doi:10.2533/chimia.2019.540 Chimia 73 (2019) 540–548 

# Mass Spectrometric Exploration of the Biochemical Basis of Living Systems

Ruedi Aebersold[§ab*] and Peter Blattmann[a]

[§]Paracelsus Prize 2018

*Abstract:* Predicting how a system behaves under changing conditions is an essential component of science and engineering. The ability to make accurate predictions about the system indicates that it is well understood and provides the opportunity to simulate the response to conditions that would be empirically difficult or impossible to test. In the life sciences, the term *systems biology* was introduced to articulate the notion that the molecular and phenotypic response of a cell or organism to perturbations is the result of interplay of a multitude of molecules. The ability to predict the behavior of such complex molecular systems remains challenging and inevitably requires the involvement of different types of models and data that support them. In this article, we discuss a range of data-driven models that have proven particularly useful for predicting the behavior of biological systems at different levels of complexity and the matching data generation methods that support them. We specifically focus on predictions based on protein or proteome data generated by mass spectrometry. We describe three case studies that represent frequently encountered situations in systems biology.

**Keywords**: Mass spectrometry · Mathematical modeling · Proteomics · SWATH/DIA · Systems Biology

***Ruedi Aebersold*** is a Swiss and Canadian scientist trained at the Biocenter, University of Basel. He completed his education at Caltech. He is a Professor at ETH Zurich and the University of Zurich. He was on the faculties of the Universities of British Columbia and Washington and co-founded, with Lee Hood and Alan Aderem, the Institute for Systems Biology in Seattle. He is on the SAB of a number of research organizations and has served as senior editor for *Molecular and Cellular Proteomics* and *Molecular Systems Biology*. He has co-founded several companies and holds several public service appointments. The research focus of his group is the proteome. The group has pioneered several widely used techniques and generated a range of open access/open source software and statistical tools that have contributed to making proteomic research results more transparent, accurate and reproducible.

## 1. Predictive Models and their Dependencies

Predicting the future is a natural human desire. In ancient times, oracles claimed to predict fate and fulfilled such desires. Once scientists began to understand the principles of natural processes such as weather, sun eclipses, or diseases, they started to make predictions about these phenomena. Predictions are inherent to the scientific method, that relies on formulating hypotheses from the current body of knowledge that are then tested in an experiment. Furthermore, predictions are important in non-scientific disciplines such as sports betting, election outcomes, or demographics. In some instances, predictions have become very accurate, such as for sun eclipses, weather for the next few days, or the behavior of engineered devices such as clocks or motors. In contrast, accurate predictions in other fields, exemplified by earthquakes or stock market changes, have remained very difficult or impossible.

In chemistry and the life sciences, the need to make predictions has a long history. Medicinal chemists have used methods like quantitative structure–activity relationship (QSAR) to predict how structural changes of a molecular entity affect its function (reviewed in Nelson and Ismail[1]). In the life sciences, particularly in medicine, significant efforts have been spent on identifying molecular or physiological markers, so-called 'biomarkers', capable of predicting the health trajectory of individuals or groups of individuals. Outstanding examples of such predictions include the Framingham risk score for cardiovascular disease, the APGAR score to assess the health of newborn babies, and PSA score indicating the likelihood of the presence of prostate cancer in a patient. Whereas the predictive accuracy of many of these tests remains far from perfect, the expectations to employ molecular patterns to support important mechanistic, epidemiological, or clinical predictions has dramatically increased over the last years. This is illustrated by the arrival of the personalized or precision medicine era, which is rooted in the expectation that multiple layers of high dimensional data, typically data collected by 'Omics technologies' on a specific person, can be computationally integrated and analyzed to make predictions that guide and improve medical therapy of the tested individual. Similarly, in basic biology, the emergence of the systems biology paradigm firmly established the need for predictive models of biological processes to provide insights into their complexities of design and operation. Outstanding examples of this type of effort include mechanistic models of the bacterial flagellar motor,[2] the circadian clock[3] or the mitotic cell cycle.[4]

*Correspondence: Prof. R. Aebersold[ab]
E-mail: aebersold@imsb.biol.ethz.ch
[a]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Otto-Stern-Weg 3, CH-8093 Zurich; [b]Faculty of Science, University of Zurich, Zurich

To make accurate predictions, essentially two main elements are required: Theory and data. Theory relies on so-called first principles and scientists, particularly theoretical physicists like Einstein over the last century, provided powerful examples how theories can advance our understanding of complex systems. The other element is data, that are acquired by careful observation of a process or system. Epidemiology, astronomy, and many other fields of the natural sciences have greatly profited from accurate observations and the ensuing data then paved the way for the discovery of the underlying principles of these processes and systems. At the end, both accurate data and a good theoretical understanding are required to be fully able to predict the behavior of a system. However, for different fields of science, the relative availability of theory and data varies considerably (Fig. 1). For example, astronomers can predict sun eclipses based on accurate data from the location of celestial bodies and because they have a profound understanding of the physical laws that define their movements. Similarly, in engineering or particle physics, a strong knowledge of the first principles enables engineers to build airplanes whose aerodynamic properties are then assessed in specific test flights or physicists to design the Large Hadron Collider experiments to test their theories. In contrast, the molecular life sciences generally lack first principles and theory. Therefore, predictive models in biology and medicine are substantially dependent on acquired data that indicate the acute state of the system. Encapsulated under the title 'Omics research', a range of technologies has been developed over the last two decades that now generate large volumes of reproducible, quantitatively accurate, and comprehensive molecular data of biological and clinical specimens. These data can be used to support data-driven, statistical predictions or to infer relationships, even in the absence of strong theory. In other fields, neither theory nor sufficient data are available to support accurate predictions. It is for example very challenging to predict how to form a successful company, how to make good laws, or how to guide political or economic decisions. In these fields, decisions are mainly guided by intuition.

Hence, different fields of science and human activity find themselves in drastically different areas in the data vs. first principle graph displayed in Fig. 1. Yet, they aim to achieve the same objective; to make accurate predictions. Fields in which first principles and theory are well developed can use the theory to generate models whose predictions can then be tested in experiments. In fields which are rich in data, the inverse approach of using data to generate models can be employed. This is, however, inherently more difficult. Regression, classification, or neural networks are groups of algorithms that have been applied to large data sets with the aim of better understanding the underlying processes and infer predictive models. These models should ideally not only explain the actual case, but be generalizable and thus applicable to other situations, even situations that are for practical reasons not experimentally testable. The recently developed Omics technologies have transformed biology and medicine into the domain of data-rich sciences. In the following paragraphs, we will discuss approaches that have been used to gain biological knowledge from molecular data in the life sciences and relate them to the methods that were used to generate data suitable for these analyses.

