 10 was chosen as this spans a good deal of the range of values in O across all sequences (the abscissa in Fig. 4B in the paper). (Decrease the value if the goal is lowering the predicted abundance.)
6: For each increased O value:
    1. find the closest guide point on O
    2. take the amino acid corresponding to this guide point (i.e. from the guide sequence)
    3. substitute the residue in s_{wt} with this amino acid

It would be perhaps improper to say this framework gives transparency to the model, as it relies on the distorting UMAP projection to approximately map the topology of the network's embedded space, based on several assumption about this high-dimensional space. Regardless, it provides a way to bypass understanding of the exact behavior of the model and perform an optimization task on protein sequence space. Of course, much of this framework is predicated on the performance of the model and how well it is able to approximate an ideal sequence-to-abundance function. Still, *in silico* results were orders of magnitude larger than random mutation (Fig. 4D in the paper). And as suggested above, there is room for improvement. The ideal version of this method would have a mapping from O back to sequence space, giving an amino acid for an arbitrary O or E point, without the need for guide sequences. Further investigating the properties outlined above more thoroughly (and rigorously) was beyond the scope of the study.

In terms of method development, my attempt was to approach the probing and mapping of the spaces learned by the model in a systematic fashion, inspired by order theory and topology, as well as the intuitive way deep models are understood to behave.

Finally, it should be noted that the framework is generally applicable to any sequence-to-reals BERT model, not relying on any domain knowledge. Quite likely it generalizes to other types of deep models as well, provided they learn a similarly-structured embedded space. While it is not clear how important the positional encoding is for the approach, this type of encoding is however a common feature in current sequence models [Vas+17].

6 | Conclusions and outlook

In this thesis I have described the data-driven approach to studying the proteome, first by outlining the philosophical outlook, a complementary approach to hypothesis-driven research, then by giving examples of its usage and the benefits of deep machine learning techniques to model complex biological systems, showing that much can be learned from only sequence data, as the carrier of both functional information and evolutionary conditioning, and finally by discussing how the derived deep models may be used to generate new hypotheses, either directly by inspecting their inner workings, or by using them to validate more mechanistic models. In the main chapters of the work, I have listed several projects that take advantage of both large amounts of experimental data and computing power, two crucial underpinnings of the approach.

The first application was the efficient implementation of an unsupervised decomposition procedure of dense mass spectrometric data, enabling its application to high-throughput protein quantification. Benefits of the pipeline are the recovery of the majority of scan signal, in contrast to existing methods relying on spectral libraries, and enabling precise quantification using standard downstream software. As a data-driven methodology, the variability in the data itself is used for the decomposition, with a limited amount of natural assumptions, such as non-negativity and the elution behavior of analytes.

The other three papers included in the thesis demonstrated that sequence alone may be used to predict different aspects of proteomic phenotype. From amino acid sequence one may predict optimum organism growth temperature and thereby learn protein features which can be repurposed to related problems, for instance predicting enzyme catalytic temperature. This type of sequence data was also used in a more abstract model to predict protein abundance, and, similarly, a deep model was trained to predict mRNA abundance from DNA sequence.

By probing these models through different techniques, insights were gained into the sequence logic that determines their predictions. The gene was shown to be a co-evolving unit, where regulatory and coding regions control the level of transcript in tandem. Moreover, a grammar of motif combinations across the regulatory regions was derived and used for engineering. As a determinant of the cell's protein content, the interactions of residues along the entire length of the amino acid sequence were shown to be relevant and hierarchically organized. Moreover, physicochemical properties and various domains were highlighted as

relevant for the predicted protein amount. The sequence features learned by the temperature-predicting model showed agreement with known factors that influence the thermal adaptation of proteins, such as hydrophobicity and secondary structure. Additionally, the most prediction-relevant domains for higher temperatures were those associated to metabolic processes and response to stress, hinting at the importance of these for thermophilic organisms.

To gain more benefit from the transcript and protein quantity predicting models described in the respective papers, I illustrated how they may be combined to yield improved predictions of protein amounts, by constructing a meta-model using predictions as surrogates of the sequence-dependent models themselves. Thus, while technical or complexity limitations can hinder development of an overarching deep model, such surrogate or, alternatively, heterogenous ensemble models may be constructed by composing partial models of the greater system.

Finally, I have described a mathematical approach to sort and manipulate the sequence representations learned by the Transformer-type deep model, allowing for an exploration of sequence space informed by the learning task that can be used to support engineering by providing optimized protein mutants. The framework relies only on the manifold hypothesis that the representations are organized in a near linear fashion in the embedded space of the deep model. It is thus usable for any similar sequence-to-value model, provided it was able to adequately learn its task.

In closing, it is enticing to think about further developments that integrate data-derived models and the overall understanding and theory building. While uncertain that a hypothetical Biological Theory of Everything may be formulated through a only handful of laws, at least with current approaches, or that its formulation would provide effective predictive models, the integrative program of systems biology and related areas provides a guide for bridging different modeling and theory-building paradigms, of which the data-driven approach is one. On the other hand, the drawback of the powerful deep learning models currently in wide use across science is their opaque or extremely complicated inner workings. However, efforts are being made to construct more transparent models, that would lend themselves to easier human interpretation. On the experimental side, technological advances are yielding ever increasing amounts of data to explore. It will be most intriguing to see whether such developments will foster new analytical frameworks or perhaps even a new calculus of deep learning to be used in building models. But the factor I suggest is ultimately essential is the collaborations and insights through multidisciplinary participation, as the nature of projects that model complex biological systems demands sustained awareness of different

fields and receptiveness to evolving techniques, almost regardless of how initial goals of such projects are satisfied. As with many things, the scientific journey is at least as important as the destination.

A | Appendix

Table A.1: CNN parameters after hyperparameter search respecting the train-test split of the mRNA data in Paper IV. The range of the hyperparameter search is the same as the one in **Paper III** (Supplementary Information). The network consists of 7 convolutional layers with batch normalization and dropout, each layer with the given number of filters, convolutional kernel size, dilation rate, and connection dropout rate.

Hyperparameter	Value
num_conv1d_layers	7
learning_rate	0.0006
beta_1	0.714
beta_2	0.714
filters_0	64
kernel_size_0	40
dilation_rate_0	3
conv1d_dropout_rate_0	0.81
filters_1	128
kernel_size_1	50
dilation_rate_1	1
conv1d_dropout_rate_1	0.5
filters_2	128
kernel_size_2	80
dilation_rate_2	5
conv1d_dropout_rate_2	0.83
filters_3	64
kernel_size_3	80
dilation_rate_3	1
conv1d_dropout_rate_3	0.565
dense_units_0	128
dense_dropout_rate_0	0.346
dense_units_1	128
dense_dropout_rate_1	0.842
filters_4	64
kernel_size_4	20
dilation_rate_4	1
conv1d_dropout_rate_4	0.1
filters_5	64
kernel_size_5	20
dilation_rate_5	1
conv1d_dropout_rate_5	0.1
filters_6	64
kernel_size_6	20
dilation_rate_6	1
conv1d_dropout_rate_6	0.1

Table A.2: Hyperparameters of the random forest surrogate model. All other parameters were left as default (scikit-learn 0.22.2).

