

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



Chapter

# Computational Methods for the Study of Peroxisomes in Health and Disease

*Naomi van Wijk and Michal Linial*

## Abstract

Peroxisomal dysfunction has been linked to severe human metabolic disorders but is also linked to human diseases, including obesity, neurodegeneration, age-related diseases, and cancer. As such, peroxisome research has significantly increased in recent years. In parallel, advances in computational methods and data processing analysis may now be used to approach unanswered questions on peroxisome regulation, mechanism, function, and biogenesis in the context of healthy and pathological phenotypes. Here, we intend to provide an overview of advanced computational methods for the analysis of imaging data, protein structure modeling, proteomics, and genomics. We provide a concise background on these approaches, with specific and relevant examples. This chapter may serve as a broad resource for the current status of technological advances, and an introduction to computational methods for peroxisome research.

**Keywords:** high-resolution microscopy, structure prediction, integrative modeling, deep learning, genomics, proteomics, mass spectrometry, Zellweger syndrome

## 1. Preface

Peroxisomes are single membrane-bound organelles found in all eukaryotic cells, with diverse functions according to the cell type and metabolic conditions. The study of peroxisomes, their processes, and regulation activity has taken flight in the last decade, thanks to the introduction of novel cellular methodologies. Simultaneously, sequencing and human genetics methods have strongly improved, and a growing number of metabolic diseases have been associated with genetic variations in peroxisomal genes. However, a deep understanding of the coordinated workings of peroxisomal genes in health and disease is still lacking.

Peroxisomes play a key role in cell metabolism and homeostasis. For example, they are involved in the  $\beta$ -oxidation of fatty acids, the formation of specific ether phospholipids, and in the dissipation of damage caused by reactive oxygen species (ROS). At the same time, peroxisomes are very dynamic organelles that use creative solutions in their biogenesis and assembly, import of large proteins, morphological changes by fusion and fission, and complex interactions with other organelles and lipid droplets. Several insightful review articles summarize the current challenges of

the metabolism and biology of peroxisomes [1, 2]. This chapter explores the use of state-of-the-art computational approaches and methods for the study of peroxisome biology at various levels. First, we discuss the basics of deep learning and machine learning algorithms. Then, we explore how deep-learning-augmented cell imaging explores peroxisomal biology. Advances in microscopy have yielded a trove of high-resolution, high-throughput imaging data from living cells that require and enable a more advanced level of analysis. We then zoom in to a higher-level resolution, focusing on structural analysis and integrative modeling of peroxisomal proteins and their protein complexes. Next, we explore high throughput methods to study the interactions between proteins and lipids at the level of the proteome and lipidome. Finally, we discuss genetic approaches considering peroxisomal dysfunction and pathology. We highlight the role of computational and bioinformatics-based approaches to major open questions in the field of peroxisome biology.

## **2. Analysis of cell biology and imaging data**

Observation remains one of the central pillars of biological research and cell biology in particular. Throughout the second half of the 20th century, electron microscopy (EM) images were fundamental to unveil the intracellular structures and organelles of a static cell at nanoscale resolution. The field of cell biology was revived by merging molecular biology techniques with *in vivo* dyes in the form of reporters and biosensors (e.g., GFP and its derivatives). The use of such fluorescence probes allowed monitoring cell dynamics under varying conditions while remaining as close-to-native state.

Current microscopy technology enables imaging at exceptional resolution, in the xy-plane varying from about 200 nm in confocal microscopy to about 10 nm in single-molecule localization microscopy [3]. Since peroxisome size ranges from 100 to 1000 nm in size, high-resolution microscopy is required for accurate and detailed imaging of protein expression and cellular localization. With the advances in high-resolution microscopy linked to automated, robotic lab support, the experimental results produce unprecedented big data of images and videos of cell dynamics. Although each technique has its advantages and limitations, the level of detail they reveal requires advanced data processing, analysis, and storage solutions. In past years, advanced algorithms, and especially deep-learning algorithms, developed for application to the biological domain, have skyrocketed.

### **2.1 Basics of deep learning**

Supervised deep-learning algorithms require an annotated dataset, training on that dataset, and the use of the trained model on unseen, new data. In imaging applications, a dataset can be augmented and diversified by providing the model with edited (e.g., zoomed or rotated) images. The model is then trained on the dataset. The classification task aims to construct a function that takes this array as input to predict a label. In neural networks, which are usually used for classification problems, the learning task then aims to minimize the “loss function” (i.e., error), by optimizing a set of parameters, or weights, that multiply input data to obtain the output data that is passed on to the next layer in the neural network. A common loss function is cross-entropy to measure the difference between a true label and the predicted one. The algorithm architecture should be chosen to minimize overfitting, that is, when the

model performs better on the training data than on a validation dataset. In the case of underfitting, the model performs poorer on the training set, resulting in suboptimal performance. For an elaborated review of the use of deep learning for cellular image analysis see ref. [4] and references within.

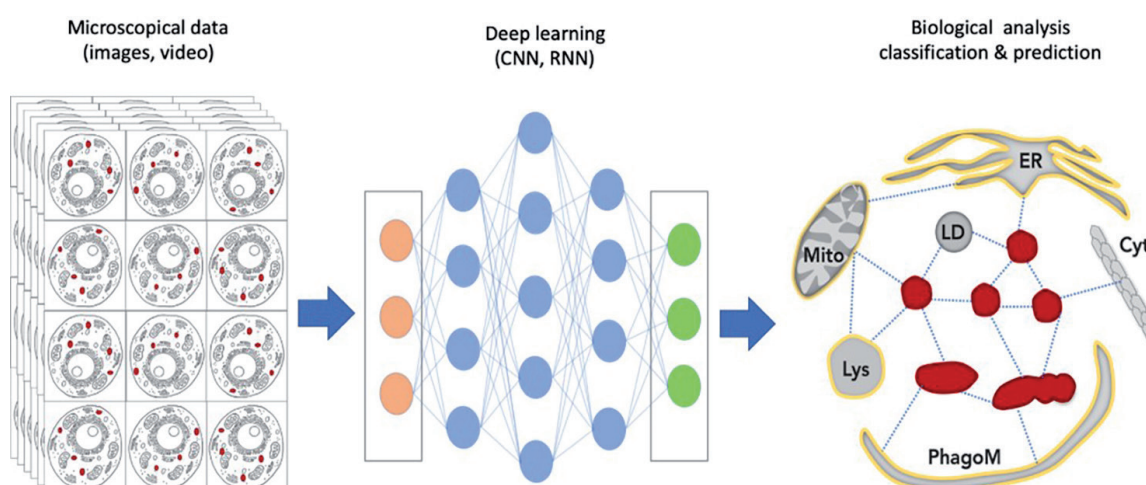
## 2.2 Advanced image analysis

The field of cellular and molecular biology had benefited from the availability of analytical tools for classical microscopy images. Major tools include CellProfiler, Microscopy Image Browser, OMERO, Fiji, and others [4]. Applying cell imaging tools according to unified standards led to the success of large-scale resources, such as the Human Proteome Atlas (HPA). HPA compiled enormous amounts of microscopical confocal images for annotating the subcellular information by using specific antibodies for most human coding genes across several human cell lines [5].

Recent years have seen the publication of updates of these tools with added components of deep-learning algorithms. Important applications of these tools include image classification, image segmentation, object tracking, and augmented microscopy. Specific examples of image analysis for the study of organelles are briefly discussed. **Figure 1** shows a scheme for addressing open questions in peroxisome biology using a large set of raw microscopic images (static and dynamic) and state-of-the-art methods of deep learning for the task of deciphering cross-organelle interactions [6, 7].

### 2.2.1 Image classification

In image classification, labels can be added to images, for example, to identify if a fluorophore-labeled protein is localized to a specific organelle or resides in the cytoplasm. For this purpose, machine-learning-based image classification has long focused on the generation of classifiers to identify changes in morphology following exposure to compounds or growth conditions or to identify changes in cell state.



**Figure 1.** A scheme for addressing open questions in peroxisome biology using a large set of raw microscopic images (static and dynamic). The image data is transformed by a convolutional neural network (CNN) whose output provides insights on large-scale cell biological challenges such as the task of organelles interactions, shape, and their dynamic crosstalk. Mito, mitochondrion; LD, lipid droplet; PhagoM, phagophore membrane; Lys, lysozyme, ER, endoplasmic reticulum; Cyt, polymerized cytoskeleton fiber (e.g., actin). Peroxisomes are colored red and membranes are colored yellow.

A popular software package called ilastik provides an addition to classical cell imaging tools with workflows for image segmentation, object classification, counting, and tracking. The pixel classification workflow produces semantic segmentation of images, attaching a user-defined class label to each pixel of the image. This step also forms the first step for object classification with morphological object features or may be used as initial input for the carving workflow. For example, ilastik was used for the reconstruction of 3D data from focused ion beam scanning EM (FIB-SEM) to segment the ER, using a pseudo-automated approach [8].