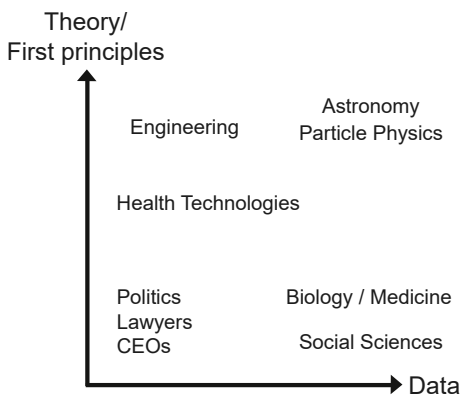


Fig. 1. The data vs. first principle graph. A detailed understanding enabling accurate prediction of processes requires both advanced theoretical knowledge (y-axis) and available accurate data (x-axis). Different scientific and non-scientific fields attaining predictions find themselves in different areas of this landscape.

## 2. Techniques to Generate Protein Data to Support Predictive, Data-driven Models

### 2.1 *General Considerations on Biomolecular Data Generation*

In the absence of a comprehensive theory to predict the behavior of biological systems, data-driven approaches need to be pursued. Over the past few decades, life science research has been transformed by the development of a wide range of techniques to reproducibly and systematically identify and quantify different types of biomolecules. The need to sequence nucleic acids at a large scale pioneered these developments and culminated in the current, powerful genomics techniques that preferentially use fluorescent labels as a readout to quantify the amplified nucleic acid molecules. In contrast, the analysis of other types of biomolecules depends on signals obtained from the native materials extracted from biological specimens. In such cases mass spectrometry is the analytical method of choice to analyze metabolites, glycans, lipids, proteins, and peptides. Because proteins are the type of biomolecule that exert most biological functions and thus define, as an ensemble, the biochemical state of a cell, we focus this manuscript on mass spectrometric methods that generate data suitable to support functional predictions of cells or tissues. The systematic, parallel analysis of many proteins extracted from a biological sample is referred to as proteomics.

### 2.2 *General Principles of Mass Spectrometric Analysis of Proteins and Proteomes*

Mass spectrometric (MS) measurements require that the tested analyte be present in the vacuum system of the mass spectrometer in ionized form. MS is a generic method and more than 100 years old,[5] but its application to proteins and peptides has a much shorter history. For the traditionally available ionization methods, it was very challenging to ionize peptides and proteins, without obliterating them. This changed in the late 1980s when almost concurrently two 'soft' ionization methods were developed that provided a rather general solution to the problem of ionizing intact peptides, proteins and other larger molecules like glycans and lipids. These two methods, electrospray ionization (ESI)[6] and matrix assisted laser desorption ionization (MALDI)[7] respectively, transformed protein and proteome research. With these ionization methods essentially any polypeptide could be ionized and, provided a suitable instrument was interfaced, transferred into the gas phase and subjected to mass spectrometric analysis.

Over the last three decades, a wide range of techniques and MS instruments for proteome research has been developed applied and reviewed.[8] Most instruments support an approach termed bottom-up proteomics that involves the transformation of proteins extracted from a biological source into peptides by the use of proteases. For technical reasons, trypsin is the most frequently used protease. The thus generated peptides are then subjected to two stages of analysis in a tandem mass spectrometer (MS and MS/MS measurement). In this process, the mass to charge ratio (m/z) of the molecular ions of a specific peptide is first determined and secondly, specific

molecular ions are isolated that are subjected to fragmentation along the peptide backbone. In the most frequent implementation of bottom-up proteomics, known as shotgun proteomics or data-dependent analysis (DDA), the mass spectrometer sequentially selects precursor ions (molecular ions of a specific peptide) for fragmentation from all the precursors detected by the instrument at a particular time point. The masses and intensities of the resulting fragment ions are then recorded, generating a fragment ion spectrum of a peptide. Importantly, the fragment ion spectrum of a peptide reflects its amino acid sequence and can be considered a unique 'fingerprint' of that specific peptide. In a subsequent computational operation, referred to as protein inference, the identified peptides are used to determine a set of proteins identified in the measurement. MS instrumentation and software tools for the analysis of fragment ion spectra progressed rapidly, so that now thousands of proteins can be routinely identified from a biological sample. With the introduction of different quantification strategies, the identified proteins can also be quantified, either in absolute terms (*i.e.* how many copies of a protein are present in a sample) or in relative terms (*i.e.* how does the abundance of a protein differ between two or more samples). In general, these techniques confidently answer the questions which proteins exist in a sample and how much of each identified protein is present.

### 2.3 *Specific Data Requirements to Support Predictive Models and the Targeted MS Methods that Match them*

Predictive models generally require input data that contain quantitative measurements of sets of proteins that describe the system under investigation in different states. Such data are best portrayed as a data matrix in which one axis represents the number of replicate measurements (*i.e.* conditions) and the other axis represents the proteins (*i.e.* features) that have been quantified in each replicate. The quality of the predictions made from such data matrices critically depends on the quantitative accuracy and the reproducibility of the measurements. High reproducibility produces data matrices with a minimal number of missing values across repeat analyses of the same sample and can help distinguish values that are missing from the matrix for technical or biological reasons.

For technical reasons that are quite well understood, it has remained challenging to generate complete and quantitatively accurate data matrices by the widely used DDA method.[9] We therefore developed and applied targeted mass spectrometry methods to consistently quantify sets of proteins across repeat analyses. The prototypical targeted MS method is called selected reaction monitoring (SRM). In SRM, ions in a window centered around the precise mass-to-charge (m/z) ratio of a targeted peptide are recursively selected for fragmentation over the chromatographic elution time of the peptide and the mass spectrometer is programmed to selectively detect specific fragment ions that are, in combination, specific for the targeted peptide. The net result of this method is a group of fragment ion chromatograms that collectively test the hypothesis that the targeted peptide is absent from the sample. Rejection of this hypothesis indicates the presence of the targeted peptide in the sample. SRM is a robust, highly reproducible and quantitatively precise method with a dynamic range of about 5 orders of magnitude. To increase the utility of this method for proteomics, we generated extensive spectral libraries that serve as prior information defining the instrument settings required to acquire the fragment ion chromatograms. These include a library containing assays for every yeast protein[10] and a library that covers more than 99% of the human proteome,[11] thus making essentially all human proteins confidently and reliably measurable by mass spectrometry. In spite of the favorable performance profile, the use of SRM in proteomics is limited by the relatively low number of peptides – in the range of tens to few hundreds – that can be quantified in a single injection. SRM is therefore well suited to support predictive models of relatively confined processes like metabolic pathways[12] (see also section 3.2 below), but is impractical to support modeling of more complex systems.