Hyperparameter	Value	Description
`n_estimators`	5000	n. of trees in the ensemble
`max_depth`	10	depth limit of trees in ensemble
`min_samples_leaf`	2	the n. of samples of lead nodes
`bootstrap`	True	whether to use bootstrapping (training set sampling)
`max_features`	sqrt	size of training set sample
`random_state`	42	random seed to reproduce results

Table A.3: Software used for the results in the thesis.

Package	Version	Use
scikit-learn	0.22.2	random forest and linear regression mdoels, data processing
tensorflow	2.3.0	CNN models
keras	1.1.2	CNN models, data processing
numpy	1.18.5	general
scipy	1.5.3	general
pandas	1.2.2	data processing and analysis
seaborn	0.11.1	plotting
matplotlib	3.3.2	plotting
upsetplot	0.4.1	plotting

References

[14] *When Is Correlation Transitive?* What's new. June 5, 2014. URL: https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/ (visited on 09/05/2021).

[Aar07] Scott Aaronson. "The Limits of Quantum Computers". In: *Computer Science – Theory and Applications*. Ed. by Volker Diekert, Mikhail V. Volkov, and Andrei Voronkov. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 4–4. ISBN: 978-3-540-74510-5. DOI: 10.1007/978-3-540-74510-5_2.

[Aku+07] Tatsuya Akutsu et al. "Control of Boolean Networks: Hardness Results and Algorithms for Tree Structured Networks". In: *Journal of Theoretical Biology* 244.4 (Feb. 21, 2007), pp. 670–679. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2006.09.023. URL: https://www.sciencedirect.com/science/article/pii/S0022519306004334 (visited on 10/21/2021).

[Alb15] Bruce Alberts. *Molecular Biology of the Cell*. Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015. 1 p. ISBN: 978-0-8153-4524-4.

[All+19] Ethan C. Alley et al. "Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning". In: *Nature Methods* 16.12 (12 Dec. 2019), pp. 1315–1322. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0598-1. URL: https://www.nature.com/articles/s41592-019-0598-1 (visited on 11/01/2021).

[Ang+16] Christof Angermueller et al. "Deep Learning for Computational Biology". In: *Molecular Systems Biology* 12.7 (July 2016), p. 878. ISSN: 1744-4292, 1744-4292, 1744-4292. DOI: 10.15252/msb.20156651. URL: http://msb.embopress.org/lookup/doi/10.15252/msb.20156651 (visited on 11/06/2017).

[Blu+21] Matthias Blum et al. "The InterPro Protein Families and Domains Database: 20 Years On". In: *Nucleic Acids Research* 49.D1 (Jan. 8, 2021), pp. D344–D354. ISSN: 0305-1048. DOI:

10.1093/nar/gkaa977. URL: https://doi.org/10.1093/nar/gkaa977 (visited on 08/29/2021).

[Bok11] Alisa Bokulich. "How Scientific Models Can Explain". In: *Synthese* 180.1 (May 1, 2011), pp. 33–45. ISSN: 1573-0964. DOI: 10.1007/s11229-009-9565-1. URL: https://doi.org/10.1007/s11229-009-9565-1 (visited on 10/23/2021).

[Bon+16] Mads T. Bonde et al. "Predictable Tuning of Protein Expression in Bacteria". In: *Nature Methods* 13.3 (3 Mar. 2016), pp. 233–236. ISSN: 1548-7105. DOI: 10.1038/nmeth.3727. URL: https://www.nature.com/articles/nmeth.3727 (visited on 10/27/2021).

[Bru+15] Roland Bruderer et al. "Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues *[S]". In: *Molecular & Cellular Proteomics* 14.5 (May 1, 2015), pp. 1400–1410. ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.M114.044305. URL: https://www.mcponline.org/article/S1535-9476(20)32885-1/abstract (visited on 11/09/2021).

[BS20] Christopher Buccitelli and Matthias Selbach. "mRNAs, Proteins and the Emerging Principles of Gene Expression Control". In: *Nature Reviews Genetics* 21.10 (10 Oct. 2020), pp. 630–644. ISSN: 1471-0064. DOI: 10.1038/s41576-020-0258-4. URL: https://www.nature.com/articles/s41576-020-0258-4 (visited on 11/12/2020).

[BV20] Rosalin Bonetta and Gianluca Valentino. "Machine Learning Techniques for Protein Function Prediction". In: *Proteins: Structure, Function, and Bioinformatics* 88.3 (2020), pp. 397–413. ISSN: 1097-0134. DOI: 10.1002/prot.25832. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25832 (visited on 11/13/2020).

[BZZ20] Filip Buric, Jan Zrimec, and Aleksej Zelezniak. "Parallel Factor Analysis Enables Quantification and Identification of Highly Convolved Data-Independent-Acquired Protein Spectra". In: *Patterns* 1.9 (Dec. 11, 2020). ISSN: 2666-3899. DOI: 10.1016/j.patter.2020.100137. URL: https://www.cell.com/patterns/abstract/S2666-3899(20)30185-9 (visited on 01/11/2021).

[CFC05] Emidio Capriotti, Piero Fariselli, and Rita Casadio. "I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure". In: *Nucleic Acids Research* 33 (suppl_2 July 1, 2005), W306–W310. ISSN: 0305-1048. DOI: 10.1093/nar/gki375. URL: https://doi.org/10.1093/nar/gki375 (visited on 11/12/2021).

[Cha+20] Sreenivas Chavali et al. "Amino Acid Homorepeats in Proteins". In: *Nature Reviews Chemistry* 4.8 (8 Aug. 2020), pp. 420–434. ISSN: 2397-3358. DOI: 10.1038/s41570-020-0204-1. URL: https://www.nature.com/articles/s41570-020-0204-1 (visited on 09/17/2020).

[CHA04] Cristian I. Castillo-Davis, Daniel L. Hartl, and Guillaume Achaz. "Cis-Regulatory and Protein Evolution in Orthologous and Duplicate Genes". In: *Genome Research* 14.8 (Aug. 1, 2004), pp. 1530–1536. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.2662504. pmid: 15256508. URL: https://genome.cshlp.org/content/14/8/1530 (visited on 11/05/2021).

[Che+10] Kevin Chen et al. "Correlating Gene Expression Variation with Cis-Regulatory Polymorphism in Saccharomyces Cerevisiae". In: *Genome Biology and Evolution* 2 (Jan. 1, 2010), pp. 697–707. ISSN: 1759-6653. DOI: 10.1093/gbe/evq054. URL: https://doi.org/10.1093/gbe/evq054 (visited on 11/05/2021).

[Col+17] Ben C. Collins et al. "Multi-Laboratory Assessment of Reproducibility, Qualitative and Quantitative Performance of SWATH-Mass Spectrometry". In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00249-5. URL: http://www.nature.com/articles/s41467-017-00249-5 (visited on 11/06/2017).

[CR18] Sean M. Cascarina and Eric D. Ross. "Proteome-Scale Relationships between Local Amino Acid Composition and Protein Fates and Functions". In: *PLOS Computational Biology* 14.9 (Sept. 24, 2018), e1006256. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006256. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006256 (visited on 08/17/2021).