In a study by Li et al. [9], deep learning was used to classify organelle morphology of chloroplasts, mitochondria, and peroxisomes in the plant model *Arabidopsis*. The authors described a deep-learning framework, DeepLearnMOR (Deep Learning of the Morphology of Organelles), that identifies organelle morphology abnormalities at 97% accuracy. In the study, a dataset of 47,000 confocal fluorescence microscopy images from *Arabidopsis* wild-type and mutant plants with abnormal division in chloroplasts, mitochondria, or peroxisomes, was used to train the model. The dataset was augmented by using rotated, flipped, and split images. The model is based on both transfer learning and convoluted neural networks and significantly outperformed conventional machine-learning methods. In deep learning, transfer learning entails training a model on a large dataset and then fine-tuning the model for a different task using a new, smaller dataset. In this framework, the model distinguished well between mitochondria and peroxisomes, despite the overlap in their sizes. The framework can be used to study subtle morphological changes to classify intact and aberrant human peroxisome morphology.

In another example, a multi-scale convolutional neural network approach was developed and trained on eight publicly available cellular imaging datasets [10]. Following training, both the binary phenotype classification task as well as a multi-label classification task, performed at least as good as state-of-the-art architectures, saving time on the manual adjustment of parameters for segmentation and feature selection, that is needed for conventional image analysis pipelines. The datasets included images of stains for various organelles and cell types. Although peroxisomes were not included in the datasets used, this approach can be used for expanding the classification to cover peroxisome phenotypes.

An additional classic classification question regards the observation that proteins are localized into multiple subcellular compartments. About half of human proteins exist in more than one organelle simultaneously, and these multi-locational proteins are likely to play critical roles in cellular functions [5]. To deal with these multilabel proteins, most existing methods converted the multilabel classification problem into  $L$  binary problems,  $L$  being the number of classes. However legitimate, such simplified approaches ignored label dependencies that actually exist among subcellular locations [11]. The difficulty of the classification part mainly lies in the multilabel nature of protein localization, as is also exemplified in ref. [10]. Even in a simplified setting of a cell line in culture, cells are at different stages of cell division and density, yielding nonuniform localization profiles.

### *2.2.2 Image segmentation*

Image segmentation is the task of identifying multiple objects or features within an image, for example, cell counting. LysoQuant is a deep-learning approach for the detection and segmentation of organelles and is available as an ImageJ plugin. Its efficacy was demonstrated using the ER as a model organelle and a polymerogenic  $\alpha$ 1-antitrypsin Z (ATZ) variant as a model disease-causing aberrant protein [12]. The

model's performance was validated on the quantification of catabolic pathways that maintain cellular homeostasis and proteostasis. The model was tested on two cell types and on the ER as a model organelle, but it may very well be applicable for use in other cell types and for the study of peroxisomes.

Due to the advancement of classification accuracy and the availability of high computational power, cell image segmentation approaches are often based on deep convoluted neural networks. One of the disadvantages of deep-learning approaches is the large amount of training data required to train them. Although software packages, such as CellProfiler, already use deep-learning models, they do not support retraining on new data, thus restricting their application domain to an available set of datasets. In contrast, U-Net is pretrained on a diverse set of data and for every new task needs only a few (<10) annotated images [13].

In addition, sequential images often differ one from the other in the sense that the number of objects belonging to each class differs from image to image, leading to an imbalance in the class weights. One approach to tackle this problem is by automatically updating the weights of the imbalanced classes by constructing a new objective function. In one recent study, a U-Net-like convoluted neural networks model is used with two updated loss functions to improve segmentation of cell organelles, including cytoplasm, plastids, nucleus, mitochondrion, and peroxisome [14]. This demonstrates the importance of adapting the loss function in deep-learning approaches for improving the success of segmentation tasks.

### *2.2.3 Object tracking*

The challenges in imaging described so far assume cells to be rather static objects. Obviously, the dynamics and heterogeneity of cells define their biology. In object tracking, objects are followed through a series of time-lapse images. This requires two tasks—object detection and object linkage. Single-particle tracking (SPT) is often the rate-limiting step in live-cell imaging studies of subcellular dynamics. Many object tracking models are based on a tracking algorithm that addresses the principal challenges of SPT, namely high particle density, particle motion heterogeneity, temporary disappearance of particles, and merging and splitting of particles. The algorithm first links particles between consecutive frames and then links the resulting track segments into complete trajectories [15]. This approach forms the basis for software such as CellProfiler and TrackMate.

TrackMate is a software that offers several detection and tracking modules that allow combining manual and automated particle tracking approaches. An openly available tool, it is available as an extension of ImageJ. Moreover, the capabilities of the software can be tailored by the user through the addition of specific tracking, detection, visualization, or analysis modules. Its data model makes it a useful tool for a wide range of tracking applications, ranging from single-particle tracking of subcellular organelles to cell lineage analysis. Importantly, the TrackMate study stresses the importance of avoiding photoinduced stress due to the continuous or repetitive illumination required for fluorescence microscopy [16].

An interesting example of the use of TrackMate for tracking, along with other image analysis methods, is presented in ref. [17]. While current imaging techniques are constrained by the small number of distinctive fluorescent labels within a single image, the use of confocal and lattice light sheet (LLS) fluorescence microscopy combined with computational sophistication allowed to track globular organelles and propose dynamic inter-organelle contacts. The study describes the frequency and

locality of two- to five-way interactions among major membrane-bound organelles (ER, Golgi, lysosome, peroxisome, mitochondria, and lipid droplet) and shows how these relationships change over time.

#### *2.2.4 Image augmentation*

Augmented microscopy is the extraction of latent information from biological images, such as the identification of the locations of cellular nuclei in bright-field images. Many augmented microscopies approach train neural networks to translate between label-free (bright-field, phase, differential interference contrast, and transmission EM) and labeled (fluorescence) images of the same cells. The ability to predict fluorescence images from grayscale data is advantageous for increased imaging speed and improved time-lapse imaging. Moreover, this neural network methodology was adapted from the classical field of imaging processing and its implementation to organelle biology takes advantage of the overwhelming amounts of grayscale images already produced by standard biology labs [18]. In this study, the prediction performance across organelles and other subcellular structures reach an accuracy of detection between 70 and 90% for cellular compartments, such as nucleoli, nuclear envelope, mitochondria, and ER [18, 19]. Undoubtedly, the model can be trained and tested for the study of peroxisomes as well.

### **3. Protein structure modeling and analysis**

Protein structure modeling and prediction are experiencing exciting times. The “protein folding problem,” which has puzzled scientists for many decades, asks to predict a protein’s structure from its primary amino acid sequence. It is thus not surprising that artificial intelligence (AI)-driven protein structure software, which includes both AlphaFold and RoseTTAFold, were announced as the Science breakthrough of the year 2021. In this chapter, we will discuss these models, followed by a discussion of how protein prediction models may be integrated into classical 3D experimental methods, such as cryo-EM, X-ray crystallography, mass spectrometry (MS), and nuclear magnetic resonance (NMR). We will then discuss approaches for modeling protein–protein interactions (PPI) and protein complexes.

#### **3.1 AlphaFold**

The Critical Assessment of protein Structure Prediction (CASP) was initiated in 1994 as a biennial open competition for advancing methods for 3D protein structure prediction. Until 2016, the average prediction accuracy score across multiple approaches was bound to 30–40 (on a scale of 100). In 2018, the AlphaFold model developed by the DeepMind company (owned by Google) scored ~55, and in 2020, AlphaFold2 reached an astonishing score of ~92. For reference, a score > 90 is roughly equivalent to the variation monitored from repeated experiments for determining protein structure [20].