To overcome this limitation, we developed SWATH-MS, a massively parallel targeted mass spectrometry technique that extends the scope of targeted measurements from hundreds of peptides *via* SRM to thousands of peptides in a single injection.[13] SWATH-MS acquires fragment ions of all peptide ions present in a user-defined (chromatographic retention time vs. precursor m/z) window of an LC-MS/MS measurement and is coupled to a targeted data analysis strategy detecting and quantifying specific peptides in a sample. The method is schematically illustrated in Fig. 2. The SWATH-MS method was initially implemented on Qq-TOF instruments of the instrument manufacturer Sciex. The method has experienced rapid uptake and is now also implemented under the more generic term data-independent analysis (DIA) on instruments of a range of suppliers and is rapidly becoming the method of choice for generating large proteomic data sets, in particular the type of large data matrices described above.

By concurrently fragmenting multiple types of precursor ions selected in extended isolation windows (typically 10–25 Da window width), data acquisition using SWATH/DIA deviates from the principle inherent in DDA mass spectrometry that a specific precursor ion is isolated prior to fragmentation and that therefore all the observed fragment ions in the spectrum can be associated with the sequenced peptide. SWATH/DIA generates convoluted fragment ion spectra where the signals are derived from multiple precursors that are concurrently selected. Consequently, these complex fragment ion spectra need to be deconvoluted to support the assignment of fragment ions to a specific peptide. We solved this challenging problem by developing the software tool OpenSWATH that effectively applies a targeted strategy that is conceptually similar to the data analysis of SRM data.[14] In effect, the tool uses prior information in the form of a spectral library to extract peak groups from the SWATH/DIA dataset that are then statistically scored for their ability to indicate the presence of the targeted peptide (Fig. 2). The signal intensity of the peak group additionally indicates the abundance of the peptide in the sample. Each sample is generally acquired in a single LC-MS/MS run and the resulting data file is stored in a computer and constitutes a permanent record at the level of fragment ion spectra that can be perpetually re-searched. At present, the SWATH/DIA method has been benchmarked in terms of its reproducibility across laboratories,[15] the consistency and accuracy of data analysis,[16] and extensive spectral library resources supporting the targeted data analysis strategy have been published and are publicly accessible.[17] Overall, the technique has reached an impressive performance profile. In a cross-lab benchmarking study 11 groups worldwide identified and quantified close to 5000 proteins with a high level of consistency and quantitative precision from 1 microgram of total peptide mass injected.[15] With recent advances in mass spectrometry, in the range of 7000 proteins could recently be reproducibly identified and quantified at a coefficient of variation of 10–15% from as little as 250 nanogram of total peptide mass injected, and even from 4 ng peptide mass, which corresponds to an estimated 10–15 HeLa cells, in excess of 1000 proteins were detectable.[18] Overall, this favorable performance profile, particularly the high level of pattern reproducibility, quantitative precision, ease of use, and relatively high sample throughput make SWATH/DIA a powerful method for the generation of data in support of predictive models.

## 3. Methods to Generate Predictive Models

### 3.1 *General Considerations on Predictive Models and their Use in Biology*

Living cells or organisms are immensely complex systems in which thousands of biochemical reactions occur in parallel at every