[Cra+20] Miles Cranmer et al. *Discovering Symbolic Models from Deep Learning with Inductive Biases*. Nov. 17, 2020. arXiv: 2006.11287 [astro-ph, physics:physics, stat]. URL: http://arxiv.org/abs/2006.11287 (visited on 10/29/2021).

[Cup+17] Josh T. Cuperus et al. "Deep Learning of the Regulatory Grammar of Yeast 5' Untranslated Regions from 500,000 Random Sequences". In: *Genome Research* 27.12 (Dec. 1, 2017), pp. 2015–2024. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.224964.117. pmid: 29097404. URL: https://genome.cshlp.org/content/27/12/2015 (visited on 10/27/2021).

[DCA19] Siva Krishna Dasari, Abbas Cheddad, and Petter Andersson. "Random Forest Surrogate Models to Support Design Space Exploration in Aerospace Use-Case". In: *Artificial Intelligence Applications and Innovations*. Ed. by John MacIntyre et al. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2019, pp. 532–544. ISBN: 978-3-030-19823-7. DOI: 10.1007/978-3-030-19823-7_45.

[Dem+19] Vadim Demichev et al. "DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput". In: *Nature Methods* (Nov. 25, 2019), pp. 1–4. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0638-x. URL: https://www.nature.com/articles/s41592-019-0638-x (visited on 11/28/2019).

[Dem+21] Vadim Demichev et al. "A Time-Resolved Proteomic and Prognostic Map of COVID-19". In: *Cell Systems* 12.8 (Aug. 18, 2021), 780–794.e7. ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.05.005. pmid: 34139154. URL: https://www.cell.com/cell-systems/abstract/S2405-4712(21)00160-5 (visited on 11/08/2021).

[Deu+15] Eric W. Deutsch et al. "Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics". In: *PROTEOMICS – Clinical Applications* 9.7-8 (2015), pp. 745–754. ISSN: 1862-8354. DOI: 10.1002/prca.201400164. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prca.201400164 (visited on 11/09/2021).

[Deu+18] Eric W. Deutsch et al. "Expanding the Use of Spectral Libraries in Proteomics". In: *Journal of Proteome Research* 17.12 (Dec. 7, 2018), pp. 4051–4060. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.8b00485. URL: https://doi.org/10.1021/acs.jproteome.8b00485 (visited on 01/11/2021).

[Deu+20] Eric W Deutsch et al. "The ProteomeXchange Consortium in 2020: Enabling 'Big Data' Approaches in Proteomics". In: *Nucleic Acids Research* 48.D1 (Jan. 8, 2020), pp. D1145–D1152. ISSN: 0305-1048. DOI: 10.1093/nar/gkz984. URL: https://doi.org/10.1093/nar/gkz984 (visited on 10/29/2021).

[Dev+19] Jacob Devlin et al. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 05/19/2021).

[Dha13] Vasant Dhar. "Data Science and Prediction". In: *Communications of the ACM* 56.12 (Dec. 1, 2013), pp. 64–73. ISSN: 0001-0782. DOI: 10.1145/2500499. URL: https://doi.org/10.1145/2500499 (visited on 10/22/2021).

[DK21] Jiali Duan and C.-C. Jay Kuo. *Bridging Gap between Image Pixels and Semantics via Supervision: A Survey*. July 29, 2021. arXiv: 2107.13757 [cs]. URL: http://arxiv.org/abs/2107.13757 (visited on 08/17/2021).

[DML19] Benjamin Dubreuil, Or Matalon, and Emmanuel D. Levy. "Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-Functional Interactions". In: *Journal of Molecular Biology* 431.24 (Dec. 6, 2019), pp. 4978–4992. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2019.08.008. URL: http://www.sciencedirect.com/science/article/pii/S0022283619305145 (visited on 06/03/2020).

[El-12] Mansi El-Mansi. *Fermentation Microbiology and Biotechnology*. Vol. 3rd ed. Boca Raton, FL: CRC Press, 2012. ISBN: 978-1-4398-5579-9. URL: https://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=416772&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[Fer+21] Mauricio Ferreira et al. "Protein Abundance Prediction Through Machine Learning Methods". In: *Journal of Molecular Biology* 433.22 (Nov. 5, 2021), p. 167267. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2021.167267. URL: https://www.sciencedirect.com/science/article/pii/S0022283621005039 (visited on 10/28/2021).

[FH20] Roman Frigg and Stephan Hartmann. "Models in Science". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/spr2020/entries/models-science/.

[GA21] Charlotte A. G. H. van Gelder and Maarten Altelaar. "Neuroproteomics of the Synapse: Subcellular Quantification of Protein Networks and Signaling Dynamics". In: *Molecular & Cellular Proteomics* 20 (Jan. 1, 2021). ISSN: 1535-9476, 1535-9484. DOI: 10.1016/j.mcpro.2021.100087. pmid: 33933679. URL: https://www.mcponline.org/article/S1535-9476(21)00060-8/abstract (visited on 08/17/2021).

[Gai+20] P. Gainza et al. "Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning". In: *Nature Methods* 17.2 (2 Feb. 2020), pp. 184–192. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0666-6. URL: https://www.nature.com/articles/s41592-019-0666-6 (visited on 06/04/2020).

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[Gil+12] Ludovic C. Gillet et al. "Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis *". In: *Molecular & Cellular Proteomics* 11.6 (June 1, 2012). ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.O111.016717. pmid: 22261725. URL: https://www.mcponline.org/article/S1535-9476(20)30442-4/abstract (visited on 11/08/2021).

[Hås90] Johan Håstad. "Tensor Rank Is NP-Complete". In: *Journal of Algorithms* 11.4 (Dec. 1, 1990), pp. 644–654. ISSN: 0196-6774. DOI: 10.1016/0196-6774(90)90014-6. URL: http://www.sciencedirect.com/science/article/pii/0196677490900146 (visited on 12/06/2019).

[HBB18] Brandon Ho, Anastasia Baryshnikova, and Grant W. Brown. "Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces Cerevisiae Proteome". In: *Cell Systems* 6.2 (Feb. 28, 2018), 192–205.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.12.004. pmid: 29361465. URL:

https://www.cell.com/cell-systems/abstract/S2405-4712(17)30546-X (visited on 01/16/2020).

[HF21] Henriette Haukedal and Kristine K. Freude. "Implications of Glycosylation in Alzheimer's Disease". In: *Frontiers in Neuroscience* 14 (2021), p. 1432. ISSN: 1662-453X. DOI: 10.3389/fnins.2020.625348. URL: https://www.frontiersin.org/article/10.3389/fnins.2020.625348 (visited on 10/12/2021).

[HI97] William E. Hart and Sorin Istrail. "Robust Proofs of NP-Hardness for Protein Folding: General Lattices and Energy Potentials". In: *Journal of Computational Biology* 4.1 (1997), pp. 1–22. URL: http://www.brown.edu/Research/Istrail_Lab/papers/robustproofs.pdf.