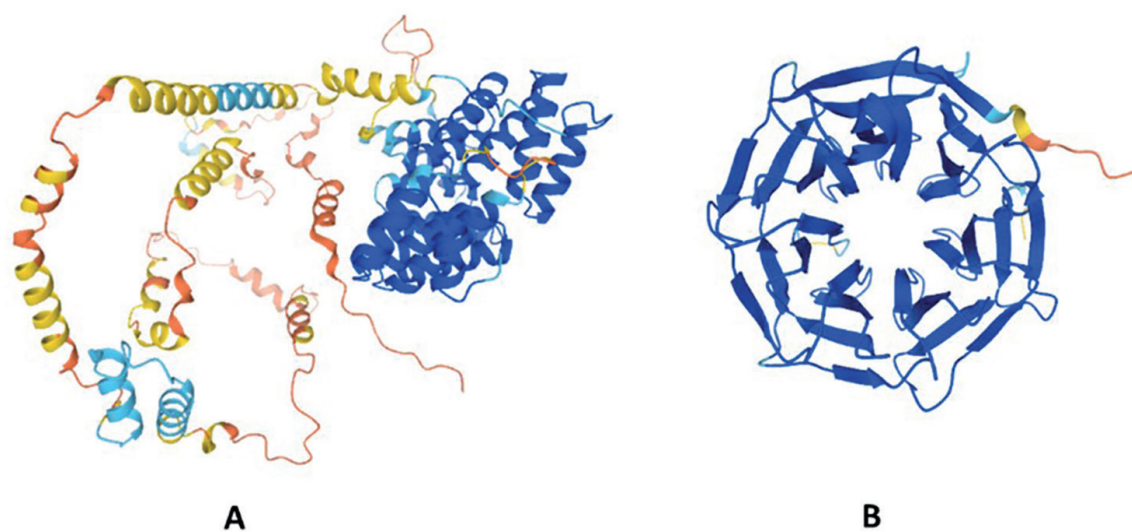
The input of the AlphaFold model is based on the primary amino acid sequence of a protein and the sequence alignment compilation of known homologs. Specifically, the amino acid sequence is used to build a multiple sequence alignment (MSA) from similar sequences found in protein sequence databases. Assuming an accurate MSA, the amino acid pairs that were co-mutated along the evolution path can be detected. A crude

structure representation, or “pair representation” is then produced based on structural templates and on the paired amino acids that are likely in contact with each other. The templates and MSA are then passed through a transformer termed “Evoformer” that takes the MSA representation and the pair representation, to refine these representations. In the final step, the “structure module” takes the refined MSA representation and pair representation to construct a refined 3D structure model of heavy atoms [21].

The predicted structures contain atomic coordinates and per-residue confidence estimates on a scale from 0 to 100, with higher scores corresponding to higher confidence. This per-residue confidence measure, pLDDT, is based on a preexisting metric used in the protein structure prediction field. Scores  $>70$  are considered residues predicted with confidence and  $>90$  are considered as very high confidence prediction (**Figure 2**). It should be noted that in multi-domain predictions, individual domains may be largely accurate while their relative position is not [22, 23].

As of January 2022, AlphaFold database provides free access to  $>360,000$  predicted structures across 21 proteomes. In comparison, the Protein Data Bank (PDB) that was announced 50 years ago contains “only” 180,000 experimentally solved 3D structures, and many of them do not cover the full length of the protein. However, the UniProtKB database includes almost 220 million entries. For many of the PEX proteins (peroxins) involved in peroxisome biogenesis, no experimentally determined structures are available [24]. However, a search for PEX proteins in the current AlphaFold database identified 28 hits from yeast, 26 from Arabidopsis, and 22 from human origin.

For example, the extensively studied peroxisomal targeting signal 1 receptor (PEX5), is represented in the PDB archive (nine entries) that mostly covers the well-folded domain at the C' terminus. The AlphaFold database, however, includes a prediction of the full-length protein, which includes a very high confidence prediction for the C' terminal domain, but also revealed some helices that were previously not solved (**Figure 2A**). It should be noted that the low-confidence regions in AlphaFold predictions largely correspond with known intrinsically disordered regions (IDRs) [25]. For the human peroxisomal targeting signal 2 receptor (PEX7) protein, no experimentally solved structures are available. However, the AlphaFold



**Figure 2.** Prediction of PEX5 (A) and PEX7 (B) structure by AlphaFold2. Model confidence—Dark blue, very high ( $pLDDT > 90$ ); light blue, confident ( $90 > pLDDT > 70$ ); yellow, low ( $70 > pLDDT > 50$ ); orange, very low ( $pLDDT < 50$ ). Note that many of the low-confidence structure predictions coincide with intrinsic disorder in the protein.



predicted protein structure is a strikingly beautiful  $\beta$ -propeller structure with seven-fold symmetry, similar to a published structure of its yeast homolog (**Figure 2B**) [26]. It is now possible to compare structural differences between the yeast and human homologs to shed light on differences in function. We anticipate that many structural and mechanistic questions regarding peroxisomal proteins can now be approached with a wealth of reliable modeled structural data.

### 3.2 Rosetta

The Rosetta software suite was initially developed in the Baker Lab at the University of Washington, Seattle, and was evolved as an active collaborative effort [27]. It includes many functionalities for macromolecular modeling. Its applications and protocols include *ab initio* modeling [28–30], including membrane protein modeling [31–33], comparative modeling using one or more known structures as templates for modeling, “fold-and-dock” application for the prediction of symmetric homooligomer structures, FlexPepDock for *ab initio* or refinement of peptide docking to a receptor, “relax” that improves protein energy landscape modeling and more.

Around the same time that the AlphaFold model was published following its winning entry at the CASP14 competition, RoseTTAFold, another extremely successful protein structure prediction model, was published. Like AlphaFold2, it uses deep neural networks to find sequence patterns in databases of similar sequences. When given a new sequence to model, RoseTTAFold proceeds along multiple tracks—one creates an MSA, another predicts pairwise interactions between amino acids within the protein, and the third constructs the 3D model structure. The program bounces among the tracks to refine the model, using the output of each one to update and refine the others [34, 35].

The trRosetta model is also Rosetta-based and incorporates restraints for prediction. The algorithm starts off with MSA for distance and contact prediction to learn probability distributions over distances between residues and determine residue orientation. The predicted distances and orientations are then used to generate 3D structures using constrained energy minimization. The lowest-energy backbone is then subjected to Rosetta full-atom relaxation to add side chains and make the structures physically plausible, and to generate the lowest-energy full-atom model. The trRosetta network is able to identify the most important residues for determining protein folding and can apply this on *de novo* designed proteins, although the model was only trained on native proteins [36].

### 3.3 Modeling of protein: protein interactions

The prediction of multi-chain protein complexes is an even greater challenge than the prediction of monomeric protein structures. In the context of peroxisome biology, the functional and organelle mysteries reside in the dynamic interaction between the membranal and matrix proteins [2]. In addition, PEX protein’s subcellular localization is governed by post-translation modifications (PTMs), such as ubiquitination. All these aspects are executed by the dynamic of protein–protein interactions (PPI).

Many docking algorithms are being used for *ab initio* or template-based modeling, several of which we will discuss briefly. Some of the first methods for multimolecular modeling, developed in the Wolfson Lab at Tel Aviv University, include CombDock, PatchDock, FlexDock, and more recently, DockStar [37]. Input for the CombDock algorithm is an ordered set of protein sub-structures, which then combinatorically

assembles the inter-contacts that define their overall organization. PatchDock and FlexDock are based on the integration of local shape complementary evidence, where both input molecules are considered rigid, or in which one of the input molecules is considered rigid and the other flexible. DockStar integrates both low-resolution information (e.g., MS) and high-resolution (e.g., X-ray, NMR, and homology modeling) experimental data, combining atomic structures and interaction data. It then calculates the optimal assembly of the individual subunits.

Another approach is used by pyDock, which uses electrostatics, desolvation energy, and in a limited manner, van der Waals forces, to score rigid-body docking poses. InterEvDock and InterEvDock2 use co-evolutionary information in docking based on rigid-body sampling. In InterEvDock2, protein sequences can be provided as input, not only 3D structures. The algorithm then first performs comparative modeling based on template search. If biological input is available such as a pair of residues known to be in contact, restraints with a tunable distance threshold can be specified for use in the docking procedure. A recent review of multi-molecular modeling approaches can be found in ref. [38].

The latest developments in AI-harnessed modeling approaches will likely become some of our most important tools for the modeling of PPI and protein complexes. AlphaFold-Multimer is an AlphaFold model trained specifically for multimeric inputs of known stoichiometry. For example, an A2B2C2 heteromer was solved with a structural prediction score of 98, virtually identical to the solved structure for this complex [39]. In addition, AlphaFold2 was shown to successfully predict peptide-protein complexes even though it was trained only on monomer chains [40]. Open questions regarding peroxisome complexes can now be approached using this updated model. For example, the human PEX2-PEX10-PEX12 proteins form a protein-ubiquitin ligase complex for which currently no structure is available. The same is true for the AAA+ ATPase heterotrimeric complex PEX1-PEX6, and more.

RoseTTAFold has comparable performance in identifying PPI to that of experimental methods, but the combination of applying the RoseTTAFold model with AlphaFold further increases identification accuracy. A combined protein interaction identification pipeline that incorporates a rapidly computable version of RoseTTAFold with the slower but more accurate AlphaFold, evaluates interactions between the 8.3 million possible pairs of yeast proteins. In total, 106 previously un-identified assemblies and 806 that were structurally uncharacterized, were modeled. These models include higher-order complexes up to pentameric assemblies [41]. This combined approach demonstrates the strength of combining various neural network-based models to maximize modeling accuracy and speed.

### **3.4 Integrative modeling and analysis**

The algorithms described above in some cases use experimentally solved structures as templates for the modeling of protein complexes. These are instances of integrative structural modeling, which involves the determination of macromolecular structures by combining experimental and computational modeling approaches [42].

Workflows for experimental methods can be improved with modeling approaches at various steps. In X-ray crystallography, protein modeling can be used to improve the determination of the protein structure, which is hampered by the “phase problem” that prevents the direct determination of the 3D structure from X-ray diffraction data. This approach was used by combining AlphaFold modeling with X-ray

diffraction data to the determination of the structure of the ORF8 protein of SARS-CoV-2 [43].