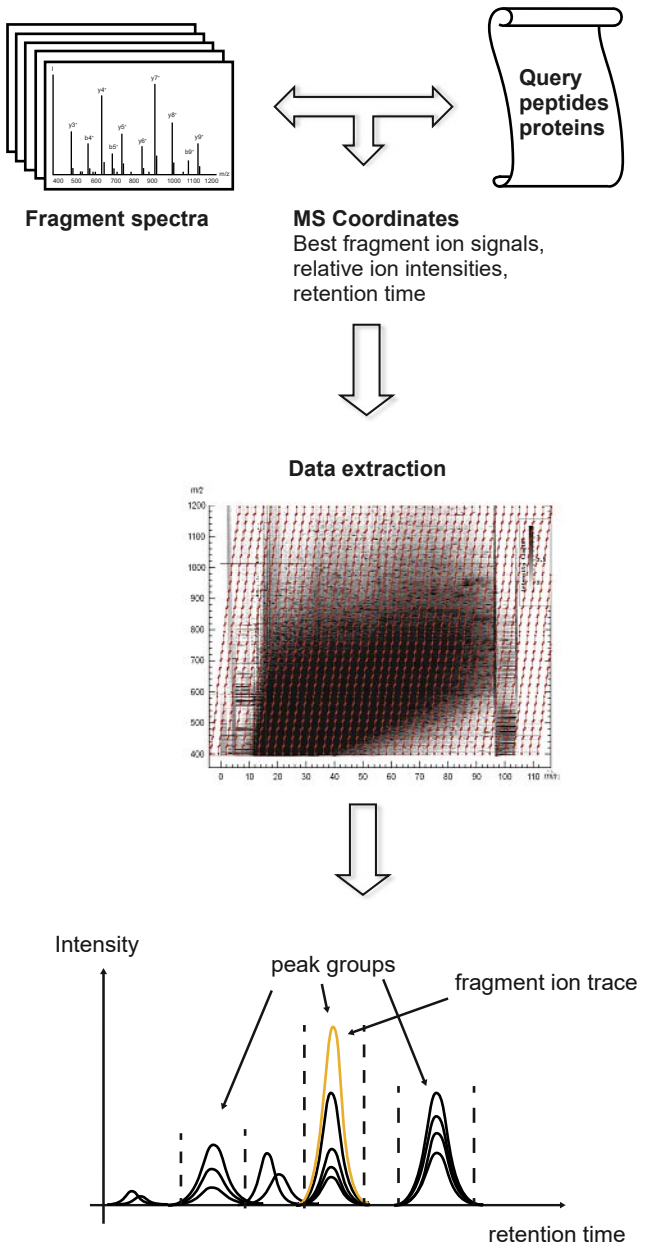


Fig. 2. The concept of SWATH-MS/DIA data generation and extraction. Data acquisition consists of recursively selecting the precursor ions in user-defined precursor selection windows, fragmentation of the selected ions and recording of the resulting convoluted fragment ions. The windows are arranged as adjacent segments so that in the overall process all the ions in a user-defined (chromatographic retention time and m/z) window are repeatedly fragmented (middle section). The SWATH-MS/DIA data analysis method consists of using query peptide information present in a spectral library to extract the fragment ion traces from the acquired convoluted fragment ion map. This query peptide information consists of the mass coordinates of the best ion signals, their relative ion intensity, and the retention time information. Several fragments from the same peptide co-elute and form a peak group that confidently identifies the targeted peptide. The intensity of the peak group indicates the peptide quantity. Statistical models accurately compute the probability that a targeted peptide has been identified in the sample.

point in time. With the emerging ability to identify and quantify the molecules that make up a cell (genes, proteins, metabolites, lipids), the next focus is to understand how a coordinated interplay of these molecules results in a specific phenotype. Specific subsystems for which extensive prior information is available have been successfully modeled and, depending on the type of data and available prior knowledge, different types of modeling approaches have been employed. These models differ in the level of detail and accuracy they explain the biological process under study. In the most detailed case, processes are explained by actual physical constants that can be estimated from solving ordinary differential equations describing *e.g.* biochemical reactions or interactions between proteins. However, such models are typically only fully defined for processes with a few interacting proteins, exemplified by the flagellar motor of bacteria.[2] A type of model often used in metabolism research or in other areas with relatively high levels of prior knowledge are constrained-based models, where specific constraints *e.g.* the stoichiometry of metabolic reactions are used to optimize a model towards a defined goal such as maximal growth. Graph-based models can be used to solve topological problems, such as the most likely path by which a signal is transmitted across a network of signaling pathways. In this case, and in general in systems biology, networks are used as a representation of the complex cellular processes or interactions. These networks, consisting of nodes connected by edges, can represent many different aspects of cellular functionalities from protein–protein interaction, signaling pathways, or how proteins affect each other (*i.e.* functional interaction) exemplified by the phosphorylation of a protein by a kinase. Assessing correlation by regression between different quantitative values is a popular method in biomedical research to reveal association between genes and any phenotypic trait (Genome-wide association studies (GWAS) and quantitative trait loci (QTL)). Finally, similarity between different genetic or proteomic profiles (*i.e.* hierarchical clustering) is often used to characterize the underlying structure of biological networks or processes and patient or disease subgroups. In the following, we show three different examples of different modeling approaches that were used to infer relationships or new biological knowledge from proteomic data acquired by the mass spectrometric methods described in section 2.

### 3.2 *Correlative Analysis of Molecular Data Predict the Functional Effect of Protein Phosphorylation*

Protein phosphorylation is a crucial post-translational modification and is one of the most important regulatory signals in cells. Phosphorylation is known to play an important role in growth factor signaling, immune activation, and regulation of homeostatic processes. Prior to the advent of mass spectrometry-based phosphoproteomics, it was very challenging to detect and quantify phosphorylation sites on a larger scale. This dramatically changed with the development of mass spectrometric methods which now support the identification of thousands of phosphorylation sites from complex samples. However, the ability to identify phosphorylation sites has far outstripped the ability to identify their functional significance. We know only for a small minority of the sites whether and how they regulate a cellular process.

In the model organism *S. cerevisiae*, 204 enzymes catalyze the central carbon and amino-acid metabolism by facilitating 168 biochemical reactions in cells. To predict the functional significance of phosphorylation sites in the metabolic system, we correlated protein abundance, phosphorylation stoichiometry, and metabolic flux through the enzyme in cells across different metabolic states. Phosphopeptide abundance data was acquired using DDA,[12b] protein abundance data by SRM,[12b] and the metabolic fluxes were estimated using constrained-based analysis.[12a] Based on the profiles, specific phosphoproteins were categorized into different cases, depending whether the phosphoprotein and protein levels changed in concordance or if only one type of biomolecules changed. For 11 out of 35 enzymes tested, sufficient data was available to correlate the level of phosphorylation and previously estimated metabolic fluxes. Out of these, the level of phosphorylation in five enzymes (Pda1, Fba1, Gpd1, Gpd2, and Pfk2) correlated with the estimated flux of the biochemical reaction they catalyze (Fig. 3), suggesting that the respective phosphorylation sites regulated this biochemical

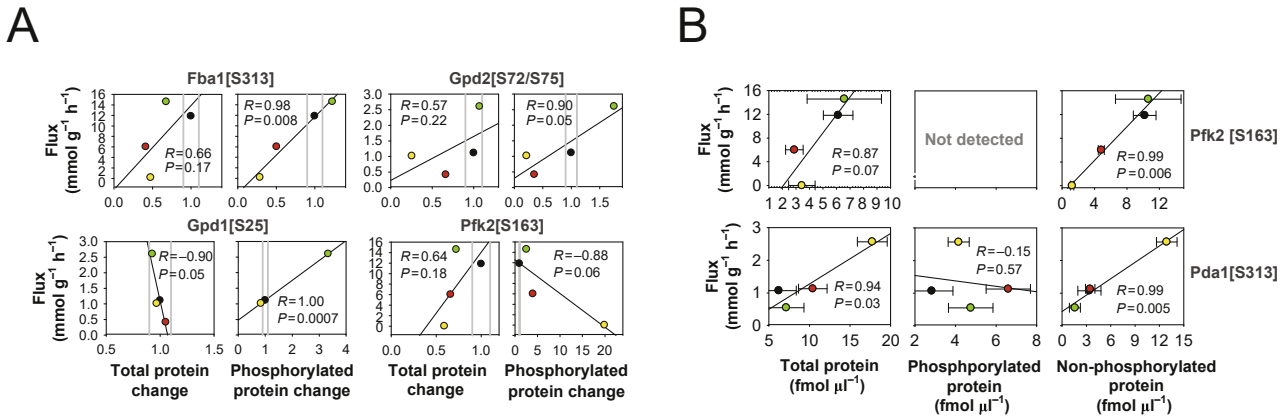


Fig. 3. Correlation analysis between total protein change, phosphorylated protein change, non-phosphorylated protein change, and metabolic flux. Significant correlation between phosphorylation of the yeast proteins Pda1, Fba1, Gpd1, Gpd2, and Pfk2 and the corresponding metabolic fluxes in different metabolic growth conditions were observed. The negative correlation for Pfk2 and Pda1 between the metabolic flux and the phosphorylated protein levels, and the positive correlation of the non-phosphorylated protein, suggested that the phosphorylated protein inhibits the function of this enzyme. Phosphorylated protein levels were acquired with DDA (A) and SRM (B), total protein levels and non-phosphorylated levels with SRM. Figure reproduced from ref. [12b].