[HK62] Michael Held and Richard M. Karp. "A Dynamic Programming Approach to Sequencing Problems". In: *Journal of the Society for Industrial and Applied Mathematics* 10.1 (Mar. 1, 1962), pp. 196–210. ISSN: 0368-4245. DOI: 10.1137/0110015. URL: https://epubs.siam.org/doi/10.1137/0110015 (visited on 11/15/2021).

[HN20] Michelle Heck and Benjamin A. Neely. "Proteomics in Non-Model Organisms: A New Analytical Frontier". In: *Journal of Proteome Research* 19.9 (Sept. 4, 2020), pp. 3595–3606. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.0c00448. URL: https://doi.org/10.1021/acs.jproteome.0c00448 (visited on 08/17/2021).

[Ho+21] J. J. David Ho et al. "Translational Remodeling by RNA-Binding Proteins and Noncoding RNAs". In: *WIREs RNA* 12.5 (2021), e1647. ISSN: 1757-7012. DOI: 10.1002/wrna.1647. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1647 (visited on 11/05/2021).

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[Htu+19] Phu Mon Htut et al. *Do Attention Heads in BERT Track Syntactic Dependencies?* Nov. 27, 2019. arXiv: 1911.12246 [cs]. URL: http://arxiv.org/abs/1911.12246 (visited on 06/30/2021).

[Iqb+05] Khalid Iqbal et al. “Tau Pathology in Alzheimer Disease and Other Tauopathies”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. The Biology and Pathobiology of Tau 1739.2 (Jan. 3, 2005), pp. 198–210. ISSN: 0925-4439. DOI: 10.1016/j.bbadis.2004.09.008. URL: https://www.sciencedirect.com/science/article/pii/S0925443904001784 (visited on 10/14/2021).

[Jac+19] Friedrich Felix Jacob et al. “Spent Yeast from Brewing Processes: A Biodiverse Starting Material for Yeast Extract Production”. In: *Fermentation* 5.2 (2 June 2019), p. 51. DOI: 10.3390/fermentation5020051. URL: https://www.mdpi.com/2311-5637/5/2/51 (visited on 06/24/2020).

[JM09] Dan Jurafsky and James H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 2. ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2009. ISBN: 0-13-187321-0. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07470a&AN=clc.5beca2d8.9971.462f.8134.6b0632966cac&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[Joh+14] Lea G. Johnsen et al. “Automated Resolution of Overlapping Peaks in Chromatographic Data”. In: *Journal of Chemometrics* 28.2 (Feb. 1, 2014), pp. 71–82. ISSN: 1099-128X. DOI: 10.1002/cem.2575. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.2575 (visited on 08/31/2018).

[Jum+21] John Jumper et al. “Highly Accurate Protein Structure Prediction with AlphaFold”. In: *Nature* 596.7873 (7873 Aug. 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 09/04/2021).

[Käl+08] Lukas Käll et al. “Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases”. In: *Journal of Proteome Research* 7.1 (Jan. 1, 2008), pp. 29–34. ISSN: 1535-3893. DOI: 10.1021/pr700600n. URL: https://doi.org/10.1021/pr700600n (visited on 12/08/2018).

[Kar11] Richard M. Karp. “Heuristic Algorithms in Computational Molecular Biology”. In: *Journal of Computer and System Sciences*. Celebrating Karp’s Kyoto Prize 77.1 (Jan. 1, 2011),

pp. 122–128. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2010.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0022000010000930 (visited on 10/21/2021).

[KB09] Tamara G. Kolda and Brett W. Bader. "Tensor Decompositions and Applications". In: *SIAM Review* 51.3 (Aug. 5, 2009), pp. 455–500. ISSN: 0036-1445. DOI: 10.1137/07070111X. URL: https://epubs.siam.org/doi/10.1137/07070111X (visited on 12/06/2019).

[KD20] Ivan V. Korendovych and William F. DeGrado. "De Novo Protein Design, a Retrospective". In: *Quarterly Reviews of Biophysics* 53 (2020). ISSN: 0033-5835, 1469-8994. DOI: 10.1017/S0033583519000131. URL: https://www.cambridge.org/core/journals/quarterly-reviews-of-biophysics/article/de-novo-protein-design-a-retrospective/FF37903868E1651D7E61A8495FB00B50 (visited on 10/15/2021).

[Kme+20] Cathleen Kmezik et al. "Multimodular Fused Acetyl–Feruloyl Esterases from Soil and Gut Bacteroidetes Improve Xylanase Depolymerization of Recalcitrant Biomass". In: *Biotechnology for Biofuels* 13.1 (Mar. 31, 2020), p. 60. ISSN: 1754-6834. DOI: 10.1186/s13068-020-01698-9. URL: https://doi.org/10.1186/s13068-020-01698-9 (visited on 10/15/2021).

[Kos+19] Jean Kossaifi et al. "TensorLy: Tensor Learning in Python". In: *Journal of Machine Learning Research* 20.26 (2019), pp. 1–6. ISSN: 1533-7928. URL: http://jmlr.org/papers/v20/18-277.html (visited on 11/10/2021).

[KP14] Sangtae Kim and Pavel A. Pevzner. "MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics". In: *Nature Communications* 5 (Oct. 31, 2014), p. 5277. ISSN: 2041-1723. DOI: 10.1038/ncomms6277. URL: https://www.nature.com/articles/ncomms6277 (visited on 07/22/2019).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 10/30/2021).

[KTN00] Sandeep Kumar, Chung-Jung Tsai, and Ruth Nussinov. "Factors Enhancing Protein Thermostability". In: *Protein Engineering, Design and Selection* 13.3 (Mar. 1, 2000), pp. 179–191. ISSN: 1741-0126. DOI: 10.1093/protein/13.3.179. URL: https://doi.org/10.1093/protein/13.3.179 (visited on 08/19/2021).

[Kul16] Jerzy K. Kulski. *Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications*. IntechOpen, Jan. 14, 2016. ISBN: 978-953-51-2240-1. DOI: 10.5772/61964. URL: https://www.intechopen.com/chapters/49602 (visited on 10/29/2021).

[LA16] Yansheng Liu and Ruedi Aebersold. "The Interdependence of Transcript and Protein Abundance: New Data–New Complexities". In: *Molecular Systems Biology* 12.1 (Jan. 1, 2016), p. 856. ISSN: 1744-4292. DOI: 10.15252/msb.20156720. URL: https://www.embopress.org/doi/full/10.15252/msb.20156720 (visited on 06/02/2020).

[Lah+17] Petri-Jaan Lahtvee et al. "Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast". In: *Cell Systems* 4.5 (May 24, 2017), 495–504.e5. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.03.003. URL: http://www.sciencedirect.com/science/article/pii/S2405471217300881 (visited on 02/03/2020).

[LBA16] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. "On the Dependency of Cellular Protein Levels on mRNA Abundance". In: *Cell* 165.3 (Apr. 21, 2016), pp. 535–550. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.03.014. pmid: 27104977. URL: https://www.cell.com/cell/abstract/S0092-8674(16)30270-7 (visited on 11/05/2021).