Cryo-EM is more and more often used for structure determination. The highest resolution structures solved with cryo-EM have been solved at less than 2 Å [44]. This improvement in resolution has made cryo-EM a likely candidate to replace X-ray crystallography as the gold standard of experimental structure determination. A variety of software is available for modeling macromolecular assemblies using cryo-EM for *de novo* modeling, fitting, and validation of the atomic model. For high-resolution structures, data integration can be used for *de novo* model building and local refinement, whereas intermediate or low-resolution structures may benefit from rigid fitting and fully integrative modeling [45]. An impressive example of integrative modeling was the publication of the full structure of the 52 MDa nuclear pore complex (NPC) from yeast. In that study, protein modeling was combined with data obtained from cryo-electron tomography (cryo-ET), a method similar to cryo-EM, chemical cross-linking mass spectrometry, *in vivo* imaging, and existing solved structures of NPC subunits, as the experimental data input [46]. The successful use of *ab initio* modeling is further illustrated by the cryoSPARC software, which makes it possible to very quickly perform unsupervised *ab initio* 3D classification to discover multiple 3D states of a protein without prior structural knowledge [47]. CryoSPARC is an end-to-end solution for cryo-EM analysis and structure refinement by combining 3D reconstruction algorithms with specially designed software.

NMR may benefit from systematic back-calculation of expectation spectra across a conformational space should then allow reconstruction of the experimental spectra. This would enable the comparison of the back-calculated and experimental data and provide us with a quantitative quality measure [48]. In addition, complex NMR results can benefit from the strength of a deep neural network-based approach such as DEEP Picker, to aid in the analysis of NMR spectra and correctly characterize overlapping peaks.

In cross-linking MS (CL-MS), an *in vitro* or *in vivo* cross-linking reaction covalently binds specific residues. The cross-linked proteins are then enzymatically digested, followed by liquid chromatography–tandem mass spectrometry (LC–MS/MS). The covalently linked peptides provide valuable information about spatially close residues, whether intramolecular or intermolecular, which are then used as constraints in structural modeling [49]. XL-MS data can be useful in the validation and refinement of predicted structures as well. The benefit of this integrative modeling approach was recently demonstrated in the structural determination of three proteins from SARS-CoV-2. In this study, docking algorithms PatchDock, CombDock, and AlphaFold2 were used to model the proteins and refine them with CL-MS data that were collected from living cells. It was found that intradomain cross-links were satisfied in most cases, whereas interdomain cross-links were often violated, demonstrating the strength of this integrative modeling approach [50].

Importantly, the integration of diverse data sources into unified predictive models is likely to advance the knowledge of the protein complex of peroxisomes. While the structural study of peroxisomal proteins is in its infancy, we demonstrate the strength of recently developed computational software, tools, and algorithms to integrate data using breakthrough AI-based structural prediction approaches and integrative modeling.

## 4. Proteomics

Although the studies of individual proteins are important, they mostly ignore the PPI, cellular and physical environment of the studied protein. In recent years, proteomics approaches have become extremely valuable for the systematic analysis and discovery of the involvement of and interaction between proteins at the cellular or organellar level. Depending on the goal of the study, proteome analysis may include the determination of the full cellular or organellar proteome or PPI. The abundance profiles of proteins throughout all fractions of the purification can be compared to the profile of known marker protein to identify proteins that co-fractionate with the organelle of interest. To identify interactions, the abundance of interactors to a tagged protein of interest is compared to the abundance of the same, untagged protein. In addition, proteome dynamics can be followed by analyzing the organellar proteome at various time points following a stimulus [51]. Here, we discuss some of the most advanced methods for the analysis and application of proteomics data, with an emphasis on mass spectrometry (MS)-based proteomics. Importantly, MS and other advanced quantitative proteomics highlight regulation that cannot be explored using nucleic acids sequencing approaches. In the context of peroxisomal biology, the information includes the presence of protein variants, PPI, subcellular localization, and the status of post-translational modifications (PTMs). Analyzing protein levels in an organelle is fundamental to exploring molecular signaling and dynamics in response to varying conditions.

### 4.1 Mass spectrometry-based proteomics

For organelle proteomics, three approaches can be taken to provide input for MS analysis, namely data-dependent acquisition, data-independent acquisition, and targeted proteomics, each of which is briefly discussed below. Following separation of the organelle, the organelle fraction can be run on SDS/PAGE to separate proteins and provide more specific input samples according to the size range of protein bands. The protein bands are then digested by specific proteases (e.g., trypsin) in-gel and analyzed. Alternatively, the organelle can be solubilized as a whole and digested in solution. The digested product is then analyzed. In addition, the organelle fraction can be fed into a workflow of liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) and subsequent data analysis [52]. A number of MS-based proteomics studies were undertaken in recent years to yield a comprehensive list of the mammalian proteome from specific tissues and cell lines [53].

#### 4.1.1 Data acquisition approaches

In data-dependent acquisition (DDA), the eluted peptides from LC are detected by the first MS step, usually within a wide range of a mass-to-charge ratio of 400–1,200  $m/z$ . The most abundant peptides are then entered into the second MS step following fractionation to yield a more informative peak spectrum [54]. However, if the number of precursor ions exceeds the number of precursor selection cycles, peptides detected in repeat analysis become irreproducible [55]. An inherent difficulty in quantitative MS proteomics is that only a few peptides are consistently identified in a complex protein mixture. The uncontrolled complexity of biological samples leads to poor reproducibility of MS-identified peptides. Characteristic

features of prototypic peptides and their physicochemical properties were the basis for developing successful computational tools for predicting peptides for any organism [56]. To improve the detectability of all peptides in a sample, in data-independent acquisition (DIA), the second MS step receives as input, not individual peaks, but instead all peaks within a defined  $m/z$  range, thus using all peptide precursors within a mass range of interest. Usually, most tryptic peptides are within the 400–1200  $m/z$  range, so all the initial peaks within this range are further specified and analyzed [57]. Finally, in targeted proteomics, selected or multiple reaction monitoring (SRM/MRM) is applied, meaning that a set of key peptides from a target list is quantitatively followed in many samples. In this case, the mass spectrometer fragments and analyzes only those peptides, increasing sensitivity and enabling the study of low-copy number proteins [58].

#### *4.1.2 Proteomics analysis platforms*

Various proteomics workflow software and pipelines are available, including Proteome Discoverer, MaxQuant, Mascot, OpenMS, CompOmics, and ProteomicsDB, as described in ref. [59]. These tools have developed drastically over the years, including more and more deep learning and advanced statistics capabilities. For example, MaxQuant is an open-source software package that supports DDA data and has an integrated peptide search engine, Andromeda. From the same developers, Perseus provides statistical tools for high-dimensional omics data analysis covering normalization, pattern recognition, time-series analysis, cross-omics comparisons, and multiple-hypothesis testing. MaxDIA was recently added to the MaxQuant pipeline and provides deep learning-based analysis of DIA data. Similarly, OpenMS is a much-used open-source software framework that enables both analyses of proteomics and metabolomics data. As such, it may be an especially relevant tool for peroxisome research. In addition, Proteome Discoverer includes deep-learning algorithms for the construction of spectral libraries as well as for the improved analysis of low-quality MS/MS spectra.

The retention time of a peptide refers to the time it takes for a peptide to elute from a liquid chromatography (LC) column prior to analysis by MS. As such, a peptide's retention time is determined by the degree to which it interacts with the column, and as such is highly reproducible under the same LC conditions. The accurate prediction of peptide retention time can be used to improve the sensitivity of peptide identification against a peptide database. Early models such as SSRCalc are based on the retention times of 2000 peptides and base predictions on peptide sequence [60, 61]. More advanced, deep learning-based methods can be used for building libraries of MS spectra to enable data analysis from DIA MS. For example, the ProteomicsDB was recently expanded with Prosit, a neural network model that accurately learns and predicts chromatographic retention time and fragment ion intensity of any peptide, both tryptic (i.e., digested by trypsin) and non-tryptic. Other examples, based on different types of neural network algorithms, include DeepRT and, more recently, DeepDIA. The latter approach generates *in silico* spectral libraries, comparable to experimental libraries, to enhance DIA analysis. The various neural network-based models require large training datasets, often in the range of 100,000 peptides or more. However, the pretrained model can then be trained on much smaller experiment-specific datasets that make the model very accurate. The main challenges in this field remain retention time prediction of modified peptides and cross-linked peptides [62].