process. In addition, targeted proteomics (SRM) was used to estimate the absolute abundance of the total, phosphorylated, and non-phosphorylated protein for Pda1 and Pfk2 from the same cell extract using heavy isotope labelled reference peptides. The good correlation between the non-phosphorylated protein levels and the estimated flux suggested that the phosphorylation inhibited the activity of these enzymes.

In conclusion, this study showed how careful measurements of protein and phosphoprotein levels by mass spectrometry, and correlation with metabolic flux data can make a prediction which phosphorylation sites are functionally important for cellular processes, thus suggesting their functional effect.

### 3.3 *Logic-based Modeling of Cell Specific Regulation of Cholesterol Homeostasis*

To make predictions in a more complex biological system, cholesterol homeostasis in human cells, we used proteomic data and logic modeling to understand the processes underlying variability of cellular drug response.[19] Whereas it has been well known that different individuals and model cell lines significantly differ in their response to drug treatment or other external perturbations, it is typically not clear which processes determine this variability, even though this knowledge is critical for translational research. To address this question, we perturbed the system that maintains cholesterol levels in human cells in four different model cell lines with the same panel of drugs (*e.g.* atorvastatin) or siRNAs and quantified the abundance of more than 3000 proteins after each of the 23 perturbations. As the goal was to quantify as many proteins as possible across the more than 280 samples of the study, the SWATH-MS/DIA approach was selected for this project. For 12 drug perturbation conditions across the cell lines (159 samples), we quantified the drug and metabolite levels in addition to the protein levels. To understand which biological processes were variable between the cell lines based on this vast quantitative data, we opted for a mathematical modeling approach that generated cell-line-specific models of the core cholesterol regulatory mechanisms. This approach, called CellNOpt, was developed by the group of Julio Saez-Rodriguez and produces mathematical models from experimental data and prior knowledge (first principles).[20] These models could then be compared to understand how the tested cells differ in the biochemical processes that regulate cholesterol levels. These models consist of nodes, representing the proteins and metabolites, and edges that represent the functional interactions between these entities. The edges are defined by three parameters: Two of these parameters define a hill-type function that describes in which manner the source node affects a target node (Fig. 4A). The third parameter is a parameter that determines how fast the source nodes affects the target node. Furthermore, we described in our study some of the metabolite interactions using mass action kinetics. By using experimental data for 15 proteins and 16 metabolites, we trained 108 parameters describing the core network of cholesterol regulation consisting of 24 protein nodes, 12 metabolite modes, 6 input nodes and 59 interactions. For each cell line, 100 models were trained based on the bootstrapped data. Hence, we obtained a distribution of 100 values for each parameter that described the bootstrapped experimental data in the best possible way (as defined by the minimal difference between levels predicted by the model after all 23 perturbations and the experimentally measured values in these conditions) (Fig. 4).

This study showed that a number of cellular processes were involved in the regulation of cholesterol levels in the tested cells and that the variability in drug response could not be reduced to a few factors (*e.g.* difference in drug uptake). By measuring both the intracellular drug concentration, we showed that sometimes cell lines that experience a lower internal drug concentration resulted in a higher phenotypic effect, suggesting that pharmacodynamic factors downstream of the drug target dominated. Furthermore, we observed a highly variable effect of transcription factors on the expression of their target proteins. That the variability was not due to the variable activation of the transcription factor could be deduced because the targets regulated by the same transcription factor did not vary in a coherent manner, suggesting that the variability was introduced downstream of the transcription factor.

In summary, this study showed how an extensive matrix of proteomic data acquired from differentially perturbed cells by SWATH/DIA, in conjunction with prior knowledge of a complex regulated cellular process could be effectively integrated into a mathematical representation of a biological cellular process. In this study, we only integrated measurements of 31 molecules, representing less than 1% of the measured proteins. Consequently, a large potential still exists to incorporate further data into a more extensive model. This is presently limited by the availability of prior biological knowledge on how exactly these protein levels are connected to the perturbations or other nodes.

### 3.4 *Statistical Modeling of the Effects of Genomic Variability on Biochemical Pathways and Phenotype*

The question how genomic variability affects phenotypes is of fundamental importance in basic biology and clinical research. The significance of this question has been further

increased by the powerful genomic technologies that now provide accurate complete genomic sequences from whole populations. Frequently, genomic variability discovered by such analyses in two or more groups, *e.g.* clinical cases and controls, are then statistically associated to identify variants that strongly associate with the observed phenotype. Whereas this type of relationship is informative, it does not indicate biochemical causality.

In an attempt to link statistical associations between genotypic variability and a numerical phenotype to the underlying biochemical mechanisms, we used SWATH/DIA proteome measurements in a Drosophila genetic reference panel (DGRP).[21] Specifically, we selected larvae and adult flies from inbred lines of the DGRP that represent the genomic variability of an outbred population. We selected 30 strains with extreme wing size phenotypes of which 15 strains had large wings and 15 strains had small wings. The wing size was determined by morphometric measurements and represents a genetically determined numerical phenotype. We then isolated wing imaginal discs which contain larval state cells that are predetermined to form the adult wings from male and female larvae of the same DGRP strains. Extracted proteins from these imaginal discs were subjected to SWATH/DIA analysis in duplicate. Overall, the study generated a data matrix consisting of 120 quantitative proteome analyses in which 6755 unique peptides representing