[LCC19] Yang-Ming Lin, Ching-Tai Chen, and Jia-Ming Chang. "MS2CNN: Predicting MS/MS Spectrum Based on Protein Sequence Using Deep Convolutional Neural Networks". In: *BMC Genomics* 20.9 (Dec. 24, 2019), p. 906. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6297-6. URL: https://doi.org/10.1186/s12864-019-6297-6 (visited on 01/13/2021).

[Leo20] Sabina Leonelli. "Scientific Research and Big Data". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab,

Stanford University, 2020. URL: https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.

[Lik09] Vladimir A. Likić. "Extraction of Pure Components from Overlapped Signals in Gas Chromatography-Mass Spectrometry (GC-MS)". In: *BioData Mining* 2.1 (Oct. 12, 2009), p. 6. ISSN: 1756-0381. DOI: 10.1186/1756-0381-2-6. pmid: 19818154.

[Lip17] Zachary C. Lipton. *The Mythos of Model Interpretability*. Mar. 6, 2017. arXiv: 1606.03490 [cs, stat]. URL: http://arxiv.org/abs/1606.03490 (visited on 08/18/2021).

[Lud+18] Christina Ludwig et al. "Data-Independent Acquisition-Based SWATH-MS for Quantitative Proteomics: A Tutorial". In: *Molecular Systems Biology* 14.8 (Aug. 1, 2018), e8126. ISSN: 1744-4292. DOI: 10.15252/msb.20178126. URL: https://www.embopress.org/doi/full/10.15252/msb.20178126 (visited on 08/07/2019).

[Ma15] Bin Ma. "Novor: Real-Time Peptide de Novo Sequencing Software". In: *Journal of The American Society for Mass Spectrometry* 26.11 (Nov. 1, 2015), pp. 1885–1894. ISSN: 1879-1123. DOI: 10.1007/s13361-015-1204-0. URL: https://doi.org/10.1007/s13361-015-1204-0 (visited on 08/05/2019).

[MB12] Susan Michaelis and Jemima Barrowman. "Biogenesis of the Saccharomyces Cerevisiae Pheromone A-Factor, from Yeast Mating to Human Disease". In: *Microbiology and Molecular Biology Reviews* 76.3 (Sept. 1, 2012), pp. 626–651. DOI: 10.1128/MMBR.00010-12. URL: https://journals.asm.org/doi/10.1128/MMBR.00010-12 (visited on 10/12/2021).

[McI+14] Sean McIlwain et al. "Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis". In: *Journal of Proteome Research* 13.10 (Oct. 3, 2014), pp. 4488–4491. ISSN: 1535-3893. DOI: 10.1021/pr500741y. URL: https://doi.org/10.1021/pr500741y (visited on 10/03/2018).

[MDR17] Miguel Correa Marrero, Aalt D. J. van Dijk, and Dick de Ridder. "Sequence-Based Analysis of Protein Degradation Rates". In: *Proteins: Structure, Function, and Bioinformatics* 85.9 (2017), pp. 1593–1601. ISSN: 1097-0134. DOI: 10.1002/prot.25323. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25323 (visited on 06/02/2020).

[Mes+21] Christoph B. Messner et al. "Ultra-Fast Proteomics with Scanning SWATH". In: *Nature Biotechnology* 39.7 (7 July 2021), pp. 846–854. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00860-4. URL: https://www.nature.com/articles/s41587-021-00860-4 (visited on 11/08/2021).

[MH08] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. ISSN: 1533-7928. URL: http://jmlr.org/papers/v9/vandermaaten08a.html (visited on 10/25/2021).

[MHM18] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Dec. 6, 2018. arXiv: 1802.03426 [cs, stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 09/15/2020).

[MMS16] H. Pezeshgi Modarres, M. R. Mofrad, and A. Sanati-Nezhad. "Protein Thermostability Engineering". In: *RSC Advances* 6.116 (Dec. 13, 2016), pp. 115252–115270. ISSN: 2046-2069. DOI: 10.1039/C6RA16992A. URL: https://pubs.rsc.org/en/content/articlelanding/2016/ra/c6ra16992a (visited on 08/21/2021).

[MR18] Thilo Muth and Bernhard Y. Renard. "Evaluating de Novo Sequencing in Proteomics: Already an Accurate Alternative to Database-Driven Peptide Identification?" In: *Briefings in Bioinformatics* 19.5 (Sept. 28, 2018), pp. 954–970. ISSN: 1467-5463. DOI: 10.1093/bib/bbx033. URL: https://academic.oup.com/bib/article/19/5/954/3076504 (visited on 09/18/2019).

[MR20] Ali Mobasheri and Stephen M. Richardson. "Cell and Gene Therapy for Spine Regeneration: Mammalian Protein Production Platforms for Overproduction of Therapeutic Proteins and Growth Factors". In: *Neurosurgery Clinics of North America*. New Technologies in Spine Surgery 31.1 (Jan. 1, 2020), pp. 131–139. ISSN: 1042-3680. DOI: 10.1016/j.nec.2019.08.015. URL: https://www.sciencedirect.com/science/article/pii/S1042368019300774 (visited on 10/18/2021).

[MV17] Miguel Martin-Perez and Judit Villén. "Determinants and Regulation of Protein Turnover in Yeast". In: *Cell Systems* 5.3 (Sept. 27, 2017), 283–294.e5. ISSN: 2405-4712. DOI: 10.

1016/j.cels.2017.08.008. pmid: 28918244. URL: https://www.cell.com/cell-systems/abstract/S2405-4712(17)30341-1 (visited on 10/12/2021).

[Nav+16] Pedro Navarro et al. "A Multicenter Study Benchmarks Software Tools for Label-Free Proteome Quantification". In: *Nature Biotechnology* 34.11 (Nov. 2016), pp. 1130–1136. ISSN: 1546-1696. DOI: 10.1038/nbt.3685. URL: https://www.nature.com/articles/nbt.3685 (visited on 08/30/2018).

[Nie17] Jens Nielsen. "Systems Biology of Metabolism". In: *Annual Review of Biochemistry* 86.1 (2017), pp. 245–275. DOI: 10.1146/annurev-biochem-061516-044757. eprint: https://doi.org/10.1146/annurev-biochem-061516-044757. URL: https://doi.org/10.1146/annurev-biochem-061516-044757.

[NK16] Jens Nielsen and Jay D. Keasling. "Engineering Cellular Metabolism". In: *Cell* 164.6 (Mar. 10, 2016), pp. 1185–1197. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.02.004. pmid: 26967285. URL: https://www.cell.com/cell/abstract/S0092-8674(16)30070-8 (visited on 10/15/2021).

[NP77] G. Nicolis and Ilya Prigogine. *Self-Organization in Nonequilibrium Systems : From Dissipative Structures to Order through Fluctuations.* 1977. ISBN: 0-471-02401-5. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07470a&AN=clc.eb2edd28.33cf.450b.bba6.0e074cb55119&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[Ono+17] Sideney Becker Onofre et al. "Chemical Composition of the Biomass of Saccharomyces Cerevisiae - (Meyen Ex E. C. Hansen, 1883) Yeast Obtained from the Beer Manufacturing Process". In: *International Journal of Environment, Agriculture and Biotechnology* 2.2 (2017), pp. 558–562. ISSN: 24561878. DOI: 10.22161/ijeab/2.2.2. URL: http://ijeab.com/upload_document/issue_files/2%20IJEAB-JAN-2017-37-Chemical%20Composition%20of%20the%20Biomass%20of%20Saccharomyces%20cerevisiae.pdf (visited on 06/24/2020).