## 4.2 Proteome-wide characterization of peroxisome proteins

The determination of the peroxisome proteome, as well as the interaction networks between them and the modifications they undergo that affect their function, can be done using a variety of computational tools and approaches. The proteome-wide study of peroxisome proteins is especially interesting to understand regulatory and functional aspects. Early experiments identified nearly all peroxisomal proteins, but it was then still challenging to discriminate between genuine peroxisomal proteins and co-isolated non-peroxisomal proteins, mostly derived from interacting organelles (e.g., mitochondria). The increasing sensitivity of MS enabled the detection of rare peroxisomal proteins, using improved experimental procedures such as organellar profiling [63, 64]. A more recent study of the proteome of HeLa cells provides a proteomic workflow for the generation of reproducible organellar maps. In the study, organellar clusters were assigned proteins based on a machine learning approach [65]. Based on this and other studies, an updated list of the mammalian peroxisomal proteome was made available in ref. [53]. However, many proteins are found in multiple organelles and peroxisome composition may vary between cell types.

An alternative approach to the MS-based proteomics approach described above is the prediction of localization to the peroxisome using peroxisomal targeting signals [66]. Matrix proteins are directed to the peroxisome with a type I or type II peroxisomal target sequence (PTS1 and PTS2) and are transported following the binding of PEX5 or PEX5/PEX7, respectively. In most cases, membrane proteins contain either of two types of PTS (mPTS-I/mPTS-II) and are inserted into the membrane by PEX19 and PEX3. Prediction algorithms for mammalian PTS1 motifs were published almost two decades ago and include PTS1-predictor, PTS1Prowler, PeroxiP, and an algorithm included in the PeroxisomeDB database. PTS1-predictor uses a position-specific scoring matrix (PSSM) or position weight matrix (PWM) approach, derived from aligned sequences of proteins known to harbor a PTS1 sequence, but also peptide sequences of proteins bound to various PEX5 homologs. Thus, the tripeptide dataset used for PTS1 prediction includes a larger amount of variations to find less probable proteins as well [67, 68]. The prediction of PTS2 proteins is much more challenging due to the small number of proteins bearing this signal peptide. A PTS2 prediction tool with a limited success rate is included in PeroxisomeDB. Finally, the In-Pero pipeline, based on a machine learning approach, was recently published for the prediction of sub-peroxisomal cellular localization of unclassified peroxisomal proteins.

The local concentration of functionally interconnected proteins yields PPI, the physical contact between two proteins, which may occur in a binary manner or may exist in multimeric complexes. Several proteome-wide PPI maps are available, including HI-II-14, BioPlex 3.0, and CoFrac. The number of protein pairs in each of these maps is well above 10,000, but the overlap between the maps is very limited [69]. The HI-II-14 map includes only binary interactions, generated from yeast-2-hybrid assays, whereas the others are based on affinity purification or co-fractionation followed by MS. An interesting development is the combination of various databases and depositories, such as the recently published MuSIC 1.0 map that integrates immunofluorescence images from the Human Protein Atlas (HPA) with affinity purification data from BioPlex 2.0 [70].

In addition to the interaction between proteins, PTMs include about 300 types of modifications, including phosphorylation, ubiquitination, glycosylation, acetylation

but also regulated peptide cleavage. The large number of potential PTMs strongly affects protein function. It is expected to increase the proteome's complexity by at least an order of magnitude. PTMs can be identified experimentally in high-throughput MS approaches or in small experiments. However, deep learning also enables accurate prediction of whether a given site can be modified (general site prediction) and if a site can be modified by a specific enzyme (enzyme-specific prediction). Multiple PTMs can be predicted using CapsNet or MusiteDeep. In addition, specific modifications can be accurately predicted with dedicated tools like DeepPhos for phosphorylation, DeepAcet for acetylation, DeepUbiquitylation for ubiquitylation, and more [62].

### **4.3 Metabolomics**

Although not strictly a part of the proteomics field of study, we briefly discuss the understudied metabolomics methods as they are of especial importance in the context of peroxisome metabolism. Metabolomics studies are commonly done using NMR or, more frequently, MS for the analysis of whole-cell or subcellular fractions. As such, much of the metabolomics pipeline is similar to that of the proteomics pipeline—gas chromatography (GC) or liquid chromatography (LC) coupled with MS or tandem MS (MS/MS).

Metabolomics data analysis can be done in an unsupervised or supervised manner. The goal of the unsupervised analysis is the grouping of features (sample, metabolites, and spectral features) according to the measured molecular data, and as such, this approach is suitable when no prior information is available about the system. In contrast, in supervised analysis, a set of features is pre-assigned to a class and is used as a training set for the method of choice to define a classifier that will be used for the classification of an unknown sample [71]. In addition, metabolomics studies can be divided into untargeted and targeted studies. Untargeted, also referred to as discovery-based metabolomics, focuses on global detection and relative quantitation of small molecules in a sample. In contrast, targeted or validation-based metabolomics focuses on measuring well-defined groups of metabolites with opportunities for absolute quantitation [72]. Several metabolomics data analysis tools are available, one of the most popular being MetaboAnalyst. This web-based tool suite covers four analysis categories, including statistical analysis, functional analysis, data integration and systems biology, data processing, and utility functions.

One of the challenges of subcellular or organellar metabolomics is the difficulty to observe metabolic fluxes between compartments. Metabolic flux studies are usually done directly by using isotopically labeled nutrients and measuring isotopically labeled metabolites to infer flux via metabolic flux analysis (MFA) or flux balance analysis (FBA) [73]. In FBA, the flow of metabolites through a metabolic network is analyzed mathematically. Two widely used computational flux inference approaches include isotope tracing coupled with MFA ( $^{13}\text{C}$ -MFA) and constraint-based reconstruction and analysis (COBRA) [74]. Spatial flux analysis was done for the mitochondria and cytosol using  $^{13}\text{C}$ -MFA: isotope tracing in intact cells and subsequent rapid fractionation and metabolism quenching, followed by LC-MS-based metabolomics. Computational deconvolution with metabolic and thermodynamic modeling was used to infer compartment-specific metabolic fluxes [75]. In COBRA, which integrates various experimental and -omics data sources to reconstruct metabolic networks, applying constraints, for example, mass conservation, maximum

reaction rates, and regulation, to construct a space of allowed network states [76]. These approaches could be used to model differences in metabolic flux in healthy and mutated peroxisome factors.

## **5. Human genetic research and peroxisomal disease**

The contribution of peroxisomes to cells and organ physiology has been extensively discussed [77]. It was shown to have an indispensable role in the condition of specific metabolic needs, which explains the importance of peroxisomes in human congenital diseases. However, it becomes apparent that the amounts and properties of this organelle with respect to other organelles (e.g., ER, mitochondria) impact other pathologies of cell homeostasis, such as neurodegeneration, obesity, and more [2, 78].

Over the last decade, our knowledge of human diseases has drastically increased due to the breakthrough in sequencing technologies [79–81]. Projects such as the 1000 Genomes Initiative, ClinVar, OMIM, gnomAD, and others, provide the genetic variation landscape across individuals and populations [82–85]. Databases from such projects and large biobanks (e.g., UK BioBank [86]) are successfully being used for linking genetics with human diseases. For example, the UK BioBank resource gathered genotyping and exome sequencing data of approximately 500,000 people, combining it with clinical and lifestyle information. The unprecedented quality of these resources, merged with computational solutions, data sharing, and standardization advanced the field of human diseases. We briefly introduce computational-based methodologies used for improving the utility of genetic variations in the case of genetic-based peroxisomal diseases.

### **5.1 Advances in human genetics research**

Mendelian diseases are caused by pathological mutations in a single gene with high penetrance. Consequently, the manifestation of the disease is determined by the simple rules of dominant or recessive inheritance. On the other hand, complex diseases result from the presence of many variants where each may carry a small effect. To study such diseases, genome-wide association studies (GWAS) have been used to connect human genetics with complex diseases. The ultimate goal of GWAS is to identify causal connections between genetic variants, traits, and phenotypes. GWAS provides a statistical value to a genotyped variant in the genome to assess the contribution of any specific variants to the studied phenotype. GWAS is useful to suggest association in cases where the sample size, the allele frequency, and the effect size of the association reach statistical significance [87]. Unfortunately, GWAS is prone to confounding factors and biases due to unresolved population structure and linkage disequilibrium (LD) [88, 89]. To overcome the drawback of GWAS due to population size, family-based studies are used as an attractive alternative. In such cases, the genetic variations of family members are determined (i.e., genotyping of parents and their inflicted child).