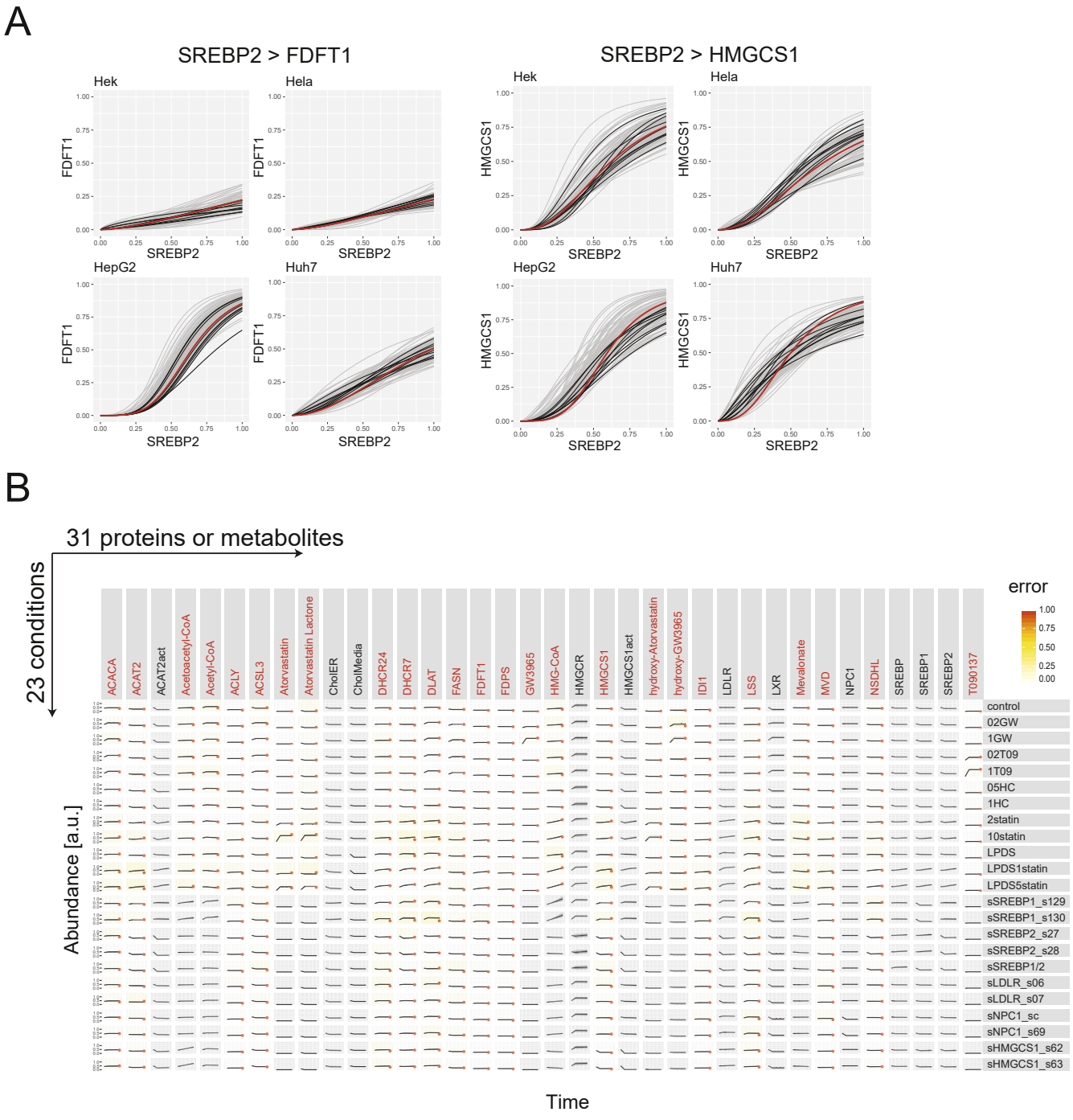


Fig. 4. A) Estimated relationships between the source and target node (*e.g.* SREBP and FDFT1 and HMGCS1 respectively) are shown as curves defined by two trained parameters. These parameters have been trained 100 times for each cell line and the parameters from the best overall solution (red), 10% best solutions (black) are shown. B) Based on the estimated parameters for all edges, this is the prediction of the overall model on how the different proteins and metabolites (x-axis) changed across the 23 different conditions (y-axis) of one cell line. The red dot at the end represents the experimentally validated abundance value for the experimentally measurable nodes (red). Figure reproduced with permission from ref. [19].