[Pea10] Judea Pearl. "Causal Inference". In: *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008.* Causality: Objectives and Assessment. PMLR, Feb. 18, 2010,

pp. 39–58. URL: https://proceedings.mlr.press/v6/pearl10a.html (visited on 10/21/2021).

[Pec+18] Ryan Peckner et al. "Specter: Linear Deconvolution for Targeted Analysis of Data-Independent Acquisition Mass Spectrometry Proteomics". In: *Nature Methods* 15.5 (5 May 2018), pp. 371–378. ISSN: 1548-7105. DOI: 10.1038/nmeth.4643. URL: https://www.nature.com/articles/nmeth.4643 (visited on 11/10/2021).

[Pin+20] Lindsay K. Pino et al. "The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics". In: *Mass Spectrometry Reviews* 39.3 (2020), pp. 229–244. ISSN: 1098-2787. DOI: 10.1002/mas.21540. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21540 (visited on 11/09/2021).

[PL15] Michael S. Packer and David R. Liu. "Methods for the Directed Evolution of Proteins". In: *Nature Reviews Genetics* 16.7 (7 July 2015), pp. 379–394. ISSN: 1471-0064. DOI: 10.1038/nrg3927. URL: https://www.nature.com/articles/nrg3927 (visited on 08/18/2021).

[PMB21] Parenti, Giancarlo, Medina, Diego L, and Ballabio, Andrea. "The Rapidly Evolving View of Lysosomal Storage Diseases". In: *EMBO Molecular Medicine* 13.2 (Feb. 5, 2021), e12836. ISSN: 1757-4676. DOI: 10.15252/emmm.202012836. URL: https://www.embopress.org/doi/full/10.15252/emmm.202012836 (visited on 10/14/2021).

[PMW19] Nishant Pappireddi, Lance Martin, and Martin Wühr. "A Review on Quantitative Multiplexed Proteomics". In: *ChemBioChem* 20.10 (2019), pp. 1210–1224. ISSN: 1439-7633. DOI: 10.1002/cbic.201800650. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201800650 (visited on 11/08/2021).

[PW16] Ky Young Park and Soo Jin Wi. "Potential of Plants to Produce Recombinant Protein Products". In: *Journal of Plant Biology* 59.6 (Dec. 1, 2016), pp. 559–568. ISSN: 1867-0725. DOI: 10.1007/s12374-016-0482-9. URL: https://doi.org/10.1007/s12374-016-0482-9 (visited on 10/18/2021).

[PY10] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1558-2191. DOI: 10.1109/TKDE.2009.191.

[Rao+19] Roshan Rao et al. *Evaluating Protein Transfer Learning with TAPE*. June 19, 2019. arXiv: 1906.08230 [cs, q-bio, stat]. URL: http://arxiv.org/abs/1906.08230 (visited on 03/23/2021).

[RB12] Steven T. Rutherford and Bonnie L. Bassler. "Bacterial Quorum Sensing: Its Role in Virulence and Possibilities for Its Control". In: *Cold Spring Harbor Perspectives in Medicine* 2.11 (Nov. 1, 2012), a012427. ISSN: , 2157-1422. DOI: 10.1101/cshperspect.a012427. pmid: 23125205. URL: http://perspectivesinmedicine.cshlp.org/content/2/11/a012427 (visited on 10/12/2021).

[Rib+19] Andrea Riba et al. "Protein Synthesis Rates and Ribosome Occupancies Reveal Determinants of Translation Elongation Rates". In: *Proceedings of the National Academy of Sciences* 116.30 (July 23, 2019), pp. 15023–15032. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1817299116. pmid: 31292258. URL: https://www.pnas.org/content/116/30/15023 (visited on 01/23/2020).

[Rit+76] Garry L. Ritter et al. "Factor Analysis of the Mass Spectra of Mixtures". In: *Analytical Chemistry* 48.3 (1976), pp. 591–595. DOI: 10.1021/ac60367a028. URL: http://dx.doi.org/10.1021/ac60367a028.

[RKA13] Philip A. Romero, Andreas Krause, and Frances H. Arnold. "Navigating the Protein Fitness Landscape with Gaussian Processes". In: *Proceedings of the National Academy of Sciences* 110.3 (Jan. 15, 2013), E193–E201. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1215251110. pmid: 23277561. URL: https://www.pnas.org/content/110/3/E193 (visited on 11/12/2021).

[RKR20] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. *A Primer in BERTology: What We Know about How BERT Works*. Nov. 9, 2020. arXiv: 2002.12327 [cs]. URL: http://arxiv.org/abs/2002.12327 (visited on 11/24/2020).

[Rös+14] Hannes L. Röst et al. "OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data". In: *Nature Biotechnology* 32.3 (3 Mar. 2014), pp. 219–223. ISSN: 1546-1696. DOI: 10.1038/nbt.2841. URL: https://www.nature.com/articles/nbt.2841 (visited on 11/09/2021).

[Ros+17] George Rosenberger et al. "Inference and Quantification of Peptidoforms in Large Sample Cohorts by SWATH-MS". In: *Nature Biotechnology* 35.8 (8 Aug. 2017), pp. 781–788. ISSN: 1546-1696. DOI: 10.1038/nbt.3908. URL: https://www.nature.com/articles/nbt.3908 (visited on 11/08/2021).

[RT18] Alessandro Raganato and Jörg Tiedemann. "An Analysis of Encoder Representations in Transformer-Based Machine Translation". In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Nov. 2018, pp. 287–297. DOI: 10.18653/v1/W18-5431. URL: https://www.aclweb.org/anthology/W18-5431 (visited on 06/30/2021).

[San+16] Laura Sanchez-Garcia et al. "Recombinant Pharmaceuticals from Microbial Cells: A 2015 Update". In: *Microbial Cell Factories* 15.1 (Feb. 9, 2016), p. 33. ISSN: 1475-2859. DOI: 10.1186/s12934-016-0437-3. URL: https://doi.org/10.1186/s12934-016-0437-3 (visited on 10/18/2021).

[SBG04] Age K. Smilde, Rasmus Bro, and Paul Geladi. *Multi-Way Analysis with Applications in the Chemical Sciences*. Wiley, 2004. ISBN: 0-471-98691-7. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07470a&AN=clc.ec728559.bf45.4f56.8dca.5893372a3ff2&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[Sch+15] Olga T Schubert et al. "Building High-Quality Assay Libraries for Targeted Analysis of SWATH MS Data". In: *Nature Protocols* 10.3 (Feb. 12, 2015), pp. 426–441. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2015.015. URL: http://www.nature.com/doifinder/10.1038/nprot.2015.015 (visited on 11/06/2017).

[Seh+21] David Sehnal et al. "Mol* Viewer: Modern Web App for 3D Visualization and Analysis of Large Biomolecular Structures". In: *Nucleic Acids Research* 49.W1 (July 2, 2021), W431–W437. ISSN: 0305-1048. DOI: 10.1093/nar/gkab314.