Another subfield in human genetics that is likely to impact modern medicine is polygenic risk scores (PRS). For PRS, genetics and data from health records are used to present a model for individuals' risk of having the disease of interest [90]. To make a meaningful prediction, the PRS model aggregates an individual's genotype information. Converting PRS to a machine learning prediction model calls for large-scale individual-level data, often using the summary statistics of GWAS results



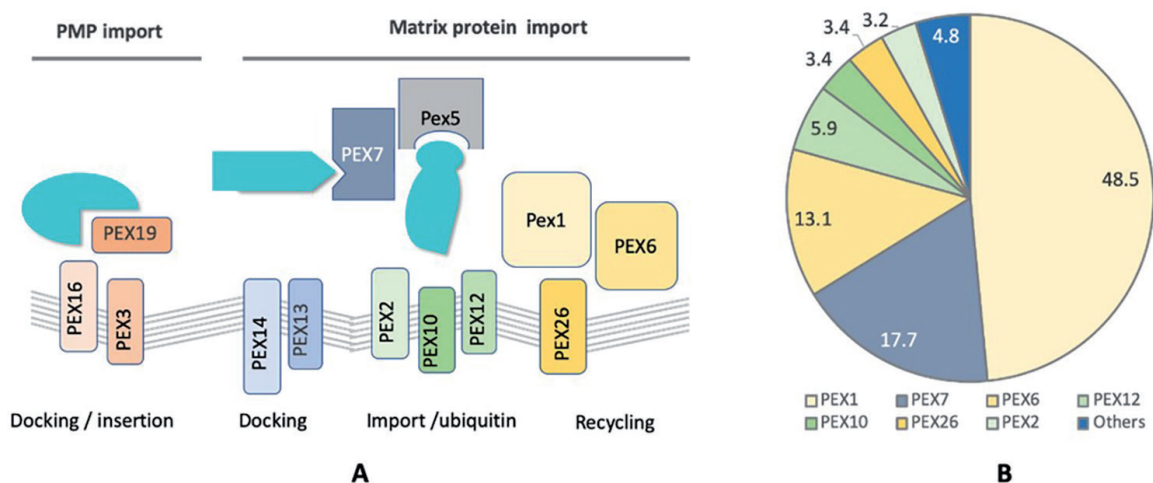
to build such a PRS model [91]. Currently, PRS is not yet typically incorporated into clinical settings and its utility remains questionable [92]. With the improvement in sequencing quality and amounts, rare variants turned out to carry special interest and importance [93, 94]. They occur at a low frequency, still exerting strong phenotypic effects. It is expected that the contribution of rare variants may be substantial in explaining disease heritability. The release of 500,000 whole-exome sequences (WES) by the UK-Biobank allows us to expand our knowledge including the genetic basis of relatively rare occurring diseases [95]. Burden tests in which genes, or defined chromosomal segments, replace individual variants as the statistical relevant unit improve genetic interpretability and utility [96]. For example, under the PWAS (proteome-based association study) methodology, coding genes are associated with a disease by quantifying the effect of genetic variation on the protein function. To assign a valid association between genetic results and the medical condition, PWAS requests that the impact of the genetic variations in the disease and healthy cohorts will be significantly different [97].

Peroxisomal disease does not always fit a trivial definition of a simple Mendelian disease, nor does it match chronic complex disease (e.g., type 2 diabetes). Peroxisomal diseases occur with a wide range of symptoms, interaction with developmental disorders, and severity. Therefore, studying the genetic basis of peroxisomal diseases covers two molecular etiologies—(i) The genetics of PEX proteins with a cellular understanding of organelle biology (e.g., peroxisome formation); (ii) Compilation of an exhaustive list of genetic variants and genes associated with peroxisomes. In this section, we only address the genetic approach associated with these two research goals.

## **5.2 Classes of peroxisomal diseases**

Peroxisomes play a critical role in a variety of metabolic processes, especially in lipid metabolism. All active cells contain many peroxisomes (100 s to 1000s), therefore displaying varying capacity toward metabolic homeostasis, oxidative stress, and lipid metabolism in general. A failure in producing functional peroxisomes is the cause of numerous genetic diseases [98]. Conversely, one or more genetic defects too may result in failure of functional peroxisome production. Peroxisomal disorders are classified into two groups: (i) Specific peroxisomal enzyme deficiencies; (ii) Peroxisome biogenesis disorders (PBDs). PBDs result from a failure in the post-translational import of proteins to the peroxisome's matrix. The underlying genetics was used to better understand peroxisome formation while focusing on the failure to regulate multiple processes in peroxisomal cellular dynamics. In this section, we briefly categorized the peroxisomal diseases that are directly attributed to the failure to produce functional organelle. These diseases are connected to defects in the recognition of newly synthesized proteins, defects in the synthesis of peroxisome membranes, and failure in the insertion into the peroxisomal membrane (**Figure 3**).

PBDs display a spectrum of related diseases that differ in their severity, clinical manifestations, and underlying genetics. PBDs with clear molecular causal genes are collectively called Zellweger spectrum syndromes (ZSS). This set includes Zellweger syndrome (ZS), infantile Refsum disease (IRD), and neonatal adrenoleukodystrophy (NALD). The class of peroxisomal diseases that are caused by enzymatic deficiencies includes, for example, rhizomelic chondrodysplasia punctata (RCDP) type 2. Altogether there are 12 known rare diseases whose protein deficiencies are known, simple biomarkers for definitive diagnosis standards are missing. The task of



**Figure 3.** Molecular basis of peroxisomal diseases. (a) Molecular view of the peroxisomal membrane with the main PEX proteins. PEX5 and PEX7 are receptors for the PTS1 and PTS2 proteins, respectively. PEX19 binds to proteins for insertion in the peroxisome membrane. Light blue indicates cargo proteins. (b) Frequency distribution of mutations in PEX proteins listed in a. the percentage is based on > 1300 patients diagnosed with peroxisome biogenesis disorder (PBDs). Adapted from [99].

diagnosis of the different peroxisomal diseases was empowered by a machine learning algorithm [100].

In most ZSS diseases, very long-chain fatty acids and branched-chain fatty acids accumulate in the plasma of the affected individuals. The accumulation of these lipids impairs multiple organs, leading to a poor prognosis. Among the ZSS diseases, Zellweger syndrome (ZS) accounts for most cases. ZS is a rare autosomal recessive disorder (1:50,000) that in most cases is caused by mutations in PEX1, a member of the AAA-type ATPase family. Mechanistically, the disease is caused by defective protein import due to the mutations in one of 13 major PEX proteins (**Figure 3a**). The partition between patients with isolated deficiencies of metabolic enzymes and those with ZSS can be clinically challenging. Thus, a set of biochemical assays were developed to monitor peroxisome functions and facilitate correct diagnosis (e.g., levels of alanine transaminase or alkaline phosphatase). In PBDs, the reported mutations occur in any of the 13 major PEX proteins at different frequencies (**Figure 3b**). The interpretation of the mutations' effect on the peroxisome function is based on the accumulated knowledge of the role of PEX genes in the biogenesis of the organelle. Altogether, most peroxisomal human diseases are associated with a failure of the PEX proteins to import peroxisomal matrix proteins. In healthy individuals, newly synthesized matrix proteins reach the peroxisome by interacting with PEX5 or PEX7, cytosolic receptors that recognize either a C'-terminal PTS1 or an N'-terminal PTS2. In humans, two alternatively spliced PEX5 isoforms coexist, but only the longer version also binds PEX7. PEX2, PEX10, and PEX12 are zinc-binding RING finger proteins that are responsible for the ubiquitination of PEX5 which is essential for its recycling. The membranous peroxisomal membrane PEX26 combined with PEX1 and PEX6 is involved in the recycling process by releasing ubiquitinated PEX5 from the membrane. The malfunction of PEX7 results in a clinical phenotype of RCDP type 1 disease.

The identification of the genes underlying PBDs initially relied on detailed studies in yeast using classical genetics complementation groups [101]. With the availability of genotyping and WES data, the field of genetic testing evolved, and currently, searching

for pathological mutations in PEX genes is used as a diagnostic service [99]. It is important to note that some of the genetic mutations are not restricted to PEX proteins but are also associated with genes involving the dynamics of peroxisomes and other organelles (e.g., mitochondrial and peroxisomal fission). Remarkably, cells with intact and functional proteins that do not reach the lumen of the peroxisome, result in severe phenotypes, presumably due to the rapid degradation of these misallocated enzymes. This is a general trend among many of the peroxisome matrix proteins.

Mutations in PEX1 are associated with the majority of ZSS cases. PEX1 contains ATP-binding motifs with ATPase activity. All the mutations reported are either defined as loss of function, but others are single missense mutations that affect the ATP binding pocket of the protein interface of PEX1 and PEX6. For a detailed genetic summary of the other PEX proteins, see ref. [99]. In contrast, X-linked adrenoleukodystrophy (ALD), an X-linked disorder, is caused by mutations in the ABCD1 gene that encodes an ABC transporter that act as a channel for the very-long-chain fatty acids entering the peroxisome.

### **5.3 Genomics tools for the study of peroxisome biology**

A valuable resource called OpenTargets integrates multiple large-scale omics data including literature, drugs, and pathways [102]. The goal of OpenTargets is to bridge between molecular targets, drugs, and human diseases. Under the term of peroxisomal diseases in OpenTargets, many of the single enzyme defects that specify alteration in the metabolic enzymes are included. The strength of such a resource is in the ability in exposed overlooked processes that are dependent on intact peroxisome function. For example, among the 35 significant gene targets, all major PEX proteins are included (**Figure 3**). However, the strong support for human diseases that involve PEX11B, indicates that peroxisomal fission is a key process for cell homeostasis as PEX11B acts in recruiting dynamin-related GTPase to the peroxisomal membrane [103].