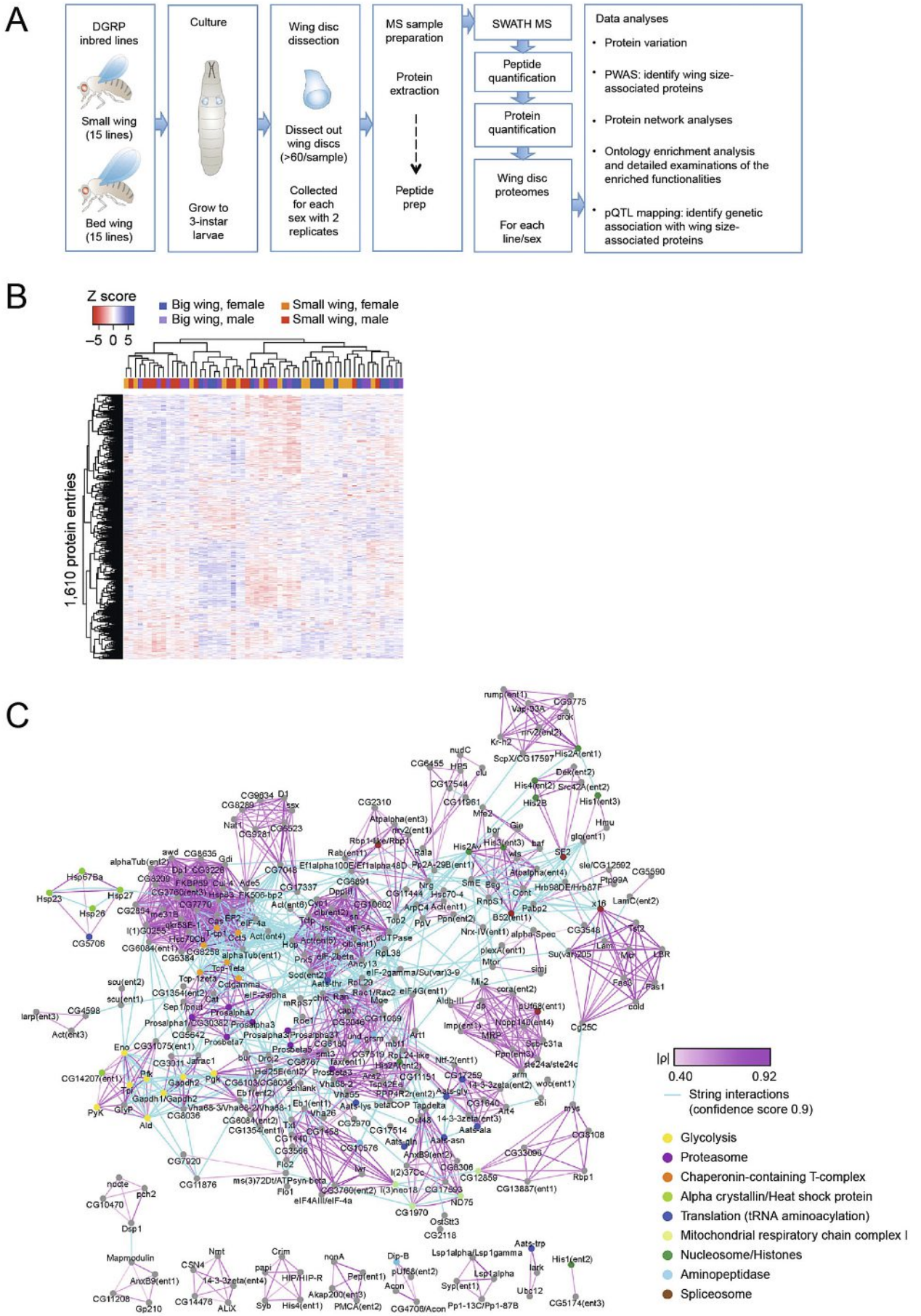


Fig. 5. A) Experimental design. Wing disks from large wing and small wing *Drosophila melanogaster* strains were isolated, proteins extracted and quantified by SWATH/DIA mass spectrometry. B) Shows the data matrix generated consisting of 1610 proteins quantified across 120 proteome measurements. Colors on top indicate the respective fly phenotypes. C) Results. Computed network of proteins for which the abundance was regulated by specific genetic loci. The thus identified proteins formed a network that was strongly enriched in few biochemical functions as indicated in the text that are strongly deterministic of wing size. Figure reproduced with permission from ref. [22].

1610 proteins were reproducibly identified and comparatively quantified across replicates.

We used this protein abundance matrix to carry out a proteome-wide association study (PWAS) in which protein abundance was associated with wing size, specifically the centroid wing size. This analysis, the first PWAS study on a complex trait, identified 34 and 304 protein entities as associated with relative (*i.e.* normalized to the body size) and absolute wing size. Overall, the data indicate that ~20% of the quantified proteins were associated with wing size and about one half correlated positively and the other half negatively.[22]

Further progressing towards investigating the biochemical mechanisms that determine wing size, we identified by hierarchical clustering protein modules consisting of proteins associated with wing size and integrated these modules with the STRING network to generate a molecular network consisting of 303 nodes connected with 1,560 edges that was highly enriched for specific biochemical functions (Fig 5C). Overall, the results indicated that processes including RNA splicing, chromatin assembly, protein folding and translation, and cytoskeletal organization correlated with the body size in general. Furthermore, the data indicated that glucose metabolism exhibits a relatively specific correlation with wing size whereas oxidative phosphorylation was negatively correlated with wing size, suggesting that relatively small shifts between respiration and glycolysis were a main factor determining wing size. Overall, this study demonstrated that statistical association between genetic variability and protein abundance identified a high number of proteins for which the abundance was determined by a specific genetic locus (pQTL, protein quantitative trait locus). Notably, these proteins collectively formed a network that predicted specific biochemical processes that determine wing size in flies, thus making a link between genetic variability, biochemical processes affected by the genetic variability, and the phenotype they determine.

## 4. Outlook/Discussion

We have outlined above three different studies from our lab where state-of-the-art proteomic data was used to increase our understanding of complex biological processes. Current mass spectrometers and new methods such as SWATH/DIA have greatly increased the ability to produce highly accurate quantitative data on protein abundance across hundreds of biological samples. Hence, the challenge lies in using this data to deduce new biological knowledge (*i.e.* theory/first principles) in order to advance our predictive capabilities (Fig. 1). This challenge is not unique to proteomics research but exists as well in other omics fields. As an example, the DREAM Challenges (*www.dreamchallenges.org*) are a collaborative, non-profit, and open science effort that have posted several similar challenges. Examples are the network inference of signaling networks or drug sensitivity prediction.[23] These studies have shown that the incorporation of prior knowledge results in better performing models, thus suggesting that increasing our knowledge on the first principles of biological systems would be helpful. Notably, these studies also showed that there is not yet an effective consensus method to infer molecular networks from large-scale data. Hence, further advances in this area are required. All of these efforts are of course only possible if accurate, meaningful, and relevant data is available. With mass spectrometry-based proteomics it is now possible to generate such data sets on protein abundance, phosphorylated protein residues, or interaction of proteins in cells or clinical samples opening up new approaches for attempting predictions about the behavior of biological systems in basic biology and clinical research.

***Acknowledgements***

We want to thank Prof. Petros Koumoutsakos for discussion and the idea for Figure 1. Work was supported by the Swiss National Science Foundation (SNSF, grant number: 31003A_166435) and the European Research Council (ERC, grant number 670821 PROTEOMICS4D)