URL: https://doi.org/10.1093/nar/gkab314 (visited on 11/15/2021).

[She02] Fred Sherman. "Getting Started with Yeast". In: *Methods in Enzymology*. Ed. by Christine Guthrie and Gerald R. Fink. Vol. 350. Guide to Yeast Genetics and Molecular and Cell Biology - Part B. Academic Press, Jan. 1, 2002, pp. 3–41. DOI: 10.1016/S0076-6879(02)50954-X. URL: http://www.sciencedirect.com/science/article/pii/S007668790250954X (visited on 01/05/2021).

[Shi+16] Jin-Hyung Shim et al. "Three-Dimensional Bioprinting of Multilayered Constructs Containing Human Mesenchymal Stromal Cells for Osteochondral Tissue Regeneration in the Rabbit Knee Joint". In: 8.1 (Feb. 2016), p. 014102. ISSN: 1758-5090. DOI: 10.1088/1758-5090/8/1/014102. URL: https://doi.org/10.1088/1758-5090/8/1/014102 (visited on 10/18/2021).

[Shi00] Hidetoshi Shimodaira. "Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function". In: *Journal of Statistical Planning and Inference* 90.2 (Oct. 1, 2000), pp. 227–244. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(00)00115-4. URL: https://www.sciencedirect.com/science/article/pii/S0378375800001154 (visited on 09/05/2021).

[SK06] Johannes Schneider and Scott Kirkpatrick. *Stochastic Optimization*. 1st ed. 2006. Scientific Computation. Springer Berlin Heidelberg, 2006. ISBN: 978-3-540-34560-2. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07472a&AN=clec.SPRINGERLINK9783540345602&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[SMG20] Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. "Parametric UMAP Embeddings for Representation and Semi-Supervised Learning". In: (Sept. 27, 2020). URL: https://arxiv.org/abs/2009.12981v4 (visited on 10/31/2021).

[Smi+14] Rob Smith et al. "Proteomics, Lipidomics, Metabolomics: A Mass Spectrometry Tutorial from a Computer Scientist's Point of View". In: *BMC Bioinformatics* 15.7 (May 28, 2014), S9. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-S7-S9.

URL: https://doi.org/10.1186/1471-2105-15-S7-S9 (visited on 03/05/2018).

[SMV09] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. "Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression". In: *Nature Biotechnology* 27.10 (10 Oct. 2009), pp. 946–950. ISSN: 1546-1696. DOI: 10.1038/nbt.1568. URL: https://www.nature.com/articles/nbt.1568 (visited on 10/27/2021).

[Spe+20] Maike Sperk et al. "Utility of Proteomics in Emerging and Re-Emerging Infectious Diseases Caused by RNA Viruses". In: *Journal of Proteome Research* 19.11 (Nov. 6, 2020), pp. 4259–4274. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.0c00380. URL: https://doi.org/10.1021/acs.jproteome.0c00380 (visited on 10/15/2021).

[SS13] Kevin Struhl and Eran Segal. "Determinants of Nucleosome Positioning". In: *Nature Structural & Molecular Biology* 20.3 (3 Mar. 2013), pp. 267–273. ISSN: 1545-9985. DOI: 10.1038/nsmb.2506. URL: https://www.nature.com/articles/nsmb.2506 (visited on 11/05/2021).

[Ste+15] Zachary D. Stephens et al. "Big Data: Astronomical or Genomical?" In: *PLOS Biology* 13.7 (July 7, 2015), e1002195. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002195. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195 (visited on 10/21/2021).

[Ste12] Eric J. Stewart. "Growing Unculturable Bacteria". In: *Journal of Bacteriology* 194.16 (Aug. 15, 2012), pp. 4151–4160. DOI: 10.1128/JB.00345-12. URL: https://journals.asm.org/doi/10.1128/JB.00345-12 (visited on 10/21/2021).

[Str14] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Boulder, UNITED STATES: Westview Press, 2014. ISBN: 978-0-8133-4911-4. URL: http://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=1181622 (visited on 10/19/2021).

[Sun+05] Jianping Sun et al. "Solution Structure of Kti11p from Saccharomyces Cerevisiae Reveals a Novel Zinc-Binding Module," in: *Biochemistry* 44.24 (June 1, 2005), pp. 8801–8809. ISSN: 0006-2960. DOI: 10.1021/bi0504714. URL:

https://doi.org/10.1021/bi0504714 (visited on 11/15/2021).

[Sur+19] Ahmet Sureyya Rifaioglu et al. "DEEPred: Automated Protein Function Prediction with Multi-Task Feed-Forward Deep Neural Networks". In: *Scientific Reports* 9.1 (1 May 14, 2019), p. 7344. ISSN: 2045-2322. DOI: 10.1038/s41598-019-43708-3. URL: https://www.nature.com/articles/s41598-019-43708-3 (visited on 11/12/2021).

[Thu94] Hugh Thurston. *Early Astronomy. [Electronic Resource]*. 1st ed. 1994. Springer Study Edition. Springer New York, 1994. ISBN: 978-1-4612-4322-9. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07472a&AN=clec.SPRINGERLINK9781461243229&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[Tin+17] Ying S. Ting et al. "PECAN: Library-Free Peptide Detection for Data-Independent Acquisition Tandem Mass Spectrometry Data". In: *Nature Methods* 14.9 (9 Sept. 2017), pp. 903–908. ISSN: 1548-7105. DOI: 10.1038/nmeth.4390. URL: https://www.nature.com/articles/nmeth.4390 (visited on 11/09/2021).

[Tou+20] V. Tournier et al. "An Engineered PET Depolymerase to Break down and Recycle Plastic Bottles". In: *Nature* 580.7802 (7802 Apr. 2020), pp. 216–219. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2149-4. URL: https://www.nature.com/articles/s41586-020-2149-4 (visited on 10/18/2021).

[Tra+17] Ngoc Hieu Tran et al. "De Novo Peptide Sequencing by Deep Learning". In: *Proceedings of the National Academy of Sciences* 114.31 (Aug. 1, 2017), pp. 8247–8252. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1705691114. pmid: 28720701. URL: https://www.pnas.org/content/114/31/8247 (visited on 11/09/2021).

[Tra+19] Ngoc Hieu Tran et al. "Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry". In: *Nature Methods* 16.1 (1 Jan. 2019), pp. 63–66. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0260-3. URL: https://www.nature.com/articles/s41592-018-0260-3 (visited on 01/13/2021).

[Tso+15] Chih-Chiang Tsou et al. “DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics”. In: *Nature Methods* 12.3 (3 Mar. 2015), pp. 258–264. ISSN: 1548-7105. DOI: 10.1038/nmeth.3255. URL: https://www.nature.com/articles/nmeth.3255 (visited on 03/25/2020).

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 04/29/2021).

[vdBer+14] Bastiaan A. van den Berg et al. “Protein Redesign by Learning from Data”. In: *Protein Engineering, Design and Selection* 27.9 (Sept. 1, 2014), pp. 281–288. ISSN: 1741-0126. DOI: 10.1093/protein/gzu031. URL: https://academic.oup.com/peds/article/27/9/281/2272557 (visited on 01/21/2020).