Based on GWAS results and support from mouse models the list of candidate genes for peroxisomal diseases keeps expanding. Currently, most of the peroxisomal mutated metabolic enzymes are listed, for example, acyl-CoA oxidase 1 (ACOX1), alanine-glyoxylate, serine-pyruvate aminotransferase (AGXT), and many more. Importantly, through an integrative approach that combines medical case studies and model organisms, several candidate genes whose function in peroxisome biology was not established are scored high, suggesting their overlooked roles in peroxisomal function, for example, UniProtKB entries E9PAM4 (Phosphatidylinositol 4-Kinase Type 2) and E9PPB4 (Peroxisomal Biogenesis Factor 19 Isoform 2). In summary, the list of variants associated with peroxisomal diseases was instrumental in shedding light on peroxisome metabolism and dynamics.

## **6. Conclusions**

The last two decades have seen a fast-growing interest in the understanding of peroxisome biology with works on peroxisome biogenesis, import mechanisms, and the characterization of peroxisome proteins. It also emphasizes the existing gap between the more advanced methodologies of protein research concerning the understudied field of lipidomics. In parallel, the power of computational approaches will become pivotal in answering still-open questions regarding regulation, metabolism, and inter-organelle communication at a high level, and the elucidation of the mechanism

and structure of peroxisome proteins and complexes at a more detailed level. We have aimed to provide an overview of some of the most important computational approaches that are likely to serve the research community, both basic and clinical, to expand its research toolbox in the study of peroxisome biology in health and disease.

## **Acknowledgements**

The study was supported by the Israel Science Foundation (ISF) grant (2753/20) and CIRD (3035000323). We thank Dan Ofer for critical reading and Prof. Maya Schuldiner and Dr. Einat Zalckvar (Weizmann Institute) for sharing their insights on peroxisome biology.

## **Conflict of interest**

The authors declare no conflict of interest.

## **Abbreviations**

ER	Endoplasmic reticulum
GWAS	Genome-wide association study
MFA	Metabolic flux analysis
MS	mass spectrometry
PEX	Peroxin
PMP	Peroxisomal membrane protein
PTS	Peroxisomal targeting signal
PPI	protein–protein interactions
PTM	post-translational modification
WES	Whole-exome sequencing
ZSS	Zellweger spectrum syndromes

IntechOpen

IntechOpen

### **Author details**


Naomi van Wijk and Michal Linial\*

Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

\*Address all correspondence to: [michall@cc.huji.ac.il](mailto:michall@cc.huji.ac.il)

### **IntechOpen**

---

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Islinger M, Grille S, Fahimi HD, Schrader M. The peroxisome: An update on mysteries. *Histochemistry and Cell Biology*. 2012;**137**(5):547-574
- [2] Islinger M, Voelkl A, Fahimi HD, Schrader M. The peroxisome: An update on mysteries 2.0. *Histochemistry and Cell Biology*. 2018;**150**(5):443-471
- [3] Lu M, Ward E, van Tartwijk FW, Kaminski CF. Advances in the study of organelle interactions and their role in neurodegenerative diseases enabled by super-resolution microscopy. *Neurobiology of Disease*. 2021;**159**:105475
- [4] Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Valen D. Deep learning for cellular image analysis. *Nature Methods*. 2019;**16**(12):1233-1246
- [5] Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. *Science*. 2017;**356**(6340):eaal3321
- [6] Fransen M, Lismont C, Walton P. The peroxisome-mitochondria connection: How and why? *IJMS*. 2017;**18**(6):1126
- [7] Shai N, Schuldiner M, Zalckvar E. No peroxisome is an island—Peroxisome contact sites. *Biochimica et Biophysica Acta (BBA): Molecular Cell Research*. 2016;**1863**(5):1061-1069
- [8] Nixon-Abell J, Obara CJ, Weigel AV, Li D, Legant WR, Xu CS, et al. Increased spatiotemporal resolution reveals highly dynamic dense tubular matrices in the peripheral ER. *Science*. 2016;**354**(6311):aaf3928
- [9] Li J, Peng J, Jiang X, Rea AC, Peng J, Hu J. DeepLearnMOR: A deep-learning framework for fluorescence image-based classification of organelle morphology. *Plant Physiology*. 2021;**186**(4):1786-1799
- [10] Godinez WJ, Hossain I, Lazic SE, Davies JW, Zhang X. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*. 2017;**33**(13):2010-2019
- [11] Xu Y-Y, Yang F, Shen H-B. Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction. *Bioinformatics*. 2016;**32**(14):2184-2192
- [12] Morone D, Marazza A, Bergmann TJ, Molinari M. Deep learning approach for quantification of organelles and misfolded polypeptide delivery within degradative compartments. *MBoC*. 2020;**31**(14):1512-1524
- [13] Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: Deep learning for cell counting, detection, and morphometry. *Nature Methods*. 2019;**16**(1):67-70
- [14] Yudistira N, Kavitha M, Itabashi T, Iwane AH, Kurita T. Prediction of sequential organelles localization under imbalance using a balanced deep U-Net. *Scientific Reports*. 2020;**10**(1):2626
- [15] Jaqaman K, Loerke D, Mettlen M, Kuwata H, Grinstein S, Schmid SL, et al. Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*. 2008;**5**(8):695-702
- [16] Tinevez J-Y, Perry N, Schindelin J, Hoopes GM, Reynolds GD, Laplantine E, et al. TrackMate: An open and extensible platform for single-particle tracking. *Methods*. 2017;**115**:80-90

- [17] Valm AM, Cohen S, Legant WR, Melunis J, Hershberg U, Wait E, et al. Applying systems-level spectral imaging and analysis to reveal the organelle interactome. *Nature*. 2017;**546**(7656):162-167
- [18] Ounkomol C, Seshamani S, Maleckar MM, Collman F, Johnson GR. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*. 2018;**15**(11):917-920
- [19] Brent R, Boucheron L. Deep learning to predict microscope images. *Nature Methods*. 2018;**15**(11):868-870
- [20] Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*. 2020;**588**(7837):203-204
- [21] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;**596**(7873):583-589
- [22] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. 2022;**50**(D1):D439-D444
- [23] Mariani V, Biasini M, Barbato A, Schwede T. IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013;**29**(21):2722-2728
- [24] Jansen RLM, Santana-Molina C, van den Noort M, Devos DP, van der Klei IJ. Comparative genomics of peroxisome biogenesis proteins: Making sense of the PEX proteins. *Frontiers in Cell and Development Biology*. 2021;**20**(9):654163
- [25] Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*. 2021;**433**(20):167208
- [26] Pan D, Nakatsu T, Kato H. Crystal structure of peroxisomal targeting signal-2 bound to its receptor complex Pex7p-Pex21p. *Nature Structural & Molecular Biology*. 2013;**20**(8):987-993
- [27] Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, et al. Modeling and docking of antibody structures with Rosetta. *Nature Protocols*. 2017;**12**(2):401-416
- [28] Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, et al. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins*. 2001;**5**:119-126
- [29] Bonneau R, Strauss CEM, Rohl CA, Chivian D, Bradley P, Malmström L, et al. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*. 2002;**322**(1):65-78
- [30] Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005;**309**(5742):1868-1871
- [31] Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass membrane protein structure prediction using Rosetta. *Proteins*. 2005;**62**(4):1010-1025
- [32] Barth P, Schonbrun J, Baker D. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;**104**(40):15682-15687
- [33] Barth P, Wallner B, Baker D. Prediction of membrane protein structures with complex topologies

using limited constraints. Proceedings of the National Academy of Sciences of the United States of America. 2009;**106**(5):1409-1414

[34] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;**373**(6557):871-876

[35] Pennisi E. Protein structure prediction now easier, faster. *Science*. 2021;**373**(6552):262-263

[36] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences of the United States of America. 2020;**117**(3):1496-1503

[37] Schneidman-Duhovny D, Wolfson HJ. Modeling of Multimolecular Complexes. *Methods in Molecular Biology*. 2020;**2112**:163-174

[38] Rosell M, Fernández-Recio J. Docking approaches for modeling multi-molecular assemblies. *Current Opinion in Structural Biology*. 2020;**64**:59-65

[39] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. *Bioinformatics*. 2021 [cited 2022 Jan 27]. Available from: <http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034>

[40] Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. Harnessing protein folding neural networks for peptide-protein docking. *Nature Communications*. 2022;**13**(1):176

[41] Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I,

Ovchinnikov S, et al. Computed structures of core eukaryotic protein complexes. *Science*. 2021;**374**(6573):eabm4805

[42] Masrati G, Landau M, Ben-Tal N, Lupas A, Kosloff M, Kosinski J. Integrative structural biology in the era of accurate structure prediction. *Journal of Molecular Biology*. 2021;**433**(20):167127

[43] Flower TG, Hurley JH. Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. *Protein Science*. 2021;**30**(4):728-734

[44] Callaway E. Revolutionary cryo-EM is taking over structural biology. *Nature*. 2020;**578**(7794):201-201

[45] Malhotra S, Träger S, Dal Peraro M, Topf M. Modelling structures in cryo-EM maps. *Current Opinion in Structural Biology*. 2019;**58**:105-114

[46] Kim SJ, Fernandez-Martinez J, Nudelman I, Shi Y, Zhang W, Raveh B, et al. Integrative structure and functional anatomy of a nuclear pore complex. *Nature*. 2018;**555**(7697):475-482

[47] Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*. 2017;**14**(3):290-296

[48] ElGamacy M, Riss M, Zhu H, Truffault V, Coles M. Mapping local conformational landscapes of proteins in solution. *Structure*. 2019;**27**(5):853-865.e5

[49] Piersimoni L, Kastritis PL, Arlt C, Sinz A. Cross-linking mass spectrometry for investigating protein conformations and protein-protein interactions: A method for all seasons. *Chemical Review*. 2021;[acs.chemrev.1c00786](https://doi.org/10.1021/acs.chemrev.1c00786).