Received: June 30, 2019

[1] M. L. Nelson, M. Y. Ismail, in 'Comprehensive Medicinal Chemistry II', Eds. J. B. Taylor, D. J. Triggle, Elsevier, Oxford, **2007**, p. 597, DOI: 10.1016/B0-08-045044-X/00221-2.
[2] H. C. Berg, *Annu. Rev. Biochem.* **2003**, *72*, 19, DOI: 10.1146/annurev.biochem.72.121801.161737.
[3] J. Yeung, F. Naef, *Trends Genet.* **2018**, *34*, 915, DOI: 10.1016/j.tig.2018.09.005.
[4] B. Stern, P. Nurse, *Trends Genet.* **1996**, *12*, 345.
[5] J. J. Thomson, *Proc. Roy. Soc. Lond. a-Conta.* **1913**, *89*, 1, DOI: 10.1098/rspa.1913.0057.
[6] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **1989**, *246*, 64, DOI: 10.1126/science.2675315.
[7] M. Karas, F. Hillenkamp, *Anal. Chem.* **1988**, *60*, 2299.
[8] R. Aebersold, M. Mann, *Nature* **2016**, *537*, 347, DOI: 10.1038/nature19949.
[9] B. Domon, R. Aebersold, *Science* **2006**, *312*, 212, DOI: 10.1126/science.1124619.
[10] a) P. Picotti, H. Lam, D. Campbell, E. W. Deutsch, H. Mirzaei, J. Ranish, B. Domon, R. Aebersold, *Nature Meth.* **2008**, *5*, 913, DOI: 10.1038/nmeth1108-913; b) P. Picotti, M. Clement-Ziza, H. Lam, D. S. Campbell, A. Schmidt, E. W. Deutsch, H. Rost, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J. J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R. L. Moritz, A. Beyer, R. Aebersold, *Nature* **2013**, *494*, 266, DOI: 10.1038/nature11835.
[11] U. Kusebauch, D. S. Campbell, E. W. Deutsch, C. S. Chu, D. A. Spicer, M. Y. Brusniak, J. Slagel, Z. Sun, J. Stevens, B. Grimes, D. Shteynberg, M. R. Hoopmann, P. Blattmann, A. V. Ratushny, O. Rinner, P. Picotti, C. Carapito, C. Y. Huang, M. Kapousouz, H. Lam, T. Tran, E. Demir, J. D. Aitchison, C. Sander, L. Hood, R. Aebersold, R. L. Moritz, *Cell* **2016**, *166*, 766, DOI: 10.1016/j.cell.2016.06.041.
[12] a) R. Costenoble, P. Picotti, L. Reiter, R. Stallmach, M. Heinemann, U. Sauer, R. Aebersold, *Mol. Sys. Biol.* **2011**, *7*, 464, DOI: 10.1038/msb.2010.122; b) A. P. Oliveira, C. Ludwig, P. Picotti, M. Kogadeeva, R. Aebersold, U. Sauer, *Mol. Sys. Biol.* **2012**, *8*, 623, DOI: 10.1038/msb.2012.55.
[13] L. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol. Cell. Proteomics MCP* **2012**, *11*, DOI: 10.1074/mcp.O111.016717.
[14] H. L. Rost, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinovic, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmstrom, L. Malmstrom, R. Aebersold, *Nature Biotechnol.* **2014**, *32*, 219, DOI: 10.1038/nbt.2841.
[15] B. C. Collins, C. L. Hunter, Y. Liu, B. Schilling, G. Rosenberger, S. L. Bader, D. W. Chan, B. W. Gibson, A. C. Gingras, J. M. Held, M. Hirayama-Kurogi, G. Hou, C. Krisp, B. Larsen, L. Lin, S. Liu, M. P. Molloy, R. L. Moritz, S. Ohtsuki, R. Schlapbach, N. Selevsek, S. N. Thomas, S. C. Tzeng, H. Zhang, R. Aebersold, *Nature Commun.* **2017**, *8*, 291, DOI: 10.1038/s41467-017-00249-5.
[16] P. Navarro, J. Kuharev, L. C. Gillet, O. M. Bernhardt, B. MacLean, H. L. Rost, S. A. Tate, C. C. Tsou, L. Reiter, U. Distler, G. Rosenberger, Y. Perez-Riverol, A. I. Nesvizhskii, R. Aebersold, S. Tenzer, *Nature Biotechnol.* **2016**, *34*, 1130, DOI: 10.1038/nbt.3685.
[17] a) G. Rosenberger, C. C. Koh, T. Guo, H. L. Rost, P. Kouvonen, B. C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O. T. Schubert, P. Faridi, H. A. Ebhardt, M. Matondo, H. Lam, S. L. Bader, D. S. Campbell, E. W. Deutsch, R. L. Moritz, S. Tate, R. Aebersold, *Sci. Data* **2014**, *1*, 140031, DOI: 10.1038/sdata.2014.31; b) P. Blattmann, V. Stutz, G. Lizzo, J. Richard, P. Gut, R. Aebersold, *Sci. Data* **2019**, *6*, 190011, DOI: 10.1038/sdata.2019.11.
[18] S. Amon, F. Meier-Abt, L. C. Gillet, S. Dimitrieva, A. P. Theocharides, M. G. Manz, R. Aebersold, *Mol. Cell. Proteomics MCP* **2019**, DOI: 10.1074/mcp.TIR119.001431.
[19] P. Blattmann, D. Henriques, M. Zimmermann, F. Frommelt, U. Sauer, J. Saez-Rodriguez, R. Aebersold, *Cell Sys.* **2017**, *5*, 604, DOI: 10.1016/j.cels.2017.11.002.
[20] C. Terfve, T. Cokelaer, D. Henriques, A. MacNamara, E. Goncalves, M. K. Morris, M. van Iersel, D. A. Lauffenburger, J. Saez-Rodriguez, *BMC Syst. Biol.* **2012**, *6*, 133, DOI: 10.1186/1752-0509-6-133.
[21] T. F. Mackay, S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. Anholt, M. Barron, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L. L. Pu, C. Qu, M. Ramia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L.

Turlapati, K. C. Worley, Y. Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, R. A. Gibbs, *Nature* **2012**, *482*, 173, DOI: 10.1038/nature10811.

[22] H. Okada, H. A. Ebhardt, S. C. Vonesch, R. Aebersold, E. Hafen, *Nature Commun.* **2016**, *7*, 12649, DOI: 10.1038/ncomms12649.

[23] a) J. C. Costello, L. M. Heiser, E. Georgii, M. Gonen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J. P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, N. D. Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, G. Stolovitzky, *Nature Biotechnol.* **2014**, *32*, 1202, DOI: 10.1038/nbt.2877; b) S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, C. K. Wong, K. Graim, A. Bivol, H. Wang, F. Zhu, B. Afsari, L. V. Danilova, A. V. Favorov, W. S. Lee, D. Taylor, C. W. Hu, B. L. Long, D. P. Noren, A. J. Bisberg, H.-D. Consortium, G. B. Mills, J. W. Gray, M. Kellen, T. Norman, S. Friend, A. A. Qutub, E. J. Fertig, Y. Guan, M. Song, J. M. Stuart, P. T. Spellman, H. Koeppl, G. Stolovitzky, J. Saez-Rodriguez, S. Mukherjee, *Nature Meth.* **2016**, *13*, 310, DOI: 10.1038/nmeth.3773.