[vDH13] JanMaarten van Dijl and Michael Hecker. “Bacillus Subtilis: From Soil Bacterium to Super-Secreting Cell Factory”. In: *Microbial Cell Factories* 12.1 (Jan. 14, 2013), p. 3. ISSN: 1475-2859. DOI: 10.1186/1475-2859-12-3. URL: https://doi.org/10.1186/1475-2859-12-3 (visited on 10/18/2021).

[Vec+18] Ignazio Vecchio et al. “The Discovery of Insulin: An Important Milestone in the History of Medicine”. In: *Frontiers in Endocrinology* 9 (2018), p. 613. ISSN: 1664-2392. DOI: 10.3389/fendo.2018.00613. URL: https://www.frontiersin.org/article/10.3389/fendo.2018.00613 (visited on 10/18/2021).

[Ven+04] John D. Venable et al. “Automated Approach for Quantitative Analysis of Complex Peptide Mixtures from Tandem Mass Spectra”. In: *Nature Methods* 1.1 (1 Oct. 2004), pp. 39–45. ISSN: 1548-7105. DOI: 10.1038/nmeth705. URL: https://www.nature.com/articles/nmeth705 (visited on 11/08/2021).

[Ver+19] Manasvi Verma et al. “A Short Translational Ramp Determines the Efficiency of Protein Synthesis”. In: *Nature Communications* 10.1 (Dec. 18, 2019), pp. 1–15. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13810-1. URL: https://www.nature.com/articles/s41467-019-13810-1 (visited on 02/03/2020).

[Vig+20] Jesse Vig et al. *BERTology Meets Biology: Interpreting Attention in Protein Language Models*. July 13, 2020. arXiv: 2006.15222 [cs, q-bio]. URL: http://arxiv.org/abs/2006.15222 (visited on 03/23/2021).

[Vig19] Jesse Vig. *A Multiscale Visualization of Attention in the Transformer Model*. June 12, 2019. arXiv: 1906.05714 [cs]. URL: http://arxiv.org/abs/1906.05714 (visited on 10/30/2021).

[VM12] Christine Vogel and Edward M. Marcotte. "Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses". In: *Nature Reviews Genetics* 13.4 (4 Apr. 2012), pp. 227–232. ISSN: 1471-0064. DOI: 10.1038/nrg3185. URL: https://www.nature.com/articles/nrg3185 (visited on 06/02/2020).

[Vow+18] Jakob Vowinckel et al. "Cost-Effective Generation of Precise Label-Free Quantitative Proteomes in High-Throughput by microLC and Data-Independent Acquisition". In: *Scientific Reports* 8.1 (Mar. 12, 2018), p. 4346. ISSN: 2045-2322. DOI: 10.1038/s41598-018-22610-4. URL: https://www.nature.com/articles/s41598-018-22610-4%C2%A0 (visited on 01/30/2019).

[Web+20] Marc Weber et al. "Impact of C-Terminal Amino Acid Composition on Protein Expression in Bacteria". In: *Molecular Systems Biology* 16.5 (May 1, 2020), e9208. ISSN: 1744-4292. DOI: 10.15252/msb.20199208. URL: https://www.embopress.org/doi/10.15252/msb.20199208 (visited on 06/04/2020).

[WJ94] Lusheng Wang and Tao Jiang. "On the Complexity of Multiple Sequence Alignment". In: *Journal of Computational Biology* 1.4 (Jan. 1, 1994), pp. 337–348. DOI: 10.1089/cmb.1994.1.337. URL: https://www.liebertpub.com/doi/10.1089/cmb.1994.1.337 (visited on 10/21/2021).

[WKW16] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. "A Survey of Transfer Learning". In: *Journal of Big Data* 3.1 (May 28, 2016), p. 9. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6. URL: https://doi.org/10.1186/s40537-016-0043-6 (visited on 09/05/2021).

[WP18] Roy S. K. Walker and Isak S. Pretorius. "Applications of Yeast Synthetic Biology Geared towards the Production of Biopharmaceuticals". In: *Genes* 9.7 (7 July 2018), p. 340. DOI: 10.3390/genes9070340. URL: https://www.mdpi.com/2073-4425/9/7/340 (visited on 10/18/2021).

[WP19] Sarah Wiegreffe and Yuval Pinter. "Attention Is Not Not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://www.aclweb.org/anthology/D19-1002 (visited on 03/24/2021).

[Yan+18] Kevin K Yang et al. "Learned Protein Embeddings for Machine Learning". In: *Bioinformatics* 34.15 (Aug. 1, 2018), pp. 2642–2648. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty178. URL: https://doi.org/10.1093/bioinformatics/bty178 (visited on 11/01/2021).

[YWA19] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. "Machine-Learning-Guided Directed Evolution for Protein Engineering". In: *Nature Methods* 16.8 (8 Aug. 2019), pp. 687–694. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0496-6. URL: https://www.nature.com/articles/s41592-019-0496-6 (visited on 08/18/2021).

[Zel+18] Aleksej Zelezniak et al. "Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts". In: *Cell Systems* 7.3 (Sept. 26, 2018), 269–283.e6. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2405471218303168 (visited on 10/29/2021).

[ZF14] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 818–833. ISBN: 978-3-319-10590-1. DOI: 10.1007/978-3-319-10590-1_53.

[Zha+] Fangfei Zhang et al. "Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020". In: *PROTEOMICS* n/a.n/a (), p. 1900276.

ISSN: 1615-9861. DOI: 10.1002/pmic.201900276. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201900276 (visited on 05/06/2020).

[ZKZ16] Bo Zhang, Lukas Käll, and Roman A. Zubarev. "DeMix-Q: Quantification-Centered Data Processing Workflow*". In: *Molecular & Cellular Proteomics* 15.4 (Apr. 1, 2016), pp. 1467–1478. ISSN: 1535-9476. DOI: 10.1074/mcp.O115.055475. URL: https://www.sciencedirect.com/science/article/pii/S1535947620336343 (visited on 11/08/2021).

[Zri+20a] Jan Zrimec et al. "Deep Learning Suggests That Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure". In: *Nature Communications* 11.1 (1 Dec. 1, 2020), p. 6141. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19921-4. URL: https://www.nature.com/articles/s41467-020-19921-4 (visited on 10/29/2021).

[Zri+20b] Jan Zrimec et al. "Plastic-Degrading Potential across the Global Microbiome Correlates with Recent Pollution Trends". In: (Dec. 15, 2020), p. 2020.12.13.422558. DOI: 10.1101/2020.12.13.422558. URL: https://www.biorxiv.org/content/10.1101/2020.12.13.422558v2 (visited on 10/18/2021).

[Zri+21] Jan Zrimec et al. "Learning the Regulatory Code of Gene Expression". In: *Frontiers in Molecular Biosciences* 8 (2021), p. 530. ISSN: 2296-889X. DOI: 10.3389/fmolb.2021.673363. URL: https://www.frontiersin.org/article/10.3389/fmolb.2021.673363 (visited on 09/07/2021).