- [50] Slavin M, Zamel J, Zohar K, Eliyahu T, Braitbard M, Brielle E, et al. Targeted in situ cross-linking mass spectrometry and integrative modeling reveal the architectures of three proteins from SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America*. 2021;**118**(34):e2103554118
- [51] Yates JR III, Gilchrist A, Howell KE, Bergeron JJM. Proteomics of organelles and large cellular structures. *Nature Reviews. Molecular Cell Biology*. 2005;**6**(9):702-714
- [52] Drissi R, Dubois M-L, Boisvert F-M. Proteomics methods for subcellular proteome analysis. *The FEBS Journal*. 2013;**280**(22):5626-5634
- [53] Yifrach E, Fischer S, Oeljeklaus S, Schuldiner M, Zalckvar E, Warscheid B. Defining the mammalian peroxisomal proteome. In: del Río LA, Schrader M, editors. *Proteomics of Peroxisomes* [Internet]. Singapore: Springer; 2018. p. 47-66
- [54] Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, et al. Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics*. 2011;**11**(4):535-553
- [55] Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*. 2017;**8**(1):291
- [56] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*. 2007;**25**(1):125-131
- [57] Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*. 2012;**11**(6)
- [58] Doerr A. Mass spectrometry-based targeted proteomics. *Nature Methods*. 2013;**10**(1):23-23
- [59] Basic Sciences Division, Universidad de Monterrey, San Pedro Garza García, N.L. Mexico, Gracia KC, Husi H, Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK, Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, UK. *Computational Approaches in Proteomics*. In: Division of Biomedical Science, University of the Highlands and Islands, UK, Husi H, editors. *Computational Biology* [Internet]. Codon Publications; 2019 [cited 2022 Jan 30]. p. 119-42. Available from: <https://exonpublications.com/index.php/exon/article/view/223>
- [60] Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, et al. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC. *Molecular & Cellular Proteomics*. 2004;**3**(9):908-919
- [61] Krokhin OV, Ying S, Cortens JP, Ghosh D, Spicer V, Ens W, et al. Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. *Analytical Chemistry*. 2006;**78**(17):6265-6269
- [62] Wen B, Zeng W, Liao Y, Shi Z, Savage SR, Jiang W, et al. Deep Learning in Proteomics. *Proteomics*. Vol. 202020. p. 1900335

- [63] Wiese S, Gronemeyer T, Ofman R, Kunze M, Grou CP, Almeida JA, et al. Proteomics characterization of mouse kidney peroxisomes by Tandem mass spectrometry and protein correlation profiling. *Molecular & Cellular Proteomics*. 2007;**6**(12):2045-2057
- [64] Andersen JS, Mann M. Organellar proteomics: Turning inventories into insights. *EMBO Reports*. 2006;**7**(9):874-879
- [65] Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife*. 2016;**9**(5):e16950
- [66] Kunze M. Predicting peroxisomal targeting signals to elucidate the peroxisomal proteome of mammals. In: del Río LA, Schrader M, editors. *Proteomics of Peroxisomes* [Internet]. Singapore: Springer. p. 157-199
- [67] Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of Molecular Biology*. 2003;**328**(3):581-592
- [68] Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F. Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *Journal of Molecular Biology*. 2003;**328**(3):567-579
- [69] Luck K, Sheynkman GM, Zhang I, Vidal M. Proteome-scale human interactomics. *Trends in Biochemical Sciences*. 2017;**42**(5):342-354
- [70] Qin Y, Huttlin EL, Winsnes CF, Gosztyla ML, Wacheul L, Kelly MR, et al. A multi-scale map of cell structure fusing protein images and interactions. *Nature*. 2021;**600**(7889):536-542
- [71] Čuperlović-Culf M, Barnett DA, Culf AS, Chute I. Cell culture metabolomics: Applications and future directions. *Drug Discovery Today*. 2010;**15**(15-16):610-621
- [72] Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies—Challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*. 2016;**27**(12):1897-1905
- [73] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nature Biotechnology*. 2010;**28**(3):245-248
- [74] Lagziel S, Lee WD, Shlomi T. Studying metabolic flux adaptations in cancer through integrated experimental-computational approaches. *BMC Biology*. 2019;**17**(1):51
- [75] Lee WD, Mukha D, Aizenshtein E, Shlomi T. Spatial-fluxomics provides a subcellular-compartmentalized view of reductive glutamine metabolism in cancer cells. *Nature Communications*. 2019;**10**(1):1351
- [76] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nature Protocols*. 2007;**2**(3):727-738
- [77] Schrader M, Costello J, Godinho LF, Islinger M. Peroxisome-mitochondria interplay and disease. *Journal of Inherited Metabolic Disease*. 2015;**38**(4):681-702
- [78] Argyriou C, D'Agostino MD, Braverman N. Peroxisome biogenesis disorders. *TRD*. 2016;**1**(2):111-144
- [79] Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. *Cell*. 2019;**177**(1):70-84

- [80] van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. *Trends in Genetics*. 2018;**34**(9):666-681
- [81] Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome sequencing: Current and future perspectives. *G3 Genes|Genomes|Genetics*. 2015;**5**(8):1543-1550
- [82] Auton A, Abecasis GR, Steering Committee, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;**526**(7571):68-74
- [83] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*. 2018;**46**(D1):D1062-D1067
- [84] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*. 2015;**43**:D789-D798
- [85] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;**581**(7809):434-443
- [86] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;**562**(7726):203-209
- [87] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*. 2012;**8**(12):e1002822
- [88] Lee JJ, McGue M, Iacono WG, Chow CC. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic Epidemiology*. 2018;**42**(8):783-795
- [89] Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. Barsh GS, editor. *PLoS Genetics*. 2018;**14**(12):e1007309
- [90] Duncan L, Shen H, Gelaye B, Meijssen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*. 2019;**10**(1):3328
- [91] Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*. 2021;**591**(7849):211-219
- [92] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews: Genetics*. 2018;**19**(9):581-590
- [93] Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden testing of rare variants identified through exome sequencing via publicly available control data. *The American Journal of Human Genetics*. 2018;**103**(4):522-534
- [94] Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nature Communications*. 2020;**11**(1):542
- [95] Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nature Communications*. 2020;**11**(1):6130

- [96] Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*. 2014;**46**(2):200-204
- [97] Brandes N, Linial N, Linial M. PWAS: Proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biology*. 2020;**21**(1):173
- [98] Braverman NE, D’Agostino MD, MacLean GE. Peroxisome biogenesis disorders: Biological, clinical and pathophysiological perspectives: Peroxisome biogenesis disorders. *Developmental Disabilities Research Reviews*. 2013;**17**(3):187-196
- [99] Waterham HR, Ebberink MS. Genetics and molecular basis of human peroxisome biogenesis disorders. *Biochimica et Biophysica Acta (BBA): Molecular Basis of Disease*. 2012;**1822**(9):1430-1441
- [100] Subhashini P, Jaya Krishna S, Usha Rani G, Sushma Chander N, Maheshwar Reddy G, Naushad SM. Application of machine learning algorithms for the differential diagnosis of peroxisomal disorders. *The Journal of Biochemistry*. 2019;**165**(1):67-73
- [101] Fujiki Y, Okumoto K, Mukai S, Honsho M, Tamura S. Peroxisome biogenesis in mammalian cells. *Frontiers in Physiology* [Internet]. 2014;5. [cited 2022 Feb 6]. Available from: <http://journal.frontiersin.org/article/10.3389/fphys.2014.00307/abstract>
- [102] Ghousaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, et al. Open targets genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*. 2021;**49**(D1):D1311-D1320
- [103] Li X, Gould SJ. The dynamin-like GTPase DLP1 is essential for peroxisome division and is recruited to peroxisomes in part by PEX11. *The Journal of Biological Chemistry*. 2003;**278**(19):17012-17020